Abstract Book - CVPR-2016 Main Conference and Workshops

This document contains paper titles, abstracts and links to original papers at the Computer Vision Foundation site of all papers presented in CVPR-2016 main conference and workshops

Abstracts were automatically extracted from PDF files. Automatic extraction failed in some cases (Paper text as embedded image, Non English font / characters, Changes to CVPR template). In such cases authors who wish to, can email me (at moshe@ben-ezra.org) a Latex Abstract and I will fix the documnet.

Hope you find it useful.

–Moshe

1 Main Conference

1.1 Panoramic Stereo Videos with a Single Camera

[Link to Paper]

We present a practical solution for generating 360 stereo panoramic videos using a single camera. Current approaches either use a moving camera that captures multiple images of a scene, which are then stitched together to form the final panorama, or use multiple cameras that are synchronized. A moving camera limits the solution to static scenes, while multi-camera solutions require dedicated calibrated setups. Our approach improves upon the existing solutions in two significant ways: It solves the problem using a single camera, thus minimizing the calibration problem and providing us the ability to convert any digital camera into a panoramic stereo capture device. It captures all the light rays required for stereo panoramas in a single frame using a compact custom designed mirror, thus making the design practical to manufacture and easier to use. We analyze several properties of the design as well as present panoramic stereo and depth estimation results.

1.2 Memory Efficient Max Flow for Multi-label Submodular MRFs

[Link to Paper]

Multi-label submodular Markov Random Fields (MRFs) have been shown to be solvable using max-flow based on an encoding of the labels proposed by Ishikawa, in which each variable Xi is represented by nodes (where is the number of labels) arranged in a column. However, this method in general requires 2 2 edges for each pair of neighbouring variables. This makes it inapplicable to realistic problems with many variables and labels, due to excessive memory requirement. In this paper, we introduce a variant of the max-flow algorithm that requires much less storage. Consequently, our algorithm makes it possible to optimally solve multi-label submodular problems involving large numbers of variables and labels on a standard computer.

1.3 Multi-Cue Zero-Shot Learning with Strong Supervision

[Link to Paper]

Scaling up visual category recognition to large numbers of classes remains challenging. A promising research direction is zero-shot learning, which does not require any training data to recognize new classes, but rather relies on some form of auxiliary information describing the new classes. Ultimately, this may allow to use textbook knowledge that humans employ to learn about new classes by transferring knowledge from classes they know well. The most successful zero-shot learning approaches currently require a particular type

of auxiliary information namely attribute annotations performed by humans that is not readily available for most classes. Our goal is to circumvent this bottleneck by substituting such annotations by extracting multiple pieces of information from multiple unstructured text sources readily available on the web. To compensate for the weaker form of auxiliary information, we incorporate stronger supervision in the form of semantic part annotations on the classes from which we transfer knowledge. We achieve our goal by a joint embedding framework that maps multiple text parts as well as multiple semantic parts into a common space. Our results consistently and significantly improve on the state-of-the-art in zero-short recognition and retrieval.

1.4 Recovering the Missing Link: Predicting Class-Attribute Associations for Unsupervised Zero-Shot Learning

[Link to Paper]

Collecting training images for all visual categories is not only expensive but also impractical. Zero-shot learning (ZSL), especially using attributes, offers a pragmatic solution to this problem. However, at test time most attributebased methods require a full description of attribute associations for each unseen class. Providing these associations is time consuming and often requires domain specific knowledge. In this work, we aim to carry out attribute-based zero-shot classification in an unsupervised manner. We propose an approach to learn relations that couples class embeddings with their corresponding attributes. Given only the name of an unseen class, the learned relationship model is used to automatically predict the classattribute associations. Furthermore, our model facilitates transferring attributes across data sets without additional effort. Integrating knowledge from multiple sources results in a significant additional improvement in performance. We evaluate on two public data sets: Animals with Attributes and aPascal/aYahoo. Our approach outperforms state-ofthe-art methods in both predicting class-attribute associations and unsupervised ZSL by a large margin. hunter ? color stripes grazer spots flippers arctic paws jungle Figure 1: Given only the name of an unseen category, here Leopard, our method automatically predicts the list of attributes (e.g. yellow, spots) associated with the class through relationships (e.g. has color, has pattern). These predicted associations are leveraged to build category classifiers for zero-shot learning.

1.5 Social LSTM: Human Trajectory Prediction in Crowded Spaces

[Link to Paper]

Pedestrians follow different trajectories to avoid obstacles and accommodate fellow pedestrians. Any autonomous vehicle navigating such a scene should be able to foresee the future

positions of pedestrians and accordingly adjust its path to avoid collisions. This problem of trajectory prediction can be viewed as a sequence generation task, where we are interested in predicting the future trajectory of people based on their past positions. Following the recent success of Recurrent Neural Network (RNN) models for sequence prediction tasks, we propose an LSTM model which can learn general human movement and predict their future trajectories. This is in contrast to traditional approaches which use hand-crafted functions such as Social forces. We demonstrate the performance of our method on several public datasets. Our model outperforms state-of-the-art methods on some of these datasets . We also analyze the trajectories predicted by our model to demonstrate the motion behaviour learned by our model.

1.6 Recognizing Emotions from Abstract Paintings using Non-Linear Matrix Completion

[Link to Paper]

Advanced computer vision and machine learning techniques tried to automatically categorize the emotions elicited by abstract paintings with limited success. Since the annotation of the emotional content is highly resourceconsuming, datasets of abstract paintings are either constrained in size or partially annotated. Consequently, it is natural to address the targeted task within a transductive framework. Intuitively, the use of multi-label classification techniques is desirable so to synergically exploit the relations between multiple latent variables, such as emotional content, technique, author, etc. A very popular approach for transductive multi-label recognition under linear classification settings is matrix completion. In this study we introduce non-linear matrix completion (NLMC), thus extending classical linear matrix completion techniques to the non-linear case. Together with the theory grounding the model, we propose an efficient optimization solver. As shown by our extensive experimental validation on two publicly available datasets, NLMC outperforms state-of-the-art methods when recognizing emotions from abstract paintings.

1.7 Unsupervised Learning from Narrated Instruction Videos

[Link to Paper]

We address the problem of automatically learning the main steps to complete a certain task, such as changing a car tire, from a set of narrated instruction videos. The contributions of this paper are three-fold. First, we develop a new unsupervised learning approach that takes advantage of the complementary nature of the input video and the associated narration. The method solves two clustering problems, one in text and one in video, applied one after each other and linked by joint constraints to obtain a single coherent sequence

of steps in both modalities. Second, we collect and annotate a new challenging dataset of real-world instruction videos from the Internet. The dataset contains about 800,000 frames for five different tasks1 that include complex interactions between people and objects, and are captured in a variety of indoor and outdoor settings. Third, we experimentally demonstrate that the proposed method can automatically discover, in an unsupervised manner, the main steps to achieve the task and locate the steps in the input videos.

1.8 Rolling shutter absolute pose problem with known vertical direction

[Link to Paper]

We present a solution to the rolling shutter (RS) absolute camera pose problem with known vertical direction. Our new solver, R5Pup, is an extension of the general minimal solution R6P, which uses a double linearized RS camera model initialized by the standard perspective P3P. Here, thanks to using known vertical directions, we avoid double linearization and can get the camera absolute pose directly from the RS model without the initialization by a standard P3P. Moreover, we need only five 2D-to-3D matches while R6P needed six such matches. We demonstrate in simulated and real experiments that our new R5Pup is robust, fast and a very practical method for absolute camera pose computation for modern cameras on mobile devices. We compare our R5Pup to the state of the art RS and perspective methods and demonstrate that it outperforms them when vertical direction is known in the range of accuracy available on modern mobile devices. We also demonstrate that when using R5Pup solver in structure from motion (SfM) pipelines, it is better to transform already reconstructed scenes into the standard position, rather than using hard constraints on the verticality of up vectors.

1.9 Action Recognition in Video Using Sparse Coding and Relative Features

[Link to Paper]

This work presents an approach to category-based action recognition in video using sparse coding techniques. The proposed approach includes two main contributions: i) A new method to handle intra-class variations by decomposing each video into a reduced set of representative atomic action acts or key-sequences, and ii) A new video descriptor, ITRA: Inter-Temporal Relational Act Descriptor, that exploits the power of comparative reasoning to capture relative similarity relations among key-sequences. In terms of the method to obtain key-sequences, we introduce a loss function that, for each video, leads to the identification of a sparse set of representative key-frames capturing both, relevant particularities arising in the input video, as well as relevant generalities arising in the complete class collection. In terms of the method to obtain the ITRA descriptor, we introduce a novel scheme

to quantify relative intra and inter-class similarities among local temporal patterns arising in the videos. The resulting ITRA descriptor demonstrates to be highly effective to discriminate among action categories. As a result, the proposed approach reaches remarkable action recognition performance on several popular benchmark datasets, outperforming alternative state-ofthe-art techniques by a large margin.

1.10 Learning to Match Aerial Images with Deep Attentive Architectures

[Link to Paper]

Image matching is a fundamental problem in Computer Vision. In the context of feature-based matching, SIFT and its variants have long excelled in a wide array of applications. However, for ultra-wide baselines, as in the case of aerial images captured under large camera rotations, the appearance variation goes beyond the reach of SIFT and RANSAC. In this paper we propose a data-driven, deep learning-based approach that sidesteps local correspondence by framing the problem as a classification task. Furthermore, we demonstrate that local correspondences can still be useful. To do so we incorporate an attention mechanism to produce a set of probable matches, which allows us to further increase performance. We train our models on a dataset of urban aerial imagery consisting of 'same' and 'different' pairs, collected for this purpose, and characterize the problem via a human study with annotations from Amazon Mechanical Turk. We demonstrate that our models outperform the state-of-the-art on ultra-wide baseline matching and approach human accuracy.

1.11 Neural Module Networks

[Link to Paper]

Visual question answering is fundamentally compositional in nature–a question like where is the dog? shares substructure with questions like what color is the dog? and where is the cat? This paper seeks to simultaneously exploit the representational capacity of deep networks and the compositional linguistic structure of questions. We describe a procedure for constructing and learning neural module networks, which compose collections of jointly-trained neural "modules" into deep networks for question answering. Our approach decomposes questions into their linguistic substructures, and uses these structures to dynamically instantiate modular networks (with reusable components for recognizing dogs, classifying colors, etc.). The resulting compound networks are jointly trained. We evaluate our approach on two challenging datasets for visual question answering, achieving state-of-the-art results on both the VQA natural image dataset and a new dataset of complex questions about abstract shapes.

1.12 Learnt Quasi-Transitive Similarity for Retrieval from Large Collections of Faces

[Link to Paper]

We are interested in identity-based retrieval of face sets from large unlabelled collections acquired in uncontrolled environments. Given a baseline algorithm for measuring the similarity of two face sets, the meta-algorithm introduced in this paper seeks to leverage the structure of the data corpus to make the best use of the available baseline. In particular, we show how partial transitivity of inter-personal similarity can be exploited to improve the retrieval of particularly challenging sets which poorly match the query under the baseline measure. We: (i) describe the use of proxy sets as a means of computing the similarity between two sets, (ii) introduce transitivity meta-features based on the similarity of salient modes of appearance variation between sets, (iii) show how quasi-transitivity can be learnt from such features without any labelling or manual intervention, and (iv) demonstrate the effectiveness of the proposed methodology through experiments on the notoriously challenging YouTube database.

1.13 NetVLAD: CNN architecture for weakly supervised place recognition

[Link to Paper]

We tackle the problem of large scale visual place recognition, where the task is to quickly and accurately recognize the location of a given query photograph. We present the following three principal contributions. First, we develop a convolutional neural network (CNN) architecture that is trainable in an end-to-end manner directly for the place recognition task. The main component of this architecture, NetVLAD, is a new generalized VLAD layer, inspired by the "Vector of Locally Aggregated Descriptors" image representation commonly used in image retrieval. The layer is readily pluggable into any CNN architecture and amenable to training via backpropagation. Second, we develop a training procedure, based on a new weakly supervised ranking loss, to learn parameters of the architecture in an end-to-end manner from images depicting the same places over time downloaded from Google Street View Time Machine. Finally, we show that the proposed architecture significantly outperforms non-learnt image representations and off-the-shelf CNN descriptors on two challenging place recognition benchmarks, and improves over current stateof-the-art compact image representations on standard image retrieval benchmarks.

1.14 3D Semantic Parsing of Large-Scale Indoor Spaces

[Link to Paper]

In this paper, we propose a method for semantic parsing the 3D point cloud of an entire building using a hierarchical approach: first, the raw data is parsed into semantically meaningful spaces (e.g. rooms, etc) that are aligned into a canonical reference coordinate system. Second, the spaces are parsed into their structural and building elements (e.g. walls, columns, etc). Performing these with a strong notation of global 3D space is the backbone of our method. The alignment in the first step injects strong 3D priors from the canonical coordinate system into the second step for discovering elements. This allows diverse challenging scenarios as man-made indoor spaces often show recurrent geometric patterns while the appearance features can change drastically. We also argue that identification of structural elements in indoor spaces is essentially a detection problem, rather than segmentation which is commonly used. We evaluated our method on a new dataset of several buildings with a covered area of over 6, 000m2 and over 215 million points, demonstrating robust results readily useful for practical applications.

1.15   Efficient Indexing of Billion-Scale datasets of deep descriptors

[Link to Paper]

Existing billion-scale nearest neighbor search systems have mostly been compared on a single dataset of a billion of SIFT vectors, where systems based on the Inverted Multi-Index (IMI) have been performing very well, achieving state-of-the-art recall in several milliseconds. SIFT-like descriptors, however, are quickly being replaced with descriptors based on deep neural networks (DNN) that provide better performance for many computer vision tasks. In this paper, we introduce a new dataset of one billion descriptors based on DNNs and reveal the relative inefficiency of IMI-based indexing for such descriptors compared to SIFT data. We then introduce two new indexing structures, the Non-Orthogonal Inverted Multi-Index (NO-IMI) and the Generalized Non-Orthogonal Inverted Multi-Index (GNO-IMI). We show that due to additional flexibility, the new structures are able to adapt to DNN descriptor distribution in a better way. In particular, extensive experiments on the new dataset demonstrate that these data structures provide considerably better trade-off between the speed of retrieval and recall, given similar amount of memory, as compared to the standard Inverted Multi-Index.

1.16   Analyzing Classifiers: Fisher Vectors and Deep Neural Networks

[Link to Paper]

Fisher vector (FV) classifiers and Deep Neural Networks (DNNs) are popular and successful algorithms for solving image classification problems. However, both are generally considered 'black box' predictors as the non-linear transformations involved have so far prevented

transparent and interpretable reasoning. Recently, a principled technique, Layer-wise Relevance Propagation (LRP), has been developed in order to better comprehend the inherent structured reasoning of complex nonlinear classification models such as Bag of Feature models or DNNs. In this paper we (1) extend the LRP framework also for Fisher vector classifiers and then use it as analysis tool to (2) quantify the importance of context for classification, (3) qualitatively compare DNNs against FV classifiers in terms of important image regions and (4) detect potential flaws and biases in data. All experiments are performed on the PASCAL VOC 2007 and ILSVRC 2012 data sets.

1.17 Multiview Image Completion with Space Structure Propagation

[Link to Paper]

We present a multiview image completion method that provides geometric consistency among different views by propagating space structures. Since a user specifies the region to be completed in one of multiview photographs casually taken in a scene, the proposed method enables us to complete the set of photographs with geometric consistency by creating or removing structures on the specified region. The proposed method incorporates photographs to estimate dense depth maps. We initially complete color as well as depth from a view, and then facilitate two stages of structure propagation and structure-guided completion. Structure propagation optimizes space topology in the scene across photographs, while structure-guide completion enhances, and completes local image structure of both depth and color in multiple photographs with structural coherence by searching nearest neighbor fields in relevant views. We demonstrate the effectiveness of the proposed method in completing multiview images.

1.18 GIFT: A Real-time and Scalable 3D Shape Search Engine

[Link to Paper]

Projective analysis is an important solution for 3D shape retrieval, since human visual perceptions of 3D shapes rely on various 2D observations from different view points. Although multiple informative and discriminative views are utilized, most projection-based retrieval systems suffer from heavy computational cost, thus cannot satisfy the basic requirement of scalability for search engines. In this paper, we present a real-time 3D shape search engine based on the projective images of 3D shapes. The real-time property of our search engine results from the following aspects: (1) efficient projection and view feature extraction using GPU acceleration; (2) the first inverted file, referred as F-IF, is utilized to speed up the procedure of multi-view matching; (3) the second inverted file (S-IF), which captures

a local distribution of 3D shapes in the feature manifold, is adopted for efficient context-based reranking. As a result, for each query the retrieval task can be finished within one second despite the necessary cost of IO overhead. We name the proposed 3D shape search engine, which combines GPU acceleration and Inverted File (Twice), as GIFT. Besides its high efficiency, GIFT also outperforms the state-of-the-art methods significantly in retrieval accuracy on various shape benchmarks and competitions.

1.19 A nonlinear regression technique for manifold valued data with applications to Medical Image Analysis

[Link to Paper]

Regression is an essential tool in Statistical analysis of data with many applications in Computer Vision, Machine Learning, Medical Imaging and various disciplines of Science and Engineering. Linear and nonlinear regression in a vector space setting has been well studied in literature. However, generalizations to manifold-valued data are only recently gaining popularity. With the exception of a few, most existing methods of regression for manifold valued data are limited to geodesic regression which is a generalization of the linear regression in vector-spaces. In this paper, we present a novel nonlinear kernel-based regression method that is applicable to manifold valued data. Our method is applicable to cases when the independent and dependent variables in the regression model are both manifold-valued or one is manifold-valued and the other is vector or scalar valued. Further, unlike most methods, our method does not require any imposed ordering on the manifold-valued data. The performance of our model is tested on a large number of real data sets acquired from Alzhiemers and movement disorder (Parkinsons and Essential Tremor) patients. We present an extensive set of results along with statistical validation and comparisons.

1.20 Marr Revisited: 2D-3D Alignment via Surface Normal Prediction

[Link to Paper]

We introduce an approach that leverages surface normal predictions, along with appearance cues, to retrieve 3D models for objects depicted in 2D still images from a large CAD object library. Critical to the success of our approach is the ability to recover accurate surface normals for objects in the depicted scene. We introduce a skip-network model built on the pre-trained Oxford VGG convolutional neural network (CNN) for surface normal prediction. Our model achieves state-of-the-art accuracy on the NYUv2 RGB-D dataset for surface normal prediction, and recovers fine object detail compared to previous methods.

Furthermore, we develop a two-stream network over the input image and predicted surface normals that jointly learns pose and style for CAD model retrieval. When using the predicted surface normals, our two-stream network matches prior work using surface normals computed from RGB-D images on the task of pose prediction, and achieves state of the art when using RGB-D input. Finally, our two-stream network allows us to retrieve CAD models that better match the style and pose of a depicted object compared with baseline approaches.

1.21 Principled Parallel Mean-Field Inference for Discrete Random Fields

[Link to Paper]

Mean-field variational inference is one of the most popular approaches to inference in discrete random fields. Standard mean-field optimization is based on coordinate descent and in many situations can be impractical. Thus, in practice, various parallel techniques are used, which either rely on ad hoc smoothing with heuristically set parameters, or put strong constraints on the type of models. In this paper, we propose a novel proximal gradientbased approach to optimizing the variational objective. It is naturally parallelizable and easy to implement. We prove its convergence, and demonstrate that, in practice, it yields faster convergence and often finds better optima than more traditional mean-field optimization techniques. Moreover, our method is less sensitive to the choice of parameters.

1.22 Simultaneous Optical Flow and Intensity Estimation from an Event Camera

[Link to Paper]

Event cameras are bio-inspired vision sensors which mimic retinas to measure per-pixel intensity change rather than outputting an actual intensity image. This proposed paradigm shift away from traditional frame cameras offers significant potential advantages: namely avoiding high data rates, dynamic range limitations and motion blur. Unfortunately, however, established computer vision algorithms may not at all be applied directly to event cameras. Methods proposed so far to reconstruct images, estimate optical flow, track a camera and reconstruct a scene come with severe restrictions on the environment or on the motion of the camera, e.g. allowing only rotation. Here, we propose, to the best of our knowledge, the first algorithm to simultaneously recover the motion field and brightness image, while the camera undergoes a generic motion through any scene. Our approach employs minimisation of a cost function that contains the asynchronous event data as well as spatial and temporal regularisation within a sliding window time interval. Our implementation relies on GPU optimisation and runs in near real-time. In a series of examples,

we demonstrate the successful operation of our framework, including in situations where conventional cameras suffer from dynamic range limitations and motion blur.

1.23 Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks

[Link to Paper]

It is well known that contextual and multi-scale representations are important for accurate visual recognition. In this paper we present the Inside-Outside Net (ION), an object detector that exploits information both inside and outside the region of interest. Contextual information outside the region of interest is integrated using spatial recurrent neural networks. Inside, we use skip pooling to extract information at multiple scales and levels of abstraction. Through extensive experiments we evaluate the design space and provide readers with an overview of what tricks of the trade are important. ION improves state-of-the-art on PASCAL VOC 2012 object detection from 73.9% to 77.9% mAP. On the new and more challenging MS COCO dataset, we improve state-of-the-art from 19.7% to 33.1% mAP. In the 2015 MS COCO Detection Challenge, our ION model won "Best Student Entry" and finished 3rd place overall. As intuition suggests, our detection results provide strong evidence that context and multi-scale representations improve small object detection.

1.24 Camera Calibration from Dynamic Silhouettes Using Motion Barcodes

[Link to Paper]

Computing the epipolar geometry between cameras with very different viewpoints is often problematic as matching points are hard to find. In these cases, it has been proposed to use information from dynamic objects in the scene for suggesting point and line correspondences. We propose a speed up of about two orders of magnitude, as well as an increase in robustness and accuracy, to methods computing epipolar geometry from dynamic silhouettes. This improvement is based on a new temporal signature: motion barcode for lines. Motion barcode is a binary temporal sequence for lines, indicating for each frame the existence of at least one foreground pixel on that line. The motion barcodes of two corresponding epipolar lines are very similar, so the search for corresponding epipolar lines can be limited only to lines having similar barcodes. The use of motion barcodes leads to increased speed, accuracy, and robustness in computing the epipolar geometry.

1.25 Towards Open Set Deep Networks

[Link to Paper]

Deep networks have produced significant gains for various visual recognition problems, leading to high impact academic and commercial applications. Recent work in deep networks highlighted that it is easy to generate images that humans would never classify as a particular object class, yet networks classify such images high confidence as that given class deep network are easily fooled with images humans do not consider meaningful. The closed set nature of deep networks forces them to choose from one of the known classes leading to such artifacts. Recognition in the real world is open set, i.e. the recognition system should reject unknown/unseen classes at test time. We present a methodology to adapt deep networks for open set recognition, by introducing a new model layer, Open-Max, which estimates the probability of an input being from an unknown class. A key element of estimating the unknown probability is adapting Meta-Recognition concepts to the activation patterns in the penultimate layer of the network. OpenMax allows rejection of "fooling" and unrelated open set images presented to the system; OpenMax greatly reduces the number of obvious errors made by a deep network. We prove that the Open-Max concept provides bounded open space risk, thereby formally providing an open set recognition solution. We evaluate the resulting open set deep networks using pre-trained networks from the Caffe Model-zoo on ImageNet 2012 validation data, and thousands of fooling and open set images. The proposed OpenMax model significantly outperforms open set recognition accuracy of basic deep networks as well as deep networks with thresholding of SoftMax probabilities.

1.26 EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild

[Link to Paper]

Research in face perception and emotion theory requires very large annotated databases of images of facial expressions of emotion. Annotations should include Action Units (AUs) and their intensities as well as emotion category. This goal cannot be readily achieved manually. Herein, we present a novel computer vision algorithm to annotate a large database of one million images of facial expressions of emotion in the wild (i.e., face images downloaded from the Internet). First, we show that this newly proposed algorithm can recognize AUs and their intensities reliably across databases. To our knowledge, this is the first published algorithm to achieve highly-accurate results in the recognition of AUs and their intensities across multiple databases. Our algorithm also runs in real-time (¿30 images/second), allowing it to work with large numbers of images and video sequences. Second, we use WordNet to download 1,000,000 images of facial expressions with associated

emotion keywords from the Internet. These images are then automatically annotated with AUs, AU intensities and emotion categories by our algorithm. The result is a highly useful database that can be readily queried using semantic descriptions for applications in computer vision, affective computing, social and cognitive psychology and neuroscience; e.g., "show me all the images with happy faces" or "all images with AU 1 at intensity c."

1.27   Non-Local Image Dehazing

[Link to Paper]

Haze limits visibility and reduces image contrast in outdoor images. The degradation is different for every pixel and depends on the distance of the scene point from the camera. This dependency is expressed in the transmission coefficients, that control the scene attenuation and amount of haze in every pixel. Previous methods solve the single image dehazing problem using various patch-based priors. We, on the other hand, propose an algorithm based on a new, non-local prior. The algorithm relies on the assumption that colors of a haze-free image are well approximated by a few hundred distinct colors, that form tight clusters in RGB space. Our key observation is that pixels in a given cluster are often non-local, i.e., they are spread over the entire image plane and are located at different distances from the camera. In the presence of haze these varying distances translate to different transmission coefficients. Therefore, each color cluster in the clear image becomes a line in RGB space, that we term a haze-line. Using these haze-lines, our algorithm recovers both the distance map and the haze-free image. The algorithm is linear in the size of the image, deterministic and requires no training. It performs well on a wide variety of images and is competitive with other stateof-the-art methods.

1.28   Linear Shape Deformation Models with Local Support using Graph-based Structured Matrix Factorisation

[Link to Paper]

Representing 3D shape deformations by highdimensional linear models has many applications in computer vision and medical imaging. Commonly, using Principal Components Analysis a low-dimensional subspace of the high-dimensional shape space is determined. However, the resulting factors (the most dominant eigenvectors of the covariance matrix) have global support, i.e. changing the coefficient of a single factor deforms the entire shape. Based on matrix factorisation with sparsity and graph-based regularisation terms, we present a method to obtain deformation factors with local support. The benefits include better flexibility and interpretability as well as the possibility of interactively deforming shapes locally. We demonstrate that for brain shapes our method outperforms the state of

the art in local support models with respect to generalisation and sparse reconstruction, whereas for body shapes our method gives more realistic deformations.

1.29   Semantic Segmentation with Boundary Neural Fields

[Link to Paper]

The state-of-the-art in semantic segmentation is currently represented by fully convolutional networks (FCNs). However, FCNs use large receptive fields and many pooling layers, both of which cause blurring and low spatial resolution in the deep layers. As a result FCNs tend to produce segmentations that are poorly localized around object boundaries. Prior work has attempted to address this issue in post-processing steps, for example using a color-based CRF on top of the FCN predictions. However, these approaches require additional parameters and low-level features that are difficult to tune and integrate into the original network architecture. Additionally, most CRFs use colorbased pixel affinities, which are not well suited for semantic segmentation and lead to spatially disjoint predictions. To overcome these problems, we introduce a Boundary Neural Field (BNF), which is a global energy model integrating FCN predictions with boundary cues. The boundary information is used to enhance semantic segment coherence and to improve object localization. Specifically, we first show that the convolutional filters of semantic FCNs provide good features for boundary detection. We then employ the predicted boundaries to define pairwise potentials in our energy. Finally, we show that our energy decomposes semantic segmentation into multiple binary problems, which can be relaxed for efficient global optimization. We report extensive experiments demonstrating that minimization of our global boundary-based energy yields results superior to prior globalization methods, both quantitatively as well as qualitatively.

1.30   Staple: Complementary Learners for Real-Time Tracking

[Link to Paper]

Correlation Filter-based trackers have recently achieved excellent performance, showing great robustness to challenging situations exhibiting motion blur and illumination changes. However, since the model that they learn depends strongly on the spatial layout of the tracked object, they are notoriously sensitive to deformation. Models based on colour statistics have complementary traits: they cope well with variation in shape, but suffer when illumination is not consistent throughout a sequence. Moreover, colour distributions alone can be insufficiently discriminative. In this paper, we show that a simple tracker combining complementary cues in a ridge regression framework can operate faster than

80 FPS and outperform not only all entries in the popular VOT14 competition, but also recent and far more sophisticated trackers according to multiple benchmarks.

1.31 CP-mtML: Coupled Projection multi-task Metric Learning for Large Scale Face Retrieval

[Link to Paper]

We propose a novel Coupled Projection multi-task Metric Learning (CP-mtML) method for large scale face retrieval. In contrast to previous works which were limited to low dimensional features and small datasets, the proposed method scales to large datasets with high dimensional face descriptors. It utilises pairwise (dis-)similarity constraints as supervision and hence does not require exhaustive class annotation for every training image. While, traditionally, multi-task learning methods have been validated on same dataset but different tasks, we work on the more challenging setting with heterogeneous datasets and different tasks. We show empirical validation on multiple face image datasets of different facial traits, e.g. identity, age and expression. We use classic Local Binary Pattern (LBP) descriptors along with the recent Deep Convolutional Neural Network (CNN) features. The experiments clearly demonstrate the scalability and improved performance of the proposed method on the tasks of identity and age based face image retrieval compared to competitive existing methods, on the standard datasets and with the presence of a million distractor face images.

1.32 3D Part-Based Sparse Tracker with Automatic Synchronization and Registration

[Link to Paper]

In this paper, we present a part-based sparse tracker in a particle filter framework where both the motion and appearance model are formulated in 3D. The motion model is adaptive and directed according to a simple yet powerful occlusion handling paradigm, which is intrinsically fused in the motion model. Also, since 3D trackers are sensitive to synchronization and registration noise in the RGB and depth streams, we propose automated methods to solve these two issues. Extensive experiments are conducted on a popular RGBD tracking benchmark, which demonstrate that our tracker can achieve superior results, outperforming many other recent and state-of-the-art RGBD trackers.

1.33 Dynamic Image Networks for Action Recognition

[Link to Paper]

We introduce the concept of dynamic image, a novel compact representation of videos useful for video analysis especially when convolutional neural networks (CNNs) are used. The dynamic image is based on the rank pooling concept and is obtained through the parameters of a ranking machine that encodes the temporal evolution of the frames of the video. Dynamic images are obtained by directly applying rank pooling on the raw image pixels of a video producing a single RGB image per video. This idea is simple but powerful as it enables the use of existing CNN models directly on video data with fine-tuning. We present an efficient and effective approximate rank pooling operator, speeding it up orders of magnitude compared to rank pooling. Our new approximate rank pooling CNN layer allows us to generalize dynamic images to dynamic feature maps and we demonstrate the power of our new representations on standard benchmarks in action recognition achieving state-of-the-art performance.

1.34 Weakly Supervised Deep Detection Networks

[Link to Paper]

Weakly supervised learning of object detection is an important problem in image understanding that still does not have a satisfactory solution. In this paper, we address this problem by exploiting the power of deep convolutional neural networks pre-trained on large-scale image-level classification tasks. We propose a weakly supervised deep detection architecture that modifies one such network to operate at the level of image regions, performing simultaneously region selection and classification. Trained as an image classifier, the architecture implicitly learns object detectors that are better than alternative weakly supervised detection systems on the PASCAL VOC data. The model, which is a simple and elegant end-to-end architecture, outperforms standard data augmentation and fine-tuning techniques for the task of image-level classification as well.

1.35 Large-Scale Semantic 3D Reconstruction: an Adaptive Multi-Resolution Model for Multi-Class Volumetric Labeling

[Link to Paper]

We propose an adaptive multi-resolution formulation of semantic 3D reconstruction. Given a set of images of a scene, semantic 3D reconstruction aims to densely reconstruct both the 3D shape of the scene and a segmentation into semantic object classes. Jointly reasoning

about shape and class allows one to take into account class-specific shape priors (e.g., building walls should be smooth and vertical, and vice versa smooth, vertical surfaces are likely to be building walls), leading to improved reconstruction results. So far, semantic 3D reconstruction methods have been limited to small scenes and low resolution, because of their large memory footprint and computational cost. To scale them up to large scenes, we propose a hierarchical scheme which refines the reconstruction only in regions that are likely to contain a surface, exploiting the fact that both high spatial resolution and high numerical precision are only required in those regions. Our scheme amounts to solving a sequence of convex optimizations while progressively removing constraints, in such a way that the energy, in each iteration, is the tightest possible approximation of the underlying energy at full resolution. In our experiments the method saves up to 98% memory and 95% computation time, without any loss of accuracy.

1.36 A Robust Multilinear Model Learning Framework for 3D Faces

[Link to Paper]

Multilinear models are widely used to represent the statistical variations of 3D human faces as they decouple shape changes due to identity and expression. Existing methods to learn a multilinear face model degrade if not every person is captured in every expression, if face scans are noisy or partially occluded, if expressions are erroneously labeled, or if the vertex correspondence is inaccurate. These limitations impose requirements on the training data that disqualify large amounts of available 3D face data from being usable to learn a multilinear model. To overcome this, we introduce the first framework to robustly learn a multilinear model from 3D face databases with missing data, corrupt data, wrong semantic correspondence, and inaccurate vertex correspondence. To achieve this robustness to erroneous training data, our framework jointly learns a multilinear model and fixes the data. We evaluate our framework on two publicly available 3D face databases, and show that our framework achieves a data completion accuracy that is comparable to state-of-the-art tensor completion methods. Our method reconstructs corrupt data more accurately than state-of-the-art methods, and improves the quality of the learned model significantly for erroneously labeled expressions.

1.37 A 3D Morphable Model learnt from 10,000 faces

[Link to Paper]

We present Large Scale Facial Model (LSFM) – a 3D Morphable Model (3DMM) automatically constructed from 9,663 distinct facial identities. To the best of our knowledge LSFM is the largest-scale Morphable Model ever constructed, containing statistical information

19

from a huge variety of the human population. To build such a large model we introduce a novel fully automated and robust Morphable Model construction pipeline. The dataset that LSFM is trained on includes rich demographic information about each subject, allowing for the construction of not only a global 3DMM but also models tailored for specific age, gender or ethnicity groups. As an application example, we utilise the proposed model to perform age classification from 3D shape alone. Furthermore, we perform a systematic analysis of the constructed 3DMMs that showcases their quality and descriptive power. The presented extensive qualitative and quantitative evaluations reveal that the proposed 3DMM achieves state-of-the-art results, outperforming existing models by a large margin. Finally, for the benefit of the research community, we make publicly available the source code of the proposed automatic 3DMM construction pipeline. In addition, the constructed global 3DMM and a variety of bespoke models tailored by age, gender and ethnicity are available on application to researchers involved in medically oriented research.

1.38 iLab-20M: A large-scale controlled object dataset to investigate deep learning

[Link to Paper]

Tolerance to image variations (e.g., translation, scale, pose, illumination, background) is an important desired property of any object recognition system, be it human or machine. Moving towards increasingly bigger datasets has been trending in computer vision especially with the emergence of highly popular deep learning models. While being very useful for learning invariance to object inter- and intra-class shape variability, these large-scale wild datasets are not very useful for learning invariance to other parameters urging researchers to resort to other tricks for training models. In this work, we introduce a large-scale synthetic dataset, which is freely and publicly available, and use it to answer several fundamental questions regarding selectivity and invariance properties of convolutional neural networks. Our dataset contains two parts: a) objects shot on a turntable: 15 categories, 8 rotation angles, 11 cameras on a semi-circular arch, 5 lighting conditions, 3 focus levels, variety of backgrounds (23.4 per instance) generating 1320 images per instance (about 22 million images in total), and b) scenes: in which a robotic arm takes pictures of objects on a 1:160 scale scene. We study: 1) invariance and selectivity of different CNN layers, 2) knowledge transfer from one object category to another, 3) systematic or random sampling of images to build a train set, 4) domain adaptation from synthetic to natural scenes, and 5) order of knowledge delivery to CNNs. We also discuss how our analyses can lead the field to develop more efficient deep learning methods.

1.39 Computational Imaging for VLBI Image Reconstruction

[Link to Paper]

Very long baseline interferometry (VLBI) is a technique for imaging celestial radio emissions by simultaneously observing a source from telescopes distributed across Earth. The challenges in reconstructing images from fine angular resolution VLBI data are immense. The data is extremely sparse and noisy, thus requiring statistical image models such as those designed in the computer vision community. In this paper we present a novel Bayesian approach for VLBI image reconstruction. While other methods often require careful tuning and parameter selection for different types of data, our method (CHIRP) produces good results under different settings such as low SNR or extended emission. The success of our method is demonstrated on realistic synthetic experiments as well as publicly available real data. We present this problem in a way that is accessible to members of the community, and provide a dataset website (vlbiimaging.csail.mit.edu) that facilitates controlled comparisons across algorithms.

1.40 Uncertainty-Driven 6D Pose Estimation of Objects and Scenes from a Single RGB Image

[Link to Paper]

In recent years, the task of estimating the 6D pose of object instances and complete scenes, i.e. camera localization, from a single input image has received considerable attention. Consumer RGB-D cameras have made this feasible, even for difficult, texture-less objects and scenes. In this work, we show that a single RGB image is sufficient to achieve visually convincing results. Our key concept is to model and exploit the uncertainty of the system at all stages of the processing pipeline. The uncertainty comes in the form of continuous distributions over 3D object coordinates and discrete distributions over object labels. We give three technical contributions. Firstly, we develop a regularized, auto-context regression framework which iteratively reduces uncertainty in object coordinate and object label predictions. Secondly, we introduce an efficient way to marginalize object coordinate distributions over depth. This is necessary to deal with missing depth information. Thirdly, we utilize the distributions over object labels to detect multiple objects simultaneously with a fixed budget of RANSAC hypotheses. We tested our system for object pose estimation and camera localization on commonly used data sets. We see a major improvement over competing systems.

1.41 A Deeper Look at Saliency: Feature Contrast, Semantics, and Beyond

[Link to Paper]

In this paper we consider the problem of visual saliency modeling, including both human gaze prediction and salient object segmentation. The overarching goal of the paper is to identify high level considerations relevant to deriving more sophisticated visual saliency models. A deep learning model based on fully convolutional networks (FCNs) is presented, which shows very favorable performance across a wide variety of benchmarks relative to existing proposals. We also demonstrate that the manner in which training data is selected, and ground truth treated is critical to resulting model behaviour. Recent efforts have explored the relationship between human gaze and salient objects, and we also examine this point further in the context of FCNs. Close examination of the proposed and alternative models serves as a vehicle for identifying problems important to developing more comprehensive models going forward.

1.42 Online Learning with Bayesian Classification Trees

[Link to Paper]

Randomized classification trees are among the most popular machine learning tools and found successful applications in many areas. Although this classifier was originally designed as offline learning algorithm, there has been an increased interest in the last years to provide an online variant. In this paper, we propose an online learning algorithm for classification trees that adheres to Bayesian principles. In contrast to state-of-the-art approaches that produce large forests with complex trees, we aim at constructing small ensembles consisting of shallow trees with high generalization capabilities. Experiments on benchmark machine learning and body part recognition datasets show superior performance over state-of-the-art approaches.

1.43 Multicamera calibration from visible and mirrored epipoles

[Link to Paper]

Multicamera rigs are used in a large number of 3D Vision applications, such as 3D modeling, motion capture or telepresence and a robust calibration is of utmost importance in order to achieve a high accuracy results. In many practical configurations the cameras in a rig are arranged in such a way, that they can observe each other, in other words a number of epipoles correspond to the real image points. In this paper we propose a solution for the automatic recovery of the external calibration of a multicamera system by enforcing only

simple geometrical constraints, arising from the epipole visibility, without using any calibration object, such as checkerboards, laser pointers or similar. Additionally, we introduce an extension of the method that handles the case of epipoles being visible in the reflection of a planar mirror, which makes the algorithm suitable for the calibration of any multicamera system, irrespective of the number of cameras and their actual mutual visibility, and furthermore we remark that it requires only one or a few images per camera and therefore features a high speed and usability. We produce an evidence of the algorithm effectiveness by presenting a wide set of tests performed on synthetic as well as real datasets and we compare the results with those obtained using a traditional LED-based algorithm. The real datasets have been captured using a multicamera Virtual Reality (VR) rig and a spherical dome configuration for 3D reconstruction.

1.44 Minimizing the Maximal Rank

[Link to Paper]

In computer vision, many problems can be formulated as finding a low rank approximation of a given matrix. Ideally, if all elements of the measurement matrix are available, this is easily solved in the L2-norm using factorization. However, in practice this is rarely the case. Lately, this problem has been addressed using different approaches, one is to replace the rank term by the convex nuclear norm, another is to derive the convex envelope of the rank term plus a data term. In the latter case, matrices are divided into sub-matrices and the envelope is computed for each subblock individually. In this paper a new convex envelope is derived which takes all sub-matrices into account simultaneously. This leads to a simpler formulation, using only one parameter to control the trade-of between rank and data fit, for applications where one seeks low rank approximations of multiple matrices with the same rank. We show in this paper how our general framework can be used for manifold denoising of several images at once, as well as just denoising one image. Experimental comparisons show that our method achieves results similar to state-of-the-art approaches while being applicable for other problems such as linear shape model estimation.

1.45 A Probabilistic Collaborative Representation based Approach for Pattern Classification

[Link to Paper]

Conventional representation based classifiers, ranging from the classical nearest neighbor classifier and nearest subspace classifier to the recently developed sparse representation based classifier (SRC) and collaborative representation based classifier (CRC), are essentially distance based classifiers. Though SRC and CRC have shown interesting classification

23

results, their intrinsic classification mechanism remains unclear. In this paper we propose a probabilistic collaborative representation framework, where the probability that a test sample belongs to the collaborative subspace of all classes can be well defined and computed. Consequently, we present a probabilistic collaborative representation based classifier (ProCRC), which jointly maximizes the likelihood that a test sample belongs to each of the multiple classes. The final classification is performed by checking which class has the maximum likelihood. The proposed ProCRC has a clear probabilistic interpretation, and it shows superior performance to many popular classifiers, including SRC, CRC and SVM. Coupled with the CNN features, it also leads to state-of-the-art classification results on a variety of challenging visual datasets.

1.46 Detection and Accurate Localization of Circular Fiducials under Highly Challenging Conditions

[Link to Paper]

Using fiducial markers ensures reliable detection and identification of planar features in images. Fiducials are used in a wide range of applications, especially when a reliable visual reference is needed, e.g., to track the camera in cluttered or textureless environments. A marker designed for such applications must be robust to partial occlusions, varying distances and angles of view, and fast camera motions. In this paper, we present a robust, highly accurate fiducial system, whose markers consist of concentric rings, along with its theoretical foundations. Relying on projective properties, it allows to robustly localize the imaged marker and to accurately detect the position of the image of the (common) circle center. We demonstrate that our system can detect and accurately localize these circular fiducials under very challenging conditions and the experimental results reveal that it outperforms other recent fiducial systems.

1.47 GOGMA: Globally-Optimal Gaussian Mixture Alignment

[Link to Paper]

Gaussian mixture alignment is a family of approaches that are frequently used for robustly solving the point-set registration problem. However, since they use local optimisation, they are susceptible to local minima and can only guarantee local optimality. Consequently, their accuracy is strongly dependent on the quality of the initialisation. This paper presents the first globally-optimal solution to the 3D rigid Gaussian mixture alignment problem under the L2 distance between mixtures. The algorithm, named GOGMA, employs a branch-and-bound approach to search the space of 3D rigid motions SE(3), guaranteeing global optimality regardless of the initialisation. The geometry of SE(3) was used to find novel

24

upper and lower bounds for the objective function and local optimisation was integrated into the scheme to accelerate convergence without voiding the optimality guarantee. The evaluation empirically supported the optimality proof and showed that the method performed much more robustly on two challenging datasets than an existing globally-optimal registration solution.

1.48 Large Scale Hard Sample Mining with Monte Carlo Tree Search

[Link to Paper]

We investigate an efficient strategy to collect false positives from very large training sets in the context of object detection. Our approach scales up the standard bootstrapping procedure by using a hierarchical decomposition of an image collection which reflects the statistical regularity of the detector's responses. Based on that decomposition, our procedure uses a Monte Carlo Tree Search to prioritize the sampling toward sub-families of images which have been observed to be rich in false positives, while maintaining a fraction of the sampling toward unexplored sub-families of images. The resulting procedure increases substantially the proportion of false positive samples among the visited ones compared to a naive uniform sampling. We apply experimentally this new procedure to face detection with a collection of 100,000 background images and to pedestrian detection with 32,000 images. We show that for two standard detectors, the proposed strategy cuts the number of images to visit by half to obtain the same amount of false positives and the same final performance.

1.49 Pedestrian Detection Inspired by Appearance Constancy and Shape Symmetry

[Link to Paper]

The discrimination and simplicity of features are very important for effective and efficient pedestrian detection. However, most state-of-the-art methods are unable to achieve good tradeoff between accuracy and efficiency. Inspired by some simple inherent attributes of pedestrians (i.e., appearance constancy and shape symmetry), we propose two new types of non-neighboring features (NNF): side-inner difference features (SIDF) and symmetrical similarity features (SSF). SIDF can characterize the difference between the background and pedestrian and the difference between the pedestrian contour and its inner part. SSF can capture the symmetrical similarity of pedestrian shape. However, it's difficult for neighboring features to have such above characterization abilities. Finally, we propose to combine both non-neighboring and neighboring features for pedestrian detection. It's found that

nonneighboring features can further decrease the average miss rate by 4.44%. Experimental results on INRIA and Caltech pedestrian datasets demonstrate the effectiveness and efficiency of the proposed method. Compared to the state-ofthe-art methods without using CNN, our method achieves the best detection performance on Caltech, outperforming the second best method (i.e., Checkerboards) by 1.63%.

1.50 Human Pose Estimation with Iterative Error Feedback

[Link to Paper]

Hierarchical feature extractors such as Convolutional Networks (ConvNets) have achieved impressive performance on a variety of classification tasks using purely feedforward processing. Feedforward architectures can learn rich representations of the input space but do not explicitly model dependencies in the output spaces, that are quite structured for tasks such as articulated human pose estimation or object segmentation. Here we propose a framework that expands the expressive power of hierarchical feature extractors to encompass both input and output spaces, by introducing top-down feedback. Instead of directly predicting the outputs in one go, we use a self-correcting model that progressively changes an initial solution by feeding back error predictions, in a process we call Iterative Error Feedback (IEF). IEF shows excellent performance on the task of articulated pose estimation in the challenging MPII and LSP benchmarks, matching the state-of-the-art without requiring ground truth scale annotation.

1.51 Learning Aligned Cross-Modal Representations from Weakly Aligned Data

[Link to Paper]

People can recognize scenes across many different modalities beyond natural images. In this paper, we investigate how to learn cross-modal scene representations that transfer across modalities. To study this problem, we introduce a new cross-modal scene dataset. While convolutional neural networks can categorize cross-modal scenes well, they also learn an intermediate representation not aligned across modalities, which is undesirable for crossmodal transfer applications. We present methods to regularize cross-modal convolutional neural networks so that they have a shared representation that is agnostic of the modality. Our experiments suggest that our scene representation can help transfer representations across modalities for retrieval. Moreover, our visualizations suggest that units emerge in the shared representation that tend to activate on consistent concepts independently of the modality. denotes equal contribution

1.52 An efficient Exact-PGA algorithm for constant curvature manifolds

[Link to Paper]

Manifold-valued datasets are widely encountered in many computer vision tasks. A non-linear analog of the PCA algorithm, called the Principal Geodesic Analysis (PGA) algorithm suited for data lying on Riemannian manifolds was reported in literature a decade ago. Since the objective function in the PGA algorithm is highly non-linear and hard to solve efficiently in general, researchers have proposed a linear approximation. Though this linear approximation is easy to compute, it lacks accuracy especially when the data exhibits a large variance. Recently, an alternative called the exact PGA was proposed which tries to solve the optimization without any linearization. For general Riemannian manifolds, though it yields a better accuracy than the original (linearized) PGA, for data that exhibit large variance, the optimization is not computationally efficient. In this paper, we propose an efficient exact PGA algorithm for constant curvature Riemannian manifolds (CCM-EPGA). The CCM-EPGA algorithm differs significantly from existing PGA algorithms in two aspects, (i) the distance between a given manifold-valued data point and the principal submanifold is computed analytically and thus no optimization is required as in the existing methods. (ii) Unlike the existing PGA algorithms, the descent into codimension-1 submanifolds does not require any optimization but is accomplished through the use of the Rimeannian inverse Exponential map and the parallel transport operations. We present theoretical and experimental results for constant curvature Riemannian manifolds depicting favorable performance of the CCM-EPGA algorithm compared to existing PGA algorithms. We also present data reconstruction from the principal components which has not been reported in literature in this setting.

1.53 BORDER: An Oriented Rectangles Approach to Texture-less Object Recognition

[Link to Paper]

This paper presents an algorithm coined BORDER (Bounding Oriented-Rectangle Descriptors for Enclosed Regions) for texture-less object recognition. By fusing a regional object encompassment concept with descriptor-based pipelines, we extend local-patches into scalable object-sized oriented rectangles for optimal object information encapsulation with minimal outliers. We correspondingly introduce a modified line-segment detection technique termed Linelets to stabilize keypoint repeatability in homogenous conditions. In addition, a unique sampling technique facilitates the incorporation of robust angle primitives to produce discriminative rotation-invariant descriptors. BORDER's high competence in object recognition particularly excels in homogenous conditions obtaining superior detection rates in the presence of high-clutter, occlusion and scale-rotation changes when compared

with modern state-of-the-art texture-less object detectors such as BOLD and LINE2D on public texture-less object databases.

1.54 We Are Humor Beings: Understanding and Predicting Visual Humor

[Link to Paper]

Humor is an integral part of human lives. Despite being tremendously impactful, it is perhaps surprising that we do not have a detailed understanding of humor yet. As interactions between humans and AI systems increase, it is imperative that these systems are taught to understand subtleties of human expressions such as humor. In this work, we are interested in the question what content in a scene causes it to be funny? As a first step towards understanding visual humor, we analyze the humor manifested in abstract scenes and design computational models for them. We collect two datasets of abstract scenes that facilitate the study of humor at both the scene-level and the object-level. We analyze the funny scenes and explore the different types of humor depicted in them via human studies. We model two tasks that we believe demonstrate an understanding of some aspects of visual humor. The tasks involve predicting the funniness of a scene and altering the funniness of a scene. We show that our models perform well quantitatively, and qualitatively through human studies. Our datasets are publicly available.

1.55 Kinematic Structure Correspondences via Hypergraph Matching

[Link to Paper]

In this paper, we present a novel framework for finding the kinematic structure correspondence between two objects in videos via hypergraph matching. In contrast to prior appearance and graph alignment based matching methods which have been applied among two similar static images, the proposed method finds correspondences between two dynamic kinematic structures of heterogeneous objects in videos. Our main contributions can be summarised as follows: (i) casting the kinematic structure correspondence problem into a hypergraph matching problem, incorporating multi-order similarities with normalising weights, (ii) a structural topology similarity measure by a new topology constrained subgraph isomorphism aggregation, (iii) a kinematic correlation measure between pairwise nodes, and (iv) a combinatorial local motion similarity measure using geodesic distance on the Riemannian manifold. We demonstrate the robustness and accuracy of our method through a number of experiments on complex articulated synthetic and real data.

1.56   Random Features for Sparse Signal Classification

[Link to Paper]

Random features is an approach for kernel-based inference on large datasets. In this paper, we derive performance guarantees for random features on signals, like images, that enjoy sparse representations and show that the number of random features required to achieve a desired approximation of the kernel similarity matrix can be significantly smaller for sparse signals. Based on this, we propose a scheme termed compressive random features that first obtains low-dimensional projections of a dataset and, subsequently, derives random features on the low-dimensional projections. This scheme provides significant improvements in signal dimensionality, computational time, and storage costs over traditional random features while enjoying similar theoretical guarantees for achieving inference performance. We support our claims by providing empirical results across many datasets.

1.57   They Are Not Equally Reliable: Semantic Event Search using Differentiated Concept Classifiers

[Link to Paper]

Complex event detection on unconstrained Internet videos has seen much progress in recent years. However, state-of-the-art performance degrades dramatically when the number of positive training exemplars falls short. Since label acquisition is costly, laborious, and time-consuming, there is a real need to consider the much more challenging semantic event search problem, where no example video is given. In this paper, we present a state-of-the-art event search system without any example videos. Relying on the key observation that events (e.g. dog show) are usually compositions of multiple mid-level concepts (e.g. "dog," "theater," and "dog jumping"), we first train a skip-gram model to measure the relevance of each concept with the event of interest. The relevant concept classifiers then cast votes on the test videos but their reliability, due to lack of labeled training videos, has been largely unaddressed. We propose to combine the concept classifiers based on a principled estimate of their accuracy on the unlabeled test videos. A novel warping technique is proposed to improve the performance and an efficient highly-scalable algorithm is provided to quickly solve the resulting optimization. We conduct extensive experiments on the latest TRECVID MEDTest 2014, MEDTest 2013 and CCV datasets, and achieve state-of-the-art performances.

1.58 Variable Aperture Light Field Photography: Overcoming the Diffraction-limited Spatio-angular Resolution Tradeoff

[Link to Paper]

Light fields have many applications in machine vision, consumer photography, robotics, and microscopy. However, the prevalent resolution limits of existing light field imaging systems hinder widespread adoption. In this paper, we analyze fundamental resolution limits of light field cameras in the diffraction limit. We propose a sequential, coded-aperture-style acquisition scheme that optimizes the resolution of a light field reconstructed from multiple photographs captured from different perspectives and fnumber settings. We also show that the proposed acquisition scheme facilitates high dynamic range light field imaging and demonstrate a proof-of-concept prototype system. With this work, we hope to advance our understanding of the resolution limits of light field photography and develop practical computational imaging systems to overcome them.

1.59 Synthesized Classifiers for Zero-Shot Learning

[Link to Paper]

Given semantic descriptions of object classes, zeroshot learning aims to accurately recognize objects of the unseen classes, from which no examples are available at the training stage, by associating them to the seen classes, from which labeled examples are provided. We propose to tackle this problem from the perspective of manifold learning. Our main idea is to align the semantic space that is derived from external information to the model space that concerns itself with recognizing visual features. To this end, we introduce a set of "phantom" object classes whose coordinates live in both the semantic space and the model space. Serving as bases in a dictionary, they can be optimized from labeled data such that the synthesized real object classifiers achieve optimal discriminative performance. We demonstrate superior accuracy of our approach over the state of the art on four benchmark datasets for zero-shot learning, including the full ImageNet Fall 2011 dataset with more than 20,000 unseen classes.

1.60 Personalizing Human Video Pose Estimation

[Link to Paper]

We propose a personalized ConvNet pose estimator that automatically adapts itself to the uniqueness of a person's appearance to improve pose estimation in long videos. We make

the following contributions: (i) we show that given a few high-precision pose annotations, e.g. from a generic ConvNet pose estimator, additional annotations can be generated throughout the video using a combination of image-based matching for temporally distant frames, and dense optical flow for temporally local frames; (ii) we develop an occlusion aware self-evaluation model that is able to automatically select the high-quality and reject the erroneous additional annotations; and (iii) we demonstrate that these high-quality annotations can be used to fine-tune a ConvNet pose estimator and thereby personalize it to lock on to key discriminative features of the person's appearance. The outcome is a substantial improvement in the pose estimates for the target video using the personalized ConvNet compared to the original generic ConvNet. Our method outperforms the state of the art (including top ConvNet methods) by a large margin on three standard benchmarks, as well as on a new challenging YouTube video dataset. Furthermore, we show that training from the automatically generated annotations can be used to improve the performance of a generic ConvNet on other benchmarks.

1.61   Object-Proposal Evaluation Protocol is 'Gameable'

[Link to Paper]

Object proposals have quickly become the de-facto preprocessing step in a number of vision pipelines (for object detection, object discovery, and other tasks). Their performance is usually evaluated on partially annotated datasets. In this paper, we argue that the choice of using a partially annotated dataset for evaluation of object proposals is problematic as we demonstrate via a thought experiment, the evaluation protocol is 'gameable', in the sense that progress under this protocol does not necessarily correspond to a "better" category independent object proposal algorithm. To alleviate this problem, we: (1) Introduce a nearly-fully annotated version of PASCAL VOC dataset, which serves as a test-bed to check if object proposal techniques are overfitting to a particular list of categories. (2) Perform an exhaustive evaluation of object proposal methods on our introduced nearly-fully annotated PASCAL dataset and perform cross-dataset generalization experiments; and (3) Introduce a diagnostic experiment to detect the bias capacity in an object proposal algorithm. This tool circumvents the need to collect a densely annotated dataset, which can be expensive and cumbersome to collect. Finally, we have released an easy-to-use toolbox which combines various publicly available implementations of object proposal algorithms which standardizes the proposal generation and evaluation so that new methods can be added and evaluated on different datasets. We hope that the results presented in the paper will motivate the community to test the category independence of various object proposal methods by carefully choosing the evaluation protocol.

1.62 ASP Vision: Optically Computing the First Layer of Convolutional Neural Networks using Angle Sensitive Pixels

[Link to Paper]

Deep learning using convolutional neural networks (CNNs) is quickly becoming the state-of-the-art for challenging computer vision applications. However, deep learning's power consumption and bandwidth requirements currently limit its application in embedded and mobile systems with tight energy budgets. In this paper, we explore the energy savings of optically computing the first layer of CNNs. To do so, we utilize bio-inspired Angle Sensitive Pixels (ASPs), custom CMOS diffractive image sensors which act similar to Gabor filter banks in the V1 layer of the human visual cortex. ASPs replace both image sensing and the first layer of a conventional CNN by directly performing optical edge filtering, saving sensing energy, data bandwidth, and CNN FLOPS to compute. Our experimental results (both on synthetic data and a hardware prototype) for a variety of vision tasks such as digit recognition, object recognition, and face identification demonstrate up to 90% reduction in image sensor power consumption and 90% reduction in data bandwidth from sensor to CPU, while achieving similar performance compared to traditional deep learning pipelines.

1.63 A New Finsler Minimal Path Model with Curvature Penalization for Image Segmentation and Closed Contour Detection

[Link to Paper]

In this paper, we propose a new curvature penalized minimal path model for image segmentation via closed contour detection based on the weighted Euler elastica curves, firstly introduced to the field of computer vision in [22]. Our image segmentation method extracts a collection of curvature penalized minimal geodesics, concatenated to form a closed contour, by connecting a set of user-specified points. Globally optimal minimal paths can be computed by solving an Eikonal equation. This first order PDE is traditionally regarded as unable to penalize curvature, which is related to the path acceleration in active contour models. We introduce here a new approach that enables finding a global minimum of the geodesic energy including a curvature term. We achieve this through the use of a novel Finsler metric adding to the image domain the orientation as an extra space dimension. This metric is non-Riemannian and asymmetric, defined on an orientation lifted space, incorporating the curvature penalty in the geodesic energy. Experiments show that the proposed Finsler minimal path model indeed outperforms state-of-the-art minimal path models in both synthetic and real images.

1.64 Attention to Scale: Scale-aware Semantic Image Segmentation

[Link to Paper]

Incorporating multi-scale features in fully convolutional neural networks (FCNs) has been a key element to achieving state-of-the-art performance on semantic image segmentation. One common way to extract multi-scale features is to feed multiple resized input images to a shared deep network and then merge the resulting features for pixelwise classification. In this work, we propose an attention mechanism that learns to softly weight the multi-scale features at each pixel location. We adapt a state-of-the-art semantic image segmentation model, which we jointly train with multi-scale input images and the attention model. The proposed attention model not only outperforms averageand max-pooling, but allows us to diagnostically visualize the importance of features at different positions and scales. Moreover, we show that adding extra supervision to the output at each scale is essential to achieving excellent performance when merging multi-scale features. We demonstrate the effectiveness of our model with extensive experiments on three challenging datasets, including PASCAL-Person-Part, PASCAL VOC 2012 and a subset of MS-COCO 2014.

1.65 Automatic Image Cropping : A Computational Complexity Study

[Link to Paper]

Attention based automatic image cropping aims at preserving the most visually important region in an image. A common task in this kind of method is to search for the smallest rectangle inside which the summed attention is maximized. We demonstrate that under appropriate formulations, this task can be achieved using efficient algorithms with low computational complexity. In a practically useful scenario where the aspect ratio of the cropping rectangle is given, the problem can be solved with a computational complexity linear to the number of image pixels. We also study the possibility of multiple rectangle cropping and a new model facilitating fully automated image cropping.

1.66 DCAN: Deep Contour-Aware Networks for Accurate Gland Segmentation

[Link to Paper]

The morphology of glands has been used routinely by pathologists to assess the malignancy degree of adenocarcinomas. Accurate segmentation of glands from histology images is a crucial step to obtain reliable morphological statistics for quantitative diagnosis. In this

paper, we proposed an efficient deep contour-aware network (DCAN) to solve this challenging problem under a unified multi-task learning framework. In the proposed network, multi-level contextual features from the hierarchical architecture are explored with auxiliary supervision for accurate gland segmentation. When incorporated with multi-task regularization during the training, the discriminative capability of intermediate features can be further improved. Moreover, our network can not only output accurate probability maps of glands, but also depict clear contours simultaneously for separating clustered objects, which further boosts the gland segmentation performance. This unified framework can be efficient when applied to large-scale histopathological data without resorting to additional steps to generate contours based on low-level cues for post-separating. Our method won the 2015 MICCAI Gland Segmentation Challenge out of 13 competitive teams, surpassing all the other methods by a significant margin.

1.67 Full Flow: Optical Flow Estimation By Global Optimization over Regular Grids

[Link to Paper]

We present a global optimization approach to optical flow estimation. The approach optimizes a classical optical flow objective over the full space of mappings between discrete grids. No descriptor matching is used. The highly regular structure of the space of mappings enables optimizations that reduce the computational complexity of the algorithm's inner loop from quadratic to linear and support efficient matching of tens of thousands of nodes to tens of thousands of displacements. We show that one-shot global optimization of a classical Horn-Schunck-type objective over regular grids at a single resolution is sufficient to initialize continuous interpolation and achieve state-of-the-art performance on challenging modern benchmarks.

1.68 Learning Online Smooth Predictors for Realtime Camera Planning using Recurrent Decision Trees

[Link to Paper]

We study the problem of online prediction for realtime camera planning, where the goal is to predict smooth trajectories that correctly track and frame objects of interest (e.g., players in a basketball game). The conventional approach for training predictors does not directly consider temporal consistency, and often produces undesirable jitter. Although post-hoc smoothing (e.g., via a Kalman filter) can mitigate this issue to some degree, it is not ideal due to overly stringent modeling assumptions (e.g., Gaussian noise). We propose a recurrent decision tree framework that can directly incorporate temporal consistency

into a data-driven predictor, as well as a learning algorithm that can efficiently learn such temporally smooth models. Our approach does not require any post-processing, making online smooth predictions much easier to generate when the noise model is unknown. We apply our approach to sports broadcasting: given noisy player detections, we learn where the camera should look based on human demonstrations. Our experiments exhibit significant improvements over conventional baselines and showcase the practicality of our approach.

1.69   Monocular 3D Object Detection for Autonomous Driving

[Link to Paper]

The goal of this paper is to perform 3D object detection from a single monocular image in the domain of autonomous driving. Our method first aims to generate a set of candidate class-specific object proposals, which are then run through a standard CNN pipeline to obtain highquality object detections. The focus of this paper is on proposal generation. In particular, we propose an energy minimization approach that places object candidates in 3D using the fact that objects should be on the ground-plane. We then score each candidate box projected to the image plane via several intuitive potentials encoding semantic segmentation, contextual information, size and location priors and typical object shape. Our experimental evaluation demonstrates that our object proposal generation approach significantly outperforms all monocular approaches, and achieves the best detection performance on the challenging KITTI benchmark, among published monocular competitors.

1.70   Robust Tensor Factorization with Unknown Noise

[Link to Paper]

Because of the limitations of matrix factorization, such as losing spatial structure information, the concept of tensor factorization has been applied for the recovery of a low dimensional subspace from high dimensional visual data. Generally, the recovery is achieved by minimizing the loss function between the observed data and the factorization representation. Under different assumptions of the noise distribution, the loss functions are in various forms, like L1 and L2 norms. However, real data are often corrupted by noise with an unknown distribution. Then any specific form of loss function for one specific kind of noise often fails to tackle such real data with unknown noise. In this paper, we propose a tensor factorization algorithm to model the noise as a Mixture of Gaussians (MoG). As MoG has the ability of universally approximating any hybrids of continuous distributions, our algorithm can effectively recover the low dimensional subspace from various forms of

noisy observations. The parameters of MoG are estimated under the EM framework and through a new developed algorithm of weighted low-rank tensor factorization (WLRTF). The effectiveness of our algorithm are substantiated by extensive experiments on both of synthetic data and real image data.

1.71 Scale-Aware Alignment of Hierarchical Image Segmentation

[Link to Paper]

Image segmentation is a key component in many computer vision systems, and it is recovering a prominent spot in the literature as methods improve and overcome their limitations. The outputs of most recent algorithms are in the form of a hierarchical segmentation, which provides segmentation at different scales in a single tree-like structure. Commonly, these hierarchical methods start from some lowlevel features, and are not aware of the scale information of the different regions in them. As such, one might need to work on many different levels of the hierarchy to find the objects in the scene. This work tries to modify the existing hierarchical algorithm by improving their alignment, that is, by trying to modify the depth of the regions in the tree to better couple depth and scale. To do so, we first train a regressor to predict the scale of regions using mid-level features. We then define the anchor slice as the set of regions that better balance between over-segmentation and undersegmentation. The output of our method is an improved hierarchy, re-aligned by the anchor slice. To demonstrate the power of our method, we perform comprehensive experiments, which show that our method, as a post-processing step, can significantly improve the quality of the hierarchical segmentation representations, and ease the usage of hierarchical image segmentation to high-level vision tasks such as object segmentation. We also prove that the improvement generalizes well across different algorithms and datasets, with a low computational cost.1

1.72 Semantic Image Segmentation with Task-Specific Edge Detection Using CNNs and a Discriminatively Trained Domain Transform

[Link to Paper]

Deep convolutional neural networks (CNNs) are the backbone of state-of-art semantic image segmentation systems. Recent work has shown that complementing CNNs with fully-connected conditional random fields (CRFs) can significantly enhance their object localization accuracy, yet dense CRF inference is computationally expensive. We propose replacing the fully-connected CRF with domain transform (DT), a modern edge-preserving filtering method in which the amount of smoothing is controlled by a reference edge map. Domain transform filtering is several times faster than dense CRF inference and we show that it

yields comparable semantic segmentation results, accurately capturing object boundaries. Importantly, our formulation allows learning the reference edge map from intermediate CNN features instead of using the image gradient magnitude as in standard DT filtering. This produces task-specific edges in an end-to-end trainable system optimizing the target semantic segmentation quality.

1.73 Similarity Learning with Spatial Constraints for Person Re-identification

[Link to Paper]

Pose variation remains one of the major factors that adversely affect the accuracy of person re-identification. Such variation is not arbitrary as body parts (e.g. head, torso, legs) have relative stable spatial distribution. Breaking down the variability of global appearance regarding the spatial distribution potentially benefits the person matching. We therefore learn a novel similarity function, which consists of multiple sub-similarity measurements with each taking in charge of a subregion. In particular, we take advantage of the recently proposed polynomial feature map to describe the matching within each subregion, and inject all the feature maps into a unified framework. The framework not only outputs similarity measurements for different regions, but also makes a better consistency among them. Our framework can collaborate local similarities as well as global similarity to exploit their complementary strength. It is flexible to incorporate multiple visual cues to further elevate the performance. In experiments, we analyze the effectiveness of the major components. The results on four datasets show significant and consistent improvements over the state-of-the-art methods.

1.74 Person Re-Identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function

[Link to Paper]

Person re-identification across cameras remains a very challenging problem, especially when there are no overlapping fields of view between cameras. In this paper, we present a novel multi-channel parts-based convolutional neural network (CNN) model under the triplet framework for person re-identification. Specifically, the proposed CNN model consists of multiple channels to jointly learn both the global full-body and local body-parts features of the input persons. The CNN model is trained by an improved triplet loss function that serves to pull the instances of the same person closer, and at the same time push the instances belonging to different persons farther from each other in the learned feature space. Extensive comparative evaluations demonstrate that our proposed

method significantly outperforms many state-of-the-art approaches, including both traditional and deep network-based ones, on the challenging i-LIDS, VIPeR, PRID2011 and CUHK01 datasets.

1.75 RIFD-CNN: Rotation-Invariant and Fisher Discriminative Convolutional Neural Networks for Object Detection

[Link to Paper]

Thanks to the powerful feature representations obtained through deep convolutional neural network (CNN), the performance of object detection has recently been substantially boosted. Despite the remarkable success, the problems of object rotation, within-class variability, and between-class similarity remain several major challenges. To address these problems, this paper proposes a novel and effective method to learn a rotation-invariant and Fisher discriminative CNN (RIFD-CNN) model. This is achieved by introducing and learning a rotation-invariant layer and a Fisher discriminative layer, respectively, on the basis of the existing high-capacity CNN architectures. Specifically, the rotation-invariant layer is trained by imposing an explicit regularization constraint on the objective function that enforces invariance on the CNN features before and after rotating. The Fisher discriminative layer is trained by imposing the Fisher discrimination criterion on the CNN features so that they have small within-class scatter but large between-class separation. In the experiments, we comprehensively evaluate the proposed method for object detection task on a public available aerial image dataset and the PASCAL VOC 2007 dataset. State-of-the-art results are achieved compared with the existing baseline methods.

1.76 Subspace Clustering with Priors via Sparse Quadratically Constrained Quadratic Programming

[Link to Paper]

This paper considers the problem of recovering a subspace arrangement from noisy samples, potentially corrupted with outliers. Our main result shows that this problem can be formulated as a convex semi-definite optimization problem subject to an additional rank constrain that involves only a very small number of variables. This is established by first reducing the problem to a quadratically constrained quadratic problem and then using its special structure to find conditions guaranteeing that a suitably built convex relaxation is indeed exact. When combined with the standard nuclear norm relaxation for rank, the results above lead to computationally efficient algorithms with optimality guarantees. A salient feature of the proposed approach is its ability to incorporate existing a-priori in-

formation about the noise, co-ocurrences, and percentage of outliers. These results are illustrated with several examples.

1.77 Two Illuminant Estimation and User Correction Preference

[Link to Paper]

This paper examines the problem of white-balance correction when a scene contains two illuminations. This is a two step process: 1) estimate the two illuminants; and 2) correct the image. Existing methods attempt to estimate a spatially varying illumination map, however, results are error prone and the resulting illumination maps are too lowresolution to be used for proper spatially varying whitebalance correction. In addition, the spatially varying nature of these methods make them computationally intensive. We show that this problem can be effectively addressed by not attempting to obtain a spatially varying illumination map, but instead by performing illumination estimation on large sub-regions of the image. Our approach is able to detect when distinct illuminations are present in the image and accurately measure these illuminants. Since our proposed strategy is not suitable for spatially varying image correction, a user study is performed to see if there is a preference for how the image should be corrected when two illuminants are present, but only a global correction can be applied. The user study shows that when the illuminations are distinct, there is a preference for the outdoor illumination to be corrected resulting in warmer final result. We use these collective findings to demonstrate an effective two illuminant estimation scheme that produces corrected images that users prefer.

1.78 Inextensible Non-Rigid Shape-from-Motion by Second-Order Cone Programming

[Link to Paper]

We present a global and convex formulation for template-less 3D reconstruction of a deforming object with the perspective camera. We show for the first time how to construct a Second-Order Cone Programming (SOCP) problem for Non-Rigid Shape-from-Motion (NRSfM) using the Maximum-Depth Heuristic (MDH). In this regard, we deviate strongly from the general trend of using affine cameras and factorization-based methods to solve NRSfM. In MDH, the points' depths are maximized so that the distance between neighbouring points in camera space are upper bounded by the geodesic distance. In NRSfM both geodesic and camera space distances are unknown. We show that, nonetheless, given point correspondences and the camera's intrinsics the whole problem is convex and solvable with SOCP. We show with extensive experiments that our method accurately reconstructs quasi-isometric surfaces from partial views under articulated and strong deformations. It

naturally handles missing correspondences, non-smooth objects and is very simple to implement compared to previous methods, with only one free parameter (the neighbourhood size).

1.79 Guaranteed Outlier Removal with Mixed Integer Linear Programs

[Link to Paper]

The maximum consensus problem is fundamentally important to robust geometric fitting in computer vision. Solving the problem exactly is computationally demanding, and the effort required increases rapidly with the problem size. Although randomized algorithms are much more efficient, the optimality of the solution is not guaranteed. Towards the goal of solving maximum consensus exactly, we present guaranteed outlier removal as a technique to reduce the runtime of exact algorithms. Specifically, before conducting global optimization, we attempt to remove data that are provably true outliers, i.e., those that do not exist in the maximum consensus set. We propose an algorithm based on mixed integer linear programming to perform the removal. The result of our algorithm is a smaller data instance that admits a much faster solution by subsequent exact algorithms, while yielding the same globally optimal result as the original problem. We demonstrate that overall speedups of up to 80% can be achieved on common vision problems1.

1.80 Canny Text Detector: Fast and Robust Scene Text Localization Algorithm

[Link to Paper]

This paper presents a novel scene text detection algorithm, Canny Text Detector, which takes advantage of the similarity between image edge and text for effective text localization with improved recall rate. As closely related edge pixels construct the structural information of an object, we observe that cohesive characters compose a meaningful word/sentence sharing similar properties such as spatial location, size, color, and stroke width regardless of language. However, prevalent scene text detection approaches have not fully utilized such similarity, but mostly rely on the characters classified with high confidence, leading to low recall rate. By exploiting the similarity, our approach can quickly and robustly localize a variety of texts. Inspired by the original Canny edge detector, our algorithm makes use of double threshold and hysteresis tracking to detect texts of low confidence. Experimental results on public datasets demonstrate that our algorithm outperforms the state-ofthe-art scene text detection methods in terms of detection rate.

1.81 Improving Person Re-identification via Pose-aware Multi-shot Matching

[Link to Paper]

Person re-identification is the problem of recognizing people across images or videos from non-overlapping views. Although there has been much progress in person re-identification for the last decade, it still remains a challenging task because of severe appearance changes of a person due to diverse camera viewpoints and person poses. In this paper, we propose a novel framework for person reidentification by analyzing camera viewpoints and person poses, so-called Pose-aware Multi-shot Matching (PaMM), which robustly estimates target poses and efficiently conducts multi-shot matching based on the target pose information. Experimental results using public person reidentification datasets show that the proposed methods are promising for person re-identification under diverse viewpoints and pose variances.

1.82 Simultaneous Estimation of Near IR BRDF and Fine-Scale Surface Geometry

[Link to Paper]

Near-Infrared (NIR) images of most materials exhibit less texture or albedo variations making them beneficial for vision tasks such as intrinsic image decomposition and structured light depth estimation. Understanding the reflectance properties (BRDF) of materials in the NIR wavelength range can be further useful for many photometric methods including shape from shading and inverse rendering. However, even with less albedo variation, many materials e.g. fabrics, leaves, etc. exhibit complex fine-scale surface detail making it hard to accurately estimate BRDF. In this paper, we present an approach to simultaneously estimate NIR BRDF and fine-scale surface details by imaging materials under different IR lighting and viewing directions. This is achieved by an iterative scheme that alternately estimates surface detail and NIR BRDF of materials. Our setup does not require complicated gantries or calibration and we present the first NIR dataset of 100 materials including a variety of fabrics (knits, weaves, cotton, satin, leather), and organic (skin, leaves, jute, trunk, fur) and inorganic materials (plastic, concrete, carpet). The NIR BRDFs measured from material samples are used with a shape-from-shading algorithm to demonstrate fine-scale reconstruction of objects from a single NIR image.

1.83 Video-Story Composition via Plot Analysis

[Link to Paper]

We address the problem of composing a story out of multiple short video clips taken by a person during an activity or experience. Inspired by plot analysis of written stories, our method generates a sequence of video clips ordered in such a way that it reflects plot dynamics and content coherency. That is, given a set of multiple video clips, our method composes a video which we call a video-story. We define metrics on scene dynamics and coherency by dense optical flow features and a patch matching algorithm. Using these metrics, we define an objective function for the video-story. To efficiently search for the best video-story, we introduce a novel Branch-and-Bound algorithm which guarantees the global optimum. We collect the dataset consisting of 23 video sets from the web, resulting in a total of 236 individual video clips. With the acquired dataset, we perform extensive user studies involving 30 human subjects by which the effectiveness of our approach is quantitatively and qualitatively verified.

1.84 Visual Tracking Using Attention-Modulated Disintegration and Integration

[Link to Paper]

In this paper, we present a novel attention-modulated visual tracking algorithm that decomposes an object into multiple cognitive units, and trains multiple elementary trackers in order to modulate the distribution of attention according to various feature and kernel types. In the integration stage it recombines the units to memorize and recognize the target object effectively. With respect to the elementary trackers, we present a novel attentional feature-based correlation filter (AtCF) that focuses on distinctive attentional features. The effectiveness of the proposed algorithm is validated through experimental comparison with state-of-theart methods on widely-used tracking benchmark datasets.

1.85 Backtracking ScSPM Image Classifier for Weakly Supervised Top-down Saliency

[Link to Paper]

Top-down saliency models produce a probability map that peaks at target locations specified by a task/goal such as object detection. They are usually trained in a supervised setting involving annotations of objects. We propose a weakly supervised top-down saliency framework using only binary labels that indicate the presence/absence of an object in an image. First, the probabilistic contribution of each image patch to the confidence of an ScSPM-based classifier produces a Reverse-ScSPM (R-ScSPM) saliency map. Neighborhood information is then incorporated through a contextual saliency map which is estimated using logistic regression learnt on patches having high R-ScSPM saliency. Both the saliency

42

maps are combined to obtain the final saliency map. We evaluate the performance of the proposed weakly supervised top-down saliency and achieves comparable performance with fully supervised approaches. Experiments are carried out on 5 challenging datasets across 3 different applications.

1.86 Structured Feature Learning for Pose Estimation

[Link to Paper]

In this paper, we propose a structured feature learning framework to reason the correlations among body joints at the feature level in human pose estimation. Different from existing approaches of modeling structures on score maps or predicted labels, feature maps preserve substantially richer descriptions of body joints. The relationships between feature maps of joints are captured with the introduced geometrical transform kernels, which can be easily implemented with a convolution layer. Features and their relationships are jointly learned in an end-to-end learning system. A bi-directional tree structured model is proposed, so that the feature channels at a body joint can well receive information from other joints. The proposed framework improves feature learning substantially. With very simple post processing, it reaches the best mean PCP on the LSP and FLIC datasets. Compared with the baseline of learning features at each joint separately with ConvNet, the mean PCP has been improved by 18% on FLIC. The code is released to the public. 1

1.87 Deep SimNets

[Link to Paper]

We present a deep layered architecture that generalizes convolutional neural networks (ConvNets). The architecture, called SimNets, is driven by two operators: (i) a similarity function that generalizes inner-product, and (ii) a log-mean-exp function called MEX that generalizes maximum and average. The two operators applied in succession give rise to a standard neuron but in "feature space". The feature spaces realized by SimNets depend on the choice of the similarity operator. The simplest setting, which corresponds to a convolution, realizes the feature space of the Exponential kernel, while other settings realize feature spaces of more powerful kernels (Generalized Gaussian, which includes as special cases RBF and Laplacian), or even dynamically learned feature spaces (Generalized Multiple Kernel Learning). As a result, the SimNet contains a higher abstraction level compared to a traditional ConvNet. We argue that enhanced expressiveness is important when the networks are small due to run-time constraints (such as those imposed by mobile applications). Empirical evaluation validates the superior expressiveness of SimNets, showing a significant gain in accuracy over ConvNets when computational resources at run-time are

limited. We also show that in large-scale settings, where computational complexity is less of a concern, the additional capacity of SimNets can be controlled with proper regularization, yielding accuracies comparable to state of the art ConvNets.

1.88   The Cityscapes Dataset for Semantic Urban Scene Understanding

[Link to Paper]

Visual understanding of complex urban street scenes is an enabling factor for a wide range of applications. Object detection has benefited enormously from large-scale datasets, especially in the context of deep learning. For semantic urban scene understanding, however, no current dataset adequately captures the complexity of real-world urban scenes. To address this, we introduce Cityscapes, a benchmark suite and large-scale dataset to train and test approaches for pixel-level and instance-level semantic labeling. Cityscapes is comprised of a large, diverse set of stereo video sequences recorded in streets from 50 different cities. 5000 of these images have high quality pixel-level annotations; 20 000 additional images have coarse annotations to enable methods that leverage large volumes of weakly-labeled data. Crucially, our effort exceeds previous attempts in terms of dataset size, annotation richness, scene variability, and complexity. Our accompanying empirical study provides an in-depth analysis of the dataset characteristics, as well as a performance evaluation of several state-of-the-art approaches based on our benchmark.

1.89   Semantic Channels for Fast Pedestrian Detection

[Link to Paper]

Pedestrian detection and semantic segmentation are high potential tasks for many real-time applications. However most of the top performing approaches provide state of art results at high computational costs. In this work we propose a fast solution for achieving state of art results for both pedestrian detection and semantic segmentation. As baseline for pedestrian detection we use sliding windows over cost efficient multiresolution filtered LUV+HOG channels. We use the same channels for classifying pixels into eight semantic classes. Using short range and long range multiresolution channel features we achieve more robust segmentation results compared to traditional codebook based approaches at much lower computational costs. The resulting segmentations are used as additional semantic channels in order to achieve a more powerful pedestrian detector. To also achieve fast pedestrian detection we employ a multiscale detection scheme based on a single flexible pedestrian model and a single image scale. The proposed solution provides competitive results on both pedestrian detection and semantic segmentation benchmarks at 8 FPS on CPU and at 15 FPS on GPU, being the fastest top performing approach.

1.90 Structure from Motion with Objects

[Link to Paper]

This paper shows for the first time that is possible to reconstruct the position of rigid objects and to jointly recover affine camera calibration solely from a set of object detections in a video sequence. In practice, this work can be considered as the extension of Tomasi and Kanade factorization method using objects. Instead of using points to form a rank constrained measurement matrix, we can form a matrix with similar rank properties using 2D object detection proposals. In detail, we first fit an ellipse onto the image plane at each bounding box as given by the object detector. The collection of all the ellipses in the dual space is used to create a measurement matrix that gives a specific rank constraint. This matrix can be factorised and metrically upgraded in order to provide the affine camera matrices and the 3D position of the objects as an ellipsoid. Moreover, we recover the full 3D quadric thus giving additional information about object occupancy and 3D pose. Finally, we also show that 2D points measurements can be seamlessly included in the framework to reduce the number of objects required. This last aspect unifies the classical point-based Tomasi and Kanade approach with objects in a unique framework. Experiments with synthetic and real data show the feasibility of our approach for the affine camera case.

1.91 Fine-grained Categorization and Dataset Bootstrapping using Deep Metric Learning with Humans in the Loop

[Link to Paper]

Existing fine-grained visual categorization methods often suffer from three challenges: lack of training data, large number of fine-grained categories, and high intraclass vs. low interclass variance. In this work we propose a generic iterative framework for fine-grained categorization and dataset bootstrapping that handles these three challenges. Using deep metric learning with humans in the loop, we learn a low dimensional feature embedding with anchor points on manifolds for each category. These anchor points capture intra-class variances and remain discriminative between classes. In each round, images with high confidence scores from our model are sent to humans for labeling. By comparing with exemplar images, labelers mark each candidate image as either a "true positive" or a "false positive." True positives are added into our current dataset and false positives are regarded as "hard negatives" for our metric learning model. Then the model is retrained with an expanded dataset and hard negatives for the next round. To demonstrate the effectiveness of the proposed framework, we bootstrap a fine-grained flower dataset with 620 categories from Instagram images. The proposed deep metric learning scheme is evaluated on both our dataset and the CUB-200-2001 Birds dataset. Experimental evaluations show significant performance gain using dataset bootstrapping and demonstrate state-of-the-art results

achieved by the proposed deep metric learning methods.

1.92   Recurrently Target-Attending Tracking

[Link to Paper]

Robust visual tracking is a challenging task in computer vision. Due to the accumulation and propagation of estimation error, model drifting often occurs and degrades the tracking performance. To mitigate this problem, in this paper we propose a novel tracking method called Recurrently Target-attending Tracking (RTT). RTT attempts to identify and exploit those reliable parts which are beneficial for the overall tracking process. To bypass occlusion and discover reliable components, multi-directional Recurrent Neural Networks (RNNs) are employed in RTT to capture long-range contextual cues by traversing a candidate spatial region from multiple directions. The produced confidence maps from the RNNs are employed to adaptively regularize the learning of discriminative correlation filters by suppressing clutter background noises while making full use of the information from reliable parts. To solve the weighted correlation filters, we especially derive an efficient closedform solution with a sharp reduction in computation complexity. Extensive experiments demonstrate that our proposed RTT is more competitive over those correlation filter based methods.

1.93   Instance-aware Semantic Segmentation via Multi-task Network Cascades

[Link to Paper]

Semantic segmentation research has recently witnessed rapid progress, but many leading methods are unable to identify object instances. In this paper, we present Multitask Network Cascades for instance-aware semantic segmentation. Our model consists of three networks, respectively differentiating instances, estimating masks, and categorizing objects. These networks form a cascaded structure, and are designed to share their convolutional features. We develop an algorithm for the nontrivial end-to-end training of this causal, cascaded structure. Our solution is a clean, single-step training framework and can be generalized to cascades that have more stages. We demonstrate state-of-the-art instance-aware semantic segmentation accuracy on PASCAL VOC. Meanwhile, our method takes only 360ms testing an image using VGG-16, which is two orders of magnitude faster than previous systems for this challenging problem. As a by product, our method also achieves compelling object detection results which surpass the competitive Fast/Faster R-CNN systems. The method described in this paper is the foundation of our submissions to the MS COCO 2015 segmentation competition, where we won the 1st place.

1.94 Rolling Shutter Camera Relative Pose: Generalized Epipolar Geometry

[Link to Paper]

The vast majority of modern consumer-grade cameras employ a rolling shutter mechanism. In dynamic geometric computer vision applications such as visual SLAM, the so-called rolling shutter effect therefore needs to be properly taken into account. A dedicated relative pose solver appears to be the first problem to solve, as it is of eminent importance to bootstrap any derivation of multi-view geometry. However, despite its significance, it has received inadequate attention to date. This paper presents a detailed investigation of the geometry of the rolling shutter relative pose problem. We introduce the rolling shutter essential matrix, and establish its link to existing models such as the push-broom cameras, summarized in a clean hierarchy of multi-perspective cameras. The generalization of well-established concepts from epipolar geometry is completed by a definition of the Sampson distance in the rolling shutter case. The work is concluded with a careful investigation of the introduced epipolar geometry for rolling shutter cameras on several dedicated benchmarks.

1.95 A Probabilistic Framework for Color-Based Point Set Registration

[Link to Paper]

In recent years, sensors capable of measuring both color and depth information have become increasingly popular. Despite the abundance of colored point set data, stateof-the-art probabilistic registration techniques ignore the available color information. In this paper, we propose a probabilistic point set registration framework that exploits available color information associated with the points. Our method is based on a model of the joint distribution of 3D-point observations and their color information. The proposed model captures discriminative color information, while being computationally efficient. We derive an EM algorithm for jointly estimating the model parameters and the relative transformations. Comprehensive experiments are performed on the Stanford Lounge dataset, captured by an RGB-D camera, and two point sets captured by a Lidar sensor. Our results demonstrate a significant gain in robustness and accuracy when incorporating color information. On the Stanford Lounge dataset, our approach achieves a relative reduction of the failure rate by 78% compared to the baseline. Furthermore, our proposed model outperforms standard strategies for combining color and 3D-point information, leading to state-of-the-art results.

1.96 Adaptive Decontamination of the Training Set: A Unified Formulation for Discriminative Visual Tracking

[Link to Paper]

Tracking-by-detection methods have demonstrated competitive performance in recent years. In these approaches, the tracking model heavily relies on the quality of the training set. Due to the limited amount of labeled training data, additional samples need to be extracted and labeled by the tracker itself. This often leads to the inclusion of corrupted training samples, due to occlusions, misalignments and other perturbations. Existing tracking-by-detection methods either ignore this problem, or employ a separate component for managing the training set. We propose a novel generic approach for alleviating the problem of corrupted training samples in tracking-bydetection frameworks. Our approach dynamically manages the training set by estimating the quality of the samples. Contrary to existing approaches, we propose a unified formulation by minimizing a single loss over both the target appearance model and the sample quality weights. The joint formulation enables corrupted samples to be downweighted while increasing the impact of correct ones. Experiments are performed on three benchmarks: OTB-2015 with 100 videos, VOT-2015 with 60 videos, and TempleColor with 128 videos. On the OTB-2015, our unified formulation significantly improves the baseline, with a gain of 3.8% in mean overlap precision. Finally, our method achieves state-of-the-art results on all three datasets.

1.97 Temporal Epipolar Regions

[Link to Paper]

Dynamic events are often photographed by a number of people from different viewpoints at different times, resulting in an unconstrained set of images. Finding the corresponding moving features in each of the images allows us to extract information about objects of interest in the scene. Computing correspondence of moving features in such a set of images is considerably more challenging than computing correspondence in video due to possible significant differences in viewpoints and inconsistent timing between image captures. The prediction methods used in video for improving robustness and efficiency are not applicable to a set of still images. In this paper we propose a novel method to predict locations of an approximately linear moving feature point, given a small subset of correspondences and the temporal order of image captures. Our method extends the use of epipolar geometry to divide images into valid and invalid regions, termed Temporal Epipolar Regions (TERs). We formally prove that the location of a feature in a new image is restricted to valid TERs. We demonstrate the effectiveness of our method in reducing the search space for correspondence on both synthetic and challenging real world data, and show the improved matching.

1.98 DeLay: Robust Spatial Layout Estimation for Cluttered Indoor Scenes

[Link to Paper]

We consider the problem of estimating the spatial layout of an indoor scene from a monocular RGB image, modeled as the projection of a 3D cuboid. Existing solutions to this problem often rely strongly on hand-engineered features and vanishing point detection, which are prone to failure in the presence of clutter. In this paper, we present a method that uses a fully convolutional neural network (FCNN) in conjunction with a novel optimization framework for generating layout estimates. We demonstrate that our method is robust in the presence of clutter and handles a wide range of highly challenging scenes. We evaluate our method on two standard benchmarks and show that it achieves state of the art results, outperforming previous methods by a wide margin.

1.99 Geometry-Informed Material Recognition

[Link to Paper]

Our goal is to recognize material categories using images and geometry information. In many applications, such as construction management, coarse geometry information is available. We investigate how 3D geometry (surface normals, camera intrinsic and extrinsic parameters) can be used with 2D features (texture and color) to improve material classification. We introduce a new dataset, GeoMat, which is the first to provide both image and geometry data in the form of: (i) training and testing patches that were extracted at different scales and perspectives from real world examples of each material category, and (ii) a large scale construction site scene that includes 160 images and over 800,000 hand labeled 3D points. Our results show that using 2D and 3D features both jointly and independently to model materials improves classification accuracy across multiple scales and viewing directions for both material patches and images of a large scale construction site scene.

1.100 Discovering the physical parts of an articulated object class from multiple videos

[Link to Paper]

We propose a motion-based method to discover the physical parts of an articulated object class (e.g. head/torso/leg of a horse) from multiple videos. The key is to find object regions that exhibit consistent motion relative to the rest of the object, across multiple videos. We can then learn a location model for the parts and segment them accurately in the individual

videos using an energy function that also enforces temporal and spatial consistency in part motion. Unlike our approach, traditional methods for motion segmentation or non-rigid structure from motion operate on one video at a time. Hence they cannot discover a part unless it displays independent motion in that particular video. We evaluate our method on a new dataset of 32 videos of tigers and horses, where we significantly outperform a recent motion segmentation method on the task of part discovery (obtaining roughly twice the accuracy).

1.101   Similarity Metric For Curved Shapes In Euclidean Space

[Link to Paper]

In this paper, we introduce a similarity metric for curved shapes that can be described, distinctively, by ordered points. The proposed method represents a given curve as a point in the deformation space, the direct product of rigid transformation matrices, such that the successive action of the matrices on a fixed starting point reconstructs the full curve. In general, both open and closed curves are represented in the deformation space modulo shape orientation and orientation preserving diffeomorphisms. The use of direct product Lie groups to represent curved shapes led to an explicit formula for geodesic curves and the formulation of a similarity metric between shapes by the L2-norm on the Lie algebra. Additionally, invariance to reparametrization or estimation of point correspondence between shapes is performed as an intermediate step for computing geodesics. Furthermore, since there is no computation of differential quantities on the curves, our representation is more robust to local perturbations and needs no pre-smoothing. We compare our method with the elastic shape metric defined through the square root velocity (SRV) mapping, and other shape matching approaches.

1.102   Structure Inference Machines: Recurrent Neural Networks for Analyzing Relations in Group Activity Recognition

[Link to Paper]

Rich semantic relations are important in a variety of visual recognition problems. As a concrete example, group activity recognition involves the interactions and relative spatial relations of a set of people in a scene. State of the art recognition methods center on deep learning approaches for training highly effective, complex classifiers for interpreting images. However, bridging the relatively low-level concepts output by these methods to interpret higher-level compositional scenes remains a challenge. Graphical models are a standard tool for this task. In this paper, we propose a method to integrate graphical models and deep neural networks into a joint framework. Instead of using a traditional

inference method, we use a sequential inference modeled by a recurrent neural network. Beyond this, the appropriate structure for inference can be learned by imposing gates on edges between nodes. Empirical results on group activity recognition demonstrate the potential of this model to handle highly structured learning tasks.

1.103   A Continuous Occlusion Model for Road Scene Understanding

[Link to Paper]

We present a physically interpretable, continuous threedimensional (3D) model for handling occlusions with applications to road scene understanding. We probabilistically assign each point in space to an object with a theoretical modeling of the reflection and transmission probabilities for the corresponding camera ray. Our modeling is unified in handling occlusions across a variety of scenarios, such as associating structure from motion (SFM) point tracks with potentially occluding objects or modeling object detection scores in applications such as 3D localization. For point track association, our model uniformly handles static and dynamic objects, which is an advantage over motion segmentation approaches traditionally used in multibody SFM. Detailed experiments on the KITTI raw dataset show the superiority of the proposed method over both state-of-the-art motion segmentation and a baseline that heuristically uses detection bounding boxes for resolving occlusions. We also demonstrate how our continuous occlusion model may be applied to the task of 3D localization in road scenes.

1.104   DeepCAMP: Deep Convolutional Action & Attribute Mid-Level Patterns

[Link to Paper]

The recognition of human actions and the determination of human attributes are two tasks that call for fine-grained classification. Indeed, often rather small and inconspicuous objects and features have to be detected to tell their classes apart. In order to deal with this challenge, we propose a novel convolutional neural network that mines mid-level image patches that are sufficiently dedicated to resolve the corresponding subtleties. In particular, we train a newly designed CNN (DeepPattern) that learns discriminative patch groups. There are two innovative aspects to this. On the one hand we pay attention to contextual information in an original fashion. On the other hand, we let an iteration of feature learning and patch clustering purify the set of dedicated patches that we use. We validate our method for action classification on two challenging datasets: PASCAL VOC 2012 Action and Stanford 40 Actions, and for attribute recognition we use the Berkeley Attributes of People dataset. Our discriminative mid-level mining CNN obtains state-of-theart results on these datasets, without a need for annotations about parts and poses.

51

1.105 Solving Temporal Puzzles

[Link to Paper]

Many physical phenomena, within short time windows, can be explained by low order differential relations. In a discrete world, these relations can be described using low order difference equations or equivalently low order auto regressive (AR) models. In this paper, based on this intuition, we propose an algorithm for solving time-sort temporal puzzles, defined as scrambled time series that need to be sorted out. We frame this problem using a mixed-integer semi definite programming formulation and show how to turn it into a mixed-integer linear programming problem, which can be solved with off-the-shelf solvers, by using the recently introduced atomic norm framework. Our experiments show the effectiveness and generality of our approach in different scenarios.

1.106 Heterogeneous Light Fields

[Link to Paper]

In contrast to traditional binocular or multi-view stereo approaches, the adequately sampled space of observations in light-field imaging allows, to obtain dense and high quality depth maps. It also extends capabilities beyond those of traditional methods. Previously, constant intensity has been assumed for estimating disparity of orientation in most approaches to analyze epipolar plane images (EPIs). Here, we introduce a modified structure tensor approach which improves depth estimation. This extension also includes a model of non-constant intensity on EPI manifolds. We derive an approach to estimate high quality depth maps in luminance-gradient light fields, as well as in color-filtered light fields. Color-filtered light fields pose particular challenges due to the fact that structures can change significantly in appearance with wavelength and can completely vanish at some wavelength. We demonstrate solutions to this challenge and obtain a dense sRGB image reconstruction in addition to dense depth maps.

1.107 Structured Regression Gradient Boosting

[Link to Paper]

xi Feature Extraction We propose a new way to train a structured output prediction model. More specifically, we train nonlinear data terms in a Gaussian Conditional Random Field (GCRF) by a generalized version of gradient boosting. The approach is evaluated on three challenging regression benchmarks: vessel detection, single image depth estimation and image inpainting. These experiments suggest that the proposed boosting framework

matches or exceeds the state-of-the-art. F1r i,m weak learner b1 b1,m(xi) y i r i,m F2r i,m b2,m(xi) hm (xi;b) weak learner b2 M (xi) F3r i,m weak learner b3 b3,m(xi)

1.108 Inverting Visual Representations with Convolutional Networks

[Link to Paper]

Feature representations, both hand-designed and learned ones, are often hard to analyze and interpret, even when they are extracted from visual data. We propose a new approach to study image representations by inverting them with an up-convolutional neural network. We apply the method to shallow representations (HOG, SIFT, LBP), as well as to deep networks. For shallow representations our approach provides significantly better reconstructions than existing methods, revealing that there is surprisingly rich information contained in these features. Inverting a deep network trained on ImageNet provides several insights into the properties of the feature representation learned by the network. Most strikingly, the colors and the rough contours of an image can be reconstructed from activations in higher network layers and even from the predicted class probabilities. HOG SIFT AlexNet-CONV3 AlexNet-FC8 Figure 1: We train convolutional networks to reconstruct images from different feature representations. Top row: Input features. Bottom row: Reconstructed image. Reconstructions from HOG and SIFT are very realistic. Reconstructions from AlexNet preserve color and rough object positions even when reconstructing from higher layers.

1.109 Recovering 6D Object Pose and Predicting Next-Best-View in the Crowd

[Link to Paper]

Object detection and 6D pose estimation in the crowd (scenes with multiple object instances, severe foreground occlusions and background distractors), has become an important problem in many rapidly evolving technological areas such as robotics and augmented reality. Single shotbased 6D pose estimators with manually designed features are still unable to tackle the above challenges, motivating the research towards unsupervised feature learning and next-best-view estimation. In this work, we present a complete framework for both single shot-based 6D object pose estimation and next-best-view prediction based on Hough Forests, the state of the art object pose estimator that performs classification and regression jointly. Rather than using manually designed features we a) propose an unsupervised feature learnt from depth-invariant patches using a Sparse Autoencoder and b) offer an extensive evaluation of various state of the art features. Furthermore, taking advantage of the clustering performed in the leaf nodes of Hough Forests, we learn to

estimate the reduction of uncertainty in other views, formulating the problem of selecting the next-best-view. To further improve pose estimation, we propose an improved joint registration and hypotheses verification module as a final refinement step to reject false detections. We provide two additional challenging datasets inspired from realistic scenarios to extensively evaluate the state of the art and our framework. One is related to domestic environments and the other depicts a bin-picking scenario mostly found in industrial settings. We show that our framework significantly outperforms state of the art both on public and on our datasets.

1.110 Longitudinal Face Modeling via Temporal Deep Restricted Boltzmann Machines

[Link to Paper]

Modeling the face aging process is a challenging task due to large and non-linear variations present in different stages of face development. This paper presents a deep model approach for face age progression that can efficiently capture the non-linear aging process and automatically synthesize a series of age-progressed faces in various age ranges. In this approach, we first decompose the longterm age progress into a sequence of short-term changes and model it as a face sequence. The Temporal Deep Restricted Boltzmann Machines based age progression model together with the prototype faces are then constructed to learn the aging transformation between faces in the sequence. In addition, to enhance the wrinkles of faces in the later age ranges, the wrinkle models are further constructed using Restricted Boltzmann Machines to capture their variations in different facial regions. The geometry constraints are also taken into account in the last step for more consistent age-progressed results. The proposed approach is evaluated using various face aging databases, i.e. FGNET, Cross-Age Celebrity Dataset (CACD) and MORPH, and our collected large-scale aging database named AginG Faces in the Wild (AGFW). In addition, when ground-truth age is not available for input image, our proposed system is able to automatically estimate the age of the input face before aging process is employed.

1.111 WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks

[Link to Paper]

In this paper, we introduce a novel framework for WEakly supervised Learning of Deep cOnvolutional neural Networks (WELDON). Our method is dedicated to automatically selecting relevant image regions from weak annotations, e.g. global image labels, and encompasses the following contributions. Firstly, WELDON leverages recent improvements on

the Multiple Instance Learning paradigm, i.e. negative evidence scoring and top instance selection. Secondly, the deep CNN is trained to optimize Average Precision, and fine-tuned on the target dataset with efficient computations due to convolutional feature sharing. A thorough experimental validation shows that WELDON outperforms state-of-the-art results on six different datasets.

1.112 Accelerated Generative Models for 3D Point Cloud Data

[Link to Paper]

Finding meaningful, structured representations of 3D point cloud data (PCD) has become a core task for spatial perception applications. In this paper we introduce a method for constructing compact generative representations of PCD at multiple levels of detail. As opposed to deterministic structures such as voxel grids or octrees, we propose probabilistic subdivisions of the data through local mixture modeling, and show how these subdivisions can provide a maximum likelihood segmentation of the data. The final representation is hierarchical, compact, parametric, and statistically derived, facilitating run-time occupancy calculations through stochastic sampling. Unlike traditional deterministic spatial subdivision methods, our technique enables dynamic creation of voxel grids according the application's best needs. In contrast to other generative models for PCD, we explicitly enforce sparsity among points and mixtures, a technique which we call expectation sparsification. This leads to a highly parallel hierarchical Expectation Maximization (EM) algorithm well-suited for the GPU and real-time execution. We explore the trade-offs between model fidelity and model size at various levels of detail, our tests showing favorable performance when compared to octree and NDT-based methods.

1.113 MCMC Shape Sampling for Image Segmentation with Nonparametric Shape Priors

[Link to Paper]

Segmenting images of low quality or with missing data is a challenging problem. Integrating statistical prior information about the shapes to be segmented can improve the segmentation results significantly. Most shape-based segmentation algorithms optimize an energy functional and find a point estimate for the object to be segmented. This does not provide a measure of the degree of confidence in that result, neither does it provide a picture of other probable solutions based on the data and the priors. With a statistical view, addressing these issues would involve the problem of characterizing the posterior densities of the shapes of the objects to be segmented. For such characterization, we propose a Markov chain Monte Carlo (MCMC) samplingbased image segmentation algorithm that

uses statistical shape priors. In addition to better characterization of the statistical structure of the problem, such an approach would also have the potential to address issues with getting stuck at local optima, suffered by existing shape-based segmentation methods. Our approach is able to characterize the posterior probability density in the space of shapes through its samples, and to return multiple solutions, potentially from different modes of a multimodal probability density, which would be encountered, e.g., in segmenting objects from multiple shape classes. We present promising results on a variety of data sets. We also provide an extension for segmenting shapes of objects with parts that can go through independent shape variations. This extension involves the use of local shape priors on object parts and provides robustness to limitations in shape training data size.

1.114 A Consensus-Based Framework for Distributed Bundle Adjustment

[Link to Paper]

In this paper we study large-scale optimization problems in multi-view geometry, in particular the Bundle Adjustment problem. In its conventional formulation, the complexity of existing solvers scale poorly with problem size, hence this component of the Structure-from-Motion pipeline can quickly become a bottle-neck. Here we present a novel formulation for solving bundle adjustment in a truly distributed manner using consensus based optimization methods. Our algorithm is presented with a concise derivation based on proximal splitting, along with a theoretical proof of convergence and brief discussions on complexity and implementation. Experiments on a number of real image datasets convincingly demonstrates the potential of the proposed method by outperforming the conventional bundle adjustment formulation by orders of magnitude.

1.115 A Paradigm for Building Generalized Models of Human Image Perception through Data Fusion

[Link to Paper]

In many sub-fields, researchers collect datasets of human ground truth that are used to create a new algorithm. For example, in research on image perception, datasets have been collected for topics such as what makes an image aesthetic or memorable. Despite high costs for human data collection, datasets are infrequently reused beyond their own fields of interest. Moreover, the algorithms built from them are domain-specific (predict a small set of attributes) and usually unconnected to one another. In this paper, we present a paradigm for building generalized and expandable models of human image perception. First, we fuse multiple fragmented and partially-overlapping datasets through data imputation. We then create a theoretically-structured statistical model of human image perception that is fit to

the fused datasets. The resulting model has many advantages. (1) It is generalized, going beyond the content of the constituent datasets, and can be easily expanded by fusing additional datasets. (2) It provides a new ontology usable as a network to expand human data in a cost-effective way. (3) It can guide the design of a generalized computational algorithm for multi-dimensional visual perception. Indeed, experimental results show that a model-based algorithm outperforms state-of-the-art methods on predicting visual sentiment, visual realism and interestingness. Our paradigm can be used in various visual tasks (e.g., video summarization).

1.116 HyperDepth: Learning Depth from Structured Light Without Matching

[Link to Paper]

Structured light sensors are popular due to their robustness to untextured scenes and multipath. These systems triangulate depth by solving a correspondence problem between each camera and projector pixel. This is often framed as a local stereo matching task, correlating patches of pixels in the observed and reference image. However, this is computationally intensive, leading to reduced depth accuracy and framerate. We contribute an algorithm for solving this correspondence problem efficiently, without compromising depth accuracy. For the first time, this problem is cast as a classification-regression task, which we solve extremely efficiently using an ensemble of cascaded random forests. Our algorithm scales in number of disparities, and each pixel can be processed independently, and in parallel. No matching or even access to the corresponding reference pattern is required at runtime, and regressed labels are directly mapped to depth. Our GPU-based algorithm runs at a 1KHz for 1.3MP input/output images, with disparity error of 0.1 subpixels. We show a prototype high framerate depth camera running at 375Hz, useful for solving tracking-related problems. We demonstrate our algorithmic performance, creating high resolution real-time depth maps that surpass the quality of current state of the art depth technologies, highlighting quantization-free results with reduced holes, edge fattening and other stereo-based depth artifacts.

1.117 Convolutional Two-Stream Network Fusion for Video Action Recognition

[Link to Paper]

Recent applications of Convolutional Neural Networks (ConvNets) for human action recognition in videos have proposed different solutions for incorporating the appearance and motion information. We study a number of ways of fusing ConvNet towers both spatially

and temporally in order to best take advantage of this spatio-temporal information. We make the following findings: (i) that rather than fusing at the softmax layer, a spatial and temporal network can be fused at a convolution layer without loss of performance, but with a substantial saving in parameters; (ii) that it is better to fuse such networks spatially at the last convolutional layer than earlier, and that additionally fusing at the class prediction layer can boost accuracy; finally (iii) that pooling of abstract convolutional features over spatiotemporal neighbourhoods further boosts performance. Based on these studies we propose a new ConvNet architecture for spatiotemporal fusion of video snippets, and evaluate its performance on standard benchmarks where this architecture achieves state-of-the-art results.

1.118   6D Dynamic Camera Relocalization from Single Reference Image

[Link to Paper]

Dynamic relocalization of 6D camera pose from single reference image is a costly and challenging task that requires delicate hand-eye calibration and precision positioning platform to do 3D mechanical rotation and translation. In this paper, we show that high-quality camera relocalization can be achieved in a much less expensive way. Based on inexpensive platform with unreliable absolute repositioning accuracy (ARA), we propose a hand-eye calibration free strategy to actively relocate camera into the same 6D pose that produces the input reference image, by sequentially correcting 3D relative rotation and translation. We theoretically prove that, by this strategy, both rotational and translational relative pose can be effectively reduced to zero, with bounded unknown hand-eye pose displacement. To conquer 3D rotation and translation ambiguity, this theoretical strategy is further revised to a practical relocalization algorithm with faster convergence rate and more reliability by jointly adjusting 3D relative rotation and translation. Extensive experiments validate the effectiveness and superior accuracy of the proposed approach on laboratory tests and challenging real-world applications.

1.119   Interactive Segmentation on RGBD Images via Cue Selection

[Link to Paper]

Interactive image segmentation is an important problem in computer vision with many applications including image editing, object recognition and image retrieval. Most existing interactive segmentation methods only operate on color images. Until recently, very few works have been proposed to leverage depth information from low-cost sensors to improve interactive segmentation. While these methods achieve better results than color-based methods, they are still limited in either using depth as an additional color channel or simply

combining depth with color in a linear way. We propose a novel interactive segmentation algorithm which can incorporate multiple feature cues like color, depth, and normals in an unified graph cut framework to leverage these cues more effectively. A key contribution of our method is that it automatically selects a single cue to be used at each pixel, based on the intuition that only one cue is necessary to determine the segmentation label locally. This is achieved by optimizing over both segmentation labels and cue labels, using terms designed to decide where both the segmentation and label cues should change. Our algorithm thus produces not only the segmentation mask but also a cue label map that indicates where each cue contributes to the final result. Extensive experiments on five large scale RGBD datasets show that our proposed algorithm performs significantly better than both other color-based and RGBD based algorithms in reducing the amount of user inputs as well as increasing segmentation accuracy.

1.120 Local Background Enclosure for RGB-D Salient Object Detection

[Link to Paper]

Recent work in salient object detection has considered the incorporation of depth cues from RGB-D images. In most cases, depth contrast is used as the main feature. However, areas of high contrast in background regions cause false positives for such methods, as the background frequently contains regions that are highly variable in depth. Here, we propose a novel RGB-D saliency feature. Local Background Enclosure (LBE) captures the spread of angular directions which are background with respect to the candidate region and the object that it is part of. We show that our feature improves over state-of-the-art RGB-D saliency approaches as well as RGB methods on the RGBD1000 and NJUDS2000 datasets.

1.121 Pairwise Linear Regression Classification for Image Set Retrieval

[Link to Paper]

This paper proposes the pairwise linear regression classification (PLRC) for image set retrieval. In PLRC, we first define a new concept of the unrelated subspace and introduce two strategies to constitute the unrelated subspace. In order to increase the information of maximizing the query set and the unrelated image set, we introduce a combination metric for two new classifiers based on two constitution strategies of the unrelated subspace. Extensive experiments on six well-known databases prove that the performance of PLRC is better than that of DLRC and several state-of-theart classifiers for different vision recognition tasks: clusterbased face recognition, video-based face recognition, object recognition and action recognition.

1.122 Discriminative Hierarchical Rank Pooling for Activity Recognition

[Link to Paper]

We present hierarchical rank pooling, a video sequence encoding method for activity recognition. It consists of a network of rank pooling functions which captures the dynamics of rich convolutional neural network features within a video sequence. By stacking non-linear feature functions and rank pooling over one another, we obtain a high capacity dynamic encoding mechanism, which is used for action recognition. We present a method for jointly learning the video representation and activity classifier parameters. Our method obtains state-of-the art results on three important activity recognition benchmarks: 76.7% on Hollywood2, 66.9% on HMDB51 and, 91.4% on UCF101. Figure 1: Illustration of hierarchical rank pooling for encoding the temporal dynamics of a video sequence.

1.123 Structured Prediction of Unobserved Voxels From a Single Depth Image

[Link to Paper]

Building a complete 3D model of a scene, given only a single depth image, is underconstrained. To gain a full volumetric model, one needs either multiple views, or a single view together with a library of unambiguous 3D models that will fit the shape of each individual object in the scene. We hypothesize that objects of dissimilar semantic classes often share similar 3D shape components, enabling a limited dataset to model the shape of a wide range of objects, and hence estimate their hidden geometry. Exploring this hypothesis, we propose an algorithm that can complete the unobserved geometry of tabletop-sized objects, based on a supervised model trained on already available volumetric elements. Our model maps from a local observation in a single depth image to an estimate of the surface shape in the surrounding neighborhood. We validate our approach both qualitatively and quantitatively on a range of indoor object collections and challenging real scenes.

1.124 DeepStereo: Learning to Predict New Views from the World's Imagery

[Link to Paper]

Deep networks have recently enjoyed enormous success when applied to recognition and classification problems in computer vision [22, 33], but their use in graphics problems has been limited ([23, 7] are notable recent exceptions). In this work, we present a novel deep architecture that performs new view synthesis directly from pixels, trained from a

large number of posed image sets. In contrast to traditional approaches, which consist of multiple complex stages of processing, each of which requires careful tuning and can fail in unexpected ways, our system is trained end-to-end. The pixels from neighboring views of a scene are presented to the network, which then directly produces the pixels of the unseen view. The benefits of our approach include generality (we only require posed image sets and can easily apply our method to different domains), and high quality results on traditionally difficult scenes. We believe this is due to the end-to-end nature of our system, which is able to plausibly generate pixels according to color, depth, and texture priors learnt automatically from the training data. We show view interpolation results on imagery from the KITTI dataset [12], from data from [1] as well as on Google Street View images. To our knowledge, our work is the first to apply deep learning to the problem of new view synthesis from sets of real-world, natural imagery.

1.125 3D Shape Attributes

[Link to Paper]

In this paper we investigate 3D attributes as a means to understand the shape of an object in a single image. To this end, we make a number of contributions: (i) we introduce and define a set of 3D Shape attributes, including planarity, symmetry and occupied space; (ii) we show that such properties can be successfully inferred from a single image using a Convolutional Neural Network (CNN); (iii) we introduce a 143K image dataset of sculptures with 2197 works over 242 artists for training and evaluating the CNN; (iv) we show that the 3D attributes trained on this dataset generalize to images of other (non-sculpture) object classes; and furthermore (v) we show that the CNN also provides a shape embedding that can be used to match previously unseen sculptures largely independent of viewpoint.

1.126 Optimal Relative Pose with Unknown Correspondences

[Link to Paper]

Previous work on estimating the epipolar geometry of two views relies on being able to reliably match feature points based on appearance. In this paper, we go one step further and show that it is feasible to compute both the epipolar geometry and the correspondences at the same time based on geometry only. We do this in a globally optimal manner. Our approach is based on an efficient branch and bound technique in combination with bipartite matching to solve the correspondence problem. We rely on several recent works to obtain good bounding functions to battle the combinatorial explosion of possible matchings. It is experimentally demonstrated that more difficult cases can be handled and that more inlier correspondences can be obtained by being less restrictive in the matching phase.

1.127 Split and Match: Example-based Adaptive Patch Sampling for Unsupervised Style Transfer

[Link to Paper]

This paper presents a novel unsupervised method to transfer the style of an example image to a source image. The complex notion of image style is here considered as a local texture transfer, eventually coupled with a global color transfer. For the local texture transfer, we propose a new method based on an adaptive patch partition that captures the style of the example image and preserves the structure of the source image. More precisely, this example-based partition predicts how well a source patch matches an example patch. Results on various images show that our method outperforms the most recent techniques.

1.128 A weighted variational model for simultaneous reflectance and illumination estimation

[Link to Paper]

We propose a weighted variational model to estimate both the reflectance and the illumination from an observed image. We show that, though it is widely adopted for ease of modeling, the log-transformed image for this task is not ideal. Based on the previous investigation of the logarithmic transformation, a new weighted variational model is proposed for better prior representation, which is imposed in the regularization terms. Different from conventional variational models, the proposed model can preserve the estimated reflectance with more details. Moreover, the proposed model can suppress noise to some extent. An alternating minimization scheme is adopted to solve the proposed model. Experimental results demonstrate the effectiveness of the proposed model with its algorithm. Compared with other variational methods, the proposed method yields comparable or better results on both subjective and objective assessments.

1.129 Exploiting Spectral-Spatial Correlation for Coded Hyperspectral Image Restoration

[Link to Paper]

Conventional scanning and multiplexing techniques for hyperspectral imaging suffer from limited temporal and/or spatial resolution. To resolve this issue, coding techniques are becoming increasingly popular in developing snapshot systems for high-resolution hyperspectral imaging. For such systems, it is a critical task to accurately restore the 3D hy-

perspectral image from its corresponding coded 2D image. In this paper, we propose an effective method for coded hyperspectral image restoration, which exploits extensive structure sparsity in the hyperspectral image. Specifically, we simultaneously explore spectral and spatial correlation via low-rank regularizations, and formulate the restoration problem into a variational optimization model, which can be solved via an iterative numerical algorithm. Experimental results using both synthetic data and real images show that the proposed method can significantly outperform the state-of-the-art methods on several popular coding-based hyperspectral imaging systems.

1.130 Occlusion Boundary Detection via Deep Exploration of Context

[Link to Paper]

Occlusion boundaries contain rich perceptual information about the underlying scene structure. They also provide important cues in many visual perception tasks such as scene understanding, object recognition, and segmentation. In this paper, we improve occlusion boundary detection via enhanced exploration of contextual information (e.g., local structural boundary patterns, observations from surrounding regions, and temporal context), and in doing so develop a novel approach based on convolutional neural networks (CNNs) and conditional random fields (CRFs). Experimental results demonstrate that our detector significantly outperforms the state-of-the-art (e.g., improving the F-measure from 0.62 to 0.71 on the commonly used CMU benchmark). Last but not least, we empirically assess the roles of several important components of the proposed detector, so as to validate the rationale behind this approach.

1.131 Semi-supervised Vocabulary-informed Learning

[Link to Paper]

Despite significant progress in object categorization, in recent years, a number of important challenges remain; mainly, ability to learn from limited labeled data and ability to recognize object classes within large, potentially open, set of labels. Zero-shot learning is one way of addressing these challenges, but it has only been shown to work with limited sized class vocabularies and typically requires separation between supervised and unsupervised classes, allowing former to inform the latter but not vice versa. We propose the notion of semi-supervised vocabulary-informed learning to alleviate the above mentioned challenges and address problems of supervised, zero-shot and open set recognition using a unified framework. Specifically, we propose a maximum margin framework for semantic manifoldbased recognition that incorporates distance constraints from (both supervised and unsupervised) vocabulary atoms, ensuring that labeled samples are projected closest

63

to their correct prototypes, in the embedding space, than to others. We show that resulting model shows improvements in supervised, zero-shot, and large open set recognition, with up to 310K class vocabulary on AwA and ImageNet datasets.

1.132   Symmetry reCAPTCHA

[Link to Paper]

This paper is a reaction to the poor performance of symmetry detection algorithms on real-world images, benchmarked since CVPR 2011. Our systematic study reveals significant difference between human labeled (reflection and rotation) symmetries on photos and the output of computer vision algorithms on the same photo set. We exploit this human-machine symmetry perception gap by proposing a novel symmetry-based Turing test. By leveraging a comprehensive user interface, we collected more than 78,000 symmetry labels from 400 Amazon Mechanical Turk raters on 1,200 photos from the Microsoft COCO dataset. Using a set of ground-truth symmetries automatically generated from noisy human labels, the effectiveness of our work is evidenced by a separate test where over 96% success rate is achieved. We demonstrate statistically significant outcomes for using symmetry perception as a powerful, alternative, image-based reCAPTCHA.

1.133   Learning Local Image Descriptors with Deep Siamese and Triplet Convolutional Networks by Minimizing Global Loss Functions

[Link to Paper]

Recent innovations in training deep convolutional neural network (ConvNet) models have motivated the design of new methods to automatically learn local image descriptors. The latest deep ConvNets proposed for this task consist of a siamese network that is trained by penalising misclassification of pairs of local image patches. Current results from machine learning show that replacing this siamese by a triplet network can improve the classification accuracy in several problems, but this has yet to be demonstrated for local image descriptor learning. Moreover, current siamese and triplet networks have been trained with stochastic gradient descent that computes the gradient from individual pairs or triplets of local image patches, which can make them prone to overfitting. In this paper, we first propose the use of triplet networks for the problem of local image descriptor learning. Furthermore, we also propose the use of a global loss that minimises the overall classification error in the training set, which can improve the generalisation capability of the model. Using the UBC benchmark dataset for comparing local image descriptors, we show that the triplet network produces a more accurate embedding than the siamese network in terms of the UBC dataset errors. Moreover, we also demonstrate that a combination of the triplet and

global losses produces the best embedding in the field, using this triplet network. Finally, we also show that the use of the central-surround siamese network trained with the global loss produces the best result of the field on the UBC dataset.

1.134   PatchBatch: a Batch Augmented Loss for Optical Flow

[Link to Paper]

We propose a new pipeline for optical flow computation, based on Deep Learning techniques. We suggest using a Siamese CNN to independently, and in parallel, compute the descriptors of both images. The learned descriptors are then compared efficiently using the L2 norm and do not require network processing of patch pairs. The success of the method is based on an innovative loss function that computes higher moments of the loss distributions for each training batch. Combined with an Approximate Nearest Neighbor patch matching method and a flow interpolation technique, state of the art performance is obtained on the most challenging and competitive optical flow benchmarks.

1.135   Virtual Worlds as Proxy for Multi-Object Tracking Analysis

[Link to Paper]

Modern computer vision algorithms typically require expensive data acquisition and accurate manual labeling. In this work, we instead leverage the recent progress in computer graphics to generate fully labeled, dynamic, and photo-realistic proxy virtual worlds. We propose an efficient real-to-virtual world cloning method, and validate our approach by building and publicly releasing a new video dataset, called "Virtual KITTI" 1, automatically labeled with accurate ground truth for object detection, tracking, scene and instance segmentation, depth, and optical flow. We provide quantitative experimental evidence suggesting that (i) modern deep learning algorithms pre-trained on real data behave similarly in real and virtual worlds, and (ii) pre-training on virtual data improves performance. As the gap between real and virtual worlds is small, virtual worlds enable measuring the impact of various weather and imaging conditions on recognition performance, all other things being equal. We show these factors may affect drastically otherwise high-performing deep models for tracking.

1.136   Just look at the image: viewpoint-specific surface normal prediction for improved multi-view reconstruction

[Link to Paper]

We present a multi-view reconstruction method that combines conventional multi-view stereo (MVS) with appearance-based normal prediction, to obtain dense and accurate 3D surface models. Reliable surface normals reconstructed from multi-view correspondence serve as training data for a convolutional neural network (CNN), which predicts continuous normal vectors from raw image patches. By training from known points in the same image, the prediction is specifically tailored to the materials and lighting conditions of the particular scene, as well as to the precise camera viewpoint. It is therefore a lot easier to learn than generic single-view normal estimation. The estimated normal maps, together with the known depth values from MVS, are integrated to dense depth maps, which in turn are fused into a 3D model. Experiments on the DTU dataset show that our method delivers 3D reconstructions with the same accuracy as MVS, but with significantly higher completeness.

1.137 Learning Attributes Equals Multi-Source Domain Generalization

[Link to Paper]

Attributes possess appealing properties and benefit many computer vision problems, such as object recognition, learning with humans in the loop, and image retrieval. Whereas the existing work mainly pursues utilizing attributes for various computer vision problems, we contend that the most basic problem–how to accurately and robustly detect attributes from images–has been left under explored. Especially, the existing work rarely explicitly tackles the need that attribute detectors should generalize well across different categories, including those previously unseen. Noting that this is analogous to the objective of multi-source domain generalization, if we treat each category as a domain, we provide a novel perspective to attribute detection and propose to gear the techniques in multi-source domain generalization for the purpose of learning cross-category generalizable attribute detectors. We validate our understanding and approach with extensive experiments on four challenging datasets and three different problems.

1.138 You Lead, We Exceed: Labor-Free Video Concept Learning by Jointly Exploiting Web Videos and Images

[Link to Paper]

(Abstract could not be extracted from PDF)

1.139 Compact Bilinear Pooling

[Link to Paper]

Bilinear models has been shown to achieve impressive performance on a wide range of visual tasks, such as semantic segmentation, fine grained recognition and face recognition. However, bilinear features are high dimensional, typically on the order of hundreds of thousands to a few million, which makes them impractical for subsequent analysis. We propose two compact bilinear representations with the same discriminative power as the full bilinear representation but with only a few thousand dimensions. Our compact representations allow back-propagation of classification errors enabling an end-to-end optimization of the visual recognition system. The compact bilinear representations are derived through a novel kernelized analysis of bilinear pooling which provide insights into the discriminative power of bilinear pooling, and a platform for further research in compact pooling methods. Experimentation illustrate the utility of the proposed representations for image classification and few-shot learning across several datasets.

1.140 Parametric Object Motion from Blur

[Link to Paper]

Motion blur can adversely affect a number of vision tasks, hence it is generally considered a nuisance. We instead treat motion blur as a useful signal that allows to compute the motion of objects from a single image. Drawing on the success of joint segmentation and parametric motion models in the context of optical flow estimation, we propose a parametric object motion model combined with a segmentation mask to exploit localized, non-uniform motion blur. Our parametric image formation model is differentiable w.r.t. the motion parameters, which enables us to generalize marginal-likelihood techniques from uniform blind deblurring to localized, non-uniform blur. A two-stage pipeline, first in derivative space and then in image space, allows to estimate both parametric object motion as well as a motion segmentation from a single image alone. Our experiments demonstrate its ability to cope with very challenging cases of object motion blur.

1.141 Image Style Transfer Using Convolutional Neural Networks

[Link to Paper]

Rendering the semantic content of an image in different styles is a difficult image processing task. Arguably, a major limiting factor for previous approaches has been the lack of image representations that explicitly represent semantic information and, thus, allow

to separate image content from style. Here we use image representations derived from Convolutional Neural Networks optimised for object recognition, which make high level image information explicit. We introduce A Neural Algorithm of Artistic Style that can separate and recombine the image content and style of natural images. The algorithm allows us to produce new images of high perceptual quality that combine the content of an arbitrary photograph with the appearance of numerous wellknown artworks. Our results provide new insights into the deep image representations learned by Convolutional Neural Networks and demonstrate their potential for high level image synthesis and manipulation.

1.142 Robust 3D Hand Pose Estimation in Single Depth Images: from Single-View CNN to Multi-View CNNs

[Link to Paper]

Articulated hand pose estimation plays an important role in human-computer interaction. Despite the recent progress, the accuracy of existing methods is still not satisfactory, partially due to the difficulty of embedded highdimensional and non-linear regression problem. Different from the existing discriminative methods that regress for the hand pose with a single depth image, we propose to first project the query depth image onto three orthogonal planes and utilize these multi-view projections to regress for 2D heat-maps which estimate the joint positions on each plane. These multi-view heat-maps are then fused to produce final 3D hand pose estimation with learned pose priors. Experiments show that the proposed method largely outperforms state-of-the-art on a challenging dataset. Moreover, a cross-dataset experiment also demonstrates the good generalization ability of the proposed method. Figure 1: Overview of our proposed multi-view regression framework. We generate heat-maps for three views by projecting 3D points onto three orthogonal planes. Three CNNs are trained in parallel to map each view's projected image to its corresponding heat-maps, which are then fused together to estimate 3D hand joint locations.

1.143 LocNet: Improving Localization Accuracy for Object Detection

[Link to Paper]

We propose a novel object localization methodology with the purpose of boosting the localization accuracy of stateof-the-art object detection systems. Our model, given a search region, aims at returning the bounding box of an object of interest inside this region. To accomplish its goal, it relies on assigning conditional probabilities to each row and column of this region, where these probabilities provide useful information regarding the location of the boundaries of the object inside the search region and allow the accurate inference

of the object bounding box under a simple probabilistic framework. For implementing our localization model, we make use of a convolutional neural network architecture that is properly adapted for this task, called LocNet. We show experimentally that LocNet achieves a very significant improvement on the mAP for high IoU thresholds on PASCAL VOC2007 test set and that it can be very easily coupled with recent stateof-the-art object detection systems, helping them to boost their performance. Finally, we demonstrate that our detection approach can achieve high detection accuracy even when it is given as input a set of sliding windows, thus proving that it is independent of box proposal methods.

1.144   Gravitational Approach for Point Set Registration

[Link to Paper]

In this paper a new astrodynamics inspired rigid point set registration algorithm is introduced – the Gravitational Approach (GA). We formulate point set registration as a modified N -body problem with additional constraints and obtain an algorithm with unique properties which is fully scalable with the number of processing cores. In GA, a template point set moves in a viscous medium under gravitational forces induced by a reference point set. Pose updates are completed by numerically solving the differential equations of Newtonian mechanics. We discuss techniques for efficient implementation of the new algorithm and evaluate it on several synthetic and real-world scenarios. GA is compared with the widely used Iterative Closest Point and the state of the art rigid Coherent Point Drift algorithms. Experiments evidence that the new approach is robust against noise and can handle challenging scenarios with structured outliers. Figure 1: Point set registration with the Gravitational Approach: template moves in the gravitational field induced by the reference. Coordinates of the template points are individually updated by solving equations of particle motion in a viscous medium, whereupon rigidity constraints are applied. Left: initial misalignment of the helix point sets [8] and the induced gravitational field; right: registration result after 150 iterations.

1.145   Blind Image Deconvolution by Automatic Gradient Activation

[Link to Paper]

Blind image deconvolution is an ill-posed inverse problem which is often addressed through the application of appropriate prior. Although some priors are informative in general, many images do not strictly conform to this, leading to degraded performance in the kernel estimation. More critically, real images may be contaminated by nonuniform noise such as saturation and outliers. Methods for removing specific image areas based on some priors have been proposed, but they operate either manually or by defining fixed criteria. We

show here that a subset of the image gradients are adequate to estimate the blur kernel robustly, no matter the gradient image is sparse or not. We thus introduce a gradient activation method to automatically select a subset of gradients of the latent image in a cutting-plane-based optimization scheme for kernel estimation. No extra assumption is used in our model, which greatly improves the accuracy and flexibility. More importantly, the proposed method affords great convenience for handling noise and outliers. Experiments on both synthetic data and real-world images demonstrate the effectiveness and robustness of the proposed method in comparison with the state-of-the-art methods.

1.146 Cross Modal Distillation for Supervision Transfer

[Link to Paper]

In this work we propose a technique that transfers supervision between images from different modalities. We use learned representations from a large labeled modality as supervisory signal for training representations for a new unlabeled paired modality. Our method enables learning of rich representations for unlabeled modalities and can be used as a pre-training procedure for new modalities with limited labeled data. We transfer supervision from labeled RGB images to unlabeled depth and optical flow images and demonstrate large improvements for both these cross modal supervision transfers.

1.147 Synthetic Data for Text Localisation in Natural Images

[Link to Paper]

In this paper we introduce a new method for text detection in natural images. The method comprises two contributions: First, a fast and scalable engine to generate synthetic images of text in clutter. This engine overlays synthetic text to existing background images in a natural way, accounting for the local 3D scene geometry. Second, we use the synthetic images to train a Fully-Convolutional Regression Network (FCRN) which efficiently performs text detection and bounding-box regression at all locations and multiple scales in an image. We discuss the relation of FCRN to the recently-introduced YOLO detector, as well as other end-toend object detection systems based on deep learning. The resulting detection network significantly out performs current methods for text detection in natural images, achieving an F-measure of 84.2% on the standard ICDAR 2013 benchmark. Furthermore, it can process 15 images per second on a GPU.

1.148 Pull the Plug? Predicting If Computers or Humans Should Segment Images

[Link to Paper]

Foreground object segmentation is a critical step for many image analysis tasks. While automated methods can produce high-quality results, their failures disappoint users in need of practical solutions. We propose a resource allocation framework for predicting how best to allocate a fixed budget of human annotation effort in order to collect higher quality segmentations for a given batch of images and automated methods. The framework is based on a proposed prediction module that estimates the quality of given algorithm-drawn segmentations. We demonstrate the value of the framework for two novel tasks related to "pulling the plug" on computer and human annotators. Specifically, we implement two systems that automatically decide, for a batch of images, when to replace 1) humans with computers to create coarse segmentations required to initialize segmentation tools and 2) computers with humans to create final, fine-grained segmentations. Experiments demonstrate the advantage of relying on a mix of human and computer efforts over relying on either resource alone for segmenting objects in three diverse datasets representing visible, phase contrast microscopy, and fluorescence microscopy images.

1.149 Video2GIF: Automatic Generation of Animated GIFs from Video

[Link to Paper]

We introduce the novel problem of automatically generating animated GIFs from video. GIFs are short looping video with no sound, and a perfect combination between image and video that really capture our attention. GIFs tell a story, express emotion, turn events into humorous moments, and are the new wave of photojournalism. We pose the question: Can we automate the entirely manual and elaborate process of GIF creation by leveraging the plethora of user generated GIF content? We propose a Robust Deep RankNet that, given a video, generates a ranked list of its segments according to their suitability as GIF. We train our model to learn what visual content is often selected for GIFs by using over 100K user generated GIFs and their corresponding video sources. We effectively deal with the noisy web data by proposing a novel adaptive Huber loss in the ranking formulation. We show that our approach is robust to outliers and picks up several patterns that are frequently present in popular animated GIFs. On our new large-scale benchmark dataset, we show the advantage of our approach over several state-of-the-art methods.

1.150 High-quality Depth from Uncalibrated Small Motion Clip

[Link to Paper]

We propose a novel approach that generates a highquality depth map from a set of images captured with a small viewpoint variation, namely small motion clip. As opposed to prior methods that recover scene geometry and camera motions using pre-calibrated cameras, we introduce a self-calibrating bundle adjustment tailored for small motion. This allows our dense stereo algorithm to produce a high-quality depth map for the user without the need for camera calibration. In the dense matching, the distributions of intensity profiles are analyzed to leverage the benefit of having negligible intensity changes within the scene due to the minuscule variation in viewpoint. The depth maps obtained by the proposed framework show accurate and extremely fine structures that are unmatched by previous literature under the same small motion configuration.

1.151 Contour detection in unstructured 3D point clouds

[Link to Paper]

We describe a method to automatically detect contours, i.e. lines along which the surface orientation sharply changes, in large-scale outdoor point clouds. Contours are important intermediate features for structuring point clouds and converting them into high-quality surface or solid models, and are extensively used in graphics and mapping applications. Yet, detecting them in unstructured, inhomogeneous point clouds turns out to be surprisingly difficult, and existing line detection algorithms largely fail. We approach contour extraction as a two-stage discriminative learning problem. In the first stage, a contour score for each individual point is predicted with a binary classifier, using a set of features extracted from the point's neighborhood. The contour scores serve as a basis to construct an overcomplete graph of candidate contours. The second stage selects an optimal set of contours from the candidates. This amounts to a further binary classification in a higher-order MRF, whose cliques encode a preference for connected contours and penalize loose ends. The method can handle point clouds ¿ 107 points in a couple of minutes, and vastly outperforms a baseline that performs Canny-style edge detection on a range image representation of the point cloud.

1.152 Proposal Flow

[Link to Paper]

Finding image correspondences remains a challenging problem in the presence of intra-class variations and large changes in scene layout. Semantic flow methods are designed to handle images depicting different instances of the same object or scene category. We introduce a novel approach to semantic flow, dubbed proposal flow, that establishes reliable correspondences using object proposals. Unlike prevailing semantic flow approaches that operate on pixels or regularly sampled local regions, proposal flow benefits from the characteristics of modern object proposals, that exhibit high repeatability at multiple scales, and can take advantage of both local and geometric consistency constraints among proposals. We also show that proposal flow can effectively be transformed into a conventional dense flow field. We introduce a new dataset that can be used to evaluate both general semantic flow techniques and region-based approaches such as proposal flow. We use this benchmark to compare different matching algorithms, object proposals, and region features within proposal flow, to the state of the art in semantic flow. This comparison, along with experiments on standard datasets, demonstrates that proposal flow significantly outperforms existing semantic flow methods in various settings.

1.153 Mirror Surface Reconstruction under an Uncalibrated Camera

[Link to Paper]

This paper addresses the problem of mirror surface reconstruction, and a solution based on observing the reflections of a moving reference plane on the mirror surface is proposed. Unlike previous approaches which require tedious work to calibrate the camera, our method can recover both the camera intrinsics and extrinsics together with the mirror surface from reflections of the reference plane under at least three unknown distinct poses. Our previous work has demonstrated that 3D poses of the reference plane can be registered in a common coordinate system using reflection correspondences established across images. This leads to a bunch of registered 3D lines formed from the reflection correspondences. Given these lines, we first derive an analytical solution to recover the camera projection matrix through estimating the line projection matrix. We then optimize the camera projection matrix by minimizing reprojection errors computed based on a cross-ratio formulation. The mirror surface is finally reconstructed based on the optimized cross-ratio constraint. Experimental results on both synthetic and real data are presented, which demonstrate the feasibility and accuracy of our method.

1.154 Understanding Real World Indoor Scenes With Synthetic Data

[Link to Paper]

Scene understanding is a prerequisite to many high level tasks for any automated intelligent machine operating in real world environments. Recent attempts with supervised learning have shown promise in this direction but also highlighted the need for enormous quantity of supervised data – performance increases in proportion to the amount of data used. However, this quickly becomes prohibitive when considering the manual labour needed to collect such data. In this work, we focus our attention on depth based semantic per-pixel labelling as a scene understanding problem and show the potential of computer graphics to generate virtually unlimited labelled data from synthetic 3D scenes. By carefully synthesizing training data with appropriate noise models we show comparable performance to state-of-theart RGBD systems on NYUv2 dataset despite using only depth data as input and set a benchmark on depth-based segmentation on SUN RGB-D dataset.

1.155 Recurrent Attention Models for Depth-Based Person Identification

[Link to Paper]

We present an attention-based model that reasons on human body shape and motion dynamics to identify individuals in the absence of RGB information, hence in the dark. Our approach leverages unique 4D spatio-temporal signatures to address the identification problem across days. Formulated as a reinforcement learning task, our model is based on a combination of convolutional and recurrent neural networks with the goal of identifying small, discriminative regions indicative of human identity. We demonstrate that our model produces state-of-the-art results on several published datasets given only depth images. We further study the robustness of our model towards viewpoint, appearance, and volumetric changes. Finally, we share insights gleaned from interpretable 2D, 3D, and 4D visualizations of our model's spatio-temporal attention.

1.156 Identifying Good Training Data for Self-Supervised Free Space Estimation

[Link to Paper]

This paper proposes a novel technique to extract training data from free space in a scene using a stereo camera. The proposed technique exploits the projection of planes in the v-disparity image paired with Bayesian linear regression to reliably identify training image pixels belonging to free space in a scene. Unlike other methods in the literature, the algorithm does not require any prior training, has only one free parameter, and is shown to provide consistent results over a variety of terrains without the need for any manual tuning. The proposed method is compared to two other data extraction methods from the

literature. Results of Support Vector classifiers using training data extracted by the proposed technique are superior in terms of quality and consistency of free space estimation. Furthermore, the computation time required by the proposed technique is shown to be smaller and more consistent than that of other training data extraction methods.

1.157   When VLAD met Hilbert

[Link to Paper]

In many challenging visual recognition tasks where training data is limited, Vectors of Locally Aggregated Descriptors (VLAD) have emerged as powerful image/video representations that compete with or outperform state-ofthe-art approaches. In this paper, we address two fundamental limitations of VLAD: its requirement for the local descriptors to have vector form and its restriction to linear classifiers due to its high-dimensionality. To this end, we introduce a kernelized version of VLAD. This not only lets us inherently exploit more sophisticated classification schemes, but also enables us to efficiently aggregate nonvector descriptors (e.g., manifold-valued data) in the VLAD framework. Furthermore, we propose an approximate formulation that allows us to accelerate the coding process while still benefiting from the properties of kernel VLAD. Our experiments demonstrate the effectiveness of our approach at handling manifold-valued data, such as covariance descriptors, on several classification tasks. Our results also evidence the benefits of our nonlinear VLAD descriptors against the linear ones in Euclidean space using several standard benchmark datasets.

1.158   FANNG: Fast Approximate Nearest Neighbour Graphs

[Link to Paper]

We present a new method for approximate nearest neighbour search on large datasets of high dimensional feature vectors, such as SIFT or GIST descriptors. Our approach constructs a directed graph that can be efficiently explored for nearest neighbour queries. Each vertex in this graph represents a feature vector from the dataset being searched. The directed edges are computed by exploiting the fact that, for these datasets, the intrinsic dimensionality of the local manifold-like structure formed by the elements of the dataset is significantly lower than the embedding space. We also provide an efficient search algorithm that uses this graph to rapidly find the nearest neighbour to a query with high probability. We show how the method can be adapted to give a strong guarantee of 100% recall where the query is within a threshold distance of its nearest neighbour. We demonstrate that our method is significantly more efficient than existing state of the art methods. In particular, our GPU implementation can deliver 90% recall for queries on a data set of 1

million SIFT descriptors at a rate of over 1.2 million queries per second on a Titan X. Finally we also demonstrate how our method scales to datasets of 5M and 20M entries.

1.159 Learning Temporal Regularity in Video Sequences

[Link to Paper]

Perceiving meaningful activities in a long video sequence is a challenging problem due to ambiguous definition of 'meaningfulness' as well as clutters in the scene. We approach this problem by learning a generative model for regular motion patterns (termed as regularity) using multiple sources with very limited supervision. Specifically, we propose two methods that are built upon the autoencoders for their ability to work with little to no supervision. We first leverage the conventional handcrafted spatio-temporal local features and learn a fully connected autoencoder on them. Second, we build a fully convolutional feed-forward autoencoder to learn both the local features and the classifiers as an end-to-end learning framework. Our model can capture the regularities from multiple datasets. We evaluate our methods in both qualitative and quantitative ways showing the learned regularity of videos in various aspects and demonstrating competitive performance on anomaly detection datasets as an application.

1.160 Learning to Co-Generate Object Proposals with a Deep Structured Network

[Link to Paper]

Generating object proposals has become a key component of modern object detection pipelines. However, most existing methods generate the object candidates independently of each other. In this paper, we present an approach to co-generating object proposals in multiple images, thus leveraging the collective power of multiple object candidates. In particular, we introduce a deep structured network that jointly predicts the objectness scores and the bounding box locations of multiple object candidates. Our deep structured network consists of a fully-connected Conditional Random Field built on top of a set of deep Convolutional Neural Networks, which learn features to model both the individual object candidates and the similarity between multiple candidates. To train our deep structured network, we develop an end-to-end learning algorithm that, by unrolling the CRF inference procedure, lets us backpropagate the loss gradient throughout the entire structured network. We demonstrate the effectiveness of our approach on two benchmark datasets, showing significant improvement over state-of-the-art object proposal algorithms.

1.161 Deep Residual Learning for Image Recognition

[Link to Paper]

Deeper neural networks are more difficult to train. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. We explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. We provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth. On the ImageNet dataset we evaluate residual nets with a depth of up to 152 layers–8 deeper than VGG nets [40] but still having lower complexity. An ensemble of these residual nets achieves 3.57% error on the ImageNet test set. This result won the 1st place on the ILSVRC 2015 classification task. We also present analysis on CIFAR-10 with 100 and 1000 layers. The depth of representations is of central importance for many visual recognition tasks. Solely due to our extremely deep representations, we obtain a 28% relative improvement on the COCO object detection dataset. Deep residual nets are foundations of our submissions to ILSVRC & COCO 2015 competitions1, where we also won the 1st places on the tasks of ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation.

1.162 Exemplar-Driven Top-Down Saliency Detection via Deep Association

[Link to Paper]

Top-down saliency detection is a knowledge-driven search task. While some previous methods aim to learn this ”knowledge” from category-specific data, others transfer existing annotations in a large dataset through appearance matching. In contrast, we propose in this paper a locateby-exemplar strategy. This approach is challenging, as we only use a few exemplars (up to 4) and the appearances among the query object and the exemplars can be very different. To address it, we design a two-stage deep model to learn the intra-class association between the exemplars and query objects. The first stage is for learning object-to-object association, and the second stage is to learn background discrimination. Extensive experimental evaluations show that the proposed method outperforms different baselines and the category-specific models. In addition, we explore the influence of exemplar properties, in terms of exemplar number and quality. Furthermore, we show that the learned model is a universal model and offers great generalization to unseen objects.

1.163 Convolutional Networks for Shape from Light Field

[Link to Paper]

Convolutional Neural Networks (CNNs) have recently been successfully applied to various Computer Vision (CV) applications. In this paper we utilize CNNs to predict depth information for given Light Field (LF) data. The proposed method learns an end-to-end mapping between the 4D light field and a representation of the corresponding 4D depth field in terms of 2D hyperplane orientations. The obtained prediction is then further refined in a post processing step by applying a higher-order regularization. Existing LF datasets are not sufficient for the purpose of the training scheme tackled in this paper. This is mainly due to the fact that the ground truth depth of existing datasets is inaccurate and/or the datasets are limited to a small number of LFs. This made it necessary to generate a new synthetic LF dataset, which is based on the raytracing software POV-Ray. This new dataset provides floating point accurate ground truth depth fields, and due to a random scene generator the dataset can be scaled as required.

1.164 Fast Temporal Activity Proposals for Efficient Detection of Human Actions in Untrimmed Videos

[Link to Paper]

In many large-scale video analysis scenarios, one is interested in localizing and recognizing human activities that occur in short temporal intervals within long untrimmed videos. Current approaches for activity detection still struggle to handle large-scale video collections and the task remains relatively unexplored. This is in part due to the computational complexity of current action recognition approaches and the lack of a method that proposes fewer intervals in the video, where activity processing can be focused. In this paper, we introduce a proposal method that aims to recover temporal segments containing actions in untrimmed videos. Building on techniques for learning sparse dictionaries, we introduce a learning framework to represent and retrieve activity proposals. We demonstrate the capabilities of our method in not only producing high quality proposals but also in its efficiency. Finally, we show the positive impact our method has on recognition performance when it is used for action detection, while running at 10FPS.

1.165 Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data

[Link to Paper]

While recent deep neural network models have achieved promising results on the image captioning task, they rely largely on the availability of corpora with paired image and sentence captions to describe objects in context. In this work, we propose the Deep Compositional Captioner (DCC) to address the task of generating descriptions of novel objects

which are not present in paired imagesentence datasets. Our method achieves this by leveraging large object recognition datasets and external text corpora and by transferring knowledge between semantically similar concepts. Current deep caption models can only describe objects contained in paired image-sentence corpora, despite the fact that they are pre-trained with large object recognition datasets, namely ImageNet. In contrast, our model can compose sentences that describe novel objects and their interactions with other objects. We demonstrate our model's ability to describe novel concepts by empirically evaluating its performance on MSCOCO and show qualitative results on ImageNet images of objects for which no paired image-sentence data exist. Further, we extend our approach to generate descriptions of objects in video clips. Our results show that DCC has distinct advantages over existing image and video captioning approaches for generating descriptions of new objects in context.

1.166 Shortlist Selection with Residual-Aware Distance Estimator for K-Nearest Neighbor Search

[Link to Paper]

In this paper, we introduce a novel shortlist computation algorithm for approximate, high-dimensional nearest neighbor search. Our method relies on a novel distance estimator: the residual-aware distance estimator, that accounts for the residual distances of data points to their respective quantized centroids, and uses it for accurate shortlist computation. Furthermore, we perform the residualaware distance estimation with little additional memory and computational cost through simple pre-computation methods for inverted index and multi-index schemes. Because it modifies the initial shortlist collection phase, our new algorithm is applicable to most inverted indexing methods that use vector quantization. We have tested the proposed method with the inverted index and multi-index on a diverse set of benchmarks including up to one billion data points with varying dimensions, and found that our method robustly improves the accuracy of shortlists (up to 127% relatively higher) over the state-of-the-art techniques with a comparable or even faster computational cost.

1.167 Scene recognition with CNNs: objects, scales and dataset bias

[Link to Paper]

Since scenes are composed in part of objects, accurate recognition of scenes requires knowledge about both scenes and objects. In this paper we address two related problems: 1) scale induced dataset bias in multi-scale convolutional neural network (CNN) architectures, and 2) how to combine effectively scene-centric and object-centric knowledge (i.e.

79

Places and ImageNet) in CNNs. An earlier attempt, Hybrid-CNN[23], showed that incorporating ImageNet did not help much. Here we propose an alternative method taking the scale into account, resulting in significant recognition gains. By analyzing the response of ImageNet-CNNs and Places-CNNs at different scales we find that both operate in different scale ranges, so using the same network for all the scales induces dataset bias resulting in limited performance. Thus, adapting the feature extractor to each particular scale (i.e. scale-specific CNNs) is crucial to improve recognition, since the objects in the scenes have their specific range of scales. Experimental results show that the recognition accuracy highly depends on the scale, and that simple yet carefully chosen multi-scale combinations of ImageNet-CNNs and Places-CNNs, can push the stateof-the-art recognition accuracy in SUN397 up to 66.26% (and even 70.17% with deeper architectures, comparable to human performance).

1.168   Learning with Side Information through Modality Hallucination

[Link to Paper]

We present a modality hallucination architecture for training an RGB object detection model which incorporates depth side information at training time. Our convolutional hallucination network learns a new and complementary RGB image representation which is taught to mimic convolutional mid-level features from a depth network. At test time images are processed jointly through the RGB and hallucination networks to produce improved detection performance. Thus, our method transfers information commonly extracted from depth training data to a network which can extract that information from the RGB counterpart. We present results on the standard NYUDv2 dataset and report improvement on the RGB detection task.

1.169   Recombinator Networks: Learning Coarse-to-Fine Feature Aggregation

[Link to Paper]

Deep neural networks with alternating convolutional, max-pooling and decimation layers are widely used in state of the art architectures for computer vision. Max-pooling purposefully discards precise spatial information in order to create features that are more robust, and typically organized as lower resolution spatial feature maps. On some tasks, such as whole-image classification, max-pooling derived features are well suited; however, for tasks requiring precise localization, such as pixel level prediction and segmentation, max-pooling destroys exactly the information required to perform well. Precise localization may be preserved by shallow convnets without pooling but at the expense of robustness. Can

we have our max-pooled multilayered cake and eat it too? Several papers have proposed summation and concatenation based methods for combining upsampled coarse, abstract features with finer features to produce robust pixel level predictions. Here we introduce another model – dubbed Recombinator Networks – where coarse features inform finer features early in their formation such that finer features can make use of several layers of computation in deciding how to use coarse features. The model is trained once, end-to-end and performs better than summation-based architectures, reducing the error from the previous state of the art on two facial keypoint datasets, AFW and AFLW, by 30% and beating the current state-of-the-art on 300W without using extra data. We improve performance even further by adding a denoising prediction model based on a novel convnet formulation.

1.170 Learning Transferrable Knowledge for Semantic Segmentation with Deep Convolutional Neural Network

[Link to Paper]

We propose a novel weakly-supervised semantic segmentation algorithm based on Deep Convolutional Neural Network (DCNN). Contrary to existing weakly-supervised approaches, our algorithm exploits auxiliary segmentation annotations available for different categories to guide segmentations on images with only image-level class labels. To make segmentation knowledge transferrable across categories, we design a decoupled encoder-decoder architecture with attention model. In this architecture, the model generates spatial highlights of each category presented in images using an attention model, and subsequently performs binary segmentation for each highlighted region using decoder. Combining attention model, the decoder trained with segmentation annotations in different categories boosts accuracy of weakly-supervised semantic segmentation. The proposed algorithm demonstrates substantially improved performance compared to the state-of-theart weakly-supervised techniques in PASCAL VOC 2012 dataset when our model is trained with the annotations in 60 exclusive categories in Microsoft COCO dataset.

1.171 An Egocentric Look at Video Photographer Identity

[Link to Paper]

Egocentric cameras are being worn by an increasing number of users, among them many security forces worldwide. GoPro cameras already penetrated the mass market, reporting substantial increase in sales every year. As headworn cameras do not capture the photographer, it may seem that the anonymity of the photographer is preserved even when the video is publicly distributed. We show that camera motion, as can be computed from the

egocentric video, provides unique identity information. The photographer can be reliably recognized from a few seconds of video captured when walking. The proposed method achieves more than 90% recognition accuracy in cases where the random success rate is only 3%. Applications can include theft prevention by locking the camera when not worn by its rightful owner. Searching video sharing services (e.g. YouTube) for egocentric videos shot by a specific photographer may also become possible. An important message in this paper is that photographers should be aware that sharing egocentric video will compromise their anonymity, even when their face is not visible.

1.172 Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification

[Link to Paper]

Convolutional Neural Networks (CNN) are state-of-theart models for many image classification tasks. However, to recognize cancer subtypes automatically, training a CNN on gigapixel resolution Whole Slide Tissue Images (WSI) is currently computationally impossible. The differentiation of cancer subtypes is based on cellular-level visual features observed on image patch scale. Therefore, we argue that in this situation, training a patch-level classifier on image patches will perform better than or similar to an image-level classifier. The challenge becomes how to intelligently combine patch-level classification results and model the fact that not all patches will be discriminative. We propose to train a decision fusion model to aggregate patch-level predictions given by patch-level CNNs, which to the best of our knowledge has not been shown before. Furthermore, we formulate a novel Expectation-Maximization (EM) based method that automatically locates discriminative patches robustly by utilizing the spatial relationships of patches. We apply our method to the classification of glioma and non-small-cell lung carcinoma cases into subtypes. The classification accuracy of our method is similar to the inter-observer agreement between pathologists. Although it is impossible to train CNNs on WSIs, we experimentally demonstrate using a comparable non-cancer dataset of smaller images that a patch-based CNN can outperform an image-based CNN.

1.173 Bottom-Up and Top-Down Reasoning with Hierarchical Rectified Gaussians

[Link to Paper]

Convolutional neural nets (CNNs) have demonstrated remarkable performance in recent history. Such approaches tend to work in a "unidirectional" bottom-up feed-forward fashion. However, practical experience and biological evidence tells us that feedback plays

a crucial role, particularly for detailed spatial understanding tasks. This work explores "bidirectional" architectures that also reason with top-down feedback: neural units are influenced by both lower and higher-level units. We do so by treating units as rectified latent variables in a quadratic energy function, which can be seen as a hierarchical Rectified Gaussian model (RGs) [39]. We show that RGs can be optimized with a quadratic program (QP), that can in turn be optimized with a recurrent neural network (with rectified linear units). This allows RGs to be trained with GPU-optimized gradient descent. From a theoretical perspective, RGs help establish a connection between CNNs and hierarchical probabilistic models. From a practical perspective, RGs are well suited for detailed spatial tasks that can benefit from top-down reasoning. We illustrate them on the challenging task of keypoint localization under occlusions, where local bottom-up evidence may be misleading. We demonstrate state-of-the-art results on challenging benchmarks.

1.174 Efficient Coarse-to-Fine PatchMatch for Large Displacement Optical Flow

[Link to Paper]

As a key component in many computer vision systems, optical flow estimation, especially with large displacements, remains an open problem. In this paper we present a simple but powerful matching method works in a coarseto-fine scheme for optical flow estimation. Inspired by the nearest neighbor field (NNF) algorithms, our approach, called CPM (Coarse-to-fine PatchMatch), blends an efficient random search strategy with the coarse-to-fine scheme for optical flow problem. Unlike existing NNF techniques, which is efficient but the results is often too noisy for optical flow caused by the lack of global regularization, we propose a propagation step with constrained random search radius between adjacent levels on the hierarchical architecture. The resulting correspondences enjoys a built-in smoothing effect, which is more suited for optical flow estimation than NNF techniques. Furthermore, our approach can also capture the tiny structures with large motions which is a problem for traditional coarse-to-fine optical flow algorithms. Interpolated by an edge-preserving interpolation method (EpicFlow), our method outperforms the state of the art on MPI-Sintel and KITTI, and runs much faster than the competing methods.

1.175 Image Deblurring using Smartphone Inertial Sensors

[Link to Paper]

Removing image blur caused by camera shake is an ill-posed problem, as both the latent image and the point spread function (PSF) are unknown. A recent approach to address

this problem is to record camera motion through inertial sensors, i.e., gyroscopes and accelerometers, and then reconstruct spatially-variant PSFs from these readings. While this approach has been effective for highquality inertial sensors, it has been infeasible for the inertial sensors in smartphones, which are of relatively low quality and present a number of challenging issues, including varying sensor parameters, high sensor noise, and calibration error. In this paper, we identify the issues that plague smartphone inertial sensors and propose a solution that successfully utilizes the sensor readings for image deblurring. With both the sensor data and the image itself, the proposed method is able to accurately estimate the sensor parameters online and also the spatially-variant PSFs for enhanced deblurring performance. The effectiveness of this technique is demonstrated in experiments on a popular mobile phone. With this approach, the quality of image deblurring can be appreciably raised on the most common of imaging devices.

1.176 Learning Structured Inference Neural Networks with Label Relations

[Link to Paper]

Images of scenes have various objects as well as abundant attributes, and diverse levels of visual categorization are possible. A natural image could be assigned with finegrained labels that describe major components, coarsegrained labels that depict high level abstraction, or a set of labels that reveal attributes. Such categorization at different concept layers can be modeled with label graphs encoding label information. In this paper, we exploit this rich information with a state-of-art deep learning framework, and propose a generic structured model that leverages diverse label relations to improve image classification performance. Our approach employs a novel stacked label prediction neural network, capturing both inter-level and intra-level label semantics. We evaluate our method on benchmark image datasets, and empirical results illustrate the efficacy of our model.

1.177 Natural Language Object Retrieval

[Link to Paper]

In this paper, we address the task of natural language object retrieval, to localize a target object within a given image based on a natural language query of the object. Natural language object retrieval differs from text-based image retrieval task as it involves spatial information about objects within the scene and global scene context. To address this issue, we propose a novel Spatial Context Recurrent ConvNet (SCRC) model as scoring function on candidate boxes for object retrieval, integrating spatial configurations and global scene-level contextual information into the network. Our model processes query text, local image descriptors, spatial configurations and global context features through a recurrent network,

84

outputs the probability of the query text conditioned on each candidate box as a score for the box, and can transfer visual-linguistic knowledge from image captioning domain to our task. Experimental results demonstrate that our method effectively utilizes both local and global information, outperforming previous baseline methods significantly on different datasets and scenarios, and can exploit large scale vision and language datasets for knowledge transfer.

1.178 Progressive Feature Matching with Alternate Descriptor Selection and Correspondence Enrichment

[Link to Paper]

We address two difficulties in establishing an accurate system for image matching. First, image matching relies on the descriptor for feature extraction, but the optimal descriptor often varies from image to image, or even patch to patch. Second, conventional matching approaches carry out geometric checking on a small set of correspondence candidates due to the concern of efficiency. It may result in restricted performance in recall. We aim at tackling the two issues by integrating adaptive descriptor selection and progressive candidate enrichment into image matching. We consider that the two integrated components are complementary: The high-quality matching yielded by adaptively selected descriptors helps in exploring more plausible candidates, while the enriched candidate set serves as a better reference for descriptor selection. It motivates us to formulate image matching as a joint optimization problem, in which adaptive descriptor selection and progressive correspondence enrichment are alternately conducted. Our approach is comprehensively evaluated and compared with the state-of-the-art approaches on two benchmarks. The promising results manifest its effectiveness.

1.179 Temporal Multimodal Learning in Audiovisual Speech Recognition

[Link to Paper]

In view of the advantages of deep networks in producing useful representation, the generated features of different modality data (such as image, audio) can be jointly learned using Multimodal Restricted Boltzmann Machines (MRBM). Recently, audiovisual speech recognition based the MRBM has attracted much attention, and the MRBM shows its effectiveness in learning the joint representation across audiovisual modalities. However, the built networks have weakness in modeling the multimodal sequence which is the natural property of speech signal. In this paper, we will introduce a novel temporal multimodal deep learning architecture, named as Recurrent Temporal Multimodal RBM (RTMRBM), that models multimodal sequences by transforming the sequence of connected MRBMs

into a probabilistic series model. Compared with existing multimodal networks, it's simple and efficient in learning temporal joint representation. We evaluate our model on audiovisual speech datasets, two public (AVLetters and AVLetters2) and one self-build. The experimental results demonstrate that our approach can obviously improve the accuracy of recognition compared with standard MRBM and the temporal model based on conditional RBM. In addition, RTMRBM still outperforms non-temporal multimodal deep networks in the presence of the weakness of long-term dependencies.

1.180 Camera Calibration from Periodic Motion of a Pedestrian

[Link to Paper]

Camera calibration directly from image sequences of a pedestrian without using any calibration object is a really challenging task and should be well solved in computer vision, especially in visual surveillance. In this paper, we propose a novel camera calibration method based on recovering the three orthogonal vanishing points (TOVPs), just using an image sequence of a pedestrian walking in a straight line, without any assumption of scenes or motions, e.g., control points with known 3D coordinates, parallel or perpendicular lines, non-natural or pre-designed special human motions, as often necessary in previous methods. The traces of shoes of a pedestrian carry more rich and easily detectable metric information than all other body parts in the periodic motion of a pedestrian, but such information is usually overlooked by previous work. In this paper, we employ the images of the toes of the shoes on the ground plane to determine the vanishing point corresponding to the walking direction, and then utilize harmonic conjugate properties in projective geometry to recover the vanishing point corresponding to the perpendicular direction of the walking direction in the horizontal plane and the vanishing point corresponding to the vertical direction. After recovering all of the TOVPs, the intrinsic and extrinsic parameters of the camera can be determined. Experiments on various scenes and viewing angles prove the feasibility and accuracy of the proposed method.

1.181 Detecting Migrating Birds at Night

[Link to Paper]

Bird migration is a critical indicator of environmental health, biodiversity, and climate change. Existing techniques for monitoring bird migration are either expensive (e.g., satellite tracking), labor-intensive (e.g., moon watching), indirect and thus less accurate (e.g., weather radar), or intrusive (e.g., attaching geolocators on captured birds). In this paper, we present a vision-based system for detecting migrating birds in flight at night. Our

system takes stereo videos of the night sky as inputs, detects multiple flying birds and estimates their orientations, speeds, and altitudes. The main challenge lies in detecting flying birds of unknown trajectories under high noise level due to the low-light environment. We address this problem by incorporating stereo constraints for rejecting physically implausible configurations and gathering evidence from two (or more) views. Specifically, we develop a robust stereo-based 3D line fitting algorithm for geometric verification and a deformable part response accumulation strategy for trajectory verification. We demonstrate the effectiveness of the proposed approach through quantitative evaluation of real videos of birds migrating at night collected with near-infrared cameras.

1.182   Homography Estimation from the Common Self-polar Triangle of Separate Ellipses

[Link to Paper]

How to avoid ambiguity is a challenging problem for conic-based homography estimation. In this paper, we address the problem of homography estimation from two separate ellipses. We find that any two ellipses have a unique common self-polar triangle, which can provide three line correspondences. Furthermore, by investigating the location features of the common self-polar triangle, we show that one vertex of the triangle lies outside of both ellipses, while the other two vertices lies inside the ellipses separately. Accordingly, one more line correspondence can be obtained from the intersections of the conics and the common self-polar triangle. Therefore, four line correspondences can be obtained based on the common self-polar triangle, which can provide enough constraints for the homography estimation. The main contributions in this paper include: (1) A new discovery on the location features of the common self-polar triangle of separate ellipses. (2) A novel approach for homography estimation. Simulate experiments and real experiments are conducted to demonstrate the feasibility and accuracy of our approach.

1.183   Learning Deep Representation for Imbalanced Classification

[Link to Paper]

Data in vision domain often exhibit highly-skewed class distribution, i.e., most data belong to a few majority classes, while the minority classes only contain a scarce amount of instances. To mitigate this issue, contemporary classification methods based on deep convolutional neural network (CNN) typically follow classic strategies such as class re-sampling or cost-sensitive training. In this paper, we conduct extensive and systematic experiments to validate the effectiveness of these classic schemes for representation learning on class-imbalanced data. We further demonstrate that more discriminative deep representation

can be learned by enforcing a deep network to maintain both intercluster and inter-class margins. This tighter constraint effectively reduces the class imbalance inherent in the local data neighborhood. We show that the margins can be easily deployed in standard deep learning framework through quintuplet instance sampling and the associated triple-header hinge loss. The representation learned by our approach, when combined with a simple k-nearest neighbor (kNN) algorithm, shows significant improvements over existing methods on both high- and low-level vision classification tasks that exhibit imbalanced class distribution.

1.184 Part-Stacked CNN for Fine-Grained Visual Categorization

[Link to Paper]

In the context of fine-grained visual categorization, the ability to interpret models as human-understandable visual manuals is sometimes as important as achieving high classification accuracy. In this paper, we propose a novel PartStacked CNN architecture that explicitly explains the finegrained recognition process by modeling subtle differences from object parts. Based on manually-labeled strong part annotations, the proposed architecture consists of a fully convolutional network to locate multiple object parts and a two-stream classification network that encodes object-level and part-level cues simultaneously. By adopting a set of sharing strategies between the computation of multiple object parts, the proposed architecture is very efficient running at 20 frames/sec during inference. Experimental results on the CUB-200-2011 dataset reveal the effectiveness of the proposed architecture, from multiple perspectives of classification accuracy, model interpretability, and efficiency. Being able to provide interpretable recognition results in realtime, the proposed method is believed to be effective in practical applications.

1.185 Sparse Coding and Dictionary Learning with Linear Dynamical Systems

[Link to Paper]

Linear Dynamical Systems (LDSs) are the fundamental tools for encoding spatio-temporal data in various disciplines. To enhance the performance of LDSs, in this paper, we address the challenging issue of performing sparse coding on the space of LDSs, where both data and dictionary atoms are LDSs. Rather than approximate the extended observability with a finite-order matrix, we represent the space of LDSs by an infinite Grassmannian consisting of the orthonormalized extended observability subspaces. Via a homeomorphic mapping, such Grassmannian is embedded into the space of symmetric matrices, where a tractable objective function can be derived for sparse coding. Then, we propose an efficient method

to learn the system parameters of the dictionary atoms explicitly, by imposing the symmetric constraint to the transition matrices of the data and dictionary systems. Moreover, we combine the state covariance into the algorithm formulation, thus further promoting the performance of the models with symmetric transition matrices. Comparative experimental evaluations reveal the superior performance of proposed methods on various tasks including video classification and tactile recognition.

1.186 Unsupervised Learning of Discriminative Attributes and Visual Representations

[Link to Paper]

Attributes offer useful mid-level features to interpret visual data. While most attribute learning methods are supervised by costly human-generated labels, we introduce a simple yet powerful unsupervised approach to learn and predict visual attributes directly from data. Given a large unlabeled image collection as input, we train deep Convolutional Neural Networks (CNNs) to output a set of discriminative, binary attributes often with semantic meanings. Specifically, we first train a CNN coupled with unsupervised discriminative clustering, and then use the cluster membership as a soft supervision to discover shared attributes from the clusters while maximizing their separability. The learned attributes are shown to be capable of encoding rich imagery properties from both natural images and contour patches. The visual representations learned in this way are also transferrable to other tasks such as object detection. We show other convincing results on the related tasks of image retrieval and classification, and contour detection. Large Small sized sized VehAicnliemal LonSgholergt slegs BiSgmsahlalpoevdal shaped Marine Aero Birds and frogs Small deer Non-stSrtariagihgtht DoSuibnlge-leli-nliended Curved Junctional

1.187 Volumetric 3D Tracking by Detection

[Link to Paper]

In this paper, we propose a new framework for 3D tracking by detection based on fully volumetric representations. On one hand, 3D tracking by detection has shown robust use in the context of interaction (Kinect) and surface tracking. On the other hand, volumetric representations have recently been proven efficient both for building 3D features and for addressing the 3D tracking problem. We leverage these benefits by unifying both families of approaches into a single, fully volumetric tracking-by-detection framework. We use a centroidal Voronoi tessellation (CVT) representation to compactly tessellate shapes with optimal discretization, construct a feature space, and perform the tracking according to the correspondences provided by trained random forests. Our results show improved

tracking and training computational efficiency and improved memory performance. This in turn enables the use of larger training databases than state of the art approaches, which we leverage by proposing a cross-tracking subject training scheme to benefit from all subject sequences for all tracking situations, thus yielding better detection and less overfitting.

1.188 Detecting Repeating Objects using Patch Correlation Analysis

[Link to Paper]

In this paper we describe a new method for detecting and counting a repeating object in an image. While the method relies on a fairly sophisticated deformable part model, unlike existing techniques it estimates the model parameters in an unsupervised fashion thus alleviating the need for a user-annotated training data and avoiding the associated specificity. This automatic fitting process is carried out by exploiting the recurrence of small image patches associated with the repeating object and analyzing their spatial correlation. The analysis allows us to reject outlier patches, recover the visual and shape parameters of the part model, and detect the object instances efficiently. In order to achieve a practical system which is able to cope with diverse images, we describe a simple and intuitive active-learning procedure that updates the object classification by querying the user on very few carefully chosen marginal classifications. Evaluation of the new method against the state-of-the-art techniques demonstrates its ability to achieve higher accuracy through a better user experience.

1.189 Coupled Harmonic Bases for Longitudinal Characterization of Brain Networks

[Link to Paper]

There is a great deal of interest in using large scale brain imaging studies to understand how brain connectivity evolves over time for an individual and how it varies over different levels/quantiles of cognitive function. To do so, one typically performs so-called tractography procedures on diffusion MR brain images and derives measures of brain connectivity expressed as graphs. The nodes correspond to distinct brain regions and the edges encode the strength of the connection. The scientific interest is in characterizing the evolution of these graphs over time or from healthy individuals to diseased. We pose this important question in terms of the Laplacian of the connectivity graphs derived from various longitudinal or disease time points – quantifying its progression is then expressed in terms of coupling the harmonic bases of a full set of Laplacians. We derive a coupled system of generalized eigenvalue problems (and corresponding numerical optimization schemes) whose

solution helps characterize the full life cycle of brain connectivity evolution in a given dataset. Finally, we show a set of results on a diffusion MR imaging dataset of middle aged people at risk for Alzheimer's disease (AD), who are cognitively healthy. In such asymptomatic adults, we find that a framework for characterizing brain connectivity evolution provides the ability to predict cognitive scores for individual subjects, and for estimating the progression of participant's brain connectivity into the future.

1.190 FireCaffe: near-linear acceleration of deep neural network training on compute clusters

[Link to Paper]

Long training times for high-accuracy deep neural networks (DNNs) impede research into new DNN architectures and slow the development of high-accuracy DNNs. In this paper we present FireCaffe, which successfully scales deep neural network training across a cluster of GPUs. We also present a number of best practices to aid in comparing advancements in methods for scaling and accelerating the training of deep neural networks. The speed and scalability of distributed algorithms is almost always limited by the overhead of communicating between servers; DNN training is not an exception to this rule. Therefore, the key consideration here is to reduce communication overhead wherever possible, while not degrading the accuracy of the DNN models that we train. Our approach has three key pillars. First, we select network hardware that achieves high bandwidth between GPU servers Infiniband or Cray interconnects are ideal for this. Second, we consider a number of communication algorithms, and we find that reduction trees are more efficient and scalable than the traditional parameter server approach. Third, we optionally increase the batch size to reduce the total quantity of communication during DNN training, and we identify hyperparameters that allow us to reproduce the small-batch accuracy while training with large batch sizes. When training GoogLeNet and Network-in-Network on ImageNet, we achieve a 47x and 39x speedup, respectively, when training on a cluster of 128 GPUs.

1.191 A Hierarchical Deep Temporal Model for Group Activity Recognition

[Link to Paper]

In group activity recognition, the temporal dynamics of the whole activity can be inferred based on the dynamics of the individual people representing the activity. We build a deep model to capture these dynamics based on LSTM (long short-term memory) models. To make use of these observations, we present a 2-stage deep temporal model for the group activity recognition problem. In our model, a LSTM model is designed to represent action

dynamics of individual people in a sequence and another LSTM model is designed to aggregate person-level information for whole activity understanding. We evaluate our model over two datasets: the Collective Activity Dataset and a new volleyball dataset. Experimental results demonstrate that our proposed model improves group activity recognition performance compared to baseline methods.

1.192 How hard can it be? Estimating the difficulty of visual search in an image

[Link to Paper]

We address the problem of estimating image difficulty defined as the human response time for solving a visual search task. We collect human annotations of image difficulty for the PASCAL VOC 2012 data set through a crowd-sourcing platform. We then analyze what human interpretable image properties can have an impact on visual search difficulty, and how accurate are those properties for predicting difficulty. Next, we build a regression model based on deep features learned with state of the art convolutional neural networks and show better results for predicting the groundtruth visual search difficulty scores produced by human annotators. Our model is able to correctly rank about 75% image pairs according to their difficulty score. We also show that our difficulty predictor generalizes well to new classes not seen during training. Finally, we demonstrate that our predicted difficulty scores are useful for weakly supervised object localization (8% improvement) and semi-supervised object classification (1% improvement).

1.193 Hedgehog Shape Priors for Multi-object Segmentation

[Link to Paper]

Star-convexity prior is popular for interactive single object segmentation due to its simplicity and amenability to binary graph cut optimization. We propose a more general multi-object segmentation approach. Moreover, each object can be constrained by a more descriptive shape prior, "hedgehog". Each hedgehog shape has its surface normals locally constrained by an arbitrary given vector field, e.g. gradient of the user-scribble distance transform. In contrast to star-convexity, the tightness of our normal constraint can be changed giving better control over allowed shapes. For example, looser constraints, i.e. wider cones of allowed normals, give more relaxed hedgehog shapes. On the other hand, the tightest constraint enforces skeleton consistency with the scribbles. In general, hedgehog shapes are more descriptive than a star, which is only a special case corresponding to a radial vector field and weakest tightness. Our approach has significantly more applications

92

than standard single star-convex segmentation, e.g. in medical data we can separate multiple non-star organs with similar appearances and weak edges. Optimization is done by our modified -expansion moves shown to be submodular for multi-hedgehog shapes.

1.194 Efficient Large-Scale Similarity Search Using Matrix Factorization

[Link to Paper]

We consider the image retrieval problem of finding the images in a dataset that are most similar to a query image. Our goal is to reduce the number of vector operations and memory for performing a search without sacrificing accuracy of the returned images. We adopt a group testing formulation and design the decoding architecture using either dictionary learning or eigendecomposition. The latter is a plausible option for small-to-medium sized problems with high-dimensional global image descriptors, whereas dictionary learning is applicable in large-scale scenarios. We evaluate our approach for global descriptors obtained from both SIFT and CNN features. Experiments with standard image search benchmarks, including the Yahoo100M dataset comprising 100 million images, show that our method gives comparable (and sometimes superior) accuracy compared to exhaustive search while requiring only 10% of the vector operations and memory. Moreover, for the same search complexity, our method gives significantly better accuracy compared to approaches based on dimensionality reduction or locality sensitive hashing.

1.195 Structured Receptive Fields in CNNs

[Link to Paper]

Learning powerful feature representations with CNNs is hard when training data are limited. Pre-training is one way to overcome this, but it requires large datasets sufficiently similar to the target domain. Another option is to design priors into the model, which can range from tuned hyperparameters to fully engineered representations like Scattering Networks. We combine these ideas into structured receptive field networks, a model which has a fixed filter basis and yet retains the flexibility of CNNs. This flexibility is achieved by expressing receptive fields in CNNs as a weighted sum over a fixed basis which is similar in spirit to Scattering Networks. The key difference is that we learn arbitrary effective filter sets from the basis rather than modeling the filters. This approach explicitly connects classical multiscale image analysis with general CNNs. With structured receptive field networks, we improve considerably over unstructured CNNs for small and medium dataset scenarios as well as over Scattering for large datasets. We validate our findings on ILSVRC2012, Cifar-10, Cifar-100 and MNIST. As a realistic small dataset example, we

show state-of-the-art classification results on popular 3D MRI brain-disease datasets where pre-training is difficult due to a lack of large public datasets in a similar domain.

1.196 Active Image Segmentation Propagation

[Link to Paper]

We propose a semi-automatic method to obtain foreground object masks for a large set of related images. We develop a stagewise active approach to propagation: in each stage, we actively determine the images that appear most valuable for human annotation, then revise the foreground estimates in all unlabeled images accordingly. In order to identify images that, once annotated, will propagate well to other examples, we introduce an active selection procedure that operates on the joint segmentation graph over all images. It prioritizes human intervention for those images that are uncertain and influential in the graph, while also mutually diverse. We apply our method to obtain foreground masks for over 1 million images. Our method yields state-of-the-art accuracy on the ImageNet and MIT Object Discovery datasets, and it focuses human attention more effectively than existing propagation strategies.

1.197 Structural-RNN: Deep Learning on Spatio-Temporal Graphs

[Link to Paper]

Deep Recurrent Neural Network architectures, though remarkably capable at modeling sequences, lack an intuitive high-level spatio-temporal structure. That is while many problems in computer vision inherently have an underlying high-level structure and can benefit from it. Spatiotemporal graphs are a popular tool for imposing such highlevel intuitions in the formulation of real world problems. In this paper, we propose an approach for combining the power of high-level spatio-temporal graphs and sequence learning success of Recurrent Neural Networks (RNNs). We develop a scalable method for casting an arbitrary spatiotemporal graph as a rich RNN mixture that is feedforward, fully differentiable, and jointly trainable. The proposed method is generic and principled as it can be used for transforming any spatio-temporal graph through employing a certain set of well defined steps. The evaluations of the proposed approach on a diverse set of problems, ranging from modeling human motion to object interactions, shows improvement over the state-of-the-art with a large margin. We expect this method to empower new approaches to problem formulation through high-level spatio-temporal graphs and Recurrent Neural Networks. Links: Web

1.198 Learning Sparse High Dimensional Filters: Image Filtering, Dense CRFs and Bilateral Neural Networks

[Link to Paper]

Bilateral filters have wide spread use due to their edgepreserving properties. The common use case is to manually choose a parametric filter type, usually a Gaussian filter. In this paper, we will generalize the parametrization and in particular derive a gradient descent algorithm so the filter parameters can be learned from data. This derivation allows to learn high dimensional linear filters that operate in sparsely populated feature spaces. We build on the permutohedral lattice construction for efficient filtering. The ability to learn more general forms of high-dimensional filters can be used in several diverse applications. First, we demonstrate the use in applications where single filter applications are desired for runtime reasons. Further, we show how this algorithm can be used to learn the pairwise potentials in densely connected conditional random fields and apply these to different image segmentation tasks. Finally, we introduce layers of bilateral filters in CNNs and propose bilateral neural networks for the use of highdimensional sparse data. This view provides new ways to encode model structure into network architectures. A diverse set of experiments empirically validates the usage of general forms of filters.

1.199 Primary Object Segmentation in Videos via Alternate Convex Optimization of Foreground and Background Distributions

[Link to Paper]

An unsupervised video object segmentation algorithm, which discovers a primary object in a video sequence automatically, is proposed in this work. We introduce three energies in terms of foreground and background probability distributions: Markov, spatiotemporal, and antagonistic energies. Then, we minimize a hybrid of the three energies to separate a primary object from its background. However, the hybrid energy is nonconvex. Therefore, we develop the alternate convex optimization (ACO) scheme, which decomposes the nonconvex optimization into two quadratic programs. Moreover, we propose the forward-backward strategy, which performs the segmentation sequentially from the first to the last frames and then vice versa, to exploit temporal correlations. Experimental results on extensive datasets demonstrate that the proposed ACO algorithm outperforms the state-of-the-art techniques significantly.

1.200 Slow and steady feature analysis: higher order temporal coherence in video

[Link to Paper]

How can unlabeled video augment visual learning? Existing methods perform "slow" feature analysis, encouraging the representations of temporally close frames to exhibit only small differences. While this standard approach captures the fact that high-level visual signals change slowly over time, it fails to capture how the visual content changes. We propose to generalize slow feature analysis to "steady" feature analysis. The key idea is to impose a prior that higher order derivatives in the learned feature space must be small. To this end, we train a convolutional neural network with a regularizer on tuples of sequential frames from unlabeled video. It encourages feature changes over time to be smooth, i.e., similar to the most recent changes. Using five diverse datasets, including unlabeled YouTube and KITTI videos, we demonstrate our method's impact on object, scene, and action recognition tasks. We further show that our features learned from unlabeled video can even surpass a standard heavily supervised pretraining approach.

1.201 Stereo Matching with Color and Monochrome Cameras in Low-light Conditions

[Link to Paper]

Consumer devices with stereo cameras have become popular because of their low-cost depth sensing capability. However, those systems usually suffer from low imaging quality and inaccurate depth acquisition under low-light conditions. To address the problem, we present a new stereo matching method with a color and monochrome camera pair. We focus on the fundamental trade-off that monochrome cameras have much better light-efficiency than color-filtered cameras. Our key ideas involve compensating for the radiometric difference between two crossspectral images and taking full advantage of complementary data. Consequently, our method produces both an accurate depth map and high-quality images, which are applicable for various depth-aware image processing. Our method is evaluated using various datasets and the performance of our depth estimation consistently outperforms state-of-the-art methods.

1.202 End-to-End Saliency Mapping via Probability Distribution Prediction

[Link to Paper]

Most saliency estimation methods aim to explicitly model low-level conspicuity cues such as edges or blobs and may additionally incorporate top-down cues using face or text detection. Data-driven methods for training saliency models using eye-fixation data are increasingly popular, particularly with the introduction of large-scale datasets and deep architectures. However, current methods in this latter paradigm use loss functions designed for classification or regression tasks whereas saliency estimation is evaluated on topographical maps. In this work, we introduce a new saliency map model which formulates a map as a generalized Bernoulli distribution. We then train a deep architecture to predict such maps using novel loss functions which pair the softmax activation function with measures designed to compute distances between probability distributions. We show in extensive experiments the effectiveness of such loss functions over standard ones on four public benchmark datasets, and demonstrate improved performance over state-of-the-art saliency methods.

1.203   Robust Multi-body Feature Tracker: A Segmentation-free Approach

[Link to Paper]

Feature tracking is a fundamental problem in computer vision, with applications in many computer vision tasks, such as visual SLAM and action recognition. This paper introduces a novel multi-body feature tracker that exploits a multi-body rigidity assumption to improve tracking robustness under a general perspective camera model. A conventional approach to addressing this problem would consist of alternating between solving two subtasks: motion segmentation and feature tracking under rigidity constraints for each segment. This approach, however, requires knowing the number of motions, as well as assigning points to motion groups, which is typically sensitive to the motion estimates. By contrast, here, we introduce a segmentationfree solution to multi-body feature tracking that bypasses the motion assignment step and reduces to solving a series of subproblems with closed-form solutions. Our experiments demonstrate the benefits of our approach in terms of tracking accuracy and robustness to noise.

1.204   What Sparse Light Field Coding Reveals about Scene Structure

[Link to Paper]

In this paper, we propose a novel method for depth estimation in light fields which employs a specifically designed sparse decomposition to leverage the depth-orientation relationship on its epipolar plane images. The proposed method learns the structure of the central view and uses this information to construct a light field dictionary for which groups of atoms

correspond to unique disparities. This dictionary is then used to code a sparse representation of the light field. Analyzing the coefficients of this representation with respect to the disparities of their corresponding atoms yields an accurate and robust estimate of depth. In addition, if the light field has multiple depth layers, such as for reflective or transparent surfaces, statistical analysis of the coefficients can be employed to infer the respective depth of the superimposed layers.

1.205 Pairwise Decomposition of Image Sequences for Active Multi-View Recognition

[Link to Paper]

A multi-view image sequence provides a much richer capacity for object recognition than from a single image. However, most existing solutions to multi-view recognition typically adopt hand-crafted, model-based geometric methods, which do not readily embrace recent trends in deep learning. We propose to bring Convolutional Neural Networks to generic multi-view recognition, by decomposing an image sequence into a set of image pairs, classifying each pair independently, and then learning an object classifier by weighting the contribution of each pair. This allows for recognition over arbitrary camera trajectories, without requiring explicit training over the potentially infinite number of camera paths and lengths. Building these pairwise relationships then naturally extends to the next-best-view problem in an active recognition framework. To achieve this, we train a second Convolutional Neural Network to map directly from an observed image to next viewpoint. Finally, we incorporate this into a trajectory optimisation task, whereby the best recognition confidence is sought for a given trajectory length. We present state-of-the-art results in both guided and unguided multi-view recognition on the ModelNet dataset, and show how our method can be used with depth images, greyscale images, or both.

1.206 DenseCap: Fully Convolutional Localization Networks for Dense Captioning

[Link to Paper]

We introduce the dense captioning task, which requires a computer vision system to both localize and describe salient regions in images in natural language. The dense captioning task generalizes object detection when the descriptions consist of a single word, and Image Captioning when one predicted region covers the full image. To address the localization and description task jointly we propose a Fully Convolutional Localization Network (FCLN) architecture that processes an image with a single, efficient forward pass, requires no external regions proposals, and can be trained end-to-end with a single round of optimization.

The architecture is composed of a Convolutional Network, a novel dense localization layer, and Recurrent Neural Network language model that generates the label sequences. We evaluate our network on the Visual Genome dataset, which comprises 94,000 images and 4,100,000 region-grounded captions. We observe both speed and accuracy improvements over baselines based on current state of the art approaches in both generation and retrieval settings.

1.207   Globally Optimal Manhattan Frame Estimation in Real-time

[Link to Paper]

Given a set of surface normals, we pose a Manhattan Frame (MF) estimation problem as a consensus set maximization that maximizes the number of inliers over the rotation search space. We solve this problem through a branchand-bound framework, which mathematically guarantees a globally optimal solution. However, the computational time of conventional branch-and-bound algorithms are intractable for real-time performance. In this paper, we propose a novel bound computation method within an efficient measurement domain for MF estimation, i.e., the extended Gaussian image (EGI). By relaxing the original problem, we can compute the bounds in real-time, while preserving global optimality. Furthermore, we quantitatively and qualitatively demonstrate the performance of the proposed method for synthetic and real-world data. We also show the versatility of our approach through two applications: extension to multiple MF estimation and video stabilization.

1.208   Large-pose Face Alignment via CNN-based Dense 3D Model Fitting

[Link to Paper]

Large-pose face alignment is a very challenging problem in computer vision, which is used as a prerequisite for many important vision tasks, e.g, face recognition and 3D face reconstruction. Recently, there have been a few attempts to solve this problem, but still more research is needed to achieve highly accurate results. In this paper, we propose a face alignment method for large-pose face images, by combining the powerful cascaded CNN regressor method and 3DMM. We formulate the face alignment as a 3DMM fitting problem, where the camera projection matrix and 3D shape parameters are estimated by a cascade of CNN-based regressors. The dense 3D shape allows us to design pose-invariant appearance features for effective CNN learning. Extensive experiments are conducted on the challenging databases (AFLW and AFW), with comparison to the state of the art.

1.209 Mixture of Bilateral-Projection Two-dimensional Probabilistic Principal Component Analysis

[Link to Paper]

The probabilistic principal component analysis (PPCA) is built upon a global linear mapping, with which it is insufficient to model complex data variation. This paper proposes a mixture of bilateral-projection probabilistic principal component analysis model (mixB2DPPCA) on 2D data. With multi-components in the mixture, this model can be seen as a 'soft' cluster algorithm and has capability of modeling data with complex structures. A Bayesian inference scheme has been proposed based on the variational EM (Expectation-Maximization) approach for learning model parameters. Experiments on some publicly available databases show that the performance of mixB2DPPCA has been largely improved, resulting in more accurate reconstruction errors and recognition rates than the existing PCA-based algorithms.

1.210 Moral Lineage Tracing

[Link to Paper]

Lineage tracing, the tracking of living cells as they move and divide, is a central problem in biological image analysis. Solutions, called lineage forests, are key to understanding how the structure of multicellular organisms emerges. We propose an integer linear program (ILP) whose feasible solutions define, for every image in a sequence, a decomposition into cells (segmentation) and, across images, a lineage forest of cells (tracing). In this ILP, path-cut inequalities enforce the morality of lineages, i.e., the constraint that cells do not merge. To find feasible solutions of this NP-hard problem, with certified bounds to the global optimum, we define efficient separation procedures and apply these as part of a branch-and-cut algorithm. To show the effectiveness of this approach, we analyze feasible solutions for real microscopy data in terms of bounds and run-time, and by their weighted edit distance to lineage forests traced by humans.

1.211 Macroscopic Interferometry: Rethinking Depth Estimation with Frequency-Domain Time-of-Flight

[Link to Paper]

A form of meter-scale, macroscopic interferometry is proposed using conventional time-of-flight (ToF) sensors. Today, ToF sensors use phase-based sampling, where the phase delay

between emitted and received, high-frequency signals encodes distance. This paper examines an alternative ToF architecture, inspired by micron-scale, microscopic interferometry, that relies only on frequency sampling: we refer to our proposed macroscopic technique as FrequencyDomain Time of Flight (FD-ToF). The proposed architecture offers several benefits over existing phase ToF systems, such as robustness to phase wrapping and implicit resolution of multi-path interference, all while capturing the same number of subframes. A prototype camera is constructed to demonstrate macroscopic interferometry at meter scale.

1.212 Answer-Type Prediction for Visual Question Answering

[Link to Paper]

Recently, algorithms for object recognition and related tasks have become sufficiently proficient that new vision tasks can now be pursued. In this paper, we build a system capable of answering open-ended text-based questions about images, which is known as Visual Question Answering (VQA). Our approach's key insight is that we can predict the form of the answer from the question. We formulate our solution in a Bayesian framework. When our approach is combined with a discriminative model, the combined model achieves state-of-the-art results on four benchmark datasets for open-ended VQA: DAQUAR, COCO-QA, The VQA Dataset, and Visual7W.

1.213 Multi-view Deep Network for Cross-view Classification

[Link to Paper]

Cross-view recognition that intends to classify samples between different views is an important problem in computer vision. The large discrepancy between different even heterogenous views make this problem quite challenging. To eliminate the complex (maybe even highly nonlinear) view discrepancy for favorable cross-view recognition, we propose a multi-view deep network (MvDN), which seeks for a non-linear discriminant and view-invariant representation shared between multiple views. Specifically, our proposed MvDN network consists of two sub-networks, view-specific sub-network attempting to remove view-specific variations and the following common sub-network attempting to obtain common representation shared by all views. As the objective of MvDN network, the Fisher loss, i.e. the Rayleigh quotient objective, is calculated from the samples of all views so as to guide the learning of the whole network. As a result, the representation from the topmost layers of the MvDN network is robust to view discrepancy, and also discriminative. The experiments of face recognition across pose and face recognition across feature

type on three datasets with 13 and 2 views respectively demonstrate the superiority of the proposed method, especially compared to the typical linear ones.

1.214 WarpNet: Weakly Supervised Matching for Single-view Reconstruction

[Link to Paper]

We present an approach to matching images of objects in fine-grained datasets without using part annotations, with an application to the challenging problem of weakly supervised single-view reconstruction. This is in contrast to prior works that require part annotations, since matching objects across class and pose variations is challenging with appearance features alone. We overcome this challenge through a novel deep learning architecture, WarpNet, that aligns an object in one image with a different object in another. We exploit the structure of the fine-grained dataset to create artificial data for training this network in an unsupervised-discriminative learning approach. The output of the network acts as a spatial prior that allows generalization at test time to match real images across variations in appearance, viewpoint and articulation. On the CUB-200-2011 dataset of bird categories, we improve the AP over an appearance-only network by 13.6%. We further demonstrate that our WarpNet matches, together with the structure of fine-grained datasets, allow single-view reconstructions with quality comparable to using annotated point correspondences.

1.215 Multi-label Ranking from Positive and Unlabeled Data

[Link to Paper]

In this paper, we specifically examine the training of a multi-label classifier from data with incompletely assigned labels. This problem is fundamentally important in many multi-label applications because it is almost impossible for human annotators to assign a complete set of labels, although their judgments are reliable. In other words, a multilabel dataset usually has properties by which (1) assigned labels are definitely positive and (2) some labels are absent but are still considered positive. Such a setting has been studied as a positive and unlabeled (PU) classification problem in a binary setting. We treat incomplete label assignment problems as a multi-label PU ranking, which is an extension of classical binary PU problems to the wellstudied rank-based multi-label classification. We derive the conditions that should be satisfied to cancel the negative effects of label incompleteness. Our experimentally obtained results demonstrate the effectiveness of these conditions.

1.216 Object Detection from Video Tubelets with Convolutional Neural Networks

[Link to Paper]

Deep Convolution Neural Networks (CNNs) have shown impressive performance in various vision tasks such as image classification, object detection and semantic segmentation. For object detection, particularly in still images, the performance has been significantly increased last year thanks to powerful deep networks (e.g. GoogleNet) and detection frameworks (e.g. Regions with CNN features (RCNN)). The lately introduced ImageNet [6] task on object detection from video (VID) brings the object detection task into the video domain, in which objects' locations at each frame are required to be annotated with bounding boxes. In this work, we introduce a complete framework for the VID task based on still-image object detection and general object tracking. Their relations and contributions in the VID task are thoroughly studied and evaluated. In addition, a temporal convolution network is proposed to incorporate temporal information to regularize the detection results and shows its effectiveness for the task. Code is available at https://github.com/myfavouritekk/vdetlib.

1.217 An Empirical Evaluation of Current Convolutional Architectures' Ability to Manage Nuisance Location and Scale Variability

[Link to Paper]

We conduct an empirical study to test the ability of convolutional neural networks (CNNs) to reduce the effects of nuisance transformations of the input data, such as location, scale and aspect ratio. We isolate factors by adopting a common convolutional architecture either deployed globally on the image to compute class posterior distributions, or restricted locally to compute class conditional distributions given location, scale and aspect ratios of bounding boxes determined by proposal heuristics. In theory, averaging the latter should yield inferior performance compared to proper marginalization. Yet empirical evidence suggests the converse, leading us to conclude that at the current level of complexity of convolutional architectures and scale of the data sets used to train them CNNs are not very effective at marginalizing nuisance variability. We also quantify the effects of context on the overall classification task and its impact on the performance of CNNs, and propose improved sampling techniques for heuristic proposal schemes that improve end-to-end performance to state-of-the-art levels. We test our hypothesis on a classification task using the ImageNet Challenge benchmark and on a wide-baseline matching task using the Oxford and Fischer's datasets.

1.218 The MegaFace Benchmark: 1 Million Faces for Recognition at Scale

[Link to Paper]

Recent face recognition experiments on a major benchmark (LFW [15]) show stunning performance-a number of algorithms achieve near to perfect score, surpassing human recognition rates. In this paper, we advocate evaluations at the million scale (LFW includes only 13K photos of 5K people). To this end, we have assembled the MegaFace dataset and created the first MegaFace challenge. Our dataset includes One Million photos that capture more than 690K different individuals. The challenge evaluates performance of algorithms with increasing numbers of "distractors" (going from 10 to 1M) in the gallery set. We present both identification and verification performance, evaluate performance with respect to pose and a persons age, and compare as a function of training data size (#photos and #people). We report results of state of the art and baseline algorithms. The MegaFace dataset, baseline code, and evaluation scripts, are all publicly released for further experimentations

1.219 Weakly Supervised Object Boundaries

[Link to Paper]

State-of-the-art learning based boundary detection methods require extensive training data. Since labelling object boundaries is one of the most expensive types of annotations, there is a need to relax the requirement to carefully annotate images to make both the training more affordable and to extend the amount of training data. In this paper we propose a technique to generate weakly supervised annotations and show that bounding box annotations alone suffice to reach high-quality object boundaries without using any object-specific boundary annotations. With the proposed weak supervision techniques we achieve the top performance on the object boundary detection task, outperforming by a large margin the current fully supervised state-of-theart methods.

1.220 In the Shadows, Shape Priors Shine: Using Occlusion to Improve Multi-Region Segmentation

[Link to Paper]

We present a new algorithm for multi-region segmentation of 2D images with objects that may partially occlude each other. Our algorithm is based on the observation that human performance on this task is based both on prior knowledge about plausible shapes and taking into account the presence of occluding objects whose shape is already known once an

occluded region is identified, the shape prior can be used to guess the shape of the missing part. We capture the former aspect using a deep learning model of shape; for the latter, we simultaneously minimize the energy of all regions and consider only unoccluded pixels for data agreement. Existing algorithms incorporating object shape priors consider every object separately in turn and can't distinguish genuine deviation from the expected shape from parts missing due to occlusion. We show that our method significantly improves on the performance of a representative algorithm, as evaluated on both preprocessed natural and synthetic images. Furthermore, on the synthetic images, we recover the ground truth segmentation with good accuracy.

1.221 Accurate Image Super-Resolution Using Very Deep Convolutional Networks

[Link to Paper]

We present a highly accurate single-image superresolution (SR) method. Our method uses a very deep convolutional network inspired by VGG-net used for ImageNet classification [19]. We find increasing our network depth shows a significant improvement in accuracy. Our final model uses 20 weight layers. By cascading small filters many times in a deep network structure, contextual information over large image regions is exploited in an efficient way. With very deep networks, however, convergence speed becomes a critical issue during training. We propose a simple yet effective training procedure. We learn residuals only and use extremely high learning rates (104 times higher than SRCNN [6]) enabled by adjustable gradient clipping. Our proposed method performs better than existing methods in accuracy and visual improvements in our results are easily noticeable.

1.222 Deeply-Recursive Convolutional Network for Image Super-Resolution

[Link to Paper]

We propose an image super-resolution method (SR) using a deeply-recursive convolutional network (DRCN). Our network has a very deep recursive layer (up to 16 recursions). Increasing recursion depth can improve performance without introducing new parameters for additional convolutions. Albeit advantages, learning a DRCN is very hard with a standard gradient descent method due to exploding/vanishing gradients. To ease the difficulty of training, we propose two extensions: recursive-supervision and skip-connection. Our method outperforms previous methods by a large margin.

1.223 Latent Variable Graphical Model Selection using Harmonic Analysis: Applications to the Human Connectome Project (HCP)

[Link to Paper]

A major goal of imaging studies such as the (ongoing) Human Connectome Project (HCP) is to characterize the structural network map of the human brain and identify its associations with covariates such as genotype, risk factors, and so on that correspond to an individual. But the set of image derived measures and the set of covariates are both large, so we must first estimate a 'parsimonious' set of relations between the measurements. For instance, a Gaussian graphical model will show conditional independences between the random variables, which can then be used to setup specific downstream analyses. But most such data involve a large list of 'latent' variables that remain unobserved, yet affect the 'observed' variables sustantially. Accounting for such latent variables is not directly addressed by standard precision matrix estimation, and is tackled via highly specialized optimization methods. This paper offers a unique harmonic analysis view of this problem. By casting the estimation of the precision matrix in terms of a composition of low-frequency latent variables and high-frequency sparse terms, we show how the problem can be formulated using a new wavelet-type expansion in non-Euclidean spaces. Our formulation poses the estimation problem in the frequency space and shows how it can be solved by a simple sub-gradient scheme. We provide a set of scientific results on 500 scans from the recently released HCP data where our algorithm recovers highly interpretable and sparse conditional dependencies between brain connectivity pathways and well-known covariates.

1.224 Learning to Select Pre-trained Deep Representations with Bayesian Evidence Framework

[Link to Paper]

We propose a Bayesian evidence framework to facilitate transfer learning from pre-trained deep convolutional neural networks (CNNs). Our framework is formulated on top of a least squares SVM (LS-SVM) classifier, which is simple and fast in both training and testing, and achieves competitive performance in practice. The regularization parameters in LS-SVM is estimated automatically without grid search and cross-validation by maximizing evidence, which is a useful measure to select the best performing CNN out of multiple candidates for transfer learning; the evidence is optimized efficiently by employing Aitken's delta-squared process, which accelerates convergence of fixed point update. The proposed Bayesian evidence framework also provides a good solution to identify the best ensemble of heterogeneous CNNs through a greedy algorithm. Our Bayesian evidence framework for transfer learning is tested on 12 visual recognition datasets and illustrates the state-of-the-art performance consistently in terms of prediction accuracy and modeling efficiency.

1.225 Reconstructing Shapes and Appearances of Thin Film Objects using RGB Images

[Link to Paper]

Reconstruction of shapes and appearances of thin film objects can be applied to many fields such as industrial inspection, biological analysis, and archaeologic research. However, it comes with many challenging issues because the appearances of thin film can change dramatically depending on view and light directions. The appearance is deeply dependent on not only the shapes but also the optical parameters of thin film. In this paper, we propose a novel method to estimate shapes and film thickness. First, we narrow down candidates of zenith angle by degree of polarization and determine it by the intensity of thin film which increases monotonically along the zenith angle. Second, we determine azimuth angle from occluding boundaries. Finally, we estimate the film thickness by comparing a look-up table of color along the thickness and zenith angle with captured images. We experimentally evaluated the accuracy of estimated shapes and appearances and found that our proposed method is effective.

1.226 Structured Feature Similarity with Explicit Feature Map

[Link to Paper]

Feature matching is a fundamental process in a variety of computer vision tasks. Beyond the standard L2 metric, various methods to measure similarity between features have been proposed mainly on the assumption that the features are defined in a histogram form. On the other hand, in a field of image quality assessment, SSIM [27] produces effective similarity between images, taking the place of L2 metric. In this paper, we propose a feature similarity measurement method based on the SSIM. Unlike the previous methods, the proposed method is built on not a histogram form but a tensor structure of a feature array extracted such as on spatial grids, in order to construct effective SSIMbased similarity measure of high robustness which is a key requirement in feature matching. In addition, we provide the explicit feature map such that the proposed similarity metric is embedded as a dot product. It contributes to significant speedup in similarity measurement as well as to feature transformation toward an effective vector form to which linear classifiers are directly applicable. In the experiments on various tasks, the proposed method exhibits favorable performance in both feature matching and classification.

1.227 POD: Discovering Primary Objects in Videos Based on Evolutionary Refinement of Object Recurrence, Background, and Primary Object Models

[Link to Paper]

A primary object discovery (POD) algorithm for a video sequence is proposed in this work, which is capable of discovering a primary object, as well as identifying noisy frames that do not contain the object. First, we generate object proposals for each frame. Then, we bisect each proposal into foreground and background regions, and extract features from each region. By superposing the foreground and background features, we build the object recurrence model, the background model, and the primary object model. We develop an iterative scheme to refine each model evolutionarily using the information in the other models. Finally, using the evolved primary object model, we select candidate proposals and locate the bounding box of a primary object by merging the proposals selectively. Experimental results on a challenging dataset demonstrate that the proposed POD algorithm extracts primary objects accurately and robustly.

1.228 Amplitude Modulated Video Camera - Light Separation in Dynamic Scenes

[Link to Paper]

Controlled light conditions improve considerably the performance of most computer vision algorithms. Dynamic light conditions create varying spatial changes in color and intensity across the scene. These condition, caused by a moving shadow for example, force developers to create algorithms which are robust to such variations. We suggest a computational camera which produces images that are not influenced by environmental variations in light conditions. The key insight is that many years ago, similar difficulties were already solved in radio communication; As a result each channel is immune to interference from other radio channels. Amplitude Modulated (AM) video camera separates the influence of a modulated light from other unknown light sources in the scene; Causing the AM video camera frame to appear the same - independent of the light conditions in which it was taken. We built a prototype of the AM video camera by using off the shelf hardware and tested it. AM video camera was used to demonstrate color constancy, shadow removal and contrast enhancement in real time. We show theoretically and empirically that: 1. the proposed system can produce images with similar noise levels as a standard camera. 2. The images created by such camera are almost completely immune to temporal, spatial and spectral changes in the background light. Figure 1. AM video camera enable working in stable illumination conditions independent of the background illumination

1.229 Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled

[Link to Paper]

This work presents a new approach to learning a framebased classifier on weakly labelled sequence data by embedding a CNN within an iterative EM algorithm. This allows the CNN to be trained on a vast number of example images when only loose sequence level information is available for the source videos. Although we demonstrate this in the context of hand shape recognition, the approach has wider application to any video recognition task where frame level labelling is not available. The iterative EM algorithm leverages the discriminative ability of the CNN to iteratively refine the frame level annotation and subsequent training of the CNN. By embedding the classifier within an EM framework the CNN can easily be trained on 1 million hand images. We demonstrate that the final classifier generalises over both individuals and data sets. The algorithm is evaluated on over 3000 manually labelled hand shape images of 60 different classes which will be released to the community. Furthermore, we demonstrate its use in continuous sign language recognition on two publicly available large sign language data sets, where it outperforms the current state-of-the-art by a large margin. To our knowledge no previous work has explored expectation maximization without Gaussian mixture models to exploit weak sequence labels for sign language recognition.

1.230 Sliced Wasserstein Kernels for Probability Distributions

[Link to Paper]

Optimal transport distances, otherwise known as Wasserstein distances, have recently drawn ample attention in computer vision and machine learning as powerful discrepancy measures for probability distributions. The recent developments on alternative formulations of the optimal transport have allowed for faster solutions to the problem and have revamped their practical applications in machine learning. In this paper, we exploit the widely used kernel methods and provide a family of provably positive definite kernels based on the Sliced Wasserstein distance and demonstrate the benefits of these kernels in a variety of learning tasks. Our work provides a new perspective on the application of optimal transport flavored distances through kernel methods in machine learning tasks.

1.231 HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection

[Link to Paper]

Almost all of the current top-performing object detection networks employ region proposals to guide the search for object instances. State-of-the-art region proposal methods usually need several thousand proposals to get high recall, thus hurting the detection efficiency. Although the latest Region Proposal Network method gets promising detection accuracy with several hundred proposals, it still struggles in small-size object detection and precise localization (e.g., large IoU thresholds), mainly due to the coarseness of its feature maps. In this paper, we present a deep hierarchical network, namely HyperNet, for handling region proposal generation and object detection jointly. Our HyperNet is primarily based on an elaborately designed Hyper Feature which aggregates hierarchical feature maps first and then compresses them into a uniform space. The Hyper Features well incorporate deep but highly semantic, intermediate but really complementary, and shallow but naturally high-resolution features of the image, thus enabling us to construct HyperNet by sharing them both in generating proposals and detecting objects via an end-to-end joint training strategy. For the deep VGG16 model, our method achieves completely leading recall and state-of-the-art object detection accuracy on PASCAL VOC 2007 and 2012 using only 100 proposals per image. It runs with a speed of 5 fps (including all steps) on a GPU, thus having the potential for real-time processing.

1.232 Prior-Less Compressible Structure from Motion

[Link to Paper]

Many non-rigid 3D structures are not modelled well through a low-rank subspace assumption. This is problematic when it comes to their reconstruction through Structure from Motion (SfM). We argue in this paper that a more expressive and general assumption can be made around compressible 3D structures. The vision community, however, has hitherto struggled to formulate effective strategies for recovering such structures after projection without the aid of additional priors (e.g. temporal ordering, rigid substructures, etc.). In this paper we present a "prior-less" approach to solve compressible SfM. Specifically, we demonstrate how the problem of SfM - assuming compressible 3D structures - can be theoretically characterized as a block sparse dictionary learning problem. We validate our approach experimentally by demonstrating reconstructions of 3D structures that are intractable using current state-of-theart low-rank SfM approaches.

1.233 Sparse Coding for Third-order Super-symmetric Tensor Descriptors with Application to Texture Recognition

[Link to Paper]

Super-symmetric tensors a higher-order extension of scatter matrices are becoming increasingly popular in machine learning and computer vision for modeling data statistics, co-occurrences, or even as visual descriptors. They were shown recently to outperform second-order approaches [18], however, the size of these tensors are exponential in the data dimensionality, which is a significant concern. In this paper, we study third-order supersymmetric tensor descriptors in the context of dictionary learning and sparse coding. For this purpose, we propose a novel non-linear third-order texture descriptor. Our goal is to approximate these tensors as sparse conic combinations of atoms from a learned dictionary. Apart from the significant benefits to tensor compression that this framework offers, our experiments demonstrate that the sparse coefficients produced by this scheme lead to better aggregation of high-dimensional data and showcase superior performance on two common computer vision tasks compared to the state of the art.

1.234 Incremental Object Discovery in Time-Varying Image Collections

[Link to Paper]

In this paper, we address the problem of object discovery in time-varying, large-scale image collections. A core part of our approach is a novel Limited Horizon Minimum Spanning Tree (LH-MST) structure that closely approximates the Minimum Spanning Tree at a small fraction of the latter's computational cost. Our proposed tree structure can be created in a local neighborhood of the matching graph during image retrieval and can be efficiently updated whenever the image database is extended. We show how the LH-MST can be used within both single-link hierarchical agglomerative clustering and the Iconoid Shift framework for object discovery in image collections, resulting in significant efficiency gains and making both approaches capable of incremental clustering with online updates. We evaluate our approach on a dataset of 500k images from the city of Paris and compare its results to the batch version of both clustering algorithms.

1.235 Visual Word2Vec (vis-w2v): Learning Visually Grounded Word Embeddings Using Abstract Scenes

[Link to Paper]

We propose a model to learn visually grounded word embeddings (vis-w2v) to capture visual notions of semantic relatedness. While word embeddings trained using text have been extremely successful, they cannot uncover notions of semantic relatedness implicit in our visual world. For instance, although "eats" and "stares at" seem unrelated in text, they share semantics visually. When people are eating something, they also tend to stare at the food. Grounding diverse relations like "eats" and "stares at" into vision remains

111

challenging, despite recent progress in vision. We note that the visual grounding of words depends on semantics, and not the literal pixels. We thus use abstract scenes created from clipart to provide the visual grounding. We find that the embeddings we learn capture fine-grained, visually grounded notions of semantic relatedness. We show improvements over text-only word embeddings (word2vec) on three tasks: common-sense assertion classification, visual paraphrasing and text-based image retrieval. Our code and datasets are available online.

1.236 Eye Tracking for Everyone

[Link to Paper]

From scientific research to commercial applications, eye tracking is an important tool across many domains. Despite its range of applications, eye tracking has yet to become a pervasive technology. We believe that we can put the power of eye tracking in everyone's palm by building eye tracking software that works on commodity hardware such as mobile phones and tablets, without the need for additional sensors or devices. We tackle this problem by introducing GazeCapture, the first large-scale dataset for eye tracking, containing data from over 1450 people consisting of almost 2.5M frames. Using GazeCapture, we train iTracker, a convolutional neural network for eye tracking, which achieves a significant reduction in error over previous approaches while running in real time (1015fps) on a modern mobile device. Our model achieves a prediction error of 1.71cm and 2.53cm without calibration on mobile phones and tablets respectively. With calibration, this is reduced to 1.34cm and 2.12cm. Further, we demonstrate that the features learned by iTracker generalize well to other datasets, achieving state-of-the-art results. The code, data, and models are available at http://gazecapture.csail.mit.edu.

1.237 Saliency Unified: A Deep Architecture for simultaneous Eye Fixation Prediction and Salient Object Segmentation

[Link to Paper]

Human eye fixations often correlate with locations of salient objects in the scene. However, only a handful of approaches have attempted to simultaneously address the related aspects of eye fixations and object saliency. In this work, we propose a deep convolutional neural network (CNN) capable of predicting eye fixations and segmenting salient objects in a unified framework. We design the initial network layers, shared between both the tasks, such that they capture the object level semantics and the global contextual aspects of saliency, while the deeper layers of the network address task specific aspects. In addition, our network captures saliency at multiple scales via inceptionstyle convolution

blocks. Our network shows a significant improvement over the current state-of-the-art for both eye fixation prediction and salient object segmentation across a number of challenging datasets.

1.238 Recurrent Attentional Networks for Saliency Detection

[Link to Paper]

Convolutional-deconvolution networks can be adopted to perform end-to-end saliency detection. But, they do not work well with objects of multiple scales. To overcome such a limitation, in this work, we propose a recurrent attentional convolutional-deconvolution network (RACDNN). Using spatial transformer and recurrent network units, RACDNN is able to iteratively attend to selected image sub-regions to perform saliency refinement progressively. Besides tackling the scale problem, RACDNN can also learn context-aware features from past iterations to enhance saliency refinement in future iterations. Experiments on several challenging saliency detection datasets validate the effectiveness of RACDNN, and show that RACDNN outperforms state-of-the-art saliency detection methods.

1.239 Efficient Intersection of Three Quadrics and Applications in Computer Vision

[Link to Paper]

In this paper, we present a new algorithm for finding all intersections of three quadrics. The proposed method is algebraic in nature and it is considerably more efficient than the Grobner basis and resultant-based solutions previously used in computer vision applications. We identify several computer vision problems that are formulated and solved as systems of three quadratic equations and for which our algorithm readily delivers considerably faster results. Also, we propose new formulations of three important vision problems: absolute camera pose with unknown focal length, generalized pose-and-scale, and hand-eye calibration with known translation. These new formulations allow our algorithm to significantly outperform the state-of-theart in speed.

1.240 ReconNet: Non-Iterative Reconstruction of Images from Compressively Sensed Measurements

[Link to Paper]

The goal of this paper is to present a non-iterative and more importantly an extremely fast algorithm to reconstruct images from compressively sensed (CS) random measurements. To this end, we propose a novel convolutional neural network (CNN) architecture which takes in CS measurements of an image as input and outputs an intermediate reconstruction. We call this network, ReconNet. The intermediate reconstruction is fed into an off-the-shelf denoiser to obtain the final reconstructed image. On a standard dataset of images we show significant improvements in reconstruction results (both in terms of PSNR and time complexity) over state-of-the-art iterative CS reconstruction algorithms at various measurement rates. Further, through qualitative experiments on real data collected using our block single pixel camera (SPC), we show that our network is highly robust to sensor noise and can recover visually better quality images than competitive algorithms at extremely low sensing rates of 0.1 and 0.04. To demonstrate that our algorithm can recover semantically informative images even at a low measurement rate of 0.01, we present a very robust proof of concept real-time visual tracking application.

1.241 Feature Space Optimization for Semantic Video Segmentation

[Link to Paper]

We present an approach to long-range spatio-temporal regularization in semantic video segmentation. Temporal regularization in video is challenging because both the cam- era and the scene may be in motion. Thus Euclidean distance in the space-time volume is not a good proxy for correspondence. We optimize the mapping of pixels to a Euclidean feature space so as to minimize distances between corresponding points. Structured prediction is performed by a dense CRF that operates on the optimized features. Experimental results demonstrate that the presented approach increases the accuracy and temporal consistency of semantic video segmentation.

1.242 When Naive Bayes Nearest Neighbors Meet Convolutional Neural Networks

[Link to Paper]

Since Convolutional Neural Networks (CNNs) have become the leading learning paradigm in visual recognition, Naive Bayes Nearest Neighbor (NBNN)-based classifiers have lost momentum in the community. This is because (1) such algorithms cannot use CNN activations as input features; (2) they cannot be used as final layer of CNN architectures for end-to-end training , and (3) they are generally not scalable and hence cannot handle big data. This paper proposes a framework that addresses all these issues, thus bringing back NBNNs on the map. We solve the first by extracting CNN activations from local patches at

multiple scale levels, similarly to [13]. We address simultaneously the second and third by proposing a scalable version of Naive Bayes Non-linear Learning (NBNL, [7]). Results obtained using pre-trained CNNs on standard scene and domain adaptation databases show the strength of our approach, opening a new season for NBNNs.

1.243 Thin-Slicing for Pose: Learning to Understand Pose without Explicit Pose Estimation

[Link to Paper]

We address the problem of learning a pose-aware, compact embedding that projects images with similar human poses to be placed close-by in the embedding space. The embedding function is built on a deep convolutional network, and trained with triplet-based rank constraints on real image data. This architecture allows us to learn a robust representation that captures differences in human poses by effectively factoring out variations in clothing, background, and imaging conditions in the wild. For a variety of pose-related tasks, the proposed pose embedding provides a cost-efficient and natural alternative to explicit pose estimation, circumventing challenges of localizing body joints. We demonstrate the efficacy of the embedding on pose-based image retrieval and action recognition problems.

1.244 One-Shot Learning of Scene Locations via Feature Trajectory Transfer

[Link to Paper]

The appearance of (outdoor) scenes changes considerably with the strength of certain transient attributes, such as "rainy", "dark" or "sunny". Obviously, this also affects the representation of an image in feature space, e.g., as activations at a certain CNN layer, and consequently impacts scene recognition performance. In this work, we investigate the variability in these transient attributes as a rich source of information for studying how image representations change as a function of attribute strength. In particular, we leverage a recently introduced dataset with fine-grain annotations to estimate feature trajectories for a collection of transient attributes and then show how these trajectories can be transferred to new image representations. This enables us to synthesize new data along the transferred trajectories with respect to the dimensions of the space spanned by the transient attributes. Applicability of this concept is demonstrated on the problem of oneshot recognition of scene locations. We show that data synthesized via feature trajectory transfer considerably boosts recognition performance, (1) with respect to baselines and (2) in combination with state-of-the-art approaches in oneshot learning.

1.245 Efficient Globally Optimal 2D-to-3D Deformable Shape Matching

[Link to Paper]

We propose the first algorithm for non-rigid 2D-to-3D shape matching, where the input is a 2D query shape as well as a 3D target shape and the output is a continuous matching curve represented as a closed contour on the 3D shape. We cast the problem as finding the shortest circular path on the product 3-manifold of the two shapes. We prove that the optimal matching can be computed in polynomial time with a (worst-case) complexity of O(mn2 log(n)), where m and n denote the number of vertices on the 2D and the 3D shape respectively. Quantitative evaluation confirms that the method provides excellent results for sketch-based deformable 3D shape retrieval.

1.246 A Comparative Study for Single Image Blind Deblurring

[Link to Paper]

Numerous single image blind deblurring algorithms have been proposed to restore latent sharp images under camera motion. However, these algorithms are mainly evaluated using either synthetic datasets or few selected real blurred images. It is thus unclear how these algorithms would perform on images acquired "in the wild" and how we could gauge the progress in the field. In this paper, we aim to bridge this gap. We present the first comprehensive perceptual study and analysis of single image blind deblurring using real-world blurred images. First, we collect a dataset of real blurred images and a dataset of synthetically blurred images. Using these datasets, we conduct a large-scale user study to quantify the performance of several representative state-of-the-art blind deblurring algorithms. Second, we systematically analyze subject preferences, including the level of agreement, significance tests of score differences, and rationales for preferring one method over another. Third, we study the correlation between human subjective scores and several full-reference and noreference image quality metrics. Our evaluation and analysis indicate the performance gap between synthetically blurred images and real blurred image and sheds light on future research in single image blind deblurring.

1.247 Saliency Guided Dictionary Learning for Weakly-Supervised Image Parsing

[Link to Paper]

In this paper, we propose a novel method to perform weakly-supervised image parsing based on the dictionary learning framework. To deal with the challenges caused by the

label ambiguities, we design a saliency guided weight assignment scheme to boost the discriminative dictionary learning. More specifically, with a collection of tagged images, the proposed method first conducts saliency detection and automatically infers the confidence for each semantic class to be foreground or background. These clues are then incorporated to learn the dictionaries, the weights, as well as the sparse representation coefficients in the meanwhile. Once obtained the coefficients of a superpixel, we use a sparse representation classifier to determine its semantic label. The approach is validated on the MSRC21, PASCAL VOC07, and VOC12 datasets. Experimental results demonstrate the encouraging performance of our approach in comparison with some state-of-the-arts.

1.248 Loss Functions for Top-k Error: Analysis and Insights

[Link to Paper]

In order to push the performance on realistic computer vision tasks, the number of classes in modern benchmark datasets has significantly increased in recent years. This increase in the number of classes comes along with increased ambiguity between the class labels, raising the question if top-1 error is the right performance measure. In this paper, we provide an extensive comparison and evaluation of established multiclass methods comparing their top-k performance both from a practical as well as from a theoretical perspective. Moreover, we introduce novel top-k loss functions as modifications of the softmax and the multiclass SVM losses and provide efficient optimization schemes for them. In the experiments, we compare on various datasets all of the proposed and established methods for top-k error optimization. An interesting insight of this paper is that the softmax loss yields competitive top-k performance for all k simultaneously. For a specific top-k error, our new topk losses lead typically to further improvements while being faster to train than the softmax.

1.249 TI-POOLING: transformation-invariant pooling for feature learning in Convolutional Neural Networks

[Link to Paper]

In this paper we present a deep neural network topology that incorporates a simple to implement transformationinvariant pooling operator (TI-POOLING). This operator is able to efficiently handle prior knowledge on nuisance variations in the data, such as rotation or scale changes. Most current methods usually make use of dataset augmentation to address this issue, but this requires larger number of model parameters and more training data, and results in significantly increased training time and larger chance of under- or overfitting. The main reason for these drawbacks is that that the learned model needs to capture

117

adequate features for all the possible transformations of the input. On the other hand, we formulate features in convolutional neural networks to be transformation-invariant. We achieve that using parallel siamese architectures for the considered transformation set and applying the TI-POOLING operator on their outputs before the fully-connected layers. We show that this topology internally finds the most optimal "canonical" instance of the input image for training and therefore limits the redundancy in learned features. This more efficient use of training data results in better performance on popular benchmark datasets with smaller number of parameters when comparing to standard convolutional neural networks with dataset augmentation and to other baselines.

1.250   Fast Algorithms for Convolutional Neural Networks

[Link to Paper]

Deep convolutional neural networks take GPU-days of computation to train on large data sets. Pedestrian detection for self driving cars requires very low latency. Image recognition for mobile phones is constrained by limited processing resources. The success of convolutional neural networks in these situations is limited by how fast we can compute them. Conventional FFT based convolution is fast for large filters, but state of the art convolutional neural networks use small, 3 3 filters. We introduce a new class of fast algorithms for convolutional neural networks using Winograd's minimal filtering algorithms. The algorithms compute minimal complexity convolution over small tiles, which makes them fast with small filters and small batch sizes. We benchmark a GPU implementation of our algorithm with the VGG network and show state of the art throughput at batch sizes from 1 to 64.

1.251   Closed-Form Training of Mahalanobis Distance for Supervised Clustering

[Link to Paper]

Clustering is the task of grouping a set of objects so that objects in the same cluster are more similar to each other than to those in other clusters. The crucial step in most clustering algorithms is to find an appropriate similarity metric, which is both challenging and problem-dependent. Supervised clustering approaches, which can exploit labeled clustered training data that share a common metric with the test set, have thus been proposed. Unfortunately, current metric learning approaches for supervised clustering do not scale to large or even medium-sized datasets. In this paper, we propose a new structured Mahalanobis Distance Metric Learning method for supervised clustering. We formulate our problem as an instance of large margin structured prediction and prove that it can be

solved very efficiently in closed-form. The complexity of our method is (in most cases) linear in the size of the training dataset. We further reveal a striking similarity between our approach and multivariate linear regression. Experiments on both synthetic and real datasets confirm several orders of magnitude speedup while still achieving state-of-the-art performance.

1.252 Conformal Surface Alignment with Optimal Mobius Search

[Link to Paper]

Deformations of surfaces with the same intrinsic shape can often be described accurately by a conformal model. A major focus of computational conformal geometry is the estimation of the conformal mapping that aligns a given pair of object surfaces. The uniformization theorem enables this task to be acccomplished in a canonical 2D domain, wherein the surfaces can be aligned using a Mobius transformation. Current algorithms for estimating Mobius transformations, however, often cannot provide satisfactory alignment or are computationally too costly. This paper introduces a novel globally optimal algorithm for estimating Mobius transformations to align surfaces that are topological discs. Unlike previous methods, the proposed algorithm deterministically calculates the best transformation, without requiring good initializations. Further, our algorithm is also much faster than previous techniques in practice. We demonstrate the efficacy of our algorithm on data commonly used in computational conformal geometry.

1.253 Fast ConvNets Using Group-wise Brain Damage

[Link to Paper]

We revisit the idea of brain damage, i.e. the pruning of the coefficients of a neural network, and suggest how brain damage can be modified and used to speedup convolutional layers in ConvNets. The approach uses the fact that many efficient implementations reduce generalized convolutions to matrix multiplications. The suggested brain damage process prunes the convolutional kernel tensor in a group-wise fashion. After such pruning, convolutions can be reduced to multiplications of thinned dense matrices, which leads to speedup. We investigate different ways to add group-wise prunning to the learning process, and show that severalfold speedups of convolutional layers can be attained using group-sparsity regularizers. Our approach can adjust the shapes of the receptive fields in the convolutional layers, and even prune excessive feature maps from ConvNets, all in data-driven way.

1.254 Automatic Content-Aware Color and Tone Stylization

[Link to Paper]

We introduce a new technique that automatically generates diverse, visually compelling stylizations for a photograph in an unsupervised manner. We achieve this by learning style ranking for a given input using a large photo collection and selecting a diverse subset of matching styles for final style transfer. We also propose an improved technique that transfers the global color and tone of the chosen exemplars to the input photograph while avoiding the common visual artifacts produced by the existing style transfer methods. Together, our style selection and transfer techniques produce compelling, artifact-free results on a wide range of input photographs, and a user study shows that our results are preferred over other techniques.

1.255 Consensus of Non-Rigid Reconstructions

[Link to Paper]

Recently, there have been many progresses for the problem of non-rigid structure reconstruction based on 2D trajectories, but it is still challenging to deal with complex deformations or restricted view ranges. Promising alternatives are the piecewise reconstruction approaches, which divide trajectories into several local parts and stitch their individual reconstructions to produce an entire 3D structure. These methods show the state-of-the-art performance, however, most of them are specialized for relatively smooth surfaces and some are quite complicated. Meanwhile, it has been reported numerously in the field of pattern recognition that obtaining consensus from many weak hypotheses can give a strong, powerful result. Inspired by these reports, in this paper, we push the concept of part-based reconstruction to the limit: Instead of considering the parts as explicitly-divided local patches, we draw a large number of small random trajectory sets. From their individual reconstructions, we pull out a statistic of each 3D point to retrieve a strong reconstruction, of which the procedure can be expressed as a sparse l1-norm minimization problem. In order to resolve the reflection ambiguity between weak (and possibly bad) reconstructions, we propose a novel optimization framework which only involves a single eigenvalue decomposition. The proposed method can be applied to any type of data and outperforms the existing methods for the benchmark sequences, even though it is composed of a few, simple steps. Furthermore, it is easily parallelizable, which is another advantage.

1.256 Deep Saliency with Encoded Low level Distance Map and High Level Features

[Link to Paper]

Recent advances in saliency detection have utilized deep learning to obtain high level features to detect salient regions in a scene. These advances have demonstrated superior results over previous works that utilize hand-crafted low level features for saliency detection. In this paper, we demonstrate that hand-crafted features can provide complementary information to enhance performance of saliency detection that utilizes only high level features. Our method utilizes both high level and low level features for saliency detection under a unified deep learning framework. The high level features are extracted using the VGG-net, and the low level features are compared with other parts of an image to form a low level distance map. The low level distance map is then encoded using a convolutional neural network(CNN) with multiple 1 1 convolutional and ReLU layers. We concatenate the encoded low level distance map and the high level features, and connect them to a fully connected neural network classifier to evaluate the saliency of a query region. Our experiments show that our method can further improve the performance of stateof-the-art deep learning-based saliency detection methods.

1.257 Laplacian Patch-Based Image Synthesis

[Link to Paper]

Patch-based image synthesis has been enriched with global optimization on the image pyramid. Successively, the gradient-based synthesis has improved structural coherence and details. However, the gradient operator is directional and inconsistent and requires computing multiple operators. It also introduces a significantly heavy computational burden to solve the Poisson equation that often accompanies artifacts in non-integrable gradient fields. In this paper, we propose a patch-based synthesis using a Laplacian pyramid to improve searching correspondence with enhanced awareness of edge structures. Contrary to the gradient operators, the Laplacian pyramid has the advantage of being isotropic in detecting changes to provide more consistent performance in decomposing the base structure and the detailed localization. Furthermore, it does not require heavy computation as it employs approximation by the differences of Gaussians. We examine the potentials of the Laplacian pyramid for enhanced edge-aware correspondence search. We demonstrate the effectiveness of the Laplacian-based approach over the state-of-the-art patchbased image synthesis methods.

1.258 Recursive Recurrent Nets with Attention Modeling for OCR in the Wild

[Link to Paper]

We present recursive recurrent neural networks with attention modeling (R2AM) for lexicon-free optical character recognition in natural scene images. The primary advantages of the proposed method are: (1) use of recursive convolutional neural networks (CNNs), which allow for parametrically efficient and effective image feature extraction; (2) an implicitly learned character-level language model, embodied in a recurrent neural network which avoids the need to use N-grams; and (3) the use of a soft-attention mechanism, allowing the model to selectively exploit image features in a coordinated way, and allowing for end-to-end training within a standard backpropagation framework. We validate our method with state-of-the-art performance on challenging benchmark datasets: Street View Text, IIIT5k, ICDAR and Synth90k.

1.259 Comparative Deep Learning of Hybrid Representations for Image Recommendations

[Link to Paper]

In many image-related tasks, learning expressive and discriminative representations of images is essential, and deep learning has been studied for automating the learning of such representations. Some user-centric tasks, such as image recommendations, call for effective representations of not only images but also preferences and intents of users over images. Such representations are termed hybrid and addressed via a deep learning approach in this paper. We design a dual-net deep network, in which the two subnetworks map input images and preferences of users into a same latent semantic space, and then the distances between images and users in the latent space are calculated to make decisions. We further propose a comparative deep learning (CDL) method to train the deep network, using a pair of images compared against one user to learn the pattern of their relative distances. The CDL embraces much more training data than naive deep learning, and thus achieves superior performance than the latter, with no cost of increasing network complexity. Experimental results with real-world data sets for image recommendations have shown the proposed dual-net network and CDL greatly outperform other stateof-the-art image recommendation solutions.

1.260 A Multi-Level Contextual Model For Person Recognition in Photo Albums

[Link to Paper]

In this work, we present a new framework for person recognition in photo albums that exploits contextual cues at multiple levels, spanning individual persons, individual photos, and photo groups. Through experiments, we show that the information available at each of these distinct contextual levels provides complementary cues as to person identities. At the person level, we leverage clothing and body appearance in addition to facial appearance, and to compensate for instances where the faces are not visible. At the photo level we leverage a learned prior on the joint distribution of identities on the same photo to guide the identity assignments. Going beyond a single photo, we are able to infer natural groupings of photos with shared context in an unsupervised manner. By exploiting this shared contextual information, we are able to reduce the identity search space and exploit higher intra-personal appearance consistency within photo groups. Our new framework enables efficient use of these complementary multi-level contextual cues to improve overall recognition rates on the photo album person recognition task, as demonstrated through state-of-theart results on a challenging public dataset. Our results outperform competing methods by a significant margin, while being computationally efficient and practical in a real world application.

1.261 Combining Markov Random Fields and Convolutional Neural Networks for Image Synthesis

[Link to Paper]

This paper studies a combination of generative Markov random field (MRF) models and discriminatively trained deep convolutional neural networks (dCNNs) for synthesizing 2D images. The generative MRF acts on higher-levels of a dCNN feature pyramid, controling the image layout at an abstract level. We apply the method to both photographic and non-photo-realistic (artwork) synthesis tasks. The MRF regularizer prevents over-excitation artifacts and reduces implausible feature mixtures common to previous dCNN inversion approaches, permitting synthezing photographic content with increased visual plausibility. Unlike standard MRF-based texture synthesis, the combined system can both match and adapt local features with considerable variability, yielding results far out of reach of classic generative MRF methods.

1.262 Conditional Graphical Lasso for Multi-label Image Classification

[Link to Paper]

Multi-label image classification aims to predict multiple labels for a single image which contains diverse content. By utilizing label correlations, various techniques have been

developed to improve classification performance. However, current existing methods either neglect image features when exploiting label correlations or lack the ability to learn image-dependent conditional label structures. In this paper, we develop conditional graphical Lasso (CGL) to handle these challenges. CGL provides a unified Bayesian framework for structure and parameter learning conditioned on image features. We formulate the multi-label prediction as CGL inference problem, which is solved by a mean field variational approach. Meanwhile, CGL learning is efficient due to a tailored proximal gradient procedure by applying the maximum a posterior (MAP) methodology. CGL performs competitively for multi-label image classification on benchmark datasets MULAN scene, PASCAL VOC 2007 and PASCAL VOC 2012, compared with the state-of-the-art multi-label classification algorithms.

1.263 Coordinating Multiple Disparity Proposals for Stereo Computation

[Link to Paper]

While great progress has been made in stereo computation over the last decades, large textureless regions remain challenging. Segment-based methods can tackle this problem properly, but their performances are sensitive to the segmentation results. In this paper, we alleviate the sensitivity by generating multiple proposals on absolute and relative disparities from multi-segmentations. These proposals supply rich descriptions of surface structures. Especially, the relative disparity between distant pixels can encode the large structure, which is critical to handle the large textureless regions. The proposals are coordinated by point-wise competition and pairwise collaboration within a MRF model. During inference, a dynamic programming is performed in different directions with various step sizes, so the longrange connections are better preserved. In the experiments, we carefully analyzed the effectiveness of the major components. Results on the 2014 Middlebury and KITTI 2015 stereo benchmark show that our method is comparable to state-of-the-art.

1.264 Deep Contrast Learning for Salient Object Detection

[Link to Paper]

Salient object detection has recently witnessed substantial progress due to powerful features extracted using deep convolutional neural networks (CNNs). However, existing CNN-based methods operate at the patch level instead of the pixel level. Resulting saliency maps are typically blurry, especially near the boundary of salient objects. Furthermore, image patches are treated as independent samples even when they are overlapping, giving rise to

significant redundancy in computation and storage. In this paper, we propose an end-to-end deep contrast network to overcome the aforementioned limitations. Our deep network consists of two complementary components, a pixel-level fully convolutional stream and a segment-wise spatial pooling stream. The first stream directly produces a saliency map with pixel-level accuracy from an input image. The second stream extracts segment-wise features very efficiently, and better models saliency discontinuities along object boundaries. Finally, a fully connected CRF model can be optionally incorporated to improve spatial coherence and contour localization in the fused result from these two streams. Experimental results demonstrate that our deep model significantly improves the state of the art.

1.265 Fast Algorithms for Linear and Kernel SVM+

[Link to Paper]

The SVM+ approach has shown excellent performance in visual recognition tasks for exploiting privileged information in the training data. In this paper, we propose two efficient algorithms for solving the linear and kernel SVM+, respectively. For linear SVM+, we absorb the bias term into the weight vector, and formulate a new optimization problem with simpler constraints in the dual form. Then, we develop an efficient dual coordinate descent algorithm to solve the new optimization problem. For kernel SVM+, we further apply the 2-loss, which leads to a simpler optimization problem in the dual form with only half of dual variables when compared with the dual form of the original SVM+ method. More interestingly, we show that our new dual problem can be efficiently solved by using the SMO algorithm of the one-class SVM problem. Comprehensive experiments on three datasets clearly demonstrate that our proposed algorithms achieve significant speed-up than the state-of-the-art solvers for linear and kernel SVM+.

1.266 Iterative Instance Segmentation

[Link to Paper]

Existing methods for pixel-wise labelling tasks generally disregard the underlying structure of labellings, often leading to predictions that are visually implausible. While incorporating structure into the model should improve prediction quality, doing so is challenging manually specifying the form of structural constraints may be impractical and inference often becomes intractable even if structural constraints are given. We sidestep this problem by reducing structured prediction to a sequence of unconstrained prediction problems and demonstrate that this approach is capable of automatically discovering priors on shape, contiguity of region predictions and smoothness of region contours from data without any a priori specification. On the instance segmentation task, this method outperforms the

state-of-the-art, achieving a mean APr of 63.6% at 50% overlap and 43.3% at 70% overlap.

1.267 Learning Weight Uncertainty with Stochastic Gradient MCMC for Shape Classification

[Link to Paper]

Learning the representation of shape cues in 2D & 3D objects for recognition is a fundamental task in computer vision. Deep neural networks (DNNs) have shown promising performance on this task. Due to the large variability of shapes, accurate recognition relies on good estimates of model uncertainty, ignored in traditional training of DNNs, typically learned via stochastic optimization. This paper leverages recent advances in stochastic gradient Markov Chain Monte Carlo (SG-MCMC) to learn weight uncertainty in DNNs. It yields principled Bayesian interpretations for the commonly used Dropout/DropConnect techniques and incorporates them into the SG-MCMC framework. Extensive experiments on 2D & 3D shape datasets and various DNN models demonstrate the superiority of the proposed approach over stochastic optimization. Our approach yields higher recognition accuracy when used in conjunction with Dropout and Batch-Normalization.

1.268 Rain Streak Removal Using Layer Priors

[Link to Paper]

This paper addresses the problem of rain streak removal from a single image. Rain streaks impair visibility of an image and introduce undesirable interference that can severely affect the performance of computer vision algorithms. Rain streak removal can be formulated as a layer decomposition problem, with a rain streak layer superimposed on a background layer containing the true scene content. Existing decomposition methods that address this problem employ either dictionary learning methods or impose a low rank structure on the appearance of the rain streaks. While these methods can improve the overall visibility, they tend to leave too many rain streaks in the background image or over-smooth the background image. In this paper, we propose an effective method that uses simple patch-based priors for both the background and rain layers. These priors are based on Gaussian mixture models and can accommodate multiple orientations and scales of the rain streaks. This simple approach removes rain streaks better than the existing methods qualitatively and quantitatively. We overview our method and demonstrate its effectiveness over prior work on a number of examples. Input Rain patch Background Rain streaks Figure 1: Upper Row: An example rain image and a zoomed-in patch that mainly contains annoying effect of rain streaks. Our method learns a rain streak layer prior on this region and used it for

layer separation. Lower Row: The background and rain layers recovered by our proposed method, respectively.

1.269 Recognizing Car Fluents

[Link to Paper]

(Abstract could not be extracted from PDF)

1.270 Rotational Crossed-Slit Light Fields

[Link to Paper]

Light fields (LFs) are image-based representation that records the radiance along all rays along every direction through every point in space. Traditionally LFs are acquired by using a 2D grid of evenly spaced pinhole cameras or by translating a pinhole camera along the 2D grid using a robot arm. In this paper, we present a novel LF sampling scheme by exploiting a special non-centric camera called the crossed-slit or XSlit camera. An XSlit camera acquires rays that simultaneously pass through two oblique slits. We show that, instead of translating the camera as in the pinhole case, we can effectively sample the LF by rotating individual or both slits while keeping the camera fixed. This leads a "fixed-location" LF acquisition scheme. We further show through theoretical analysis and experiments that the resulting XSlit LFs provide several advantages: they provide more dense spatial-angular sampling, are amenable multi-view stereo matching and volumetric reconstruction, and can synthesize unique refocusing effects.

1.271 Simultaneous Clustering and Model Selection for Tensor Affinities

[Link to Paper]

Estimating the number of clusters remains a difficult model selection problem. We consider this problem in the domain where the affinity relations involve groups of more than two nodes. Building on the previous formulation for the pairwise affinity case, we exploit the mathematical structures in the higher order case. We express the original minimal-rank and positive semi-definite (PSD) constraints in a form amenable for numerical implementation, as the original constraints are either intractable or even undefined in general in the higher order case. To scale to large problem sizes, we also propose an alternative formulation, so that it can be efficiently solved via stochastic optimization in an online fashion. We evaluate our algorithm with different applications to demonstrate its superiority, and show it can adapt to varying levels of unbalancedness of clusters.

1.272 TGIF: A New Dataset and Benchmark on Animated GIF Description

[Link to Paper]

With the recent popularity of animated GIFs on social media, there is need for ways to index them with rich metadata. To advance research on animated GIF understanding, we collected a new dataset, Tumblr GIF (TGIF), with 100K animated GIFs from Tumblr and 120K natural language descriptions obtained via crowdsourcing. The motivation for this work is to develop a testbed for image sequence description systems, where the task is to generate natural language descriptions for animated GIFs or video clips. To ensure a high quality dataset, we developed a series of novel quality controls to validate free-form text input from crowdworkers. We show that there is unambiguous association between visual content and natural language descriptions in our dataset, making it an ideal benchmark for the visual content captioning task. We perform extensive statistical analyses to compare our dataset to existing image and video description datasets. Next, we provide baseline results on the animated GIF description task, using three representative techniques: nearest neighbor, statistical machine translation, and recurrent neural networks. Finally, we show that models fine-tuned from our animated GIF description dataset can be helpful for automatic movie description.

1.273 Unsupervised Learning of Edges

[Link to Paper]

Data-driven approaches for edge detection have proven effective and achieve top results on modern benchmarks. However, all current data-driven edge detectors require manual supervision for training in the form of hand-labeled region segments or object boundaries. Specifically, human annotators mark semantically meaningful edges which are subsequently used for training. Is this form of strong, highlevel supervision actually necessary to learn to accurately detect edges? In this work we present a simple yet effective approach for training edge detectors without human supervision. To this end we utilize motion, and more specifically, the only input to our method is noisy semi-dense matches between frames. We begin with only a rudimentary knowledge of edges (in the form of image gradients), and alternate between improving motion estimation and edge detection in turn. Using a large corpus of video data, we show that edge detectors trained using our unsupervised scheme approach the performance of the same methods trained with full supervision (within 3-5%). Finally, we show that when using a deep network for the edge detector, our approach provides a novel pre-training scheme for object detection.

1.274 VLAD3: Encoding Dynamics of Deep Features for Action Recognition

[Link to Paper]

Previous approaches to action recognition with deep features tend to process video frames only within a small temporal region, and do not model long-range dynamic information explicitly. However, such information is important for the accurate recognition of actions, especially for the discrimination of complex activities that share sub-actions, and when dealing with untrimmed videos. Here, we propose a representation, VLAD for Deep Dynamics (VLAD3), that accounts for different levels of video dynamics. It captures short-term dynamics with deep convolutional neural network features, relying on linear dynamic systems (LDS) to model medium-range dynamics. To account for long-range inhomogeneous dynamics, a VLAD descriptor is derived for the LDS and pooled over the whole video, to arrive at the final VLAD3 representation. An extensive evaluation was performed on Olympic Sports, UCF101 and THUMOS15, where the use of the VLAD3 representation leads to stateof-the-art results.

1.275 Weakly Supervised Object Localization with Progressive Domain Adaptation

[Link to Paper]

We address the problem of weakly supervised object localization where only image-level annotations are available for training. Many existing approaches tackle this problem through object proposal mining. However, a substantial amount of noise in object proposals causes ambiguities for learning discriminative object models. Such approaches are sensitive to model initialization and often converge to an undesirable local minimum. In this paper, we address this problem by progressive domain adaptation with two main steps: classification adaptation and detection adaptation. In classification adaptation, we transfer a pre-trained network to our multi-label classification task for recognizing the presence of a certain object in an image. In detection adaptation, we first use a mask-out strategy to collect class-specific object proposals and apply multiple instance learning to mine confident candidates. We then use these selected object proposals to fine-tune all the layers, resulting in a fully adapted detection network. We extensively evaluate the localization performance on the PASCAL VOC and ILSVRC datasets and demonstrate significant performance improvement over the state-of-the-art methods.

1.276 Reversible Recursive Instance-level Object Segmentation

[Link to Paper]

In this work, we propose a novel Reversible Recursive Instance-level Object Segmentation (R2-IOS) framework to address the challenging instance-level object segmentation task. R2-IOS consists of a reversible proposal refinement sub-network that predicts bounding box offsets for refining the object proposal locations, and an instance-level segmentation sub-network that generates the foreground mask of the dominant object instance in each proposal. By being recursive, R2-IOS iteratively optimizes the two subnetworks during joint training, in which the refined object proposals and improved segmentation predictions are alternately fed into each other to progressively increase the network capabilities. By being reversible, the proposal refinement sub-network adaptively determines an optimal number of refinement iterations required for each proposal during both training and testing. Furthermore, to handle multiple overlapped instances within a proposal, an instance-aware denoising autoencoder is introduced into the segmentation sub-network to distinguish the dominant object from other distracting instances. Extensive experiments on the challenging PASCAL VOC 2012 benchmark well demonstrate the superiority of R2-IOS over other state-of-the-art methods. In particular, the APr over 20 classes at 0.5 IoU achieves 66.7%, which significantly outperforms the results of 58.7% by PFN [17] and 46.3% by [22].

1.277 Semantic Object Parsing with Local-Global Long Short-Term Memory

[Link to Paper]

Semantic object parsing is a fundamental task for understanding objects in detail in computer vision community, where incorporating multi-level contextual information is critical for achieving such fine-grained pixel-level recognition. Prior methods often leverage the contextual information through post-processing predicted confidence maps. In this work, we propose a novel deep Local-Global Long Short-Term Memory (LG-LSTM) architecture to seamlessly incorporate short-distance and long-distance spatial dependencies into the feature learning over all pixel positions. In each LG-LSTM layer, local guidance from neighboring positions and global guidance from the whole image are imposed on each position to better exploit complex local and global contextual information. Individual LSTMs for distinct spatial dimensions are also utilized to intrinsically capture various spatial layouts of semantic parts in the images, yielding distinct hidden and memory cells of each position for each dimension. In our parsing approach, several LG-LSTM layers are stacked and appended to the intermediate convolutional layers to directly enhance visual features, allowing network parameters to be learned in an end-to-end way. The long chains of sequential computation by stacked LG-LSTM layers also enable each pixel to sense a much larger region for inference benefiting from the memorization of previous dependencies in all positions along all dimensions. Comprehensive evaluations on three public datasets well demonstrate the significant superiority of our LG-LSTM over other state-of-the-art

methods.

1.278 A Hierarchical Pose-Based Approach to Complex Action Understanding Using Dictionaries of Actionlets and Motion Poselets

[Link to Paper]

In this paper, we introduce a new hierarchical model for human action recognition using body joint locations. Our model can categorize complex actions in videos, and perform spatio-temporal annotations of the atomic actions that compose the complex action being performed. That is, for each atomic action, the model generates temporal action annotations by estimating its starting and ending times, as well as, spatial annotations by inferring the human body parts that are involved in executing the action. Our model includes three key novel properties: (i) it can be trained with no spatial supervision, as it can automatically discover active body parts from temporal action annotations only; (ii) it jointly learns flexible representations for motion poselets and actionlets that encode the visual variability of body parts and atomic actions; (iii) a mechanism to discard idle or non-informative body parts which increases its robustness to common pose estimation errors. We evaluate the performance of our method using multiple action recognition benchmarks. Our model consistently outperforms baselines and state-of-the-art action recognition methods.

1.279 Deep Structured Scene Parsing by Learning with Image Descriptions

[Link to Paper]

This paper addresses a fundamental problem of scene understanding: How to parse the scene image into a structured configuration (i.e., a semantic object hierarchy with object interaction relations) that finely accords with human perception. We propose a deep architecture consisting of two networks: i) a convolutional neural network (CNN) extracting the image representation for pixelwise object labeling and ii) a recursive neural network (RNN) discovering the hierarchical object structure and the inter-object relations. Rather than relying on elaborative user annotations (e.g., manually labeling semantic maps and relations), we train our deep model in a weakly-supervised manner by leveraging the descriptive sentences of the training images. Specifically, we decompose each sentence into a semantic tree consisting of nouns and verb phrases, and facilitate these trees discovering the configurations of the training images. Once these scene configurations are determined, then the parameters of both the CNN and RNN are updated accordingly by back propagation. The entire model training is accomplished through an Expectation-Maximization method. Extensive experiments suggest that our model is capable of producing meaningful

and structured scene configurations and achieving more favorable scene labeling performance on PASCAL VOC 2012 over other state-of-theart weakly-supervised methods.

1.280 Efficient Piecewise Training of Deep Structured Models for Semantic Segmentation

[Link to Paper]

Recent advances in semantic image segmentation have mostly been achieved by training deep convolutional neural networks (CNNs). We show how to improve semantic segmentation through the use of contextual information; specifically, we explore 'patch-patch' context between image regions, and 'patch-background' context. For learning from the patch-patch context, we formulate Conditional Random Fields (CRFs) with CNN-based pairwise potential functions to capture semantic correlations between neighboring patches. Efficient piecewise training of the proposed deep structured model is then applied to avoid repeated expensive CRF inference for back propagation. For capturing the patch-background context, we show that a network design with traditional multi-scale image input and sliding pyramid pooling is effective for improving performance. Our experimental results set new state-of-the-art performance on a number of popular semantic segmentation datasets, including NYUDv2, PASCAL VOC 2012, PASCAL-Context, and SIFT-flow. In particular, we achieve an intersection-overunion score of 78.0 on the challenging PASCAL VOC 2012 dataset.

1.281 Learning Compact Binary Descriptors with Unsupervised Deep Neural Networks

[Link to Paper]

In this paper, we propose a new unsupervised deep learning approach called DeepBit to learn compact binary descriptor for efficient visual object matching. Unlike most existing binary descriptors which were designed with random projections or linear hash functions, we develop a deep neural network to learn binary descriptors in an unsupervised manner. We enforce three criterions on binary codes which are learned at the top layer of our network: 1) minimal loss quantization, 2) evenly distributed codes and 3) uncorrelated bits. Then, we learn the parameters of the networks with a back-propagation technique. Experimental results on three different visual analysis tasks including image matching, image retrieval, and object recognition clearly demonstrate the effectiveness of the proposed approach.

1.282 ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation

[Link to Paper]

Large-scale data is of crucial importance for learning semantic segmentation models, but annotating per-pixel masks is a tedious and inefficient procedure. We note that for the topic of interactive image segmentation, scribbles are very widely used in academic research and commercial software, and are recognized as one of the most userfriendly ways of interacting. In this paper, we propose to use scribbles to annotate images, and develop an algorithm to train convolutional networks for semantic segmentation supervised by scribbles. Our algorithm is based on a graphical model that jointly propagates information from scribbles to unmarked pixels and learns network parameters. We present competitive object semantic segmentation results on the PASCAL VOC dataset by using scribbles as annotations. Scribbles are also favored for annotating stuff (e.g., water, sky, grass) that has no well-defined shape, and our method shows excellent results on the PASCALCONTEXT dataset thanks to extra inexpensive scribble annotations. Our scribble annotations on PASCAL VOC are available at http://research.microsoft.com/en-us/um/ people/jifdai/downloads/scribble_sup.

1.283 Visualizing and Understanding Deep Texture Representations

[Link to Paper]

A number of recent approaches have used deep convolutional neural networks (CNNs) to build texture representations. Nevertheless, it is still unclear how these models represent texture and invariances to categorical variations. This work conducts a systematic evaluation of recent CNN-based texture descriptors for recognition and attempts to understand the nature of invariances captured by these representations. First we show that the recently proposed bilinear CNN model [25] is an excellent generalpurpose texture descriptor and compares favorably to other CNN-based descriptors on various texture and scene recognition benchmarks. The model is translationally invariant and obtains better accuracy on the ImageNet dataset without requiring spatial jittering of data compared to corresponding models trained with spatial jittering. Based on recent work [13, 28] we propose a technique to visualize pre-images, providing a means for understanding categorical properties that are captured by these representations. Finally, we show preliminary results on how a unified parametric model of texture analysis and synthesis can be used for attribute-based image manipulation, e.g. to make an image more swirly, honeycombed, or knitted. The source code and additional visualizations are available at http://vis-www.cs.umass.edu/texture.

1.284 The Multiverse Loss for Robust Transfer Learning

[Link to Paper]

Deep learning techniques are renowned for supporting effective transfer learning. However, as we demonstrate, the transferred representations support only a few modes of separation and much of its dimensionality is unutilized. In this work, we suggest to learn, in the source domain, multiple orthogonal classifiers. We prove that this leads to a reduced rank representation, which, however, supports more discriminative directions. Interestingly, the softmax probabilities produced by the multiple classifiers are likely to be identical. Experimental results, on CIFAR-100 and LFW, further demonstrate the effectiveness of our method.

1.285 DHSNet: Deep Hierarchical Saliency Network for Salient Object Detection

[Link to Paper]

Traditional1 salient object detection models often use hand-crafted features to formulate contrast and various prior knowledge, and then combine them artificially. In this work, we propose a novel end-to-end deep hierarchical saliency network (DHSNet) based on convolutional neural networks for detecting salient objects. DHSNet first makes a coarse global prediction by automatically learning various global structured saliency cues, including global contrast, objectness, compactness, and their optimal combination. Then a novel hierarchical recurrent convolutional neural network (HRCNN) is adopted to further hierarchically and progressively refine the details of saliency maps step by step via integrating local context information. The whole architecture works in a global to local and coarse to fine manner. DHSNet is directly trained using whole images and corresponding ground truth saliency masks. When testing, saliency maps can be generated by directly and efficiently feedforwarding testing images through the network, without relying on any other techniques. Evaluations on four benchmark datasets and comparisons with other 11 state-of-the-art algorithms demonstrate that DHSNet not only shows its significant superiority in terms of performance, but also achieves a real-time speed of 23 FPS on modern GPUs.

1.286 DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations

[Link to Paper]

Recent advances in clothes recognition have been driven by the construction of clothes datasets. Existing datasets are limited in the amount of annotations and are difficult to cope with the various challenges in real-world applications. In this work, we introduce Deep-Fashion1, a large-scale clothes dataset with comprehensive annotations. It contains over 800,000 images, which are richly annotated with massive attributes, clothing landmarks, and correspondence of images taken under different scenarios including store, street snapshot, and consumer. Such rich annotations enable the development of powerful algorithms in clothes recognition and facilitating future researches. To demonstrate the advantages of DeepFashion, we propose a new deep model, namely FashionNet, which learns clothing features by jointly predicting clothing attributes and landmarks. The estimated landmarks are then employed to pool or gate the learned features. It is optimized in an iterative manner. Extensive experiments demonstrate the effectiveness of FashionNet and the usefulness of DeepFashion.

1.287 Deep Relative Distance Learning: Tell the Difference Between Similar Vehicles

[Link to Paper]

The growing explosion in the use of surveillance cameras in public security highlights the importance of vehicle search from a large-scale image or video database. However, compared with person re-identification or face recognition, vehicle search problem has long been neglected by researchers in vision community. This paper focuses on an interesting but challenging problem, vehicle re-identification (a.k.a precise vehicle search). We propose a Deep Relative Distance Learning (DRDL) method which exploits a two-branch deep convolutional network to project raw vehicle images into an Euclidean space where distance can be directly used to measure the similarity of arbitrary two vehicles. To further facilitate the future research on this problem, we also present a carefully-organized largescale image database "VehicleID", which includes multiple images of the same vehicle captured by different realworld cameras in a city. We evaluate our DRDL method on our VehicleID dataset and another recently-released vehicle model classification dataset "CompCars" in three sets of experiments: vehicle re-identification, vehicle model verification and vehicle retrieval. Experimental results show that our method can achieve promising results and outperforms several state-of-the-art approaches.

1.288 Deep Supervised Hashing for Fast Image Retrieval

[Link to Paper]

In this paper, we present a new hashing method to learn compact binary codes for highly efficient image retrieval on large-scale datasets. While the complex image appearance variations still pose a great challenge to reliable retrieval, in light of the recent progress of Convolutional Neural Networks (CNNs) in learning robust image representation on various vision tasks, this paper proposes a novel Deep Supervised Hashing (DSH) method to learn compact similarity-preserving binary code for the huge body of image data. Specifically, we devise a CNN architecture that takes pairs of images (similar/dissimilar) as training inputs and encourages the output of each image to approximate discrete values (e.g. +1/-1). To this end, a loss function is elaborately designed to maximize the discriminability of the output space by encoding the supervised information from the input image pairs, and simultaneously imposing regularization on the real-valued outputs to approximate the desired discrete values. For image retrieval, new-coming query images can be easily encoded by propagating through the network and then quantizing the network outputs to binary codes representation. Extensive experiments on two large scale datasets CIFAR-10 and NUS-WIDE show the promising performance of our method compared with the state-of-the-arts.

1.289 Highway Vehicle Counting in Compressed Domain

[Link to Paper]

This paper presents a highway vehicle counting method in compressed domain, aiming at achieving acceptable estimation performance approaching the pixel-domain methods. Such a task essentially is challenging because the available information (e.g. motion vector) to describe vehicles in videos is quite limited and inaccurate, and the vehicle count in realistic traffic scenes always varies greatly. To tackle this issue, we first develop a batch of low-level features, which can be extracted from the encoding metadata of videos, to mitigate the informational insufficiency of compressed videos. Then we propose a Hierarchical Classification based Regression (HCR) model to estimate the vehicle count from features. HCR hierarchically divides the traffic scenes into different cases according to vehicle density, such that the broad-variation characteristics of traffic scenes can be better approximated. Finally, we evaluated the proposed method on the real highway surveillance videos. The results show that our method is very competitive to the pixeldomain methods, which can reach similar performance along with its lower complexity.

1.290 Layered Scene Decomposition via the Occlusion-CRF

[Link to Paper]

This paper addresses the challenging problem of perceiving the hidden or occluded geometry of the scene depicted in any given RGBD image. Unlike other image labeling problems such as image segmentation where each pixel needs to be assigned a single label, layered decomposition requires us to assign multiple labels to pixels. We propose a novel "Occlusion-CRF" model that allows for the integration of sophisticated priors to regularize the solution space and enables the automatic inference of the layer decomposition. We use a generalization of the Fusion Move algorithm to perform Maximum a Posterior (MAP) inference on the model that can handle the large label sets needed to represent multiple surface assignments to each pixel. We have evaluated the proposed model and the inference algorithm on many RGBD images of cluttered indoor scenes. Our experiments show that not only is our model able to explain occlusions but it also enables automatic inpainting of occluded/invisible surfaces.

1.291 Learning Relaxed Deep Supervision for Better Edge Detection

[Link to Paper]

(Abstract could not be extracted from PDF)

1.292 Manifold SLIC: A Fast Method to Compute Content-Sensitive Superpixels

[Link to Paper]

Superpixels are perceptually meaningful atomic regions that can effectively capture image features. Among various methods for computing uniform superpixels, simple linear iterative clustering (SLIC) is popular due to its simplicity and high performance. In this paper, we extend SLIC to compute content-sensitive superpixels, i.e., small superpixels in content-dense regions (e.g., with high intensity or color variation) and large superpixels in content-sparse regions. Rather than the conventional SLIC method that clusters pixels in 5, we map the image to a 2-dimensional manifold 5, whose area elements are a good measure of the content density in . We propose an efficient method to compute restricted centroidal Voronoi tessellation (RCVT) – a uniform tessellation – on , which induces the content-sensitive superpixels in . Unlike other algorithms that characterize content-sensitivity by geodesic distances, manifold SLIC tackles the problem by measuring areas of Voronoi cells on , which can be computed at a very low cost. As a result, it runs 10 times faster than the state-of-the-art content-sensitive superpixels algorithm. We evaluate manifold SLIC and seven representative methods on the BSDS500 benchmark and observe that our method outperforms the existing methods.

1.293 Multi-scale Patch Aggregation (MPA) for Simultaneous Detection and Segmentation

[Link to Paper]

Aiming at simultaneous detection and segmentation (SDS), we propose a proposal-free framework, which detect and segment object instances via mid-level patches. We design a unified trainable network on patches, which is followed by a fast and effective patch aggregation algorithm to infer object instances. Our method benefits from end-to-end training. Without object proposal generation, computation time can also be reduced. In experiments, our method yields results 62.1% and 61.8% in terms of mAP r on VOC2012 segmentation val and VOC2012 SDS val, which are stateof-the-art at the time of submission. We also report results on Microsoft COCO test-std/test-dev dataset in this paper.

1.294 Multilinear Hyperplane Hashing

[Link to Paper]

Hashing has become an increasingly popular technique for fast nearest neighbor search. Despite its successful progress in classic pointto-point search, there are few studies regarding point-to-hyperplane search, which has strong practical capabilities of scaling up applications like active learning with SVMs. Existing hyperplane hashing methods enable the fast search based on randomly generated hash codes, but still suffer from a low collision probability and thus usually require long codes for a satisfying performance. To overcome this problem, this paper proposes a multilinear hyperplane hashing that generates a hash bit using multiple linear projections. Our theoretical analysis shows that with an even number of random linear projections, the multilinear hash function possesses strong locality sensitivity to hyperplane queries. To leverage its sensitivity to the angle distance, we further introduce an angular quantization based learning framework for compact multilinear hashing, which considerably boosts the search performance with less hash bits. Experiments with applications to large-scale (up to one million) active learning on two datasets demonstrate the overall superiority of the proposed approach.

1.295 Structural Correlation Filter for Robust Visual Tracking

[Link to Paper]

In this paper, we propose a novel structural correlation filter (SCF) model for robust visual tracking. The proposed SCF model takes part-based tracking strategies into account in a

correlation filter tracker, and exploits circular shifts of all parts for their motion modeling to preserve target object structure. Compared with existing correlation filter trackers, our proposed tracker has several advantages: (1) Due to the part strategy, the learned structural correlation filters are less sensitive to partial occlusion, and have computational efficiency and robustness. (2) The learned filters are able to not only distinguish the parts from the background as the traditional correlation filters, but also exploit the intrinsic relationship among local parts via spatial constraints to preserve object structure. (3) The learned correlation filters not only make most parts share similar motion, but also tolerate outlier parts that have different motion. Both qualitative and quantitative evaluations on challenging benchmark image sequences demonstrate that the proposed SCF tracking algorithm performs favorably against several state-of-the-art methods.

1.296 Progressive Prioritized Multi-view Stereo

[Link to Paper]

This work proposes a progressive patch based multiview stereo algorithm able to deliver a dense point cloud at any time. This enables an immediate feedback on the reconstruction process in a user centric scenario. With increasing processing time, the model is improved in terms of resolution and accuracy. The algorithm explicitly handles input images with varying effective scale and creates visually pleasing point clouds. A priority scheme assures that the limited computational power is invested in scene parts, where the user is most interested in or the overall error can be reduced the most. The architecture of the proposed pipeline allows fast processing times in large scenes using a pure open-source CPU implementation. We show the performance of our algorithm on challenging standard datasets as well as on real-world scenes and compare it to the baseline.

1.297 Needle-Match: Reliable Patch Matching under High Uncertainty

[Link to Paper]

Reliable patch-matching forms the basis for many algorithms (super-resolution, denoising, inpainting, etc.) However, when the image quality deteriorates (by noise, blur or geometric distortions), the reliability of patch-matching deteriorates as well. Matched patches in the degraded image, do not necessarily imply similarity of the underlying patches in the (unknown) high-quality image. This restricts the applicability of patch-based methods. In this paper we present a patch representation called "Needle", which consists of small multi-scale versions of the patch and its immediate surrounding region. While the patch at the finest image scale is severely degraded, the degradation decreases dramatically in coarser needle scales, revealing reliable information for matching. We show that the Needle is

robust to many types of image degradations, leads to matches faithful to the underlying high-quality patches, and to improvement in existing patch-based methods.

1.298 Adaptive Object Detection Using Adjacency and Zoom Prediction

[Link to Paper]

State-of-the-art object detection systems rely on an accurate set of region proposals. Several recent methods use a neural network architecture to hypothesize promising object locations. While these approaches are computationally efficient, they rely on fixed image regions as anchors for predictions. In this paper we propose to use a search strategy that adaptively directs computational resources to sub-regions likely to contain objects. Compared to methods based on fixed anchor locations, our approach naturally adapts to cases where object instances are sparse and small. Our approach is comparable in terms of accuracy to the state-of-the-art Faster R-CNN approach while using two orders of magnitude fewer anchors on average. Code is publicly available.

1.299 Coherent Parametric Contours for Interactive Video Object Segmentation

[Link to Paper]

Interactive video segmentation systems aim at producing sub-pixel-level object boundaries for visual effect applications. Recent approaches mainly focus on using sparse user input (i.e. scribbles) for efficient segmentation; however, the quality of the final object boundaries is not satisfactory for the following reasons: (1) the boundary on each frame is often not accurate; (2) boundaries across adjacent frames wiggle around inconsistently, causing temporal flickering; and (3) there is a lack of direct user control for fine tuning. We propose Coherent Parametric Contours, a novel video segmentation propagation framework that addresses all the above issues. Our approach directly models the object boundary using a set of parametric curves, providing direct user controls for manual adjustment. A spatiotemporal optimization algorithm is employed to produce object boundaries that are spatially accurate and temporally stable. We show that existing evaluation datasets are limited and demonstrate a new set to cover the common cases in professional rotoscoping. A new metric for evaluating temporal consistency is proposed. Results show that our approach generates higher quality, more coherent segmentation results than previous methods.

1.300 Tensor Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Tensors via Convex Optimization

[Link to Paper]

This paper studies the Tensor Robust Principal Component (TRPCA) problem which extends the known Robust PCA [4] to the tensor case. Our model is based on a new tensor Singular Value Decomposition (t-SVD) [14] and its induced tensor tubal rank and tensor nuclear norm. Consider that we have a 3-way tensor X Rn1n2n3 such that X = L0 + S0, where L0 has low tubal rank and S0 is sparse. Is that possible to recover both components? In this work, we prove that under certain suitable assumptions, we can recover both the low-rank and the sparse components exactly by simply solving a convex program whose objective is a weighted combination of the tensor nuclear norm and the 1-norm, i.e., min L + E 1, s.t. X = L + E, L,E where = 1/ max(n1, n2)n3. Interestingly, TRPCA involves RPCA as a special case when n3 = 1 and thus it is a simple and elegant tensor extension of RPCA. Also numerical experiments verify our theory and the application for the image denoising demonstrates the effectiveness of our method.

1.301 A Hole Filling Approach based on Background Reconstruction for View Synthesis in 3D Video

[Link to Paper]

The depth image based rendering (DIBR) plays a key role in 3D video synthesis, by which other virtual views can be generated from a 2D video and its depth map. However, in the synthesis process, the background occluded by the foreground objects might be exposed in the new view, resulting in some holes in the synthetized video. In this paper, a hole filling approach based on background reconstruction is proposed, in which the temporal correlation information in both the 2D video and its corresponding depth map are exploited to construct a background video. To construct a clean background video, the foreground objects are detected and removed. Also motion compensation is applied to make the background reconstruction model suitable for moving camera scenario. Each frame is projected to the current plane where a modified Gaussian mixture model is performed. The constructed background video is used to eliminate the holes in the synthetized video. Our experimental results have indicated that the proposed approach has better quality of the synthetized 3D video compared with the other methods.

1.302 Efficient Deep Learning for Stereo Matching

[Link to Paper]

In the past year, convolutional neural networks have been shown to perform extremely well for stereo estimation. However, current architectures rely on siamese networks which exploit concatenation followed by further processing layers, requiring a minute of GPU computation per image pair. In contrast, in this paper we propose a matching network which is able to produce very accurate results in less than a second of GPU computation. Towards this goal, we exploit a product layer which simply computes the inner product between the two representations of a siamese architecture. We train our network by treating the problem as multi-class classification, where the classes are all possible disparities. This allows us to get calibrated scores, which result in much better matching performance when compared to existing approaches. Figure 1: To learn informative image patch representations we employ a siamese network which extracts marginal distributions over all possible disparities for each pixel.

1.303   Going Deeper into First-Person Activity Recognition

[Link to Paper]

We bring together ideas from recent work on feature design for egocentric action recognition under one framework by exploring the use of deep convolutional neural networks (CNN). Recent work has shown that features such as hand appearance, object attributes, local hand motion and camera ego-motion are important for characterizing first-person actions. To integrate these ideas under one framework, we propose a twin stream network architecture, where one stream analyzes appearance information and the other stream analyzes motion information. Our appearance stream encodes prior knowledge of the egocentric paradigm by explicitly training the network to segment hands and localize objects. By visualizing certain neuron activation of our network, we show that our proposed architecture naturally learns features that capture object attributes and hand-object configurations. Our extensive experiments on benchmark egocentric action datasets show that our deep architecture enables recognition rates that significantly outperform state-of-the-art techniques an average 6.6% increase in accuracy over all datasets. Furthermore, by learning to recognize objects, actions and activities jointly, the performance of individual recognition tasks also increase by 30% (actions) and 14% (objects). We also include the results of extensive ablative analysis to highlight the importance of network design decisions.

1.304   Group MAD Competition - A New Methodology to Compare Objective Image Quality Models

[Link to Paper]

Objective image quality assessment (IQA) models aim to automatically predict human visual perception of image quality and are of fundamental importance in the field of image processing and computer vision. With an increasing number of IQA models proposed, how to fairly compare their performance becomes a major challenge due to the enormous size of image space and the limited resource for subjective testing. The standard approach in literature is to compute several correlation metrics between subjective mean opinion scores (MOSs) and objective model predictions on several well-known subject-rated databases that contain distorted images generated from a few dozens of source images, which however provide an extremely limited representation of real-world images. Moreover, most IQA models developed on these databases often involve machine learning and/or manual parameter tuning steps to boost their performance, and thus their generalization capabilities are questionable. Here we propose a novel methodology to compare IQA models. We first build a database that contains 4,744 source natural images, together with 94,880 distorted images created from them. We then propose a new mechanism, namely group MAximum Differentiation (gMAD) competition, which automatically selects subsets of image pairs from the database that provide the strongest test to let the IQA models compete with each other. Subjective testing on the selected subsets reveals the relative performance of the IQA models and provides useful insights on potential ways to improve them. We report the gMAD competition results between 16 well-known IQA models, but the framework is extendable, allowing future IQA models to be added into the competition.

1.305 Learning Activity Progression in LSTMs for Activity Detection and Early Detection

[Link to Paper]

In this work we improve training of temporal deep models to better learn activity progression for activity detection and early detection tasks. Conventionally, when training a Recurrent Neural Network, specifically a Long Short Term Memory (LSTM) model, the training loss only considers classification error. However, we argue that the detection score of the correct activity category, or the detection score margin between the correct and incorrect categories, should be monotonically non-decreasing as the model observes more of the activity. We design novel ranking losses that directly penalize the model on violation of such monotonicities, which are used together with classification loss in training of LSTM models. Evaluation on ActivityNet shows significant benefits of the proposed ranking losses in both activity detection and early detection tasks. Figure 1: We study two problems: activity detection and early detection. For activity detection, we detect the category of the activity and its start and end point. For early detection, we need to detect the category and the start point of an activity after observing only a fraction of the activity. This example sequence contains the activity using ATM.

143

1.306 Bilateral Space Video Segmentation

[Link to Paper]

In this work, we propose a novel approach to video segmentation that operates in bilateral space. We design a new energy on the vertices of a regularly sampled spatiotemporal bilateral grid, which can be solved efficiently using a standard graph cut label assignment. Using a bilateral formulation, the energy that we minimize implicitly approximates long-range, spatio-temporal connections between pixels while still containing only a small number of variables and only local graph edges. We compare to a number of recent methods, and show that our approach achieves state-of-the-art results on multiple benchmarks in a fraction of the runtime. Furthermore, our method scales linearly with image size, allowing for interactive feedback on real-world high resolution video. Figure 1. Example results of our bilateral space video segmentation which automatically propagates a user provided mask on the first frame (left column) through the complete video (remaining columns). Thanks to the efficiency of our method, errors in the later frames can be easily fixed in an interactive manner.

1.307 Multiple Models Fitting as a Set Coverage Problem

[Link to Paper]

This paper deals with the extraction of multiple models from noisy or outlier-contaminated data. We cast the multi-model fitting problem in terms of set coverage, deriving a simple and effective method that generalizes Ransac to multiple models and deals with intersecting structures and outliers in a straightforward and principled manner, while avoiding the typical shortcomings of sequential approaches and those of clustering. The method compares favorably against the state-of-the-art on simulated and publicly available real data-sets.

1.308 Regularizing Long Short Term Memory with 3D Human-Skeleton Sequences for Action Recognition

[Link to Paper]

This paper argues that large-scale action recognition in video can be greatly improved by providing an additional modality in training data namely, 3D human-skeleton sequences aimed at complementing poorly represented or missing features of human actions in the training videos. For recognition, we use Long Short Term Memory (LSTM) grounded via

a deep Convolutional Neural Network (CNN) onto the video. Training of LSTM is regularized using the output of another encoder LSTM (eLSTM) grounded on 3D human-skeleton training data. For such regularized training of LSTM, we modify the standard backpropagation through time (BPTT) in order to address the wellknown issues with gradient descent in constraint optimization. Our evaluation on three benchmark datasets Sports1M, HMDB-51, and UCF101 shows accuracy improvements from 1.7% up to 14.8% relative to the state of the art.

1.309 Composition-preserving Deep Photo Aesthetics Assessment

[Link to Paper]

Photo aesthetics assessment is challenging. Deep convolutional neural network (ConvNet) methods have recently shown promising results for aesthetics assessment. The performance of these deep ConvNet methods, however, is often compromised by the constraint that the neural network only takes the fixed-size input. To accommodate this requirement, input images need to be transformed via cropping, scaling, or padding, which often damages image composition, reduces image resolution, or causes image distortion, thus compromising the aesthetics of the original images. In this paper, we present a composition-preserving deep ConvNet method that directly learns aesthetics features from the original input images without any image transformations. Specifically, our method adds an adaptive spatial pooling layer upon the regular convolution and pooling layers to directly handle input images with original sizes and aspect ratios. To allow for multi-scale feature extraction, we develop the Multi-Net Adaptive Spatial Pooling ConvNet architecture which consists of multiple sub-networks with different adaptive spatial pooling sizes and leverage a scenebased aggregation layer to effectively combine the predictions from multiple sub-networks. Our experiments on the large-scale aesthetics assessment benchmark (AVA [29]) demonstrate that our method can significantly improve the state-of-the-art results in photo aesthetics assessment.

1.310 Affinity CNN: Learning Pixel-Centric Pairwise Relations for Figure/Ground Embedding

[Link to Paper]

Spectral embedding provides a framework for solving perceptual organization problems, including image segmentation and figure/ground organization. From an affinity matrix describing pairwise relationships between pixels, it clusters pixels into regions, and, using a complex-valued extension, orders pixels according to layer. We train a convolutional neural network (CNN) to directly predict the pairwise relationships that define this affinity

145

matrix. Spectral embedding then resolves these predictions into a globallyconsistent segmentation and figure/ground organization of the scene. Experiments demonstrate significant benefit to this direct coupling compared to prior works which use explicit intermediate stages, such as edge detection, on the pathway from image to affinities. Our results suggest spectral embedding as a powerful alternative to the conditional random field (CRF)-based globalization schemes typically coupled to deep neural networks.

1.311 What Players do with the Ball: A Physically Constrained Interaction Modeling

[Link to Paper]

Tracking the ball is critical for video-based analysis of team sports. However, it is difficult, especially in lowresolution images, due to the small size of the ball, its speed that creates motion blur, and its often being occluded by players. In this paper, we propose a generic and principled approach to modeling the interaction between the ball and the players while also imposing appropriate physical constraints on the ball's trajectory. We show that our approach, formulated in terms of a Mixed Integer Program, is more robust and more accurate than several state-of-the-art approaches on real-life volleyball, basketball, and soccer sequences.

1.312 Generation and Comprehension of Unambiguous Object Descriptions

[Link to Paper]

We propose a method that can generate an unambiguous description (known as a referring expression) of a specific object or region in an image, and which can also comprehend or interpret such an expression to infer which object is being described. We show that our method outperforms previous methods that generate descriptions of objects without taking into account other potentially ambiguous objects in the scene. Our model is inspired by recent successes of deep learning methods for image captioning, but while image captioning is difficult to evaluate, our task allows for easy objective evaluation. We also present a new large-scale dataset for referring expressions, based on MSCOCO. We have released the dataset and a toolbox for visualization and evaluation, see https://github.com/ mjhucla/Google_Refexp_toolbox.

1.313 Geospatial Correspondences for Multimodal Registration

[Link to Paper]

146

The growing availability of very high resolution (¡1 m/pixel) satellite and aerial images has opened up unprecedented opportunities to monitor and analyze the evolution of land-cover and land-use across the world. To do so, images of the same geographical areas acquired at different times and, potentially, with different sensors must be efficiently parsed to update maps and detect land-cover changes. However, a naive transfer of ground truth labels from one location in the source image to the corresponding location in the target image is generally not feasible, as these images are often only loosely registered (with up to 50m of non-uniform errors). Furthermore, land-cover changes in an area over time must be taken into account for an accurate ground truth transfer. To tackle these challenges, we propose a mid-level sensor-invariant representation that encodes image regions in terms of the spatial distribution of their spectral neighbors. We incorporate this representation in a Markov Random Field to simultaneously account for nonlinear mis-registrations and enforce locality priors to find matches between multi-sensor images. We show how our approach can be used to assist in several multimodal land-cover update and change detection problems.

1.314 Pose-Aware Face Recognition in the Wild

[Link to Paper]

We propose a method to push the frontiers of unconstrained face recognition in the wild, focusing on the problem of extreme pose variations. As opposed to current techniques which either expect a single model to learn pose invariance through massive amounts of training data, or which normalize images to a single frontal pose, our method explicitly tackles pose variation by using multiple posespecific models and rendered face images. We leverage deep Convolutional Neural Networks (CNNs) to learn discriminative representations we call Pose-Aware Models (PAMs) using 500K images from the CASIA WebFace dataset. We present a comparative evaluation on the new IARPA Janus Benchmark A (IJB-A) and PIPA datasets. On these datasets PAMs achieve remarkably better performance than commercial products and surprisingly also outperform methods that are specifically fine-tuned on the target dataset.

1.315 Deep Exemplar 2D-3D Detection by Adapting from Real to Rendered Views

[Link to Paper]

This paper presents an end-to-end convolutional neural network (CNN) for 2D-3D exemplar detection. We demonstrate that the ability to adapt the features of natural images to better align with those of CAD rendered views is critical to the success of our technique.

We show that the adaptation can be learned by compositing rendered views of textured object models on natural images. Our approach can be naturally incorporated into a CNN detection pipeline and extends the accuracy and speed benefits from recent advances in deep learning to 2D-3D exemplar detection. We applied our method to two tasks: instance detection, where we evaluated on the IKEA dataset [36], and object category detection, where we out-perform Aubry et al. [3] for "chair" detection on a subset of the Pascal VOC dataset.

1.316   Reinforcement Learning for Visual Object Detection

[Link to Paper]

One of the most widely used strategies for visual object detection is based on exhaustive spatial hypothesis search. While methods like sliding windows have been successful and effective for many years, they are still brute-force, independent of the image content and the visual category being searched. In this paper we present principled sequential models that accumulate evidence collected at a small set of image locations in order to detect visual objects effectively. By formulating sequential search as reinforcement learning of the search policy (including the stopping condition), our fully trainable model can explicitly balance for each class, specifically, the conflicting goals of exploration sampling more image regions for better accuracy , and exploitation stopping the search efficiently when sufficiently confident about the target's location. The methodology is general and applicable to any detector response function. We report encouraging results in the PASCAL VOC 2012 object detection test set showing that the proposed methodology achieves almost two orders of magnitude speed-up over sliding window methods.

1.317   Hierarchical Gaussian Descriptor for Person Re-Identification

[Link to Paper]

Describing the color and textural information of a person image is one of the most crucial aspects of person reidentification. In this paper, we present a novel descriptor based on a hierarchical distribution of pixel features. A hierarchical covariance descriptor has been successfully applied for image classification. However, the mean information of pixel features, which is absent in covariance, tends to be major discriminative information of person images. To solve this problem, we describe a local region in an image via hierarchical Gaussian distribution in which both means and covariances are included in their parameters. More specifically, we model the region as a set of multiple Gaussian distributions in which each Gaussian represents the appearance of a local patch. The characteristics of the set of Gaussians are again described by another Gaussian distribution. In both steps, unlike the

hierarchical covariance descriptor, the proposed descriptor can model both the mean and the covariance information of pixel features properly. The results of experiments conducted on five databases indicate that the proposed descriptor exhibits remarkably high performance which outperforms the state-ofthe-art descriptors for person re-identification.

1.318 HD Maps: Fine-grained Road Segmentation by Parsing Ground and Aerial Images

[Link to Paper]

In this paper we present an approach to enhance existing maps with fine grained segmentation categories such as parking spots and sidewalk, as well as the number and location of road lanes. Towards this goal, we propose an efficient approach that is able to estimate these fine grained categories by doing joint inference over both, monocular aerial imagery, as well as ground images taken from a stereo camera pair mounted on top of a car. Important to this is reasoning about the alignment between the two types of imagery, as even when the measurements are taken with sophisticated GPS+IMU systems, this alignment is not sufficiently accurate. We demonstrate the effectiveness of our approach on a new dataset which enhances KITTI [8] with aerial images taken with a camera mounted on an airplane and flying around the city of Karlsruhe, Germany.

1.319 A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation

[Link to Paper]

Recent work has shown that optical flow estimation can be formulated as a supervised learning task and can be successfully solved with convolutional networks. Training of the so-called FlowNet was enabled by a large synthetically generated dataset. The present paper extends the concept of optical flow estimation via convolutional networks to disparity and scene flow estimation. To this end, we propose three synthetic stereo video datasets with sufficient realism, variation, and size to successfully train large networks. Our datasets are the first large-scale datasets to enable training and evaluation of scene flow methods. Besides the datasets, we present a convolutional network for real-time disparity estimation that provides state-of-the-art results. By combining a flow and disparity estimation network and training it jointly, we demonstrate the first scene flow estimation with a convolutional network. Figure 1. Our datasets provide over 35000 stereo frames with dense ground truth for optical flow, disparity and disparity change, as well as other data such as object segmentation.

1.320 Recurrent Convolutional Network for Video-based Person Re-Identification

[Link to Paper]

In this paper we propose a novel recurrent neural network architecture for video-based person re-identification. Given the video sequence of a person, features are extracted from each frame using a convolutional neural network that incorporates a recurrent final layer, which allows information to flow between time-steps. The features from all timesteps are then combined using temporal pooling to give an overall appearance feature for the complete sequence. The convolutional network, recurrent layer, and temporal pooling layer, are jointly trained to act as a feature extractor for video-based re-identification using a Siamese network architecture. Our approach makes use of colour and optical flow information in order to capture appearance and motion information which is useful for video re-identification. Experiments are conduced on the iLIDS-VID and PRID-2011 datasets to show that this approach outperforms existing methods of video-based re-identification.

1.321 From Keyframes to Key Objects: Video Summarization by Representative Object Proposal Selection

[Link to Paper]

We propose to summarize a video into a few key objects by selecting representative object proposals generated from video frames. This representative selection problem is formulated as a sparse dictionary selection problem, i.e., choosing a few representatives object proposals to reconstruct the whole proposal pool. Compared with existing sparse dictionary selection based representative selection methods, our new formulation can incorporate object proposal priors and locality prior in the feature space when selecting representatives. Consequently it can better locate key objects and suppress outlier proposals. We convert the optimization problem into a proximal gradient problem and solve it by the fast iterative shrinkage thresholding algorithm (FISTA). Experiments on synthetic data and real benchmark datasets show promising results of our key object summarization apporach in video content mining and search. Comparisons with existing representative selection approaches such as K-mediod, sparse dictionary selection and density based selection validate that our formulation can better capture the key video objects despite appearance variations, cluttered backgrounds and camera motions.

1.322 Uncalibrated Photometric Stereo by Stepwise Optimization Using Principal Components of Isotropic BRDFs

[Link to Paper]

The uncalibrated photometric stereo problem for nonLambertian surfaces is challenging because of the large number of unknowns and its ill-posed nature stemming from unknown reflectance functions. We propose a model that represents various isotropic reflectance functions by using the principal components of items in a dataset, and formulate the uncalibrated photometric stereo as a regression problem. We then solve it by stepwise optimization utilizing principal components in order of their eigenvalues. We have also developed two techniques that lead to convergence and highly accurate reconstruction, namely (1) a coarse-to-fine approach with normal grouping, and (2) a randomized multipoint search. Our experimental results with synthetic data showed that our method significantly outperformed previous methods. We also evaluated the algorithm in terms of real image data, where it gave good reconstruction results.

1.323 Approximate Log-Hilbert-Schmidt Distances between Covariance Operators for Image Classification

[Link to Paper]

This paper presents a novel framework for visual object recognition using infinite-dimensional covariance operators of input features, in the paradigm of kernel methods on infinite-dimensional Riemannian manifolds. Our formulation provides a rich representation of image features by exploiting their non-linear correlations, using the power of kernel methods and Riemannian geometry. Theoretically, we provide an approximate formulation for the Log-HilbertSchmidt distance between covariance operators that is efficient to compute and scalable to large datasets. Empirically, we apply our framework to the task of image classification on eight different, challenging datasets. In almost all cases, the results obtained outperform other state of the art methods, demonstrating the competitiveness and potential of our framework.

1.324 Cross-stitch Networks for Multi-task Learning

[Link to Paper]

Multi-task learning in Convolutional Networks has displayed remarkable success in the field of recognition. This success can be largely attributed to learning shared representations from multiple supervisory tasks. However, existing multi-task approaches rely on

enumerating multiple network architectures specific to the tasks at hand, that do not generalize. In this paper, we propose a principled approach to learn shared representations in ConvNets using multitask learning. Specifically, we propose a new sharing unit: "cross-stitch" unit. These units combine the activations from multiple networks and can be trained end-to-end. A network with cross-stitch units can learn an optimal combination of shared and task-specific representations. Our proposed method generalizes across multiple tasks and shows dramatically improved performance over baseline methods for categories with few training examples. Attributes Detection Surface Normals Sem. Segmentation Has saddle Four legs Object location Surface orientation Pixel labels Figure 1: Given an input image, one can leverage multiple related properties to improve performance by using a multi-task learning framework. In this paper, we propose cross-stitch units, a principled way to use such a multi-task framework for ConvNets.

1.325 Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels

[Link to Paper]

When human annotators are given a choice about what to label in an image, they apply their own subjective judgments on what to ignore and what to mention. We refer to these noisy "human-centric" annotations as exhibiting human reporting bias. Examples of such annotations include image tags and keywords found on photo sharing sites, or in datasets containing image captions. In this paper, we use these noisy annotations for learning visually correct image classifiers. Such annotations do not use consistent vocabulary, and miss a significant amount of the information present in an image; however, we demonstrate that the noise in these annotations exhibits structure and can be modeled. We propose an algorithm to decouple the human reporting bias from the correct visually grounded labels. Our results are highly interpretable for reporting "what's in the image" versus "what's worth saying." We demonstrate the algorithm's efficacy along a variety of metrics and datasets, including MS COCO and Yahoo Flickr 100M. We show significant improvements over traditional algorithms for both image classification and image captioning, doubling the performance of existing methods in some cases.

1.326 Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks

[Link to Paper]

Automatic detection and classification of dynamic hand gestures in real-world systems intended for human computer interaction is challenging as: 1) there is a large diversity

in how people perform gestures, making detection and classification difficult; 2) the system must work online in order to avoid noticeable lag between performing a gesture and its classification; in fact, a negative lag (classification before the gesture is finished) is desirable, as feedback to the user can then be truly instantaneous. In this paper, we address these challenges with a recurrent three-dimensional convolutional neural network that performs simultaneous detection and classification of dynamic hand gestures from multi-modal data. We employ connectionist temporal classification to train the network to predict class labels from inprogress gestures in unsegmented input streams. In order to validate our method, we introduce a new challenging multimodal dynamic hand gesture dataset captured with depth, color and stereo-IR sensors. On this challenging dataset, our gesture recognition system achieves an accuracy of 83.8%, outperforms competing state-of-the-art algorithms, and approaches human accuracy of 88.4%. Moreover, our method achieves state-of-the-art performance on SKIG and ChaLearn2014 benchmarks.

1.327 SublabelAccurate Relaxation of Nonconvex Energies

[Link to Paper]

We propose a novel spatially continuous framework for convex relaxations based on functional lifting. Our method can be interpreted as a sublabelaccurate solution to multilabel problems. We show that previously proposed functional lifting methods optimize an energy which is linear between two labels and hence require (often infinitely) many labels for a faithful approximation. In contrast, the proposed formulation is based on a piecewise convex approximation and therefore needs far fewer labels see Fig. 1. In comparison to recent MRF-based approaches, our method is formulated in a spatially continuous setting and shows less grid bias. Moreover, in a local sense, our formulation is the tightest possible convex relaxation. It is easy to implement and allows an efficient primal-dual optimization on GPUs. We show the effectiveness of our approach on several computer vision problems.

1.328 DeepFool: a simple and accurate method to fool deep neural networks

[Link to Paper]

State-of-the-art deep neural networks have achieved impressive results on many image classification tasks. However, these same architectures have been shown to be unstable to small, well sought, perturbations of the images. Despite the importance of this phenomenon, no effective methods have been proposed to accurately compute the robustness of state-of-the-art deep classifiers to such perturbations on large-scale datasets. In this

paper, we fill this gap and propose the DeepFool algorithm to efficiently compute perturbations that fool deep networks, and thus reliably quantify the robustness of these classifiers. Extensive experimental results show that our approach outperforms recent methods in the task of computing adversarial perturbations and making classifiers more robust.1

1.329 Active Learning for Delineation of Curvilinear Structures

[Link to Paper]

Many recent delineation techniques owe much of their increased effectiveness to path classification algorithms that make it possible to distinguish promising paths from others. The downside of this development is that they require annotated training data, which is tedious to produce. In this paper, we propose an Active Learning approach that considerably speeds up the annotation process. Unlike standard ones, it takes advantage of the specificities of the delineation problem. It operates on a graph and can reduce the training set size by up to 80% without compromising the reconstruction quality. We will show that our approach outperforms conventional ones on various biomedical and natural image datasets, thus showing that it is broadly applicable. (a) (b) Figure 1: Images of two different neural structures obtained using confocal microscopy. The enormous variety of curvilinear structures requires problem-specific training datasets even in case of the same modality.

1.330 Using Self-Contradiction to Learn Confidence Measures in Stereo Vision

[Link to Paper]

Learned confidence measures gain increasing importance for outlier removal and quality improvement in stereo vision. However, acquiring the necessary training data is typically a tedious and time consuming task that involves manual interaction, active sensing devices and/or synthetic scenes. To overcome this problem, we propose a new, flexible, and scalable way for generating training data that only requires a set of stereo images as input. The key idea of our approach is to use different view points for reasoning about contradictions and consistencies between multiple depth maps generated with the same stereo algorithm. This enables us to generate a huge amount of training data in a fully automated manner. Among other experiments, we demonstrate the potential of our approach by boosting the performance of three learned confidence measures on the KITTI2012 dataset by simply training them on a vast amount of automatically generated training data rather than a limited amount of laser ground truth data.

1.331 Information Bottleneck Learning Using Privileged Information for Visual Recognition

[Link to Paper]

We explore the visual recognition problem from a main data view when an auxiliary data view is available during training. This is important because it allows improving the training of visual classifiers when paired additional data is cheaply available, and it improves the recognition from multi-view data when there is a missing view at testing time. The problem is challenging because of the intrinsic asymmetry caused by the missing auxiliary view during testing. We account for such view during training by extending the information bottleneck method, and by combining it with risk minimization. In this way, we establish an information theoretic principle for leaning any type of visual classifier under this particular setting. We use this principle to design a large-margin classifier with an efficient optimization in the primal space. We extensively compare our method with the state-of-the-art on different visual recognition datasets, and with different types of auxiliary data, and show that the proposed framework has a very promising potential.

1.332 A Task-Oriented Approach for Cost-sensitive Recognition

[Link to Paper]

With the recent progress in visual recognition, we have already started to see a surge of vision related real-world applications. These applications, unlike general scene understanding, are task oriented and require specific information from visual data. Considering the current growth in new sensory devices, feature designs, feature learning methods, and algorithms, the search in the space of features and models becomes combinatorial. In this paper, we propose a novel cost-sensitive task-oriented recognition method that is based on a combination of linguistic semantics and visual cues. Our task-oriented framework is able to generalize to unseen tasks for which there is no training data and outperforms state-of-the-art cost-based recognition baselines on our new task-based dataset.

1.333 Newtonian Image Understanding: Unfolding the Dynamics of Objects in Static Images

[Link to Paper]

In this paper, we study the challenging problem of predicting the dynamics of objects in static images. Given a query object in an image, our goal is to provide a physical understanding of the object in terms of the forces acting upon it and its long term motion as

response to those forces. Direct and explicit estimation of the forces and the motion of objects from a single image is extremely challenging. We define intermediate physical abstractions called Newtonian scenarios and introduce Newtonian Neural Network (N 3) that learns to map a single image to a state in a Newtonian scenario. Our evaluations show that our method can reliably predict dynamics of a query object from a single image. In addition, our approach can provide physical reasoning that supports the predicted dynamics in terms of velocity and force vectors. To spur research in this direction we compiled Visual Newtonian Dynamics (VIND) dataset that includes more than 6000 videos aligned with Newtonian scenarios represented using game engines, and more than 4500 still images with their ground truth dynamics.

1.334 Kernel Approximation via Empirical Orthogonal Decomposition for Unsupervised Feature Learning

[Link to Paper]

Kernel approximation methods are important tools for various machine learning problems. There are two major methods used to approximate the kernel function: the Nystrom method and the random features method. However, the Nystrom method requires relatively high-complexity post-processing to calculate a solution and the random features method does not provide sufficient generalization performance. In this paper, we propose a method that has good generalization performance without high-complexity postprocessing via empirical orthogonal decomposition using the probability distribution estimated from training data. We provide a bound for the approximation error of the proposed method. Our experiments show that the proposed method is better than the random features method and comparable with the Nystrom method in terms of the approximation error and classification accuracy. We also show that hierarchical feature extraction using our kernel approximation demonstrates better performance than the existing methods.

1.335 Blockout: Dynamic Model Selection for Hierarchical Deep Networks

[Link to Paper]

Most deep architectures for image classificationeven those that are trained to classify a large number of diverse categorieslearn shared image representations with a single model. Intuitively, however, categories that are more similar should share more information than those that are very different. While hierarchical deep networks address this problem by learning separate features for subsets of related categories, current implementations require simplified models using fixed architectures specified via heuristic clustering methods. Instead, we propose Blockout, a method for regularization and model selection that

simultaneously learns both the model architecture and parameters. A generalization of Dropout, our approach gives a novel parametrization of hierarchical architectures that allows for structure learning via back-propagation. To demonstrate its utility, we evaluate Blockout on the CIFAR and ImageNet datasets, demonstrating improved classification accuracy, better regularization performance, faster training, and the clear emergence of hierarchical network structures.

1.336 Deep Decision Network for Multi-Class Image Classification

[Link to Paper]

In this paper, we present a novel Deep Decision Network (DDN) that provides an alternative approach towards building an efficient deep learning network. During the learning phase, starting from the root network node, DDN automatically builds a network that splits the data into disjoint clusters of classes which would be handled by the subsequent expert networks. This results in a tree-like structured network driven by the data. The proposed method provides an insight into the data by identifying the group of classes that are hard to classify and require more attention when compared to others. DDN also has the ability to make early decisions thus making it suitable for timesensitive applications. We validate DDN on two publicly available benchmark datasets: CIFAR-10 and CIFAR-100 and it yields state-of-the-art classification performance on both the datasets. The proposed algorithm has no limitations to be applied to any generic classification problems.

1.337 Temporally coherent 4D reconstruction of complex dynamic scenes

[Link to Paper]

This paper presents an approach for reconstruction of 4D temporally coherent models of complex dynamic scenes. No prior knowledge is required of scene structure or camera calibration allowing reconstruction from multiple moving cameras. Sparse-to-dense temporal correspondence is integrated with joint multi-view segmentation and reconstruction to obtain a complete 4D representation of static and dynamic objects. Temporal coherence is exploited to overcome visual ambiguities resulting in improved reconstruction of complex scenes. Robust joint segmentation and reconstruction of dynamic objects is achieved by introducing a geodesic star convexity constraint. Comparative evaluation is performed on a variety of unstructured indoor and outdoor dynamic scenes with hand-held cameras and multiple people. This demonstrates reconstruction of complete temporally coherent 4D scene models with improved nonrigid object segmentation and shape reconstruction.

1.338 Sample and Filter: Nonparametric Scene Parsing via Efficient Filtering

[Link to Paper]

Scene parsing has attracted a lot of attention in computer vision. While parametric models have proven effective for this task, they cannot easily incorporate new training data. By contrast, nonparametric approaches, which bypass any learning phase and directly transfer the labels from the training data to the query images, can readily exploit new labeled samples as they become available. Unfortunately, because of the computational cost of their label transfer procedures, state-of-the-art nonparametric methods typically filter out most training images to only keep a few relevant ones to label the query. As such, these methods throw away many images that still contain valuable information and generally obtain an unbalanced set of labeled samples. In this paper, we introduce a nonparametric approach to scene parsing that follows a sample-andfilter strategy. More specifically, we propose to sample labeled superpixels according to an image similarity score, which allows us to obtain a balanced set of samples. We then formulate label transfer as an efficient filtering procedure, which lets us exploit more labeled samples than existing techniques. Our experiments evidence the benefits of our approach over state-of-the-art nonparametric methods on two benchmark datasets.

1.339 G-CNN: an Iterative Grid Based Object Detector

[Link to Paper]

We introduce G-CNN, an object detection technique based on CNNs which works without proposal algorithms. G-CNN starts with a multi-scale grid of fixed bounding boxes. We train a regressor to move and scale elements of the grid towards objects iteratively. G-CNN models the problem of object detection as finding a path from a fixed grid to boxes tightly surrounding the objects. G-CNN with around 180 boxes in a multi-scale grid performs comparably to Fast R-CNN which uses around 2K bounding boxes generated with a proposal technique. This strategy makes detection faster by removing the object proposal stage as well as reducing the number of boxes to be processed.

1.340 A Holistic Approach to Cross-Channel Image Noise Modeling and its Application to Image Denoising

[Link to Paper]

Modelling and analyzing noise in images is a fundamental task in many computer vision systems. Traditionally, noise has been modelled per color channel assuming that the color channels are independent. Although the color channels can be considered as mutually independent in camera RAW images, signals from different color channels get mixed during the imaging process inside the camera due to gamut mapping, tone-mapping, and compression. We show the influence of the in-camera imaging pipeline on noise and propose a new noise model in the 3D RGB space to accounts for the color channel mix-ups. A data-driven approach for determining the parameters of the new noise model is introduced as well as its application to image denoising. The experiments show that our noise model represents the noise in regular JPEG images more accurately compared to the previous models and is advantageous in image denoising.

1.341 Learning Multi-Domain Convolutional Neural Networks for Visual Tracking

[Link to Paper]

We propose a novel visual tracking algorithm based on the representations from a discriminatively trained Convolutional Neural Network (CNN). Our algorithm pretrains a CNN using a large set of videos with tracking groundtruths to obtain a generic target representation. Our network is composed of shared layers and multiple branches of domain-specific layers, where domains correspond to individual training sequences and each branch is responsible for binary classification to identify target in each domain. We train each domain in the network iteratively to obtain generic target representations in the shared layers. When tracking a target in a new sequence, we construct a new network by combining the shared layers in the pretrained CNN with a new binary classification layer, which is updated online. Online tracking is performed by evaluating the candidate windows randomly sampled around the previous target state. The proposed algorithm illustrates outstanding performance in existing tracking benchmarks.

1.342 Globally Optimal Rigid Intensity Based Registration: A Fast Fourier Domain Approach

[Link to Paper]

(Abstract could not be extracted from PDF)

1.343 Single image object modeling based on BRDF and r-surfaces learning

[Link to Paper]

A methodology for 3D surface modeling from a single image is proposed. The principal novelty is concave and specular surface modeling without any externally imposed prior. The main idea of the method is to use BRDFs and generated rendered surfaces, to transfer the normal field, computed for the generated samples, to the unknown surface. The transferred information is adequate to blow and sculpt the segmented image mask in to a bas-relief of the object. The object surface is further refined basing on a photo-consistency formulation that relates for error minimization the original image and the modeled object.

1.344 A Field Model for Repairing 3D Shapes

[Link to Paper]

This paper proposes a field model for repairing 3D shapes constructed from multi-view RGB data. Specifically, we represent a 3D shape in a Markov random field (MRF) in which the geometric information is encoded by random binary variables and the appearance information is retrieved from a set of RGB images captured at multiple viewpoints. The local priors in the MRF model capture the local structures of object shapes and are learnt from 3D shape templates using a convolutional deep belief network. Repairing a 3D shape is formulated as the maximum a posteriori (MAP) estimation in the corresponding MRF. Variational mean field approximation technique is adopted for the MAP estimation. The proposed method was evaluated on both artificial data and real data obtained from reconstruction of practical scenes. Experimental results have shown the robustness and efficiency of the proposed method in repairing noisy and incomplete 3D shapes.

1.345 RAW Image Reconstruction using a Self-Contained sRGB-JPEG Image with only 64 KB Overhead

[Link to Paper]

Most camera images are saved as 8-bit standard RGB (sRGB) compressed JPEGs. Even when JPEG compression is set to its highest quality, the encoded sRGB image has been significantly processed in terms of color and tone manipulation. This makes sRGB-JPEG images undesirable for many computer vision tasks that assume a direct relationship between pixel values and incoming light. For such applications, the RAW image format is preferred,

as RAW represents a minimally processed, sensor-specific RGB image with higher dynamic range that is linear with respect to scene radiance. The drawback with RAW images, however, is that they require large amounts of storage and are not well-supported by many imaging applications. To address this issue, we present a method to encode the necessary metadata within an sRGB image to reconstruct a high-quality RAW image. Our approach requires no calibration of the camera and can reconstruct the original RAW to within 0.3% error with only a 64 KB overhead for the additional data. More importantly, our output is a fully selfcontained 100% complainant sRGB-JPEG file that can be used as-is, not affecting any existing image workflow - the RAW image can be extracted when needed, or ignored otherwise. We detail our approach and show its effectiveness against competing strategies.

1.346 Progressively Parsing Interactional Objects for Fine Grained Action Detection

[Link to Paper]

Fine grained video action analysis often requires reliable detection and tracking of various interacting objects and human body parts, denoted as Interactional Object Parsing. However, most of the previous methods based on either independent or joint object detection might suffer from high model complexity and challenging image content, e.g., illumination/pose/appearance/scale variation, motion, and occlusion etc. In this work, we propose an end-to-end system based on recurrent neural network to perform frame by frame interactional object parsing, which can alleviate the difficulty through an incremental/progressive manner. Our key innovation is that: instead of jointly outputting all object detections at once, for each frame we use a set of long-short term memory (LSTM) nodes to incrementally refine the detections. After passing through each LSTM node, more object detections are consolidated and thus more contextual information could be utilized to localize more difficult objects. The object parsing results are further utilized to form object specific action representation for fine grained action detection. Extensive experiments on two benchmark fine grained activity datasets demonstrate that our proposed algorithm achieves better interacting object detection performance, which in turn boosts the action recognition performance over the state-of-the-art.

1.347 Object Tracking via Dual Linear Structured SVM and Explicit Feature Map

[Link to Paper]

Structured support vector machine (SSVM) based methods have demonstrated encouraging performance in recent object tracking benchmarks. However, the complex and expensive optimization limits their deployment in real-world applications. In this paper, we present a simple yet efficient dual linear SSVM (DLSSVM) algorithm to enable fast learning and execution during tracking. By analyzing the dual variables, we propose a primal classifier update formula where the learning step size is computed in closed form. This online learning method significantly improves the robustness of the proposed linear SSVM with lower computational cost. Second, we approximate the intersection kernel for feature representations with an explicit feature map to further improve tracking performance. Finally, we extend the proposed DLSSVM tracker with multi-scale estimation to address the "drift" problem. Experimental results on large benchmark datasets with 50 and 100 video sequences show that the proposed DLSSVM tracking algorithm achieves state-of-the-art performance.

1.348 Ordinal Regression with Multiple Output CNN for Age Estimation

[Link to Paper]

To address the non-stationary property of aging patterns, age estimation can be cast as an ordinal regression problem. However, the processes of extracting features and learning a regression model are often separated and optimized independently in previous work. In this paper, we propose an End-to-End learning approach to address ordinal regression problems using deep Convolutional Neural Network, which could simultaneously conduct feature learning and regression modeling. In particular, an ordinal regression problem is transformed into a series of binary classification sub-problems. And we propose a multiple output CNN learning algorithm to collectively solve these classification sub-problems, so that the correlation between these tasks could be explored. In addition, we publish an Asian Face Age Dataset (AFAD) containing more than 160K facial images with precise age ground-truths, which is the largest public age dataset to date. To the best of our knowledge, this is the first work to address ordinal regression problems by using CNN, and achieves the state-of-the-art performance on both the MORPH and AFAD datasets.

1.349 Image Question Answering using Convolutional Neural Network with Dynamic Parameter Prediction

[Link to Paper]

We tackle image question answering (ImageQA) problem by learning a convolutional neural network (CNN) with a dynamic parameter layer whose weights are determined adaptively based on questions. For the adaptive parameter prediction, we employ a separate

parameter prediction network, which consists of gated recurrent unit (GRU) taking a question as its input and a fully-connected layer generating a set of candidate weights as its output. However, it is challenging to construct a parameter prediction network for a large number of parameters in the fully-connected dynamic parameter layer of the CNN. We reduce the complexity of this problem by incorporating a hashing technique, where the candidate weights given by the parameter prediction network are selected using a predefined hash function to determine individual weights in the dynamic parameter layer. The proposed network–joint network with the CNN for ImageQA and the parameter prediction network– is trained end-to-end through back-propagation, where its weights are initialized using a pre-trained CNN and GRU. The proposed algorithm illustrates the state-of-the-art performance on all available public ImageQA benchmarks.

1.350   Efficiently Creating 3D Training Data for Fine Hand Pose Estimation

[Link to Paper]

While many recent hand pose estimation methods critically rely on a training set of labelled frames, the creation of such a dataset is a challenging task that has been overlooked so far. As a result, existing datasets are limited to a few sequences and individuals, with limited accuracy, and this prevents these methods from delivering their full potential. We propose a semi-automated method for efficiently and accurately labeling each frame of a hand depth video with the corresponding 3D locations of the joints: The user is asked to provide only an estimate of the 2D reprojections of the visible joints in some reference frames, which are automatically selected to minimize the labeling work by efficiently optimizing a sub-modular loss function. We then exploit spatial, temporal, and appearance constraints to retrieve the full 3D poses of the hand over the complete sequence. We show that this data can be used to train a recent state-of-the-art hand pose estimation method, leading to increased accuracy.

1.351   Fast Detection of Curved Edges at Low SNR

[Link to Paper]

Detecting edges is a fundamental problem in computer vision with many applications, some involving very noisy images. While most edge detection methods are fast, they perform well only on relatively clean images. Unfortunately, sophisticated methods that are robust to high levels of noise are quite slow. In this paper we develop a novel multiscale method to detect curved edges in noisy images. Even though our algorithm searches for edges over an exponentially large set of candidate curves, its runtime is nearly linear in the total number of image pixels. As we demonstrate experimentally, our algorithm is orders of

magnitude faster than previous methods designed to deal with high noise levels. At the same time it obtains comparable and often superior results to existing methods on a variety of challenging noisy images. Figure 1. A noisy binary pattern with three curved fibers at high noise level (left, SNR=1) or low noise level (middle, SNR=2). The clean pattern is shown on the right. While the long fiber is noticeable already at the low SNR. the two shorter fibers can be spotted only at the higher SNR.

1.352   Do It Yourself Hyperspectral Imaging with Everyday Digital Cameras

[Link to Paper]

Capturing hyperspectral images requires expensive and specialized hardware that is not readily accessible to most users. Digital cameras, on the other hand, are significantly cheaper in comparison and can be easily purchased and used. In this paper, we present a framework for reconstructing hyperspectral images by using multiple consumer-level digital cameras. Our approach works by exploiting the different spectral sensitivities of different camera sensors. In particular, due to the differences in spectral sensitivities of the cameras, different cameras yield different RGB measurements for the same spectral signal. We introduce an algorithm that is able to combine and convert these different RGB measurements into a single hyperspectral image for both indoor and outdoor scenes. This camera-based approach allows hyperspectral imaging at a fraction of the cost of most existing hyperspectral hardware. We validate the accuracy of our reconstruction against ground truth hyperspectral images (using both synthetic and real cases) and show its usage on relighting applications.

1.353   Recognizing Activities of Daily Living with a Wrist-mounted Camera

[Link to Paper]

We present a novel dataset and a novel algorithm for recognizing activities of daily living (ADL) from a first-person wearable camera. Handled objects are crucially important for egocentric ADL recognition. For specific examination of objects related to users' actions separately from other objects in an environment, many previous works have addressed the detection of handled objects in images captured from head-mounted and chest-mounted cameras. Nevertheless, detecting handled objects is not always easy because they tend to appear small in images. They can be occluded by a user's body. As described herein, we mount a camera on a user's wrist. A wrist-mounted camera can capture handled objects at a large scale, and thus it enables us to skip the object detection process. To compare a wrist-mounted camera and a head-mounted camera, we also developed a novel and publicly available dataset 1 that includes videos and annotations of daily activities captured

simultaneously by both cameras. Additionally, we propose a discriminative video representation that retains spatial and temporal information after encoding the frame descriptors extracted by convolutional neural networks (CNN).

1.354   Improved Hamming Distance Search using Variable Length Hashing

[Link to Paper]

This paper addresses the problem of ultra-large-scale search in Hamming spaces. There has been considerable research on generating compact binary codes in vision, for example for visual search tasks. However the issue of efficient searching through huge sets of binary codes remains largely unsolved. To this end, we propose a novel, unsupervised approach to thresholded search in Hamming space, supporting long codes (e.g. 512-bits) with a wide-range of Hamming distance radii. Our method is capable of working efficiently with billions of codes delivering between one to three orders of magnitude acceleration, as compared to prior art. This is achieved by relaxing the equal-size constraint in the Multi-Index Hashing approach, leading to multiple hash-tables with variable length hash-keys. Based on the theoretical analysis of the retrieval probabilities of multiple hash-tables we propose a novel search algorithm for obtaining a suitable set of hash-key lengths. The resulting retrieval mechanism is shown empirically to improve the efficiency over the state-of-the-art, across a range of datasets, bit-depths and retrieval thresholds.

1.355   Real-Time Depth Refinement for Specular Objects

[Link to Paper]

The introduction of consumer RGB-D scanners set off a major boost in 3D computer vision research. Yet, the precision of existing depth scanners is not accurate enough to recover fine details of a scanned object. While modern shading based depth refinement methods have been proven to work well with Lambertian objects, they break down in the presence of specularities. We present a novel shape from shading framework that addresses this issue and enhances both diffuse and specular objects' depth profiles. We take advantage of the built-in monochromatic IR projector and IR images of the RGB-D scanners and present a lighting model that accounts for the specular regions in the input image. Using this model, we reconstruct the depth map in real-time. Both quantitative tests and visual evaluations prove that the proposed method produces state of the art depth reconstruction results.

1.356 Trust No One: Low Rank Matrix Factorization Using Hierarchical RANSAC

[Link to Paper]

In this paper we present a system for performing low rank matrix factorization. Low-rank matrix factorization is an essential problem in many areas, including computer vision with applications in affine structure-from-motion, photometric stereo, and non-rigid structure from motion. We specifically target structured data patterns, with outliers and large amounts of missing data. Using recently developed characterizations of minimal solutions to matrix factorization problems with missing data, we show how these can be used as building blocks in a hierarchical system that performs bootstrapping on all levels. This gives a robust and fast system, with state-of-the-art performance.

1.357 Factors in Finetuning Deep Model for Object Detection with Long-tail Distribution

[Link to Paper]

Finetuning from a pretrained deep model is found to yield state-of-the-art performance for many vision tasks. This paper investigates many factors that influence the performance in finetuning for object detection. There is a longtailed distribution of sample numbers for classes in object detection. Our analysis and empirical results show that classes with more samples have higher impact on the feature learning. And it is better to make the sample number more uniform across classes. Generic object detection can be considered as multiple equally important tasks. Detection of each class is a task. These classes/tasks have their individuality in discriminative visual appearance representation. Taking this individuality into account, we cluster objects into visually similar class groups and learn deep representations for these groups separately. A hierarchical feature learning scheme is proposed. In this scheme, the knowledge from the group with large number of classes is transferred for learning features in its subgroups. Finetuned on the GoogLeNet model, experimental results show 4.7% absolute mAP improvement of our approach on the ImageNet object detection dataset without increasing much computational cost at the testing stage. Code is available on www.ee.cuhk.edu.hk/ wlouyang/ projects/ImageNetFactors/CVPR16.html

1.358 ForgetMeNot: Memory-Aware Forensic Facial Sketch Matching

[Link to Paper]

We investigate whether it is possible to improve the performance of automated facial forensic sketch matching by learning from examples of facial forgetting over time. Forensic facial

sketch recognition is a key capability for law enforcement, but remains an unsolved problem. It is extremely challenging because there are three distinct contributors to the domain gap between forensic sketches and photos: The well-studied sketch-photo modality gap, and the less studied gaps due to (i) the forgetting process of the eye-witness and (ii) their inability to elucidate their memory. In this paper, we address the memory problem head on by introducing a database of 400 forensic sketches created at different time-delays. Based on this database we build a model to reverse the forgetting process. Surprisingly, we show that it is possible to systematically "un-forget" facial details. Moreover, it is possible to apply this model to dramatically improve forensic sketch recognition in practice: we achieve the state of the art results when matching 195 benchmark forensic sketches against corresponding photos and a 10,030 mugshot database.

1.359   Visually Indicated Sounds

[Link to Paper]

Objects make distinctive sounds when they are hit or scratched. These sounds reveal aspects of an object's material properties, as well as the actions that produced them. In this paper, we propose the task of predicting what sound an object makes when struck as a way of studying physical interactions within a visual scene. We present an algorithm that synthesizes sound from silent videos of people hitting and scratching objects with a drumstick. This algorithm uses a recurrent neural network to predict sound features from videos and then produces a waveform from these features with an example-based synthesis procedure. We show that the sounds predicted by our model are realistic enough to fool participants in a "real or fake" psychophysical experiment, and that they convey significant information about material properties and physical interactions.

1.360   Discriminative Invariant Kernel Features: A Bells-and-Whistles-Free Approach to Unsupervised Face Recognition and Pose Estimation

[Link to Paper]

We propose an explicitly discriminative and 'simple' approach to generate invariance to nuisance transformations modeled as unitary. In practice, the approach works well to handle non-unitary transformations as well. Our theoretical results extend the reach of a recent theory of invariance to discriminative and kernelized features based on unitary kernels. As a special case, a single common framework can be used to generate subject-specific pose-invariant features for face recognition and vice-versa for pose estimation. We show that our main proposed method (DIKF) can perform well under very challenging large-scale semisynthetic face matching and pose estimation protocols with unaligned faces using no

167

landmarking whatsoever. We additionally benchmark on CMU MPIE and outperform previous work in almost all cases on off-angle face matching while we are on par with the previous state-of-the-art on the LFW unsupervised and image-restricted protocols, without any low-level image descriptors other than raw-pixels.

1.361 Blind Image Deblurring Using Dark Channel Prior

[Link to Paper]

We present a simple and effective blind image deblurring method based on the dark channel prior. Our work is inspired by the interesting observation that the dark channel of blurred images is less sparse. While most image patches in the clean image contain some dark pixels, these pixels are not dark when averaged with neighboring highintensity pixels during the blur process. This change in the sparsity of the dark channel is an inherent property of the blur process, which we both prove mathematically and validate using training data. Therefore, enforcing the sparsity of the dark channel helps blind deblurring on various scenarios, including natural, face, text, and low-illumination images. However, sparsity of the dark channel introduces a non-convex non-linear optimization problem. We introduce a linear approximation of the min operator to compute the dark channel. Our look-up-table-based method converges fast in practice and can be directly extended to non-uniform deblurring. Extensive experiments show that our method achieves state-of-the-art results on deblurring natural images and compares favorably methods that are well-engineered for specific scenarios.

1.362 Hierarchical Recurrent Neural Encoder for Video Representation with Application to Captioning

[Link to Paper]

Recently, deep learning approach, especially deep Convolutional Neural Networks (ConvNets), have achieved overwhelming accuracy with fast processing speed for image classification. Incorporating temporal structure with deep ConvNets for video representation becomes a fundamental problem for video content analysis. In this paper, we propose a new approach, namely Hierarchical Recurrent Neural Encoder (HRNE), to exploit temporal information of videos. Compared to recent video representation inference approaches, this paper makes the following three contributions. First, our HRNE is able to efficiently exploit video temporal structure in a longer range by reducing the length of input information flow, and compositing multiple consecutive inputs at a higher level. Second, computation operations are significantly lessened while attaining more nonlinearity. Third, HRNE is able to uncover temporal transitions between frame chunks with different granularities, i.e. it

can model the temporal transitions between frames as well as the transitions between segments. We apply the new method to video captioning where temporal information plays a crucial role. Experiments demonstrate that our method outperforms the state-of-the-art on video captioning benchmarks.

1.363 Jointly Modeling Embedding and Translation to Bridge Video and Language

[Link to Paper]

Automatically describing video content with natural language is a fundamental challenge of computer vision. Recurrent Neural Networks (RNNs), which models sequence dynamics, has attracted increasing attention on visual interpretation. However, most existing approaches generate a word locally with the given previous words and the visual content, while the relationship between sentence semantics and visual content is not holistically exploited. As a result, the generated sentences may be contextually correct but the semantics (e.g., subjects, verbs or objects) are not true. This paper presents a novel unified framework, named Long Short-Term Memory with visual-semantic Embedding (LSTM-E), which can simultaneously explore the learning of LSTM and visual-semantic embedding. The former aims to locally maximize the probability of generating the next word given previous words and visual content, while the latter is to create a visual-semantic embedding space for enforcing the relationship between the semantics of the entire sentence and visual content. The experiments on YouTube2Text dataset show that our proposed LSTM-E achieves to-date the best published performance in generating natural sentences: 45.3% and 31.0% in terms of BLEU@4 and METEOR, respectively. Superior performances are also reported on two movie description datasets (M-VAD and MPII-MD). In addition, we demonstrate that LSTM-E outperforms several state-of-the-art techniques in predicting Subject-Verb-Object (SVO) triplets.

1.364 Robust Kernel Estimation with Outliers Handling for Image Deblurring

[Link to Paper]

Estimating blur kernels from real world images is a challenging problem as the linear image formation assumption does not hold when significant outliers, such as saturated pixels and non-Gaussian noise, are present. While some existing non-blind deblurring algorithms can deal with outliers to a certain extent, few blind deblurring methods are developed to well estimate the blur kernels from the blurred images with outliers. In this paper, we present an algorithm to address this problem by exploiting reliable edges and removing outliers in

the intermediate latent images, thereby estimating blur kernels robustly. We analyze the effects of outliers on kernel estimation and show that most state-of-the-art blind deblurring methods may recover delta kernels when blurred images contain significant outliers. We propose a robust energy function which describes the properties of outliers for the final latent image restoration. Furthermore, we show that the proposed algorithm can be applied to improve existing methods to deblur images with outliers. Extensive experiments on different kinds of challenging blurry images with significant amount of outliers demonstrate the proposed algorithm performs favorably against the state-of-the-art methods.

1.365 Shallow and Deep Convolutional Networks for Saliency Prediction

[Link to Paper]

The prediction of salient areas in images has been traditionally addressed with hand-crafted features based on neuroscience principles. This paper, however, addresses the problem with a completely data-driven approach by training a convolutional neural network (convnet). The learning process is formulated as a minimization of a loss function that measures the Euclidean distance of the predicted saliency map with the provided ground truth. The recent publication of large datasets of saliency prediction has provided enough data to train end-to-end architectures that are both fast and accurate. Two designs are proposed: a shallow convnet trained from scratch, and a another deeper solution whose first three layers are adapted from another network trained for classification. To the authors' knowledge, these are the first end-to-end CNNs trained and tested for the purpose of saliency prediction.

1.366 Soft-Segmentation Guided Object Motion Deblurring

[Link to Paper]

Object motion blur is a challenging problem as the foreground and the background in the scenes undergo different types of image degradation due to movements in various directions and speed. Most object motion deblurring methods address this problem by segmenting blurred images into regions where different kernels are estimated and applied for restoration. Segmentation on blurred images is difficult due to ambiguous pixels between regions, but it plays an important role for object motion deblurring. To address these problems, we propose a novel model for object motion deblurring. The proposed model is developed based on a maximum a posterior formulation in which soft-segmentation is incorporated for object layer estimation. We propose an efficient algorithm to jointly estimate object segmentation and camera motion where each layer can be deblurred well under the

guidance of the soft-segmentation. Experimental results demonstrate that the proposed algorithm performs favorably against the state-of-the-art object motion deblurring methods on challenging scenarios.

1.367 We don't need no bounding-boxes: Training object class detectors using only human verification

[Link to Paper]

Training object class detectors typically requires a large set of images in which objects are annotated by boundingboxes. However, manually drawing bounding-boxes is very time consuming. We propose a new scheme for training object detectors which only requires annotators to verify bounding-boxes produced automatically by the learning algorithm. Our scheme iterates between re-training the detector, re-localizing objects in the training images, and human verification. We use the verification signal both to improve re-training and to reduce the search space for re-localisation, which makes these steps different to what is normally done in a weakly supervised setting. Extensive experiments on PASCAL VOC 2007 show that (1) using human verification to update detectors and reduce the search space leads to the rapid production of high-quality bounding-box annotations; (2) our scheme delivers detectors performing almost as good as those trained in a fully supervised setting, without ever drawing any bounding-box; (3) as the verification task is very quick, our scheme substantially reduces total annotation time by a factor 6-9.

1.368 Isometric Non-Rigid Shape-from-Motion in Linear Time

[Link to Paper]

We study Isometric Non-Rigid Shape-from-Motion (IsoNRSfM): given multiple intrinsically calibrated monocular images, we want to reconstruct the time-varying 3D shape of an object undergoing isometric deformations. We show that Iso-NRSfM is solvable from the warps (the inter-image geometric transformations). We propose a new theoretical framework based on Riemmanian manifolds to represent the unknown 3D surfaces, as embeddings of the camera's retinal planes. This allows us to use the manifolds' metric tensor and Christoffel Symbol fields, which we prove are related across images by simple rules depending only on the warps. This forms a set of important theoretical results. Using the infinitesimal planarity formulation, it then allows us to derive a system of two quartics in two variables for each image pair. The sum-of-squares of these polynomials is independent of the number of images and can be solved globally, forming a well-posed problem for N 3 images, whose solution directly leads to the surface's normal field. The proposed method

outperforms existing work in terms of accuracy and computation cost on synthetic and real datasets.

1.369 Efficient and Robust Color Consistency for Community Photo Collections

[Link to Paper]

We present an efficient technique to optimize color consistency of a collection of images depicting a common scene. Our method first recovers sparse pixel correspondences in the input images and stacks them into a matrix with many missing entries. We show that this matrix satisfies a rank two constraint under a simple color correction model. These parameters can be viewed as pseudo white balance and gamma correction parameters for each input image. We present a robust low-rank matrix factorization method to estimate the unknown parameters of this model. Using them, we improve color consistency of the input images or perform color transfer with any input image as the source. Our approach is insensitive to outliers in the pixel correspondences thereby precluding the need for complex pre-processing steps. We demonstrate high quality color consistency results on large photo collections of popular tourist landmarks and personal photo collections containing images of people.

1.370 Egocentric Future Localization

[Link to Paper]

We presents a method for future localization: to predict plausible future trajectories of ego-motion in egocentric stereo images. Our paths avoid obstacles, move between objects, even turn around a corner into space behind objects. As a byproduct of the predicted trajectories, we discover the empty space occluded by foreground objects. One key innovation is the creation of an EgoRetinal map, akin to an illustrated tourist map, that 'rearranges' pixels taking into accounts depth information, the ground plane, and body motion direction, so that it allows motion planning and perception of objects on one image space. We learn to plan trajectories directly on this EgoRetinal map using first person experience of walking around in a variety of scenes. In a testing phase, given an novel scene, we find multiple hypotheses of future trajectories from the learned experience. We refine them by minimizing a cost function that describes compatibility between the obstacles in the EgoRetinal map and trajectories. We quantitatively evaluate our method to show predictive validity and apply to various real world daily activities including walking, shopping, and social interactions.

1.371 Force from Motion: Decoding Physical Sensation in a First Person Video

[Link to Paper]

A first-person video can generate powerful physical sensations of action in an observer. In this paper, we focus on a problem of Force from Motion–decoding the sensation of 1) passive forces such as the gravity, 2) the physical scale of the motion (speed) and space, and 3) active forces exerted by the observer such as pedaling a bike or banking on a ski turn. The sensation of gravity can be observed in a natural image. We learn this image cue for predicting a gravity direction in a 2D image and integrate the prediction across images to estimate the 3D gravity direction using structure from motion. The sense of physical scale is revealed to us when the body is in a dynamically balanced state. We compute the unknown physical scale of 3D reconstructed camera motion by leveraging the torque equilibrium at a banked turn that relates the centripetal force, gravity, and the body leaning angle. The active force and torque governs 3D egomotion through the physics of rigid body dynamics. Using an inverse dynamics optimization, we directly minimize 2D reprojection error (in video) with respect to 3D world structure, active forces, and additional passive forces such as air drag and friction force. We use structure from motion with the physical scale and gravity direction as an initialization of our bundle adjustment for force estimation. Our method shows quantitatively equivalent reconstruction comparing to IMU measurements in terms of gravity and scale recovery and outperforms method based on 2D optical flow for an active action recognition task. We apply our method to first person videos of mountain biking, urban bike racing, skiing, speedflying with parachute, and wingsuit flying where inertial measurements are not accessible.

1.372 Context Encoders: Feature Learning by Inpainting

[Link to Paper]

We present an unsupervised visual feature learning algorithm driven by context-based pixel prediction. By analogy with auto-encoders, we propose Context Encoders a convolutional neural network trained to generate the contents of an arbitrary image region conditioned on its surroundings. In order to succeed at this task, context encoders need to both understand the content of the entire image, as well as produce a plausible hypothesis for the missing part(s). When training context encoders, we have experimented with both a standard pixel-wise reconstruction loss, as well as a reconstruction plus an adversarial loss. The latter produces much sharper results because it can better handle multiple modes in the output. We found that a context encoder learns a representation that captures not just appearance but also the semantics of visual structures. We quantitatively demonstrate the effectiveness of our learned features for CNN pre-training on classification, detection, and

segmentation tasks. Furthermore, context encoders can be used for semantic inpainting tasks, either stand-alone or as initialization for non-parametric methods.

1.373   Unsupervised Cross-Dataset Transfer Learning for Person Re-identification

[Link to Paper]

Most existing person re-identification (Re-ID) approaches follow a supervised learning framework, in which a large number of labelled matching pairs are required for training. This severely limits their scalability in realworld applications. To overcome this limitation, we develop a novel cross-dataset transfer learning approach to learn a discriminative representation. It is unsupervised in the sense that the target dataset is completely unlabelled. Specifically, we present an multi-task dictionary learning method which is able to learn a dataset-shared but targetdata-biased representation. Experimental results on five benchmark datasets demonstrate that the method significantly outperforms the state-of-the-art.

1.374   A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation

[Link to Paper]

Over the years, datasets and benchmarks have proven their fundamental importance in computer vision research, enabling targeted progress and objective comparisons in many fields. At the same time, legacy datasets may impend the evolution of a field due to saturated algorithm performance and the lack of contemporary, high quality data. In this work we present a new benchmark dataset and evaluation methodology for the area of video object segmentation. The dataset, named DAVIS (Densely Annotated VIdeo Segmentation), consists of fifty high quality, Full HD video sequences, spanning multiple occurrences of common video object segmentation challenges such as occlusions, motionblur and appearance changes. Each video is accompanied by densely annotated, pixel-accurate and per-frame ground truth segmentation. In addition, we provide a comprehensive analysis of several state-of-the-art segmentation approaches using three complementary metrics that measure the spatial extent of the segmentation, the accuracy of the silhouette contours and the temporal coherence. The results uncover strengths and weaknesses of current approaches, opening up promising directions for future works.

1.375 PSyCo: Manifold Span Reduction for Super Resolution

[Link to Paper]

The main challenge in Super Resolution (SR) is to discover the mapping between the low- and high-resolution manifolds of image patches, a complex ill-posed problem which has recently been addressed through piecewise linear regression with promising results. In this paper we present a novel regression-based SR algorithm that benefits from an extended knowledge of the structure of both manifolds. We propose a transform that collapses the 16 variations induced from the dihedral group of transforms (i.e. rotations, vertical and horizontal reflections) and antipodality (i.e. diametrically opposed points in the unitary sphere) into a single primitive. The key idea of our transform is to study the different dihedral elements as a group of symmetries within the high-dimensional manifold. We obtain the respective set of mirror-symmetry axes by means of a frequency analysis of the dihedral elements, and we use them to collapse the redundant variability through a modified symmetry distance. The experimental validation of our algorithm shows the effectiveness of our approach, which obtains competitive quality with a dictionary of as little as 32 atoms (reducing other methods' dictionaries by at least a factor of 32) and further pushing the state-of-the-art with a 1024 atoms dictionary.

1.376 Determining occlusions from space and time image reconstructions

[Link to Paper]

The problem of localizing occlusions between consecutive frames of a video is important but rarely tackled on its own. In most works, it is tightly interleaved with the computation of accurate optical flows, which leads to a delicate chicken-and-egg problem. With this in mind, we propose a novel approach to occlusion detection where visibility or not of a point in next frame is formulated in terms of visual reconstruction. The key issue is now to determine how well a pixel in the first image can be "reconstructed" from co-located colors in the next image. We first exploit this reasoning at the pixel level with a new detection criterion. Contrary to the ubiquitous displaced-framedifference and forward-backward flow vector matching, the proposed alternative does not critically depend on a precomputed, dense displacement field, while being shown to be more effective. We then leverage this local modeling within an energy-minimization framework that delivers occlusion maps. An easy-to-obtain collection of parametric motion models is exploited within the energy to provide the required level of motion information. Our approach outperforms state-of-the-art detection methods on the challenging MPI Sintel dataset.

1.377 Efficient Point Process Inference for Large-scale Object Detection

[Link to Paper]

We tackle the problem of large-scale object detection in images, where the number of objects can be arbitrarily large, and can exhibit significant overlap/occlusion. A successful approach to modelling the large-scale nature of this problem has been via point process density functions which jointly encode object qualities and spatial interactions. But the corresponding optimisation problem is typically difficult or intractable, and many of the best current methods rely on Monte Carlo Markov Chain (MCMC) simulation, which converges slowly in a large solution space. We propose an efficient point process inference for largescale object detection using discrete energy minimization. In particular, we approximate the solution space by a finite set of object proposals and cast the point process density function to a corresponding energy function of binary variables whose values indicate which object proposals are accepted. We resort to the local submodular approximation (LSA) based trust-region optimisation to find the optimal solution. Furthermore we analyse the error of LSA approximation, and show how to adjust the point process energy to dramatically speed up the convergence without harming the optimality. We demonstrate the superior efficiency and accuracy of our method using a variety of large-scale object detection applications such as crowd human detection, birds, cells counting/localization.

1.378 Automated 3D Face Reconstruction from Multiple Images using Quality Measures

[Link to Paper]

Automated 3D reconstruction of faces from images is challenging if the image material is difficult in terms of pose, lighting, occlusions and facial expressions, and if the initial 2D feature positions are inaccurate or unreliable. We propose a method that reconstructs individual 3D shapes from multiple single images of one person, judges their quality and then combines the best of all results. This is done separately for different regions of the face. The core element of this algorithm and the focus of our paper is a quality measure that judges a reconstruction without information about the true shape. We evaluate different quality measures, develop a method for combining results, and present a complete processing pipeline for automated reconstruction.

1.379 DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation

[Link to Paper]

This paper considers the task of articulated human pose estimation of multiple people in real world images. We propose an approach that jointly solves the tasks of detection and pose estimation: it infers the number of persons in a scene, identifies occluded body parts, and disambiguates body parts between people in close proximity of each other. This joint formulation is in contrast to previous strategies, that address the problem by first detecting people and subsequently estimating their body pose. We propose a partitioning and labeling formulation of a set of body-part hypotheses generated with CNN-based part detectors. Our formulation, an instance of an integer linear program, implicitly performs non-maximum suppression on the set of part candidates and groups them to form configurations of body parts respecting geometric and appearance constraints. Experiments on four different datasets demonstrate state-of-the-art results for both single person and multi person pose estimation1.

1.380 CNN-N-Gram for Handwriting Word Recognition

[Link to Paper]

Given an image of a handwritten word, a CNN is employed to estimate its n-gram frequency profile, which is the set of n-grams contained in the word. Frequencies for unigrams, bigrams and trigrams are estimated for the entire word and for parts of it. Canonical Correlation Analysis is then used to match the estimated profile to the true profiles of all words in a large dictionary. The CNN that is used employs several novelties such as the use of multiple fully connected branches. Applied to all commonly used handwriting recognition benchmarks, our method outperforms, by a very large margin, all existing methods.

1.381 Do computational models differ systematically from human object perception?

[Link to Paper]

Recent advances in neural networks have revolutionized computer vision, but these algorithms are still outperformed by humans. Could this performance gap be due to systematic differences between object representations in humans and machines? To answer this question we collected a large dataset of 26,675 perceived dissimilarity measurements from 2,801 visual objects across 269 human subjects, and used this dataset to train and test leading computational models. The best model (a combination of all models) accounted for 68% of the explainable variance. Importantly, all computational models showed systematic deviations from perception: (1) They underestimated perceptual distances between

objects with symmetry or large area differences; (2) They overestimated perceptual distances between objects with shared features. Our results reveal critical elements missing in computer vision algorithms and point to explicit encoding of these properties in higher visual areas in the brain.

1.382   Hedged Deep Tracking

[Link to Paper]

In recent years, several methods have been developed to utilize hierarchical features learned from a deep convolutional neural network (CNN) for visual tracking. However, as features from a certain CNN layer characterize an object of interest from only one aspect or one level, the performance of such trackers trained with features from one layer (usually the second to last layer) can be further improved. In this paper, we propose a novel CNN based tracking framework, which takes full advantage of features from different CNN layers and uses an adaptive Hedge method to hedge several CNN based trackers into a single stronger one. Extensive experiments on a benchmark dataset of 100 challenging image sequences demonstrate the effectiveness of the proposed algorithm compared to several state-of-theart trackers.

1.383   Hierarchically Gated Deep Networks for Semantic Segmentation

[Link to Paper]

Semantic segmentation aims to parse the scene structure of images by annotating the labels to each pixel so that images can be segmented into different regions. While image structures usually have various scales, it is difficult to use a single scale to model the spatial contexts for all individual pixels. Multi-scale Convolutional Neural Networks (CNNs) and their variants have made striking success for modeling the global scene structure for an image. However, they are limited in labeling fine-grained local structures like pixels and patches, since spatial contexts might be blindly mixed up without appropriately customizing their scales. To address this challenge, we develop a novel paradigm of multiscale deep network to model spatial contexts surrounding different pixels at various scales. It builds multiple layers of memory cells, learning feature representations for individual pixels at their customized scales by hierarchically absorbing relevant spatial contexts via memory gates between layers. Such Hierarchically Gated Deep Networks (HGDNs) can customize a suitable scale for each pixel, thereby delivering better performance on labeling scene structures of various scales. We conduct the experiments on two datasets, and show competitive results compared with the other multi-scale deep networks on the semantic segmentation task.

178

1.384 TenSR: Multi-Dimensional Tensor Sparse Representation

[Link to Paper]

The conventional sparse model relies on data representation in the form of vectors. It represents the vector-valued or vectorized one dimensional (1D) version of an signal as a highly sparse linear combination of basis atoms from a large dictionary. The 1D modeling, though simple, ignores the inherent structure and breaks the local correlation inside multidimensional (MD) signals. It also dramatically increases the demand of memory as well as computational resources especially when dealing with high dimensional signals. In this paper, we propose a new sparse model TenSR based on tensor for MD data representation along with the corresponding MD sparse coding and MD dictionary learning algorithms. The proposed TenSR model is able to well approximate the structure in each mode inherent in MD signals with a series of adaptive separable structure dictionaries via dictionary learning. The proposed MD sparse coding algorithm by proximal method further reduces the computational cost significantly. Experimental results with real world MD signals, i.e. 3D Multi-spectral images, show the proposed TenSR greatly reduces both the computational and memory costs with competitive performance in comparison with the state-of-the-art sparse representation methods. We believe our proposed TenSR model is a promising way to empower the sparse representation especially for large scale high order signals.

1.385 Volumetric and Multi-View CNNs for Object Classification on 3D Data

[Link to Paper]

3D shape models are becoming widely available and easier to capture, making available 3D information crucial for progress in object classification. Current state-of-theart methods rely on CNNs to address this problem. Recently, we witness two types of CNNs being developed: CNNs based upon volumetric representations versus CNNs based upon multi-view representations. Empirical results from these two types of CNNs exhibit a large gap, indicating that existing volumetric CNN architectures and approaches are unable to fully exploit the power of 3D representations. In this paper, we aim to improve both volumetric CNNs and multi-view CNNs according to extensive analysis of existing approaches. To this end, we introduce two distinct network architectures of volumetric CNNs. In addition, we examine multi-view CNNs, where we introduce multiresolution filtering in 3D. Overall, we are able to outperform current state-of-the-art methods for both volumetric CNNs and multi-view CNNs. We provide extensive experiments designed to evaluate underlying design choices, thus providing a better understanding of the space of methods available for object classification on 3D data.

1.386 3D Reconstruction of Transparent Objects with Position-Normal Consistency

[Link to Paper]

Estimating the shape of transparent and refractive objects is one of the few open problems in 3D reconstruction. Under the assumption that the rays refract only twice when traveling through the object, we present the first approach to simultaneously reconstructing the 3D positions and normals of the object's surface at both refraction locations. Our acquisition setup requires only two cameras and one monitor, which serves as the light source. After acquiring the ray-ray correspondences between each camera and the monitor, we solve an optimization function which enforces a new position-normal consistency constraint. That is, the 3D positions of surface points shall agree with the normals required to refract the rays under Snell's law. Experimental results using both synthetic and real data demonstrate the robustness and accuracy of the proposed approach.

1.387 Less is more: zero-shot learning from online textual documents with noise suppression

[Link to Paper]

Classifying a visual concept merely from its associated online textual source, such as a Wikipedia article, is an attractive research topic in zero-shot learning because it alleviates the burden of manually collecting semantic attributes. Recent work has pursued this approach by exploring various ways of connecting the visual and text domains. In this paper, we revisit this idea by going further to consider one important factor: the textual representation is usually too noisy for the zero-shot learning application. This observation motivates us to design a simple yet effective zero-shot learning method that is capable of suppressing noise in the text. Specifically, we propose an l2,1-norm based objective function which can simultaneously suppress the noisy signal in the text and learn a function to match the text document and visual features. We also develop an optimization algorithm to efficiently solve the resulting problem. By conducting experiments on two large datasets, we demonstrate that the proposed method significantly outperforms those competing methods which rely on online information sources but with no explicit noise suppression. Furthermore, we make an in-depth analysis of the proposed method and provide insight as to what kind of information in documents is useful for zero-shot learning.

1.388 Joint Training of Cascaded CNN for Face Detection

[Link to Paper]

Cascade has been widely used in face detection, where classifier with low computation cost can be firstly used to shrink most of the background while keeping the recall. The cascade in detection is popularized by seminal Viola-Jones framework and then widely used in other pipelines, such as DPM and CNN. However, to our best knowledge, most of the previous detection methods use cascade in a greedy manner, where previous stages in cascade are fixed when training a new stage. So optimizations of different CNNs are isolated. In this paper, we propose joint training to achieve end-to-end optimization for CNN cascade. We show that the back propagation algorithm used in training CNN can be naturally used in training CNN cascade. We present how jointly training can be conducted on naive CNN cascade and more sophisticated region proposal network (RPN) and fast R-CNN. Experiments on face detection benchmarks verify the advantages of the joint training.

1.389 Equiangular Kernel Dictionary Learning with Applications to Dynamic Texture Analysis

[Link to Paper]

Most existing dictionary learning algorithms consider a linear sparse model, which often cannot effectively characterize the nonlinear properties present in many types of visual data, e.g. dynamic texture (DT). Such nonlinear properties can be exploited by the so-called kernel sparse coding. This paper proposed an equiangular kernel dictionary learning method with optimal mutual coherence to exploit the nonlinear sparsity of high-dimensional visual data. Two main issues are addressed in the proposed method: (1) coding stability for redundant dictionary of infinite-dimensional space; and (2) computational efficiency for computing kernel matrix of training samples of high-dimensional data. The proposed kernel sparse coding method is applied to dynamic texture analysis with both local DT pattern extraction and global DT pattern characterization. The experimental results showed its performance gain over existing methods.

1.390 Object Co-segmentation via Graph Optimized-Flexible Manifold Ranking

[Link to Paper]

Aiming at automatically discovering the common objects contained in a set of relevant images and segmenting them as foreground simultaneously, object co-segmentation has become an active research topic in recent years. Although a number of approaches have been proposed to address this problem, many of them are designed with the misleading assumption, unscalable prior, or low flexibility and thus still suffer from certain limitations,

which reduces their capability in the real-world scenarios. To alleviate these limitations, we propose a novel two-stage co-segmentation framework, which introduces the weak background prior to establish a globally close-loop graph to represent the common object and union background separately. Then a novel graph optimized-flexible manifold ranking algorithm is proposed to flexibly optimize the graph connection and node labels to co-segment the common objects. Experiments on three image datasets demonstrate that our method outperforms other state-of-the-art methods.

1.391 Sparse Coding for Classification via Discrimination Ensemble

[Link to Paper]

Discriminative sparse coding has emerged as a promising technique in image analysis and recognition, which couples the process of classifier training and the process of dictionary learning for improving the discriminability of sparse codes. Many existing approaches consider only a simple single linear classifier whose discriminative power is rather weak. In this paper, we proposed a discriminative sparse coding method which jointly learns a dictionary for sparse coding and an ensemble classifier for discrimination. The ensemble classifier is composed of a set of linear predictors and constructed via both subsampling on data and subspace projection on sparse codes. The advantages of the proposed method over the existing ones are multi-fold: better discriminability of sparse codes, weaker dependence on peculiarities of training data, and more expressibility of classifier for classification. These advantages are also justified in the experiments, as our method outperformed several recent methods in several recognition tasks.

1.392 Unbiased Photometric Stereo for Colored Surfaces: A Variational Approach

[Link to Paper]

3D shape recovery using photometric stereo (PS) gained increasing attention in the computer vision community in the last three decades due to its ability to recover the thinnest geometric structures. Yet, the reliabiliy of PS for color images is difficult to guarantee, because existing methods are usually formulated as the sequential estimation of the colored albedos, the normals and the depth. Hence, the overall reliability depends on that of each subtask. In this work we propose a new formulation of color photometric stereo, based on image ratios, that makes the technique independent from the albedos. This allows the unbiased 3Dreconstruction of colored surfaces in a single step, by solving a system of linear PDEs using a variational approach.

1.393 From Dusk till Dawn: Modeling in the Dark

[Link to Paper]

Internet photo collections naturally contain a large variety of illumination conditions, with the largest difference between day and night images. Current modeling techniques do not embrace the broad illumination range often leading to reconstruction failure or severe artifacts. We present an algorithm that leverages the appearance variety to obtain more complete and accurate scene geometry along with consistent multi-illumination appearance information. The proposed method relies on automatic scene appearance grouping, which is used to obtain separate dense 3D models. Subsequent model fusion combines the separate models into a complete and accurate reconstruction of the scene. In addition, we propose a method to derive the appearance information for the model under the different illumination conditions, even for scene parts that are not observed under one illumination condition. To achieve this, we develop a cross-illumination color transfer technique. We evaluate our method on a large variety of landmarks from across Europe reconstructed from a database of 7.4M images.

1.394 4527

[Link to Paper]

1.395 3D Action Recognition from Novel Viewpoints

[Link to Paper]

We propose a human pose representation model that transfers human poses acquired from different unknown views to a view-invariant high-level space. The model is a deep convolutional neural network and requires a large corpus of multiview training data which is very expensive to acquire. Therefore, we propose a method to generate this data by fitting synthetic 3D human models to real motion capture data and rendering the human poses from numerous viewpoints. While learning the CNN model, we do not use action labels but only the pose labels after clustering all training poses into k clusters. The proposed model is able to generalize to real depth images of unseen poses without the need for re-training or fine-tuning. Real depth videos are passed through the model frame-wise to extract viewinvariant features. For spatio-temporal representation, we propose group sparse Fourier Temporal Pyramid which robustly encodes the action specific most discriminative output features of the proposed human pose model. Experiments on two multiview and three single-view benchmark datasets show that the proposed method dramatically outperforms existing state-of-the-art in action recognition.

1.396 Detecting events and key actors in multi-person videos

[Link to Paper]

Multi-person event recognition is a challenging task, often with many people active in the scene but only a small subset contributing to an actual event. In this paper, we propose a model which learns to detect events in such videos while automatically "attending" to the people responsible for the event. Our model does not use explicit annotations regarding who or where those people are during training and testing. In particular, we track people in videos and use a recurrent neural network (RNN) to represent the track features. We learn time-varying attention weights to combine these features at each time-instant. The attended features are then processed using another RNN for event detection/classification. Since most video datasets with multiple people are restricted to a small number of videos, we also collected a new basketball dataset comprising 257 basketball games with 14K event annotations corresponding to 11 event classes. Our model outperforms state-of-the-art methods for both event classification and detection on this new dataset. Additionally, we show that the attention mechanism is able to consistently localize the relevant players.

1.397 Dense Monocular Depth Estimation in Complex Dynamic Scenes

[Link to Paper]

We present an approach to dense depth estimation from a single monocular camera that is moving through a dynamic scene. The approach produces a dense depth map from two consecutive frames. Moving objects are reconstructed along with the surrounding environment. We provide a novel motion segmentation algorithm that segments the optical flow field into a set of motion models, each with its own epipolar geometry. We then show that the scene can be reconstructed based on these motion models by optimizing a convex program. The optimization jointly reasons about the scales of different objects and assembles the scene in a common coordinate frame, determined up to a global scale. Experimental results demonstrate that the presented approach outperforms prior methods for monocular depth estimation in dynamic scenes. Figure 1: Given two frames from a monocular video of a dynamic scene captured by a single moving camera, our approach computes a dense depth map that reproduces the spatial layout of the scene, including the moving objects. Top: input frames. The white vehicle is approaching the camera, while the camera itself undergoes forward translation and in-plane rotation. Bottom: the estimated depth map.

1.398 Theory and Practice of Structure-from-Motion using Affine Correspondences

[Link to Paper]

Affine Correspondences (ACs) are more informative than Point Correspondences (PCs) that are used as input in mainstream algorithms for Structure-from-Motion (SfM). Since ACs enable to estimate models from fewer correspondences, its use can dramatically reduce the number of combinations during the iterative step of sample-and-test that exists in most SfM pipelines. However, using ACs instead of PCs as input for SfM passes by fully understanding the relations between ACs and multi-view geometry, as well as by establishing practical, effective AC-based algorithms. This article is a step forward into this direction, by providing a clear account about how ACs constrain the two-view geometry, and by proposing new algorithms for plane segmentation and visual odometry that compare favourably with respect to methods relying in PCs.

1.399 MDL-CW: A Multimodal Deep Learning Framework with Cross Weights

[Link to Paper]

Deep learning has received much attention as of the most powerful approaches for multimodal representation learning in recent years. An ideal model for multimodal data can reason about missing modalities using the available ones, and usually provides more information when multiple modalities are being considered. All the previous deep models contain separate modality-specific networks and find a shared representation on top of those networks. Therefore, they only consider high level interactions between modalities to find a joint representation for them. In this paper, we propose a multimodal deep learning framework (MDLCW) that exploits the cross weights between representation of modalities, and try to gradually learn interactions of the modalities in a deep network manner (from low to high level interactions). Moreover, we theoretically show that considering these interactions provide more intra-modality information, and introduce a multi-stage pre-training method that is based on the properties of multi-modal data. In the proposed framework, as opposed to the existing deep methods for multi-modal data, we try to reconstruct the representation of each modality at a given level, with representation of other modalities in the previous layer. Extensive experimental results show that the proposed model outperforms state-of-the-art information retrieval methods for both image and text queries on the PASCAL-sentence and SUNAttribute databases.

1.400 CoMaL: Good Features to Match on Object Boundaries

[Link to Paper]

Traditional Feature Detectors and Trackers use information aggregation in 2D patches to detect and match discriminative patches. However, this information does not remain the same at object boundaries when there is object motion against a significantly varying background. In this paper, we propose a new approach for feature detection, tracking and re-detection that gives significantly improved results at the object boundaries. We utilize level lines or iso-intensity curves that often remain stable and can be reliably detected even at the object boundaries, which they often trace. Stable portions of long level lines are detected and points of high curvature are detected on such curves for corner detection. Further, this level line is used to separate the portions belonging to the two objects, which is then used for robust matching of such points. While such CoMaL (Corners on Maximally-stable Level Line Segments) points were found to be much more reliable at the object boundary regions, they perform comparably at the interior regions as well. This is illustrated in exhaustive experiments on realworld datasets.

1.401 You Only Look Once: Unified, Real-Time Object Detection

[Link to Paper]

We present YOLO, a new approach to object detection. Prior work on object detection re-purposes classifiers to perform detection. Instead, we frame object detection as a regression problem to spatially separated bounding boxes and associated class probabilities. A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation. Since the whole detection pipeline is a single network, it can be optimized end-to-end directly on detection performance. Our unified architecture is extremely fast. Our base YOLO model processes images in real-time at 45 frames per second. A smaller version of the network, Fast YOLO, processes an astounding 155 frames per second while still achieving double the mAP of other real-time detectors. Compared to state-of-the-art detection systems, YOLO makes more localization errors but is less likely to predict false positives on background. Finally, YOLO learns very general representations of objects. It outperforms other detection methods, including DPM and R-CNN, when generalizing from natural images to other domains like artwork.

1.402 Learning Deep Representations of Fine-Grained Visual Descriptions

[Link to Paper]

State-of-the-art methods for zero-shot visual recognition formulate learning as a joint embedding problem of images and side information. In these formulations the current best complement to visual features are attributes: manuallyencoded vectors describing shared characteristics among categories. Despite good performance, attributes have limitations: (1) finer-grained recognition requires commensurately more attributes, and (2) attributes do not provide a natural language interface. We propose to overcome these limitations by training neural language models from scratch; i.e. without pre-training and only consuming words and characters. Our proposed models train end-to-end to align with the fine-grained and category-specific content of images. Natural language provides a flexible and compact way of encoding only the salient visual aspects for distinguishing categories. By training on raw text, our model can do inference on raw text as well, providing humans a familiar mode both for annotation and retrieval. Our model achieves strong performance on zero-shot text-based image retrieval and significantly outperforms the attribute-based state-of-the-art for zero-shot classification on the CaltechUCSD Birds 200-2011 dataset.

1.403 Deep Reflectance Maps

[Link to Paper]

Undoing the image formation process and therefore decomposing appearance into its intrinsic properties is a challenging task due to the under-constrained nature of this inverse problem. While significant progress has been made on inferring shape, materials and illumination from images only, progress in an unconstrained setting is still limited. We propose a convolutional neural architecture to estimate reflectance maps of specular materials in natural lighting conditions. We achieve this in an end-to-end learning formulation that directly predicts a reflectance map from the image itself. We show how to improve estimates by facilitating additional supervision in an indirect scheme that first predicts surface orientation and afterwards predicts the reflectance map by a learning-based sparse data interpolation. In order to analyze performance on this difficult task, we propose a new challenge of Specular MAterials on SHapes with complex IllumiNation (SMASHINg) using both synthetic and real images. Furthermore, we show the application of our method to a range of image editing tasks on real images.

1.404 Three-Dimensional Object Detection and Layout Prediction using Clouds of Oriented Gradients

[Link to Paper]

We develop new representations and algorithms for three-dimensional (3D) object detection and spatial layout prediction in cluttered indoor scenes. RGB-D images are traditionally described by local geometric features of the 3D point cloud. We propose a cloud of oriented gradient (COG) descriptor that links the 2D appearance and 3D pose of object categories, and thus accurately models how perspective projection affects perceived image boundaries. We also propose a "Manhattan voxel" representation which better captures the 3D room layout geometry of common indoor environments. Effective classification rules are learned via a structured prediction framework that accounts for the intersection-over-union overlap of hypothesized 3D cuboids with human annotations, as well as orientation estimation errors. Contextual relationships among categories and layout are captured via a cascade of classifiers, leading to holistic scene hypotheses with improved accuracy. Our model is learned solely from annotated RGB-D images, without the benefit of CAD models, but nevertheless its performance substantially exceeds the state-of-the-art on the SUN RGB-D database. Avoiding CAD models allows easier learning of detectors for many object categories.

1.405   From Bows to Arrows: Rolling Shutter Rectification of Urban Scenes

[Link to Paper]

The rule of perspectivity that 'straight-lines-mustremain-straight' is easily inflected in CMOS cameras by distortions introduced by motion. Lines can be rendered as curves due to the row-wise exposure mechanism known as rolling shutter (RS). We solve the problem of correcting distortions arising from handheld cameras due to RS effect from a single image free from motion blur with special relevance to urban scenes. We develop a procedure to extract prominent curves from the RS image since this is essential for deciphering the varying row-wise motion. We pose an optimization problem with line desirability costs based on straightness, angle, and length, to resolve the geometric ambiguities while estimating the camera motion based on a rotation-only model assuming known camera intrinsic matrix. Finally, we rectify the RS image based on the estimated camera trajectory using inverse mapping. We show rectification results for RS images captured using mobile phone cameras. We also compare our single image method against existing video and nonblind RS rectification methods that typically require multiple images.

1.406   Joint Probabilistic Matching Using m-Best Solutions

[Link to Paper]

(Abstract could not be extracted from PDF)

1.407 Learning Action Maps of Large Environments via First-Person Vision

[Link to Paper]

When people observe and interact with physical spaces, they are able to associate functionality to regions in the environment. Our goal is to automate dense functional understanding of large spaces by leveraging sparse activity demonstrations recorded from an ego-centric viewpoint. The method we describe enables functionality estimation in large scenes where people have behaved, as well as novel scenes where no behaviors are observed. Our method learns and predicts "Action Maps", which encode the ability for a user to perform activities at various locations. With the usage of an egocentric camera to observe human activities, our method scales with the size of the scene without the need for mounting multiple static surveillance cameras and is well-suited to the task of observing activities up-close. We demonstrate that by capturing appearance-based attributes of the environment and associating these attributes with activity demonstrations, our proposed mathematical framework allows for the prediction of Action Maps in new environments. Additionally, we offer a preliminary glance of the applicability of Action Maps by demonstrating a proof-ofconcept application in which they are used in concert with activity detections to perform localization.

1.408 Temporal Action Detection using a Statistical Language Model

[Link to Paper]

While current approaches to action recognition on presegmented video clips already achieve high accuracies, temporal action detection is still far from comparably good results. Automatically locating and classifying the relevant action segments in videos of varying lengths proves to be a challenging task. We propose a novel method for temporal action detection including statistical length and language modeling to represent temporal and contextual structure. Our approach aims at globally optimizing the joint probability of three components, a length and language model and a discriminative action model, without making intermediate decisions. The problem of finding the most likely action sequence and the corresponding segment boundaries in an exponentially large search space is addressed by dynamic programming. We provide an extensive evaluation of each model component on Thumos 14, a large action detection dataset, and report state-of-the-art results on three datasets.

1.409 The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes

[Link to Paper]

Vision-based semantic segmentation in urban scenarios is a key functionality for autonomous driving. Recent revolutionary results of deep convolutional neural networks (DCNNs) foreshadow the advent of reliable classifiers to perform such visual tasks. However, DCNNs require learning of many parameters from raw images; thus, having a sufficient amount of diverse images with class annotations is needed. These annotations are obtained via cumbersome, human labour which is particularly challenging for semantic segmentation since pixel-level annotations are required. In this paper, we propose to use a virtual world to automatically generate realistic synthetic images with pixel-level annotations. Then, we address the question of how useful such data can be for semantic segmentation in particular, when using a DCNN paradigm. In order to answer this question we have generated a synthetic collection of diverse urban images, named SYNTHIA, with automatically generated class annotations. We use SYNTHIA in combination with publicly available real-world urban images with manually provided annotations. Then, we conduct experiments with DCNNs that show how the inclusion of SYNTHIA in the training stage significantly improves performance on the semantic segmentation task.

1.410 Information-Driven Adaptive Structured-Light Scanners

[Link to Paper]

Sensor planning and active sensing, long studied in robotics, adapt sensor parameters to maximize a utility function while constraining resource expenditures. Here we consider information gain as the utility function. While these concepts are often used to reason about 3D sensors, these are usually treated as a predefined, black-box, component. In this paper we show how the same principles can be used as part of the 3D sensor. We describe the relevant generative model for structured-light 3D scanning and show how adaptive pattern selection can maximize information gain in an open-loop-feedback manner. We then demonstrate how different choices of relevant variable sets (corresponding to the subproblems of locatization and mapping) lead to different criteria for pattern selection and can be computed in an online fashion. We show results for both subproblems with several pattern dictionary choices and demonstrate their usefulness for pose estimation and depth acquisition.

1.411 Adaptive 3D Face Reconstruction from Unconstrained Photo Collections

[Link to Paper]

Given a collection of "in-the-wild" face images captured under a variety of unknown pose, expression, and illumination conditions, this paper presents a method for reconstructing a 3D face surface model of an individual along with albedo information. Motivated by the success of recent face reconstruction techniques on large photo collections, we extend prior work to adapt to low quality photo collections with fewer images. We achieve this by fitting a 3D Morphable Model to form a personalized template and developing a novel photometric stereo formulation, under a coarse-to-fine scheme. Superior experimental results are reported on synthetic and real-world photo collections.

1.412 Some like it hot - visual guidance for preference prediction

[Link to Paper]

For people first impressions of someone are of determining importance. They are hard to alter through further information. This begs the question if a computer can reach the same judgement. Earlier research has already pointed out that age, gender, and average attractiveness can be estimated with reasonable precision. We improve the stateof-the-art, but also predict - based on someone's known preferences - how much that particular person is attracted to a novel face. Our computational pipeline comprises a face detector, convolutional neural networks for the extraction of deep features, standard support vector regression for gender, age and facial beauty, and - as the main novelties - visual regularized collaborative filtering to infer interperson preferences as well as a novel regression technique for handling visual queries without rating history. We validate the method using a very large dataset from a dating site as well as images from celebrities. Our experiments yield convincing results, i.e. we predict 76% of the ratings correctly solely based on an image, and reveal some sociologically relevant conclusions. We also validate our collaborative filtering solution on the standard MovieLens rating dataset, augmented with movie posters, to predict an individuals movie rating. We demonstrate our algorithms on howhot.io which went viral around the Internet with more than 50 million pictures evaluated in the first month.

1.413 Monocular Depth Estimation Using Neural Regression Forest

[Link to Paper]

This paper presents a novel deep architecture, called neural regression forest (NRF), for depth estimation from a single image. NRF combines random forests and convolutional neural networks (CNNs). Scanning windows extracted from the image represent samples which are passed down the trees of NRF for predicting their depth. At every tree node, the sample is filtered with a CNN associated with that node. Results of the convolutional filtering are passed to left and right children nodes, i.e., corresponding CNNs, with a Bernoulli probability, until the leaves, where depth estimations are made. CNNs at every node are designed to have fewer parameters than seen in recent work, but their stacked processing along a path in the tree effectively amounts to a deeper CNN. NRF allows for parallelizable training of all "shallow" CNNs, and efficient enforcing of smoothness in depth estimation results. Our evaluation on the benchmark Make3D and NYUv2 datasets demonstrates that NRF outperforms the state of the art, and gracefully handles gradually decreasing training datasets.

1.414 Convexity Shape Constraints for Image Segmentation

[Link to Paper]

Segmenting an image into multiple components is a central task in computer vision. In many practical scenarios, prior knowledge about plausible components is available. Incorporating such prior knowledge into models and algorithms for image segmentation is highly desirable, yet can be non-trivial. In this work, we introduce a new approach that allows, for the first time, to constrain some or all components of a segmentation to have convex shapes. Specifically, we extend the Minimum Cost Multicut Problem by a class of constraints that enforce convexity. To solve instances of this NP-hard integer linear program to optimality, we separate the proposed constraints in the branch-and-cut loop of a state-of-the-art ILP solver. Results on photographs and micrographs demonstrate the effectiveness of the approach as well as its advantages over the state-of-the-art heuristic.

1.415 Large-Scale Location Recognition and the Geometric Burstiness Problem

[Link to Paper]

Visual location recognition is the task of determining the place depicted in a query image from a given database of geo-tagged images. Location recognition is often cast as an image retrieval problem and recent research has almost exclusively focused on improving the chance that a relevant database image is ranked high enough after retrieval. The implicit assumption is that the number of inliers found by spatial verification can be used to distinguish between a related and an unrelated database photo with high precision. In this paper,

we show that this assumption does not hold for large datasets due to the appearance of geometric bursts, i.e., sets of visual elements appearing in similar geometric configurations in unrelated database photos. We propose algorithms for detecting and handling geometric bursts. Although conceptually simple, using the proposed weighting schemes dramatically improves the recall that can be achieved when high precision is required compared to the standard re-ranking based on the inlier count. Our approach is easy to implement and can easily be integrated into existing location recognition systems.

1.416 Sparse to Dense 3D Reconstruction from Rolling Shutter Images

[Link to Paper]

It is well known that the rolling shutter effect in images captured with a moving rolling shutter camera causes inaccuracies to 3D reconstructions. The problem is further aggravated with weak visual connectivity from wide baseline images captured with a fast moving camera. In this paper, we propose and implement a pipeline for sparse to dense 3D construction with wide baseline images captured from a fast moving rolling shutter camera. Specifically, we propose a cost function for Bundle Adjustment (BA) that models the rolling shutter effect, incorporates GPS/INS readings, and enforces pairwise smoothness between neighboring poses. We optimize over the 3D structures, camera poses and velocities. We also introduce a novel interpolation scheme for the rolling shutter plane sweep stereo algorithm that allows us to achieve a 7 speed up in the depth map computations for dense reconstruction without losing accuracy. We evaluate our proposed pipeline over a 2.6km image sequence captured with a rolling shutter camera mounted on a moving car.

1.417 Semantic 3D Reconstruction with Continuous Regularization and Ray Potentials Using a Visibility Consistency Constraint

[Link to Paper]

We propose an approach for dense semantic 3D reconstruction which uses a data term that is defined as potentials over viewing rays, combined with continuous surface area penalization. Our formulation is a convex relaxation which we augment with a crucial non-convex constraint that ensures exact handling of visibility. To tackle the non-convex minimization problem, we propose a majorizeminimize type strategy which converges to a critical point. We demonstrate the benefits of using the non-convex constraint experimentally. For the geometry-only case, we set a new state of the art on two datasets of the commonly used Middlebury multi-view stereo benchmark. Moreover, our general-purpose formulation directly reconstructs thin objects, which are usually treated with specialized

algorithms. A qualitative evaluation on the dense semantic 3D reconstruction task shows that we improve significantly over previous methods.

1.418 Single Image Camera Calibration with Lenticular Arrays for Augmented Reality

[Link to Paper]

We consider the problem of camera pose estimation for a scenario where the camera may have continuous and unknown changes in its focal length. Understanding frame by frame changes in camera focal length is vital to accurately estimating camera pose and vital to accurately rendering virtual objects in a scene with the correct perspective. However, most approaches to camera calibration require geometric constraints from many frames or the observation of a 3D calibration object – both of which may not be feasible in augmented reality settings. This paper introduces a calibration object based on a flat lenticular array that creates a color coded light-field whose observed color changes depending on the angle from which it is viewed. We derive an approach to estimate the focal length of the camera and the relative pose of an object from a single image. We characterize the performance of camera calibration across various focal lengths and camera models, and we demonstrate the advantages of the focal length estimation in rendering a virtual object in a video with constant zooming.

1.419 Structure-from-Motion Revisited

[Link to Paper]

Incremental Structure-from-Motion is a prevalent strategy for 3D reconstruction from unordered image collections. While incremental reconstruction systems have tremendously advanced in all regards, robustness, accuracy, completeness, and scalability remain the key problems towards building a truly general-purpose pipeline. We propose a new SfM technique that improves upon the state of the art to make a further step towards this ultimate goal. The full reconstruction pipeline is released to the public as an open-source implementation.

1.420 Instance-level video segmentation from object tracks

[Link to Paper]

We address the problem of segmenting multiple object instances in complex videos. Our method does not require manual pixel-level annotation for training, and relies instead on

readily-available object detectors or visual object tracking only. Given object bounding boxes at input, we cast video segmentation as a weakly-supervised learning problem. Our proposed objective combines (a) a discriminative clustering term for background segmentation, (b) a spectral clustering one for grouping pixels of same object instances, and (c) linear constraints enabling instance-level segmentation. We propose a convex relaxation of this problem and solve it efficiently using the Frank-Wolfe algorithm. We report results and compare our method to several baselines on a new video dataset for multi-instance person segmentation.

1.421 Robust, Real-Time 3D Tracking of Multiple Objects with Similar Appearances

[Link to Paper]

This paper proposes a novel method for tracking multiple moving objects and recovering their three-dimensional (3D) models separately using multiple calibrated cameras. For robustly tracking objects with similar appearances, the proposed method uses geometric information regarding 3D scene structure rather than appearance. A major limitation of previous techniques is foreground confusion, in which the shapes of objects and/or ghosting artifacts are ignored and are hence not appropriately specified in foreground regions. To overcome this limitation, our method classifies foreground voxels into targets (objects and artifacts) in each frame using a novel, probabilistic two-stage framework. This is accomplished by step-wise application of a track graph describing how targets interact and the maximum a posteriori expectation-maximization algorithm for the estimation of target parameters. We introduce mixture models with semiparametric component distributions regarding 3D target shapes. In order to not confuse artifacts with objects of interest, we automatically detect and track artifacts based on a closed-world assumption. Experimental results show that our method outperforms state-of-the-art trackers on seven public sequences while achieving real-time performance.

1.422 Optical Flow with Semantic Segmentation and Localized Layers

[Link to Paper]

Existing optical flow methods make generic, spatially homogeneous, assumptions about the spatial structure of the flow. In reality, optical flow varies across an image depending on object class. Simply put, different objects move differently. Here we exploit recent advances in static semantic scene segmentation to segment the image into objects of different types. We define different models of image motion in these regions depending on the type of object. For example, we model the motion on roads with homographies, vegetation

with spatially smooth flow, and independently moving objects like cars and planes with affine motion plus deviations. We then pose the flow estimation problem using a novel formulation of localized layers, which addresses limitations of traditional layered models for dealing with complex scene motion. Our semantic flow method achieves the lowest error of any published monocular method in the KITTI-2015 flow benchmark and produces qualitatively better flow and segmentation than recent top methods on a wide range of natural videos.

1.423 Estimating Sparse Signals with Smooth Support via Convex Programming and Block Sparsity

[Link to Paper]

Conventional algorithms for sparse signal recovery and sparse representation rely on 1-norm regularized variational methods. However, when applied to the reconstruction of sparse images, i.e., images where only a few pixels are non-zero, simple 1-norm-based methods ignore potential correlations in the support between adjacent pixels. In a number of applications, one is interested in images that are not only sparse, but also have a support with smooth (or contiguous) boundaries. Existing algorithms that take into account such a support structure mostly rely on nonconvex methods and–as a consequence–do not scale well to high-dimensional problems and/or do not converge to global optima. In this paper, we explore the use of new block 1-norm regularizers, which enforce image sparsity while simultaneously promoting smooth support structure. By exploiting the convexity of our regularizers, we develop new computationally-efficient recovery algorithms that guarantee global optimality. We demonstrate the efficacy of our regularizers on a variety of imaging tasks including compressive image recovery, image restoration, and robust PCA.

1.424 NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis

[Link to Paper]

Recent approaches in depth-based human activity analysis achieved outstanding performance and proved the effectiveness of 3D representation for classification of action classes. Currently available depth-based and RGB+Dbased action recognition benchmarks have a number of limitations, including the lack of training samples, distinct class labels, camera views and variety of subjects. In this paper we introduce a large-scale dataset for RGB+D human action recognition with more than 56 thousand video samples and 4 million frames, collected from 40 distinct subjects. Our dataset contains 60 different action classes including daily, mutual, and health-related actions. In addition, we propose a new recurrent

196

neural network structure to model the long-term temporal correlation of the features for each body part, and utilize them for better action classification. Experimental results show the advantages of applying deep learning methods over state-of-the-art handcrafted features on the suggested cross-subject and crossview evaluation criteria for our dataset. The introduction of this large scale dataset will enable the community to apply, develop and adapt various data-hungry learning techniques for the task of depth-based and RGB+D-based human activity analysis.

1.425 Refining Architectures of Deep Convolutional Neural Networks

[Link to Paper]

Deep Convolutional Neural Networks (CNNs) have recently evinced immense success for various image recognition tasks [11, 27]. However, a question of paramount importance is somewhat unanswered in deep learning research - is the selected CNN optimal for the dataset in terms of accuracy and model size? In this paper, we intend to answer this question and introduce a novel strategy that alters the architecture of a given CNN for a specified dataset, to potentially enhance the original accuracy while possibly reducing the model size. We use two operations for architecture refinement, viz. stretching and symmetrical splitting. Stretching increases the number of hidden units (nodes) in a given CNN layer, while a symmetrical split of say K between two layers separates the input and output channels into K equal groups, and connects only the corresponding input-output channel groups. Our procedure starts with a pre-trained CNN for a given dataset, and optimally decides the stretch and split factors across the network to refine the architecture. We empirically demonstrate the necessity of the two operations. We evaluate our approach on two natural scenes attributes datasets, SUN Attributes [16] and CAMIT-NSAD [20], with architectures of GoogleNet and VGG-11, that are quite contrasting in their construction. We justify our choice of datasets, and show that they are interestingly distinct from each other, and together pose a challenge to our architectural refinement algorithm. Our results substantiate the usefulness of the proposed method.

1.426 Min Norm Point Algorithm for Higher Order MRF-MAP Inference

[Link to Paper]

Many tasks in computer vision and machine learning can be modelled as the inference problems in an MRF-MAP formulation and can be reduced to minimizing a submodular function. Using higher order clique potentials to model complex dependencies between pixels improves the performance but the current state of the art inference algorithms fail to scale for larger clique sizes. We adapt a well known Min Norm Point algorithm from

mathematical optimization literature to exploit the sum of submodular structure found in the MRF-MAP formulation. Unlike some contemporary methods, we do not make any assumptions (other than submodularity) on the type of the clique potentials. Current state of the art inference algorithms for general submodular function takes many hours for problems with clique size 16, and fail to scale beyond. On the other hand, our algorithm is highly efficient and can perform optimal inference in few seconds even on clique size an order of magnitude larger. The proposed algorithm can even scale to clique sizes of many hundreds, unlocking the usage of really large size cliques for MRF-MAP inference problems in computer vision. We demonstrate the efficacy of our approach by experimenting on synthetic as well as real datasets.

1.427 Slicing Convolutional Neural Network for Crowd Video Understanding

[Link to Paper]

Learning and capturing both appearance and dynamic representations are pivotal for crowd video understanding. Convolutional Neural Networks (CNNs) have shown its remarkable potential in learning appearance representations from images. However, the learning of dynamic representation, and how it can be effectively combined with appearance features for video analysis, remains an open problem. In this study, we propose a novel spatio-temporal CNN, named Slicing CNN (S-CNN), based on the decomposition of 3D feature maps into 2D spatio- and 2D temporal-slices representations. The decomposition brings unique advantages: (1) the model is capable of capturing dynamics of different semantic units such as groups and objects, (2) it learns separated appearance and dynamic representations while keeping proper interactions between them, and (3) it exploits the selectiveness of spatial filters to discard irrelevant background clutter for crowd understanding. We demonstrate the effectiveness of the proposed S-CNN model on the WWW crowd video dataset for attribute recognition and observe significant performance improvements to the state-of-the-art methods (62.55% from 51.84% [21]).

1.428 Ambiguity Helps: Classification with Disagreements in Crowdsourced Annotations

[Link to Paper]

Imagine we show an image to a person and ask her/him to decide whether the scene in the image is warm or not warm, and whether it is easy or not to spot a squirrel in the image. For exactly the same image, the answers to those questions are likely to differ from person to person. This is because the task is inherently ambiguous. Such an ambiguous, therefore

challenging, task is pushing the boundary of computer vision in showing what can and can not be learned from visual data. Crowdsourcing has been invaluable for collecting annotations. This is particularly so for a task that goes beyond a clear-cut dichotomy as multiple human judgments per image are needed to reach a consensus. This paper makes conceptual and technical contributions. On the conceptual side, we define disagreements among annotators as privileged information about the data instance. On the technical side, we propose a framework to incorporate annotation disagreements into the classifiers. The proposed framework is simple, relatively fast, and outperforms classifiers that do not take into account the disagreements, especially if tested on high confidence annotations.

1.429 Learning from the Mistakes of Others: Matching Errors in Cross-Dataset Learning

[Link to Paper]

Can we learn about object classes in images by looking at a collection of relevant 3D models? Or if we want to learn about human (inter-)actions in images, can we benefit from videos or abstract illustrations that show these actions? A common aspect of these settings is the availability of additional or privileged data that can be exploited at training time and that will not be available and not of interest at test time. We seek to generalize the learning with privileged information (LUPI) framework, which requires additional information to be defined per image, to the setting where additional information is a data collection about the task of interest. Our framework minimizes the distribution mismatch between errors made in images and in privileged data. The proposed method is tested on four publicly available datasets: Image+ClipArt, Image+3Dobject, and Image+Video. Experimental results reveal that our new LUPI paradigm naturally addresses the cross-dataset learning.

1.430 The Next Best Underwater View

[Link to Paper]

To image in high resolution large and occlusion-prone scenes, a camera must move above and around. Degradation of visibility due to geometric occlusions and distances is exacerbated by scattering, when the scene is in a participating medium. Moreover, underwater and in other media, artificial lighting is needed. Overall, data quality depends on the observed surface, medium and the timevarying poses of the camera and light source (C&L). This work proposes to optimize C&L poses as they move, so that the surface is scanned efficiently and the descattered recovery has the highest quality. The work generalizes the next best view concept of robot vision to scattering media and cooperative movable lighting.

It also extends descattering to platforms that move optimally. The optimization criterion is information gain, taken from information theory. We exploit the existence of a prior rough 3D model, since underwater such a model is routinely obtained using sonar. We demonstrate this principle in a scaled-down setup. C(t + 1) L(t + 1) C(t) L(t) Figure 1. In a scattering medium at time t, the camera and lightsource poses are represented by C(t), L(t) respectively. The next best underwater view task optimizes the poses C(t+1), L(t+1) for object reconstruction.

1.431 Object Skeleton Extraction in Natural Images by Fusing Scale-associated Deep Side Outputs

[Link to Paper]

Object skeleton is a useful cue for object detection, complementary to the object contour, as it provides a structural representation to describe the relationship among object parts. While object skeleton extraction in natural images is a very challenging problem, as it requires the extractor to be able to capture both local and global image context to determine the intrinsic scale of each skeleton pixel. Existing methods rely on per-pixel based multi-scale feature computation, which results in difficult modeling and high time consumption. In this paper, we present a fully convolutional network with multiple scale-associated side outputs to address this problem. By observing the relationship between the receptive field sizes of the sequential stages in the network and the skeleton scales they can capture, we introduce a scale-associated side output to each stage. We impose supervision to different stages by guiding the scale-associated side outputs toward groundtruth skeletons of different scales. The responses of the multiple scaleassociated side outputs are then fused in a scale-specific way to localize skeleton pixels with multiple scales effectively. Our method achieves promising results on two skeleton extraction datasets, and significantly outperforms other competitors.

1.432 A Benchmark Dataset and Evaluation for Non-Lambertian and Uncalibrated Photometric Stereo

[Link to Paper]

Recent progress on photometric stereo extends the technique to deal with general materials and unknown illumination conditions. However, due to the lack of suitable benchmark data with ground truth shapes (normals), quantitative comparison and evaluation is difficult to achieve. In this paper, we first survey and categorize existing methods using a photometric stereo taxonomy emphasizing on non-Lambertian and uncalibrated methods.

We then introduce the 'DiLiGenT' photometric stereo image dataset with calibrated Directional Lightings, objects of General reflectance, and 'ground Truth' shapes (normals). Based on our dataset, we quantitatively evaluate state-of-the-art photometric stereo methods for general non-Lambertian materials and unknown lightings to analyze their strengths and limitations.

1.433 Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network

[Link to Paper]

Recently, several models based on deep neural networks have achieved great success in terms of both reconstruction accuracy and computational performance for single image super-resolution. In these methods, the low resolution (LR) input image is upscaled to the high resolution (HR) space using a single filter, commonly bicubic interpolation, before reconstruction. This means that the super-resolution (SR) operation is performed in HR space. We demonstrate that this is sub-optimal and adds computational complexity. In this paper, we present the first convolutional neural network (CNN) capable of real-time SR of 1080p videos on a single K2 GPU. To achieve this, we propose a novel CNN architecture where the feature maps are extracted in the LR space. In addition, we introduce an efficient sub-pixel convolution layer which learns an array of upscaling filters to upscale the final LR feature maps into the HR output. By doing so, we effectively replace the handcrafted bicubic filter in the SR pipeline with more complex upscaling filters specifically trained for each feature map, whilst also reducing the computational complexity of the overall SR operation. We evaluate the proposed approach using images and videos from publicly available datasets and show that it performs significantly better (+0.15dB on Images and +0.39dB on Videos) and is an order of magnitude faster than previous CNN-based methods.

1.434 Robust Scene Text Recognition with Automatic Rectification

[Link to Paper]

Recognizing text in natural images is a challenging task with many unsolved problems. Different from those in documents, words in natural images often possess irregular shapes, which are caused by perspective distortion, curved character placement, etc. We propose RARE (Robust text recognizer with Automatic REctification), a recognition model that is robust to irregular text. RARE is a speciallydesigned deep neural network, which consists of a Spatial Transformer Network (STN) and a Sequence Recognition Network (SRN). In testing, an image is firstly rectified via a predicted Thin-Plate-Spline (TPS) transformation,

into a more "readable" image for the following SRN, which recognizes text through a sequence recognition approach. We show that the model is able to recognize several types of irregular text, including perspective text and curved text. RARE is end-to-end trainable, requiring only images and associated text labels, making it convenient to train and deploy the model in practical systems. State-of-the-art or highly-competitive performance achieved on several benchmarks well demonstrates the effectiveness of the proposed model.

1.435 Shape Analysis with Hyperbolic Wasserstein Distance

[Link to Paper]

Shape space is an active research field in computer vision study. The shape distance defined in a shape space may provide a simple and refined index to represent a unique shape. Wasserstein distance defines a Riemannian metric for the Wasserstein space. It intrinsically measures the similarities between shapes and is robust to image noise. Thus it has the potential for the 3D shape indexing and classification research. While the algorithms for computing Wasserstein distance have been extensively studied, most of them only work for genus-0 surfaces. This paper proposes a novel framework to compute Wasserstein distance between general topological surfaces with hyperbolic metric. The computational algorithms are based on Ricci flow, hyperbolic harmonic map, and hyperbolic power Voronoi diagram and the method is general and robust. We apply our method to study human facial expression, longitudinal brain cortical morphometry with normal aging, and cortical shape classification in Alzheimer's disease (AD). Experimental results demonstrate that our method may be used as an effective shape index, which outperforms some other standard shape measures in our AD versus healthy control classification study.

1.436 Tensor Power Iteration for Multi-Graph Matching

[Link to Paper]

Due to its wide range of applications, matching between two graphs has been extensively studied and remains an active topic. By contrast, it is still under-exploited on how to jointly match multiple graphs, partly due to its intrinsic combinatorial intractability. In this work, we address this challenging problem in a principled way under the rank-1 tensor approximation framework. In particular, we formulate multi-graph matching as a combinational optimization problem with two main ingredients: unary matching over graph vertices and structure matching over graph edges, both of which across multiple graphs. Then we propose an efficient power iteration solution for the resulting NP-hard optimization problem. The proposed algorithm has several advantages: 1) the intrinsic matching consistency across multiple graphs based on the high-order tensor optimization; 2) the free

employment of powerful high-order node affinity; 3) the flexible integration between various types of node affinities and edge/hyper-edge affinities. Experiments on diverse and challenging datasets validate the effectiveness of the proposed approach in comparison with state-ofthe-arts.

1.437 Gradient-Domain Image Reconstruction Framework with Intensity-Range and Base-Structure Constraints

[Link to Paper]

This paper presents a novel unified gradient-domain image reconstruction framework with intensity-range constraint and base-structure constraint. The existing method for manipulating base structures and detailed textures are classifiable into two major approaches: i) gradient-domain and ii) layer-decomposition. To generate detail-preserving and artifact-free output images, we combine the benefits of the two approaches into the proposed framework by introducing the intensity-range constraint and the base-structure constraint. To preserve details of the input image, the proposed method takes advantage of reconstructing the output image in the gradient domain, while the output intensity is guaranteed to lie within the specified intensity range, e.g. 0-to-255, by the intensity-range constraint. In addition, the reconstructed image lies close to the base structure by the base-structure constraint, which is effective for restraining artifacts. Experimental results show that the proposed framework is effective for various applications such as tone mapping, seamless image cloning, detail enhancement, and image restoration.

1.438 Where To Look: Focus Regions for Visual Question Answering

[Link to Paper]

We present a method that learns to answer visual questions by selecting image regions relevant to the text-based query. Our method maps textual queries and visual features from various regions into a shared space where they are compared for relevance with an inner product. Our method exhibits significant improvements in answering questions such as "what color," where it is necessary to evaluate a specific location, and "what room," where it selectively identifies informative image regions. Our model is tested on the recently released VQA [1] dataset, which features free-form human-annotated questions and answers. Is it raining? What color is the walk light?

1.439 Automating Carotid Intima-Media Thickness Video Interpretation with Convolutional Neural Networks

Cardiovascular disease (CVD) is the leading cause of mortality yet largely preventable, but the key to prevention is to identify at-risk individuals before adverse events. For predicting individual CVD risk, carotid intima-media thickness (CIMT), a noninvasive ultrasound method, has proven to be valuable, offering several advantages over CT coronary artery calcium score. However, each CIMT examination includes several ultrasound videos, and interpreting each of these CIMT videos involves three operations: (1) select three end-diastolic ultrasound frames (EUF) in the video, (2) localize a region of interest (ROI) in each selected frame, and (3) trace the lumen-intima interface and the media-adventitia interface in each ROI to measure CIMT. These operations are tedious, laborious, and time consuming, a serious limitation that hinders the widespread utilization of CIMT in clinical practice. To overcome this limitation, this paper presents a new system to automate CIMT video interpretation. Our extensive experiments demonstrate that the suggested system performs reliably. The reliable performance is attributable to our unified framework based on convolutional neural networks (CNNs) coupled with our informative image representation and effective post-processing of the CNN outputs, which are uniquely designed for each of the above three operations.

1.440 Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation

Despite the recent advances in automatically describing image contents, their applications have been mostly limited to image caption datasets containing natural images (e.g., Flickr 30k, MSCOCO). In this paper, we present a deep learning model to efficiently detect a disease from an image and annotate its contexts (e.g., location, severity and the affected organs). We employ a publicly available radiology dataset of chest x-rays and their reports, and use its image annotations to mine disease names to train convolutional neural networks (CNNs). In doing so, we adopt various regularization techniques to circumvent the large normalvs-diseased cases bias. Recurrent neural networks (RNNs) are then trained to describe the contexts of a detected disease, based on the deep CNN features. Moreover, we introduce a novel approach to use the weights of the already trained pair of CNN/RNN on the domain-specific image/text dataset, to infer the joint image/text contexts for composite image labeling. Significantly improved image annotation results are demonstrated using the recurrent neural cascade model by taking the joint image/text contexts into account.

1.441 Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs

[Link to Paper]

We address temporal action localization in untrimmed long videos. This is important because videos in real applications are usually unconstrained and contain multiple action instances plus video content of background scenes or other activities. To address this challenging issue, we exploit the effectiveness of deep networks in temporal action localization via three segment-based 3D ConvNets: (1) a proposal network identifies candidate segments in a long video that may contain actions; (2) a classification network learns one-vs-all action classification model to serve as initialization for the localization network; and (3) a localization network fine-tunes the learned classification network to localize each action instance. We propose a novel loss function for the localization network to explicitly consider temporal overlap and achieve high temporal localization accuracy. In the end, only the proposal network and the localization network are used during prediction. On two largescale benchmarks, our approach achieves significantly superior performances compared with other state-of-the-art systems: mAP increases from 1.7% to 7.4% on MEX-action2 and increases from 15.0% to 19.0% on THUMOS 2014.

1.442 Training Region-based Object Detectors with Online Hard Example Mining

[Link to Paper]

The field of object detection has made significant advances riding on the wave of region-based ConvNets, but their training procedure still includes many heuristics and hyperparameters that are costly to tune. We present a simple yet surprisingly effective online hard example mining (OHEM) algorithm for training region-based ConvNet detectors. Our motivation is the same as it has always been detection datasets contain an overwhelming number of easy examples and a small number of hard examples. Automatic selection of these hard examples can make training more effective and efficient. OHEM is a simple and intuitive algorithm that eliminates several heuristics and hyperparameters in common use. But more importantly, it yields consistent and significant boosts in detection performance on benchmarks like PASCAL VOC 2007 and 2012. Its effectiveness increases as datasets become larger and more difficult, as demonstrated by the results on the MS COCO dataset. Moreover, combined with complementary advances in the field, OHEM leads to state-of-the-art results of 78.9% and 76.3% mAP on PASCAL VOC 2007 and 2012 respectively.

1.443 DAG-Recurrent Neural Networks For Scene Labeling

[Link to Paper]

In image labeling, local representations for image units are usually generated from their surrounding image patches, thus long-range contextual information is not effectively encoded. In this paper, we introduce recurrent neural networks (RNNs) to address this issue. Specifically, directed acyclic graph RNNs (DAG-RNNs) are proposed to process DAG-structured images, which enables the network to model long-range semantic dependencies among image units. Our DAG-RNNs are capable of tremendously enhancing the discriminative power of local representations, which significantly benefits the local classification. Meanwhile, we propose a novel class weighting function that attends to rare classes, which phenomenally boosts the recognition accuracy for non-frequent classes. Integrating with convolution and deconvolution layers, our DAG-RNNs achieve new state-of-the-art results on the challenging SiftFlow, CamVid and Barcelona benchmarks.

1.444 LOMo: Latent Ordinal Model for Facial Analysis in Videos

[Link to Paper]

We study the problem of facial analysis in videos. We propose a novel weakly supervised learning method that models the video event (expression, pain etc.) as a sequence of automatically mined, discriminative sub-events (e.g. onset and offset phase for smile, brow lower and cheek raise for pain). The proposed model is inspired by the recent works on Multiple Instance Learning and latent SVM/HCRF it extends such frameworks to model the ordinal or temporal aspect in the videos, approximately. We obtain consistent improvements over relevant competitive baselines on four challenging and publicly available video based facial analysis datasets for prediction of expression, clinical pain and intent in dyadic conversations. In combination with complimentary features, we report state-of-the-art results on these datasets.

1.445 Fashion Style in 128 Floats: Joint Ranking and Classification using Weak Data for Feature Extraction

[Link to Paper]

We propose a novel approach for learning features from weakly-supervised data by joint ranking and classification. In order to exploit data with weak labels, we jointly train a feature extraction network with a ranking loss and a classification network with a cross-entropy loss. We obtain highquality compact discriminative features with few parameters,

206

learned on relatively small datasets without additional annotations. This enables us to tackle tasks with specialized images not very similar to the more generic ones in existing fully-supervised datasets. We show that the resulting features in combination with a linear classifier surpass the state-of-the-art on the Hipster Wars dataset despite using features only 0.3% of the size. Our proposed features significantly outperform those obtained from networks trained on ImageNet, despite being 32 times smaller (128 singleprecision floats), trained on noisy and weakly-labeled data, and using only 1.5% of the number of parameters.1.

1.446 A Multi-Stream Bi-Directional Recurrent Neural Network for Fine-Grained Action Detection

[Link to Paper]

We present a multi-stream bi-directional recurrent neural network for fine-grained action detection. Recently, twostream convolutional neural networks (CNNs) trained on stacked optical flow and image frames have been successful for action recognition in videos. Our system uses a tracking algorithm to locate a bounding box around the person, which provides a frame of reference for appearance and motion and also suppresses background noise that is not within the bounding box. We train two additional streams on motion and appearance cropped to the tracked bounding box, along with full-frame streams. Our motion streams use pixel trajectories of a frame as raw features, in which the displacement values corresponding to a moving scene point are at the same spatial position across several frames. To model long-term temporal dynamics within and between actions, the multi-stream CNN is followed by a bi-directional Long Short-Term Memory (LSTM) layer. We show that our bi-directional LSTM network utilizes about 8 seconds of the video sequence to predict an action label. We test on two action detection datasets: the MPII Cooking 2 Dataset, and a new MERL Shopping Dataset that we introduce and make available to the community with this paper. The results demonstrate that our method significantly outperforms state-of-the-art action detection methods on both datasets.

1.447 First Person Action Recognition Using Deep Learned Descriptors

[Link to Paper]

We focus on the problem of wearer's action recognition in first person a.k.a. egocentric videos. This problem is more challenging than third person activity recognition due to unavailability of wearer's pose and sharp movements in the videos caused by the natural head motion of the wearer. Carefully crafted features based on hands and objects cues for the problem have been shown to be successful for limited targeted datasets. We propose

207

convolutional neural networks (CNNs) for end to end learning and classification of wearer's actions. The proposed network makes use of egocentric cues by capturing hand pose, head motion and saliency map. It is compact. It can also be trained from relatively small number of labeled egocentric videos that are available. We show that the proposed network can generalize and give state of the art performance on various disparate egocentric action datasets.

1.448 Learning to Localize Little Landmarks

[Link to Paper]

We interact everyday with tiny objects such as the door handle of a car or the light switch in a room. These little landmarks are barely visible and hard to localize in images. We describe a method to find such landmarks by finding a sequence of latent landmarks, each with a prediction model. Each latent landmark predicts the next in sequence, and the last localizes the target landmark. For example, to find the door handle of a car, our method learns to start with a latent landmark near the wheel, as it is globally distinctive; subsequent latent landmarks use the context from the earlier ones to get closer to the target. Our method is supervised solely by the location of the little landmark and displays strong performance on more difficult variants of established tasks and on two new tasks.

1.449 Track and Transfer: Watching Videos to Simulate Strong Human Supervision for Weakly-Supervised Object Detection

[Link to Paper]

The status quo approach to training object detectors requires expensive bounding box annotations. Our framework takes a markedly different direction: we transfer tracked object boxes from weakly-labeled videos to weakly-labeled images to automatically generate pseudo ground-truth boxes, which replace manually annotated bounding boxes. We first mine discriminative regions in the weakly-labeled image collection that frequently/rarely appear in the positive/negative images. We then match those regions to videos and retrieve the corresponding tracked object boxes. Finally, we design a hough transform algorithm to vote for the best box to serve as the pseudo GT for each image, and use them to train an object detector. Together, these lead to state-of-the-art weakly-supervised detection results on the PASCAL 2007 and 2010 datasets.

1.450 DeepHand: Robust Hand Pose Estimation by Completing a Matrix Imputed with Deep Features

[Link to Paper]

We propose DeepHand to estimate the 3D pose of a hand using depth data from commercial 3D sensors. We discriminatively train convolutional neural networks to output a low dimensional activation feature given a depth map. This activation feature vector is representative of the global or local joint angle parameters of a hand pose. We efficiently identify 'spatial' nearest neighbors to the activation feature, from a database of features corresponding to synthetic depth maps, and store some 'temporal' neighbors from previous frames. Our matrix completion algorithm uses these 'spatio-temporal' activation features and the corresponding known pose parameter values to estimate the unknown pose parameters of the input feature vector. Our database of activation features supplements large viewpoint coverage and our hierarchical estimation of pose parameters is robust to occlusions. We show that our approach compares favorably to state-of-the-art methods while achieving real time performance (32 FPS) on a standard computer.

1.451 Gradual DropIn of Layers to Train Very Deep Neural Networks

[Link to Paper]

We introduce the concept of dynamically growing a neural network during training. In particular, an untrainable deep network starts as a trainable shallow network and newly added layers are slowly, organically added during training, thereby increasing the network's depth. This is accomplished by a new layer, which we call DropIn. The DropIn layer starts by passing the output from a previous layer (effectively skipping over the newly added layers), then increasingly including units from the new layers for both feedforward and backpropagation. We show that deep networks, which are untrainable with conventional methods, will converge with DropIn layers interspersed in the architecture. In addition, we demonstrate that DropIn provides regularization during training in an analogous way as dropout. Experiments are described with the MNIST dataset and various expanded LeNet architectures, CIFAR10 dataset with its architecture expanded from 3 to 11 layers, and on the ImageNet dataset with the AlexNet architecture expanded to 13 layers and the VGG 16-layer architecture.

1.452 BoxCars: 3D Boxes as CNN Input for Improved Fine-Grained Vehicle Recognition

[Link to Paper]

We are dealing with the problem of fine-grained vehicle make & model recognition and verification. Our contribution is showing that extracting additional data from the video stream besides the vehicle image itself and feeding it into the deep convolutional neural network boosts the recognition performance considerably. This additional information includes: 3D vehicle bounding box used for "unpacking" the vehicle image, its rasterized low-resolution shape, and information about the 3D vehicle orientation. Experiments show that adding such information decreases classification error by 26 % (the accuracy is improved from 0.772 to 0.832) and boosts verification average precision by 208 % (0.378 to 0.785) compared to baseline pure CNN without any input modifications. Also, the pure baseline CNN outperforms the recent state of the art solution by 0.081. We provide an annotated set "BoxCars" of surveillance vehicle images augmented by various automatically extracted auxiliary information. Our approach and the dataset can considerably improve the performance of traffic surveillance systems.

1.453 Solving Small-piece Jigsaw Puzzles by Growing Consensus

[Link to Paper]

In this paper, we present a novel computational puzzle solver for square-piece image jigsaw puzzles with no prior information such as piece orientation, anchor pieces or resulting dimension of the puzzle. By "piece" we mean a square dxd block of pixels, where we investigate pieces as small as 7x7 pixels. To reconstruct such challenging puzzles, we aim to search for piece configurations which maximize the size of consensus (i.e. grid or loop) configurations which represent a geometric consensus or agreement among pieces. Pieces are considered for addition to the existing assemblies if these pieces increase the size of the consensus configurations. In contrast to previous puzzle solvers which goal for assemblies maximizing compatibility measures between all pairs of pieces and thus depend heavily on the pairwise compatibility measure used, our new approach reduces the dependency on the pairwise compatibility measures which become increasingly uninformative at small scales and instead exploits geometric agreement among pieces. Our contribution also includes an improved pairwise compatibility measure which exploits directional derivative information along adjoining boundaries of the pieces. For the challenging unknown orientation piece puzzles where the size of pieces is small, we reduce assembly error by up to 75% compared with previous algorithms for standard datasets.

1.454 Deep Metric Learning via Lifted Structured Feature Embedding

[Link to Paper]

Learning the distance metric between pairs of examples is of great importance for learning and visual recognition. With the remarkable success from the state of the art convolutional neural networks, recent works [1, 31] have shown promising results on discriminatively training the networks to learn semantic feature embeddings where similar examples are mapped close to each other and dissimilar examples are mapped farther apart. In this paper, we describe an algorithm for taking full advantage of the training batches in the neural network training by lifting the vector of pairwise distances within the batch to the matrix of pairwise distances. This step enables the algorithm to learn the state of the art feature embedding by optimizing a novel structured prediction objective on the lifted problem. Additionally, we collected Stanford Online Products dataset: 120k images of 23k classes of online products for metric learning. Our experiments on the CUB-200-2011 [37], CARS196 [19], and Stanford Online Products datasets demonstrate significant improvement over existing deep feature embedding methods on all experimented embedding sizes with the GoogLeNet [33] network. The source code and the dataset are available at: https://github.com/rksltnl/ Deep-Metric-Learning-CVPR16.

1.455 Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images

[Link to Paper]

We focus on the task of amodal 3D object detection in RGB-D images, which aims to produce a 3D bounding box of an object in metric form at its full extent. We introduce Deep Sliding Shapes, a 3D ConvNet formulation that takes a 3D volumetric scene from a RGB-D image as input and outputs 3D object bounding boxes. In our approach, we propose the first 3D Region Proposal Network (RPN) to learn objectness from geometric shapes and the first joint Object Recognition Network (ORN) to extract geometric features in 3D and color features in 2D. In particular, we handle objects of various sizes by training an amodal RPN at two different scales and an ORN to regress 3D bounding boxes. Experiments show that our algorithm outperforms the state-of-the-art by 13.8 in mAP and is 200 faster than the original Sliding Shapes.

1.456 Predicting the Where and What of actors and actions through Online Action Localization

[Link to Paper]

This paper proposes a novel approach to tackle the challenging problem of 'online action localization' which entails predicting actions and their locations as they happen in a video. Typically, action localization or recognition is performed in an offline manner where all

the frames in the video are processed together and action labels are not predicted for the future. This disallows timely localization of actions - an important consideration for surveillance tasks. In our approach, given a batch of frames from the immediate past in a video, we estimate pose and oversegment the current frame into superpixels. Next, we discriminatively train an actor foreground model on the superpixels using the pose bounding boxes. A Conditional Random Field with superpixels as nodes, and edges connecting spatio-temporal neighbors is used to obtain action segments. The action confidence is predicted using dynamic programming on SVM scores obtained on short segments of the video, thereby capturing sequential information of the actions. The issue of visual drift is handled by updating the appearance model and pose refinement in an online manner. Lastly, we introduce a new measure to quantify the performance of action prediction (i.e. online action localization), which analyzes how the prediction accuracy varies as a function of observed portion of the video. Our experiments suggest that despite using only a few frames to localize actions at each time instant, we are able to predict the action and obtain competitive results to state-of-the-art offline methods.

1.457 Scene Labeling Using Sparse Precision Matrix

[Link to Paper]

Scene labeling task is to segment the image into meaningful regions and categorize them into classes of objects which comprised the image. Commonly used methods typically find the local features for each segment and label them using classifiers. Afterwards, labeling is smoothed in order to make sure that neighboring regions receive similar labels. However, these methods ignore expressive connections between labels and non-local dependencies among regions. In this paper, we propose to use a sparse estimation of precision matrix (also called concentration matrix), which is the inverse of covariance matrix of data obtained by graphical lasso to find interaction between labels and regions. To do this, we formulate the problem as an energy minimization over a graph, whose structure is captured by applying sparse constraint on the elements of the precision matrix. This graph encodes (or represents) only significant interactions and avoids a fully connected graph, which is typically used to reflect the long distance associations. We use local and global information to achieve better labeling. We assess our approach on three datasets and obtained promising results.

1.458 End-to-end people detection in crowded scenes

[Link to Paper]

Current people detectors operate either by scanning an image in a sliding window fashion or by classifying a discrete set of proposals. We propose a model that is based on decoding an image into a set of people detections. Our system takes an image as input and directly outputs a set of distinct detection hypotheses. Because we generate predictions jointly, common post-processing steps such as nonmaximum suppression are unnecessary. We use a recurrent LSTM layer for sequence generation and train our model end-to-end with a new loss function that operates on sets of detections. We demonstrate the effectiveness of our approach on the challenging task of detecting people in crowded scenes1.

1.459 Robust Visual Place Recognition with Graph Kernels

[Link to Paper]

A novel method for visual place recognition is introduced and evaluated, demonstrating robustness to perceptual aliasing and observation noise. This is achieved by increasing discrimination through a more structured representation of visual observations. Estimation of observation likelihoods are based on graph kernel formulations, utilizing both the structural and visual information encoded in covisibility graphs. The proposed probabilistic model is able to circumvent the typically difficult and expensive posterior normalization procedure by exploiting the information available in visual observations. Furthermore, the place recognition complexity is independent of the size of the map. Results show improvements over the state-of-theart on a diverse set of both public datasets and novel experiments, highlighting the benefit of the approach.

1.460 Material Classification Using Raw Time-of-Flight Measurements

[Link to Paper]

We propose a material classification method using raw time-of-flight (ToF) measurements. ToF cameras capture the correlation between a reference signal and the temporal response of material to incident illumination. Such measurements encode unique signatures of the material, i.e. the degree of subsurface scattering inside a volume. Subsequently, it offers an orthogonal domain of feature representation compared to conventional spatial and angular reflectance-based approaches. We demonstrate the effectiveness, robustness, and efficiency of our method through experiments and comparisons of real-world materials.

1.461 What if we do not have multiple videos of the same action? – Video Action Localization Using Web Images

[Link to Paper]

This paper tackles the problem of spatio-temporal action localization in a video, without assuming the availability of multiple videos or any prior annotations. Action is localized by employing images downloaded from internet using action name. Given web images, we first dampen image noise using random walk and evade distracting backgrounds within images using image action proposals. Then, given a video, we generate multiple spatio-temporal action proposals. We suppress camera and background generated proposals by exploiting optical flow gradients within proposals. To obtain the most action representative proposals, we propose to reconstruct action proposals in the video by leveraging the action proposals in images. Moreover, we preserve the temporal smoothness of the video and reconstruct all proposal bounding boxes jointly using the constraints that push the coefficients for each bounding box toward a common consensus, thus enforcing the coefficient similarity across multiple frames. We solve this optimization problem using variant of two-metric projection algorithm. Finally, the video proposal that has the lowest reconstruction cost and is motion salient is used to localize the action. Our method is not only applicable to the trimmed videos, but it can also be used for action localization in untrimmed videos, which is a very challenging problem. We present extensive experiments on trimmed as well as untrimmed datasets to validate the effectiveness of the proposed approach.

1.462 ProNet: Learning to Propose Object-specific Boxes for Cascaded Neural Networks

[Link to Paper]

This paper aims to classify and locate objects accurately and efficiently, without using bounding box annotations. It is challenging as objects in the wild could appear at arbitrary locations and in different scales. In this paper, we propose a novel classification architecture ProNet based on convolutional neural networks. It uses computationally efficient neural networks to propose image regions that are likely to contain objects, and applies more powerful but slower networks on the proposed regions. The basic building block is a multi-scale fully-convolutional network which assigns object confidence scores to boxes at different locations and scales. We show that such networks can be trained effectively using image-level annotations, and can be connected into cascades or trees for efficient object classification. ProNet outperforms previous state-of-the-art significantly on PASCAL VOC 2012 and MS COCO datasets for object classification and point-based localization.

1.463 Sparsifying Neural Network Connections for Face Recognition

[Link to Paper]

This paper proposes to learn high-performance deep ConvNets with sparse neural connections, referred to as sparse ConvNets, for face recognition. The sparse ConvNets are learned in an iterative way, each time one additional layer is sparsified and the entire model is retrained given the initial weights learned in previous iterations. One important finding is that directly training the sparse ConvNet from scratch failed to find good solutions for face recognition, while using a previously learned denser model to properly initialize a sparser model is critical to continue learning effective features for face recognition. This paper also proposes a new neural correlation-based weight selection criterion and empirically verifies its effectiveness in selecting informative connections from previously learned models in each iteration. When taking a moderately sparse structure (26%-76% of weights in the dense model), the proposed sparse ConvNet model significantly improves the face recognition performance of the previous state-of-theart DeepID2+ models given the same training data, while it keeps the performance of the baseline model with only 12% of the original parameters.

1.464 Rethinking the Inception Architecture for Computer Vision

[Link to Paper]

Convolutional networks are at the core of most stateof-the-art computer vision solutions for a wide variety of tasks. Since 2014 very deep convolutional networks started to become mainstream, yielding substantial gains in various benchmarks. Although increased model size and computational cost tend to translate to immediate quality gains for most tasks (as long as enough labeled data is provided for training), computational efficiency and low parameter count are still enabling factors for various use cases such as mobile vision and big-data scenarios. Here we are exploring ways to scale up networks in ways that aim at utilizing the added computation as efficiently as possible by suitably factorized convolutions and aggressive regularization. We benchmark our methods on the ILSVRC 2012 classification challenge validation set demonstrate substantial gains over the state of the art: 21.2% top-1 and 5.6% top-5 error for single frame evaluation using a network with a computational cost of 5 billion multiply-adds per inference and with using less than 25 million parameters. With an ensemble of 4 models and multi-crop evaluation, we report 3.5% top-5 error and 17.3% top-1 error on the validation set and 3.6% top-5 error on the official test set.

1.465 Using Spatial Order to Boost the Elimination of Incorrect Feature Matches

[Link to Paper]

Correctly matching feature points in a pair of images is an important preprocessing step for many computer vision applications. In this paper we propose an efficient method for estimating the number of correct matches without explicitly computing them. In addition, our method estimates the region of overlap between the images. To this end, we propose to analyze the set of matches using the spatial order of the features, as projected to the x-axis of the image. The set of features in each image is thus represented by a sequence. This reduces the analysis of the matching problem to the analysis of the permutation between the sequences. Using the Kendall distance metric between permutations and natural assumptions on the distribution of the correct and incorrect matches, we show how to estimate the abovementioned values. We demonstrate the usefulness of our method in two applications: (i) a new halting condition for RANSAC based epipolar geometry estimation methods that considerably reduce the running time, and (ii) discarding spatially unrelated image pairs in the Structure-fromMotion pipeline. Furthermore, our analysis can be used to compute the probability that a given match is correct based on the estimated number of correct matches and the rank of the features within the sequences. Our experiments on a large number of synthetic and real data demonstrate the effectiveness of our method. For example, the running time of the image matching stage in the Structure-from-Motion pipeline may be reduced by about 99% while preserving about 80% of the correctly matched feature points.

1.466 Fits Like a Glove: Rapid and Reliable Hand Shape Personalization

[Link to Paper]

We present a fast, practical method for personalizing a hand shape basis to an individual user's detailed hand shape using only a small set of depth images. To achieve this, we minimize an energy based on a sum of render-and-compare cost functions called the golden energy. However, this energy is only piecewise continuous, due to pixels crossing occlusion boundaries, and is therefore not obviously amenable to efficient gradient-based optimization. A key insight is that the energy is the combination of a smooth low-frequency function with a high-frequency, low-amplitude, piecewisecontinuous function. A central finite difference approximation with a suitable step size can therefore jump over the discontinuities to obtain a good approximation to the energy's low-frequency behavior, allowing efficient gradient-based optimization. Experimental results quantitatively demonstrate for the first time that detailed personalized models improve the accuracy of hand tracking and achieve competitive results in both tracking and model registration. Shape Space Model Fit to Data Frame 1 Frame 2 Frame F Initialization Optimized Model near depth data Model in front

216

of depth data Model behind depth data Figure 1: We show how to fit a deformable hand shape basis model [17] to a small set of depth images. Our method jointly optimizes over the shape RK and F poses f to maximize the model's alignment to the data in F depth images. The initial hand poses are automatically determined by a hand tracker that uses the mean shape mean, but there is clearly poor alignment between model and data. After our optimization to obtain personalized shape personalized, the alignment is much better, with remaining errors largely due to sensor noise.

1.467 Proximal Riemannian Pursuit for Large-scale Trace-norm Minimization

[Link to Paper]

Trace-norm regularization plays an important role in many areas such as computer vision and machine learning. When solving general large-scale trace-norm regularized problems, existing methods may be computationally expensive due to many high-dimensional truncated singular value decompositions (SVDs) or the unawareness of matrix ranks. In this paper, we propose a proximal Riemannian pursuit (PRP) paradigm which addresses a sequence of trace-norm regularized subproblems defined on nonlinear matrix varieties. To address the subproblem, we extend the proximal gradient method on vector space to nonlinear matrix varieties, in which the SVDs of intermediate solutions are maintained by cheap low-rank QR decompositions, therefore making the proposed method more scalable. Empirical studies on several tasks, such as matrix completion and low-rank representation based subspace clustering, demonstrate the competitive performance of the proposed paradigms over existing methods.

1.468 Recovering Transparent Shape from Time-of-Flight Distortion

[Link to Paper]

This paper presents a method for recovering shape and normal of a transparent object from a single viewpoint using a Time-of-Flight (ToF) camera. Our method is built upon the fact that the speed of light varies with the refractive index of the medium and therefore the depth measurement of a transparent object with a ToF camera may be distorted. We show that, from this ToF distortion, the refractive light path can be uniquely determined by estimating a single parameter. We estimate this parameter by introducing a surface normal consistency between the one determined by a light path candidate and the other computed from the corresponding shape. The proposed method is evaluated by both simulation and real-world experiments and shows faithful transparent shape recovery.

1.469 Large Scale Semi-supervised Object Detection using Visual and Semantic Knowledge Transfer

[Link to Paper]

Deep CNN-based object detection systems have achieved remarkable success on several large-scale object detection benchmarks. However, training such detectors requires a large number of labeled bounding boxes, which are more difficult to obtain than image-level annotations. Previous work addresses this issue by transforming image-level classifiers into object detectors. This is done by modeling the differences between the two on categories with both imagelevel and bounding box annotations, and transferring this information to convert classifiers to detectors for categories without bounding box annotations. We improve this previous work by incorporating knowledge about object similarities from visual and semantic domains during the transfer process. The intuition behind our proposed method is that visually and semantically similar categories should exhibit more common transferable properties than dissimilar categories, e.g. a better detector would result by transforming the differences between a dog classifier and a dog detector onto the cat class, than would by transforming from the violin class. Experimental results on the challenging ILSVRC2013 detection dataset demonstrate that each of our proposed object similarity based knowledge transfer methods outperforms the baseline methods. We found strong evidence that visual similarity and semantic relatedness are complementary for the task, and when combined notably improve detection, achieving state-of-the-art detection performance in a semi-supervised setting.

1.470 Joint Recovery of Dense Correspondence and Cosegmentation in Two Images

[Link to Paper]

We propose a new technique to jointly recover cosegmentation and dense per-pixel correspondence in two images. Our method parameterizes the correspondence field using piecewise similarity transformations and recovers a mapping between the estimated common "foreground" regions in the two images allowing them to be precisely aligned. Our formulation is based on a hierarchical Markov random field model with segmentation and transformation labels. The hierarchical structure uses nested image regions to constrain inference across multiple scales. Unlike prior hierarchical methods which assume that the structure is given, our proposed iterative technique dynamically recovers the structure along with the labeling. This joint inference is performed in an energy minimization framework using iterated graph cuts. We evaluate our method on a new dataset of 400 image pairs with manually obtained ground truth, where it outperforms state-of-the-art methods designed specifically for either cosegmentation or correspondence estimation.

1.471 Siamese Instance Search for Tracking

[Link to Paper]

In this paper we present a tracker, which is radically different from state-of-the-art trackers: we apply no model updating, no occlusion detection, no combination of trackers, no geometric matching, and still deliver state-of-theart tracking performance, as demonstrated on the popular online tracking benchmark (OTB) and six very challenging YouTube videos. The presented tracker simply matches the initial patch of the target in the first frame with candidates in a new frame and returns the most similar patch by a learned matching function. The strength of the matching function comes from being extensively trained generically, i.e., without any data of the target, using a Siamese deep neural network, which we design for tracking. Once learned, the matching function is used as is, without any adapting, to track previously unseen targets. It turns out that the learned matching function is so powerful that a simple tracker built upon it, coined Siamese INstance search Tracker, SINT, which only uses the original observation of the target from the first frame, suffices to reach state-of-theart performance. Further, we show the proposed tracker even allows for target re-identification after the target was absent for a complete video shot.

1.472 MovieQA: Understanding Stories in Movies through Question-Answering

[Link to Paper]

We introduce the MovieQA dataset which aims to evaluate automatic story comprehension from both video and text. The dataset consists of 14,944 questions about 408 movies with high semantic diversity. The questions range from simpler "Who" did "What" to "Whom", to "Why" and "How" certain events occurred. Each question comes with a set of five possible answers; a correct one and four deceiving answers provided by human annotators. Our dataset is unique in that it contains multiple sources of information video clips, plots, subtitles, scripts, and DVS [32]. We analyze our data through various statistics and methods. We further extend existing QA techniques to show that question-answering with such open-ended semantics is hard. We make this data set public along with an evaluation benchmark to encourage inspiring work in this challenging domain.

1.473 Direct Prediction of 3D Body Poses from Motion Compensated Sequences

[Link to Paper]

We propose an efficient approach to exploiting motion information from consecutive frames of a video sequence to recover the 3D pose of people. Previous approaches typically compute candidate poses in individual frames and then link them in a post-processing step to resolve ambiguities. By contrast, we directly regress from a spatio-temporal volume of bounding boxes to a 3D pose in the central frame. We further show that, for this approach to achieve its full potential, it is essential to compensate for the motion in consecutive frames so that the subject remains centered. This then allows us to effectively overcome ambiguities and improve upon the state-of-the-art by a large margin on the Human3.6m, HumanEva, and KTH Multiview Football 3D human pose estimation benchmarks.

1.474   Face2Face: Real-time Face Capture and Reenactment of RGB Videos

[Link to Paper]

We present a novel approach for real-time facial reenactment of a monocular target video sequence (e.g., Youtube video). The source sequence is also a monocular video stream, captured live with a commodity webcam. Our goal is to animate the facial expressions of the target video by a source actor and re-render the manipulated output video in a photo-realistic fashion. To this end, we first address the under-constrained problem of facial identity recovery from monocular video by non-rigid model-based bundling. At run time, we track facial expressions of both source and target video using a dense photometric consistency measure. Reenactment is then achieved by fast and efficient deformation transfer between source and target. The mouth interior that best matches the re-targeted expression is retrieved from the target sequence and warped to produce an accurate fit. Finally, we convincingly re-render the synthesized target face on top of the corresponding video stream such that it seamlessly blends with the real-world illumination. We demonstrate our method in a live setup, where Youtube videos are reenacted in real time.

1.475   Augmented Blendshapes for Real-time Simultaneous 3D Head Modeling and Facial Motion Capture

[Link to Paper]

We propose a method to build in real-time animated 3D head models using a consumer-grade RGB-D camera. Our framework is the first one to provide simultaneously comprehensive facial motion tracking and a detailed 3D model of the user's head. Anyone's head can be instantly reconstructed and his facial motion captured without requiring any training or pre-scanning. The user starts facing the camera with a neutral expression in the first frame, but is free to move, talk and change his face expression as he wills otherwise. The facial motion is tracked using a blendshape representation while the fine geometric

details are captured using a Bump image mapped over the template mesh. We propose an efficient algorithm to grow and refine the 3D model of the head on-the-fly and in real-time. We demonstrate robust and high-fidelity simultaneous facial motion tracking and 3D head modeling results on a wide range of subjects with various head poses and facial expressions. Our proposed method offers interesting possibilities for animation production and 3D video telecommunications.

1.476 Seeing Behind the Camera: Identifying the Authorship of a Photograph

[Link to Paper]

We introduce the novel problem of identifying the photographer behind a photograph. To explore the feasibility of current computer vision techniques to address this problem, we created a new dataset of over 180,000 images taken by 41 well-known photographers. Using this dataset, we examined the effectiveness of a variety of features (low and high-level, including CNN features) at identifying the photographer. We also trained a new deep convolutional neural network for this task. Our results show that high-level features greatly outperform low-level features. We provide qualitative results using these learned models that give insight into our method's ability to distinguish between photographers, and allow us to draw interesting conclusions about what specific photographers shoot. We also demonstrate two applications of our method.

1.477 Seven ways to improve example-based single image super resolution

[Link to Paper]

In this paper we present seven techniques that everybody should know to improve example-based single image super resolution (SR): 1) augmentation of data, 2) use of large dictionaries with efficient search structures, 3) cascading, 4) image self-similarities, 5) back projection refinement, 6) enhanced prediction by consistency check, and 7) context reasoning. We validate our seven techniques on standard SR benchmarks (i.e. Set5, Set14, B100) and methods (i.e. A+, SRCNN, ANR, Zeyde, Yang) and achieve substantial improvements. The techniques are widely applicable and require no changes or only minor adjustments of the SR methods. Moreover, our Improved A+ (IA) method sets new stateof-the-art results outperforming A+ by up to 0.9dB on average PSNR whilst maintaining a low time complexity. +0.6 +0.3 +0.9dB +0.8 +0.7 Improved methods +0.8 Figure 1. We largely improve (red) over the original examplebased single image super-resolution methods (blue), i.e. our IA method is 0.9dB better than A+ [31] and 2dB better than Yang et al. [40]. Results reported on Set5, 3. Details in Section 4.1.

1.478 Consistency of silhouettes and their duals

[Link to Paper]

Silhouettes provide rich information on threedimensional shape, since the intersection of the associated visual cones generates the "visual hull", which encloses and approximates the original shape. However, not all silhouettes can actually be projections of the same object in space: this simple observation has implications in object recognition and multi-view segmentation, and has been (often implicitly) used as a basis for camera calibration. In this paper, we investigate the conditions for multiple silhouettes, or more generally arbitrary closed image sets, to be geometrically "consistent". We present this notion as a natural generalization of traditional multi-view geometry, which deals with consistency for points. After discussing some general results, we present a "dual" formulation for consistency, that gives conditions for a family of planar sets to be sections of the same object. Finally, we introduce a more general notion of silhouette "compatibility" under partial knowledge of the camera projections, and point out some possible directions for future research.

1.479 Aggregating Image and Text Quantized Correlated Components

[Link to Paper]

Cross-modal tasks occur naturally for multimedia content that can be described along two or more modalities like visual content and text. Such tasks require to "translate" information from one modality to another. Methods like kernelized canonical correlation analysis (KCCA) attempt to solve such tasks by finding aligned subspaces in the description spaces of different modalities. Since they favor correlations against modality-specific information, these methods have shown some success in both cross-modal and bi-modal tasks. However, we show that a direct use of the subspace alignment obtained by KCCA only leads to coarse translation abilities. To address this problem, we first put forward a new representation method that aggregates information provided by the projections of both modalities on their aligned subspaces. We further suggest a method relying on neighborhoods in these subspaces to complete uni-modal information. Our proposal exhibits state-of-the-art results for bi-modal classification on Pascal VOC07 and improves it by over 60% for cross-modal retrieval on FlickR 8K/30K. image feature space document img txt KCCA text feature space space of correlated components, quantized Figure 1. Visual and textual contents of a document are projected onto a common space that has been previously quantized. Both points, corresponding to the same document, are encoded according to a common vocabulary before their aggregation.

1.480 Deep Canonical Time Warping

[Link to Paper]

Machine learning algorithms for the analysis of timeseries often depend on the assumption that the utilised data are temporally aligned. Any temporal discrepancies arising in the data is certain to lead to ill-generalisable models, which in turn fail to correctly capture the properties of the task at hand. The temporal alignment of time-series is thus a crucial challenge manifesting in a multitude of applications. Nevertheless, the vast majority of algorithms oriented towards the temporal alignment of time-series are applied directly on the observation space, or utilise simple linear projections. Thus, they fail to capture complex, hierarchical non-linear representations which may prove to be beneficial towards the task of temporal alignment, particularly when dealing with multi-modal data (e.g., aligning visual and acoustic information). To this end, we present the Deep Canonical Time Warping (DCTW), a method which automatically learns complex non-linear representations of multiple time-series, generated such that (i) they are highly correlated, and (ii) temporally in alignment. By means of experiments on four real datasets, we show that the representations learnt via the proposed DCTW significantly outperform state-of-the-art methods in temporal alignment, elegantly handling scenarios with highly heterogeneous features, such as the temporal alignment of acoustic and visual features.

1.481 Mnemonic Descent Method: A recurrent process applied for end-to-end face alignment

[Link to Paper]

Cascaded regression has recently become the method of choice for solving non-linear least squares problems such as deformable image alignment. Given a sizeable training set, cascaded regression learns a set of generic rules that are sequentially applied to minimise the least squares problem. Despite the success of cascaded regression for problems such as face alignment and head pose estimation, there are several shortcomings arising in the strategies proposed thus far. Specifically, (a) the regressors are learnt independently, (b) the descent directions may cancel one another out and (c) handcrafted features (e.g., HoGs, SIFT etc.) are mainly used to drive the cascade, which may be sub-optimal for the task at hand. In this paper, we propose a combined and jointly trained convolutional recurrent neural network architecture that allows the training of an end-to-end to system that attempts to alleviate the aforementioned drawbacks. The recurrent module facilitates the joint optimisation of the regressors by assuming the cascades form a nonlinear dynamical system, in effect fully utilising the information between all cascade levels by introducing a memory unit that shares information across all levels. The convolutional module allows

the network to extract features that are specialised for the task at hand and are experimentally shown to outperform hand-crafted features. We show that the application of the proposed architecture for the problem of face alignment results in a strong improvement over the current state-of-the-art.

1.482 Learning Cross-Domain Landmarks for Heterogeneous Domain Adaptation

[Link to Paper]

While domain adaptation (DA) aims to associate the learning tasks across data domains, heterogeneous domain adaptation (HDA) particularly deals with learning from cross-domain data which are of different types of features. In other words, for HDA, data from source and target domains are observed in separate feature spaces and thus exhibit distinct distributions. In this paper, we propose a novel learning algorithm of Cross-Domain Landmark Selection (CDLS) for solving the above task. With the goal of deriving a domain-invariant feature subspace for HDA, our CDLS is able to identify representative cross-domain data, including the unlabeled ones in the target domain, for performing adaptation. In addition, the adaptation capabilities of such cross-domain landmarks can be determined accordingly. This is the reason why our CDLS is able to achieve promising HDA performance when comparing to state-of-the-art HDA methods. We conduct classification experiments using data across different features, domains, and modalities. The effectiveness of our proposed method can be successfully verified.

1.483 Video Segmentation via Object Flow

[Link to Paper]

Video object segmentation is challenging due to fast moving objects, deforming shapes, and cluttered backgrounds. Optical flow can be used to propagate an object segmentation over time but, unfortunately, flow is often inaccurate, particularly around object boundaries. Such boundaries are precisely where we want our segmentation to be accurate. To obtain accurate segmentation across time, we propose an efficient algorithm that considers video segmentation and optical flow estimation simultaneously. For video segmentation, we formulate a principled, multiscale, spatio-temporal objective function that uses optical flow to propagate information between frames. For optical flow estimation, particularly at object boundaries, we compute the flow independently in the segmented regions and recompose the results. We call the process object flow and demonstrate the effectiveness of jointly optimizing optical flow and video segmentation using an iterative scheme. Exper-

iments on the SegTrack v2 and YoutubeObjects datasets show that the proposed algorithm performs favorably against the other state-of-the-art methods.

1.484 Real-Time Salient Object Detection with a Minimum Spanning Tree

[Link to Paper]

In this paper, we present a real-time salient object detection system based on the minimum spanning tree. Due to the fact that background regions are typically connected to the image boundaries, salient objects can be extracted by computing the distances to the boundaries. However, measuring the image boundary connectivity efficiently is a challenging problem. Existing methods either rely on superpixel representation to reduce the processing units or approximate the distance transform. Instead, we propose an exact and iteration free solution on a minimum spanning tree. The minimum spanning tree representation of an image inherently reveals the object geometry information in a scene. Meanwhile, it largely reduces the search space of shortest paths, resulting an efficient and high quality distance transform algorithm. We further introduce a boundary dissimilarity measure to compliment the shortage of distance transform for salient object detection. Extensive evaluations show that the proposed algorithm achieves the leading performance compared to the state-of-the-art methods in terms of efficiency and accuracy.

1.485 Self-Adaptive Matrix Completion for Heart Rate Estimation from Face Videos under Realistic Conditions

[Link to Paper]

Recent studies in computer vision have shown that, while practically invisible to a human observer, skin color changes due to blood flow can be captured on face videos and, surprisingly, be used to estimate the heart rate (HR). While considerable progress has been made in the last few years, still many issues remain open. In particular, stateof-the-art approaches are not robust enough to operate in natural conditions (e.g. in case of spontaneous movements, facial expressions, or illumination changes). Opposite to previous approaches that estimate the HR by processing all the skin pixels inside a fixed region of interest, we introduce a strategy to dynamically select face regions useful for robust HR estimation. Our approach, inspired by recent advances on matrix completion theory, allows us to predict the HR while simultaneously discover the best regions of the face to be used for estimation. Thorough experimental evaluation conducted on public benchmarks suggests that the proposed approach significantly outperforms state-ofthe-art HR estimation methods in naturalistic conditions.

1.486 Patches, Planes and Probabilities: A Non-local Prior for Volumetric 3D Reconstruction

[Link to Paper]

In this paper, we propose a non-local structured prior for volumetric multi-view 3D reconstruction. Towards this goal, we present a novel Markov random field model based on ray potentials in which assumptions about large 3D surface patches such as planarity or Manhattan world constraints can be efficiently encoded as probabilistic priors. We further derive an inference algorithm that reasons jointly about voxels, pixels and image segments, and estimates marginal distributions of appearance, occupancy, depth, normals and planarity. Key to tractable inference is a novel hybrid representation that spans both voxel and pixel space and that integrates non-local information from 2D image segmentations in a principled way. We compare our non-local prior to commonly employed local smoothness assumptions and a variety of state-of-the-art volumetric reconstruction baselines on challenging outdoor scenes with textureless and reflective surfaces. Our experiments indicate that regularizing over larger distances has the potential to resolve ambiguities where local regularizers fail.

1.487 Deep Gaussian Conditional Random Field Network: A Model-based Deep Network for Discriminative Denoising

[Link to Paper]

We propose a novel end-to-end trainable deep network architecture for image denoising based on a Gaussian Conditional Random Field (GCRF) model. In contrast to the existing discriminative denoising methods that train a separate model for each individual noise level, the proposed deep network explicitly models the input noise variance and hence is capable of handling a range of noise levels. Our deep network, which we refer to as deep GCRF network, consists of two sub-networks: (i) a parameter generation network that generates the pairwise potential parameters based on the noisy input image, and (ii) an inference network whose layers perform the computations involved in an iterative GCRF inference procedure. We train two deep GCRF networks (each network operates over a range of noise levels: one for low input noise levels and one for high input noise levels) discriminatively by maximizing the peak signal-to-noise ratio measure. Experiments on Berkeley segmentation and PASCALVOC datasets show that the proposed approach produces results on par with the stateof-the-art without training a separate network for each individual noise level.

1.488 Gaussian Conditional Random Field Network for Semantic Segmentation

[Link to Paper]

In contrast to the existing approaches that use discrete Conditional Random Field (CRF) models, we propose to use a Gaussian CRF model for the task of semantic segmentation. We propose a novel deep network, which we refer to as Gaussian Mean Field (GMF) network, whose layers perform mean field inference over a Gaussian CRF. The proposed GMF network has the desired property that each of its layers produces an output that is closer to the maximum a posteriori solution of the Gaussian CRF compared to its input. By combining the proposed GMF network with deep Convolutional Neural Networks (CNNs), we propose a new end-to-end trainable Gaussian conditional random field network. The proposed Gaussian CRF network is composed of three sub-networks: (i) a CNN-based unary network for generating unary potentials, (ii) a CNN-based pairwise network for generating pairwise potentials, and (iii) a GMF network for performing Gaussian CRF inference. When trained end-to-end in a discriminative fashion, and evaluated on the challenging PASCALVOC 2012 segmentation dataset, the proposed Gaussian CRF network outperforms various recent semantic segmentation approaches that combine CNNs with discrete CRF models.

1.489 Rolling Rotations for Recognizing Human Actions from 3D Skeletal Data

[Link to Paper]

Recently, skeleton-based human action recognition has been receiving significant attention from various research communities due to the availability of depth sensors and real-time depth-based 3D skeleton estimation algorithms. In this work, we use rolling maps for recognizing human actions from 3D skeletal data. The rolling map is a welldefined mathematical concept that has not been explored much by the vision community. First, we represent each skeleton using the relative 3D rotations between various body parts. Since 3D rotations are members of the special orthogonal group SO3, our skeletal representation becomes a point in the Lie group SO3 . . . SO3, which is also a Riemannian manifold. Then, using this representation, we model human actions as curves in this Lie group. Since classification of curves in this non-Euclidean space is a difficult task, we unwrap the action curves onto the Lie algebra so3 . . . so3 (which is a vector space) by combining the logarithm map with rolling maps, and perform classification in the Lie algebra. Experimental results on three action datasets show that the proposed approach performs equally well or better when compared to state-of-the-art.

1.490 Piecewise-planar 3D approximation from wide-baseline stereo

[Link to Paper]

This paper approximates the 3D geometry of a scene by a small number of 3D planes. The method is especially suited to man-made scenes, and only requires two calibrated wide-baseline views as inputs. It relies on the computation of a dense but noisy 3D point cloud, as for example obtained by matching DAISY descriptors [35] between the views. It then segments one of the two reference images, and adopts a multi-model fitting process to assign a 3D plane to each region, when the region is not detected as occluded. A pool of 3D plane hypotheses is first derived from the 3D point cloud, to include planes that reasonably approximate the part of the 3D point cloud observed from each reference view between randomly selected triplets of 3D points. The hypothesis-to-region assignment problem is then formulated as an energy-minimization problem, which simultaneously optimizes an original data-fidelity term, the assignment smoothness over neighboring regions, and the number of assigned planar proxies. The synthesis of intermediate viewpoints demonstrates the effectiveness of our 3D reconstruction, and thereby the relevance of our proposed data fidelity-metric.

1.491 Noisy Label Recovery for Shadow Detection in Unfamiliar Domains

[Link to Paper]

Recent shadow detection algorithms have shown initial success on small datasets of images from specific domains. However, shadow detection on broader image domains is still challenging due to the lack of annotated training data. This is due to the intense manual labor in annotating shadow data. In this paper we propose "lazy annotation", an efficient annotation method where an annotator only needs to mark the important shadow areas and some non-shadow areas. This yields data with noisy labels that are not yet useful for training a shadow detector. We address the problem of label noise by jointly learning a shadow region classifier and recovering the labels in the training set. We consider the training labels as unknowns and formulate the label recovery problem as the minimization of the sum of squared leave-one-out errors of a Least Squares SVM, which can be efficiently optimized. Experimental results show that a classifier trained with recovered labels achieves comparable performance to a classifier trained on the properly annotated data. These results suggest a feasible approach to address the task of detecting shadows in an unfamiliar domain: collecting and lazily annotating some images from the new domain for training. As will be demonstrated, this approach outperforms methods that rely on precisely annotated but less relevant datasets. Initial results suggest more general applicability.

1.492 Spatiotemporal Bundle Adjustment for Dynamic 3D Reconstruction

Bundle adjustment jointly optimizes camera intrinsics and extrinsics and 3D point triangulation to reconstruct a static scene. The triangulation constraint however is invalid for moving points captured in multiple unsynchronized videos and bundle adjustment is not purposed to estimate the temporal alignment between cameras. In this paper, we present a spatiotemporal bundle adjustment approach that jointly optimizes four coupled subproblems: estimating camera intrinsics and extrinsics, triangulating 3D static points, as well as subframe temporal alignment between cameras and estimating 3D trajectories of dynamic points. Key to our joint optimization is the careful integration of physics-based motion priors within the reconstruction pipeline, validated on a large motion capture corpus. We present an end-to-end pipeline that takes multiple uncalibrated and unsynchronized video streams and produces a dynamic reconstruction of the event. Because the videos are aligned with sub-frame precision, we reconstruct 3D trajectories of unconstrained outdoor activities at much higher temporal resolution than the input videos.

1.493 Predicting When Saliency Maps are Accurate and Eye Fixations Consistent

Many computational models of visual attention use image features and machine learning techniques to predict eye fixation locations as saliency maps. Recently, the success of Deep Convolutional Neural Networks (DCNNs) for object recognition has opened a new avenue for computational models of visual attention due to the tight link between visual attention and object recognition. In this paper, we show that using features from DCNNs for object recognition we can make predictions that enrich the information provided by saliency models. Namely, we can estimate the reliability of a saliency model from the raw image, which serves as a meta-saliency measure that may be used to select the best saliency algorithm for an image. Analogously, the consistency of the eye fixations among subjects, i.e. the agreement between the eye fixation locations of different subjects, can also be predicted and used by a designer to assess whether subjects reach a consensus about salient image locations.

1.494 Anticipating Visual Representations from Unlabeled Video

Anticipating actions and objects before they start or appear is a difficult problem in computer vision with several real-world applications. This task is challenging partly because it requires leveraging extensive knowledge of the world that is difficult to write down. We believe that a promising resource for efficiently learning this knowledge is through readily available unlabeled video. We present a framework that capitalizes on temporal structure in unlabeled video to learn to anticipate human actions and objects. The key idea behind our approach is that we can train deep networks to predict the visual representation of images in the future. Visual representations are a promising prediction target because they encode images at a higher semantic level than pixels yet are automatic to compute. We then apply recognition algorithms on our predicted representation to anticipate objects and actions. We experimentally validate this idea on two datasets, anticipating actions one second in the future and objects five seconds in the future.

1.495 Predicting Motivations of Actions by Leveraging Text

[Link to Paper]

Understanding human actions is a key problem in computer vision. However, recognizing actions is only the first step of understanding what a person is doing. In this paper, we introduce the problem of predicting why a person has performed an action in images. This problem has many applications in human activity understanding, such as anticipating or explaining an action. To study this problem, we introduce a new dataset of people performing actions annotated with likely motivations. However, the information in an image alone may not be sufficient to automatically solve this task. Since humans can rely on their lifetime of experiences to infer motivation, we propose to give computer vision systems access to some of these experiences by using recently developed natural language models to mine knowledge stored in massive amounts of text. While we are still far away from fully understanding motivation, our results suggest that transferring knowledge from language into vision can help machines understand why people in images might be performing an action.

1.496 Motion from Structure (MfS): Searching for 3D Objects in Cluttered Point Trajectories

[Link to Paper]

Object detection has been a long standing problem in computer vision, and state-of-the-art approaches rely on the use of sophisticated features and/or classifiers. However, these learning-based approaches heavily depend on the quality and quantity of labeled data, and do not generalize well to extreme poses or textureless objects. In this work, we explore

the use of 3D shape models to detect objects in videos in an unsupervised manner. We call this problem Motion from Structure (MfS): given a set of point trajectories and a 3D model of the object of interest, find a subset of trajectories that correspond to the 3D model and estimate its alignment (i.e., compute the motion matrix). MfS is related to Structure from Motion (SfM) and motion segmentation problems: unlike SfM, the structure of the object is known but the correspondence between the trajectories and the object is unknown; unlike motion segmentation, the MfS problem incorporates 3D structure, providing robustness to tracking mismatches and outliers. Experiments illustrate how our MfS algorithm outperforms alternative approaches in both synthetic data and real videos extracted from YouTube.

1.497 Copula Ordinal Regression for Joint Estimation of Facial Action Unit Intensity

[Link to Paper]

Joint modeling of the intensity of facial action units (AUs) from face images is challenging due to the large number of AUs (30+) and their intensity levels (6). This is in part due to the lack of suitable models that can efficiently handle such a large number of outputs/classes simultaneously, but also due to the lack of labelled target data. For this reason, majority of the methods proposed so far resort to independent classifiers for the AU intensity. This is suboptimal for at least two reasons: the facial appearance of some AUs changes depending on the intensity of other AUs, and some AUs co-occur more often than others. Encoding this is expected to improve the estimation of target AU intensities, especially in the case of noisy image features, head-pose variations and imbalanced training data. To this end, we introduce a novel modeling framework, Copula Ordinal Regression (COR), that leverages the power of copula functions and CRFs, to detangle the probabilistic modeling of AU dependencies from the marginal modeling of the AU intensity. Consequently, the COR model achieves the joint learning and inference of intensities of multiple AUs, while being computationally tractable. We show on two challenging datasets of naturalistic facial expressions that the proposed approach consistently outperforms (i) independent modeling of AU intensities, and (ii) the state-ofthe-art approach for the target task.

1.498 Actionness Estimation Using Hybrid Fully Convolutional Networks

[Link to Paper]

Actionness [3] was introduced to quantify the likelihood of containing a generic action instance at a specific location. Accurate and efficient estimation of actionness is important in video analysis and may benefit other relevant tasks such as action recognition and action

detection. This paper presents a new deep architecture for actionness estimation, called hybrid fully convolutional network (HFCN), which is composed of appearance FCN (A-FCN) and motion FCN (M-FCN). These two FCNs leverage the strong capacity of deep models to estimate actionness maps from the perspectives of static appearance and dynamic motion, respectively. In addition, the fully convolutional nature of H-FCN allows it to efficiently process videos with arbitrary sizes. Experiments are conducted on the challenging datasets of Stanford40, UCF Sports, and JHMDB to verify the effectiveness of H-FCN on actionness estimation, which demonstrate that our method achieves superior performance to previous ones. Moreover, we apply the estimated actionness maps on action proposal generation and action detection. Our actionness maps advance the current state-of-the-art performance of these tasks substantially.

1.499   Actions Transformations

[Link to Paper]

What defines an action like "kicking ball"? We argue that the true meaning of an action lies in the change or transformation an action brings to the environment. In this paper, we propose a novel representation for actions by modeling an action as a transformation which changes the state of the environment before the action happens (precondition) to the state after the action (effect). Motivated by recent advancements of video representation using deep learning, we design a Siamese network which models the action as a transformation on a high-level feature space. We show that our model gives improvements on standard action recognition datasets including UCF101 and HMDB51. More importantly, our approach is able to generalize beyond learned action categories and shows significant performance improvement on cross-category generalization on our new ACT dataset.

1.500   CNN-RNN: A Unified Framework for Multi-label Image Classification

[Link to Paper]

While deep convolutional neural networks (CNNs) have shown a great success in single-label image classification, it is important to note that real world images generally contain multiple labels, which could correspond to different objects, scenes, actions and attributes in an image. Traditional approaches to multi-label image classification learn independent classifiers for each category and employ ranking or thresholding on the classification results. These techniques, although working well, fail to explicitly exploit the label dependencies in an image. In this paper, we utilize recurrent neural networks (RNNs) to address this problem. Combined with CNNs, the proposed CNN-RNN framework learns a joint image-label embedding to characterize the semantic label dependency as well as

the image-label relevance, and it can be trained end-to-end from scratch to integrate both information in a unified framework. Experimental results on public benchmark datasets demonstrate that the proposed architecture achieves better performance than the state-of-the-art multi-label classification models.

1.501 Constructing Canonical Regions for Fast and Effective View Selection

[Link to Paper]

In view selection, little work has been done for optimizing the search process; views must be densely distributed and checked individually. Thus, evaluating poor views wastes much time, and a poor view may even be misidentified as a best one. In this paper, we propose a search strategy by identifying the regions that are very likely to contain best views, referred to as canonical regions. It is by decomposing the model under investigation into meaningful parts, and using the canonical views of these parts to generate canonical regions. Applying existing view selection methods in the canonical regions can not only accelerate the search process but also guarantee the quality of obtained views. As a result, when our canonical regions are used for searching N-best views during comprehensive model analysis, we can attain greater search speed and reduce the number of views required. Experimental results show the effectiveness of our method.

1.502 Context-Aware Gaussian Fields for Non-rigid Point Set Registration

[Link to Paper]

Point set registration (PSR) is a fundamental problem in computer vision and pattern recognition, and it has been successfully applied to many applications. Although widely used, existing PSR methods cannot align point sets robustly under degradations, such as deformation, noise, occlusion, outlier, rotation, and multi-view changes. This paper proposes context-aware Gaussian fields (CA-LapGF) for nonrigid PSR subject to global rigid and local non-rigid geometric constraints, where a laplacian regularized term is added to preserve the intrinsic geometry of the transformed set. CA-LapGF uses a robust objective function and the quasi-Newton algorithm to estimate the likely correspondences, and the non-rigid transformation parameters between two point sets iteratively. The CA-LapGF can estimate non-rigid transformations, which are mapped to reproducing kernel Hilbert spaces, accurately and robustly in the presence of degradations. Experimental results on synthetic and real images reveal that how CA-LapGF outperforms state-of-the-art algorithms for non-rigid PSR.

1.503   D3: Deep Dual-Domain Based Fast Restoration of JPEG-Compressed Images

[Link to Paper]

In this paper, we design a Deep Dual-Domain (D3) based fast restoration model to remove artifacts of JPEG compressed images. It leverages the large learning capacity of deep networks, as well as the problem-specific expertise that was hardly incorporated in the past design of deep architectures. For the latter, we take into consideration both the prior knowledge of the JPEG compression scheme, and the successful practice of the sparsity-based dual-domain approach. We further design the One-Step Sparse Inference (1-SI) module, as an efficient and lightweighted feed-forward approximation of sparse coding. Extensive experiments verify the superiority of the proposed D3 model over several state-of-the-art methods. Specifically, our best model is capable of outperforming the latest deep model for around 1 dB in PSNR, and is 30 times faster.

1.504   Depth from Semi-Calibrated Stereo and Defocus

[Link to Paper]

In this work, we propose a multi-camera system where we combine a main high-quality camera with two low-res auxiliary cameras. The auxiliary cameras are well calibrated and act as a passive depth sensor by generating disparity maps. The main camera has an interchangeable lens and can produce good quality images at high resolution. Our goal is, given the low-res depth map from the auxiliary cameras, generate a depth map from the viewpoint of the main camera. The advantage of our system, compared to other systems such as light-field cameras or RGBD sensors, is the ability to generate a high-resolution color image with a complete depth map, without sacrificing resolution and with minimal auxiliary hardware. Since the main camera has an interchangeable lens, it cannot be calibrated beforehand, and directly applying stereo matching on it and either of the auxiliary cameras often leads to unsatisfactory results. Utilizing both the calibrated cameras at once, we propose a novel approach to better estimate the disparity map of the main camera. Then by combining the defocus cue of the main camera, the disparity map can be further improved. We demonstrate the performance of our algorithm on various scenes.

1.505   Dictionary Pair Classifier Driven Convolutional Neural Networks for Object Detection

[Link to Paper]

Feature representation and object category classification are two key components of most object detection methods. While significant improvements have been achieved for deep feature representation learning, traditional SVM/softmax classifiers remain the dominant methods for the final object category classification. However, SVM/softmax classifiers lack the capacity of explicitly exploiting the complex structure of deep features, as they are purely discriminative methods. The recently proposed discriminative dictionary pair learning (DPL) model involves a fidelity term to minimize the reconstruction loss and a discrimination term to enhance the discriminative capability of the learned dictionary pair, and thus is appropriate for balancing the representation and discrimination to boost object detection performance. In this paper, we propose a novel object detection system by unifying DPL with the convolutional feature learning. Specifically, we incorporate DPL as a Dictionary Pair Classifier Layer (DPCL) into the deep architecture, and develop an end-to-end learning algorithm for optimizing the dictionary pairs and the neural networks simultaneously. Moreover, we design a multi-task loss for guiding our model to accomplish the three correlated tasks: objectness estimation, categoryness computation, and bounding box regression. From the extensive experiments on PASCAL VOC 2007/2012 benchmarks, our approach demonstrates the effectiveness to substantially improve the performances over the popular existing object detection frameworks (e.g., R-CNN [13] and FRCN [12]), and achieves new state-of-the-arts.

1.506 Event-specific Image Importance

[Link to Paper]

When creating a photo album of an event, people typically select a few important images to keep or share. There is some consistency in the process of choosing the important images, and discarding the unimportant ones. Modeling this selection process will assist automatic photo selection and album summarization. In this paper, we show that the selection of important images is consistent among different viewers, and that this selection process is related to the event type of the album. We introduce the concept of event-specific image importance. We collected a new event album dataset with human annotation of the relative image importance with each event album. We also propose a Convolutional Neural Network (CNN) based method to predict the image importance score of a given event album, using a novel rank loss function and a progressive training scheme. Results demonstrate that our method significantly outperforms various baseline methods.

1.507 GraB: Visual Saliency via Novel Graph Model and Background Priors

[Link to Paper]

We propose an unsupervised bottom-up saliency detection approach by exploiting novel graph structure and background priors. The input image is represented as an undirected graph with superpixels as nodes. Feature vectors are extracted from each node to cover regional color, contrast and texture information. A novel graph model is proposed to effectively capture local and global saliency cues. To obtain more accurate saliency estimations, we optimize the saliency map by using a robust background measure. Comprehensive evaluations on benchmark datasets indicate that our algorithm universally surpasses state-of-the-art unsupervised solutions and performs favorably against supervised approaches.

1.508 Improving Human Action Recognition by Non-action Classification

[Link to Paper]

In this paper we consider the task of recognizing human actions in realistic video where human actions are dominated by irrelevant factors. We first study the benefits of removing non-action video segments, which are the ones that do not portray any human action. We then learn a nonaction classifier and use it to down-weight irrelevant video segments. The non-action classifier is trained using ActionThread, a dataset with shot-level annotation for the occurrence or absence of a human action. The non-action classifier can be used to identify non-action shots with high precision and subsequently used to improve the performance of action recognition systems.

1.509 Joint Learning of Single-image and Cross-image Representations for Person Re-identification

[Link to Paper]

Person re-identification has been usually solved as either the matching of single-image representation (SIR) or the classification of cross-image representation (CIR). In this work, we exploit the connection between these two categories of methods, and propose a joint learning framework to unify SIR and CIR using convolutional neural network (CNN). Specifically, our deep architecture contains one shared sub-network together with two sub-networks that extract the SIRs of given images and the CIRs of given image pairs, respectively. The SIR sub-network is required to be computed once for each image (in

both the probe and gallery sets), and the depth of the CIR sub-network is required to be minimal to reduce computational burden. Therefore, the two types of representation can be jointly optimized for pursuing better matching accuracy with moderate computational cost. Furthermore, the representations learned with pairwise comparison and triplet comparison objectives can be combined to improve matching performance. Experiments on the CUHK03, CUHK01 and VIPeR datasets show that the proposed method can achieve favorable accuracy while compared with state-of-the-arts.

1.510 Learning Deep Structure-Preserving Image-Text Embeddings

[Link to Paper]

This paper proposes a method for learning joint embeddings of images and text using a two-branch neural network with multiple layers of linear projections followed by nonlinearities. The network is trained using a largemargin objective that combines cross-view ranking constraints with within-view neighborhood structure preservation constraints inspired by metric learning literature. Extensive experiments show that our approach gains significant improvements in accuracy for image-to-text and textto-image retrieval. Our method achieves new state-of-theart results on the Flickr30K and MSCOCO image-sentence datasets and shows promise on the new task of phrase localization on the Flickr30K Entities dataset.

1.511 Mining 3D Key-Pose-Motifs for Action Recognition

[Link to Paper]

Recognizing an action from a sequence of 3D skeletal poses is a challenging task. First, different actors may perform the same action in various styles. Second, the estimated poses are sometimes inaccurate. These challenges can cause large variations between instances of the same class. Third, the datasets are usually small, with only a few actors performing few repetitions of each action. Hence training complex classifiers risks over-fitting the data. We address this task by mining a set of key-pose-motifs for each action class. A key-pose-motif contains a set of ordered poses, which are required to be close but not necessarily adjacent in the action sequences. The representation is robust to style variations. The key-pose-motifs are represented in terms of a dictionary using soft-quantization to deal with inaccuracies caused by quantization. We propose an efficient algorithm to mine key-pose-motifs taking into account of these probabilities. We classify a sequence by matching it to the motifs of each class and selecting the class that maximizes the matching score. This simple classifier obtains state-ofthe-art performance on two benchmark datasets.

237

1.512 Mining Discriminative Triplets of Patches for Fine-Grained Classification

[Link to Paper]

Fine-grained classification involves distinguishing between similar sub-categories based on subtle differences in highly localized regions; therefore, accurate localization of discriminative regions remains a major challenge. We describe a patch-based framework to address this problem. We introduce triplets of patches with geometric constraints to improve the accuracy of patch localization, and automatically mine discriminative geometrically-constrained triplets for classification. The resulting approach only requires object bounding boxes. Its effectiveness is demonstrated using four publicly available fine-grained datasets, on which it outperforms or achieves comparable performance to the state-of-the-art in classification.

1.513 Modality and Component Aware Feature Fusion for RGB-D Scene Classification

[Link to Paper]

While convolutional neural networks (CNN) have been excellent for object recognition, the greater spatial variability in scene images typically meant that the standard full-image CNN features are suboptimal for scene classification. In this paper, we investigate a framework allowing greater spatial flexibility, in which the Fisher vector (FV) encoded distribution of local CNN features, obtained from a multitude of region proposals per image, is considered instead. The CNN features are computed from an augmented pixel-wise representation comprising multiple modalities of RGB, HHA and surface normals, as extracted from RGB-D data. More significantly, we make two postulates: (1) component sparsity – that only a small variety of region proposals and their corresponding FV GMM components contribute to scene discriminability, and (2) modal non-sparsity – within these discriminative components, all modalities have important contribution. In our framework, these are implemented through regularization terms applying group lasso to GMM components and exclusive group lasso across modalities. By learning and combining regressors for both proposal-based FV features and global CNN features, we were able to achieve state-of-the-art scene classification performance on the SUNRGBD Dataset and NYU Depth Dataset V2.

1.514 Online Reconstruction of Indoor Scenes from RGB-D Streams

[Link to Paper]

A system capable of performing robust online volumetric reconstruction of indoor scenes based on input from a handheld RGB-D camera is presented. Our system is powered by a two-pass reconstruction scheme. The first pass tracks camera poses at video rate and simultaneously constructs a pose graph on-the-fly. The tracker operates in real-time, which allows the reconstruction results to be visualized during the scanning process. Live visual feedbacks makes the scanning operation fast and intuitive. Upon termination of scanning, the second pass takes place to handle loop closures and reconstruct the final model using globally refined camera trajectories. The system is online with low delay and returns a dense model of sufficient accuracy. The beauty of this system lies in its speed, accuracy, simplicity and ease of implementation when compared to previous methods. We demonstrate the performance of our system on several real-world scenes and quantitatively assess the modeling accuracy with respect to ground truth models obtained from a LIDAR scanner.

1.515 PPP: Joint Pointwise and Pairwise Image Label Prediction

[Link to Paper]

Pointwise label and pairwise label are both widely used in computer vision tasks. For example, supervised image classification and annotation approaches use pointwise label, while attribute-based image relative learning often adopts pairwise labels. These two types of labels are often considered independently and most existing efforts utilize them separately. However, pointwise labels in image classification and tag annotation are inherently related to the pairwise labels. For example, an image labeled with "coast" and annotated with "beach, sea, sand, sky" is more likely to have a higher ranking score in terms of the attribute "open"; while "men shoes" ranked highly on the attribute "formal" are likely to be annotated with "leather, lace up" than "buckle, fabric". The existence of potential relations between pointwise labels and pairwise labels motivates us to fuse them together for jointly addressing related vision tasks. In particular, we provide a principled way to capture the relations between class labels, tags and attributes; and propose a novel framework PPP(Pointwise and Pairwise image label Prediction), which is based on overlapped group structure extracted from the pointwise-pairwise-label bipartite graph. With experiments on benchmark datasets, we demonstrate that the proposed framework achieves superior performance on three vision tasks compared to the state-of-the-art methods.

1.516 RAID-G: Robust Estimation of Approximate Infinite Dimensional Gaussian with Application to Material Recognition

[Link to Paper]

Infinite dimensional covariance descriptors can provide richer and more discriminative information than their low dimensional counterparts. In this paper, we propose a novel image descriptor, namely, robust approximate infinite dimensional Gaussian (RAID-G). The challenges of RAID-G mainly lie on two aspects: (1) description of infinite dimensional Gaussian is difficult due to its non-linear Riemannian geometric structure and the infinite dimensional setting, hence effective approximation is necessary; (2) traditional maximum likelihood estimation (MLE) is not robust to high (even infinite) dimensional covariance matrix in Gaussian setting. To address these challenges, explicit feature mapping (EFM) is first introduced for effective approximation of infinite dimensional Gaussian induced by additive kernel function, and then a new regularized MLE method based on von Neumann divergence is proposed for robust estimation of covariance matrix. The EFM and proposed regularized MLE allow a closed-form of RAID-G, which is very efficient and effective for high dimensional features. We extend RAID-G by using the outputs of deep convolutional neural networks as original features, and apply it to material recognition. Our approach is evaluated on five material benchmarks and one fine-grained benchmark. It achieves 84.9% accuracy on FMD and 86.3% accuracy on UIUC material database, which are much higher than state-of-the-arts.

1.517   Recurrent Face Aging

[Link to Paper]

Modeling the aging process of human face is important for cross-age face verification and recognition. In this paper, we introduce a recurrent face aging (RFA) framework based on a recurrent neural network which can identify the ages of people from 0 to 80. Due to the lack of labeled face data of the same person captured in a long range of ages, traditional face aging models usually split the ages into discrete groups and learn a one-step face feature transformation for each pair of adjacent age groups. However, those methods neglect the in-between evolving states between the adjacent age groups and the synthesized faces often suffer from severe ghosting artifacts. Since human face aging is a smooth progression, it is more appropriate to age the face by going through smooth transition states. In this way, the ghosting artifacts can be effectively eliminated and the intermediate aged faces between two discrete age groups can also be obtained. Towards this target, we employ a twolayer gated recurrent unit as the basic recurrent module whose bottom layer encodes a young face to a latent representation and the top layer decodes the representation to a corresponding older face. The experimental results demonstrate our proposed RFA provides better aging faces over other state-of-the-art age progression methods.

1.518 Relaxation-Based Preprocessing Techniques for Markov Random Field Inference

[Link to Paper]

Markov Random Fields (MRFs) are a widely used graphical model, but the inference problem is NP-hard. For first-order MRFs with binary labels, Dead End Elimination (DEE) [7] and QPBO [2, 14] can find the optimal labeling for some variables; the much harder case of larger label sets has been addressed by Kovtun [16, 17] and related methods [12, 23, 24, 25], which impose substantial computational overhead. We describe an efficient algorithm to correctly label a subset of the variables for arbitrary MRFs, with particularly good performance on binary MRFs. We propose a sufficient condition to check if a partial labeling is optimal, which is a generalization of DEE's purely local test. We give a hierarchy of relaxations that provide larger optimal partial labelings at the cost of additional computation. Empirical studies were conducted on several benchmarks, using expansion moves [4] for inference. Our algorithm runs in a few seconds, and improves the speed of MRF inference with expansion moves by a factor of 1.5 to 12.

1.519 Removing Clouds and Recovering Ground Observations in Satellite Image Sequences via Temporally Contiguous Robust Matrix Completion

[Link to Paper]

We consider the problem of removing and replacing clouds in satellite image sequences, which has a wide range of applications in remote sensing. Our approach first detects and removes the cloud-contaminated part of the image sequences. It then recovers the missing scenes from the clean parts using the proposed "TECROMAC" (TEmporally Contiguous RObust MAtrix Completion) objective. The objective function balances temporal smoothness with a low rank solution while staying close to the original observations. The matrix whose the rows are pixels and columns are days corresponding to the image, has low-rank because the pixels reflect land-types such as vegetation, roads and lakes and there are relatively few variations as a result. We provide efficient optimization algorithms for TECROMAC, so we can exploit images containing millions of pixels. Empirical results on real satellite image sequences, as well as simulated data, demonstrate that our approach is able to recover underlying images from heavily cloud-contaminated observations.

1.520 STCT: Sequentially Training Convolutional Networks for Visual Tracking

[Link to Paper]

Due to the limited amount of training samples, finetuning pre-trained deep models online is prone to overfitting. In this paper, we propose a sequential training method for convolutional neural networks (CNNs) to effectively transfer pre-trained deep features for online applications. We regard a CNN as an ensemble with each channel of the output feature map as an individual base learner. Each base learner is trained using different loss criterions to reduce correlation and avoid over-training. To achieve the best ensemble online, all the base learners are sequentially sampled into the ensemble via important sampling. To further improve the robustness of each base learner, we propose to train the convolutional layers with random binary masks, which serves as a regularization to enforce each base learner to focus on different input features. The proposed online training method is applied to visual tracking problem by transferring deep features trained on massive annotated visual data and is shown to significantly improve tracking performance. Extensive experiments are conducted on two challenging benchmark data set and demonstrate that our tracking algorithm can outperform state-of-the-art methods with a considerable margin.

1.521 SVBRDF-Invariant Shape and Reflectance Estimation from Light-Field Cameras

[Link to Paper]

Light-field cameras have recently emerged as a powerful tool for one-shot passive 3D shape capture. However, obtaining the shape of glossy objects like metals, plastics or ceramics remains challenging, since standard Lambertian cues like photo-consistency cannot be easily applied. In this paper, we derive a spatially-varying (SV)BRDF-invariant theory for recovering 3D shape and reflectance from lightfield cameras. Our key theoretical insight is a novel analysis of diffuse plus single-lobe SVBRDFs under a light-field setup. We show that, although direct shape recovery is not possible, an equation relating depths and normals can still be derived. Using this equation, we then propose using a polynomial (quadratic) shape prior to resolve the shape ambiguity. Once shape is estimated, we also recover the reflectance. We present extensive synthetic data on the entire MERL BRDF dataset, as well as a number of real examples to validate the theory, where we simultaneously recover shape and BRDFs from a single image taken with a Lytro Illum camera.

1.522 Studying Very Low Resolution Recognition Using Deep Networks

[Link to Paper]

Visual recognition research often assumes a sufficient resolution of the region of interest (ROI). That is usually violated in practice, inspiring us to explore the Very Low Resolution Recognition (VLRR) problem. Typically, the ROI in a VLRR problem can be smaller than 16 16 pixels, and is challenging to be recognized even by human experts. We attempt to solve the VLRR problem using deep learning methods. Taking advantage of techniques primarily in super resolution, domain adaptation and robust regression, we formulate a dedicated deep learning method and demonstrate how these techniques are incorporated step by step. Any extra complexity, when introduced, is fully justified by both analysis and simulation results. The resulting Robust Partially Coupled Networks achieves feature enhancement and recognition simultaneously. It allows for both the flexibility to combat the LR-HR domain mismatch, and the robustness to outliers. Finally, the effectiveness of the proposed models is evaluated on three different VLRR tasks, including face identification, digit recognition and font recognition, all of which obtain very impressive performances.

1.523 Supervised Quantization for Similarity Search

[Link to Paper]

In this paper, we address the problem of searching for semantically similar images from a large database. We present a compact coding approach, supervised quantization. Our approach simultaneously learns feature selection that linearly transforms the database points into a low-dimensional discriminative subspace, and quantizes the data points in the transformed space. The optimization criterion is that the quantized points not only approximate the transformed points accurately, but also are semantically separable: the points belonging to a class lie in a cluster that is not overlapped with other clusters corresponding to other classes, which is formulated as a classification problem. The experiments on several standard datasets show the superiority of our approach over the state-of-the art supervised hashing and unsupervised quantization algorithms.

1.524 The Global Patch Collider

[Link to Paper]

This paper proposes a novel extremely efficient, fullyparallelizable, task-specific algorithm for the computation of global point-wise correspondences in images and videos. Our al-

gorithm, the Global Patch Collider, is based on detecting unique collisions between image points using a collection of learned tree structures that act as conditional hash functions. In contrast to conventional approaches that rely on pairwise distance computation, our algorithm isolates distinctive pixel pairs that hit the same leaf during traversal through multiple learned tree structures. The split functions stored at the intermediate nodes of the trees are trained to ensure that only visually similar patches or their geometric or photometric transformed versions fall into the same leaf node. The matching process involves passing all pixel positions in the images under analysis through the tree structures. We then compute matches by isolating points that uniquely collide with each other ie. fell in the same empty leaf in multiple trees. Our algorithm is linear in the number of pixels but can be made constant time on a parallel computation architecture as the tree traversal for individual image points is decoupled. We demonstrate the efficacy of our method by using it to perform optical flow matching and stereo matching on some challenging benchmarks. Experimental results show that not only is our method extremely computationally efficient, but it is also able to match or outperform state of the art methods that are much more complex.

1.525 Walk and Learn: Facial Attribute Representation Learning from Egocentric Video and Contextual Data

[Link to Paper]

The way people look in terms of facial attributes (ethnicity, hair color, facial hair, etc.) and the clothes or accessories they wear (sunglasses, hat, hoodies, etc.) is highly dependent on geo-location and weather condition, respectively. This work explores, for the first time, the use of this contextual information, as people with wearable cameras walk across different neighborhoods of a city, in order to learn a rich feature representation for facial attribute classification, without the costly manual annotation required by previous methods. By tracking the faces of casual walkers on more than 40 hours of egocentric video, we are able to cover tens of thousands of different identities and automatically extract nearly 5 million pairs of images connected by or from different face tracks, along with their weather and location context, under pose and lighting variations. These image pairs are then fed into a deep network that preserves similarity of images connected by the same track, in order to capture identity-related attribute features, and optimizes for location and weather prediction to capture additional facial attribute features. Finally, the network is fine-tuned with manually annotated samples. We perform an extensive experimental analysis on wearable data and two standard benchmark datasets based on web images (LFWA and CelebA). Our method outperforms by a large margin a network trained from scratch. Moreover, even without using manually annotated identity labels for pre-training as in previous methods, our approach achieves results that are better than the state of the art.

1.526 What's Wrong with that Object? Identifying Images of Unusual Objects by Modelling the Detection Score Distribution

[Link to Paper]

This paper studies the challenging problem of identifying unusual instances of known objects in images within an "open world" setting. That is, we aim to find objects that are members of a known class, but which are not typical of that class. Thus the "unusual object" should be distinguished from both the "regular object" and the "other objects". Such unusual objects may be of interest in many applications such as surveillance or quality control. We propose to identify unusual objects by inspecting the distribution of object detection scores at multiple image regions. The key observation motivating our approach is that "regular object" images, "unusual object" images and "other objects" images exhibit different region-level scores in terms of both the score values and the spatial distributions. To model these distributions we propose to use Gaussian Processes (GP) to construct two separate generative models, one for the "regular object" and the other for the "other objects". More specifically, we design a new covariance function to simultaneously model the detection score at a single location and the score dependencies between multiple regions. We demonstrate that the proposed approach outperforms comparable methods on a new large dataset constructed for the purpose.

1.527 Cataloging Public Objects Using Aerial and Street-Level Images Urban Trees

[Link to Paper]

Each corner of the inhabited world is imaged from multiple viewpoints with increasing frequency. Online map services like Google Maps or Here Maps provide direct access to huge amounts of densely sampled, georeferenced images from street view and aerial perspective. There is an opportunity to design computer vision systems that will help us search, catalog and monitor public infrastructure, buildings and artifacts. We explore the architecture and feasibility of such a system. The main technical challenge is combining test time information from multiple views of each geographic location (e.g., aerial and street views). We implement two modules: det2geo, which detects the set of locations of objects belonging to a given category, and geo2cat, which computes the fine-grained category of the object at a given location. We introduce a solution that adapts state-ofthe-art CNN-based object detectors and classifiers. We test our method on "Pasadena Urban Trees", a new dataset of 80,000 trees with geographic and species annotations, and show that combining multiple views significantly improves both tree detection and tree species classification, rivaling human performance. Street View Images Aerial Images Map Data Multiview Detection & Recognition Southern Magnolia Fine-Grained Geographic Tree Catalog Figure 1. Overview

of proposed automated public tree cataloguing system from online maps. Aerial images and street view panoramas along with semantic map data are downloaded for some geographical region. Category detection and fine-grained classification algorithms are trained from human-annotated exemplars. Detection, classification and geolocation information is computed automatically from multiple street view images and aerial images and combined with map data to achieve a geolocated fine-grained catalog. The image shows a catalog of location and species of trees in a medium-sized city.

1.528 Convolutional Pose Machines

[Link to Paper]

Pose Machines provide a sequential prediction framework for learning rich implicit spatial models. In this work we show a systematic design for how convolutional networks can be incorporated into the pose machine framework for learning image features and image-dependent spatial models for the task of pose estimation. The contribution of this paper is to implicitly model long-range dependencies between variables in structured prediction tasks such as articulated pose estimation. We achieve this by designing a sequential architecture composed of convolutional networks that directly operate on belief maps from previous stages, producing increasingly refined estimates for part locations, without the need for explicit graphical model-style inference. Our approach addresses the characteristic difficulty of vanishing gradients during training by providing a natural learning objective function that enforces intermediate supervision, thereby replenishing back-propagated gradients and conditioning the learning procedure. We demonstrate state-of-the-art performance and outperform competing methods on standard benchmarks including the MPII, LSP, and FLIC datasets.

1.529 Dense Human Body Correspondences Using Convolutional Networks

[Link to Paper]

We propose a deep learning approach for finding dense correspondences between 3D scans of people. Our method requires only partial geometric information in the form of two depth maps or partial reconstructed surfaces, works for humans in arbitrary poses and wearing any clothing, does not require the two people to be scanned from similar viewpoints, and runs in real time. We use a deep convolutional neural network to train a feature descriptor on depth map pixels, but crucially, rather than training the network to solve the shape correspondence problem directly, we train it to solve a body region classification problem, modified to increase the smoothness of the learned descriptors near region boundaries. This approach ensures that nearby points on the human body are nearby in feature space,

and vice versa, rendering the feature descriptor suitable for computing dense correspondences between the scans. We validate our method on real and synthetic data for both clothed and unclothed humans, and show that our correspondences are more robust than is possible with state-of-the-art unsupervised methods, and more accurate than those found using methods that require full watertight 3D geometry.

1.530 Region Ranking SVM for Image Classification

[Link to Paper]

The success of an image classification algorithm largely depends on how it incorporates local information in the global decision. Popular approaches such as averagepooling and max-pooling are suboptimal in many situations. In this paper we propose Region Ranking SVM (RRSVM), a novel method for pooling local information from multiple regions. RRSVM exploits the correlation of local regions in an image, and it jointly learns a region evaluation function and a scheme for integrating multiple regions. Experiments on PASCAL VOC 2007, VOC 2012, and ILSVRC2014 datasets show that RRSVM outperforms the methods that use the same feature type and extract features from the same set of local regions. RRSVM achieves similar to or better than the state-of-the-art performance on all datasets.

1.531 Trace Quotient Meets Sparsity: A Method for Learning Low Dimensional Image Representations

[Link to Paper]

This paper presents an algorithm that allows to learn low dimensional representations of images in an unsupervised manner. The core idea is to combine two criteria that play important roles in unsupervised representation learning, namely sparsity and trace quotient. The former is known to be a convenient tool to identify underlying factors, and the latter is known as a disentanglement of underlying discriminative factors. In this work, we develop a generic cost function for learning jointly a sparsifying dictionary and a dimensionality reduction transformation. It leads to several counterparts of classic low dimensional representation methods, such as Principal Component Analysis, Local Linear Embedding, and Laplacian Eigenmap. Our proposed optimisation algorithm leverages the efficiency of geometric optimisation on Riemannian manifolds and a closed form solution to the elastic net problem.

1.532 Latent Factor Guided Convolutional Neural Networks for Age-Invariant Face Recognition

[Link to Paper]

While considerable progresses have been made on face recognition, age-invariant face recognition (AIFR) still remains a major challenge in real world applications of face recognition systems. The major difficulty of AIFR arises from the fact that the facial appearance is subject to significant intra-personal changes caused by the aging process over time. In order to address this problem, we propose a novel deep face recognition framework to learn the ageinvariant deep face features through a carefully designed CNN model. To the best of our knowledge, this is the first attempt to show the effectiveness of deep CNNs in advancing the state-of-the-art of AIFR. Extensive experiments are conducted on several public domain face aging datasets (MORPH Album2, FGNET, and CACD-VS) to demonstrate the effectiveness of the proposed model over the state-ofthe-art. We also verify the excellent generalization of our new model on the famous LFW dataset.

1.533 Efficient Large-scale Approximate Nearest Neighbor Search on the GPU

[Link to Paper]

We present a new approach for efficient approximate nearest neighbor (ANN) search in high dimensional spaces, extending the idea of Product Quantization. We propose a two level product and vector quantization tree that reduces the number of vector comparisons required during tree traversal. Our approach also includes a novel highly parallelizable re-ranking method for candidate vectors by efficiently reusing already computed intermediate values. Due to its small memory footprint during traversal the method lends itself to an efficient, parallel GPU implementation. This Product Quantization Tree (PQT) approach significantly outperforms recent state of the art methods for high dimensional nearest neighbor queries on standard reference datasets. Ours is the first work that demonstrates GPU performance superior to CPU performance on high dimensional, large scale ANN problems in time-critical real-world applications, like loop-closing in videos.

1.534 Robust Light Field Depth Estimation for Noisy Scene with Occlusion

[Link to Paper]

Light field depth estimation is an essential part of many light field applications. Numerous algorithms have been developed using various light field characteristics. However, conven-

tional methods fail when handling noisy scene with occlusion. To remedy this problem, we present a light field depth estimation method which is more robust to occlusion and less sensitive to noise. Novel data costs using angular entropy metric and adaptive defocus response are introduced. Integration of both data costs improves the occlusion and noise invariant capability significantly. Cost volume filtering and graph cut optimization are utilized to improve the accuracy of the depth map. Experimental results confirm that the proposed method is robust and achieves high quality depth maps in various scenes. The proposed method outperforms the state-of-the-art light field depth estimation methods in qualitative and quantitative evaluation.

1.535 Spatially Binned ROC: A Comprehensive Saliency Metric

[Link to Paper]

A recent trend in saliency algorithm development is large-scale benchmarking and algorithm ranking with ground truth provided by datasets of human fixations. In order to accommodate the strong bias humans have toward central fixations, it is common to replace traditional ROC metrics with a shuffled ROC metric which uses randomly sampled fixations from other images in the database as the negative set. However, the shuffled ROC introduces a number of problematic elements, including a fundamental assumption that it is possible to separate visual salience and image spatial arrangement. We argue that it is more informative to directly measure the effect of spatial bias on algorithm performance rather than try to correct for it. To capture and quantify these known sources of bias, we propose a novel metric for measuring saliency algorithm performance: the spatially binned ROC (spROC). This metric provides direct insight into the spatial biases of a saliency algorithm without sacrificing the intuitive raw performance evaluation of traditional ROC measurements. By quantitatively measuring the bias in saliency algorithms, researchers will be better equipped to select and optimize the most appropriate algorithm for a given task. We use a baseline measure of inherent algorithm bias to show that Adaptive Whitening Saliency (AWS) [14], Attention by Information Maximization (AIM) [8], and Dynamic Visual Attention (DVA) [20] provide the least spatially biased results, suiting them for tasks in which there is no information about the underlying spatial bias of the stimuli, whereas algorithms such as Graph Based Visual Saliency (GBVS) [18] and ContextAware Saliency (CAS) [15] have a significant inherent central bias.

1.536 Regularity-Driven Building Facade Matching between Aerial and Street Views

[Link to Paper]

We present an approach for detecting and matching building facades between aerial view and street-view images. We exploit the regularity of urban scene facades as captured by their lattice structures and deduced from median-tiles' shape context, color, texture and spatial similarities. Our experimental results demonstrate effective matching of oblique and partially-occluded facades between aerial and ground views. Quantitative comparisons for automated urban scene facade matching from three cities show superior performance of our method over baseline SIFT, Root-SIFT and the more sophisticated ScaleSelective Self-Similarity and Binary Coherent Edge descriptors. We also illustrate regularity-based applications of occlusion removal from street views and higher-resolution texture-replacement in aerial views.

1.537 Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources

[Link to Paper]

We propose a method for visual question answering which combines an internal representation of the content of an image with information extracted from a general knowledge base to answer a broad range of image-based questions. This allows more complex questions to be answered using the predominant neural network-based approach than has previously been possible. It particularly allows questions to be asked about the contents of an image, even when the image itself does not contain the whole answer. The method constructs a textual representation of the semantic content of an image, and merges it with textual information sourced from a knowledge base, to develop a deeper understanding of the scene viewed. Priming a recurrent neural network with this combined information, and the submitted question, leads to a very flexible visual question answering approach. We are specifically able to answer questions posed in natural language, that refer to information not contained in the image. We demonstrate the effectiveness of our model on two publicly available datasets, Toronto COCO-QA [23] and VQA [1] and show that it produces the best reported results in both cases.

1.538 Constrained Deep Transfer Feature Learning and its Applications

[Link to Paper]

Feature learning with deep models has achieved impressive results for both data representation and classification for various vision tasks. Deep feature learning, however, typically requires a large amount of training data, which may not be feasible for some application domains. Transfer learning can be one of the approaches to alleviate this problem by transferring data from data-rich source domain to data-scarce target domain. Existing

transfer learning methods typically perform one-shot transfer learning and often ignore the specific properties that the transferred data must satisfy. To address these issues, we introduce a constrained deep transfer feature learning method to perform simultaneous transfer learning and feature learning by performing transfer learning in a progressively improving feature space iteratively in order to better narrow the gap between the target domain and the source domain for effective transfer of the data from source domain to target domain. Furthermore, we propose to exploit the target domain knowledge and incorporate such prior knowledge as constraint during transfer learning to ensure that the transferred data satisfies certain properties of the target domain. To demonstrate the effectiveness of the proposed constrained deep transfer feature learning method, we apply it to thermal feature learning for eye detection by transferring from the visible domain. We also applied the proposed method for cross-view facial expression recognition as a second application. The experimental results demonstrate the effectiveness of the proposed method for both applications.

1.539 Constrained Joint Cascade Regression Framework for Simultaneous Facial Action Unit Recognition and Facial Landmark Detection

[Link to Paper]

Cascade regression framework has been shown to be effective for facial landmark detection. It starts from an initial face shape and gradually predicts the face shape update from the local appearance features to generate the facial landmark locations in the next iteration until convergence. In this paper, we improve upon the cascade regression framework and propose the Constrained Joint Cascade Regression Framework (CJCRF) for simultaneous facial action unit recognition and facial landmark detection, which are two related face analysis tasks, but are seldomly exploited together. In particular, we first learn the relationships among facial action units and face shapes as a constraint. Then, in the proposed constrained joint cascade regression framework, with the help from the constraint, we iteratively update the facial landmark locations and the action unit activation probabilities until convergence. Experimental results demonstrate that the intertwined relationships of facial action units and face shapes boost the performances of both facial action unit recognition and facial landmark detection. The experimental results also demonstrate the effectiveness of the proposed method comparing to the state-of-the-art works.

1.540 Harnessing Object and Scene Semantics for Large-Scale Video Understanding

[Link to Paper]

Large-scale action recognition and video categorization are important problems in computer vision. To address these problems, we propose a novel object- and scene-based semantic fusion network and representation. Our semantic fusion network combines three streams of information using a three-layer neural network: (i) frame-based low-level CNN features, (ii) object features from a state-of-the-art large-scale CNN object-detector trained to recognize 20K classes, and (iii) scene features from a state-of-the-art CNN scene-detector trained to recognize 205 scenes. The trained network achieves improvements in supervised activity and video categorization in two complex large-scale datasets ActivityNet and FCVID, respectively. Further, by examining and back propagating information through the fusion network, semantic relationships (correlations) between video classes and objects/scenes can be discovered. These video class-object/video class-scene relationships can in turn be used as semantic representation for the video classes themselves. We illustrate effectiveness of this semantic representation through experiments on zero-shot action/video classification and clustering.

1.541 Quantized Convolutional Neural Networks for Mobile Devices

[Link to Paper]

Recently, convolutional neural networks (CNN) have demonstrated impressive performance in various computer vision tasks. However, high performance hardware is typically indispensable for the application of CNN models due to the high computation complexity, which prohibits their further extensions. In this paper, we propose an efficient framework, namely Quantized CNN, to simultaneously speed-up the computation and reduce the storage and memory overhead of CNN models. Both filter kernels in convolutional layers and weighting matrices in fully-connected layers are quantized, aiming at minimizing the estimation error of each layer's response. Extensive experiments on the ILSVRC-12 benchmark demonstrate 4 6 speed-up and 15 20 compression with merely one percentage loss of classification accuracy. With our quantized CNN model, even mobile devices can accurately classify images within one second.

1.542 What Value Do Explicit High Level Concepts Have in Vision to Language Problems?

[Link to Paper]

Much recent progress in Vision-to-Language (V2L) problems has been achieved through a combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). This approach does not explicitly represent high-level semantic concepts, but rather seeks to progress directly from image features to text. In this paper we investigate

whether this direct approach succeeds due to, or despite, the fact that it avoids the explicit representation of high-level information. We propose a method of incorporating high-level concepts into the successful CNN-RNN approach, and show that it achieves a significant improvement on the state-of-the-art in both image captioning and visual question answering. We also show that the same mechanism can be used to introduce external semantic information and that doing so further improves performance. We achieve the best reported results on both image captioning and VQA on several benchmark datasets, and provide an analysis of the value of explicit high-level concepts in V2L problems.

1.543 Latent Embeddings for Zero-shot Classification

[Link to Paper]

We present a novel latent embedding model for learning a compatibility function between image and class embeddings, in the context of zero-shot classification. The proposed method augments the state-of-the-art bilinear compatibility model by incorporating latent variables. Instead of learning a single bilinear map, it learns a collection of maps with the selection, of which map to use, being a latent variable for the current image-class pair. We train the model with a ranking based objective function which penalizes incorrect rankings of the true class for a given image. We empirically demonstrate that our model improves the state-of-the-art for various class embeddings consistently on three challenging publicly available datasets for the zero-shot setting. Moreover, our method leads to visually highly interpretable results with clear clusters of different fine-grained object properties that correspond to different latent variable maps.

1.544 Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification

[Link to Paper]

Learning generic and robust feature representations with data from multiple domains for the same problem is of great value, especially for the problems that have multiple datasets but none of them are large enough to provide abundant data variations. In this work, we present a pipeline for learning deep feature representations from multiple domains with Convolutional Neural Networks (CNNs). When training a CNN with data from all the domains, some neurons learn representations shared across several domains, while some others are effective only for a specific one. Based on this important observation, we propose a Domain Guided Dropout algorithm to improve the feature learning procedure. Experiments show the effectiveness of our pipeline and the proposed algorithm. Our

methods on the person re-identification problem outperform stateof-the-art methods on multiple datasets by large margins.

1.545 Track and Segment: An Iterative Unsupervised Approach for Video Object Proposals

[Link to Paper]

We present an unsupervised approach that generates a diverse, ranked set of bounding box and segmentation video object proposals–spatio-temporal tubes that localize the foreground objects–in an unannotated video. In contrast to previous unsupervised methods that either track regions initialized in an arbitrary frame or train a fixed model over a cluster of regions, we instead discover a set of easy-togroup instances of an object and then iteratively update its appearance model to gradually detect harder instances in temporally-adjacent frames. Our method first generates a set of spatio-temporal bounding box proposals, and then refines them to obtain pixel-wise segmentation proposals. We demonstrate state-of-the-art segmentation results on the SegTrack v2 dataset, and bounding box tracking results that perform competitively to state-of-the-art supervised tracking methods.

1.546 DisturbLabel: Regularizing CNN on the Loss Layer

[Link to Paper]

During a long period of time we are combating overfitting in the CNN training process with model regularization, including weight decay, model averaging, data augmentation, etc. In this paper, we present DisturbLabel, an extremely simple algorithm which randomly replaces a part of labels as incorrect values in each iteration. Although it seems weird to intentionally generate incorrect training labels, we show that DisturbLabel prevents the network training from over-fitting by implicitly averaging over exponentially many networks which are trained with different label sets. To the best of our knowledge, DisturbLabel serves as the first work which adds noises on the loss layer. Meanwhile, DisturbLabel cooperates well with Dropout to provide complementary regularization functions. Experiments demonstrate competitive recognition results on several popular image recognition datasets.

1.547 InterActive: Inter-Layer Activeness Propagation

[Link to Paper]

An increasing number of computer vision tasks can be tackled with deep features, which are the intermediate outputs of a pre-trained Convolutional Neural Network. Despite the astonishing performance, deep features extracted from low-level neurons are still below satisfaction, arguably because they cannot access the spatial context contained in the higher layers. In this paper, we present InterActive, a novel algorithm which computes the activeness of neurons and network connections. Activeness is propagated through a neural network in a top-down manner, carrying highlevel context and improving the descriptive power of lowlevel and mid-level neurons. Visualization indicates that neuron activeness can be interpreted as spatial-weighted neuron responses. We achieve state-of-the-art classification performance on a wide range of image datasets.

1.548 Learned Binary Spectral Shape Descriptor for 3D Shape Correspondence

[Link to Paper]

Dense 3D shape correspondence is an important problem in computer vision and computer graphics. Recently, the local shape descriptor based 3D shape correspondence approaches have been widely studied, where the local shape descriptor is a real-valued vector to characterize the geometrical structure of the shape. Different from these realvalued local shape descriptors, in this paper, we propose to learn a novel binary spectral shape descriptor with the deep neural network for 3D shape correspondence. The binary spectral shape descriptor can require less storage space and enable fast matching. First, based on the eigenvectors of the Laplace-Beltrami operator, we construct a neural network to form a nonlinear spectral representation to characterize the shape. Then, for the defined positive and negative points on the shapes, we train the constructed neural network by minimizing the errors between the outputs and their corresponding binary descriptors, minimizing the variations of the outputs of the positive points and maximizing the variations of the outputs of the negative points, simultaneously. Finally, we binarize the output of the neural network to form the binary spectral shape descriptor for shape correspondence. The proposed binary spectral shape descriptor is evaluated on the SCAPE and TOSCA 3D shape datasets for shape correspondence. The experimental results demonstrate the effectiveness of the proposed binary shape descriptor for the shape correspondence task.

1.549 Multispectral Images Denoising by Intrinsic Tensor Sparsity Regularization

[Link to Paper]

Multispectral images (MSI) can help deliver more faithful representation for real scenes than the traditional image system, and enhance the performance of many computer vision tasks. In real cases, however, an MSI is always corrupted by various noises. In this paper, we propose a new tensor-based denoising approach by fully considering two intrinsic characteristics underlying an MSI, i.e., the global correlation along spectrum (GCS) and nonlocal self-similarity across space (NSS). In specific, we construct a new tensor sparsity measure, called intrinsic tensor sparsity (ITS) measure, which encodes both sparsity insights delivered by the most typical Tucker and CANDECOMP/PARAFAC (CP) low-rank decomposition for a general tensor. Then we build a new MSI denoising model by applying the proposed ITS measure on tensors formed by non-local similar patches within the MSI. The intrinsic GCS and NSS knowledge can then be efficiently explored under the regularization of this tensor sparsity measure to finely rectify the recovery of a MSI from its corruption. A series of experiments on simulated and real MSI denoising problems show that our method outperforms all state-of-the-arts under comprehensive quantitative performance measures.

1.550 Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer

[Link to Paper]

Semantic annotations are vital for training models for object recognition, semantic segmentation or scene understanding. Unfortunately, pixelwise annotation of images at very large scale is labor-intensive and only little labeled data is available, particularly at instance level and for street scenes. In this paper, we propose to tackle this problem by lifting the semantic instance labeling task from 2D into 3D. Given reconstructions from stereo or laser data, we annotate static 3D scene elements with rough bounding primitives and develop a model which transfers this information into the image domain. We leverage our method to obtain 2D labels for a novel suburban video dataset which we have collected, resulting in 400k semantic and instance image annotations. A comparison of our method to state-ofthe-art label transfer baselines reveals that 3D information enables more efficient annotation while at the same time resulting in improved accuracy and time-coherent labels.

1.551 Logistic Boosting Regression for Label Distribution Learning

[Link to Paper]

Label Distribution Learning (LDL) is a general learning framework which includes both single label and multi-label learning as its special cases. One of the main assumptions

made in traditional LDL algorithms is the derivation of the parametric model as the maximum entropy model. While it is a reasonable assumption without additional information, there is no particular evidence supporting it in the problem of LDL. Alternatively, using a general LDL model family to approximate this parametric model can avoid the potential influence of the specific model. In order to learn this general model family, this paper uses a method called Logistic Boosting Regression (LogitBoost) which can be seen as an additive weighted function regression from the statistical viewpoint. For each step, we can fit individual weighted regression function (base learner) to realize the optimization gradually. The base learners are chosen as weighted regression tree and vector tree, which constitute two algorithms named LDLogitBoost and AOSO-LDLogitBoost in this paper. Experiments on facial expression recognition, crowd opinion prediction on movies and apparent age estimation show that LDLogitBoost and AOSO-LDLogitBoost can achieve better performance than traditional LDL algorithms as well as other LogitBoost algorithms.

1.552 Actor-Action Semantic Segmentation with Grouping Process Models

[Link to Paper]

Actor-action semantic segmentation made an important step toward advanced video understanding: what action is happening; who is performing the action; and where is the action happening in space-time. Current methods based on layered CRFs for this problem are local and unable to capture the long-ranging interactions of video parts. We propose a new model that combines the labeling CRF with a supervoxel hierarchy, where supervoxels at various scales provide cues for possible groupings of nodes in the CRF to encourage adaptive and long-ranging interactions. The new model defines a dynamic and continuous process of information exchange: the CRF influences what supervoxels in the hierarchy are active, and these active supervoxels, in turn, affect the connectivities in the CRF; we hence call it a grouping process model. By further incorporating the video-level recognition, the proposed method achieves a large margin of 60% relative improvement over the state of the art on the recent A2D large-scale video labeling dataset, which demonstrates the effectiveness of our modeling.

1.553 Deep Interactive Object Selection

[Link to Paper]

Interactive object selection is a very important research problem and has many applications. Previous algorithms require substantial user interactions to estimate the foreground and background distributions. In this paper, we present a novel deep-learning-based algorithm which has much better understanding of objectness and can reduce user interactions

to just a few clicks. Our algorithm transforms user-provided positive and negative clicks into two Euclidean distance maps which are then concatenated with the RGB channels of images to compose (image, user interactions) pairs. We generate many of such pairs by combining several random sampling strategies to model users' click patterns and use them to finetune deep Fully Convolutional Networks (FCNs). Finally the output probability maps of our FCN-8s model is integrated with graph cut optimization to refine the boundary segments. Our model is trained on the PASCAL segmentation dataset and evaluated on other datasets with different object classes. Experimental results on both seen and unseen objects demonstrate that our algorithm has a good generalization ability and is superior to all existing interactive object selection approaches.

1.554   Discriminatively Embedded K-Means for Multi-view Clustering

[Link to Paper]

In real world applications, more and more data, for example, image/video data, are high dimensional and represented by multiple views which describe different perspectives of the data. Efficiently clustering such data is a challenge. To address this problem, this paper proposes a novel multi-view clustering method called Discriminatively Embedded K-Means (DEKM), which embeds the synchronous learning of multiple discriminative subspaces into multiview K-Means clustering to construct a unified framework, and adaptively control the intercoordinations between these subspaces simultaneously. In this framework, we firstly design a weighted multi-view Linear Discriminant Analysis (LDA), and then develop an unsupervised optimization scheme to alternatively learn the common clustering indicator, multiple discriminative subspaces and weights for heterogeneous features with convergence. Comprehensive evaluations on three benchmark datasets and comparisons with several state-of-the-art multi-view clustering algorithms demonstrate the superiority of the proposed work.

1.555   MSR-VTT: A Large Video Description Dataset for Bridging Video and Language

[Link to Paper]

While there has been increasing interest in the task of describing video with natural language, current computer vision algorithms are still severely limited in terms of the variability and complexity of the videos and their associated language that they can recognize. This is in part due to the simplicity of current benchmarks, which mostly focus on specific fine-grained domains with limited videos and simple descriptions. While researchers have

provided several benchmark datasets for image captioning, we are not aware of any large-scale video description dataset with comprehensive categories yet diverse video content. In this paper we present MSR-VTT (standing for "MSRVideo to Text") which is a new large-scale video benchmark for video understanding, especially the emerging task of translating video to text. This is achieved by collecting 257 popular queries from a commercial video search engine, with 118 videos for each query. In its current version, MSR-VTT provides 10K web video clips with 41.2 hours and 200K clip-sentence pairs in total, covering the most comprehensive categories and diverse visual content, and representing the largest dataset in terms of sentence and vocabulary. Each clip is annotated with about 20 natural sentences by 1,327 AMT workers. We present a detailed analysis of MSR-VTT in comparison to a complete set of existing datasets, together with a summarization of different state-of-the-art video-to-text approaches. We also provide an extensive evaluation of these approaches on this dataset, showing that the hybrid Recurrent Neural Networkbased approach, which combines single-frame and motion representations with soft-attention pooling strategy, yields the best generalization capability on MSR-VTT.

1.556 Multi-view People Tracking via Hierarchical Trajectory Composition

[Link to Paper]

This paper presents a hierarchical composition approach for multi-view object tracking. The key idea is to adaptively exploit multiple cues in both 2D and 3D, e.g., ground occupancy consistency, appearance similarity, motion coherence etc., which are mutually complementary while tracking the humans of interests over time. While feature online selection has been extensively studied in the past literature, it remains unclear how to effectively schedule these cues for the tracking purpose especially when encountering various challenges, e.g. occlusions, conjunctions, and appearance variations. To do so, we propose a hierarchical composition model and re-formulate multi-view multi-object tracking as a problem of compositional structure optimization. We setup a set of composition criteria, each of which corresponds to one particular cue. The hierarchical composition process is pursued by exploiting different criteria, which impose constraints between a graph node and its offsprings in the hierarchy. We learn the composition criteria using MLE on annotated data and efficiently construct the hierarchical graph by an iterative greedy pursuit algorithm. In the experiments, we demonstrate superior performance of our approach on three public datasets, one of which is newly created by us to test various challenges in multi-view multi-object tracking.

1.557 Accumulated Stability Voting: A Robust Descriptor from Descriptors of Multiple Scales

[Link to Paper]

This paper proposes a novel local descriptor through accumulated stability voting (ASV). The stability of feature dimensions is measured by their differences across scales. To be more robust to noise, the stability is further quantized by thresholding. The principle of maximum entropy is utilized for determining the best thresholds for maximizing discriminant power of the resultant descriptor. Accumulating stability renders a real-valued descriptor and it can be converted into a binary descriptor by an additional thresholding process. The real-valued descriptor attains high matching accuracy while the binary descriptor makes a good compromise between storage and accuracy. Our descriptors are simple yet effective, and easy to implement. In addition, our descriptors require no training. Experiments on popular benchmarks demonstrate the effectiveness of our descriptors and their superiority to the state-of-the-art descriptors.

1.558 CRAFT Objects from Images

[Link to Paper]

Object detection is a fundamental problem in image understanding. One popular solution is the R-CNN framework [15] and its fast versions [14, 27]. They decompose the object detection problem into two cascaded easier tasks: 1) generating object proposals from images, 2) classifying proposals into various object categories. Despite that we are handling with two relatively easier tasks, they are not solved perfectly and there's still room for improvement. In this paper, we push the "divide and conquer" solution even further by dividing each task into two sub-tasks. We call the proposed method "CRAFT" (Cascade Regionproposal-network And FasT-rcnn), which tackles each task with a carefully designed network cascade. We show that the cascade structure helps in both tasks: in proposal generation, it provides more compact and better localized object proposals; in object classification, it reduces false positives (mainly between ambiguous categories) by capturing both inter- and intra-category variances. CRAFT achieves consistent and considerable improvement over the state-ofthe-art on object detection benchmarks like PASCAL VOC 07/12 and ILSVRC.

1.559 Efficient 3D Room Shape Recovery from a Single Panorama

[Link to Paper]

We propose a method to recover the shape of a 3D room from a full-view indoor panorama. Our algorithm can automatically infer a 3D shape from a collection of partially oriented superpixel facets and line segments. The core part of the algorithm is a constraint graph, which includes lines and superpixels as vertices, and encodes their geometric relations as edges. A novel approach is proposed to perform 3D reconstruction based on the constraint graph by solving all the geometric constraints as constrained linear least-squares. The selected constraints used for reconstruction are identified using an occlusion detection method with a Markov random field. Experiments show that our method can recover room shapes that can not be addressed by previous approaches. Our method is also efficient, that is, the inference time for each panorama is less than 1 minute.

1.560 End-to-End Learning of Deformable Mixture of Parts and Deep Convolutional Neural Networks for Human Pose Estimation

[Link to Paper]

Recently, Deep Convolutional Neural Networks (DCNNs) have been applied to the task of human pose estimation, and have shown its potential of learning better feature representations and capturing contextual relationships. However, it is difficult to incorporate domain prior knowledge such as geometric relationships among body parts into DCNNs. In addition, training DCNN-based body part detectors without consideration of global body joint consistency introduces ambiguities, which increases the complexity of training. In this paper, we propose a novel end-to-end framework for human pose estimation that combines DCNNs with the expressive deformable mixture of parts. We explicitly incorporate domain prior knowledge into the framework, which greatly regularizes the learning process and enables the flexibility of our framework for loopy models or tree-structured models. The effectiveness of jointly learning a DCNN with a deformable mixture of parts model is evaluated through intensive experiments on several widely used benchmarks. The proposed approach significantly improves the performance compared with state-of-the-art approaches, especially on benchmarks with challenging articulations.

1.561 Exploit All the Layers: Fast and Accurate CNN Object Detector with Scale Dependent Pooling and Cascaded Rejection Classifiers

[Link to Paper]

In this paper, we investigate two new strategies to detect objects accurately and efficiently using deep convolutional neural network: 1) scale-dependent pooling and 2) layerwise cascaded rejection classifiers. The scale-dependent pooling (SDP) improves detection accuracy by exploiting appropriate convolutional features depending on the scale of can-

didate object proposals. The cascaded rejection classifiers (CRC) effectively utilize convolutional features and eliminate negative object proposals in a cascaded manner, which greatly speeds up the detection while maintaining high accuracy. In combination of the two, our method achieves significantly better accuracy compared to other state-of-the-arts in three challenging datasets, PASCAL object detection challenge, KITTI object detection benchmark and newly collected Inner-city dataset, while being more efficient.

1.562 Exploit Bounding Box Annotations for Multi-label Object Recognition

[Link to Paper]

Convolutional neural networks (CNNs) have shown great performance as general feature representations for object recognition applications. However, for multi-label images that contain multiple objects from different categories, scales and locations, global CNN features are not optimal. In this paper, we incorporate local information to enhance the feature discriminative power. In particular, we first extract object proposals from each image. With each image treated as a bag and object proposals extracted from it treated as instances, we transform the multi-label recognition problem into a multi-class multi-instance learning problem. Then, in addition to extracting the typical CNN feature representation from each proposal, we propose to make use of ground-truth bounding box annotations (strong labels) to add another level of local information by using nearest-neighbor relationships of local regions to form a multi-view pipeline. The proposed multi-view multi-instance framework utilizes both weak and strong labels effectively, and more importantly it has the generalization ability to even boost the performance of unseen categories by partial strong labels from other categories. Our framework is extensively compared with state-of-the-art handcrafted feature based methods and CNN based methods on two multi-label benchmark datasets. The experimental results validate the discriminative power and the generalization ability of the proposed framework. With strong labels, our framework is able to achieve state-of-the-art results in both datasets.

1.563 Joint Unsupervised Learning of Deep Representations and Image Clusters

[Link to Paper]

In this paper, we propose a recurrent framework for joint unsupervised learning of deep representations and image clusters. In our framework, successive operations in a clustering algorithm are expressed as steps in a recurrent process, stacked on top of representations output by a Convolutional Neural Network (CNN). During training, image clusters and representations are updated jointly: image clustering is conducted in the forward pass,

while representation learning in the backward pass. Our key idea behind this framework is that good representations are beneficial to image clustering and clustering results provide supervisory signals to representation learning. By integrating two processes into a single model with a unified weighted triplet loss function and optimizing it endto-end, we can obtain not only more powerful representations, but also more precise image clusters. Extensive experiments show that our method outperforms the stateof-the-art on image clustering across a variety of image datasets. Moreover, the learned representations generalize well when transferred to other tasks. The source code can be downloaded from https://github.com/ jwyang/joint-unsupervised-learning.

1.564 Multivariate Regression on the Grassmannian for Predicting Novel Domains

[Link to Paper]

We study the problem of predicting how to recognise visual objects in novel domains with neither labelled nor unlabelled training data. Domain adaptation is now an established research area due to its value in ameliorating the issue of domain shift between train and test data. However, it is conventionally assumed that domains are discrete entities, and that at least unlabelled data is provided in testing domains. In this paper, we consider the case where domains are parametrised by a vector of continuous values (e.g., time, lighting or view angle). We aim to use such domain metadata to predict novel domains for recognition. This allows a recognition model to be pre-calibrated for a new domain in advance (e.g., future time or view angle) without waiting for data collection and re-training. We achieve this by posing the problem as one of multivariate regression on the Grassmannian, where we regress a domain's subspace (point on the Grassmannian) against an independent vector of domain parameters. We derive two novel methodologies to achieve this challenging task: a direct kernel regression from RM G, and an indirect method with better extrapolation properties. We evaluate our methods on two crossdomain visual recognition benchmarks, where they perform close to the upper bound of full data domain adaptation. This demonstrates that data is not necessary for domain adaptation if a domain can be parametrically described.

1.565 Object Contour Detection with a Fully Convolutional Encoder-Decoder Network

[Link to Paper]

We develop a deep learning algorithm for contour detection with a fully convolutional encoder-decoder network. Different from previous low-level edge detection, our algo-

rithm focuses on detecting higher-level object contours. Our network is trained end-to-end on PASCAL VOC with refined ground truth from inaccurate polygon annotations, yielding much higher precision in object contour detection than previous methods. We find that the learned model generalizes well to unseen object classes from the same supercategories on MS COCO and can match state-of-the-art edge detection on BSDS500 with fine-tuning. By combining with the multiscale combinatorial grouping algorithm, our method can generate high-quality segmented object proposals, which significantly advance the state-of-the-art on PASCAL VOC (improving average recall from 0.62 to 0.67) with a relatively small amount of candidates (1660 per image).

1.566 On Benefits of Selection Diversity via Bilevel Exclusive Sparsity

[Link to Paper]

Sparse feature (dictionary) selection is critical for various tasks in computer vision, machine learning, and pattern recognition to avoid overfitting. While extensive research efforts have been conducted on feature selection using sparsity and group sparsity, we note that there has been a lack of development on applications where there is a particular preference on diversity. That is, the selected features are expected to come from different groups or categories. This diversity preference is motivated from many real-world applications such as advertisement recommendation, privacy image classification, and design of survey. In this paper, we proposed a general bilevel exclusive sparsity formulation to pursue the diversity by restricting the overall sparsity and the sparsity in each group. To solve the proposed formulation that is NP hard in general, a heuristic procedure is proposed. The main contributions in this paper include: 1) A linear convergence rate is established for the proposed algorithm; 2) The provided theoretical error bound improves the approaches such as L1 norm and L0 types methods which only use the overall sparsity and the quantitative benefits of using the diversity sparsity is provided. To the best of our knowledge, this is the first work to show the theoretical benefits of using the diversity sparsity; 3) Extensive empirical studies are provided to validate the proposed formulation, algorithm, and theory.

1.567 Robust Optical Flow Estimation of Double-Layer Images under Transparency or Reflection

[Link to Paper]

This paper deals with a challenging, frequently encountered, yet not properly investigated problem in two-frame optical flow estimation. That is, the input frames are compounds of two imaging layers one desired background layer of the scene, and one distracting,

possibly moving layer due to transparency or reflection. In this situation, the conventional brightness constancy constraint the cornerstone of most existing optical flow methods will no longer be valid. In this paper, we propose a robust solution to this problem. The proposed method performs both optical flow estimation, and image layer separation. It exploits a generalized double-layer brightness consistency constraint connecting these two tasks, and utilizes the priors for both of them. Experiments on both synthetic data and real images have confirmed the efficacy of the proposed method. To the best of our knowledge, this is the first attempt towards handling generic optical flow fields of two-frame images containing transparency or reflection.

1.568   Semantic Filtering

[Link to Paper]

Edge-preserving image operations aim at smoothing an image without blurring the edges. Many excellent edgepreserving filtering techniques have been proposed recently to reduce the computational complexity or/and separate different scale structures. They normally adopt a userselected scale measurement to control the detail/texture smoothing. However, natural photos contain objects of different sizes which cannot be described by a single scale measurement. On the other hand, edge/contour detection/analysis is closely related to edge-preserving filtering and has achieved significant progress recently. Nevertheless, most of the state-of-the-art filtering techniques ignore the success in this area. Inspired by the fact that learning-based edge detectors/classifiers significantly outperform traditional manually-designed detectors, this paper proposes a learning-based edge-preserving filtering technique. It synergistically combines the efficiency of the recursive filter and the effectiveness of the recent edge detector for scale-aware edge-preserving filtering. Unlike previous filtering methods, the propose filter can efficiently extract subjectively-meaningful structures from natural scenes containing multiple-scale objects.

1.569   Stacked Attention Networks for Image Question Answering

[Link to Paper]

This paper presents stacked attention networks (SANs) that learn to answer natural language questions from images. SANs use semantic representation of a question as query to search for the regions in an image that are related to the answer. We argue that image question answering (QA) often requires multiple steps of reasoning. Thus, we develop a multiple-layer SAN in which we query an image multiple times to infer the answer progressively. Experiments conducted on four image QA data sets demonstrate that the proposed SANs significantly outperform previous state-of-the-art approaches. The visualization of

the attention layers illustrates the progress that the SAN locates the relevant visual clues that lead to the answer of the question layer-by-layer.

1.570 WIDER FACE: A Face Detection Benchmark

[Link to Paper]

Face detection is one of the most studied topics in the computer vision community. Much of the progresses have been made by the availability of face detection benchmark datasets. We show that there is a gap between current face detection performance and the real world requirements. To facilitate future face detection research, we introduce the WIDER FACE dataset1, which is 10 times larger than existing datasets. The dataset contains rich annotations, including occlusions, poses, event categories, and face bounding boxes. Faces in the proposed dataset are extremely challenging due to large variations in scale, pose and occlusion, as shown in Fig. 1. Furthermore, we show that WIDER FACE dataset is an effective training source for face detection. We benchmark several representative detection systems, providing an overview of state-of-the-art performance and propose a solution to deal with large scale variation. Finally, we discuss common failure cases that worth to be further investigated.

1.571 Highlight Detection with Pairwise Deep Ranking for First-Person Video Summarization

[Link to Paper]

The emergence of wearable devices such as portable cameras and smart glasses makes it possible to record life logging first-person videos. Browsing such long unstructured videos is time-consuming and tedious. This paper studies the discovery of moments of user's major or special interest (i.e., highlights) in a video, for generating the summarization of first-person videos. Specifically, we propose a novel pairwise deep ranking model that employs deep learning techniques to learn the relationship between highlight and non-highlight video segments. A two-stream network structure by representing video segments from complementary information on appearance of video frames and temporal dynamics across frames is developed for video highlight detection. Given a long personal video, equipped with the highlight detection model, a highlight score is assigned to each segment. The obtained highlight segments are applied for summarization in two ways: video timelapse and video skimming. The former plays the highlight (non-highlight) segments at low (high) speed rates, while the latter assembles the sequence of segments with the highest scores. On 100 hours of first-person videos for 15 unique sports categories, our highlight detection achieves the improvement over the state-of-the-art RankSVM method by 10.5% in terms of

266

accuracy. Moreover, our approaches produce video summary with better quality by a user study from 35 human subjects.

1.572 A Dual-Source Approach for 3D Pose Estimation from a Single Image

[Link to Paper]

One major challenge for 3D pose estimation from a single RGB image is the acquisition of sufficient training data. In particular, collecting large amounts of training data that contain unconstrained images and are annotated with accurate 3D poses is infeasible. We therefore propose to use two independent training sources. The first source consists of images with annotated 2D poses and the second source consists of accurate 3D motion capture data. To integrate both sources, we propose a dual-source approach that combines 2D pose estimation with efficient and robust 3D pose retrieval. In our experiments, we show that our approach achieves state-of-the-art results and is even competitive when the skeleton structure of the two sources differ substantially.

1.573 Situation Recognition: Visual Semantic Role Labeling for Image Understanding

[Link to Paper]

This paper introduces situation recognition, the problem of producing a concise summary of the situation an image depicts including: (1) the main activity (e.g., clipping), (2) the participating actors, objects, substances, and locations (e.g., man, shears, sheep, wool, and field) and most importantly (3) the roles these participants play in the activity (e.g., the man is clipping, the shears are his tool, the wool is being clipped from the sheep, and the clipping is in a field). We use FrameNet, a verb and role lexicon developed by linguists, to define a large space of possible situations and collect a large-scale dataset containing over 500 activities, 1,700 roles, 11,000 objects, 125,000 images, and 200,000 unique situations. We also introduce structured prediction baselines and show that, in activity-centric images, situation-driven prediction of objects and activities outperforms independent object and activity recognition.

1.574 End-to-end Learning of Action Detection from Frame Glimpses in Videos

[Link to Paper]

In this work we introduce a fully end-to-end approach for action detection in videos that learns to directly predict the temporal bounds of actions. Our intuition is that the process of detecting actions is naturally one of observation and refinement: observing moments in video, and refining hypotheses about when an action is occurring. Based on this insight, we formulate our model as a recurrent neural network-based agent that interacts with a video over time. The agent observes video frames and decides both where to look next and when to emit a prediction. Since backpropagation is not adequate in this non-differentiable setting, we use REINFORCE to learn the agent's decision policy. Our model achieves state-of-the-art results on the THUMOS'14 and ActivityNet datasets while observing only a fraction (2% or less) of the video frames.

1.575   Automatic Fence Segmentation in Videos of Dynamic Scenes

[Link to Paper]

We present a fully automatic approach to detect and segment fence-like occluders from a video clip. Unlike previous approaches that usually assume either static scenes or cameras, our method is capable of handling both dynamic scenes and moving cameras. Under a bottom-up framework, it first clusters pixels into coherent groups using color and motion features. These pixel groups are then analyzed in a fully connected graph, and labeled as either fence or non-fence using graph-cut optimization. Finally, we solve a dense Conditional Random Filed (CRF) constructed from multiple frames to enhance both spatial accuracy and temporal coherence of the segmentation. Once segmented, one can use existing hole-filling methods to generate a fencefree output. Extensive evaluation suggests that our method outperforms previous automatic and interactive approaches on complex examples captured by mobile devices.

1.576   Learning to Assign Orientations to Feature Points

[Link to Paper]

We show how to train a Convolutional Neural Network to assign a canonical orientation to feature points given an image patch centered on the feature point. Our method improves feature point matching upon the state-of-the art and can be used in conjunction with any existing rotation sensitive descriptors. To avoid the tedious and almost impossible task of finding a target orientation to learn, we propose to use Siamese networks which implicitly find the optimal orientations during training. We also propose a new type of activation function for Neural Networks that generalizes the popular ReLU, maxout, and PReLU activation functions. This novel activation performs better for our task. We validate the effectiveness of our method extensively with four existing datasets, including two

non-planar datasets, as well as our own dataset. We show that we outperform the state-of-the-art without the need of retraining for each dataset. Reference SIFT Our method Reference SIFT Our method

1.577 Kernel Sparse Subspace Clustering on Symmetric Positive Definite Manifolds

[Link to Paper]

(Abstract could not be extracted from PDF)

1.578 Recognizing Micro-Actions and Reactions from Paired Egocentric Videos

[Link to Paper]

We aim to understand the dynamics of social interactions between two people by recognizing their actions and reactions using a head-mounted camera. Our work will impact several first-person vision tasks that need the detailed understanding of social interactions, such as automatic video summarization of group events and assistive systems. To recognize micro-level actions and reactions, such as slight shifts in attention, subtle nodding, or small hand actions, where only subtle body motion is apparent, we propose to use paired egocentric videos recorded by two interacting people. We show that the first-person and second-person points-of-view features of two people, enabled by paired egocentric videos, are complementary and essential for reliably recognizing micro-actions and reactions. We also build a new dataset of dyadic (two-persons) interactions that comprises more than 1000 pairs of egocentric videos to enable systematic evaluations on the task of micro-action and reaction recognition.

1.579 Visual Path Prediction in Complex Scenes with Crowded Moving Objects

[Link to Paper]

This paper proposes a novel path prediction algorithm for progressing one step further than the existing works focusing on single target path prediction. In this paper, we consider moving dynamics of co-occurring objects for path prediction in a scene that includes crowded moving objects. To solve this problem, we first suggest a two-layered probabilistic model to find major movement patterns and their cooccurrence tendency. By utilizing the unsupervised learning results from the model, we present an algorithm to find the future location of any target object. Through extensive qualitative/quantitative experiments, we

show that our algorithm can find a plausible future path in complex scenes with a large number of moving objects.

1.580 Online Multi-Object Tracking via Structural Constraint Event Aggregation

[Link to Paper]

Multi-object tracking (MOT) becomes more challenging when objects of interest have similar appearances. In that case, the motion cues are particularly useful for discriminating multiple objects. However, for online 2D MOT in scenes acquired from moving cameras, observable motion cues are complicated by global camera movements and thus not always smooth or predictable. To deal with such unexpected camera motion for online 2D MOT, a structural motion constraint between objects has been utilized thanks to its robustness to camera motion. In this paper, we propose a new data association method that effectively exploits structural motion constraints in the presence of large camera motion. In addition, to further improve the robustness of data association against mis-detections and false positives, a novel event aggregation approach is developed to integrate structural constraints in assignment costs for online MOT. Experimental results on a large number of datasets demonstrate the effectiveness of the proposed algorithm for online 2D MOT.

1.581 Image Captioning with Semantic Attention

[Link to Paper]

Automatically generating a natural language description of an image has attracted interests recently both because of its importance in practical applications and because it connects two major artificial intelligence fields: computer vision and natural language processing. Existing approaches are either top-down, which start from a gist of an image and convert it into words, or bottom-up, which come up with words describing various aspects of an image and then combine them. In this paper, we propose a new algorithm that combines both approaches through a model of semantic attention. Our algorithm learns to selectively attend to semantic concept proposals and fuse them into hidden states and outputs of recurrent neural networks. The selection and fusion form a feedback connecting the top-down and bottom-up computation. We evaluate our algorithm on two public benchmarks: Microsoft COCO and Flickr30K. Experimental results show that our algorithm significantly outperforms the state-of-the-art approaches consistently across different evaluation metrics.

1.582 Oracle Based Active Set Algorithm for Scalable Elastic Net Subspace Clustering

[Link to Paper]

State-of-the-art subspace clustering methods are based on expressing each data point as a linear combination of other data points while regularizing the matrix of coefficients with 1, 2 or nuclear norms. 1 regularization is guaranteed to give a subspace-preserving affinity (i.e., there are no connections between points from different subspaces) under broad theoretical conditions, but the clusters may not be connected. 2 and nuclear norm regularization often improve connectivity, but give a subspace-preserving affinity only for independent subspaces. Mixed 1, 2 and nuclear norm regularizations offer a balance between the subspacepreserving and connectedness properties, but this comes at the cost of increased computational complexity. This paper studies the geometry of the elastic net regularizer (a mixture of the 1 and 2 norms) and uses it to derive a provably correct and scalable active set method for finding the optimal coefficients. Our geometric analysis also provides a theoretical justification and a geometric interpretation for the balance between the connectedness (due to 2 regularization) and subspace-preserving (due to 1 regularization) properties for elastic net subspace clustering. Our experiments show that the proposed active set method not only achieves state-of-the-art clustering performance, but also efficiently handles large-scale datasets.

1.583 Scalable Sparse Subspace Clustering by Orthogonal Matching Pursuit

[Link to Paper]

Subspace clustering methods based on 1, 2 or nuclear norm regularization have become very popular due to their simplicity, theoretical guarantees and empirical success. However, the choice of the regularizer can greatly impact both theory and practice. For instance, 1 regularization is guaranteed to give a subspace-preserving affinity (i.e., there are no connections between points from different subspaces) under broad conditions (e.g., arbitrary subspaces and corrupted data). However, it requires solving a large scale convex optimization problem. On the other hand, 2 and nuclear norm regularization provide efficient closed form solutions, but require very strong assumptions to guarantee a subspace-preserving affinity, e.g., independent subspaces and uncorrupted data. In this paper we study a subspace clustering method based on orthogonal matching pursuit. We show that the method is both computationally efficient and guaranteed to give a subspace-preserving affinity under broad conditions. Experiments on synthetic data verify our theoretical analysis, and applications in handwritten digit and face clustering show that our approach

achieves the best trade off between accuracy and efficiency. Moreover, our approach is the first one to handle 100,000 data points.

1.584 Top-push Video-based Person Re-identification

[Link to Paper]

Most existing person re-identification (re-id) models focus on matching still person images across disjoint camera views. Since only limited information can be exploited from still images, it is hard (if not impossible) to overcome the occlusion, pose and camera-view change, and lighting variation problems. In comparison, video-based re-id methods can utilize extra space-time information, which contains much more rich cues for matching to overcome the mentioned problems. However, we find that when using video-based representation, some inter-class difference can be much more obscure than the one when using still-imagebased representation, because different people could not only have similar appearance but also have similar motions and actions which are hard to align. To solve this problem, we propose a top-push distance learning model (TDL), in which we integrate a top-push constrain for matching video features of persons. The top-push constraint enforces the optimization on top-rank matching in re-id, so as to make the matching model more effective towards selecting more discriminative features to distinguish different persons. Our experiments show that the proposed video-based reid framework outperforms the state-of-the-art video-based re-id methods.

1.585 Groupwise Tracking of Crowded Similar-Appearance Targets from Low-Continuity Image Sequences

[Link to Paper]

Automatic tracking of large-scale crowded targets are of particular importance in many applications, such as crowded people/vehicle tracking in video surveillance, fiber tracking in materials science, and cell tracking in biomedical imaging. This problem becomes very challenging when the targets show similar appearance and the interslice/inter-frame continuity is low due to sparse sampling, camera motion and target occlusion. The main challenge comes from the step of association which aims at matching the predictions and the observations of the multiple targets. In this paper we propose a new groupwise method to explore the target group information and employ the within-group correlations for association and tracking. In particular, the within-group association is modeled by a nonrigid 2D ThinPlate transform and a sequence of group shrinking, group growing and group merging operations are then developed to refine the composition of each group. We apply

the proposed method to track large-scale fibers from microscopy material images and compare its performance against several other multi-target tracking methods. We also apply the proposed method to track crowded people from videos with poor inter-frame continuity.

1.586 Learning Reconstruction-based Remote Gaze Estimation

[Link to Paper]

It is a challenging problem to accurately estimate gazes from low-resolution eye images that do not provide fine and detailed features for eyes. Existing methods attempt to establish the mapping between the visual appearance space to the gaze space. Different from the direct regression approach, the reconstruction-based approach represents appearance and gaze via local linear reconstruction in their own spaces. A common treatment is to use the same local reconstruction in the two spaces, i.e., the reconstruction weights in the appearance space are transferred to the gaze space for gaze reconstruction. However, this questionable treatment is taken for granted but has never been justified, leading to significant errors in gaze estimation. This paper is focused on the study of this fundamental issue. It shows that the distance metric in the appearance space needs to be adjusted, before the same reconstruction can be used. A novel method is proposed to learn the metric, such that the affinity structure of the appearance space under this new metric is as close as possible to the affinity structure of the gaze space under the normal Euclidean metric. Furthermore, the local affinity structure invariance is utilized to further regularize the solution to the reconstruction weights, so as to obtain a more robust and accurate solution. Effectiveness of the proposed method is validated and demonstrated through extensive experiments on different subjects.

1.587 Sketch Me That Shoe

[Link to Paper]

We investigate the problem of fine-grained sketch-based image retrieval (SBIR), where free-hand human sketches are used as queries to perform instance-level retrieval of images. This is an extremely challenging task because (i) visual comparisons not only need to be fine-grained but also executed cross-domain, (ii) free-hand (finger) sketches are highly abstract, making fine-grained matching harder, and most importantly (iii) annotated cross-domain sketch-photo datasets required for training are scarce, challenging many state-of-the-art machine learning techniques. In this paper, for the first time, we address all these challenges, providing a step towards the capabilities that would underpin a commercial

273

sketch-based image retrieval application. We introduce a new database of 1,432 sketch-photo pairs from two categories with 32,000 fine-grained triplet ranking annotations. We then develop a deep tripletranking model for instance-level SBIR with a novel data augmentation and staged pre-training strategy to alleviate the issue of insufficient fine-grained training data. Extensive experiments are carried out to contribute a variety of insights into the challenges of data sufficiency and over-fitting avoidance when training deep networks for finegrained cross-domain ranking tasks.

1.588 The Solution Path Algorithm for Identity-Aware Multi-Object Tracking

[Link to Paper]

We propose an identity-aware multi-object tracker based on the solution path algorithm. Our tracker not only produces identity-coherent trajectories based on cues such as face recognition, but also has the ability to pinpoint potential tracking errors. The tracker is formulated as a quadratic optimization problem with 0 norm constraints, which we propose to solve with the solution path algorithm. The algorithm successively solves the same optimization problem but under different p norm constraints, where p gradually decreases from 1 to 0. Inspired by the success of the solution path algorithm in various machine learning tasks, this strategy is expected to converge to a better local minimum than directly minimizing the hardly solvable 0 norm or the roughly approximated 1 norm constraints. Furthermore, the acquired solution path complies with the "decision making process" of the tracker, which provides more insight to locating potential tracking errors. Experiments show that not only is our proposed tracker effective, but also the solution path enables automatic pinpointing of potential tracking failures, which can be readily utilized in an active learning framework to improve identity-aware multi-object tracking.

1.589 Video Paragraph Captioning Using Hierarchical Recurrent Neural Networks

[Link to Paper]

We present an approach that exploits hierarchical Recurrent Neural Networks (RNNs) to tackle the video captioning problem, i.e., generating one or multiple sentences to describe a realistic video. Our hierarchical framework contains a sentence generator and a paragraph generator. The sentence generator produces one simple short sentence that describes a specific short video interval. It exploits both temporal- and spatial-attention mechanisms to selectively focus on visual elements during generation. The paragraph generator captures the inter-sentence dependency by taking as input the sentential embedding produced by

the sentence generator, combining it with the paragraph history, and outputting the new initial state for the sentence generator. We evaluate our approach on two large-scale benchmark datasets: YouTubeClips and TACoS-MultiLevel. The experiments demonstrate that our approach significantly outperforms the current state-of-the-art methods with BLEU@4 scores 0.499 and 0.305 respectively.

1.590 Temporal Action Localization with Pyramid of Score Distribution Features

[Link to Paper]

We investigate the feature design and classification architectures in temporal action localization. This application focuses on detecting and labeling actions in untrimmed videos, which brings more challenge than classifying presegmented videos. The major difficulty for action localization is the uncertainty of action occurrence and utilization of information from different scales. Two innovations are proposed to address this issue. First, we propose a Pyramid of Score Distribution Feature (PSDF) to capture the motion information at multiple resolutions centered at each detection window. This novel feature mitigates the influence of unknown action position and duration, and shows significant performance gain over previous detection approaches. Second, inter-frame consistency is further explored by incorporating PSDF into the state-of-the-art Recurrent Neural Networks, which gives additional performance gain in detecting actions in temporally untrimmed videos. We tested our action localization framework on the THUMOS'15 and MPII Cooking Activities Dataset, both of which show a large performance improvement over previous attempts.

1.591 Joint Unsupervised Deformable Spatio-Temporal Alignment of Sequences

[Link to Paper]

Typically, the problems of spatial and temporal alignment of sequences are considered disjoint. That is, in order to align two sequences, a methodology that (non)-rigidly aligns the images is first applied, followed by temporal alignment of the obtained aligned images. In this paper, we propose the first, to the best of our knowledge, methodology that can jointly spatio-temporally align two sequences, which display highly deformable texture-varying objects. We show that by treating the problems of deformable spatial and temporal alignment jointly, we achieve better results than considering the problems independent. Furthermore, we show that deformable spatio-temporal alignment of faces can be performed in an unsupervised manner (i.e., without employing face trackers or building person-specific deformable models).

1.592 Metric Learning as Convex Combinations of Local Models with Generalization Guarantees

[Link to Paper]

Over the past ten years, metric learning allowed the improvement of numerous machine learning approaches that manipulate distances or similarities. In this field, local metric learning has been shown to be very efficient, especially to take into account non linearities in the data and better capture the peculiarities of the application of interest. However, it is well known that local metric learning (i) can entail overfitting and (ii) face difficulties to compare two instances that are assigned to two different local models. In this paper, we address these two issues by introducing a novel metric learning algorithm that linearly combines local models (C2LM). Starting from a partition of the space in regions and a model (a score function) for each region, C2LM defines a metric between points as a weighted combination of the models. A weight vector is learned for each pair of regions, and a spatial regularization ensures that the weight vectors evolve smoothly and that nearby models are favored in the combination. The proposed approach has the particularity of working in a regression setting, of working implicitly at different scales, and of being generic enough so that it is applicable to similarities and distances. We prove theoretical guarantees of the approach using the framework of algorithmic robustness. We carry out experiments with datasets using both distances (perceptual color distances, using Mahalanobis-like distances) and similarities (semantic word similarities, using bilinear forms), showing that C2LM consistently improves regression accuracy even in the case where the amount of training data is small.

1.593 Detecting Vanishing Points using Global Image Context in a Non-Manhattan World

[Link to Paper]

We propose a novel method for detecting horizontal vanishing points and the zenith vanishing point in man-made environments. The dominant trend in existing methods is to first find candidate vanishing points, then remove outliers by enforcing mutual orthogonality. Our method reverses this process: we propose a set of horizon line candidates and score each based on the vanishing points it contains. A key element of our approach is the use of global image context, extracted with a deep convolutional network, to constrain the set of candidates under consideration. Our method does not make a Manhattan-world assumption and can operate effectively on scenes with only a single horizontal vanishing point. We evaluate our approach on three benchmark datasets and achieve state-ofthe-art performance on each. In addition, our approach is significantly faster than the previous best method.

1.594 Beyond F-formations: Determining Social Involvement in Free Standing Conversing Groups from Static Images

[Link to Paper]

In this paper, we present the first attempt to analyse differing levels of social involvement in free standing conversing groups (or the so-called F-formations) from static images. In addition, we enrich state-of-the-art F-formation modelling by learning a frustum of attention that accounts for the spatial context. That is, F-formation configurations vary with respect to the arrangement of furniture and the non-uniform crowdedness in the space during mingling scenarios. The majority of prior works have considered the labelling of conversing group as an objective task, requiring only a single annotator. However, we show that by embracing the subjectivity of social involvement, we not only generate a richer model of the social interactions in a scene but also significantly improve F-formation detection. We carry out extensive experimental validation of our proposed approach by collecting a novel set of multi-annotator labels of involvement on the publicly available Idiap Poster Data; the only multi-annotator labelled database of free standing conversing groups that is currently available.

1.595 Collaborative Quantization for Cross-Modal Similarity Search

[Link to Paper]

Cross-modal similarity search is a problem about designing a search system supporting querying across content modalities, e.g., using an image to search for texts or using a text to search for images. This paper presents a compact coding solution for efficient search, with a focus on the quantization approach which has already shown the superior performance over the hashing solutions in the single-modal similarity search. We propose a cross-modal quantization approach, which is among the early attempts to introduce quantization into cross-modal search. The major contribution lies in jointly learning the quantizers for both modalities through aligning the quantized representations for each pair of image and text belonging to a document. In addition, our approach simultaneously learns the common space for both modalities in which quantization is conducted to enable efficient and effective search using the Euclidean distance computed in the common space with fast distance table lookup. Experimental results compared with several competitive algorithms over three benchmark datasets demonstrate that the proposed approach achieves the state-of-the-art performance.

1.596 Efficient Temporal Sequence Comparison and Classification using Gram Matrix Embeddings On a Riemannian Manifold

[Link to Paper]

In this paper we propose a new framework to compare and classify temporal sequences. The proposed approach captures the underlying dynamics of the data while avoiding expensive estimation procedures, making it suitable to process large numbers of sequences. The main idea is to first embed the sequences into a Riemannian manifold by using positive definite regularized Gram matrices of their Hankelets. The advantages of the this approach are: 1) it allows for using non-Euclidean similarity functions on the Positive Definite matrix manifold, which capture better the underlying geometry than directly comparing the sequences or their Hankel matrices; and 2) Gram matrices inherit desirable properties from the underlying Hankel matrices: their rank measure the complexity of the underlying dynamics, and the order and coefficients of the associated regressive models are invariant to affine transformations and varying initial conditions. The benefits of this approach are illustrated with extensive experiments in 3D action recognition using 3D joints sequences. In spite of its simplicity, the performance of this approach is competitive or better than using state-of-art approaches for this problem. Further, these results hold across a variety of metrics, supporting the idea that the improvement stems from the embedding itself, rather than from using one of these metrics.

1.597 Efficient Training of Very Deep Neural Networks for Supervised Hashing

[Link to Paper]

In this paper, we propose training very deep neural networks (DNNs) for supervised learning of hash codes. Existing methods in this context train relatively "shallow" networks limited by the issues arising in back propagation (e.g. vanishing gradients) as well as computational efficiency. We propose a novel and efficient training algorithm inspired by alternating direction method of multipliers (ADMM) that overcomes some of these limitations. Our method decomposes the training process into independent layer-wise local updates through auxiliary variables. Empirically we observe that our training algorithm always converges and its computational complexity is linearly proportional to the number of edges in the networks. Empirically we manage to train DNNs with 64 hidden layers and 1024 nodes per layer for supervised hashing in about 3 hours using a single GPU. Our proposed very deep supervised hashing (VDSH) method significantly outperforms the state-of-theart on several benchmark datasets.

1.598 Embedding Label Structures for Fine-Grained Feature Representation

[Link to Paper]

(Abstract could not be extracted from PDF)

1.599 Fast Zero-Shot Image Tagging

[Link to Paper]

The well-known word analogy experiments show that the recent word vectors capture fine-grained linguistic regularities in words by linear vector offsets, but it is unclear how well the simple vector offsets can encode visual regularities over words. We study a particular image-word relevance relation in this paper. Our results show that the word vectors of relevant tags for a given image rank ahead of the irrelevant tags, along a principal direction in the word vector space. Inspired by this observation, we propose to solve image tagging by estimating the principal direction for an image. Particularly, we exploit linear mappings and nonlinear deep neural networks to approximate the principal direction from an input image. We arrive at a quite versatile tagging model. It runs fast given a test image, in constant time w.r.t. the training set size. It not only gives superior performance for the conventional tagging task on the NUSWIDE dataset, but also outperforms competitive baselines on annotating images with previously unseen tags.

1.600 Functional Faces: Groupwise Dense Correspondence using Functional Maps

[Link to Paper]

In this paper we present a method for computing dense correspondence between a set of 3D face meshes using functional maps. The functional maps paradigm brings with it a number of advantages for face correspondence. First, it allows us to combine various notions of correspondence. We do so by proposing a number of face-specific functions, suited to either within- or between-subject correspondence. Second, we propose a groupwise variant of the method allowing us to compute cycle-consistent functional maps between all faces in a training set. Since functional maps are of much lower dimension than point-to-point correspondences, this is feasible even when the input meshes are very high resolution. Finally, we show how a functional map provides a geometric constraint that can be used to filter feature matches between non-rigidly deforming surfaces.

1.601 How Far are We from Solving Pedestrian Detection?

[Link to Paper]

Encouraged by the recent progress in pedestrian detection, we investigate the gap between current state-of-the-art methods and the "perfect single frame detector". We enable our analysis by creating a human baseline for pedestrian detection (over the Caltech dataset), and by manually clustering the recurrent errors of a top detector. Our results characterise both localisation and background-versusforeground errors. To address localisation errors we study the impact of training annotation noise on the detector performance, and show that we can improve even with a small portion of sanitised training data. To address background/foreground discrimination, we study convnets for pedestrian detection, and discuss which factors affect their performance. Other than our in-depth analysis, we report top performance on the Caltech dataset, and provide a new sanitised set of training and test annotations.

1.602 In Defense of Sparse Tracking: Circulant Sparse Tracker

[Link to Paper]

Sparse representation has been introduced to visual tracking by finding the best target candidate with minimal reconstruction error within the particle filter framework. However, most sparse representation based trackers have high computational cost, less than promising tracking performance, and limited feature representation. To deal with the above issues, we propose a novel circulant sparse tracker (CST), which exploits circulant target templates. Because of the circulant structure property, CST has the following advantages: (1) It can refine and reduce particles using circular shifts of target templates. (2) The optimization can be efficiently solved entirely in the Fourier domain. (3) High dimensional features can be embedded into CST to significantly improve tracking performance without sacrificing much computation time. Both qualitative and quantitative evaluations on challenging benchmark sequences demonstrate that CST performs better than all other sparse trackers and favorably against state-of-the-art methods. Figure 1. Comparison of our CST tracker with state-of-the-art methods (KCF, Struck, SCM, ASLA, L1APG, and TLD) on two videos from a visual tracking benchmark [32]. On the Jogging sequence, only TLD and CST can track well when partial occlusion happens. On the Tiger sequence, our CST can track the target throughout the whole sequence, while other trackers suffer from drift. The sparse trackers (ASLA, L1APG, SCM) fail to track these sequences. Overall, the proposed CST method performs favorably against the state-of-the-art trackers.

1.603 Instance-Level Segmentation for Autonomous Driving with Deep Densely Connected MRFs

[Link to Paper]

Our aim is to provide a pixel-wise instance-level labeling of a monocular image in the context of autonomous driving. We build on recent work [32] that trained a convolutional neural net to predict instance labeling in local image patches, extracted exhaustively in a stride from an image. A simple Markov random field model using several heuristics was then proposed in [32] to derive a globally consistent instance labeling of the image. In this paper, we formulate the global labeling problem with a novel densely connected Markov random field and show how to encode various intuitive potentials in a way that is amenable to efficient mean field inference [15]. Our potentials encode the compatibility between the global labeling and the patch-level predictions, contrast-sensitive smoothness as well as the fact that separate regions form different instances. Our experiments on the challenging KITTI benchmark [8] demonstrate that our method achieves a significant performance boost over the baseline [32].

1.604 Joint Multiview Segmentation and Localization of RGB-D Images using Depth-Induced Silhouette Consistency

[Link to Paper]

In this paper, we propose an RGB-D camera localization approach which takes an effective geometry constraint, i.e. silhouette consistency, into consideration. Unlike existing approaches which usually assume the silhouettes are provided, we consider more practical scenarios and generate the silhouettes for multiple views on the fly. To obtain a set of accurate silhouettes, precise camera poses are required to propagate segmentation cues across views. To perform better localization, accurate silhouettes are needed to constrain camera poses. Therefore the two problems are intertwined with each other and require a joint treatment. Facilitated by the available depth, we introduce a simple but effective silhouette consistency energy term that binds traditional appearance-based multiview segmentation cost and RGB-D frame-to-frame matching cost together. Optimization of the problem w.r.t. binary segmentation masks and camera poses naturally fits in the graph cut minimization framework and the Gauss-Newton non-linear least-squares method respectively. Experiments show that the proposed approach achieves state-of-the-arts performance on both tasks of image segmentation and camera localization.

1.605 Learning a Discriminative Null Space for Person Re-identification

[Link to Paper]

Most existing person re-identification (re-id) methods focus on learning the optimal distance metrics across camera views. Typically a person's appearance is represented using features of thousands of dimensions, whilst only hundreds of training samples are available due to the difficulties in collecting matched training images. With the number of training samples much smaller than the feature dimension, the existing methods thus face the classic small sample size (SSS) problem and have to resort to dimensionality reduction techniques and/or matrix regularisation, which lead to loss of discriminative power. In this work, we propose to overcome the SSS problem in re-id distance metric learning by matching people in a discriminative null space of the training data. In this null space, images of the same person are collapsed into a single point thus minimising the within-class scatter to the extreme and maximising the relative between-class separation simultaneously. Importantly, it has a fixed dimension, a closed-form solution and is very efficient to compute. Extensive experiments carried out on five person re-identification benchmarks including VIPeR, PRID2011, CUHK01, CUHK03 and Market1501 show that such a simple approach beats the state-of-the-art alternatives, often by a big margin.

1.606 Multi-Oriented Text Detection with Fully Convolutional Networks

[Link to Paper]

In this paper, we propose a novel approach for text detection in natural images. Both local and global cues are taken into account for localizing text lines in a coarse-to-fine procedure. First, a Fully Convolutional Network (FCN) model is trained to predict the salient map of text regions in a holistic manner. Then, text line hypotheses are estimated by combining the salient map and character components. Finally, another FCN classifier is used to predict the centroid of each character, in order to remove the false hypotheses. The framework is general for handling text in multiple orientations, languages and fonts. The proposed method consistently achieves the state-of-the-art performance on three text detection benchmarks: MSRA-TD500, ICDAR2015 and ICDAR2013.

1.607 Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis

[Link to Paper]

Emotion is expressed in multiple modalities, yet most research has considered at most one or two. This stems in part from the lack of large, diverse, well-annotated, multimodal databases with which to develop and test algorithms. We present a well-annotated, multimodal, multidimensional spontaneous emotion corpus of 140 participants. Emotion inductions were highly varied. Data were acquired from a variety of sensors of the face that included high-resolution 3D dynamic imaging, high-resolution 2D video, and thermal (infrared) sensing, and contact physiological sensors that included electrical conductivity of the skin, respiration, blood pressure, and heart rate. Facial expression was annotated for both the occurrence and intensity of facial action units from 2D video by experts in the Facial Action Coding System (FACS). The corpus further includes derived features from 3D, 2D, and IR (infrared) sensors and baseline results for facial expression and action unit detection. The entire corpus will be made available to the research community.

1.608 Occlusion-free Face Alignment: Deep Regression Networks Coupled with De-corrupt AutoEncoders

[Link to Paper]

Face alignment or facial landmark detection plays an important role in many computer vision applications, e.g., face recognition, facial expression recognition, face animation, etc. However, the performance of face alignment system degenerates severely when occlusions occur. In this work, we propose a novel face alignment method, which cascades several Deep Regression networks coupled with De-corrupt Autoencoders (denoted as DRDA) to explicitly handle partial occlusion problem. Different from the previous works that can only detect occlusions and discard the occluded parts, our proposed de-corrupt autoencoder network can automatically recover the genuine appearance for the occluded parts and the recovered parts can be leveraged together with those non-occluded parts for more accurate alignment. By coupling de-corrupt autoencoders with deep regression networks, a deep alignment model robust to partial occlusions is achieved. Besides, our method can localize occluded regions rather than merely predict whether the landmarks are occluded. Experiments on two challenging occluded face datasets demonstrate that our method significantly outperforms the state-of-the-art methods.

1.609 Online Collaborative Learning for Open-Vocabulary Visual Classifiers

[Link to Paper]

We focus on learning open-vocabulary visual classifiers, which scale up to a large portion of natural language vocabulary (e.g., over tens of thousands of classes). In particular, the

training data are large-scale weakly labeled Web images since it is difficult to acquire sufficient well-labeled data at this category scale. In this paper, we propose a novel online learning paradigm towards this challenging task. Different from traditional N-way independent classifiers that generally fail to handle the extremely sparse and inter-related labels, our classifiers learn from continuous label embeddings discovered by collaboratively decomposing the sparse image-label matrix. Leveraging on the structure of the proposed collaborative learning formulation, we develop an efficient online algorithm that can jointly learn the label embeddings and visual classifiers. The algorithm can learn over 30,000 classes of 1,000 training images within 1 second on a standard GPU. Extensively experimental results on four benchmarks demonstrate the effectiveness of our method.

1.610 Pairwise Matching through Max-Weight Bipartite Belief Propagation

[Link to Paper]

Feature matching is a key problem in computer vision and pattern recognition. One way to encode the essential interdependence between potential feature matches is to cast the problem as inference in a graphical model, though recently alternatives such as spectral methods, or approaches based on the convex-concave procedure have achieved the state-of-the-art. Here we revisit the use of graphical models for feature matching, and propose a belief propagation scheme which exhibits the following advantages: (1) we explicitly enforce one-to-one matching constraints; (2) we offer a tighter relaxation of the original cost function than previous graphical-model-based approaches; and (3) our sub-problems decompose into max-weight bipartite matching, which can be solved efficiently, leading to orders-of-magnitude reductions in execution time. Experimental results show that the proposed algorithm produces results superior to those of the current state-of-the-art.

1.611 Picking Deep Filter Responses for Fine-grained Image Recognition

[Link to Paper]

Recognizing fine-grained sub-categories such as birds and dogs is extremely challenging due to the highly localized and subtle differences in some specific parts. Most previous works rely on object / part level annotations to build part-based representation, which is demanding in practical applications. This paper proposes an automatic finegrained recognition approach which is free of any object / part annotation at both training and testing stages. Our method explores a unified framework based on two steps of deep filter response picking. The first picking step is to find distinctive filters which respond to specific patterns significantly and consistently, and learn a set of part detectors via iteratively alternating between new positive sample mining and part model retraining. The second picking step

is to pool deep filter responses via spatially weighted combination of Fisher Vectors. We conditionally pick deep filter responses to encode them into the final representation, which considers the importance of filter responses themselves. Integrating all these techniques produces a much more powerful framework, and experiments conducted on CUB-2002011 and Stanford Dogs demonstrate the superiority of our proposed algorithm over the existing methods.

1.612 ReD-SFA: Relation Discovery Based Slow Feature Analysis for Trajectory Clustering

[Link to Paper]

For spectral embedding/clustering, it is still an open problem on how to construct an relation graph to reflect the intrinsic structures in data. In this paper, we proposed an approach, named Relation Discovery based Slow Feature Analysis (ReD-SFA), for feature learning and graph construction simultaneously. Given an initial graph with only a few nearest but most reliable pairwise relations, new reliable relations are discovered by an assumption of reliability preservation, i.e., the reliable relations will preserve their reliabilities in the learnt projection subspace. We formulate the idea as a cross entropy (CE) minimization problem to reduce the discrepancy between two Bernoulli distributions parameterized by the updated distances and the existing relation graph respectively. Furthermore, to overcome the imbalanced distribution of samples, a Boosting-like strategy is proposed to balance the discovered relations over all clusters. To evaluate the proposed method, extensive experiments are performed with various trajectory clustering tasks, including motion segmentation, time series clustering and crowd detection. The results demonstrate that ReDSFA can discover reliable intra-cluster relations with high precision, and competitive clustering performance can be achieved in comparison with state-of-the-art.

1.613 Real-time Action Recognition with Enhanced Motion Vector CNNs

[Link to Paper]

The deep two-stream architecture [23] exhibited excellent performance on video based action recognition. The most computationally expensive step in this approach comes from the calculation of optical flow which prevents it to be real-time. This paper accelerates this architecture by replacing optical flow with motion vector which can be obtained directly from compressed videos without extra calculation. However, motion vector lacks fine structures, and contains noisy and inaccurate motion patterns, leading to the evident degradation of recognition performance. Our key insight for relieving this problem is that optical flow

and motion vector are inherent correlated. Transferring the knowledge learned with optical flow CNN to motion vector CNN can significantly boost the performance of the latter. Specifically, we introduce three strategies for this, initialization transfer, supervision transfer and their combination. Experimental results show that our method achieves comparable recognition performance to the state-of-the-art, while our method can process 390.7 frames per second, which is 27 times faster than the original two-stream method.

1.614 SPDA-CNN: Unifying Semantic Part Detection and Abstraction for Fine-grained Recognition

[Link to Paper]

Most convolutional neural networks (CNNs) lack midlevel layers that model semantic parts of objects. This limits CNN-based methods from reaching their full potential in detecting and utilizing small semantic parts in recognition. Introducing such mid-level layers can facilitate the extraction of part-specific features which can be utilized for better recognition performance. This is particularly important in the domain of fine-grained recognition. In this paper, we propose a new CNN architecture that integrates semantic part detection and abstraction (SPDACNN) for fine-grained classification. The proposed network has two sub-networks: one for detection and one for recognition. The detection sub-network has a novel top-down proposal method to generate small semantic part candidates for detection. The classification sub-network introduces novel part layers that extract features from parts detected by the detection sub-network, and combine them for recognition. As a result, the proposed architecture provides an end-to-end network that performs detection, localization of multiple semantic parts, and whole object recognition within one framework that shares the computation of convolutional filters. Our method outperforms state-of-theart methods with a large margin for small parts detection (e.g. our precision of 93.40% vs the best previous precision of 74.00% for detecting the head on CUB-2011). It also compares favorably to the existing state-of-the-art on finegrained classification, e.g. it achieves 85.14% accuracy on CUB-2011.

1.615 Sample-Specific SVM Learning for Person Re-identification

[Link to Paper]

Person re-identification addresses the problem of matching people across disjoint camera views and extensive efforts have been made to seek either the robust feature representation or the discriminative matching metrics. However, most existing approaches focus on learning a fixed distance metric for all instance pairs, while ignoring the individuality of each person. In this paper, we formulate the person re-identification problem as

286

an imbalanced classification problem and learn a classifier specifically for each pedestrian such that the matching model is highly tuned to the individual's appearance. To establish correspondence between feature space and classifier space, we propose a Least Square Semi-Coupled Dictionary Learning (LSSCDL) algorithm to learn a pair of dictionaries and a mapping function efficiently. Extensive experiments on a series of challenging databases demonstrate that the proposed algorithm performs favorably against the state-of-the-art approaches, especially on the rank-1 recognition rate.

1.616 SemiContour: A Semi-supervised Learning Approach for Contour Detection

[Link to Paper]

Supervised contour detection methods usually require many labeled training images to obtain satisfactory performance. However, a large set of annotated data might be unavailable or extremely labor intensive. In this paper, we investigate the usage of semi-supervised learning (SSL) to obtain competitive detection accuracy with very limited training data (three labeled images). Specifically, we propose a semi-supervised structured ensemble learning approach for contour detection built on structured random forests (SRF). To allow SRF to be applicable to unlabeled data, we present an effective sparse representation approach to capture inherent structure in image patches by finding a compact and discriminative low-dimensional subspace representation in an unsupervised manner, enabling the incorporation of abundant unlabeled patches with their estimated structured labels to help SRF perform better node splitting. We re-examine the role of sparsity and propose a novel and fast sparse coding algorithm to boost the overall learning efficiency. To the best of our knowledge, this is the first attempt to apply SSL for contour detection. Extensive experiments on the BSDS500 segmentation dataset and the NYU Depth dataset demonstrate the superiority of the proposed method.

1.617 Single-Image Crowd Counting via Multi-Column Convolutional Neural Network

[Link to Paper]

This paper aims to develop a method than can accurately estimate the crowd count from an individual image with arbitrary crowd density and arbitrary perspective. To this end, we have proposed a simple but effective Multi-column Convolutional Neural Network (MCNN) architecture to map the image to its crowd density map. The proposed MCNN allows the input image to be of arbitrary size or resolution. By utilizing filters with receptive fields of different sizes, the features learned by each column CNN are adaptive to variations in

287

people/head size due to perspective effect or image resolution. Furthermore, the true density map is computed accurately based on geometry-adaptive kernels which do not need knowing the perspective map of the input image. Since exiting crowd counting datasets do not adequately cover all the challenging situations considered in our work, we have collected and labelled a large new dataset that includes 1198 images with about 330,000 heads annotated. On this challenging new dataset, as well as all existing datasets, we conduct extensive experiments to verify the effectiveness of the proposed model and method. In particular, with the proposed simple MCNN model, our method outperforms all existing methods. In addition, experiments show that our model, once trained on one dataset, can be readily transferred to a new dataset.

1.618 SketchNet: Sketch Classification with Web Images

[Link to Paper]

In this study, we present a weakly supervised approach that discovers the discriminative structures of sketch images, given pairs of sketch images and web images. In contrast to traditional approaches that use global appearance features or relay on keypoint features, our aim is to automatically learn the shared latent structures that exist between sketch images and real images, even when there are significant appearance differences across its relevant real images. To accomplish this, we propose a deep convolutional neural network, named SketchNet. We firstly develop a triplet composed of sketch, positive and negative real image as the input of our neural network. To discover the coherent visual structures between the sketch and its positive pairs, we introduce the softmax as the loss function. Then a ranking mechanism is introduced to make the positive pairs obtain a higher score comparing over negative ones to achieve robust representation. Finally, we formalize above-mentioned constrains into the unified objective function, and create an ensemble feature representation to describe the sketch images. Experiments on the TU-Berlin sketch benchmark demonstrate the effectiveness of our model and show that deep feature representation brings substantial improvements over other state-of-the-art methods on sketch classification.

1.619 Summary Transfer: Exemplar-based Subset Selection for Video Summarization

[Link to Paper]

Video summarization has unprecedented importance to help us digest, browse, and search today's ever-growing video collections. We propose a novel subset selection technique that leverages supervision in the form of humancreated summaries to perform automatic

keyframe-based video summarization. The main idea is to nonparametrically transfer summary structures from annotated videos to unseen test videos. We show how to extend our method to exploit semantic side information about the video's category/genre to guide the transfer process by those training videos semantically consistent with the test input. We also show how to generalize our method to subshot-based summarization, which not only reduces computational costs but also provides more flexible ways of defining visual similarity across subshots spanning several frames. We conduct extensive evaluation on several benchmarks and demonstrate promising results, outperforming existing methods in several settings.

1.620 Unconstrained Salient Object Detection via Proposal Subset Optimization

[Link to Paper]

We aim at detecting salient objects in unconstrained images. In unconstrained images, the number of salient objects (if any) varies from image to image, and is not given. We present a salient object detection system that directly outputs a compact set of detection windows, if any, for an input image. Our system leverages a Convolutional-NeuralNetwork model to generate location proposals of salient objects. Location proposals tend to be highly overlapping and noisy. Based on the Maximum a Posteriori principle, we propose a novel subset optimization framework to generate a compact set of detection windows out of noisy proposals. In experiments, we show that our subset optimization formulation greatly enhances the performance of our system, and our system attains 16-34% relative improvement in Average Precision compared with the state-of-the-art on three challenging salient object datasets.

1.621 Yin and Yang: Balancing and Answering Binary Visual Questions

[Link to Paper]

The complex compositional structure of language makes problems at the intersection of vision and language challenging. But language also provides a strong prior that can result in good superficial performance, without the underlying models truly understanding the visual content. This can hinder progress in pushing state of art in the computer vision aspects of multi-modal AI. In this paper, we address binary Visual Question Answering (VQA) on abstract scenes. We formulate this problem as visual verification of concepts inquired in the questions. Specifically, we convert the question to a tuple that concisely summarizes the visual concept to be detected in the image. If the concept can be found in the image, the answer to the question is "yes", and otherwise "no". Abstract scenes play

two roles (1) They allow us to focus on the highlevel semantics of the VQA task as opposed to the low-level recognition problems, and perhaps more importantly, (2) They provide us the modality to balance the dataset such that language priors are controlled, and the role of vision is essential. In particular, we collect fine-grained pairs of scenes for every question, such that the answer to the question is "yes" for one scene, and "no" for the other for the exact same question. Indeed, language priors alone do not perform better than chance on our balanced dataset. Moreover, our proposed approach matches the performance of a state-of-the-art VQA approach on the unbalanced dataset, and outperforms it on the balanced dataset.

1.622 Zero-Shot Learning via Joint Latent Similarity Embedding

[Link to Paper]

Zero-shot recognition (ZSR) deals with the problem of predicting class labels for target domain instances based on source domain side information (e.g. attributes) of unseen classes. We formulate ZSR as a binary prediction problem. Our resulting classifier is class-independent. It takes an arbitrary pair of source and target domain instances as input and predicts whether or not they come from the same class, i.e. whether there is a match. We model the posterior probability of a match since it is a sufficient statistic and propose a latent probabilistic model in this context. We develop a joint discriminative learning framework based on dictionary learning to jointly learn the parameters of our model for both domains, which ultimately leads to our class-independent classifier. Many of the existing embedding methods can be viewed as special cases of our probabilistic model. On ZSR our method shows 4.90% improvement over the state-of-the-art in accuracy averaged across four benchmark datasets. We also adapt ZSR method for zero-shot retrieval and show 22.45% improvement accordingly in mean average precision (mAP).

1.623 Deep Region and Multi-label Learning for Facial Action Unit Detection

[Link to Paper]

Region learning (RL) and multi-label learning (ML) have recently attracted increasing attentions in the field of facial Action Unit (AU) detection. Knowing that AUs are active on sparse facial regions, RL aims to identify these regions for a better specificity. On the other hand, a strong statistical evidence of AU correlations suggests that ML is a natural way to model the detection task. In this paper, we propose Deep Region and Multi-label Learning (DRML), a unified deep network that simultaneously addresses these two problems. One crucial aspect in DRML is a novel region layer that uses feed-forward functions to induce

important facial regions, forcing the learned weights to capture structural information of the face. Our region layer serves as an alternative design between locally connected layers (i.e., confined kernels to individual pixels) and conventional convolution layers (i.e., shared kernels across an entire image). Unlike previous studies that solve RL and ML alternately, DRML by construction addresses both problems, allowing the two seemingly irrelevant problems to interact more directly. The complete network is end-to-end trainable, and automatically learns representations robust to variations inherent within a local region. Experiments on BP4D and DISFA benchmarks show that DRML performs the highest average F1-score and AUC within and across datasets in comparison with alternative methods.

1.624 Facial Expression Intensity Estimation Using Ordinal Information

[Link to Paper]

Previous studies on facial expression analysis have been focused on recognizing basic expression categories. There is limited amount of work on the continuous expression intensity estimation, which is important for detecting and tracking emotion change. Part of the reason is the lack of labeled data with annotated expression intensity since expression intensity annotation requires expertise and is time consuming. In this work, we treat the expression intensity estimation as a regression problem. By taking advantage of the natural onset-apex-offset evolution pattern of facial expression, the proposed method can handle different amounts of annotations to perform frame-level expression intensity estimation. In fully supervised case, all the frames are provided with intensity annotations. In weakly supervised case, only the annotations of selected key frames are used. While in unsupervised case, expression intensity can be estimated without any annotations. An efficient optimization algorithm based on Alternating Direction Method of Multipliers (ADMM) is developed for solving the optimization problem associated with parameter learning. We demonstrate the effectiveness of proposed method by comparing it against both fully supervised and unsupervised approaches on benchmark facial expression datasets.

1.625 A Direct Least-Squares Solution to the PnP Problem with Unknown Focal Length

[Link to Paper]

In this work, we propose a direct least-squares solution to the perspective-n-point (PnP) pose estimation problem of a partially uncalibrated camera, whose intrinsic parameters except the focal length are known. The basic idea is to construct a proper objective function with respect to the target variables and extract all its stationary points so as to find the

global minimum. The advantages of our proposed solution over existing ones are that (i) the objective function is directly built upon the imaging equation, such that all the 3D-to-2D correspondences contribute equally to the minimized error, and that (ii) the proposed solution is noniterative, in the sense that the stationary points are retrieved by means of eigenvalue factorization and the common iterative refinement step is not needed. In addition, the proposed solution has O(n) complexity, and can be used to handle both planar and nonplanar 3D points. Experimental results show that the proposed solution is much more accurate than the existing state-of-the-art solutions, and is even comparable to the maximum likelihood estimation by minimizing the reprojection error.

1.626 Improving the Robustness of Deep Neural Networks via Stability Training

[Link to Paper]

In this paper we address the issue of output instability of deep neural networks: small perturbations in the visual input can significantly distort the feature embeddings and output of a neural network. Such instability affects many deep architectures with state-of-the-art performance on a wide range of computer vision tasks. We present a general stability training method to stabilize deep networks against small input distortions that result from various types of common image processing, such as compression, rescaling, and cropping. We validate our method by stabilizing the stateof-the-art Inception architecture [11] against these types of distortions. In addition, we demonstrate that our stabilized model gives robust state-of-the-art performance on largescale near-duplicate detection, similar-image ranking, and classification on noisy datasets.

1.627 Cascaded Interactional Targeting Network for Egocentric Video Analysis

[Link to Paper]

Knowing how hands move and what object is being manipulated are two key sub-tasks for analyzing first-person (egocentric) action. However, lack of fully annotated hand data as well as imprecise foreground segmentation make either sub-task challenging. This work aims to explicitly address these two issues via introducing a cascaded interactional targeting (i.e., infer both hand and active object regions) deep neural network. Firstly, a novel EM-like learning framework is proposed to train the pixel-level deep convolutional neural network (DCNN) by seamlessly integrating weakly supervised data (i.e., massive bounding box annotations) with a small set of strongly supervised data (i.e., fully annotated hand segmentation maps) to achieve stateof-the-art hand segmentation performance. Secondly,

the resulting high-quality hand segmentation maps are further paired with the corresponding motion maps and object feature maps, in order to explore the contextual information among object, motion and hand to generate interactional foreground regions (operated objects). The resulting interactional target maps (hand + active object) from our cascaded DCNN are further utilized to form discriminative action representation. Experiments show that our framework has achieved the state-of-the-art egocentric action recognition performance on the benchmark dataset Activities of Daily Living (ADL).

1.628 Estimating Correspondences of Deformable Objects "In-the-wild"

[Link to Paper]

During the past few years we have witnessed the development of many methodologies for building and fitting Statistical Deformable Models (SDMs). The construction of accurate SDMs requires careful annotation of images with regards to a consistent set of landmarks. However, the manual annotation of a large amount of images is a tedious, laborious and expensive procedure. Furthermore, for several deformable objects, e.g. human body, it is difficult to define a consistent set of landmarks, and, thus, it becomes impossible to train humans in order to accurately annotate a collection of images. Nevertheless, for the majority of objects, it is possible to extract the shape by object segmentation or even by shape drawing. In this paper, we show for the first time, to the best of our knowledge, that it is possible to construct SDMs by putting object shapes in dense correspondence. Such SDMs can be built with much less effort for a large battery of objects. Additionally, we show that, by sampling the dense model, a part-based SDM can be learned with its parts being in correspondence. We employ our framework to develop SDMs of human arms and legs, which can be used for the segmentation of the outline of the human body, as well as to provide better and more consistent annotations for body joints. (a) MPII (b) Fashion (c) FLIC Figure 1: Examples of inconsistent annotations of human pose among different datasets. Blue markers denote the original annotations. The arrows and green markers show the correct location at which the points should be annotated.

1.629 Fine-grained Image Classification by Exploring Bipartite-Graph Labels

[Link to Paper]

Given a food image, can a fine-grained object recognition engine tell "which restaurant which dish" the food belongs to? Such ultra-fine grained image recognition is the key for many applications like search by images, but it is very challenging because it needs to discern subtle difference between classes while dealing with the scarcity of training

data. Fortunately, the ultra-fine granularity naturally brings rich relationships among object classes. This paper proposes a novel approach to exploit the rich relationships through bipartite-graph labels (BGL). We show how to model BGL in an overall convolutional neural networks and the resulting system can be optimized through back-propagation. We also show that it is computationally efficient in inference thanks to the bipartite structure. To facilitate the study, we construct a new food benchmark dataset, which consists of 37,885 food images collected from 6 restaurants and totally 975 menus. Experimental results on this new food and three other datasets demonstrate BGL advances previous works in fine-grained object recognition. An online demo is available at http: //www.f-zhou.com/fg_demo/.

1.630 Learning Deep Features for Discriminative Localization

[Link to Paper]

In this work, we revisit the global average pooling layer proposed in [13], and shed light on how it explicitly enables the convolutional neural network (CNN) to have remarkable localization ability despite being trained on imagelevel labels. While this technique was previously proposed as a means for regularizing training, we find that it actually builds a generic localizable deep representation that exposes the implicit attention of CNNs on an image. Despite the apparent simplicity of global average pooling, we are able to achieve 37.1% top-5 error for object localization on ILSVRC 2014 without training on any bounding box annotation.We demonstrate in a variety of experiments that our network is able to localize the discriminative image regions despite just being trained for solving classification task1.

1.631 Learning Dense Correspondence via 3D-guided Cycle Consistency

[Link to Paper]

Discriminative deep learning approaches have shown impressive results for problems where human-labeled ground truth is plentiful, but what about tasks where labels are difficult or impossible to obtain? This paper tackles one such problem: establishing dense visual correspondence across different object instances. For this task, although we do not know what the ground-truth is, we know it should be consistent across instances of that category. We exploit this consistency as a supervisory signal to train a convolutional neural network to predict cross-instance correspondences between pairs of images depicting objects of the same category. For each pair of training images we find an appropriate 3D CAD model and render two synthetic views to link in with the pair, establishing a correspondence flow 4-cycle. We use ground-truth synthetic-to-synthetic correspondences, provided

by the rendering engine, to train a ConvNet to predict synthetic-to-real, real-to-real and realto-synthetic correspondences that are cycle-consistent with the ground-truth. At test time, no CAD models are required. We demonstrate that our end-to-end trained ConvNet supervised by cycle-consistency outperforms stateof-the-art pairwise matching methods in correspondencerelated tasks.

1.632 Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video

[Link to Paper]

This paper addresses the challenge of 3D full-body human pose estimation from a monocular image sequence. Here, two cases are considered: (i) the image locations of the human joints are provided and (ii) the image locations of joints are unknown. In the former case, a novel approach is introduced that integrates a sparsity-driven 3D geometric prior and temporal smoothness. In the latter case, the former case is extended by treating the image locations of the joints as latent variables to take into account considerable uncertainties in 2D joint locations. A deep fully convolutional network is trained to predict the uncertainty maps of the 2D joint locations. The 3D pose estimates are realized via an Expectation-Maximization algorithm over the entire sequence, where it is shown that the 2D joint location uncertainties can be conveniently marginalized out during inference. Empirical evaluation on the Human3.6M dataset shows that the proposed approaches achieve greater 3D pose estimation accuracy over state-of-the-art baselines. Further, the proposed approach outperforms a publicly available 2D pose estimation baseline on the challenging PennAction dataset.

1.633 A Key Volume Mining Deep Framework for Action Recognition

[Link to Paper]

1.634 A Text Detection System for Natural Scenes with Convolutional Feature Learning and Cascaded Classification

[Link to Paper]

We propose a system that finds text in natural scenes using a variety of cues. Our novel data-driven method incorporates coarse-to-fine detection of character pixels using convolutional features (Text-Conv), followed by extracting connected components (CCs) from

characters using edge and color features, and finally performing a graph-based segmentation of CCs into words (Word-Graph). For TextConv, the initial detection is based on convolutional feature maps similar to those used in Convolutional Neural Networks (CNNs), but learned using Convolutional k-means. Convolution masks defined by local and neighboring patch features are used to improve detection accuracy. The WordGraph algorithm uses contextual information to both improve word segmentation and prune false character/word detections. Different definitions for foreground (text) regions are used to train the detection stages, some based on bounding box intersection, and others on bounding box and pixel intersection. Our system obtains pixel, character, and word detection f-measures of 93.14%, 90.26%, and 86.77% respectively for the ICDAR 2015 Robust Reading Focused Scene Text dataset, out-performing state-of-the-art systems. This approach may work for other detection targets with homogenous color in natural scenes.

1.635 Beyond Local Search: Tracking Objects Everywhere with Instance-Specific Proposals

[Link to Paper]

Most tracking-by-detection methods employ a local search window around the predicted object location in the current frame assuming the previous location is accurate, the trajectory is smooth, and the computational capacity permits a search radius that can accommodate the maximum speed yet small enough to reduce mismatches. These, however, may not be valid always, in particular for fast and irregularly moving objects. Here, we present an object tracker that is not limited to a local search window and has ability to probe efficiently the entire frame. Our method generates a small number of "high-quality" proposals by a novel instance-specific objectness measure and evaluates them against the object model that can be adopted from an existing tracking-by-detection approach as a core tracker. During the tracking process, we update the object model concentrating on hard false-positives supplied by the proposals, which help suppressing distractors caused by difficult background clutters, and learn how to re-rank proposals according to the object model. Since we reduce significantly the number of hypotheses the core tracker evaluates, we can use richer object descriptors and stronger detector. Our method outperforms most recent state-of-the-art trackers on popular tracking benchmarks, and provides improved robustness for fast moving objects as well as for ultra lowframe-rate videos.

1.636 Discriminative Multi-modal Feature Fusion for RGBD Indoor Scene Recognition

[Link to Paper]

RGBD scene recognition has attracted increasingly attention due to the rapid development of depth sensors and their wide application scenarios. While many research has been conducted, most work used hand-crafted features which are difficult to capture high-level semantic structures. Recently, the feature extracted from deep convolutional neural network has produced state-of-the-art results for various computer vision tasks, which inspire researchers to explore incorporating CNN learned features for RGBD scene understanding. On the other hand, most existing work combines rgb and depth features without adequately exploiting the consistency and complementary information between them. Inspired by some recent work on RGBD object recognition using multi-modal feature fusion, we introduce a novel discriminative multi-modal fusion framework for rgbd scene recognition for the first time which simultaneously considers the inter- and intra-modality correlation for all samples and meanwhile regularizing the learned features to be discriminative and compact. The results from the multimodal layer can be back-propagated to the lower CNN layers, hence the parameters of the CNN layers and multimodal layers are updated iteratively until convergence. Experiments on the recently proposed large scale SUN RGB-D datasets show that our method achieved the state-of-the-art without any image segmentation.

1.637 Face Alignment Across Large Poses: A 3D Solution

[Link to Paper]

Face alignment, which fits a face model to an image and extracts the semantic meanings of facial pixels, has been an important topic in CV community. However, most algorithms are designed for faces in small to medium poses (below 45), lacking the ability to align faces in large poses up to 90. The challenges are three-fold: Firstly, the commonly used landmark-based face model assumes that all the landmarks are visible and is therefore not suitable for profile views. Secondly, the face appearance varies more dramatically across large poses, ranging from frontal view to profile view. Thirdly, labelling landmarks in large poses is extremely challenging since the invisible landmarks have to be guessed. In this paper, we propose a solution to the three problems in an new alignment framework, called 3D Dense Face Alignment (3DDFA), in which a dense 3D face model is fitted to the image via convolutional neutral network (CNN). We also propose a method to synthesize large-scale training samples in profile views to solve the third problem of data labelling. Experiments on the challenging AFLW database show that our approach achieves significant improvements over state-of-the-art methods.

1.638 From Noise Modeling to Blind Image Denoising

[Link to Paper]

Traditional image denoising algorithms always assume the noise to be homogeneous white Gaussian distributed. However, the noise on real images can be much more complex empirically. This paper addresses this problem and proposes a novel blind image denoising algorithm which can cope with real-world noisy images even when the noise model is not provided. It is realized by modeling image noise with mixture of Gaussian distribution (MoG) which can approximate large varieties of continuous distributions. As the number of components for MoG is unknown practically, this work adopts Bayesian nonparametric technique and proposes a novel Low-rank MoG filter (LR-MoG) to recover clean signals (patches) from noisy ones contaminated by MoG noise. Based on LR-MoG, a novel blind image denoising approach is developed. To test the proposed method, this study conducts extensive experiments on synthesis and real images. Our method achieves the state-ofthe-art performance consistently.

1.639 Inferring Forces and Learning Human Utilities From Videos

[Link to Paper]

We propose a notion of affordance that takes into account physical quantities generated when the human body interacts with real-world objects, and introduce a learning framework that incorporates the concept of human utilities, which in our opinion provides a deeper and finer-grained account not only of object affordance but also of people's interaction with objects. Rather than defining affordance in terms of the geometric compatibility between body poses and 3D objects, we devise algorithms that employ physicsbased simulation to infer the relevant forces/pressures acting on body parts. By observing the choices people make in videos (particularly in selecting a chair in which to sit) our system learns the comfort intervals of the forces exerted on body parts (while sitting). We account for people's preferences in terms of human utilities, which transcend comfort intervals to account also for meaningful tasks within scenes and spatiotemporal constraints in motion planning, such as for the purposes of robot task planning.

1.640 Traffic-Sign Detection and Classification in the Wild

[Link to Paper]

Although promising results have been achieved in the areas of traffic-sign detection and classification, few works have provided simultaneous solutions to these two tasks for realistic real world images. We make two contributions to this problem. Firstly, we have created a large traffic-sign benchmark from 100000 Tencent Street View panoramas, going beyond previous benchmarks. It provides 100000 images containing 30000 traffic-sign instances. These images cover large variations in illuminance and weather conditions. Each

traffic-sign in the benchmark is annotated with a class label, its bounding box and pixel mask. We call this benchmark Tsinghua-Tencent 100K. Secondly, we demonstrate how a robust end-to-end convolutional neural network (CNN) can simultaneously detect and classify trafficsigns. Most previous CNN image processing solutions target objects that occupy a large proportion of an image, and such networks do not work well for target objects occupying only a small fraction of an image like the traffic-signs here. Experimental results show the robustness of our network and its superiority to alternatives. The benchmark, source code and the CNN model introduced in this paper is publicly available1.

1.641 Unconstrained Face Alignment via Cascaded Compositional Learning

[Link to Paper]

We present a practical approach to address the problem of unconstrained face alignment for a single image. In our unconstrained problem, we need to deal with large shape and appearance variations under extreme head poses and rich shape deformation. To equip cascaded regressors with the capability to handle global shape variation and irregular appearance-shape relation in the unconstrained scenario, we partition the optimisation space into multiple domains of homogeneous descent, and predict a shape as a composition of estimations from multiple domain-specific regressors. With a specially formulated learning objective and a novel tree splitting function, our approach is capable of estimating a robust and meaningful composition. In addition to achieving state-of-the-art accuracy over existing approaches, our framework is also an efficient solution (350 FPS), thanks to the on-the-fly domain exclusion mechanism and the capability of leveraging the fast pixel feature.

1.642 Visual7W: Grounded Question Answering in Images

[Link to Paper]

We have seen great progress in basic perceptual tasks such as object recognition and detection. However, AI models still fail to match humans in high-level vision tasks due to the lack of capacities for deeper reasoning. Recently the new task of visual question answering (QA) has been proposed to evaluate a model's capacity for deep image understanding. Previous works have established a loose, global association between QA sentences and images. However, many questions and answers, in practice, relate to local regions in the images. We establish a semantic link between textual descriptions and image regions by object-level grounding. It enables a new type of QA with visual answers, in addition to textual answers used in previous work. We study the visual QA tasks in a grounded setting with a large collection of 7W multiple-choice QA pairs. Furthermore, we evaluate human

performance and several baseline models on the QA tasks. Finally, we propose a novel LSTM model with spatial attention to tackle the 7W QA tasks.

1.643 Fast Training of Triplet-based Deep Binary Embedding Networks

[Link to Paper]

In this paper, we aim to learn a mapping (or embedding) from images to a compact binary space in which Hamming distances correspond to a ranking measure for the image retrieval task. We make use of a triplet loss because this has been shown to be most effective for ranking problems. However, training in previous works can be prohibitively expensive due to the fact that optimization is directly performed on the triplet space, where the number of possible triplets for training is cubic in the number of training examples. To address this issue, we propose to formulate high-order binary codes learning as a multi-label classification problem by explicitly separating learning into two interleaved stages. To solve the first stage, we design a large-scale high-order binary codes inference algorithm to reduce the high-order objective to a standard binary quadratic problem such that graph cuts can be used to efficiently infer the binary codes which serve as the labels of each training datum. In the second stage we propose to map the original image to compact binary codes via carefully designed deep convolutional neural networks (CNNs) and the hashing function fitting can be solved by training binary CNN classifiers. An incremental/interleaved optimization strategy is proffered to ensure that these two steps are interactive with each other during training for better accuracy. We conduct experiments on several benchmark datasets, which demonstrate both improved training time (by as much as two orders of magnitude) as well as producing state-of-theart hashing for various retrieval tasks.

2 Computer Vision in Vehicle Technology

2.1 DR(eye)VE: a Dataset for Attention-Based Tasks with Applications to Autonomous and Assisted Driving

[Link to Paper]

Autonomous and assisted driving are undoubtedly hot topics in computer vision. However, the driving task is extremely complex and a deep understanding of drivers' behavior is still lacking. Several researchers are now investigating the attention mechanism in order to define computational models for detecting salient and interesting objects in the scene. Nevertheless, most of these models only refer to bottom up visual saliency and are focused on still images. Instead, during the driving experience the temporal nature and peculiarity of the task influence the attention mechanisms, leading to the conclusion that real life driving data is mandatory. In this paper we propose a novel and publicly available dataset acquired during actual driving. Our dataset, composed by more than 500,000 frames, contains drivers' gaze fixations and their temporal integration providing task-specific saliency maps. Geo-referenced locations, driving speed and course complete the set of released data. To the best of our knowledge, this is the first publicly available dataset of this kind and can foster new discussions on better understanding, exploiting and reproducing the driver's attention process in the autonomous and assisted cars of future generations.

2.2 DeepLanes: End-To-End Lane Position Estimation using Deep Neural Networks

[Link to Paper]

Camera-based lane detection algorithms are one of the key enablers for many semi-autonomous and fullyautonomous systems, ranging from lane keep assist to level5 automated vehicles. Positioning a vehicle between lane boundaries is the core navigational aspect of a self-driving car. Even though this should be trivial, given the clarity of lane markings on most standard roadway systems, the process is typically mired with tedious pre-processing and computational effort. We present an approach to estimate lane positions directly using a deep neural network that operates on images from laterally-mounted down-facing cameras. To create a diverse training set, we present a method to generate semi-artificial images. Besides the ability to distinguish whether there is a lane-marker present or not, the network is able to estimate the position of a lane marker with sub-centimeter accuracy at an average of 100 frames/s on an embedded automotive platform, requiring no pre- or post-processing. This system can be used not only to estimate lane position for navigation, but also provide an efficient way to validate the robustness of driver-assist features which depend on lane information.

2.3 The HCI Benchmark Suite: Stereo And Flow Ground Truth With Uncertainties for Urban Autonomous Driving

[Link to Paper]

Recent advances in autonomous driving require more and more highly realistic reference data, even for difficult situations such as low light and bad weather. We present a new stereo and optical flow dataset to complement existing benchmarks. It was specifically designed to be representative for urban autonomous driving, including realistic, systematically varied radiometric and geometric challenges which were previously unavailable. The accuracy of the ground truth is evaluated based on Monte Carlo simulations yielding full, per-pixel distributions. Interquartile ranges are used as uncertainty measure to create binary masks for arbitrary accuracy thresholds and show that we achieved uncertainties better than those reported for comparable outdoor benchmarks. Binary masks for all dynamically moving regions are supplied with estimated stereo and flow values. An initial public benchmark dataset of 55 manually selected sequences between 19 and 100 frames long are made available in a dedicated website featuring interactive tools for database search, visualization, comparison and benchmarking.

2.4 Multiple Scale Faster-RCNN Approach to Driver's Cell-phone Usage and Hands on Steering Wheel Detection

[Link to Paper]

In this paper, we present an advanced deep learning based approach to automatically determine whether a driver is using a cell-phone as well as detect if his/her hands are on the steering wheel (i.e. counting the number of hands on the wheel). To robustly detect small objects such as hands, we propose Multiple Scale Faster-RCNN (MSFRCNN) approach that uses a standard Region Proposal Network (RPN) generation and incorporates feature maps from shallower convolution feature maps, i.e. conv3 and conv4, for ROI pooling. In our driver distraction detection framework, we first make use of the proposed MS-FRCNN to detect individual objects, namely, a hand, a cell-phone, and a steering wheel. Then, the geometric information is extracted to determine if a cell-phone is being used or how many hands are on the wheel. The proposed approach is demonstrated and evaluated on the Vision for Intelligent Vehicles and Applications (VIVA) Challenge database and the challenging Strategic Highway Research Program (SHRP2) face view videos that was acquired to monitor drivers under naturalistic driving conditions. The experimental results show that our method archives better performance than Faster R-CNN on both hands on wheel detection and cell-phone usage detection while remaining at similar testing cost. Compare to the state-of-the-art cell-phone usage detection, our approach obtains higher accuracy, is

less time consuming and is independent to landmarking. The groundtruth database will be publicly available.

2.5 Monocular Long-term Target Following on UAVs

[Link to Paper]

In this paper, we investigate the challenging long-term visual tracking problem and its implementation on Unmanned Aerial Vehicles (UAVs). By exploiting the inherent correlation between Frequency tracker And Spatial detector, we propose a novel tracking algorithm, denoted as FAST. As can be theoretically and analytically shown, the superior performance of FAST originates from: 1) robustness by transforming from frequency tracker to spatial detector, FAST owns comprehensive detector to cover consequential temporal variance/invariance information that inherently retained in tracker; 2) efficiency the coarse-tofine redetection scheme avoids the training of extra classifier and exhaustive search of location and scale. Experiments testified on tracking benchmarks demonstrate the impressive performance of FAST. In particular, we successfully implement FAST on quadrotor platform to tackle with indoor and outdoor practical scenarios, achieving real-time, automatic, smooth, and long-term target following on UAVs.

2.6 UAV-based Autonomous Image Acquisition with Multi-View Stereo Quality Assurance by Confidence Prediction

[Link to Paper]

In this paper we present an autonomous system for acquiring close-range high-resolution images that maximize the quality of a later-on 3D reconstruction with respect to coverage, ground resolution and 3D uncertainty. In contrast to previous work, our system uses the already acquired images to predict the confidence in the output of a dense multi-view stereo approach without executing it. This confidence encodes the likelihood of a successful reconstruction with respect to the observed scene and potential camera constellations. Our prediction module runs in real-time and can be trained without any externally recorded ground truth. We use the confidence prediction for on-site quality assurance and for planning further views that are tailored for a specific multi-view stereo approach with respect to the given scene. We demonstrate the capabilities of our approach with an autonomous Unmanned Aerial Vehicle (UAV) in a challenging outdoor scenario.

2.7 A low-cost mirror-based active perception system for effective collision free underwater robotic navigation

[Link to Paper]

This ongoing research work presents a servo actuated mirror-based design that allows a fixed front-view visual system mounted in an underwater vehicle to extend its field of view by controlling its gaze. We are interested in the autonomous underwater exploration of coral reefs. This type of exploration must involve a cautious and collision-free navigation to avoid damaging the marine ecosystem. Generally, vision systems of underwater vehicles are carefully isolated with mechanical seals to prevent the water from entering. However, this fact causes a strictly dependence between the angle of view of the camera and the pose of the vehicle. Furthermore, the addition of a system to control camera orientation may result in a significantly reduction of useful load capacity and the movement of the vision system could carry undesirable trusting effects, especially at higher speeds. Our design of servo actuated mirror system changes the angle of view of the camera in two degrees of freedom: pan and tilt, and reaches viewing angles from the sides, bottom top and even rear views of the robot, thus enabling a more effective navigation with obstacle avoidance.

2.8 Mobile Device Based Outdoor Navigation With On-line Learning Neural Network: a Comparison with Convolutional Neural Network

[Link to Paper]

Outdoor navigation is challenging with its dynamic environments and huge appearance variances. Traditional autonomous navigation systems construct 3D driving scenes to recognize open and occupied voxels by using laser range scanners, which are not available on mobile devices. Existing image-based navigation methods, on the other hand, are costly in computation and thus cannot be deployed onto a mobile device. To overcome these difficulties, we present an on-line learning neural network for real-time outdoor navigation using only the computational resources available on a standard android mobile device (i.e. camera, GPS, and no cloud back-end). The network is trained to recognize the most relevant object in current navigation setting and make corresponding decisions (i.e. adjust direction, avoid obstacles, and follow GPS). The network is compared with state of the art image classifier, the Convolutional Neural Network, in various aspects (i.e. network size, number of updates, convergence speed and final performance). Comparisons show that our network requires a minimal number of updates and converges significantly faster to better performance. The network successfully navigated in regular long-duration testing in novel settings and blindfolded testing under sunny and cloudy weather conditions.

3 Biometrics

3.1 Grouper: Optimizing Crowdsourced Face Annotations

[Link to Paper]

This study focuses on the problem of extracting consistent and accurate face bounding box annotations from crowdsourced workers. Aiming to provide benchmark datasets for facial recognition training and testing, we create a 'gold standard' set against which consolidated face bounding box annotations can be evaluated. An evaluation methodology based on scores for several features of bounding box annotations is presented and is shown to predict consolidation performance using information gathered from crowdsourced annotations. Based on this foundation, we present "Grouper," a method leveraging density-based clustering to consolidate annotations by crowd workers. We demonstrate that the proposed consolidation scheme, which should be extensible to any number of region annotation consolidations, improves upon metadata released with the IARPA Janus Benchmark-A. Finally, we compare FR performance using the originally provided IJB-A annotations and Grouper and determine that similarity to the gold standard as measured by our evaluation metric does predict recognition performance.

3.2 Soft-margin learning for multiple feature-kernel combinations with Domain Adaptation, for recognition in surveillance face datasets

[Link to Paper]

Face recognition (FR) is the most preferred mode for biometric-based surveillance, due to its passive nature of detecting subjects, amongst all different types of biometric traits. FR under surveillance scenario does not give satisfactory performance due to low contrast, noise and poor illumination conditions on probes, as compared to the training samples. A state-of-the-art technology, Deep Learning, even fails to perform well in these scenarios. We propose a novel soft-margin based learning method for multiple feature-kernel combinations, followed by feature transformed using Domain Adaptation, which outperforms many recent state-of-the-art techniques, when tested using three real-world surveillance face datasets.

3.3 What do you do when you know that you don't know?

[Link to Paper]

Real-world biometrics recognition problems often have two unknowns: the person be recognized, as well as a hidden unknown - missing data. If we choose to ignore data that is occasionally missing, we sacrifice accuracy. In this paper, we present a novel technique

to address the problem of handling missing data in biometrics systems without having to make implicit assumptions on the distribution of the underlying data. We introduce the concept of "operational adaptation" for biometric systems and formalize the problem. We present a solution for handling missing data based on refactoring on Support Vector Machines for large scale face recognition tasks. We also develop a general approach to estimating SVM refactoring risk. We present experiments on large-scale face recognition based on describable visual attributes on LFW dataset. Our approach consistently outperforms state-of-the-art methods designed to handle missing data. Figure 1. A system for describable visual attributes for faces based on [30] and extended for open-set recognition is shown above. In the image, green text is a positive attribute, red text is negative attributes and blue color signifies unknown/missing attribute. In the above images, the left image shows how bad lighting/featuredetection led to "UNKNOWN" labels for Asian, While attributes. The example on the right shows occlusion leading to the no beard attribute being labeled "UNKNOWN". Handling such unknowns at run-time, in a learning-based system, poses considerable operational challenges. This paper is about what the system should do when it knows that it does not know about some features.

3.4 Frequency map by Structure Tensor in Logarithmic Scale Space and Forensic Fingerprints

[Link to Paper]

Increasingly, absolute frequency and orientation maps are needed, e.g. for forensics. We introduce a non-linear scale space via the logarithm of trace of the Structure Tensor. Therein, frequency estimation becomes an orientation estimation problem. We show that this offers significant advantages, including construction of efficient isotropic estimations of dense maps of frequency. In fingerprints, both maps are shown to improve each other in an enhancement scheme via Gabor filtering. We suggest a novel continuous ridge counting method, relying only on dense absolute frequency and orientation maps, without ridge detection, thinning, etc. Furthermore, we present new evidence that frequency maps are useful attributes of minutiae. We verify that the suggested method compares favorably with state of the art using forensic fingerprints as test bed, and test images where the ground truth is known. In evaluations, we use public data sets and published methods only.

3.5 A Comparison of Human and Automated Face Verification Accuracy on Unconstrained Image Sets

[Link to Paper]

Automatic face recognition technologies have seen significant improvements in performance due to a combination of advances in deep learning and availability of larger datasets for training deep networks. Since recognizing faces is a task that humans are believed to be very good at, it is only natural to compare the relative performance of automated face recognition and humans when processing fully unconstrained facial imagery. In this work, we expand on previous studies of the recognition accuracy of humans and automated systems by performing several novel analyses utilizing unconstrained face imagery. We examine the impact on performance when human recognizers are presented with varying amounts of imagery per subject, immutable attributes such as gender, and circumstantial attributes such as occlusion, illumination, and pose. Results indicate that humans greatly outperform state of the art automated face recognition algorithms on the challenging IJB-A dataset.

3.6 Simultaneous Semi-Coupled Dictionary Learning for Matching RGBD data

[Link to Paper]

Matching with hidden information which is available only during training and not during testing has recently become an important research problem. Matching data from two different modalities, known as cross-modal matching is another challenging problem due to the large variations in the data coming from different modalities. Often, these are treated as two independent problems. But for applications like matching RGBD data, when only one modality is available during testing, it can reduce to either of the two problems. In this work, we propose a framework which can handle both these scenarios seamlessly with applications to matching RGBD data of Lambertian objects. The proposed approach jointly uses the RGB and depth data to learn an illumination invariant canonical version of the objects. Dictionaries are learnt for the RGB, depth and the canonical data, such that the transformed sparse coefficients of the RGB and the depth data is equal to that of the canonical data. Given RGB or depth data, their sparse coefficients corresponding to their canonical version is computed which can be directly used for matching using a Mahalanobis metric. Extensive experiments on three datasets, EURECOM, VAP RGB-D-T and Texas 3D Face Recognition database show the effectiveness of the proposed framework.

3.7 Deep Tattoo Recognition

[Link to Paper]

Tattoo is a soft biometric that indicates discriminative characteristics of a person such as beliefs and personalities. Automatic detection and recognition of tattoo images is a difficult problem. We present deep convolutional neural network-based methods for automatic

matching of tattoo images based on the AlexNet and Siamese networks. Furthermore, we show that rather than using a simple contrastive loss function, triplet loss function can significantly improve the performance of a tattoo matching system. Extensive experiments on a recently introduced Tatt-C dataset show that our method is able to capture the meaningful structure of tattoos and performs significantly better than many competitive tattoo recognition algorithms.

3.8 Latent Fingerprint Image Segmentation using Fractal Dimension Features and Weighted Extreme Learning Machine Ensemble

[Link to Paper]

Latent fingerprints are fingerprints unintentionally left at a crime scene. Due to the poor quality and often complex image background and overlapping patterns characteristic of latent fingerprint images, separating the fingerprint region-of-interest from complex image background and overlapping patterns is a very challenging problem. In this paper, we propose a latent fingerprint segmentation algorithm based on fractal dimension features and weighted extreme learning machine. We build feature vectors from the local fractal dimension features and use them as input to a weighted extreme learning machine ensemble classifier. The patches are classified into fingerprint and nonfingerprint classes. We evaluated the proposed segmentation algorithm by comparing the results with the published results from the state of the art latent fingerprint segmentation algorithms. The experimental results of our proposed approach show significant improvement in both the false detection rate (FDR) and overall segmentation accuracy compared to the existing approaches.

3.9 A Comprehensive Analysis of Deep Learning Based Representation for Face Recognition

[Link to Paper]

Deep learning based approaches have been dominating the face recognition field due to the significant performance improvement they have provided on the challenging wild datasets. These approaches have been extensively tested on such unconstrained datasets, on the Labeled Faces in the Wild and YouTube Faces, to name a few. However, their capability to handle individual appearance variations caused by factors such as head pose, illumination, occlusion, and misalignment has not been thoroughly assessed till now. In this paper, we present a comprehensive study to evaluate the performance of deep learning based face representation under several conditions including the varying head pose angles, upper and lower face occlusion, changing illumination of different strengths, and misalignment due to

erroneous facial feature localization. Two successful and publicly available deep learning models, namely VGG-Face and Lightened CNN have been utilized to extract face representations. The obtained results show that although deep learning provides a powerful representation for face recognition, it can still benefit from preprocessing, for example, for pose and illumination normalization in order to achieve better performance under various conditions. Particularly, if these variations are not included in the dataset used to train the deep learning model, the role of preprocessing becomes more crucial. Experimental results also show that deep learning based representation is robust to misalignment and can tolerate facial feature localization errors up to 10% of the interocular distance.

3.10 Implementation of Fixed-Length Template Protection Based on Homomorphic Encryption with Application to Signature Biometrics

[Link to Paper]

Given the high sensitivity of biometric data, any information leakage poses severe security and privacy risks. This fact raises the need to protect biometric templates so that no information can be learned from them, preserving at the same time the unprotected system's performance and speed. We propose a new efficient biometric template protection scheme based on homomorphic probabilistic encryption for fixed-length templates, where only encrypted data is handled. Under a fully reproducible experimental framework, we analyse different distance measures for the particular case of on-line signature, showing that all requirements described in the ISO/IEC 24745 standard on biometric information protection are met with no performance degradation and at a low computational cost. Furthermore, the proposed approach is robust to hill-climbing and inverse-biometrics attacks.

3.11 Pooling Faces: Template based Face Recognition with Pooled Face Images

[Link to Paper]

We propose a novel approach to template based face recognition. Our dual goal is to both increase recognition accuracy and reduce the computational and storage costs of template matching. To do this, we leverage on an approach which was proven effective in many other domains, but, to our knowledge, never fully explored for face images: average pooling of face photos. We show how (and why!) the space of a template's images can be partitioned and then pooled based on image quality and head pose and the effect this has on accuracy and template size. We perform extensive tests on the IJB-A and Janus CS2 template based face identification and verification benchmarks. These show that not only does our approach outperform published state of the art despite requiring far fewer cross

311

template comparisons, but also, surprisingly, that image pooling performs on par with deep feature pooling.

3.12 A Polarimetric Thermal Database for Face Recognition Research

[Link to Paper]

We present a polarimetric thermal face database, the first of its kind, for face recognition research. This database was acquired using a polarimetric longwave infrared imager, specifically a division-of-time spinning achromatic retarder system. A corresponding set of visible spectrum imagery was also collected, to facilitate crossspectrum (also referred to as heterogeneous) face recognition research. The database consists of imagery acquired at three distances under two experimental conditions: neutral/baseline condition, and expressions condition. Annotations (spatial coordinates of key fiducial points) are provided for all images. Cross-spectrum face recognition performance on the database is benchmarked using three techniques: partial least squares, deep perceptual mapping, and coupled neural networks.

3.13 DeepGender: Occlusion and Low Resolution Robust Facial Gender Classification via Progressively Trained Convolutional Neural Networks with Attention

[Link to Paper]

In this work, we have undertaken the task of occlusion and low-resolution robust facial gender classification. Inspired by the trainable attention model via deep architecture, and the fact that the periocular region is proven to be the most salient region for gender classification purposes, we are able to design a progressive convolutional neural network training paradigm to enforce the attention shift during the learning process. The hope is to enable the network to attend to particular high-profile regions (e.g. the periocular region) without the need to change the network architecture itself. The network benefits from this attention shift and becomes more robust towards occlusions and low-resolution degradations. With the progressively trained CNN models, we have achieved better gender classification results on the large-scale PCSO mugshot database with 400K images under occlusion and low-resolution settings, compared to the one undergone traditional training. In addition, our progressively trained network is sufficiently generalized so that it can be robust to occlusions of arbitrary types and at arbitrary locations, as well as low resolution.

3.14 Feature Vector Compression based on Least Error Quantization

[Link to Paper]

We propose a distinctive feature vector compression method based on least error quantization. This method can be applied to several biometrics methods using feature vectors, and allows us to significantly reduce the memory size of feature vectors without degrading the recognition performance. In this paper, we prove that minimizing quantization error between the compressed and original vectors is most effective to control the performance in face recognition. A conventional method uses non-uniform quantizer which minimizes the quantization error in terms of L2-distance. However, face recognition methods often use metrics other than L2-distance. Our method can calculate the quantized vectors in arbitrary metrics such as Lp-distance (0 ¡ p) and the quantized subspace basis. Furthermore, we also propose a fast algorithm calculating Lp-distances between two quantized vectors without decoding them. We evaluate the performance of our method on FERET, LFW and large face datasets with LBP (Lp-distance), Mutual Subspace Method and deep feature. The results show that the recognition rate using the quantized feature vectors is as accurate as that of the method using the original vectors even though the memory size of the vectors is reduced to 1/5 - 1/10. In particular, applying our method to the state-of-the-art feature, we are able to obtain the high performance feature whose size is very small.

3.15 Offline Signature Verification Based on Bag-of-Visual Words Model Using KAZE Features and Weighting Schemes

[Link to Paper]

The familiar use of handwritten signatures in various applications (e.g., credit card authentication) increases the need for automated verification methods. However, there is still room for improvement in the performance of automated systems under various writing conditions compared to human beings, especially forensic document examiners (FDEs). Furthermore, even with modern techniques, obtaining as much information as possible from the limited samples available remains challenging task. Therefore, further research is required to improve the performance of automated systems. In this study, to improve the performance of offline signature verification, a new approach based on a bag-of-visual words (BoVW) model is adopted. The novelty features of the proposed approach are following: 1) considering the cognitive processing of visual information by FDEs to improve the performance of offline signature verification, 2) using an approach based on the BoVW model to implement the FDEs' cognitive process for feature extraction, 3) incorporating weighting schemes based on term frequency-inverse document frequency to enhance the discriminative power of each visual word, 4) adopting KAZE features in the BoVW model to consider the contour information of strokes more effectively, and 5) detecting the KAZE

features in both the strokes and background space to introduce not only the stroke itself but also the various relations between strokes. The promising performance of the proposed approach is shown by using an evaluation method with a popular CEDAR signature dataset.

3.16   Deep Secure Encoding for Face Template Protection

[Link to Paper]

In this paper we present a framework for secure identification using deep neural networks, and apply it to the task of template protection for face password authentication. We use deep convolutional neural networks (CNNs) to learn a mapping from face images to maximum entropy binary (MEB) codes. The mapping is robust enough to tackle the problem of exact matching, yielding the same code for new samples of a user as the code assigned during training. These codes are then hashed using any hash function that follows the random oracle model (like SHA-512) to generate protected face templates. The algorithm makes no unrealistic assumptions and offers high template security, cancelability, and matching performance comparable to the state-of-the-art. The efficacy of the approach is shown on CMU-PIE, Extended Yale B, and Multi-PIE face databases. We achieve high (95%) genuine accept rates (GAR) at zero false accept rate (FAR) while maintaining a high level of template security.

3.17   CALIPER: Continuous Authentication Layered with Integrated PKI Encoding Recognition

[Link to Paper]

Architectures relying on continuous authentication require a secure way to challenge the user's identity without trusting that the Continuous Authentication Subsystem (CAS) has not been compromised, i.e., that the response to the layer which manages service/application access is not fake. In this paper, we introduce the CALIPER protocol, in which a separate Continuous Access Verification Entity (CAVE) directly challenges the user's identity in a continuous authentication regime. Instead of simply returning authentication probabilities or confidence scores, CALIPER's CAS uses live hard and soft biometric samples from the user to extract a cryptographic private key embedded in a challenge posed by the CAVE. The CAS then uses this key to sign a response to the CAVE. CALIPER supports multiple modalities, key lengths, and security levels and can be applied in two scenarios: One where the CAS must authenticate its user to a CAVE running on a remote server (device-server) for access to remote application data, and another where the CAS must authenticate its user to a locally running trusted computing module (TCM) for access to local application

data (device-TCM). We further demonstrate that CALIPER can leverage device hardware resources to enable privacy and security even when the device's kernel is compromised, and we show how this authentication protocol can even be expanded to obfuscate direct kernel object manipulation (DKOM) malwares.

3.18 PARAPH: Presentation Attack Rejection by Analyzing Polarization Hypotheses

[Link to Paper]

PARAPH Image TNLC Shu-ering with only 1 Alternate polarizer Polariza;on Person For applications such as airport border control, biometric technologies that can process many capture subjects quickly, efficiently, with weak supervision, and with minimal discomfort are desirable. Facial recognition is particularly appealing because it is minimally invasive yet offers relatively good recognition performance. Unfortunately, the combination of weak supervision and minimal invasiveness makes even highly accurate facial recognition systems susceptible to spoofing via presentation attacks. Thus, there is great demand for an effective and low cost system capable of rejecting such attacks. To this end we introduce PARAPH a novel hardware extension that exploits different measurements of light polarization to yield an image space in which presentation media are readily discernible from Bona Fide facial characteristics. The PARAPH system is inexpensive with an added cost of less than 10 US dollars. The system makes two polarization measurements in rapid succession, allowing them to be approximately pixel-aligned, with a frame rate limited by the camera, not the system. There are no moving parts above the molecular level, due to the efficient use of twisted nematic liquid crystals. We present evaluation images using three presentation attack media next to an actual face high quality photos on glossy and matte paper and a video of the face on an LCD. In each case, the actual face in the image generated by PARAPH is structurally discernible from the presentations, which appear either as noise (print attacks) or saturated images (replay attacks).

3.19 Gaussian Conditional Random Fields for Face Recognition

[Link to Paper]

We propose a Gaussian Conditional Random Field (GCRF) approach to modeling the nonstationary distortions that are introduced from changing facial expressions during acquisition. While previous work employed a Gaussian Markov Random Field (GMRF) to perform deformation tolerant matching of periocular images, we show that the approach is not well-suited for facial images, which can contain significantly larger and more complex deformations across the image. Like the GMRF, the GCRF tries to find the maximum

315

scoring assignment between a match pair in the presence of non-stationary deformations. However, unlike the GMRF, the GCRF directly computes the posterior probability that the observed deformation is consistent with the distortions exhibited in other authentic match pairs. The difference is the inclusion of a derived mapping between an input comparison and output deformation score. We evaluate performance on the CMU Multi-PIE facial dataset across all sessions and expressions, finding that the GCRF is significantly more effective at capturing naturally occurring large deformations than the previous GMRF approach.

3.20 GMM-SVM Fingerprint Verification Based on Minutiae Only

[Link to Paper]

Most fingerprint recognition systems use minutiae information, which is an unordered collection of minutiae locations and orientations. Template protection algorithms such as fuzzy commitment and other modern cryptographic alternatives based on homomorphic encryption require a fixed size binary template. However, such a template is not directly applicable to fingerprint minutiae representation which by its nature is of variable size. In this study, we introduce a novel method to represent a minutiae set with a rotation invariant fixed-length vector. We represent each minutia according to its geometric relation with neighbors and use Gaussian mixture model (GMM) to model its feature distribution. A two-class linear SVM is used to create a model template for the enrollment fingerprint sample, which discriminates impressions of the same finger from other fingers. We evaluated the verification performance of our method on the FVC2002DB1 database.

3.21 Real-Time Face Identification via CNN and Boosted Hashing Forest

[Link to Paper]

The family of real-time face representations is obtained via Convolutional Network with Hashing Forest (CNHF). We learn the CNN, then transform CNN to the multiple convolution architecture and finally learn the output hashing transform via new Boosted Hashing Forest (BHF) technique. This BHF generalizes the Boosted SSC approach for hashing learning with joint optimization of face verification and identification. CNHF is trained on CASIA-WebFace dataset and evaluated on LFW dataset. We code the output of single CNN with 97% on LFW. For Hamming embedding we get CBHF-200 bit (25 byte) code with 96.3% and 2000-bit code with 98.14% on LFW. CNHF with 20007-bit hashing trees achieves 93% rank1 on LFW relative to basic CNN 89.9% rank-1. CNHF generates templates at the rate of 40+ fps with CPU Core i7 and 120+ fps with GPU GeForce GTX 650.

3.22 Two-Stream CNNs for Gesture-Based Verification and Identification: Learning User Style

[Link to Paper]

Recently, gestures have been proposed as an alternative biometric modality to traditional biometrics such as face, fingerprint, iris and gait. As a biometric, gesture is a short body motion that contains static anatomical information and changing behavioral (dynamic) information. We consider two types of gestures: full-body gestures, such as a wave of the arms, and hand gestures, such as a subtle curl of the fingers and palm. Most prior work in this area evaluates gestures in the context of a "password," where each user has a single, chosen gesture motion. Contrary to prior work, we instead aim to learn a user's gesture "style" from a set of training gestures. We use two-stream convolutional neural networks, a form of deep learning, to learn this gesture style. First, we evaluate the generalization performance during testing of our approach against gestures or users that have not been seen during training. Then, we study the importance of dynamics by suppressing dynamic information in training and testing. We find that we are able to outperform state-of-the-art methods in identification and verification for two biometrics-oriented gesture datasets for body and in-air hand gestures. BodyLogin: Full-body gestures (captured with Kinect v1) Handlogin: In-air hand gestures (captured with Kinect v2)

3.23 Weakly Supervised Facial Analysis with Dense Hyper-column Features

[Link to Paper]

Weakly supervised methods have recently become one of the most popular machine learning methods since they are able to be used on large-scale datasets without the critical requirement of richly annotated data. In this paper, we present a novel, self-taught, discriminative facial feature analysis approach in the weakly supervised framework. Our method can find regions which are discriminative across classes yet consistent within a class and can solve many face related problems. The proposed method first trains a deep face model with high discriminative capability to extract facial features. The hypercolumn features are then used to give pixel level representation for better classification performance along with discriminative region detection. In addition, calibration approaches are proposed to enable the system to deal with multi-class and mixed-class problems. The system is also able to detect multiple discriminative regions from one image. Our uniform method is able to achieve competitive results in various face analysis applications, such as occlusion detection, face recognition, gender classification, twins verification and facial attractiveness analysis.

317

3.24 Heterogeneous Face Recognition using Inter-Session Variability Modelling

[Link to Paper]

The task of Heterogeneous Face Recognition consists in to match face images that were sensed in different modalities, such as sketches to photographs, thermal images to photographs or near infrared to photographs. In this preliminary work we introduce a novel and generic approach based on Inter-session Variability Modelling to handle this task. The experimental evaluation conducted with two different image modalities showed an average rank-1 identification rates of 96.93% and 72.39% for the CUHK-CUFS (Sketches) and CASIA NIR-VIS 2.0 (Near infra-red) respectively. This work is totally reproducible and all the source code for this approach is made publicly available. Face recognition has existed as a field of research for more than 30 years and has been particularly active since the early 1990s. Researchers of many different fields (from psychology, pattern recognition, neuroscience, computer graphics and computer vision) have attempted to create and understand the face recognition task [31]. One of the most challenging tasks in automated face recognition is the matching between face images acquired in heterogeneous environments. Use-cases can cover matching of faces in unconstrained scenarios (e.g. at a distance), with long time lapse between the probe and the gallery and faces sensed in different modalities, such as thermal infrared or near infrared images (NIR) against visible spectra images (VIS). Successful solutions to heterogeneous face recognition can extend the reach of these systems to covert scenarios, such as recognition at a distance or at night-time, or even in situations where no face even exists (forensic sketch recognition). The key difficulty in matching faces from heterogeneous conditions is that images of the same subject may differ in appearance due to changes in image modality (e.g. between VIS images and NIR images, between VIS images and sketches images) introducing high within class variations. With these variations, a direct comparison between samples generally results in poor matching accuracy [8]. Heterogeneous face recognition algorithms must develop facial representations invariant to these changes. This work proposes to approach the problem of Hetererogeneous Face Recognition (HFR) as a Session Varibility task, modelling the within-class variability using Gaussian Mixture Models (GM M). Experiments carried out with the CASIA NIR-VIS 2.0 Database and CUHK-Face Sketch Database (CUFS) shown competitive results with the current state-of-the-art results. Another contribution of this work is with respect to reproducibility. All the source code used to generate the results and plots are freely available for download. The documentation is done in such way that other researchers are able to reproduce them. The organization of the paper is the following. In Section 1 we present the prior work for heterogeneous face recognition. In Section 2 the proposed approach is presented in details. In Section 3 the experimental setup and results are presented. Finally in Section 4 the conclusions and future work are presented.

4   Perception Beyond the Visible Spectrum

4.1 Learning cross-spectral similarity measures with deep convolutional neural networks

[Link to Paper]

The simultaneous use of images from different spectra can be helpful to improve the performance of many computer vision tasks. The core idea behind the usage of crossspectral approaches is to take advantage of the strengths of each spectral band providing a richer representation of a scene, which cannot be obtained with just images from one spectral band. In this work we tackle the cross-spectral image similarity problem by using Convolutional Neural Networks (CNNs). We explore three different CNN architectures to compare the similarity of cross-spectral image patches. Specifically, we train each network with images from the visible and the near-infrared spectrum, and then test the result with two public cross-spectral datasets. Experimental results show that CNN approaches outperform the current state-of-art on both cross-spectral datasets. Additionally, our experiments show that some CNN architectures are capable of generalizing between different crossspectral domains.

4.2 A Novel Benchmark RGBD Dataset for Dormant Apple Trees and its Application to Automatic Pruning

[Link to Paper]

Dormant pruning is a necessary procedure in the field of specialty crop production. In order to mitigate the need of huge labor, automation of this pruning process has become a topic of utmost importance in the field of horticulture. 3D modeling and reconstruction is a major step in such robotics precision agriculture. In this paper, we introduce a new public dataset which can be used for reconstructing dormant apple trees. Our dataset comprises of 9 different apple trees in both indoor and outdoor evironment. The images are collected using a portable Kinect2 sensor. To the best of our knowledge, this is the first publicly available dataset for the application like 3D modeling of dormant trees. We hope that the dataset will provide the entire research community working towards mechanizing dormant pruning a baseline benchmark for evaluating different 3D reconstruction and modeling algorithms.

4.3 Evaluation of Feature Channels for Correlation-Filter-Based Visual Object Tracking in Infrared Spectrum

[Link to Paper]

Correlation filters for visual object tracking in visible imagery has been well-studied. Most of the correlation-filterbased methods use either raw image intensities or feature maps of gradient orientations or color channels. However, well-known features designed for visible spectrum may not be ideal for infrared object tracking, since infrared and visible spectra have dissimilar characteristics in general. We assess the performance of two state-of-the-art correlationfilter-based object tracking methods on Linkoping Thermal InfraRed (LTIR) dataset of medium wave and longwave infrared videos, using deep convolutional neural networks (CNN) features as well as other traditional hand-crafted descriptors. The deep CNN features are trained on an infrared dataset consisting of 16K objects for a supervised classification task. The highest performance in terms of the overlap metric is achieved when these deep CNN features are utilized in a correlation-filter-based tracker.

4.4 Real-Time Physiological Measurement and Visualization Using a Synchronized Multi-Camera System

[Link to Paper]

Remote physiological measurement has widespread implications in healthcare and affective computing. This paper presents an efficient system for remotely measuring heart rate and heart rate variability using multiple low-cost digital cameras in real-time. We combine an RGB camera, monochrome camera with color filter and a thermal camera to recover the blood volume pulse (BVP). We show that using multiple cameras in synchrony yields the most accurate recovery of the BVP signal. The RGB combination is not optimal. We show that the thermal camera improves performance of measurement under dynamic ambient lighting but the thermal camera alone is not enough and accuracy can be improved by adding more spectral channels. We present a real-time prototype that allows accurate physiological measurement combined with a novel user interface to visualize changes in heart rate and heart rate variability. Finally, we propose how this system might be used for applications such as patient monitoring.

4.5 Scale Invariant Human Action Detection from Depth Cameras using Class Templates

[Link to Paper]

We consider the problem of detecting and localizing a human action from continuous action video from depth cameras. We believe that this problem is more challenging than the problem of traditional action recognition as we do not have the information about the starting and ending frames of an action class. Another challenge which makes the problem difficult, is the latency in detection of actions. In this paper, we introduce a greedy

approach to detect the action class, invariant of their temporal scale in the testing sequences using class templates and basic skeleton based feature representation from the depth stream data generated using Microsoft Kinect. We evaluate the proposed method on the standard G3D and UTKinect-Action datasets consisting of five and ten actions, respectively. Our results demonstrate that the proposed approach performs well for action detection and recognition under different temporal scales, and is able to outperform the state of the art methods at low latency.

4.6 A Modular NMF Matching Algorithm for Radiation Spectra

[Link to Paper]

In real-world object identification systems, the operational mission may change from day to day. For example, a target recognition system may be searching for heavy armor one day, and surface-to-air assets the next, or a radiation detection system may be interested in detecting medical isotopes in one instance, and special nuclear material in another. To accommodate this "mission of the day" type scenario, the underlying object database must be flexible and able to adjust to changing target sets. Traditional dimensionality reduction algorithms rely on a single basis set that is derived from the complete set of objects of interest, making missionspecific adjustment a significant task. In this work, we describe a method that uses many limited-size individual basis sets to represent objects of interest instead of a single unifying basis set. Thus, only the objects of interest for the mission at hand are used at any given time, and additional objects can be added to the system simply by training a basis for the new object. We demonstrate the modular identification system on the problem of identifying radioisotopes from their gamma ray spectra using nonnegative matrix factorization.

4.7 A Novel Visualization Tool for Evaluating the Accuracy of 3D Sensing and Reconstruction Algorithms for Automatic Dormant Pruning Applications

[Link to Paper]

Agricultural tree pruning is one of the most essential procedures in specialty crop production. During winter, skilled workers have to remove certain primary branches of the dormant trees in order to increase the productivity of those trees. The process is particularly challenging as it needs huge workforce and costs, besides, seasonal availability of skilled workers. In this paper, we introduce a novel multifunctional graphical user interface which operates in two different modes, namely, manual and automatic, and serves the

following purposes - (1) validating 3D reconstruction algorithms, (2) automating the decision making process for selecting candidate branches to be pruned, (3) providing an easy interface with any robotic pruners, (4) training unskilled seasonal pruners, and (5) disseminating the knowledge. The paper describes the operating modes of the proposed tool and highlights various features and functionalities of the software by conducting experiments on an exemplar apple tree.

4.8   Non-planar Infrared-Visible Registration for Uncalibrated Stereo Pairs

[Link to Paper]

Thermal infrared-visible video registration for nonplanar scenes is a new area in visual surveillance. It allows the combination of information from two spectra for better human detection and segmentation. In this paper, we present a novel online framework for visible and thermal infrared registration for non-planar scenes that includes foreground segmentation, feature matching, rectification and disparity calculation. Our proposed approach is based on sparse correspondences of contour points. The key ideas of the proposed framework are the removal of spurious regions at the beginning of videos and a registration methodology for non-planar scenes. Besides, a new non-planar dataset with an associated evaluation protocol is also proposed as a standard assessment. We evaluate our method on both public planar and non-planar datasets. Experimental results reveal that the proposed method can not only successfully handle non-planar scenes but also gets state-of-the-art results on planar ones.

4.9   Seeing the Forest from the Trees: A Holistic Approach to Near-infrared Heterogeneous Face Recognition

[Link to Paper]

Heterogeneous face recognition is the problem of identifying a person from a face image acquired with a nontraditional sensor by matching it to a visible gallery. Most approaches to this problem involve modeling the relationship between corresponding images from the visible and sensing domains. This is typically done at the patch level and/or with shallow models with the aim to prevent overfitting. In this work, rather than modeling local patches or using a simple model, we propose to use a complex, deep model to learn the relationship between the entirety of cross-modal face images. We describe a deep convolutional neural network based method that leverages a large visible image face dataset to prevent overfitting. We present experimental results on two benchmark datasets showing its effectiveness.

4.10 Distinguishing Weather Phenomena from Bird Migration Patterns in Radar Imagery

[Link to Paper]

Data archived by the United States radar network for weather surveillance is useful in studying ecological phenomena such as the migration patterns of birds. However, all such methods require a manual screening stage from domain experts to eliminate radar signatures of weather phenomena, since the radar beam picks up both biological and non-biological targets. Automating this screening step would be of significant help to the large-scale study of ecological phenomenon from radar data. We apply several techniques to this novel task, comparing the performance of Convolutional Neural Networks (CNNs) models against a baseline of the Fisher Vector model on SIFT descriptors. We compare the performance of deeper and shallower network architectures, deep texture models versus the regular CNN model and the effect of fine-tuning ImageNet pre-trained networks on radar imagery. Fine-tuning the networks on the radar imagery provides a significant boost, and we achieve an accuracy of 94.4% on a dataset of 13,194 radar scans, 3,799 of which contained rain.

4.11 Adaptive Object Classification using complex SAR Signatures

[Link to Paper]

This paper addresses the problem associated with the classification of signatures of objects obtained by coherent sensors whereby the signatures are complex valued. Individual phase and amplitude component of a signature are combined optimally and the resulting fused signature is used in a sparsity-based learning classifier. The results of application of this approach are then compared with the corresponding results using only the amplitudes of the signatures. To test the concept public-domain radar signatures of several land vehicles obtained at different aspect angles are used. The performance improvement, based on confusion matrices, is shown to be significant when both phase and amplitudes are used

5 DeepVision: Deep Learning in Computer Vision

5.1 Learning by tracking: Siamese CNN for robust target association

[Link to Paper]

This paper introduces a novel approach to the task of data association within the context of pedestrian tracking, by introducing a two-stage learning scheme to match pairs of detections. First, a Siamese convolutional neural network (CNN) is trained to learn descriptors encoding local spatio-temporal structures between the two input image patches, aggregating pixel values and optical flow information. Second, a set of contextual features derived from the position and size of the compared input patches are combined with the CNN output by means of a gradient boosting classifier to generate the final matching probability. This learning approach is validated by using a linear programming based multi-person tracker showing that even a simple and efficient tracker may outperform much more complex models when fed with our learned matching probabilities. Results on publicly available sequences show that our method meets state-of-the-art standards in multiple people tracking. Detections at frame N Siamese CNN + Gradient Boosting Linear Programming Detections at frame M Predicted associations Final trajectories

5.2 Adversarial Diversity and Hard Positive Generation

[Link to Paper]

State-of-the-art deep neural networks suffer from a fundamental problem they misclassify adversarial examples formed by applying small perturbations to inputs. In this paper, we present a new psychometric perceptual adversarial similarity score (PASS) measure for quantifying adversarial images, introduce the notion of hard positive generation, and use a diverse set of adversarial perturbations not just the closest ones for data augmentation. We introduce a novel hot/cold approach for adversarial example generation, which provides multiple possible adversarial perturbations for every single image. The perturbations generated by our novel approach often correspond to semantically meaningful image structures, and allow greater flexibility to scale perturbation-amplitudes, which yields an increased diversity of adversarial images. We present adversarial images on several network topologies and datasets, including LeNet on the MNIST dataset, and GoogLeNet and ResidualNet on the ImageNet dataset. Finally, we demonstrate on LeNet and GoogLeNet that fine-tuning with a diverse set of hard positives improves the robustness of these networks compared to training with prior methods of generating adversarial images.

5.3 Faster R-CNN Features for Instance Search

[Link to Paper]

Image representations derived from pre-trained Convolutional Neural Networks (CNNs) have become the new state of the art in computer vision tasks such as instance retrieval. This work explores the suitability for instance retrieval of image- and region-wise representations pooled from an object detection CNN such as Faster R-CNN. We take advantage of the object proposals learned by a Region Proposal Network (RPN) and their associated CNN features to build an instance search pipeline composed of a first filtering stage followed by a spatial reranking. We further investigate the suitability of Faster R-CNN features when the network is fine-tuned for the same objects one wants to retrieve. We assess the performance of our proposed system with the Oxford Buildings 5k, Paris Buildings 6k and a subset of TRECVid Instance Search 2013, achieving competitive results.

5.4 Deep End2End Voxel2Voxel Prediction

[Link to Paper]

Over the last few years deep learning methods have emerged as one of the most prominent approaches for video analysis. However, so far their most successful applications have been in the area of video classification and detection, i.e., problems involving the prediction of a single class label or a handful of output variables per video. Furthermore, while deep networks are commonly recognized as the best models to use in these domains, there is a widespread perception that in order to yield successful results they often require time-consuming architecture search, manual tweaking of parameters and computationally intensive preprocessing or post-processing methods. In this paper we challenge these views by presenting a deep 3D convolutional architecture trained end to end to perform voxel-level prediction, i.e., to output a variable at every voxel of the video. Most importantly, we show that the same exact architecture can be used to achieve competitive results on three widely different voxel-prediction tasks: video semantic segmentation, optical flow estimation, and video coloring. The three networks learned on these problems are trained from raw video without any form of preprocessing and their outputs do not require post-processing to achieve outstanding performance. Thus, they offer an efficient alternative to traditional and much more computationally expensive methods in these video domains.

5.5 Rich Image Captioning in the Wild

[Link to Paper]

We present an image caption system that addresses new challenges of automatically describing images in the wild. The challenges include generating high quality caption with respect to human judgments, out-of-domain data handling, and low latency required in

327

many applications. Built on top of a state-of-the-art framework, we developed a deep vision model that detects a broad range of visual concepts, an entity recognition model that identifies celebrities and landmarks, and a confidence model for the caption output. Experimental results show that our caption engine outperforms previous state-of-the-art systems significantly on both in-domain dataset (i.e. MS COCO) and out-of-domain datasets. We also make the system publicly accessible as a part of the Microsoft Cognitive Services. "Sasha Obama, Malia Obama, Michelle Obama, Peng Liyuan et al. posing for a picture with Forbidden City in the background."

5.6 ReSeg: A Recurrent Neural Network-based Model for Semantic Segmentation

[Link to Paper]

We propose a structured prediction architecture, which exploits the local generic features extracted by Convolutional Neural Networks and the capacity of Recurrent Neural Networks (RNN) to retrieve distant dependencies. The proposed architecture, called ReSeg, is based on the recently introduced ReNet model for image classification. We modify and extend it to perform the more challenging task of semantic segmentation. Each ReNet layer is composed of four RNN that sweep the image horizontally and vertically in both directions, encoding patches or activations, and providing relevant global information. Moreover, ReNet layers are stacked on top of pre-trained convolutional layers, benefiting from generic local features. Upsampling layers follow ReNet layers to recover the original image resolution in the final predictions. The proposed ReSeg architecture is efficient, flexible and suitable for a variety of semantic segmentation tasks. We evaluate ReSeg on several widely-used semantic segmentation datasets: Weizmann Horse, Oxford Flower, and CamVid; achieving stateof-the-art performance. Results show that ReSeg can act as a suitable architecture for semantic segmentation tasks, and may have further applications in other structured prediction problems. The source code and model hyperparameters are available on https://github.com/fvisin/reseg. Dipartimento di Elettronica Informazione e Bioingegneria, Politecnico di Milano, Milan, 20133, Italy Montreal Institute for Learning Algorithms (MILA), University of Montreal, Montreal, QC, H3T 1J4, Canada Courant Institute and Center for Data Science, New York University, New York, NY 10012, United States CIFAR Senior Fellow

5.7 Joint Learning of Convolutional Neural Networks and Temporally Constrained Metrics for Tracklet Association

[Link to Paper]

328

In this paper, we study the challenging problem of multiobject tracking in a complex scene captured by a single camera. Different from the existing tracklet associationbased tracking methods, we propose a novel and efficient way to obtain discriminative appearance-based tracklet affinity models. Our proposed method jointly learns the convolutional neural networks (CNNs) and temporally constrained metrics. In our method, a siamese convolutional neural network (CNN) is first pre-trained on the auxiliary data. Then the siamese CNN and temporally constrained metrics are jointly learned online to construct the appearance-based tracklet affinity models. The proposed method can jointly learn the hierarchical deep features and temporally constrained segment-wise metrics under a unified framework. For reliable association between tracklets, a novel loss function incorporating temporally constrained multi-task learning mechanism is proposed. By employing the proposed method, tracklet association can be accomplished even in challenging situations. Moreover, a largescale dataset with 40 fully annotated sequences is created to facilitate the tracking evaluation. Experimental results on five public datasets and the new large-scale dataset show that our method outperforms several state-of-the-art approaches in multi-object tracking.

6 Egocentric Vision

6.1 Body Part Based Re-identification from an Egocentric Perspective

[Link to Paper]

With the spread of wearable cameras, many consumer applications ranging from social tagging to video summarization would greatly benefit from people re-identification methods capable of dealing with the egocentric perspective. In this regard, first-person camera views present such a unique setting that traditional re-identification methods results in poor performance when applied to this scenario. In this paper, we present a simple but effective solution that overcomes the limitations of traditional approaches by dividing people images into meaningful body parts. Furthermore, by taking into account human gaze information concerning where people look at when trying to recognize a person, we devise a meaningful way to weight the contributions of different bodyparts. Experimental results validate the proposal on a novel egocentric re-identification dataset, the first of its kind, showing that the performance increases when compared to current state of the art on egocentric sequences is significant.

6.2 A Pointing Gesture Based Egocentric Interaction System: Dataset, Approach and Application

[Link to Paper]

With the heated trend of augmented reality (AR) and popularity of smart head-mounted devices, the development of natural human device interaction is important, especially the hand gesture based interaction. This paper presents a solution for the point gesture based interaction in the egocentric vision and its application. Firstly, a dataset named EgoFinger is established focusing on the pointing gesture for the egocentric vision. We discuss the dataset collection detail and as well the comprehensive analysis of this dataset, including background and foreground color distribution, hand occurrence likelihood, scale and pointing angle distribution of hand and finger, and the manual labeling error analysis. The analysis shows that the dataset covers substantial data samples in various environments and dynamic hand shapes. Furthermore, we propose a two-stage Faster R-CNN based hand detection and dual-target fingertip detection framework. Comparing with state-of-art tracking and detection algorithm, it performs the best in both hand and fingertip detection. With the large-scale dataset, we achieve fingertip detection error at about 12.22 pixels in 640px 480px video frame. Finally, using the fingertip detection result, we design and implement an input system for the egocentric vision, i.e., Ego-Air-Writing. By considering the fingertip as a pen, the user with wearable glass can write character in the air and interact with system using simple hand gestures.

6.3 Discovering Objects of Joint Attention via First-Person Sensing

[Link to Paper]

The goal of this work is to discover objects of joint attention, i.e., objects being viewed by multiple people using head-mounted cameras and eye trackers. Such objects of joint attention are expected to act as an important cue for understanding social interactions in everyday scenes. To this end, we develop a commonality-clustering method tailored to first-person videos combined with points-of-gaze sources. The proposed method uses multiscale spatiotemporal tubes around points of gaze as a candidate of objects, making it possible to deal with various sizes of objects observed in the first-person videos. We also introduce a new dataset of multiple pairs of first-person videos and points-of-gaze data. Our experimental results show that our approach can outperform several state-of-the-art commonality-clustering methods. Person 1's first person video 12 45 Person 2's first person video 12 45 3 6 3 6

6.4 Multimodal Multi-stream Deep Learning for Egocentric Activity Recognition

[Link to Paper]

In this paper, we propose a multimodal multi-stream deep learning framework to tackle the egocentric activity recognition problem, using both the video and sensor data. First, we experiment and extend a multi-stream Convolutional Neural Network to learn the spatial and temporal features from egocentric videos. Second, we propose a multistream Long Short-Term Memory architecture to learn the features from multiple sensor streams (accelerometer, gyroscope, etc.). Third, we propose to use a two-level fusion technique and experiment different pooling techniques to compute the prediction results. Experimental results using a multimodal egocentric dataset show that our proposed method can achieve very encouraging performance, despite the constraint that the scale of the existing egocentric datasets is still quite limited.

7 Embedded Vision

7.1 Embedded Vision System for Atmospheric Turbulence Mitigation

[Link to Paper]

Outdoor surveillance systems that involve farfield operations often encounter atmospheric turbulence perturbations due to a series of randomized reflections and refraction effecting incoming light rays. The resulting distortions make it hard to discriminate between true moving objects and turbulence induced motion. Current algorithms are not effective in detecting true moving objects in the scene and also rely on computationally complex warping methods. In this paper, we describe a real time embedded solution connected with traditional cameras to both rectify turbulence distortions and reliably detect and track true moving targets. Our comparisons with other methods shows better turbulence rectification with less false and miss detections. FPGADSP based embedded realization of our algorithm achieves nearly 15x speed-up along with lesser memory requirement over a quad core PC implementation. The proposed system is suitable for persistence surveillance systems and optical sight devices.

7.2 A Scalable High-Performance Hardware Architecture for Real-Time Stereo Vision by Semi-Global Matching

[Link to Paper]

Perceiving distance from two camera images, a task called stereo vision, is fundamental for many applications in robotics or automation. However, algorithms that compute this information at high accuracy have a high computational complexity. One such algorithm, Semi Global Matching (SGM), performs well in many stereo vision benchmarks, while maintaining a manageable computational complexity. Nevertheless, CPU and GPU implementations of this algorithm often fail to achieve real-time processing of camera images, especially in power-constrained embedded environments. This work presents a novel architecture to calculate disparities through SGM. The proposed architecture is highly scalable and applicable for low-power embedded as well as high-performance multicamera high-resolution applications.

7.3 Visual Monocular Obstacle Avoidance for Small Unmanned Vehicles

[Link to Paper]

This paper presents and extensively evaluates a visual obstacle avoidance method using frames of a single camera, intended for application on small devices (ground or aerial

robots or even smartphones). It is based on image region classification using so called relative focus maps, it does not require a priori training, and it is applicable in both indoor and outdoor environments, which we demonstrate through evaluations using both simulated and real data.

7.4 Embedded Computing Framework for Vision-based Real-time Surround Threat Analysis and Driver Assistance

[Link to Paper]

In this paper, we present a distributed embedded vision system that enables surround scene analysis and vehicle threat estimation. The proposed system analyzes the surroundings of the ego-vehicle using four cameras, each connected to a separate embedded processor. Each processor runs a set of optimized vision-based techniques to detect surrounding vehicles, so that the entire system operates at real-time speeds. This setup has been demonstrated on multiple vehicle testbeds with high levels of robustness under real-world driving conditions and is scalable to additional cameras. Finally, we present a detailed evaluation which shows over 95% accuracy and operation at nearly 15 frames per second. Figure 1: Surround scene assessment using cameras and embedded processors proposed.

7.5 3DCapture: 3D Reconstruction for a Smartphone

[Link to Paper]

We propose a method of reconstruction of 3D representation (a mesh with a texture) of an object on a smartphone with a monocular camera. The reconstruction consists of two parts real-time scanning around the object and postprocessing. At the scanning stage IMU sensors data are acquired along with tracks of features in video. A special care is taken to comply with 360 scan requirement. All these data are used to build a camera trajectory using bundle adjustment techniques after scanning is completed. This trajectory is used in calculation of depth maps, which then are used to construct a polygonal mesh with overlaid textures. The proposed method ensures tracking at 30 fps on a modern smartphone while the post-processing part is completed within 1 minute using an OpenCL compatible mobile GPU. In addition, we show that with a few modifications this algorithm can be adopted for human face reconstruction.

7.6 Embedded Motion Detection via Neural Response Mixture Background Modeling

[Link to Paper]

Recent studies have shown that deep neural networks (DNNs) can outperform state-of-the-art algorithms for a multitude of computer vision tasks. However, the ability to leverage DNNs for near real-time performance on embedded systems have been all but impossible so far without requiring specialized processors or GPUs. In this paper, we present a new motion detection algorithm that leverages the power of DNNs while maintaining low computational complexity needed for near real-time embedded performance without specialized hardware. The proposed Neural Response Mixture (NeRM) model leverages rich deep features extracted from the neural responses of an efficient, stochastically-formed deep neural network (StochasticNet) for constructing Gaussian mixture models to detect motion in a scene. NeRM was implemented embedded on an Axis surveillance camera, and results demonstrated that the proposed NeRM approach can achieve strong motion detection accuracy while operating at near real-time performance.

7.7 Real-time, Embedded Scene Invariant Crowd Counting Using Scale-Normalized Histogram of Moving Gradients (HoMG)

[Link to Paper]

Automated crowd counting has garnered significant interest for video surveillance. This paper proposes a novel scene invariant crowd counting algorithm designed for high accuracy yet low computational complexity in order to facilitate widespread use in real-time embedded video analytics systems. A novel low-complexity, scale-normalized feature called Histogram of Moving Gradients (HoMG) is introduced for highly effective spatiotemporal representation of crowds within a video. Real-time crowd region detection is achieved via boosted cascade of weak classifiers based on HoMG features. Based on the detected crowd regions, linear support vector regression (SVR) of crowd-region HoMG features is introduced for real-time crowd counting. Experimental results using a multi-scene crowd dataset show that the proposed algorithm outperforms state-of-the-art crowd counting algorithms while embedded on modern surveillance cameras. Thus demonstrating the efficacy of the proposed method for accurate, real-time, embedded crowd analysis.

7.8 Vision based autonomous orientational control for aerial manipulation via on-board FPGA

[Link to Paper]

336

We describe an FPGA-based on-board control system for autonomous orientation of an aerial robot to assist aerial manipulation tasks. The system is able to apply yaw control to aid an operator to precisely position a drone when it is nearby a bar-like object. This is achieved by applying parallel Hough transform enhanced with a novel image space separation method, enabling highly reliable results in various circumstances combined with high performance. The feasibility of this approach is shown by applying the system to a multi-rotor aerial robot equipped with an upward directed robotic hand on top of the airframe developed for high altitude manipulation tasks. In order to grasp a barlike object, orientation of the bar object is observed from the image data obtained by a monocular camera mounted on the robot. This data is then analyzed by the on-board FPGA system to control yaw angle of the aerial robot. In experiments, reliable yaw-orientation control of the aerial robot is achieved.

7.9 A Visual Attention Algorithm Designed for Coupled Oscillator Acceleration

[Link to Paper]

We present a new top-down and bottom-up saliency algorithm designed to exploit the capabilities of coupled oscillators: an ultra-low-power, high performance, non-boolean computer architecture designed to serve as a special purpose embedded accelerator for vision applications. To do this, we extend a widely used neuromorphic bottom-up saliency pipeline by introducing a top-down channel which looks for objects of a particular type. The proposed channel relies on a segmentation of the input image to identify exemplar object segments resembling those encountered in training. The channel leverages pre-computed bottom-up feature maps to produce a novel scale-invariant descriptor for each segment with little computational overhead. We also introduce a new technique to automatically determine exemplar segments during training, without the need for annotations per segment. We evaluate our method on both NeoVision2 DARPA challenge datasets, illustrating significant gains in performance compared to all baseline approaches.

7.10 Approximated Prediction Strategy for Reducing Power Consumption of Convolutional Neural Network Processor

[Link to Paper]

Convolutional neural network (CNN) is becoming popular because of its great ability for accurate image recognition. However, the computational cost is extremely high, which increases power consumption of embedded CV systems. This paper proposes an efficient computing method, LazyConvPool (LCP), and its hardware architecture to reduce power

consumption of CNN-based image recognition. The LCP exploits redundancy of operations in CNN and only executes essential convolutions by an approximated prediction technique. We also propose Sign Connect, which is a low computational-cost approximated prediction without any multiplications. The experimental evaluation using image classification dataset shows that the proposed method reduces the power consumption by 17.8%20.2% and energy consumption by 11.4%14.1% while retaining recognition performance.

7.11   A Diverse Low Cost High Performance Platform for Advanced Driver Assistance System (ADAS) Applications

[Link to Paper]

Advanced driver assistance systems (ADAS) are becoming more and more popular. Lot of the ADAS applications such as Lane departure warning (LDW), Forward Collision Warning (FCW), Automatic Cruise Control (ACC), Auto Emergency Braking (AEB), Surround View (SV) that were present only in high-end cars in the past have trickled down to the low and mid end vehicles. Lot of these applications are also mandated by safety authorities such as EUNCAP and NHTSA. In order to make these applications affordable in the low and mid end vehicles, it is important to have a cost effective, yet high performance and low power solution. Texas Instruments (TI's) TDA3x is an ideal platform which addresses these needs. This paper illustrates mapping of different algorithms such as SV, LDW, Object detection (OD), Structure From Motion (SFM) and Camera-Monitor Systems (CMS) to the TDA3x device, thereby demonstrating its compute capabilities. We also share the performance for these embedded vision applications, showing that TDA3x is an excellent high performance device for ADAS applications.

8 Biomedical Image Registration

8.1 Multimodal Whole Brain Registration: MRI and High Resolution Histology

[Link to Paper]

Three-dimensional brain imaging through cutting-edge MRI technology allows assessment of physical and chemical tissue properties at sub-millimeter resolution. In order to improve brain understanding as part of diagnostic tasks using MRI images, other imaging modalities to obtain deep cerebral structures and cytoarchitectural boundaries have been investigated. Under availability of postmortem samples, the fusion of MRI to brain histology supports more accurate description of neuroanatomical structures since it preserves microscopic entities and reveal fine anatomical details, unavailable otherwise. Nonetheless, histological processing causes severe tissue deformation and loss of the brain original 3D conformation, preventing direct comparisons between MRI and histology. This paper proposes an interactive computational pipeline designed to register multimodal brain data and enable direct histology-MRI correlation. Our main contribution is to develop schemes for brain data fusion, distortion corrections, using appropriate diffeomorphic mappings to align the 3D histological and MRI volumes. We describe our pipeline and preliminary developments of scalable processing schemes for highresolution images. Tests consider a postmortem human brain, and include qualitatively and quantitatively results, such as 3D visualizations and the Dice coefficient (DC) be- tween brain structures. Preliminary results show promising DC values when comparing our scheme results to manually labeled neuroanatomical regions defined by a neurosurgeon on MRI and histology data sets. DC was computed for the left caudade gyrus (LC), right hippocampus (RH) and lateral ventricles (LV).

8.2 The design of S U P E Relastix a unifying framework for a wide range of image registration methodologies

[Link to Paper]

A large diversity of image registration methodologies has emerged from the research community. The scattering of methods over toolboxes impedes rigorous comparison to select the appropriate method for a given application. Toolboxes typically tailor their implementations to a mathematical registration paradigm, which makes internal functionality nonexchangeable. Subsequently, this forms a barrier for adoption of registration technology in the clinic. We therefore propose a unifying, role-based software design that can integrate a broad range of functional registration components. These components can be configured into an algorithmic network via a single highlevel user interface. A generic component handshake mechanism provides users feedback on incompatibilities. We demonstrate the

viability of our design by incorporating two paradigms from different code bases. The implementation is done in C++ and is available as open source. The progress of embedding more paradigms can be followed via https://github.com/kaspermarstal/SuperElastix

8.3 Registration of Developmental Image Sequences with Missing Data

[Link to Paper]

Longitudinal image registration is commonly used to establish spatial correspondence between images when investigating temporal changes in brain morphology. Most image registration methods have been developed to align images that are similar in appearance or structure. If such similarity is not given (e.g., in the case of neurodevelopmental studies, which is the target application of this paper), (i) local similarity measures, (ii) metamorphosis approaches, or (iii) methods modeling longitudinal intensity change can be used. Methods modeling longitudinal intensity change have the advantage of not treating images as independent static samples. However, missing or incomplete data can lead to poor model estimation and, in turn, poor registration. Therefore, incomplete longitudinal data sets are often excluded from analysis. Here, we propose a method to build a longitudinal atlas of intensity change and incorporate it as a prior into an existing model-based registration method. We show that using the prior can guide the deformable registration of longitudinal images of brain development with missing data and produce comparable registration results to complete data sets.

8.4 Multi-Atlas based pseudo-CT Synthesis using Multimodal Image Registration and Local Atlas Fusion Strategies

[Link to Paper]

The synthesis of pseudo-CT images is of particular interest in the development of hybrid PET-MRI devices and MRI-guided radiotherapy. These images can be used for attenuation correction during PET image reconstruction. Furthermore, using MRI-based radiotherapy planning would enable a more accurate dosimetry planning due to the superior soft tissue contrast of the scans. The previously proposed methods for pseudo-CT synthesis are characterised by mainly two drawbacks. First, most proposed methods are limited to the head and neck region and therefore not feasible in case of whole body applications. Second, the presence of aligned training pairs of both MRI and CT scans for a number of subjects is assumed. In this work, we present preliminary results for atlas-based approaches using multiple CT atlas scans (from different patients) to synthesise a pseudo-CT image for a new patient using only their MRI data. This application requires accurate and robust deformable multimodal registration. We employed a recent discrete optimisation registration

framework together with a self-similarity-based metric to accurately match the CT atlases to the anatomy of the patient. The registered atlases are then jointly combined by means of local fusion strategies. We apply our method to different 3D whole body MRI scans and a total of 18 3D whole body CT atlases. In addition to intensity fusion, the proposed methods can also be used for label fusion. Since evaluation based directly on synthesised intensity values is problematic, we use the Dice overlap after the fusion of segmentation labels as a proxy measure. Our proposed new method, which uses MIND descriptors for multimodal label fusion shows overall the best results.

8.5 How to build an average model when samples are variably incomplete? Application to fossil data

[Link to Paper]

In paleontology, incomplete samples with small or large missing parts are frequently encountered. For example, dental crowns, which are widely studied in paleontology because of their potential interest in taxonomic and phylogenetic analyses, are nearly systematically affected by a variable degree of wear that alters considerably their shape. It is then difficult to compute a significant reference surface model based on classical methods which are used to build atlases from set of samples. In this paper, we present a general approach to deal with the problem of estimating an average model from a set of incomplete samples. Our method is based on a state-of-the-art non-rigid surface registration algorithm. In a first step, we detect missing parts which allows one to focus only on the common parts to get an accurate registration result. In a second step, we try to build average model of the missing parts by using information which is available in a subset of the samples. We specifically apply our method on teeth, and more precisely on the surface in between dentine and enamel tissues (EDJ). We investigate the robustness and accuracy properties of the methods on a set of artificial samples representing a high degree of incompleteness. We compare the reconstructed complete shape to a ground-truth dataset. We then show some results on real data.

8.6 Accurate small deformation exponential approximant to integrate large velocity fields:

[Link to Paper]

One of the basic components of diffeomorphic image registration algorithms based on velocity fields is the numerical method used to integrate velocity parameters and obtain spatial transformations as displacement fields. When the input velocity field does not depend on the time parameter, the solution is often referred to as the Lie exponential

of the velocity field. In this work, we present an integration method for its numerical computation based both on a generalization of the scaling and squaring algorithm and on a class of numerical integrators aimed to solve systems of ordinary differential equations called exponential integrators. This new method led to the introduction of three numerical integrators, and the subsequent validation are performed on synthetic deformations and real medical images.

8.7 Image Registration for Placenta Reconstruction

[Link to Paper]

In this paper we introduce a method to handle the challenges posed by image registration for placenta reconstruction from fetoscopic video as used in the treatment of Twinto-Twin Transfusion Syndrome (TTTS). Panorama reconstruction of the placenta greatly supports the surgeon in obtaining a complete view of the placenta to localize vascular anastomoses. The found shunts can subsequently be blocked by coagulation in the correct order. By using similarity learning in training a Convolutional Neural Network we created a novel feature extraction method, allowing robust matching of keypoints for image registration and therefore taking the most critical step in placenta reconstruction from fetoscopic video. The fetoscopic video we used for our experiments was acquired from a training simulator for TTTS surgery. We compared our method with state-of-the-art methods. The matching performance of our method is up to three times better while the mean projection error is reduced with 64% for the registered images. Our image registration method provides the ground work for a complete panorama reconstruction of the placenta.

8.8 Total correlation-based groupwise image registration for quantitative MRI

[Link to Paper]

In quantitative magnetic resonance imaging (qMRI), quantitative tissue properties can be estimated by fitting a signal model to the voxel intensities of a series of images acquired with different settings. To obtain reliable quantitative measures, it is necessary that the qMRI images are spatially aligned so that a given voxel corresponds in all images to the same anatomical location. The objective of the present study is to describe and evaluate a novel automatic groupwise registration technique using a dissimilarity metric based on an approximated form of total correlation. The proposed registration method is applied to five qMRI datasets of various anatomical locations, and the obtained registration performances are compared to these of a conventional pairwise registration based on mutual information. The results show that groupwise total correlation yields better registration

343

performances than pairwise mutual information. This study also establishes that the formulation of approximated total correlation is quite analogous to two other groupwise metrics based on principal component analysis (PCA). Registration performances of total correlation and these two PCA-based techniques are therefore compared. The results show that total correlation yields performances that are analogous to these of the PCAbased techniques. However, compared to these PCA-based metrics, total correlation has two main advantages. Firstly, it is directly derived from a multivariate form of mutual information, while the PCA-based metrics were obtained empirically. Secondly, total correlation has the advantage of requiring no user-defined parameter.

8.9 Sparse Kernel Machines for Discontinuous Registration and Nonstationary Regularization

[Link to Paper]

We present a novel approach where we address image registration with the concept of a sparse kernel machine. We formulate the registration problem as a regularized minimization functional where a reproducing kernel Hilbert space is used as transformation model. The regularization comprises a sparsity inducing l1-type norm and a well known l2 norm. We prove a representer theorem for this type of functional to guarantee a finite dimensional solution. The presented method brings the advantage of flexibly defining the admissible transformations by choosing a positive definite kernel jointly with an efficient sparse representation of the solution. As such, we introduce a new type of kernel function, which enables discontinuities in the transformation and simultaneously has nice interpolation properties. In addition, location-dependent smoothness is achieved within the same framework to further improve registration results. Finally, we make use of an adaptive grid refinement scheme to optimize on multiple scales and for a finer control point grid at locations of high gradients. We evaluate our new method with a public thoracic 4DCT dataset.

8.10 Optimal estimation of diffusion in DW-MRI by high-order MRF-based joint deformable registration and diffusion modeling

[Link to Paper]

Over the last years, the apparent diffusion coefficient (ADC), computed from diffusion-weighted magnetic resonance (DW-MR) images, has become an important imaging biomarker for evaluating and managing patients with neoplastic or cerebrovascular disease. Standard methods for the calculation of ADC ignore the presence of noise and motion between successive (in time) DW-MR images acquired by changing the b-value. In order to accurately

quantify the diffusion process during image acquisition, we introduce a method based on a high-order Markov Random Field (MRF) formulation that jointly registers the DWMR images and models the spatiotemporal diffusion. Spatial smoothness on the ADC map, as well as spatiotemporal deformation smoothness, is imposed towards producing anatomically meaningful representations. The high-order dependencies in our MRF model are handled through Dual Decomposition. Performance of registration is compared to a state-of-the art registration approach in terms of obtained fitting error of the diffusion model in the core of the tumor. Preliminary results reveal a marginally better performance of our method when compared against the standard ADC map used in clinical practice, which indicates its potential as a means for extracting imaging biomarkers.

8.11   SimpleElastix: A user-friendly, multi-lingual library for medical image registration

[Link to Paper]

In this paper we present SimpleElastix, an extension of SimpleITK designed to bring the Elastix medical image registration library to a wider audience. Elastix is a modular collection of robust C++ image registration algorithms that is widely used in the literature. However, its command-line interface introduces overhead during prototyping, experimental setup, and tuning of registration algorithms. By integrating Elastix with SimpleITK, Elastix can be used as a native library in Python, Java, R, Octave, Ruby, Lua, Tcl and C# on Linux, Mac and Windows. This allows Elastix to intregrate naturally with many development environments so the user can focus more on the registration problem and less on the underlying C++ implementation. As means of demonstration, we show how to register MR images of brains and natural pictures of faces using minimal amount of code. SimpleElastix is open source, licensed under the permissive Apache License Version 2.0 and available at https://github.com/kaspermarstal/SimpleElastix.

8.12   A Fast DRR Generation Scheme for 3D-2D Image Registration Based on the Block Projection Method

[Link to Paper]

In three-dimensional to two-dimensional (3D-2D) image registration, DRR (digitally reconstructed radiograph) generation is often a bottleneck in computation. In this article, a novel fast DRR generation scheme is proposed based on the recently introduced Block Projection method and Slab algorithm that reuse building blocks of DRRs previously generated for known poses. The scheme is flexible as exemplified in pose grid design and slab binding, and upper bounds in projection error exist and can be estimated. Experiments

were conducted to evaluate DRR quality and sensitivity to pose difference; computing time and error bounds were reported. The results showed that on a conventional computer the proposed scheme generated high quality, pose-preserving DRRs of size 512512 in 6 ms with slab binding, demonstrating its potential to be a viable solution to fast, high quality DRR generation for 3D-2D image registration.

8.13 Current- and Varifold-Based Registration of Lung Vessel and Airway Trees

[Link to Paper]

Registering lung CT images is an important problem for many applications including tracking lung motion over the breathing cycle, tracking anatomical and function changes over time, and detecting abnormal mechanical properties of the lung. This paper compares and contrasts current- and varifold-based diffeomorphic image registration approaches for registering tree-like structures of the lung. In these approaches, curve-like structures in the lung–for example, the skeletons of vessels and airways segmentation–are represented by currents or varifolds in the dual space of a Reproducing Kernel Hilbert Space (RKHS). Current and varifold representations are discretized and are parameterized via of a collection of momenta. A momenta corresponds to a line segment via the coordinates of the center of the line segment and the tangent direction of the line segment at the center. A varifold-based registration approach is similar to currents except that two varifold representations are aligned independent of the tangent vector orientation. An advantage of varifolds over currents is that the orientation of the tangent vectors can be difficult to determine especially when the vessel and airway trees are not connected. In this paper, we examine the image registration sensitivity and accuracy of current- and varifold-based registration as a function of the number and location of momentum used to represent tree like-structures in the lung. The registrations presented in this paper were generated using the Deformetrica software package ([Durrleman et al. 2014]). Keywords: Diffeomorphic image registration, currents, varifolds, momenta, Reproducing Kernel Hilbert Space (RKHS)

8.14 Reducing Drift in Mosaicing Slit-Lamp Retinal Images

[Link to Paper]

The construction of seamless and accurate mosaics from long slit-lamp retinal video sequences is an important and challenging task in navigated Pan-Retinal Photocoagulation. The main difficulty is accumulated registration drift due to the small number of features away from the optic nerve and the distortion induced by the geometry of the eye and the

contact lens. We present a new approach to reduce the drift. Our main idea is to create long-term high precision point correspondences by associating a simple global model with local correction and perform key-frame based Bundle Adjustment. We evaluate the method's performance compared to state-of-the-art. The results obtained with our method show significantly lower accumulated error.

8.15 Fast Deformable Image Registration with Non-Smooth Dual Optimization

[Link to Paper]

Optimization techniques have been widely used in deformable registration, allowing for the incorporation of similarity metrics with regularization mechanisms. These regularization mechanisms are designed to mitigate the effects of trivial solutions to ill-posed registration problems and to otherwise ensure the resulting deformation fields are well-behaved. This paper introduces a novel deformable registration (DR) algorithm, RANCOR, which uses iterative convexification to address DR problems under nonsmooth total-variation regularization. Initial comparative results against four state-of-the-art registration algorithms and under smooth regularization, respectively, are presented using the Internet Brain Segmentation Repository (IBSR) database.

8.16 Discrete Optimisation for Group-wise Cortical Surface Atlasing

[Link to Paper]

This paper presents a novel method for cortical surface atlasing. Group-wise registration is performed through a discrete optimisation framework that seeks to simultaneously improve pairwise correspondences between surface feature sets, whilst minimising a global cost relating to the rank of the feature matrix. It is assumed that when fully aligned, features will be highly linearly correlated, and thus have low rank. The framework is regularised through use of multi-resolution control point grids and higher-order smoothness terms, calculated by considering deformation strain for displacements of triplets of points. Accordingly the discrete framework is solved through high-order clique reduction. The framework is tested on cortical folding based alignment, using data from the Human Connectome Project. Preliminary results indicate that group-wise alignment improves folding correspondences, relative to registration between all pair-wise combinations, and registration to a global average template.

8.17 Tumor Growth Estimation via Registration of DCE-MRI Derived Tumor Specific Descriptors

[Link to Paper]

Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) provides information on changes occurring during tumor growth in the tumor micro-environment and vasculature. In the present paper, tumor voxel-wise estimates of tumor descriptors including total cell number, proliferative cell number, hypoxic cell number, necrotic cell number and oxygen level derived from DCE-MRI data are used to guide the deformable registration of subsequent time points over the tumor growth cycle, evaluating their predictive value for tumor growth. The analysis of three preclinical colon carcinoma longitudinal cases shows that using physiologically meaningful measures of tumor as guidance information can improve non-rigid registration of longitudinal tumor imaging data when compared to a stateof-the-art local correlation coefficient Demons approach. Moreover, using the determinant of the Jacobian of the estimated displacement field as an indicator of volume change allows us to observe a correlation between the tumor descriptor values and tumor growth, especially when maps of hypoxic cells and level of oxygen were used to aid registration. To the best of our knowledge, this work demonstrates for the first time the feasibility of using biologically meaningful tumor descriptors (total cell number, proliferative cell number, hypoxic cell number, necrotic cell number and oxygen level) derived from DCE-MRI to aid non-rigid registration of longitudinal tumor data as well as to estimate tumor growth.

8.18 Editorial for the Workshop on Biomedical Image Registration (WBIR) 2016

[Link to Paper]

8.19 Population Shape Collapse in Large Deformation Registration of MR Brain Images

[Link to Paper]

This paper examines the shape collapse problem that occurs when registering a pair of images or a population of images of the brain to a reference (target) image coordinate system using diffeomorphic image registration. Shape collapse occurs when a foreground or background structure in an image with non-zero volume is transformed into a set of zero or near zero volume as measured on a discrete voxel lattice in the target image coordinate system. Shape collapse may occur during image registration when the moving image has a

348

structure that is either missing or does not sufficiently overlap the corresponding structure in the target image[4]. Such a problem is common in image registration algorithms with large degrees of freedom such as many diffeomorphic image registration algorithms. Shape collapse is a concern when mapping functional data. For example, loss of signal may occur when mapping functional data such as fMRI, PET, SPECT using a transformation with a shape collapse if the functional signal occurs at the collapse region. This paper proposes an novel shape collapse measurement algorithm to detect the regions of shape collapse after image registration in pairwise registration. We further compute the shape collapse for a population of pairwise transformations such as occurs when registering many images to a common atlas coordinate system. Experiments are presented using the SyN diffeomorphic image registration algorithm. We demonstrate how changing the input parameters to the SyN registration algorithm can mitigate some of the collapse image registration artifacts. Keywords: shape collapse, diffeomorphic image registration

8.20 Graph Cuts-Based Registration Revisited: A Novel Approach for Lung Image Registration Using Supervoxels and Image-Guided Filtering

[Link to Paper]

This work revisits the concept of graph cuts as an efficient optimization technique in image registration. Previously, due to the computational burden involved, the use of graph cuts in this context has been mainly limited to 2D applications. Here we show how combining graph cuts with supervoxels, resulting in a sparse, yet meaningful graph-based image representation, can overcome previous limitations. Additionally, we show that a relaxed graph representation of the image allows for 'sliding' motion modeling and provides anatomically plausible estimation of the deformations. This is achieved by using image-guided filtering of the estimated sparse deformation field. We evaluate our method on a publicly available CT lung data set and show that our new approach compares very favourably with state-of-the-art in continuous and discrete image registration. Keywords: graph cuts, discrete optimization, image registration, supervoxels, image-guided filtering

8.21 A Combined EM and Visual Tracking Probabilistic Model for Robust Mosaicking: Application to Fetoscopy.

[Link to Paper]

Twin-to-Twin Transfusion Syndrome (TTTS) is a progressive pregnancy complication in which inter-twin vascular connections in the shared placenta result in a blood flow imbalance between the twins. The most effective therapy is to sever these connections by laser photo-coagulation. However, the limited field of view of the fetoscope hinders their

identification. A potential solution is to augment the surgeon's view by creating a mosaic image of the placenta. State-of-the-art mosaicking methods use feature-based approaches, which have three main limitations: (i) they are not robust against corrupt data e.g. blurred frames, (ii) temporal information is not used, (iii) the resulting mosaic suffers from drift. We introduce a probabilistic temporal model that incorporates electromagnetic and visual tracking data to achieve a robust mosaic with reduced drift. By assuming planarity of the imaged object, the nRT decomposition can be used to parametrize the state vector. Finally, we tackle the non-linear nature of the problem in a numerically stable manner by using the Square Root Unscented Kalman Filter. We show an improvement in performance in terms of robustness as well as a reduction of the drift in comparison to state-of-the-art methods in synthetic, phantom and ex vivo datasets.

8.22 Graph-Constrained Surface Registration Based on Tutte Embedding

[Link to Paper]

This work presents an efficient method to compute the registration between surfaces with consistent graph constraints based on Tutte graph embedding. Most natural objects have consistent anatomical structures, extracted as isomorphic feature graphs. For genus zero surfaces with 1 boundaries, the graphs are planar and usually 3-connected. By using Tutte embedding, each feature graph is embedded as a convex subdivision of a planar convex domain. Using the convex subdivision as constraint, surfaces are mapped onto convex subdivision domains and the registration is then computed over them. The computation is based on constrained harmonic maps to minimize the stretching energy, where curvy graph constraints become linear ones. This method is theoretically rigorous. The algorithm solves sparse linear systems and is computationally efficient and robust. The resulting mappings are proved to be unique and diffeomorphic. Experiments on various facial surface data demonstrate its efficiency and practicality.

8.23 Effects of Resolution and Registration Algorithm on the Accuracy of EPI vNavs for Real Time Head Motion Correction in MRI

[Link to Paper]

Low-resolution, EPI-based Volumetric Navigators (vNavs) have been used as a prospective motion-correction system in a variety of MRI neuroimaging pulse sequences. The use of low-resolution volumes represents a trade-off between motion tracking accuracy and acquisition time. However, this means that registration must be accurate on the order of 0.2 voxels or less to be effective for motion correction. While vNavs have shown promising results in clinical and research use, the choice of navigator and registration algorithm have

not previously been systematically evaluated. In this work we experimentally evaluate the accuracy of vNavs, and possible design choices for future improvements to the system, using real human data. We acquired navigator volumes at three isotropic resolutions (6.4 mm, 8 mm, and 10 mm) with known rotations and translations. The vNavs were then rigidly registered using trilinear, tricubic, and cubic B-spline interpolation. We demonstrate a novel refactoring of the cubic B-spline algorithm that stores pre-computed coefficients to reduce the per-interpolation time to be identical to tricubic interpolation. Our results show that increasing vNav resolution improves registration accuracy, and that cubic B-splines provide the highest registration accuracy at all vNav resolutions. Our results also suggest that the time required by vNavs may be reduced by imaging at 10 mm resolution, without substantial cost in registration accuracy.

8.24 Tissue-Volume Preserving Deformable Image Registration for 4DCT Pulmonary Images

[Link to Paper]

We propose a 4D (three spatial dimensions plus time) tissue-volume preserving non-rigid image registration algorithm for pulmonary 4D computed tomography (4DCT) data sets to provide relevant information for radiation therapy and estimate pulmonary ventilation. The sum of squared tissue volume difference (SSTVD) similarity cost takes into account the CT intensity changes of spatially corresponding voxels, which is caused by the variations of fraction of tissue within voxels throughout the respiratory cycle. The proposed 4D SSTVD registration scheme considers the entire dynamic 4D data set simultaneously, using both spatial and temporal information. We employed a uniform 4D cubic B-spline parametrization of the transform and a temporally extended linear elasticity regularization of deformation field to ensure temporal smoothness and thus biological plausibility of estimated deformation. We used a multiresolution multi-grid registration framework with limitedmemory Broyden Fletcher Goldfarb Shanno (L-BFGS) optimization procedure for rapid convergence and limited memory consumption. We conducted experiments using synthetic 2D+t images and clinical 4DCT pulmonary data sets and evaluated accuracy and temporal smoothness of the proposed method via manually annotated landmarks. Keywords: 4D image registration, groupwise image registration, sum of squared tissue volume difference, mass preservation, temporal smoothness, radiation therapy, treatment planning

8.25 Registering Retinal Vessel Images from Local to Global via Multiscale and Multicycle Features

[Link to Paper]

351

We propose a comprehensive method using multiscale and multicycle features for retinal vessel image registration with a local and global strategy. The multiscale vessel maps generated by multiwavelet kernels and multiscale hierarchical decomposition contain segmentation results at varying image resolutions in different levels of vessel details. Then the multicycle feature composed of various combinations of cycle structures with different numbers of vertices is extracted. The cycle structure consisting of vessel bifurcation points, crossover points of arteries and veins, and the connected vessels can be found by our Angle-based Depth-First Search (ADFS) algorithm. Local initial registration is implemented by the matched Cycle-Vessel feature points and global final registration is completed by the Cycle-VesselBifurcation feature points using similarity transformation. Finally, our Skeleton Alignment Error Measure (SAEM) is calculated for optimal scale and cycle feature selection, yielding the best registration result intelligently. Experimental results show that our method outperforms state-ofthe-art methods on retinal vessel image registration using different features in terms of accuracy and robustness.

9 Computational Cameras and Displays

9.1 Power-efficient Cameras Using Natural Image Statistics

[Link to Paper]

Motivated by recent results on compressed sensing cameras we consider cameras that perform an analog linear transformation on the signal, followed by scalar quantization. Specifically we ask: is it better to use compressed sensing (is an under-sampling random matrix) or direct sensing (is the sparsifying basis)? We compare the two approaches using their energy-distortion tradeoffs: assuming most of the energy consumed by such systems is in the ADC and the energy of the quantizer doubles with each bit, which system will give lower distortion for the same energy consumption? We present analytic expressions for the energy-distortion curves for three signal models: signals residing in a known subspace, sparse signals and power-law signals. For all of these models, our analysis shows that direct sensing results in lower distortion for a given energy consumption. We also present simulation results for natural images showing that direct sensing of Haar wavelet coefficients is preferable for these signals. Given the assumptions of our model, direct sensing of Haar wavelets can achieve high quality imaging (PSNR of 40 dB) with 6% the power consumption of standard cameras using 8 bits per channel.

9.2 Avoiding the Deconvolution: Framework Oriented Color Transfer for Enhancing Low-Light Images

[Link to Paper]

In this paper we introduce a novel color transfer method to address the underexposed image amplification problem. Targeted scenario implies a dual acquisition, containing a normally exposed, possibly blurred, image and an underexposed/low-light but sharp one. The problem of enhancing the low-light image is addressed as a color transfer problem. To properly solve the color transfer, the scene is split into perceptual frameworks and we propose a novel piece-wise approximation. The proposed method is shown to lead to robust results from both an objective and a subjective point of view.

9.3 Time-Offset Conversations on a Life-Sized Automultiscopic Projector Array

[Link to Paper]

We present a system for creating and displaying interactive life-sized 3D digital humans based on pre-recorded interviews. We use 30 cameras and an extensive list of questions to record a large set of video responses. Users access videos through a natural conversation

interface that mimics face-to-face interaction. Recordings of answers, listening and idle behaviors are linked together to create a persistent visual image of the person throughout the interaction. The interview subjects are rendered using flowed light fields and shown life-size on a special rear-projection screen with an array of 216 video projectors. The display allows multiple users to see different 3D perspectives of the subject in proper relation to their viewpoints, without the need for stereo glasses. The display is effective for interactive conversations since it provides 3D cues such as eye gaze and spatial hand gestures.

9.4 Depth Camera Based on Color-Coded Aperture

[Link to Paper]

In this paper we present a single-lens single-frame passive depth sensor based on conventional imaging system with minor hardware modifications. It is based on colorcoded aperture approach and has high light-efficiency which allows capturing images even with handheld devices with small cameras. The sensor measures depth in millimeters in the whole frame, in contrast to prior-art approaches. Contributions of this paper are: (1) introduction of novel light-efficient coded aperture designs and corresponding algorithm modification; (2) depth sensor calibration procedure and disparity to depth conversion method; (3) a number of color-coded aperture based depth sensor implementations including a DSLR based prototype, a smartphone based prototype and a compact camera based prototype; (4) applications including real-time 3D scene reconstruction and depth based image effects.

9.5 Strategies For Resolving Camera Metamers Using 3+1 Channel

[Link to Paper]

This paper discusses the problem of reducing camera metamerism by using a fourth spectral channel in addition to the typical red, green, and blue channels of the tristimulus sensor used in the commercial consumer cameras. Specifically, we consider three options for this fourth channel. The first option is to use a color channel from another camera to reduce metamerism. The second option is to use a color channel from an image captured by the same camera but with a color filter as the fourth channel. The third option is to design a specific spectral channel to be fabricated with the existing camera sensor. This option uses the metameric black space to design the channel. The commercial cameras' original metamerism is typically more than 20%, as observed in a dataset of 335 spectral images captured in 5 different indoor illuminations. Our results show that the third option is the best since it reduces metamerism down to about 5%. Among the first and second options, the first option is more effective and it reduces metamerism down to about 15%.

The channel designed using the third option can be used for advanced applications such as distinguishing objects with different spectral reflectances but similar colors.

9.6   SparkleGeometry: Glitter Imaging for 3D Point Tracking

[Link to Paper]

We consider a geometric inference problem for an imaging system consisting of a camera that views the world through a planar, rectangular sheet of glitter. We describe a procedure to calibrate this imaging geometry as a generalized camera which characterizes the subset of the light field viewed through each piece of glitter. We propose an easy to construct physical prototype and characterize its performance for estimating the 3D position of a moving point light source just by viewing the changing sparkle patterns visible on the glitter sheet.

10 Large Scale 3D Data: Acquisition, Modeling and Analysis

10.1   Comprehensive Automated 3D Urban Environment Modelling Using Terrestrial Laser Scanning Point Cloud

[Link to Paper]

In this paper we present a novel street scene modelling framework, which takes advantage of 3D point cloud captured by a high definition LiDAR laser scanner. We propose an automatic and robust approach to detect, segment and classify urban objects from point clouds hence reconstructing a comprehensive 3D urban environment model. Our system first automatically segments grounds point cloud. Then building facades will be detected by using binary range image processing. Remained point cloud will be grouped into voxels and subsequently transformed into super voxels. Local 3D features are extracted from super voxels and classified by trained boosted decision trees with semantic classes e.g. tree, pedestrian, and car. Given labeled point cloud the proposed algorithm reconstructs the realistic model in two phases. Firstly building facades will be rendered by ShadVis algorithm. In the second step we apply a novel and fast method for fitting the solid predefined template mesh models to non-building labeled point cloud. The proposed method is evaluated both quantitatively and qualitatively on a challenging TLS NAVTEQ True databases.

10.2   RGBD Datasets: Past, Present and Future

[Link to Paper]

Since the launch of the Microsoft Kinect, scores of RGBD datasets have been released. These have propelled advances in areas from reconstruction to gesture recognition. In this paper we explore the field, reviewing datasets across eight categories: semantics, object pose estimation, camera tracking, scene reconstruction, object tracking, human actions, faces and identification. By extracting relevant information in each category we help researchers to find appropriate data for their needs, and we consider which datasets have succeeded in driving computer vision forward and why. Finally, we examine the future of RGBD datasets. We identify key areas which are currently underexplored, and suggest that future directions may include synthetic data and dense reconstructions of static and dynamic scenes.

10.3   Real time complete dense depth reconstruction for a monocular camera

[Link to Paper]

In this paper, we aim to solve the problem of estimating complete dense depth maps from a monocular moving camera. By 'complete', we mean depth information is estimated for every pixel and detailed reconstruction is achieved. Although this problem has previously been attempted, the accuracy of complete dense depth reconstruction is a remaining problem. We propose a novel system which produces accurate complete dense depth map. The new system consists of two subsystems running in separated threads, namely, dense mapping and sparse patch-based tracking. For dense mapping, a new projection error computation method is proposed to enhance the gradient component in estimated depth maps. For tracking, a new sparse patch-based tracking method estimates camera pose by minimizing a normalized error term. The experiments demonstrate that the proposed method obtains improved performance in terms of completeness and accuracy compared to three state-ofthe-art dense reconstruction methods VSFM+CMVC, LSDSLAM and RE-MODE.

10.4 Fast and Accurate Registration of Structured Point Clouds with Small Overlaps

[Link to Paper]

To perform registration of structured point clouds with large rotation and small overlaps, this paper presents an algorithm based on the direction angles and the projection information of dense points. This algorithm fully employs the geometric information of structured environment. It consists of two parts: rotation estimation and translation estimation. For rotation estimation, a direction angle is defined for a point cloud and then the rotation matrix is obtained by comparing the difference between the distributions of angles. For translation estimation, the point clouds are projected onto three orthogonal planes and then a correlation operation is performed on the projection images to calculate the translation vector. Experiments have been conducted on several datasets. Experimental results demonstrate that the proposed algorithm outperforms the state-of-the-art approaches in terms of both accuracy and efficiency.

11 ChaLearn Looking at People and Faces of the World: Face Analysis Workshop and Challenge

11.1 Apparent Age Estimation from Face Images Combining General and Children-Specialized Deep Learning Models

[Link to Paper]

This work describes our solution in the second edition of the ChaLearn LAP competition on Apparent Age Estimation. Starting from a pretrained version of the VGG-16 convolutional neural network for face recognition, we train it on the huge IMDB-Wiki dataset for biological age estimation and then fine-tune it for apparent age estimation using the relatively small competition dataset. We show that the precise age estimation of children is the cornerstone of the competition. Therefore, we integrate a separate "children" VGG-16 network for apparent age estimation of children between 0 and 12 years old in our final solution. The "children" network is fine-tuned from the "general" one. We employ different age encoding strategies for training "general" and "children" networks: the soft one (label distribution encoding) for the "general" network and the strict one (0/1 classification encoding) for the "children" network. Finally, we highlight the importance of the state-of-the-art face detection and face alignment for the final apparent age estimation. Our resulting solution wins the 1st place in the competition significantly outperforming the runner-up.

11.2 Facial Attributes Classification using Multi-Task Representation Learning

[Link to Paper]

This paper presents a new approach for facial attribute classification using a multi-task learning approach. Unlike other approaches that uses hand engineered features, our model learns a shared feature representation that is wellsuited for multiple attribute classification. Learning a joint feature representation enables interaction between different tasks. For learning this shared feature representation we use a Restricted Boltzmann Machine (RBM) based model, enhanced with a factored multi-task component to become Multi-Task Restricted Boltzmann Machine (MT-RBM). Our approach operates directly on faces and facial landmark points to learn a joint feature representation over all the available attributes. We use an iterative learning approach consisting of a bottom-up/top-down pass to learn the shared representation of our multi-task model and at inference we use a bottom-up pass to predict the different tasks. Our approach is not restricted to any type of attributes, however, for this paper we focus only on facial attributes. We evaluate our approach on three publicly available datasets, the Celebrity Faces (CelebA), the Multi-task Facial Landmarks (MTFL), and the ChaLearn challenge dataset. We show superior classification performance improvement over the state-of-the-art.

11.3 ChaLearn Looking at People and Faces of the World: Face Analysis Workshop and Challenge 2016

[Link to Paper]

We present the 2016 ChaLearn Looking at People and Faces of the World Challenge and Workshop, which ran three competitions on the common theme of face analysis from still images. The first one, Looking at People, addressed age estimation, while the second and third competitions, Faces of the World, addressed accessory classification and smile and gender classification, respectively. We present two crowd-sourcing methodologies used to collect manual annotations. A custom-build application was used to collect and label data about the apparent age of people (as opposed to the real age). For the Faces of the World data, the citizen-science Zooniverse platform was used. This paper summarizes the three challenges and the data used, as well as the results achieved by the participants of the competitions. Details of the ChaLearn LAP FotW competitions can be found at http://gesture. chalearn.org.

11.4 Kernel ELM and CNN based Facial Age Estimation

[Link to Paper]

We propose a two-level system for apparent age estimation from facial images. Our system first classifies samples into overlapping age groups. Within each group, the apparent age is estimated with local regressors, whose outputs are then fused for the final estimate. We use a deformable parts model based face detector, and features from a pretrained deep convolutional network. Kernel extreme learning machines are used for classification. We evaluate our system on the ChaLearn Looking at People 2016 - Apparent Age Estimation challenge dataset, and report 0.3740 normal score on the sequestered test set.

11.5 Inferring Visual Persuasion via Body Language, Setting, and Deep Features

[Link to Paper]

The computer vision community has reached a point when it can start considering high-level reasoning tasks such as the "communicative intents" of images, or in what light an image portrays its subject. For example, an image might imply that a politician is competent, trustworthy, or energetic. We explore a variety of features for predicting these communicative intents. We study a number of facial expressions and body poses as cues for the implied nuances of the politician's personality. We also examine how the setting of

362

an image (e.g. kitchen or hospital) influences the audience's perception of the portrayed politician. Finally, we improve the performance of an existing approach on this problem, by learning intermediate cues using convolutional neural networks. We show state of the art results on the Visual Persuasion dataset of Joo et al. [11].

11.6   Deep Age Distribution Learning for Apparent Age Estimation

[Link to Paper]

Apparent age estimation has attracted more and more researchers since its potential applications in the real world. Apparent age estimation differs from chronological age estimation that in apparent age estimation each facial image is labelled by multiple individuals, the mean age is the ground truth age and the uncertainty is introduced by the standard deviation. In this paper, we propose a novel method called Deep Age Distribution Learning(DADL) to deal with such situation. According to the given mean age and the standard deviation, we generate a Gaussian age distribution for each facial image as the training target instead of the single age. DADL first detects the facial region in image and aligns the facial image. Then, it uses deep Convolutional Neural Network(CNN) pre-trained based on the VGGFace and fine-tuned on the age dataset to extract the predicted age distribution. Finally it uses ensemble method to get the result. Our DADL method got a good performance in ChaLearn Looking at People 2016-Track 1: Age Estimation and ranked the 2nd place among 105 registered participants.

11.7   Person-independent 3D Gaze Estimation using Face Frontalization

[Link to Paper]

Person-independent and pose-invariant estimation of eye-gaze is important for situation analysis and for automated video annotation. We propose a fast cascade regression based method that first estimates the location of a dense set of markers and their visibility, then reconstructs face shape by fitting a part-based 3D model. Next, the reconstructed 3D shape is used to estimate a canonical view of the eyes for 3D gaze estimation. The model operates in a feature space that naturally encodes local ordinal properties of pixel intensities leading to photometric invariant estimation of gaze. To evaluate the algorithm in comparison with alternative approaches, three publicly-available databases were used, Boston University Head Tracking, Multi-View Gaze and CAVE Gaze datasets. Precision for head pose and gaze averaged 4 degrees or less for pitch, yaw, and roll. The algorithm outperformed alternative methods in both datasets.

11.8 Dominant Codewords Selection with Topic Model for Action Recognition

[Link to Paper]

In this paper, we propose a framework for recognizing human activities that uses only intopic dominant codewords and a mixture of intertopic vectors. Latent Dirichlet allocation (LDA) is used to develop approximations of human motion primitives; these are mid-level representations, and they adaptively integrate dominant vectors when classifying human activities. In LDA topic modeling, action videos (documents) are represented by a bag-of-words (input from a dictionary), and these are based on improved dense trajectories ([18]). The output topics correspond to human motion primitives, such as finger moving or subtle leg motion. We eliminate the impurities, such as missed tracking or changing light conditions, in each motion primitive. The assembled vector of motion primitives is an improved representation of the action. We demonstrate our method on four different datasets.

11.9 DeepBE: Learning Deep Binary Encoding for Multi-Label Classification

[Link to Paper]

The track 2 and track 3 of ChaLearn 2016 can be considered as Multi-Label Classification problems. We present a framework of learning deep binary encoding (DeepBE) to deal with multi-label problems by transforming multi-labels to single labels. The transformation of DeepBE is in a hidden pattern, which can be well addressed by deep convolutions neural networks (CNNs). Furthermore, we adopt an ensemble strategy to enhance the learning robustness. This strategy is inspired by its effectiveness in fine-grained image recognition (FGIR) problem, while most of face related tasks such as track 2 and track 3 are also FGIR problems. By DeepBE, we got 5.45% and 10.84% mean square error for track 2 and track 3 respectively. Additionally, we proposed an algorithm adaption method to treat the multiple labels of track 2 directly and got 6.84% mean square error.

11.10 Apparent Age Estimation Using Ensemble of Deep Learning Models

[Link to Paper]

In this paper, we address the problem of apparent age estimation. Different from estimating the real age of individuals, in which each face image has a single age label, in this problem, face images have multiple age labels, corresponding to the ages perceived by the annotators, when they look at these images. This provides an intriguing computer vision

problem, since in generic image or object classification tasks, it is typical to have a single ground truth label per class. To account for multiple labels per image, instead of using average age of the annotated face image as the class label, we have grouped the face images that are within a specified age range. Using these age groups and their age-shifted groupings, we have trained an ensemble of deep learning models. Before feeding an input face image to a deep learning model, five facial landmark points are detected and used for 2-D alignment. We have employed and fine tuned convolutional neural networks (CNNs) that are based on VGG-16 [24] architecture and pretrained on the IMDB-WIKI dataset [22]. The outputs of these deep learning models are then combined to produce the final estimation. Proposed method achieves 0.3668 error in the final ChaLearn LAP 2016 challenge test set [5].

11.11 Structured Output SVM Prediction of Apparent Age, Gender and Smile From Deep Features

[Link to Paper]

We propose structured output SVM for predicting the apparent age as well as gender and smile from a single face image represented by deep features. We pose the problem of apparent age estimation as an instance of the multi-class structured output SVM classifier followed by a softmax expected value refinement. The gender and smile predictions are treated as binary classification problems. The proposed solution first detects the face in the image and then extracts deep features from the cropped image around the detected face. We use a convolutional neural network with VGG-16 architecture [25] for learning deep features. The network is pretrained on the ImageNet [24] database and then fine-tuned on IMDB-WIKI [21] and ChaLearn 2015 LAP datasets [8]. We validate our methods on the ChaLearn 2016 LAP dataset [9]. Our structured output SVMs are trained solely on ChaLearn 2016 LAP data. We achieve excellent results for both apparent age prediction and gender and smile classification.

11.12 ChaLearn Looking at People RGB-D Isolated and Continuous Datasets for Gesture Recognition

[Link to Paper]

In this paper, we present two large video multi-modal datasets for RGB and RGB-D gesture recognition: the ChaLearn LAP RGB-D Isolated Gesture Dataset (IsoGD) and the Continuous Gesture Dataset (ConGD). Both datasets are derived from the ChaLearn Gesture Dataset (CGD) that has a total of more than 50000 gestures for the "one-shot-learning" competition. To increase the potential of the old dataset, we designed new well curated

datasets composed of 249 gesture labels, and including 47933 gestures manually labeled the begin and end frames in sequences. Using these datasets we will open two competitions on the CodaLab platform so that researchers can test and compare their methods for "user independent" gesture recognition. The first challenge is designed for gesture spotting and recognition in continuous sequences of gestures while the second one is designed for gesture classification from segmented data. The baseline method based on the bag of visual words model is also presented.

11.13 Gender and Smile Classification using Deep Convolutional Neural Networks

[Link to Paper]

Facial gender and smile classification in unconstrained environment is challenging due to the invertible and large variations of face images. In this paper, we propose a deep model composed of GNet and SNet for these two tasks. We leverage the multi-task learning and the general-to-specific fine-tuning scheme to enhance the performance of our model. Our strategies exploit the inherent correlation between face identity, smile, gender and other face attributes to relieve the problem of over-fitting on small training set and improve the classification performance. We also propose the tasks-aware face cropping scheme to extract attribute-specific regions. The experimental results on the ChaLearn 16 FotW dataset for gender and smile classification demonstrate the effectiveness of our proposed methods. Input Image Face Detection Stage 1 Facial Landmarks Localization Face Cropping Stage 2 00000000

11.14 Identifying Same Persons from Temporally Synchronized Videos Taken by Multiple Wearable Cameras

[Link to Paper]

Video-based human action recognition benefits from multiple cameras which can provide temporally synchronized, multi-view videos. Cross-video person identification, i.e., determining whether at a given time, persons tracked in different videos are the same person or not, is a key step to integrate such multi-view information for collaborative action recognition. For fixed cameras, this step is relatively easy since different cameras can be precalibrated. In this paper, we study cross-video person identification for wearable cameras, which are constantly moving with the wearers. Specifically, we take the tracked persons from different videos to be the same person if their 3D poses are the same, given that these videos are synchronized. We adapt an existing algorithm to estimate the tracked person's 3D poses in each 2D video using motionbased features. Experiments show that, although

3D pose estimation is not perfect, the proposed method can still lead to better cross-video person identification than using appearance information.

12 Visual Analysis of Satellite to Street Imagery

12.1 Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks

[Link to Paper]

We propose a deep Convolutional Neural Network (CNN) for land cover mapping in remote sensing images, with a focus on urban areas. In remote sensing, class imbalance represents often a problem for tasks like land cover mapping, as small objects get less prioritised in an effort to achieve the best overall accuracy. We propose a novel approach to achieve high overall accuracy, while still achieving good accuracy for small objects. Quantifying the uncertainty on a pixel scale is another challenge in remote sensing, especially when using CNNs. In this paper we use recent advances in measuring uncertainty for CNNs and evaluate their quality both qualitatively and quantitatively in a remote sensing context. We demonstrate our ideas on different deep architectures including patch-based and so-called pixel-to-pixel approaches, as well as their combination, by classifying each pixel in a set of aerial images covering Vaihingen, Germany. The results show that we obtain an overall classification accuracy of 87%. The corresponding F1score for the small object class "car" is 80.6%, which is higher than state-of-the art for this dataset.

12.2 Automatic Alignment of Indoor and Outdoor Building Models using 3D Line Segments

[Link to Paper]

This paper presents an approach for automatically aligning the non-overlapping interior and exterior parts of a 3D building model computed from image based 3D reconstructions. We propose a method to align the 3D reconstructions by identifying corresponding 3D structures that are part of the interior and exterior model (e.g. openings like windows). In this context, we point out the potential of using 3D line segments to enrich the information of point clouds generated by SfMs and show how this can be used for interpreting the scene and matching individual reconstructions. (a) building exterior

12.3 The TUM-DLR Multimodal Earth Observation Evaluation Benchmark

[Link to Paper]

We present a new dataset for development, benchmarking, and evaluation of remote sensing and earth observation approaches with special focus on converging perspectives. In

order to provide data with different modalities, we observed the same scene using satellites, airplanes, unmanned aerial vehicles (UAV), and smartphones. The dataset is further complemented by ground-truth information and baseline results for different application scenarios. The provided data can be freely used by anybody interested in remote sensing and earth observation and will be continuously augmented and updated.

13   Performance Evaluation of Tracking and Surveillance

13.1 Realtime Anomaly Detection using Trajectory-level Crowd Behavior Learning

[Link to Paper]

We present an algorithm for realtime anomaly detection in low to medium density crowd videos using trajectorylevel behavior learning. Our formulation combines online tracking algorithms from computer vision, non-linear pedestrian motion models from crowd simulation, and Bayesian learning techniques to automatically compute the trajectory-level pedestrian behaviors for each agent in the video. These learned behaviors are used to segment the trajectories and motions of different pedestrians or agents and detect anomalies. We demonstrate the interactive performance on the PETS 2016 ARENA dataset as well as indoor and outdoor crowd video benchmarks consisting of tens of human agents. Website: http://gamma.cs.unc.edu/RCrowdT/ Anomaly/

13.2 Channel Coded Distribution Field Tracking for Thermal Infrared Imagery

[Link to Paper]

We address short-term, single-object tracking, a topic that is currently seeing fast progress for visual video, for the case of thermal infrared (TIR) imagery. The fast progress has been possible thanks to the development of new template-based tracking methods with online template updates, methods which have not been explored for TIR tracking. Instead, tracking methods used for TIR are often subject to a number of constraints, e.g., warm objects, low spatial resolution, and static camera. As TIR cameras become less noisy and get higher resolution these constraints are less relevant, and for emerging civilian applications, e.g., surveillance and automotive safety, new tracking methods are needed. Due to the special characteristics of TIR imagery, we argue that template-based trackers based on distribution fields should have an advantage over trackers based on spatial structure features. In this paper, we propose a templatebased tracking method (ABCD) designed specifically for TIR and not being restricted by any of the constraints above. In order to avoid background contamination of the object template, we propose to exploit background information for the online template update and to adaptively select the object region used for tracking. Moreover, we propose a novel method for estimating object scale change. The proposed tracker is evaluated on the VOT-TIR2015 and VOT2015 datasets using the VOT evaluation toolkit and a comparison of relative ranking of all common participating trackers in the challenges is provided. Further, the proposed tracker, ABCD, and the VOT-TIR2015 winner SRDCFir are evaluated on maritime data. Experimental results show that the ABCD tracker performs particularly well on thermal infrared sequences.

13.3 Saliency-based Detection for Maritime Object Tracking

[Link to Paper]

This paper presents a new method for object detection and tracking based on visual saliency as a way of mitigating against challenges present in maritime environments. Object detection is based on adaptive hysteresis thresholding of a saliency map generated with a modified version of the Boolean Map Saliency (BMS) approach. We show that the modification reduces false positives by suppressing detection of wakes and surface glint. Tracking is performed by matching detections frame to frame and smoothing trajectories with a Kalman filter. The proposed approach is evaluated on the PETS 2016 challenge dataset on detecting and tracking boats around a vessel at sea.

13.4 Abnormal Event Recognition: A Hybrid Approach Using Semantic Web Technologies

[Link to Paper]

Video surveillance systems generated about 65% of the Universe Big Data in 2015. The development of systems for intelligent analysis of such a large amount of data is among the most investigated topics in the academia and commercial world. Recent outcomes in knowledge management and computational intelligence demonstrate the effectiveness of semantic technologies in several fields like image and text analysis, hand writing and speech recognition. In this paper a solution that, starting from the output of a people tracking algorithm, is able to recognize simple events (person falling to the ground) and complex ones (person aggression) is presented. The proposed solution uses semantic web technologies for automatically annotating the output produced by the tracking algorithm; a sets of rules for reasoning on these annotated data are also proposed. Such rules allow to define complex analytics functions demonstrating the effectiveness of hybrid approaches for event recognition.

13.5 Online Multi-Object Tracking based on Hierarchical Association Framework

[Link to Paper]

Online multi-object tracking is one of the crucial tasks in time-critical computer vision applications. In this paper, the problem of online multi-object tracking in complex scenes from a single, static, un-calibrated camera is addressed. In complex scenes, it is still challenging due to frequent and prolonged occlusions, abrupt motion change of objects, unre-

liable detections, and so on. To handle these difficulties, this paper proposes a four-stage hierarchical association framework based on online tracking-bydetection strategy. For this framework, tracks and detections are divided into several groups depending on several cues obtained from association results with the proposed track confidence. In each association stage, different sets of tracks and detections are associated to handle the following problems simultaneously: track generation, progressive trajectory construction, track drift and fragmentation. The experimental results show the robustness and effectiveness of the proposed method compared with other state-of-the-art methods.

13.6 PETS 2016: Dataset and Challenge

[Link to Paper]

This paper describes the datasets and computer vision challenges that form part of the PETS 2016 workshop. PETS 2016 addresses the application of on-board multi sensor surveillance for protection of mobile critical assets. The sensors (visible and thermal cameras) are mounted on the asset itself and surveillance is performed around the asset. Two datasets are provided: (1) a multi sensor dataset as used for the PETS2014 challenge which addresses protection of trucks (the ARENA Dataset); and (2) a new dataset - the IPATCH Dataset - addressing the application of multi sensor surveillance to protect a vessel at sea from piracy. The dataset specifically addresses several vision challenges set in the PETS 2016 workshop, and corresponding to different steps in a video understanding system: Low-Level Video Analysis (object detection and tracking), Mid-Level Video Analysis ('simple' event detection: the behaviour recognition of a single actor) and High-Level Video Analysis ('complex' event detection: the behaviour and interaction recognition of several actors).

13.7 Semantic Modelling for Behaviour Characterisation and Threat Detection

[Link to Paper]

Threat detection in computer vision can be achieved by extraction of behavioural cues. To achieve recognition of such cues, we propose to work with Semantic Models of behaviours. Semantic Models correspond to the translation of Low-Level information (tracking information) into HighLevel semantic description. The model is then similar to a naturally spoken description of the event. We have built semantic models for the behaviours and threats addressed in the PETS 2016 IPATCH dataset. Semantic models can trigger a threat alarm by themselves or give situation awareness. We describe in this paper how seman-

tic models are built from Low-Level trajectory features and how they are recognised. The current results are promising.

13.8 Robust Visual Tracking with Deep Convolutional Neural Network based Object Proposals on PETS

[Link to Paper]

Tracking by detection based object tracking methods encounter numerous complications including object appearance changes, size and shape deformations, partial and full occlusions, which make online adaptation of classifiers and object models a substantial challenge. In this paper, we employ an object proposal network that generates a small yet refined set of bounding box candidates to mitigate the this object model refitting problem by concentrating on hard negatives when we update the classifier. This helps improving the discriminative power as hard negatives are likely to be due to background and other distractions. Another intuition is that, in each frame, applying the classifier only on the refined set of object-like candidates would be sufficient to eliminate most of the false positives. Incorporating an object proposal makes the tracker robust against shape deformations since they are handled naturally by the proposal stage. We demonstrate evaluations on the PETS 2016 dataset and compare with the state-of-theart trackers. Our method provides the superior results.

14 Observing and Understanding Hands in Action

14.1 Learning Marginalization through Regression for Hand Orientation Inference

[Link to Paper]

We present a novel marginalization method for multilayered Random Forest based hand orientation regression. The proposed model is composed of two layers, where the first layer consists of a marginalization weights regressor while the second layer contains expert regressors trained on subsets of our hand orientation dataset. We use a latent variable space to divide our dataset into subsets. Each expert regressor gives a posterior probability for assigning a given latent variable to the input data. Our main contribution comes from the regression based marginalization of these posterior probabilities. We use a Kullback-Leibler divergence based optimization for estimating the weights that are used to train our marginalization weights regressor. In comparison to the state-of-the-art of both hand orientation inference and multi-layered Random Forest marginalization, our proposed method proves to be more robust.

14.2 Skeleton-based Dynamic hand gesture recognition

[Link to Paper]

In this paper, a new skeleton-based approach is proposed for 3D hand gesture recognition. Specifically, we exploit the geometric shape of the hand to extract an effective descriptor from hand skeleton connected joints returned by the Intel RealSense depth camera. Each descriptor is then encoded by a Fisher Vector representation obtained using a Gaussian Mixture Model. A multi-level representation of Fisher Vectors and other skeleton-based geometric features is guaranteed by a temporal pyramid to obtain the final feature vector, used later to achieve the classification by a linear SVM classifier. The proposed approach is evaluated on a challenging hand gesture dataset containing 14 gestures, performed by 20 participants performing the same gesture with two different numbers of fingers. Experimental results show that our skeleton-based approach consistently achieves superior performance over a depth-based approach.

14.3 Effectiveness of Grasp Attributes and Motion-Constraints for Fine-Grained Recognition of Object Manipulation Actions

[Link to Paper]

In this work, we consider the problem of recognition of object manipulation actions. This is a challenging task for real everyday actions, as the same object can be grasped and

moved in different ways depending on its functions and geometric constraints of the task. We propose to leverage grasp and motion-constraints information, using a suitable representation, to recognize and understand action intention with different objects. We also provide an extensive experimental evaluation on the recent Yale Human Grasping dataset consisting of large set of 455 manipulation actions. The evaluation involves a) Different contemporary multi-class classifiers, and binary classifiers with one-vsone multi-class voting scheme, and b) Differential comparisons results based on subsets of attributes involving information of grasp and motion-constraints. Our results clearly demonstrate the usefulness of grasp characteristics and motion-constraints, to understand actions intended with an object.

14.4   Hidden Hands: Tracking Hands with an Occlusion Aware Tracker

[Link to Paper]

This work presents an occlusion aware hand tracker to reliably track both hands of a person using a monocular RGB camera. To demonstrate its robustness, we evaluate the tracker on a challenging, occlusion-ridden naturalistic driving dataset, where hand motions of a driver are to be captured reliably. The proposed framework additionally encodes and learns tracklets corresponding to complex (yet frequently occurring) hand interactions offline, and makes an informed choice during data association. This provides positional information of the left and right hands with no intrusion (through complete or partial occlusions) over long, unconstrained video sequences in an online manner. The tracks thus obtained may find use in domains such as human activity analysis, gesture recognition, and higher-level semantic categorization.

378

15   Differential Geometry in Computer Vision and Machine Learning

15.1 A Riemannian Framework for Statistical Analysis of Topological Persistence Diagrams

[Link to Paper]

Topological data analysis is becoming a popular way to study high dimensional feature spaces without any contextual clues or assumptions. This paper concerns itself with one popular topological feature, which is the number of d-dimensional holes in the dataset, also known as the Betti-d number. The persistence of the Betti numbers over various scales is encoded into a persistence diagram (PD), which indicates the birth and death times of these holes as scale varies. A common way to compare PDs is by a pointto-point matching, which is given by the n-Wasserstein metric. However, a big drawback of this approach is the need to solve correspondence between points before computing the distance; for n points, the complexity grows according to $O(n3)$. Instead, we propose to use an entirely new framework built on Riemannian geometry, that models PDs as 2D probability density functions that are represented in the square-root framework on a Hilbert Sphere. The resulting space is much more intuitive with closed form expressions for common operations. The distance metric is 1) correspondence-free and also 2) independent of the number of points in the dataset. The complexity of computing distance between PDs now grows according to $O(K2)$, for a K K discretization of $[0, 1]2$. This also enables the use of existing machinery in differential geometry towards statistical analysis of PDs such as computing the mean, geodesics, classification etc. We report competitive results with the Wasserstein metric, at a much lower computational load, indicating the favorable properties of the proposed approach.

15.2 The Assignment Manifold: A Smooth Model for Image Labeling

[Link to Paper]

We introduce a novel geometric approach to the image labeling problem. A general objective function is defined on a manifold of stochastic matrices, whose elements assign prior data that are given in any metric space, to observed image measurements. The corresponding Riemannian gradient flow entails a set of replicator equations, one for each data point, that are spatially coupled by geometric averaging on the manifold. Starting from uniform assignments at the barycenter as natural initialization, the flow terminates at some global maximum, each of which corresponds to an image labeling that uniquely assigns the prior data. No tuning parameters are involved, except for two parameters setting the spatial scale of geometric averaging and scaling globally the numerical range of features, respectively. Our geometric variational approach can be implemented with sparse interior-point numerics in terms of parallel multiplicative updates that converge efficiently.

15.3 Fast Dynamic Programming for Elastic Registration of Curves

[Link to Paper]

Curve registration problems in data analysis and computer vision can often be reduced to the problem of matching two functions defined on an interval. Dynamic Programming (DP) is an effective approach to solve this problem. In this paper, we propose a DP algorithm that runs in O(N) time to compute optimal diffeomorphisms for elastic registration of curves with N nodes. This algorithm contrasts favorably with other DP algorithms used for this problem: the commonly used algorithm of quadratic time complexity, and the algorithm that guarantees a globally optimal solution with O(N 4) time complexity. Key to our computational efficiency is the savings achieved by reducing our search space, focusing on thin strips around graphs of estimates of optimal diffeomorphism. Estimates and strips are obtained with a multigrid approach: an optimal diffeomorphism obtained from a lower resolution grid using DP is progressively projected to ones of higher resolution until full resolution is attained. Additionally, our DP algorithm is designed so that it can handle nonuniformly discretized curves. This enables us to realize further savings in computations, since in the case of complicated curves requiring large numbers of nodes for a high-fidelity representation, we can distribute curve nodes adaptively, focusing nodes in parts of high variation. We demonstrate effectiveness of our DP algorithm on several registration problems in elastic shape analysis, and functional data analysis.

15.4 Testing Stationarity of Brain Functional Connectivity Using Change-Point Detection in fMRI Data

[Link to Paper]

This paper studies two questions: (1) Does the functional connectivity (FC) in a human brain remain stationary during performance of a task? (2) If it is non-stationary, how can one evaluate and estimate dynamic FC? The framework presented here relies on pre-segmented brain regions to represent instantaneous FC as symmetric, positive-definite matrices (SPDMs), with entries denoting covariances of fMRI signals across regions. The time series of such SPDMs is tested for change point detection using two important ideas: (1) a convenient Riemannian structure on the space of SPDMs for calculating geodesic distances and sample statistics, and (2) a graph-based approach, for testing similarity of distributions, that uses pairwise distances and a minimal spanning tree. This hypothesis test results in a temporal segmentation of observation interval into parts with stationary connectivity and an estimation of graph displaying FC during each such interval. We demonstrate these ideas using fMRI data from HCP database.

15.5   Consensus-Based Image Segmentation via Topological Persistence

[Link to Paper]

Image segmentation is one of the most important lowlevel operation in image processing and computer vision. It is unlikely for a single algorithm with a fixed set of parameters to segment various images successfully due to variations between images. However, it can be observed that the desired segmentation boundaries are often detected more consistently than other boundaries in the output of stateof-the-art segmentation results. In this paper, we propose a new approach to capture the consensus of information from a set of segmentations generated by varying parameters of different algorithms. The probability of a segmentation curve being present is estimated based on our probabilistic image segmentation model. A connectivity probability map is constructed and persistent segments are extracted by applying topological persistence to the probability map. Finally, a robust segmentation is obtained with the detection of certain segmentation curves guaranteed. The experiments demonstrate our algorithm is able to consistently capture the curves present within the segmentation set.

15.6   Partial Matchings and Growth Mapped Evolutions in Shape Spaces

[Link to Paper]

The definition of shape spaces as homogeneous spaces under the action of a group of diffeomorphisms equipped with a right invariant metric has been successful in providing theoretically sound and numerically efficient tools for registering and comparing shapes in the context of computational anatomy and leading to the so called diffeomorphometry. However, when considering not only shapes but shape evolutions or growth modeling, what could be the equivalent shape evolution spaces if any and what can be the natural group actions ? This paper proposes a principled framework in this direction on stratified shapes.

15.7   Robust Domain Adaptation on the L1-Grassmannian Manifold

[Link to Paper]

Domain adaptation aims to remedy the loss in classification performance that often occurs due to domain shifts between training and testing datasets. This problem is known as the dataset bias attributed to variations across datasets. Domain adaptation methods on Grassmann manifolds are among the most popular, including Geodesic Subspace Sampling and Geodesic Flow Kernel. Grassmann learning facilitates compact characterization

by generating linear subspaces and representing them as points on the manifold. However, Grassmannian construction is based on PCA which is sensitive to outliers. This motivates us to find linear projections that are robust to noise, outliers, and dataset idiosyncrasies. Hence, we combine L1-PCA and Grassmann manifolds to perform robust domain adaptation. We present empirical results to validate improvements and robustness for domain adaptation in object class recognition across datasets.

15.8 Human object interaction recognition using rate-invariant shape analysis of inter joint distances trajectories

[Link to Paper]

Human action recognition has emerged as one of the most challenging and active areas of research in the computer vision domain. In addition to pose variation and scale variability, high complexity of human motions and the variability of object interactions represent additional significant challenges. In this paper, we present an approach for human-object interaction modeling and classification. Towards that goal, we adopt relevant frame-level features; the inter-joint distances and joints-object distances. These proposed features are efficiently insensitive to position and pose variation. The evolution of the these distances in time is modeled by trajectories in a high dimension space and a shape analysis framework is used to model and compare the trajectories corresponding to human-object interaction in a Riemannian manifold. The experiments conducted following state-of-the-art settings and results demonstrate the strength of the proposed method. Using only the skeletal information, we achieve state-of-the-art classification results on the benchmark dataset.

15.9 A Statistical Framework for Elastic Shape Analysis of Spatio-Temporal Evolutions of Planar Closed Curves

[Link to Paper]

We propose a new statistical framework for spatiotemporal modeling of elastic planar, closed curves. This approach combines two recent frameworks for elastic functional data analysis and elastic shape analysis. The proposed trajectory registration framework enables matching and averaging to quantify spatio-temporal deformations while taking into account their dynamic specificities. A key ingredient of this framework is a tracking method that optimizes the evolution of curves extracted from sequences of consecutive images to estimate the spatio-temporal deformation fields. Automatic estimation of such deformations including spatial changes (strain) and dynamic temporal changes (phase) was tested on simulated examples and real myocardial trajectories. Experimental results show signifi-

cant improvements in the spatio-temporal structure of trajectory comparisons and averages using the proposed framework.

15.10 On Time-series Topological Data Analysis: New Data and Opportunities

[Link to Paper]

This work introduces a new dataset and framework for the exploration of topological data analysis (TDA) techniques applied to time-series data. We examine the end-toend TDA processing pipeline for persistent homology applied to time-delay embeddings of time series embeddings that capture the underlying system dynamics from which time series data is acquired. In particular, we consider stability with respect to time series length, the approximation accuracy of sparse filtration methods, and the discriminating ability of persistence diagrams as a feature for learning. We explore these properties across a wide range of time-series datasets spanning multiple domains for single source multi-segment signals as well as multi-source single segment signals. Our analysis and dataset captures the entire TDA processing pipeline and includes time-delay embeddings, persistence diagrams, topological distance measures, as well as kernels for similarity learning and classification tasks for a broad set of time-series data sources. We outline the TDA framework and rationale behind the dataset and provide insights into the role of TDA for time-series analysis as well as opportunities for new work.

15.11 Riemannian Geometric Approaches for Measuring Movement Quality

[Link to Paper]

A growing set of applications in home-based interactive physical therapy require the ability to monitor, inform and assess the quality of everyday movements. Interactive therapy requires both real-time feedback of movement quality, as well as summative feedback of quality over a period of time. Obtaining labeled data from trained experts is the main limitation, since it is both expensive and time consuming. Motivated by recent studies in motor-control, we propose an unsupervised approach that measures movement quality of simple actions by considering the deviation of a trajectory from an ideal movement path in the configuration space. We use two different configuration spaces to demonstrate this idea the product space $S1 \times S1$ to model the interaction of two joint angles, and $SE(3) \times SE(3)$ to model the movement of two joints, for two different applications in movement quality estimation. We also describe potential applications of these ideas to assess quality in real-time.

15.12 Bayesian Model-Based Automatic Landmark Detection for Planar Curves

[Link to Paper]

Identifying landmarks, points of interest on a shape, is crucial for many statistical shape analysis applications. Landmark-based methods dominate early literature; more recently, a method combining continuous shape outlines with landmark constraints was proposed. Unfortunately, methods requiring landmark specification depend on the number selected and their locations; such annotations are tedious for large datasets and subject to human interpretation. This work provides a Bayesian model-based method for automatic landmark selection, based on good approximations of landmark set interpolations. We outline an appropriate prior and likelihood, allowing for efficient posterior inference on landmark locations. The model allows for location uncertainty quantification, an important inferential procedure for further analysis. A method for selecting an appropriate number of landmarks is also discussed. Applications include a simulated example, shapes from the MPEG-7 dataset, and mice vertebrae.

15.13 A Survey on Rotation Optimization in Structure from Motion

[Link to Paper]

We consider the problem of robust rotation optimization in Structure from Motion applications. A number of different approaches have been recently proposed, with solutions that are at times incompatible, and at times complementary. The goal of this paper is to survey and compare these ideas in a unified manner, and to benchmark their robustness against the presence of outliers. In all, we have tested more than forty variants of a these methods (including novel ones), and we find the best performing combination.

15.14 Differential Geometry boosts Convolutional Neural Networks for Object Detection

[Link to Paper]

Convolutional neural networks (CNNs) have had dramatic success in appearance based object recognition tasks such as the ImageNet visual recognition challenge [8]. However, their application to object recognition and detection thus far has focused largely on intensity or color images as inputs. Motivated by demonstrations that depth can enhance the performance of CNN-based approaches [2][5], in this paper we consider the benefits of also including differential geometric shape features. This elementary idea of using zeroth order (depth), first-order (surface normal) and second-order (surface curvature)

features in a principled manner boosts the performance of a CNN that has been pretrained on a color image database. Notably, in an object detection task involving 19 categories we achieve 39.30% accuracy on the NYUv2 dataset, which is a 10.4% improvement over the current state-of-the-art accuracy of 35.6% using the method in [5]. In the simpler scenario of turntable style object recognition, our experiments on the University of Washington (UW) RGB-D dataset yield an accuracy of 88.7% correct recognition over 51 object categories, where the best competing result is 87.5% [2]. Taken together, our results provide strong evidence that the abstraction of surface shape benefits object detection and recognition.

16 Robust Features for Computer Vision

16.1 Improving Gradient Histogram Based Descriptors for Pedestrian Detection in Datasets with Large Variations.

[Link to Paper]

Gradient histogram based descriptors, that are constructed using the gradient magnitudes as votes to orientation bins, are successfully used for Pedestrian Detection. However, their performance is hampered when presented with datasets having many variations in properties such as appearance, texture, scale, background, and object pose. Such variations can be reduced by smoothing the images. But, the performance of the descriptors, and their classifiers is affected negatively by this, due to the loss of important gradients along with the noisy ones. In this work, we show that the ranks of gradient magnitudes stay resilient to such a smoothing. We show that a combination of image smoothing and the ranks of gradient magnitudes yields good detection performances, especially when the variations in a dataset are large or the number of training samples is less. Experiments on the challenging Caltech and Daimler Pedestrian datasets, and the Inria Person dataset illustrate these findings.

16.2 Do We Need Binary Features for 3D Reconstruction?

[Link to Paper]

Binary features have been incrementally popular in the past few years due to their low memory footprints and the efficient computation of Hamming distance between binary descriptors. They have been shown with promising results on some real time applications, e.g., SLAM, where the matching operations are relative few. However, in computer vision, there are many applications such as 3D reconstruction requiring lots of matching operations between local features. Therefore, a natural question is that is the binary feature still a promising solution to this kind of applications? To get the answer, this paper conducts a comparative study of binary features and their matching methods on the context of 3D reconstruction in a recently proposed large scale mutliview stereo dataset. Our evaluations reveal that not all binary features are capable of this task. Most of them are inferior to the classical SIFT based method in terms of reconstruction accuracy and completeness with a not significant better computational performance.

16.3 Background Subtraction using Local SVD Binary Pattern

[Link to Paper]

Background subtraction is a basic problem for change detection in videos and also the first step of high-level computer vision applications. Most background subtraction methods rely on color and texture feature. However, due to illuminations changes in different scenes and affections of noise pixels, those methods often resulted in high false positives in a complex environment. To solve this problem, we propose an adaptive background subtraction model which uses a novel Local SVD Binary Pattern (named LSBP) feature instead of simply depending on color intensity. This feature can describe the potential structure of the local regions in a given image, thus, it can enhance the robustness to illumination variation, noise, and shadows. We use a sample consensus model which is well suited for our LSBP feature. Experimental results on CDnet 2012 dataset demonstrate that our background subtraction method using LSBP feature is more effective than many state-of-the-art methods.

16.4 Deeply Exploit Depth Information for Object Detection

[Link to Paper]

This paper addresses the issue on how to more effectively coordinate the depth with RGB aiming at boosting the performance of RGB-D object detection. Particularly, we investigate two primary ideas under the CNN model: property derivation and property fusion. Firstly, we propose that the depth can be utilized not only as a type of extra information besides RGB but also to derive more visual properties for comprehensively describing the objects of interest. So a two-stage learning framework consisting of property derivation and fusion is constructed. Here the properties can be derived either from the provided color/depth or their pairs (e.g. the geometry contour adopted in this paper). Secondly, we explore the fusion method of different properties in feature learning, which is boiled down to, under the CNN model, from which layer the properties should be fused together. The analysis shows that different semantic properties should be learned separately and combined before passing into the final classifier. Actually, such a detection way is in accordance with the mechanism of the primary neural cortex (V1) in brain. We experimentally evaluate the proposed method on the challenging dataset, and have achieved state-of-the-art performance.

16.5 Texture Complexity based Redundant Regions Ranking for Object Proposal

[Link to Paper]

Object proposal has been successfully applied in recent visual object detection approaches and shown improved computational efficiency. The purpose of object proposal is to use as

few as regions to cover as many as objects. In this paper, we propose a strategy named Texture Complexity based Redundant Regions Ranking (TCR) for object proposal. Our approach first produces rich but redundant regions using a color segmentation approach, i.e. Selective Search. It then uses Texture Complexity (TC) based on complete contour number and Local Binary Pattern (LBP) entropy to measure the objectness score of each region. By ranking based on the TC, it is expected that as many as true object regions are preserved, while the number of the regions is significantly reduced. Experimental results on the PASCAL VOC 2007 dataset show that the proposed TCR significantly improves the baseline approach by increasing AUC (area under recall curve) from 0.39 to 0.48. It also outperforms the state-of-the-art with AUC and uses fewer detection proposals to achieve comparable recall rates.

16.6 The Best of Both Worlds: Combining Data-independent and Data-driven Approaches for Action Recognition

[Link to Paper]

Motivated by the success of CNNs in object recognition on images, researchers are striving to develop CNN equivalents for learning video features. However, learning video features globally has proven to be quite a challenge due to the difficulty of getting enough labels, processing large-scale video data, and representing motion information. Therefore, we propose to leverage effective techniques from both data-driven and data-independent approaches to improve action recognition system. Our contribution is three-fold. First, we explicitly show that local handcrafted features and CNNs share the same convolution-pooling network structure. Second, we propose to use independent subspace analysis (ISA) to learn descriptors for state-of-the-art handcrafted features. Third, we enhance ISA with two new improvements, which make our learned descriptors significantly outperform the handcrafted ones. Experimental results on standard action recognition benchmarks show competitive performance.

16.7 Euclidean and Hamming Embedding for Image Patch Description with Convolutional Networks

[Link to Paper]

Local feature descriptors represent image patches as floating-point or binary arrays for computer vision tasks. In this paper, we propose to train Euclidean and Hamming embedding for image patch description with triplet convolutional networks. Thanks to the learning ability of deep ConvNets, the trained local feature generation method, which is called

Deeply Learned Feature Transform (DELFT), has good distinctiveness and robustness. Evaluated on the UBC benchmark, we get the state-of-the-art results using floating-point and binary features. Also, the learned features can cooperate with existing nearest neighbor search algorithms in Euclidean and Hamming space. In addition, a new benchmark is constructed to facilitate future related research, which contains 40 million image patches, corresponding to 6.7 million 3D points, being 25 times larger than existing dataset. The distinctiveness and robustness of the proposed method are demonstrated in the experimental results.

16.8 Unsupervised Robust Feature-based Partition Ensembling to Discover Categories

[Link to Paper]

The design of novel robust image descriptors is still a formidable problem. Different features, with different capabilities, are introduced every year. However, to explore how to combine them is also a fundamental task. This paper proposes two novel strategies for aggregating different featurebased image partitions to tackle the challenging problem of discovering objects in unlabeled image collections. Inspired by consensus clustering models, we introduce the Aggregated Partition (AP) approach, which, starting from a set of weak input partitions, builds a final partition where the disagreements with the input partitions are optimized. We then generalize the AP formulation and derive the Selective AP, which automatically identifies the subset of features and partitions that further improves the precision of the final partition. Experiments on three challenging datasets show how our methods are able to consistently outperform competing methods, reporting state-of-the-art results.

16.9 Efficient Deep Feature Learning and Extraction via StochasticNets

[Link to Paper]

Deep neural networks are a powerful tool for feature learning and extraction. One area worth exploring in feature extraction using deep neural networks is efficient neural connectivity formation for faster feature learning and extraction. Motivated by findings of stochastic synaptic connectivity formation in the brain as well as the brain's uncanny ability to efficiently represent information, we propose the efficient learning and extraction of features via StochasticNets, where sparsely-connected deep neural networks can be formed via stochastic connectivity between neurons. Experimental results show that features learned using deep convolutional StochasticNets, with fewer neural connections than conventional

deep convolutional neural networks, can allow for better or comparable classification accuracy than conventional deep neural networks. Finally, it was also shown that significant gains in feature extraction speed can be achieved in embedded applications using StochasticNets.

16.10   Learning Discriminative Features with Class Encoder

[Link to Paper]

Deep neural networks usually benefit from unsupervised pre-training, e.g. auto-encoders. However, the classifier further needs supervised fine-tuning methods for good discrimination. Besides, due to the limits of full-connection, the application of auto-encoders is usually limited to small, well aligned images. In this paper, we incorporate the supervised information to propose a novel formulation, namely class-encoder, whose training objective is to reconstruct a sample from another one of which the labels are identical. Class-encoder aims to minimize the intra-class variations in the feature space, and to learn a good discriminative manifolds on a class scale. We impose the class-encoder as a constraint into the softmax for better supervised training, and extend the reconstruction on feature-level to tackle the parameter size issue and translation issue. The experiments show that the class-encoder helps to improve the performance on benchmarks of classification and face recognition. This could also be a promising direction for fast training of face recognition models.

16.11   Fast Image Gradients Using Binary Feature Convolutions

[Link to Paper]

The recent increase in popularity of binary feature descriptors has opened the door to new lightweight computer vision applications. Most research efforts thus far have been dedicated to the introduction of new large-scale binary features, which are primarily used for keypoint description and matching. In this paper, we show that the side products of small-scale binary feature computations can efficiently filter images and estimate image gradients. The improved efficiency of low-level operations can be especially useful in time-constrained applications. Through our experiments, we show that efficient binary feature convolutions can be used to mimic various image processing operations, and even outperform Sobel gradient estimation in the edge detection problem, both in terms of speed and F-Measure.

16.12 Deep Features or Not: Temperature and Time Prediction in Outdoor Scenes

[Link to Paper]

In this paper, we study the effectiveness of features from Convolutional Neural Networks (CNN) for predicting the ambient temperature as well as the time of the year in an outdoor scene. We follow the benchmark provided by Glasner et al. [3] one of whose findings was that simple handcrafted features are better than the deep features (from fully connected layers) for temperature prediction. As in their work, we use the VGG-16 architecture for our CNNs, pretrained for classification on ImageNet. Our main findings on the temperature prediction task are as follows. (i) The pooling layers provide better features than the fully connected layers. (ii) The quality of the features improves little with fine-tuning of the CNN on training data. (iii) Our best setup significantly improves over the results from Glasner et al. showing that the deep features are successful in turning a camera into a crude temperature sensor. Moreover, we validate our findings also for time prediction and achieve accurate season, month, week, time of the day, and hour prediction.

16.13 Robust 2DPCA and Its Application

[Link to Paper]

Two-dimensional Principal Component Analysis (2DPCA) has been widely used for face image representation and recognition. However, 2DPCA, which is based on F-norm square, is sensitive to the presence of outliers. To enhance the robustness of 2DPCA model, we proposed a novel Robust 2DPCA objective function, called R-2DPCA. The criterion of R-2DPCA is maximizing the covariance of data in the projected subspace, while minimizing the reconstruction error of data. In addition, we use the efficient non-greedy optimization algorithms solving our objective function. Extensive experiments are done on the AR, CMU-PIE, Extended Yale B face image databases, and results illustrate that our method is more effective and robust than other robust 2DPCA algorithms, such as L1-2DPCA, L1-2DPCA-S, and N-2DPCA.

16.14 Embedding Sequential Information into Spatiotemporal Features for Action Recognition

[Link to Paper]

In this paper, we introduce a novel framework for videobased action recognition, which incorporates the sequential information with the spatiotemporal features. Specifically, the

393

spatiotemporal features are extracted from the sliced clips of videos, and then a recurrent neural network is applied to embed the sequential information into the final feature representation of the video. In contrast to most current deep learning methods for the video-based tasks, our framework incorporates both long-term dependencies and spatiotemporal information of the clips in the video. To extract the spatiotemporal features from the clips, both dense trajectories (DT) and a newly proposed 3D neural network, C3D, are applied in our experiments. Our proposed framework is evaluated on the benchmark datasets of UCF101 and HMDB51, and achieves comparable performance compared with the state-of-the-art results.

16.15   Generating Discriminative Object Proposals via Submodular Ranking

[Link to Paper]

A multi-scale greedy-based object proposal generation approach is presented. Based on the multi-scale nature of objects in images, our approach is built on top of a hierarchical segmentation. We first identify the representative and diverse exemplar clusters within each scale by using a diversity ranking algorithm. Object proposals are obtained by selecting a subset from the multi-scale segment pool via maximizing a submodular objective function, which consists of a weighted coverage term, a single-scale diversity term and a multi-scale reward term. The weighted coverage term forces the selected set of object proposals to be representative and compact; the single-scale diversity term encourages choosing segments from different exemplar clusters so that they will cover as many object patterns as possible; the multi-scale reward term encourages the selected proposals to be discriminative and selected from multiple layers generated by the hierarchical image segmentation. The experimental results on the Berkeley Segmentation Dataset and PASCAL VOC2012 segmentation dataset demonstrate the accuracy and efficiency of our object proposal model. Additionally, we validate our object proposals in simultaneous segmentation and detection and outperform the state-of-art performance.

17 Automatic Traffic Surveillance

17.1 Fast Classification of Empty and Occupied Parking Spaces Using Integral Channel Features

[Link to Paper]

In this paper we present a novel, fast and accurate system for detecting the presence of cars in parking lots. The system is based on fast integral channel features and machine learning. The methods are well suited for running embedded on low performance platforms. The methods are tested on a database of nearly 700,000 images of parking spaces, where 48.5% are occupied and the rest are free. The experimental evaluation shows improved robustness in comparison to the baseline methods for the dataset.

17.2 Towards Semantic Understanding of Surrounding Vehicular Maneuvers: A Panoramic Vision-Based Framework for Real-World Highway Studies

[Link to Paper]

This paper proposes the use of multiple low-cost visual sensors to obtain a surround view of the ego-vehicle for semantic understanding. A multi-perspective view will assist the analysis of naturalistic driving studies (NDS), by automating the task of data reduction of the observed sequences into events. A user-centric vision-based framework is presented using a vehicle detector and tracker in each separate perspective. Multi-perspective trajectories are estimated and analyzed to extract 14 different events, including potential dangerous behaviors such as overtakes and cut-ins. The system is tested on ten sequences of real-world data collected on U.S. highways. The results show the potential use of multiple low-cost visual sensors for semantic understanding around the ego-vehicle.

17.3 The CountingApp, or How to Count Vehicles in 500 Hours of Video

[Link to Paper]

This paper proposes a new method for counting vehicles based on video tracking. The process consists of two main steps: tracking vehicles and processing the output with minimal user input, separating the vehicle positions into sets of trajectories, which correspond to the paths drivers can take. The method allows to rapidly analyze videos from road sections and intersections, and yields detailed results in the form of turning movement counts. A large dataset of five hundred hours of traffic videos was processed using this method and the results are promising as mean absolute percentage error (MAPE) can get as low as 14 %

depending on the conditions and the quality of the video capture. This paper also discusses the factors that affect counting performance and how to improve counting accuracy.

17.4 Analyzing wheels of vehicles in motion using laser scanning

[Link to Paper]

In this paper, we discuss a lidar-based wheel-width measuring system. Trucks are used for an enormous part of day to day freight delivery across the world, and cause significant wear to the road they use. Road planners and construction engineers rely on traffic statistics to properly design roads. Currently, no system is able to provide them with numbers on the wheel-widths of the vehicles using a particular stretch of road. We present a system which uses a horizontal lidar measuring in a plane close and nearly parallel to the road surface. Input from this is used to detect and analyze tires. A vertical lidar detects passing vehicles so individual tires can be combined into full vehicle models. The system detects 58% of passing vehicles, but correctly counts the number of axles on 85% of detected vehicles. More than 90% of of the axles are correctly classified according to the number of mounted tired (single or dual).

17.5 Semantic Depth Map Fusion for Moving Vehicle Detection in Aerial Video

[Link to Paper]

Wide area motion imagery from an aerial platform offers a compelling advantage in providing a global picture of traffic flows for transportation and urban planning that is complementary to the information from a network of ground-based sensors and instrumented vehicles. We propose an automatic moving vehicle detection system for wide area aerial video based on semantic fusion of motion information with projected building footprint information to significantly reduce the false alarm rate in urban scenes with many tall structures. Motion detections are obtained using the flux tensor and combined with a scene level depth mask to identify tall structures using height information derived from a dense 3D point cloud estimated using multiview stereo from the same source imagery or a prior model. The trace of the flux tensor provides robust spatio-temporal information of moving edges including the motion of tall structures caused by parallax effects. The parallax induced motions are filtered out by incorporating building depth maps obtained from dense urban 3D point clouds. Using a level-set based geodesic active contours framework, the coarse thresholded tall structures depth masks evolved and stopped at the actual building boundaries. Experiments are carried out on a cropped 2k 2k region of interest for 200 frames from Albuquerque urban aerial imagery. An average precision of 83% and

recall of 76% have been reported using an object-level detection performance evaluation method.

17.6   Vehicle Re-Identification for Automatic Video Traffic Surveillance

[Link to Paper]

This paper proposes an approach to the vehicle reidentification problem in a multiple camera system. We focused on the re-identification itself assuming that the vehicle detection problem is already solved including extraction of a full-fledged 3D bounding box. The re-identification problem is solved by using color histograms and histograms of oriented gradients by a linear regressor. The features are used in separate models in order to get the best results in the shortest CPU computation time. The proposed method works with a high accuracy (60 % true positives retrieved with 10 % false positive rate on a challenging subset of the test data) in 85 milliseconds of the CPU (Core i7) computation time per one vehicle re-identification assuming the fullHD resolution video input. The applications of this work include finding important parameters such as travel time, traffic flow, or traffic information in a distributed traffic surveillance and monitoring system.

18 Computer Vision for Microscopy Image Analysis

18.1 Analysing the Structure of Collagen Fibres in SBFSEM Images

[Link to Paper]

Collagen fibres form important structures in tissue, and are essential for force transmission, scaffolding and cell addition. Each fibre is long and thin, and large numbers group together into complex networks of bundles, which are little studied as yet. Serial block-face scanning electron microscopy (SBFSEM) can be used to image tissues containing the fibres, but analysing the images manually is almost impossible - there can be over 30,000 fibres in each image slice, and many hundreds of individual image slices in a volume. We describe a system for automatically identifying and tracking the individual fibres, allowing analysis of their paths, how they form bundles and how individual fibres weave from one bundle to another.

18.2 3D structure modeling of dense capillaries by multi-objects tracking

[Link to Paper]

A newly developed imaging technique called light-sheet laser microscopy imaging can visualize the detailed 3D structures of capillaries. Capillaries form complicated network structures in the obtained data, and this makes it difficult to model vessel structures by existing methods that implicitly assume simple tree structures for blood vessels. To cope with such dense capillaries with network structures, we propose to track the flow of blood vessels along a base-axis using a multiple-object tracking framework. We first track multiple blood vessels in cross-sectional images along a single axis to make the trajectories of blood vessels, and then connect these blood vessels to reveal their entire structures. This framework is efficient to track densely distributed vessels since it uses only a single cross-sectional plane. The network structure is then generated in the postprocessing by connecting blood vessels on the basis of orientations of the trajectories. The results of experiments using a challenging real data-set demonstrate the efficacy of the proposed method, which are capable of modeling dense capillaries.

18.3 Four Dimensional Image Registration For Intravital Microscopy

[Link to Paper]

Increasingly the behavior of living systems is being evaluated using intravital microscopy since it provides subcellular resolution of biological processes in an intact living organism. Intravital microscopy images are frequently confounded by motion resulting from animal

respiration and heartbeat. In this paper we describe an image registration method capable of correcting motion artifacts in three dimensional fluorescence microscopy images collected over time. Our method uses 3D B-Spline non-rigid registration using a coarse-to-fine strategy to register stacks of images collected at different time intervals and 4D rigid registration to register 3D volumes over time. The results show that our proposed method has the ability of correcting global motion artifacts of sample tissues in four dimensional space, thereby revealing the motility of individual cells in the tissue.

18.4 Multi-View Multi-Modal Feature Embedding for Endomicroscopy Mosaic Classification

[Link to Paper]

Probe-based confocal laser endomicroscopy (pCLE) is an emerging tool for epithelial cancer diagnosis, which enables in vivo microscopic imaging during endoscopic procedures. As a new technique, definite clinical diagnosis is still referenced to the gold standard histology images. In this paper, we propose a Multi-View Multi-Modal Embedding framework (MVMME) to learn representative features for pCLE videos exploiting both pCLE mosaic and histology images. Each pCLE mosaic is represented by multiple feature representations including SIFT, Texton and HoG. A latent space is discovered by embedding the visual features from both mosaics and histology images in a supervised scheme. The features extracted from the latent spaces can make use of multi-modal imaging sources that are more discriminative than unimodal features from mosaics alone. The experiments based on real pCLE datasets demonstrate that our approach outperforms, with statistical significance, several single-view or single-modal methods. A binary classification accuracy of 96% has been achieved. (a) (b)

18.5 Spatially Aware Dictionary Learning and Coding for Fossil Pollen Identification

[Link to Paper]

We propose a robust approach for performing automatic species-level recognition of fossil pollen grains in microscopy images that exploits both global shape and local texture characteristics in a patch-based matching methodology. We introduce a novel criteria for selecting meaningful and discriminative exemplar patches. We optimize this function during training using a greedy submodular function optimization framework that gives a near-optimal solution with bounded approximation error. We use these selected exemplars as a dictionary basis and propose a spatially-aware sparse coding method to match testing images for identification while maintaining global shape correspondence. To accelerate the coding process

for fast matching, we introduce a relaxed form that uses spatiallyaware soft-thresholding during coding. Finally, we carry out an experimental study that demonstrates the effectiveness and efficiency of our exemplar selection and classification mechanisms, achieving 86.13% accuracy on a difficult fine-grained species classification task distinguishing three types of fossil spruce pollen.1

18.6 Segmentation of Overlapping Cervical Cells in Microscopic Images with Superpixel Partitioning and Cell-wise Contour Refinement

[Link to Paper]

Segmentation of cervical cells in microscopic images is an important task for computer-aided diagnosis of cervical cancer. However, their segmentation is challenging due to inhomogeneous cell cytoplasm and the overlap between the cells. In this paper, we propose an automatic segmentation method for multiple overlapping cervical cells in microscopic images using superpixel partitioning and cell-wise contour refinement. First, the cell masses are detected by superpixel generation and triangle thresholding. Then, nuclei of cells are extracted by local thresholding and outlier removal. Finally, cell cytoplasm is initially segmented by superpixel partitioning and refined by cell-wise contour refinement with graph cuts. In experiments, our method showed competitive performances in two public challenge data sets compared to the state-of-the-art methods. (a) Input images (b) Nuclei (c) Cytoplasm Figure 1: Examples of microscopic images for overlapping cervical cells, and nucleus and cytoplasm segmentation.

18.7 Cluster Sensing Superpixel and Grouping

[Link to Paper]

Superpixel algorithms have shown significant potential in computer vision applications since they can be used to accelerate other computationally demanding algorithms. However, in contrast to the original purpose of superpixels, many upper layer methods still suffer from computational problems when incorporating superpixel for speedup. In this paper, we present a cluster sensing superpixel (CSS) method to efficiently generate superpixel bricks. Based on the insight of pixel density, cluster centers generally have properties of representativeness (i.e., local maximal pixel density) and isolation (i.e., large distance from other cluster centers). Our CSS method efficiently identifies ideal cluster centers via utilizing pixel density. We also integrate superpixel cues into a bipartite graph segmentation framework and apply it to microscopy image segmentation. Extensive experiments show that our CSS method achieves impressive efficiency, being approximately five times faster than the state-of-the-art methods and having comparable performance in terms of

the standard metrics. Application on microscopy image segmentation also benefits our efficient implementation.

18.8 Methodology for Increasing the Measurement Accuracy of Image Features

[Link to Paper]

We present an optimization methodology for improving the measurement accuracy of image features for low signal to noise ratio (SNR) images. By superimposing known background noise with high quality images in various proportions, we produce a degraded image set spanning a range of SNRs with reference feature values established from the unmodified high quality images. We then experiment with a variety of image processing spatial filters applied to the degraded images and identify which filter produces an image whose feature values most closely correspond to the reference values. When using the best combination of three filters and six kernel sizes for each feature, the average correlation of feature values between the degraded and high quality images increased from 0.6 (without filtering) to 0.92 (with feature-specific filters), a 53% improvement. Selecting a single filter is more practical than having a separate filter per feature. However, this results in a 1.95% reduction in correlation and a 10% increase in feature residual root mean square error compared to selecting the optimal filter and kernel size per feature. We quantified the tradeoff between a practical solution for all features and featurespecific solution to support decision making.

18.9 ICORD: Intelligent Collection of Redundant Data A Dynamic System for Crowdsourcing Cell Segmentations Accurately and Efficiently

[Link to Paper]

Segmentation is a fundamental step in analyzing biological structures in microscopy images. When state-of-the-art automated methods are found to produce inaccurate boundaries, interactive segmentation can be effective. Since the inclusion of domain experts is typically expensive and does not scale, crowdsourcing has been considered. Due to concerns about the quality of crowd work, quality control methods that rely on a fixed number of redundant annotations have been used. We here introduce a collection strategy that dynamically assesses the quality of crowd work. We propose ICORD (Intelligent Collection Of Redundant annotation Data), a system that predicts the accuracy of a segmented region from analysis of (1) its geometric and intensity-based features and (2) the crowd worker's behavioral features. Based on this score, ICORD dynamically determines if the annotation accuracy is satisfactory or if a higher-quality annotation should be sought out in another

403

round of crowdsourcing. We tested ICORD on phase contrast and fluorescence images of 270 cells. We compared the performance of ICORD and a popular baseline method for which we aggregated 1,350 crowd-drawn cell segmentations. Our results show that ICORD collects annotations both accurately and efficiently. Accuracy levels are within 3 percentage points of those of the baseline. More importantly, due to its dynamic nature, ICORD vastly outperforms the baseline method with respect to efficiency. ICORD only uses between 27% and 50% of the resources, i.e., collection time and cost, that the baseline method requires.

18.10 3D Convolutional Networks-Based Mitotic Event Detection in Time-Lapse Phase Contrast Microscopy Image Sequences of Stem Cell Populations

[Link to Paper]

In this paper, we propose a straightforward and effective method for mitotic event detection in time-lapse phase contrast microscopy image sequences of stem cell populations. Different from most of recent methods leveraging temporal modeling to learn the latent dynamics within one mitotic event, we mainly target on the data-driven spatiotemporal visual feature learning for mitotic event representation to bypass the difficulties in both robust hand-crafted feature designing and complicated temporal dynamic learning. Specially, we design the architecture of the convolutional neural networks with 3D filters to extract the holistic feature of the volumetric region where individual mitosis event occurs. Then, the extracted features can be directly fed into the off-the-shelf classifiers for model learning or inference. Moreover, we prepare a novel and challenging dataset for mitosis detection. The comparison experiments demonstrate the superiority of the proposed method.

18.11 Unsupervised segmentation of cervical cell images using Gaussian Mixture Model

[Link to Paper]

Cervical cancer is one of the leading causes of cancer death in women. Screening at early stages using the popular Pap smear test has been demonstrated to reduce fatalities significantly. Cost effective, automated screening methods can significantly improve the adoption of these tests worldwide. Automated screening involves image analysis of cervical cells. Gaussian Mixture Models (GMM) are widely used in image processing for segmentation which is a crucial step in image analysis. In our proposed method, GMM is implemented to segment cell regions to identify cellular features such as nucleus, cytoplasm while addressing shortcomings of existing methods. This method is combined with shape based

identification of nucleus to increase the accuracy of nucleus segmentation. This enables the algorithm to accurately trace the cells and nucleus contours from the pap smear images that contain cell clusters. The method also accounts for inconsistent staining, if any. The results that are presented shows that our proposed method performs well even in challenging conditions.

18.12 Neuron Segmentation based on CNN with Semi-supervised Regularization

[Link to Paper]

Neuron segmentation in two-photon microscopy images is a critical step to investigate neural network activities in vivo. However, it still remains as a challenging problem due to the image qualities, which largely results from the non-linear imaging mechanism and 3D imaging diffusion. To address these issues, we proposed a novel framework by incorporating the convolutional neural network (CNN) with a semi-supervised regularization term, which reduces the human efforts in labeling without sacrificing the performance. Specifically, we generate a putative label for each unlabeled sample regularized with a graph-smooth term, which are used as if they were true labels. A CNN model is therefore trained in a supervised fashion with labeled and unlabeled data simultaneously, which is used to detect neuron regions in 2D images. Afterwards, neuron segmentation in a 3D volume is conducted by associating the corresponding neuron regions in each image. Experiments on real-world datasets demonstrate that our approach outperforms neuron segmentation based on the graph-based semisupervised learning, the supervised CNN and variants of the semi-supervised CNN. powerful and attractive technique to interrogate local neural circuits [22] by using nonlinear light-matter interactions of multiple photons [8]. It enables neuroscientists to measure non-invasive neural network dynamics in live animals with spatial resolution sufficient to discern individual neurons, dendrites, and synapses [12]. Among the tasks of twophoton microscopy image analysis, neuron segmentation is a critical step in that it lays a foundation for the subsequential neuron activity observation. Besides, it is imperative to develop automatic neuron segmentation algorithms with dramatically increasing number of neurons (up to 105 to 106 everyday). Currently, neuron segmentation in two-photon microscopy images remains as a challenging problem because the scattering of the laser and nonlinear imaging mechanism make it difficult to produce sharp images when a large magnification is needed. Besides, neuron segmentation is also challenged by the issue of 3D segmentation in the twophoton microscope volumes because of the diffusion when the lens moves up and down, as a sample image shown in Fig. 1.

19 Context-Based Affect Recognition and Affective Face In-the-Wild

19.1 Facial expression recognition in the wild using improved dense trajectories and Fisher vector encoding

[Link to Paper]

Improved dense trajectory features have been successfully used in video-based action recognition problems, but their application to face processing is more challenging. In this paper, we propose a novel system that deals with the problem of emotion recognition in real-world videos, using improved dense trajectory, LGBP-TOP, and geometric features. In the proposed system, we detect the face and facial landmarks from each frame of a video using a combination of two recent approaches, and register faces by means of Procrustes analysis. The improved dense trajectory and geometric features are encoded using Fisher vectors and classification is achieved by extreme learning machines. We evaluate our method on the extended Cohn-Kanade (CK+) and EmotiW 2015 Challenge databases. We obtain state-ofthe-art results in both databases.

19.2 Gaussian Process Domain Experts for Model Adaptation in Facial Behavior Analysis

[Link to Paper]

We present a novel approach for supervised domain adaptation that is based upon the probabilistic framework of Gaussian processes (GPs). Specifically, we introduce domain-specific GPs as local experts for facial expression classification from face images. The adaptation of the classifier is facilitated in probabilistic fashion by conditioning the target expert on multiple source experts. Furthermore, in contrast to existing adaptation approaches, we also learn a target expert from available target data solely. Then, a single and confident classifier is obtained by combining the predictions from multiple experts based on their confidence. Learning of the model is efficient and requires no retraining/reweighting of the source classifiers. We evaluate the proposed approach on two publicly available datasets for multi-class (MultiPIE) and multi-label (DISFA) facial expression classification. To this end, we perform adaptation of two contextual factors: 'where' (view) and 'who' (subject). We show in our experiments that the proposed approach consistently outperforms both source and target classifiers, while using as few as 30 target examples. It also outperforms the state-of-the-art approaches for supervised domain adaptation.

19.3 Fusing Aligned and Non-Aligned Face Information for Automatic Affect Recognition in the Wild: A Deep Learning Approach

[Link to Paper]

Face alignment can fail in real-world conditions, negatively impacting the performance of automatic facial expression recognition (FER) systems. In this study, we assume a realistic situation including non-alignable faces due to failures in facial landmark detection. Our proposed approach fuses information about non-aligned and aligned facial states, in order to boost FER accuracy and efficiency. Six experimental scenarios using discriminative deep convolutional neural networks (DCNs) are compared, and causes for performance differences are identified. To handle non-alignable faces better, we further introduce DCNs that learn a mapping from non-aligned facial states to aligned ones, alignment-mapping networks (AMNs). We show that AMNs represent geometric transformations of face alignment, providing features beneficial for FER. Our automatic system based on ensembles of the discriminative DCNs and the AMNs achieves impressive results on a challenging database for FER in the wild.

19.4 Towards an "In-the-Wild" Emotion Dataset Using a Game-based Framework

[Link to Paper]

In order to create an "in-the-wild" dataset of facial emotions with large number of balanced samples, this paper proposes a game-based data collection framework. The framework mainly include three components: a game engine, a game interface, and a data collection and evaluation module. We use a deep learning approach to build an emotion classifier as the game engine. Then a emotion web game to allow gamers to enjoy the games, while the data collection module obtains automatically-labelled emotion images. Using our game, we have collected more than 15,000 images within a month of the test run and built an emotion dataset "GaMo". To evaluate the dataset, we compared the performance of two deep learning models trained on both GaMo and CIFE. The results of our experiments show that because of being large and balanced, GaMo can be used to build a more robust emotion detector than the emotion detector trained on CIFE, which was used in the game engine to collect the face images.

19.5 A 3D Mask Face Anti-spoofing Database with Real World Variations

[Link to Paper]

3D mask face spoofing attack becomes new challenge and attracts more research interests in recent years. However, due to the deficiency number and limited variations of database, there are few methods be proposed to aim on it. Meanwhile, most of existing databases only concentrate on the anti-spoofing of different kinds of attacks and ignore the environmental changes in real world applications. In this paper, we build a new 3D mask

anti-spoofing database with more variations to simulate the real world scenario. The proposed database contains 12 masks from two companies with different appearance quality. 7 cameras from the stationary and mobile devices and 6 lighting settings that cover typical illumination conditions are also included. Therefore, each subject contains 42 (7 cameras * 6 lightings) genuine and 42 mask sequences and the total size is 1008 videos. Through the benchmark experiments, directions of the future study are pointed out. We plan to release the database as an platform to evaluate methods under different variations.

19.6 Extended DISFA Dataset: Investigating Posed and Spontaneous Facial Expressions

[Link to Paper]

Automatic facial expression recognition (FER) is an important component of affect-aware technologies. Because of the lack of labeled spontaneous data, majority of existing automated FER systems were trained on posed facial expressions; however in real-world applications we deal with (subtle) spontaneous facial expression. This paper introduces an extension of DISFA, a previously released and well-accepted face dataset. Extended DISFA (DISFA+) has the following features: 1) it contains a large set of posed and spontaneous facial expressions data for a same group of individuals, 2) it provides the manually labeled framebased annotations of 5-level intensity of twelve FACS facial actions, 3) it provides meta data (i.e. facial landmark points in addition to the self-report of each individual regarding every posed facial expression). This paper introduces and employs DISFA+, to analyze and compare temporal patterns and dynamic characteristics of posed and spontaneous facial expressions.

19.7 Facial Expression Recognition from World Wild Web

[Link to Paper]

Recognizing facial expression in a wild setting has remained a challenging task in computer vision. The World Wide Web is a good source of facial images which most of them are captured in uncontrolled conditions. In fact, the Internet is a Word Wild Web of facial images with expressions. This paper presents the results of a new study on collecting, annotating, and analyzing wild facial expressions from the web. Three search engines were queried using 1250 emotion related keywords in six different languages and the retrieved images were mapped by two annotators to six basic expressions and neutral. Deep neural networks and noise modeling were used in three different training scenarios to find how accurately facial expressions can be recognized when trained on noisy images collected from the web using query terms (e.g. happy face, laughing man, etc)? The results of our

experiments show that deep neural networks can recognize wild facial expressions with an accuracy of 82.12%.

19.8 Automatic Recognition of Emotions and Membership in Group Videos

[Link to Paper]

Automatic affect analysis and understanding has become a well established research area in the last two decades. However, little attention has been paid to the analysis of the affect expressed in group settings, either in the form of affect expressed by the whole group collectively or affect expressed by each individual member of the group. This paper presents a framework which, in group settings automatically classifies the affect expressed by each individual group member along both arousal and valence dimensions. We first introduce a novel Volume Quantised Local Zernike Moments Fisher Vectors (vQLZM-FV) descriptor to represent the facial behaviours of individuals in the spatio-temporal domain and then propose a method to recognize the group membership of each individual (i.e., which group the individual in question is part of) by using their face and body behavioural cues. We conduct a set of experiments on a newly collected dataset that contains fourteen recordings of four groups, each consisting of four people watching affective movie stimuli. Our experimental results show that (1) the proposed vQLZM-FV outperforms the other feature representations in affect recognition, and (2) group membership can be recognized using the non-verbal face and body features, indicating that individuals influence each other's behaviours within a group setting.

19.9 Sequential Face Alignment via Person-Specific Modeling in the Wild

[Link to Paper]

Sequential face alignment, in essence, deals with nonrigid deformation that changes over time. Although numerous methods have been proposed to show impressive success on still images, many of them still suffer from limited performance when it comes to sequential alignment in wild scenarios, e.g., involving large pose/expression variations and partial occlusions. The underlying reason is that they usually perform sequential alignment by independently applying models trained offline in each frame in a tracking-by-detection manner but completely ignoring temporal constraints that become available in sequence. To address this issue, we propose to exploit incremental learning for person-specific alignment. Our approach takes advantage of part-based representation and cascade regression for robust and efficient alignment on each frame. More importantly, it incrementally updates the representation subspace and simultaneously adapts the cascade regressors in parallel using a unified framework. Person-specific modeling is eventually achieved on the

fly while the drifting issue is significantly alleviated by erroneous detection using both part and holistic descriptors. Extensive experiments on both controlled and in-the-wild datasets demonstrate the superior performance of our approach compared with the state of the arts in terms of fitting accuracy and efficiency.

19.10 Towards Facial Expression Recognition in the Wild: A New Database and Deep Recognition System

[Link to Paper]

Automatic facial expression recognition (FER) plays an important role in many fields. However, most existing FER techniques are devoted to the tasks in the constrained conditions, which are different from actual emotions. To simulate the spontaneous expression, the number of samples in acted databases is usually small, which limits the ability of facial expression classification. In this paper, a novel database for natural facial expression is constructed leveraging the social images and then a deep model is trained based on the naturalistic dataset. An amount of social labeled images are obtained from the image search engines by using specific keywords. The algorithms of junk image cleansing are then utilized to remove the mislabeled images. Based on the collected images, the deep convolutional neural networks are learned to recognize these spontaneous expressions. Experiments show the advantages of the constructed dataset and deep approach.

19.11 A Framework for Joint Estimation and Guided Annotation of Facial Action Unit Intensity

[Link to Paper]

Manual annotation of facial action units (AUs) is highly tedious and time-consuming. Various methods for automatic coding of AUs have been proposed, however, their performance is still far below of that attained by expert human coders. Several attempts have been made to leverage these methods to reduce the burden of manual coding of AU activations (presence/absence). Nevertheless, this has not been exploited in the context of AU intensity coding, which is a far more difficult task. To this end, we propose an expertdriven probabilistic approach for joint modeling and estimation of AU intensities. Specifically, we introduce a Conditional Random Field model for joint estimation of the AU intensity that updates its predictions in an iterative fashion by relying on expert knowledge of human coders. We show in our experiments on two publicly available datasets of AU intensity (DISFA and FERA2015) that the AU coding process can significantly be facilitated by the proposed approach, allowing human coders to faster make decisions about target AU intensity.

411

19.12 Facial Affect "in-the-wild": A survey and a new database

[Link to Paper]

Well-established databases and benchmarks have been developed in the past 20 years for automatic facial behaviour analysis. Nevertheless, for some important problems regarding analysis of facial behaviour, such as (a) estimation of affect in a continuous dimensional space (e.g., valence and arousal) in videos displaying spontaneous facial behaviour and (b) detection of the activated facial muscles (i.e., facial action unit detection), to the best of our knowledge, well-established in-the-wild databases and benchmarks do not exist. That is, the majority of the publicly available corpora for the above tasks contain samples that have been captured in controlled recording conditions and/or captured under a very specific milieu. Arguably, in order to make further progress in automatic understanding of facial behaviour, datasets that have been captured in inthe-wild and in various milieus have to be developed. In this paper, we survey the progress that has been recently made on understanding facial behaviour in-the-wild, the datasets that have been developed so far and the methodologies that have been developed, paying particular attention to deep learning techniques for the task. Finally, we make a significant step further and propose a new comprehensive benchmark for training methodologies, as well as assessing the performance of facial affect/behaviour analysis/understanding in-the-wild. To the best of our knowledge, this is the first time that such a benchmark for valence and arousal "in-the-wild" is presented.

19.13 Recurrent Convolutional Neural Network Regression for Continuous Pain Intensity Estimation in Video

[Link to Paper]

Automatic pain intensity estimation possesses a significant position in healthcare and medical field. Traditional static methods prefer to extract features from frames separately in a video, which would result in unstable changes and peaks among adjacent frames. To overcome this problem, we propose a real-time regression framework based on the recurrent convolutional neural network for automatic frame-level pain intensity estimation. Given vector sequences of AAM-warped facial images, we used a slidingwindow strategy to obtain fixed-length input samples for the recurrent network. We then carefully design the architecture of the recurrent network to output continuousvalued pain intensity. The proposed end-to-end pain intensity regression framework can predict the pain intensity of each frame by considering a sufficiently large historical frames while limiting the scale of the parameters within the model. Our method achieves promising results regarding both accuracy and running speed on the published UNBCMcMaster Shoulder Pain Expression Archive Database.

20 Moving Cameras Meet Video Surveillance: From Body-Borne Cameras to Drones

20.1 Covert Video Classification by Codebook Growing Pattern

[Link to Paper]

Recent advances in visual data acquisition and Internet technologies make it convenient and popular to collect and share videos. These activities, however, also raise the issue of privacy invasion. One potential privacy threat is the unauthorized capture and/or sharing of covert videos, which are recorded without the awareness of the subject(s) in the video. Automatic classification of such videos can provide an important basis toward addressing relevant privacy issues. The task is very challenging due to the large intra-class variation and between-class similarity, since there is no limit in the content of a covert video and it may share very similar content with a regular video. The challenge brings troubles when applying existing content-based video analysis methods to covert video classification. In this paper, we propose a novel descriptor, codebook growing pattern (CGP), which is derived from latent Dirichlet allocation (LDA) over optical flows. Given an input video V , we first represent it with a sequence of histograms of optical flow (HOF). After that, these HOFs are fed into LDA to dynamically generate the codebook for V . The CGP descriptor is then defined as the growing codebook sizes in the LDA procedure. CGP fits naturally for covert video representation since (1) optical flows can capture the camera motion that characterizes the covert video acquisition, and (2) CGP by itself is insensitive to video content. To evaluate the proposed approach, we collected a large covert video dataset, the first such dataset to our knowledge, and tested the proposed method on the dataset. The results show clearly the effectiveness of the proposed approach in comparison with other state-of-the-art video classification algorithms.

20.2 Detecting Anomalous Objects on Mobile Platforms

[Link to Paper]

We present an approach where a robot patrols a fixed path through an environment, autonomously locating suspicious or anomalous objects. To learn, the robot patrols this environment building a dictionary describing what is present. The dictionary is built by clustering features from a deep neural network. The objects present vary depending on the scene, which means that an object that is anomalous in one scene may be completely normal in another. To reason about this, the robot uses a computational cognitive model to learn the dictionary elements that are typically found in each scene. Once the dictionary and model has been built, the robot can patrol the environment matching objects against the dictionary, and querying the model to find the most likely objects present and to determine which objects (if any) are anomalous. We demonstrate our approach by patrolling two indoor and one outdoor environments. Hallway Normal Cubicles Open Door

20.3 Surveillance (oversight), Sousveillance (undersight), and Metaveillance (seeing sight itself)

[Link to Paper]

Surveillance is an established practice that generally involves fixed cameras attached to fixed inanimate objects, or PTZ (Pan Tilt Zoom) cameras at a fixed position. Sur-veillance only provides part of the veillance story, and often only captures a partial truth. Further advances in miniaturization, together with wireless communication technologies, are giving rise to kinematic veillance ("kineveillance"): wearable, portable, and mobile cameras, as well as unpersoned aerial vehicles (UAVs). These additional veillances give us a more complete picture: multiple viewpoints from multiple entities bring us closer to the truth. In contrast to the extensive mathematical and conceptual framework developed around surveillance (e.g. background subtraction, frame-differencing, etc.), now that surveillance is no longer the only veillance, we need new mathematical and conceptual understandings of imaging and image processing. More importantly we need new tools for understanding the many veillances and how they interact. Therefore this paper introduces metaveillance and the veillance wavefunction for metasensing: the sensing of sensors and the sensing of their capacity to sense.

20.4 Robust Detection of Moving Vehicles in Wide Area Motion Imagery

[Link to Paper]

Multiple object tracking in Wide Area Motion Imagery (WAMI) data is usually based on initial detections coming from background subtraction or frame differencing. However, these methods are prone to produce split and merged detections. Appearance based vehicle detection can be an alternative but is not well-suited for WAMI data since classifier models are of weak discriminative power for vehicles in top view at low resolution. We introduce a moving vehicle detection algorithm that combines 2-frame differencing with a vehicle appearance model to improve object detection. Our main contributions are (1) integration of robust vehicle detection with split/merge handling and (2) estimation of assignment likelihoods between object hypotheses in consecutive frames using an appearance based similarity measure. Without using any prior knowledge, we achieve state-of-the-art detection rates and produce tracklets that considerably simplify the data association problem for multiple object tracking.

20.5 Real-time Vehicle Tracking in Aerial Video using Hyperspectral Features

[Link to Paper]

Vehicle tracking from a moving aerial platform poses a number of unique challenges including the small number of pixels representing a vehicle, large camera motion, and parallax error. This paper considers a multi-modal sensor to design a real-time persistent aerial tracking system. Wide field of view (FOV) panchromatic imagery is used to remove global camera motion whereas narrow FOV hyperspectral image is used to detect the target of interest (TOI). Hyperspectral features provide distinctive information to reject objects with different reflectance characteristics from the TOI. This way the density of detected vehicles is reduced, which increases tracking consistency. Finally, we use a spatial data based classifier to remove spurious detections. With such framework, parallax effect in non-planar scenes is avoided. The proposed tracking system is evaluated in a dense, synthetic scene and outperforms other state-of-theart traditional and aerial object trackers.