Person Re-identification for Improved Multi-person Multi-camera Tracking by Continuous Entity Association

Neeti Narayan[1], Nishant Sankaran[1], Devansh Arpit[2], Karthik Dantu[1], Srirangaraj Setlur[1] and Venu Govindaraju[1]
[1]University at Buffalo, SUNY
{neetinar, n6, kdantu, setlur, govind}@buffalo.edu
[2]Université de Montréal
devansharpit@gmail.com

Abstract

We present a novel approach to person tracking within the context of entity association. In large-scale distributed multi-camera systems, person re-identification is a challenging computer vision task as the problem is two-fold: detecting entities through identification and recognition techniques; and connecting entities temporally by associating them in often crowded environments. Since tracking essentially involves linking detections, we can reformulate it purely as a re-identification task. The inherent advantage of such a reformulation lies in the ability of the tracking algorithm to effectively handle temporal discontinuities in multi-camera environments. To accomplish this, we model human appearance, face biometric and location constraints across cameras. We do not make restrictive assumptions such as number of people in a scene. Our approach is validated by using a simple and efficient inference algorithm. Results on two publicly available datasets, CamNeT and DukeMTMC, are significantly better compared to other existing methods.

1. Introduction

With the increase in the number of deployed surveillance cameras, there is an increase in the workload of video operators to manually analyze and understand video content to monitor long term activity and behavior characterization of people in a scene. Automated analysis of large amounts of data can not only process the data faster but significantly improve the quality of surveillance. Automatic re-identification and tracking in dense crowds will allow continuous monitoring and analysis of events without relying on constant human-interaction.

Tracking multiple people across multiple cameras is not a trivial task, especially in complex and crowded scenarios with frequent occlusions and interaction of individuals. The task itself constitutes modeling vast variety of data present in videos that may include long-term occlusion, different scene illuminations, camera properties or varying number of people. In recent years, we can find a lot of literature on human detection techniques [30, 11, 22] which enable tracking-by-detection as a useful tracking strategy. The main idea is to find person detections, estimate the motion patterns of all targets and link tracklets across time to form trajectories. However, the linking step, called data association, is particularly challenging due to frequent occlusions, spurious detections and dense crowds. Researchers developed more sophisticated models such as optimization based on a discrete-continuous energy [2], modeling social and grouping behavior [15], and integration of additional motion constraints such as local flow descriptors [5] to deal with these issues. However, the underlying features used in these methods for data association limit the accuracy and do not work well in crowded environments.

Since multi-target multi-camera tracking involves having to continuously employ motion based tracking within cameras and a separate process of re-identification for persons crossing the camera boundaries, our proposed approach combines the two acts of tracking within and across cameras and reformulates it as a single problem of continuous re-identification. This unification of the two disjoint tasks presents a much clearer and simpler solution which has the advantage of not requiring temporally contiguous sequences of video frames for tracking. The approach therefore can handle discretization of continuous video segments to points of significant changes in activity of targets (like persons moving from one corner to the center of the camera or exit the camera to reappear later in another camera) and automatically work with time-based sparseness in video data. Our method infers the camera topology inherently from the motion patterns observed in the environment and also exploits appearance based features to improve the performance of data associations for re-identification.

The availability of large annotated data and recent rise of deep learning has led us to adapt the Convolutional Neural Network (CNN) learning methodology to multi-camera tracking. Our feature learning framework comprises three parts: (a) Appearance features are extracted from person detection bounding box; (b) Face-biometric features are extracted from face bounding box; and (c) Transition probabilities within and across cameras are learned. We model these constraints and use them to understand and associate detections in real world environments. This is valuable for a range of applications including surveillance, activity recognition and behavior characterization.

Given the learned, pairwise data association score, we link all available detections across frames by computing most probable associations using the proposed inference algorithm which is optimally equivalent to a greedy algorithm. The evaluation is performed on two challenging datasets, CamNeT [32] and DukeMTMC [24].

Our main contributions are as follows:

1. The notion of person tracking is reformulated as an association problem. We explicitly address the influence of human appearance, biometric and location information on human re-identification.

2. Our tracking algorithm can handle temporal gaps in the input video and we show how making the best possible associations is equivalent to a greedy algorithm.

2. Related Work

2.1. Person Re-identification

Person re-identification or consistent labeling, i.e. the capability of associating together the views of the same person captured in different places or at different times is an open problem. Much of the re-identification research in computer vision is applicable to static images [17, 23, 18]. Zheng et al. [33] used group context by proposing ratio-occurrence descriptors to capture groups. Few methods have been developed for videos [27, 6] but they focus on non-crowd surveillance scenarios, where observations are sparse and appearance is distinctive. The spatio-temporal relationships across cameras [3, 28] have also been used for human re-identification. Some researches like Chen et al. [4] have made use of prior knowledge about camera topology to learn spatio-temporal relationships and appearance relationships across networked cameras. Likewise, Mazzon et al. [20] use prior knowledge about spatial location of cameras to model potential paths a person can choose to follow. However, it is not always possible to obtain camera topology information. In this sense, our approach is more robust as it does not depend on the availability of camera topology information.

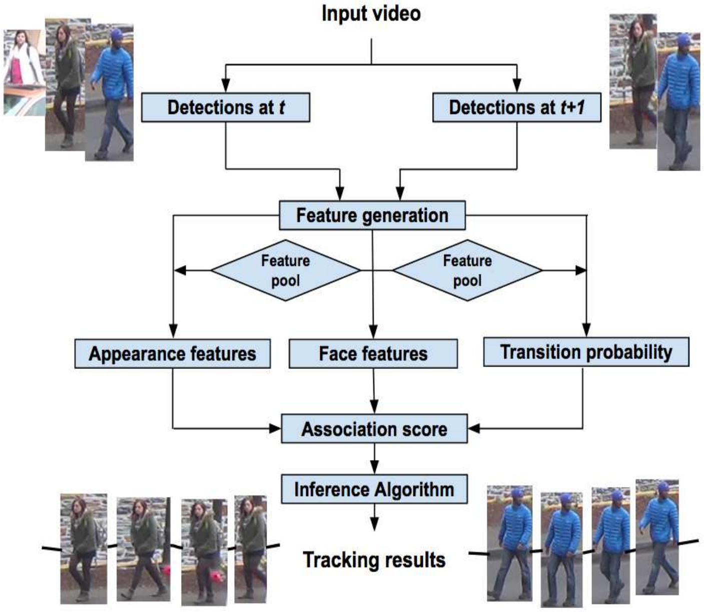

Figure 1. Block diagram of our proposed method

2.2. Person Tracking

Multi-object tracking methods have been reviewed intensively by W. Luo et al. [19]. However, multi-person tracking still remains a challenging problem, particularly in crowded environments. In [12, 21, 13], trajectories are clustered as a means to learn motion patterns. Another approach is to develop more complex motion models to better predict future trajectories, notably models that consider human-human interactions [2]. However, an issue with such models is that they handcraft features for each external influence such as occlusion, or walking in groups; making such an approach less scalable. We also observe diminishing returns when taking into account the added complexity and resulting computation.

2.3. Deep Learning

Convolutional Neural Networks (CNN) are a popular choice for end-to-end learning of image representations due to their accuracy and scalability. Many researchers have used convolutional architectures for supervised learning tasks and assessing a pair of images for different applications including face verification [29] and optical flow estimation [9]. In [17, 31], deep learning approach is used for re-identification. Much of the research on re-identification using deep learning focuses on finding an improved network architecture, an effective set of features, or similarity function for comparing features. J. Fan et al. [8] used CNNs to predict the location and scale of an individual for tracking. Given the location, scale, current and previous image frames, they train CNNs offline to learn spatial and temporal features. Similarly, A. Alahi et al. in [1], predict trajectories of people based on their past positions using Long-Short Term Memory networks (LSTM). However, they assume that the number of people in a scene is known a priori. In our approach, we do not make such an assumption.

3. Our Approach

In this section, we discuss the proposed steps involved in effectively learning detection associations and tracking people. The block diagram of our model is shown in Figure 1. First, we detect entities at each timestamp. It should be noted that while the general assumption of having contiguous frames for tracking is beneficial, it is not a necessity and the system we propose can handle temporal gaps in the video stream. Next, we extract modalities including appearance, face and location from these detections. Then, the association probability matrix is constructed based on the current and previous timestamp detections and pairwise association scores. Finally, the most probable associations are linked using the proposed inference algorithm to form trajectories.

3.1. Multimodal Inference Cues

We employ appearance, face and probable destination cues, designed to be applicable for re-identification. In this section, we discuss the feature extraction methodology. Features are extracted individually for each timestamp for all detections across cameras.

3.1.1 Appearance features

We first extract appearance-based attributes from person detections. They capture the individual traits and characteristics in the form of appearance. A common denominator for image representations are Convolutional Neural Networks (CNN). We use AlexNet model [14] that is pre-trained on ImageNet [25] as appearance feature extractor. This is done by removing the top output layer and using the activations from the last fully connected layer as features (length of 4096).

The AlexNet architecture comprises of five convolutional layers, three fully-connected layers, and three max-pooling layers following the first, second and fifth convolutional layers. The first convolutional layer has 96 filters of size 11 × 11, the second layer has 256 filters of size 5 × 5, the third, fourth and fifth layers are connected to one another without any intervening pooling and have 384, 384 and 256 filters of size 3 × 3 respectively. A fully-connected layer L learns a non-linear function $y_i^L = f(Wy_i^{L-1} + b)$, where y_i^L, W and b are the hidden representation of input x_i, weights and bias respectively, and f is the rectified linear unit activation for hidden layers i.e. $f(x) = max(0, x)$.

3.1.2 Face features

Face biometric is an established biometric for identity recognition and verification. Face modality can be used for the purpose of re-identification as it is inherently a contactless biometric and can be extracted from a distance. We use VGG-16 model [26] that is pre-trained on ImageNet for extracting facial features from face bounding box. This is done by removing the top output layer and using the activations from the last fully connected layer as face features (length of 4096).

VGG-16 is a convolutional neural network; its architecture consists of thirteen convolutional layers and three fully-connected layers. The filters are of size 3 × 3. Pooling is applied between the convolutional layers with 2 × 2 pixel window, with stride 2. Mean subtraction on the training set is used as a pre-processing step.

3.1.3 Location transition

Here, we describe the location constraint, which is linear in nature and predicts the most probable paths within and across cameras. For re-identification and tracking in multiple cameras, the knowledge about probable destination acts as a prior for a person's location in another camera view. We model the transition probability distributions by learning repetitive patterns that occur in camera networks. It is likely that individuals exiting a camera view from a particular grid space will enter another camera view from another specific grid space.

We model state transition probability distribution as a Markov chain. Each camera view is divided into n states. Assuming there are k cameras, the total number of states $N = n \times k$. A Markov chain is characterized by an $N \times N$ transition probability matrix P, each entry is in the interval $[0, 1]$ and the sum of the entries in each row add up to 1.

$$\forall S_i, S_j, P_{S_i,S_j} \in [0, 1] \tag{1}$$

$$\forall S_i, \sum_{j=1}^{N} P_{S_i,S_j} = 1 \tag{2}$$

Thus, by Markov property, we estimate the probability distribution of transition between states S_i and S_j as:

$$\begin{aligned} P(S_i, S_j) &= \Pr(X_t = S_j | X_{t-1} = S_i) \\ &= \frac{|X_t = S_j \wedge X_{t-1} = S_i|}{\sum_{k=1}^{N} |X_t = S_k \wedge X_{t-1} = S_i|} \end{aligned} \tag{3}$$

3.2. Inference Algorithm

At every timestep, the problem of re-identification can be expressed in terms of an association matrix where each row

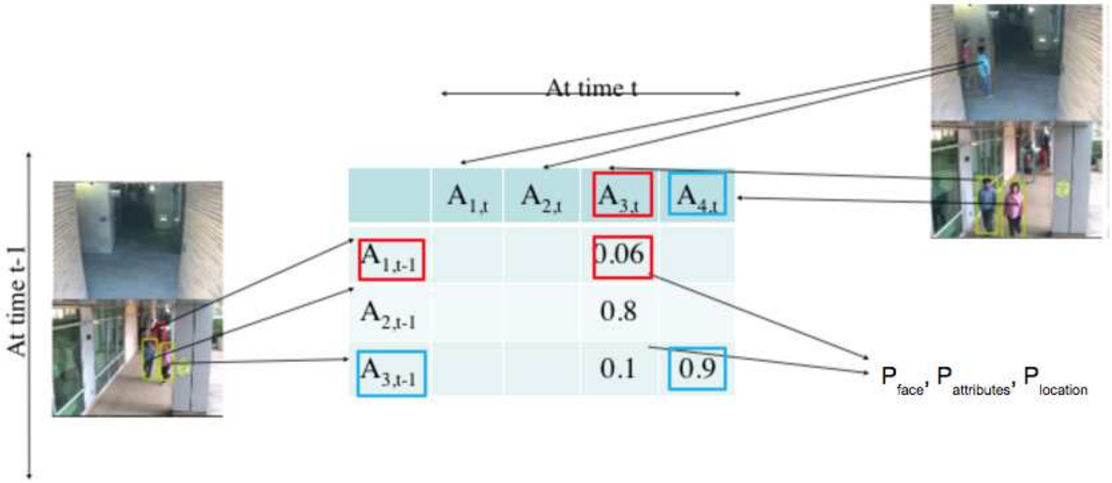

Figure 2. Inference Algorithm

represents a previously seen entity and the columns hold currently active entities. The task of making the best possible associations for every row to a column based on features or attributes of the concerned entity can be formulated as a linear programming problem as below:

$$\max_{\mathbf{W}} \mathbf{P}.\mathbf{W}$$
$$s.t \;\; \mathbf{W} \in [0,1], \mathbf{W1} = \mathbf{1}, \; \mathbf{1}^T\mathbf{W} = \mathbf{1} \tag{4}$$

where **P** is the association matrix or the probability matrix that stores the matching probabilities of the entities being associated and **W** is the weight matrix to be optimized. Figure 2 depicts how the proposed inference algorithm works on the association matrix **P**. The matching probabilities in the association matrix are the cosine distances of each mid-level attribute and face features computed using the pre-trained AlexNet and VGG-16 models respectively as described in the previous section or the location score which is the transition probability modeling the likely movement pattern between entities.

The constraint $\mathbf{W1} = \mathbf{1}$ acts to normalize the matching probabilities across the columns and enforces them to sum to 1 for every previous entity. From the formulation of this constraint, it is apparent that there will only be one maxima for every previous entity's set of association probabilities. This implies that each previous entity would only be associated to at most one of the current entities. So choosing values for the weight matrix **W** essentially reduces to assigning a value of 1 for the best associations and consequently, computing the most probable associations is optimally equivalent to a greedy approach of selecting the largest matching probability sequentially.

Dataset	Reference	Year
UCY	[16]	2007
ETHZ	[7]	2008
VIPeR	[10]	2008
CamNeT	[32]	2015
DukeMTMC	[24]	2016

Table 1. Datasets

4. Experiments and Results

4.1. Datasets

During the past few years, many datasets have been collected for the purpose of person re-identification. Mainly due to the tedious and time-consuming task of video annotation, only limited amount of labelled data for tracking in multi-camera setups is publicly available today. Table 1 shows a list of popular datasets. Since our constraints depend on time and motion information in the videos, many commonly evaluated datasets such as VIPeR [10] and ETHZ [7] cannot be used. We evaluate the proposed approach on CamNeT and DukeMTMC dataset. In this section, we introduce the datasets and the ground truth that was generated for evaluation.

4.1.1 CamNeT

CamNeT is a non-overlapping camera network tracking data set for multi-target tracking. It has more than 1600 frames, each of resolution 640 by 480 pixels, 20-30fps video taken by 8 cameras which cover both indoor and outdoor scenes at a university. The paths of around 10 to 25 people are predefined while several unknown persons move through the scene. There are 6 scenarios, each of which

Feature	CamNeT: Score(%)	DukeMTMC: Score(%)
Attribute	99.37	99.99
Face	96.56	92.07
Location transition	98.28	98.73

Table 2. AUC Scores

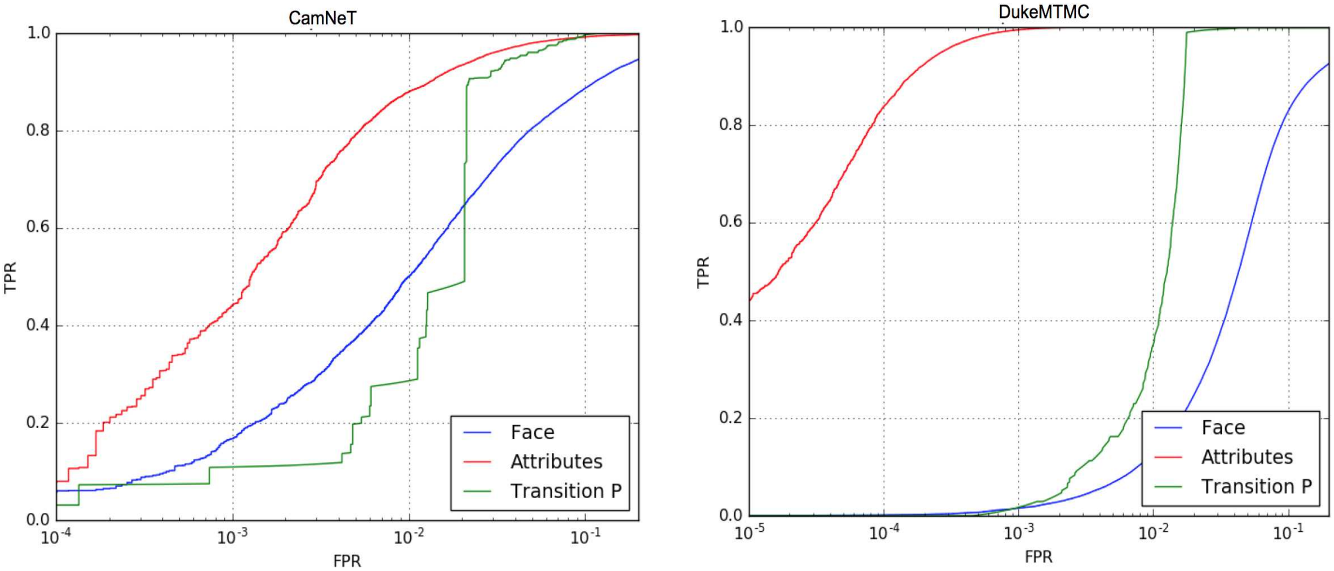

Figure 3. Performance accuracy on CamNeT (left) and DukeMTMC (right) datasets.

lasts at least 5 minutes and is within the view of 5 to 8 cameras. For our experiments, we use Scenario 1. Since the ground truth lacks unique identities, we had to perform manual tagging and rectify erroneous tracking ground truth. Annotated data including person identity and timestamp are provided to the CamNeT dataset maintainers to be hosted alongside the dataset in their webpage.

4.1.2 DukeMTMC

DukeMTMC dataset was released recently to help accelerate performance of multi-target, multi-camera tracking systems. It has more than 2 million frames of high resolution 1080p, 60fps video, observing more than 2700 identities and includes surveillance footage from 8 cameras with approximately 85 minutes of videos for each camera. Since only the ground truth for training data is made available so far, we report results on training set only. We select cameras 1 and 3 for appearance-based multi-camera tracking experiments.

4.2. Evaluation Metric

Since the task in hand is a continuous entity association problem, existing tracking evaluation metric (Multiple Object Tracking Accuracy) is not suitable. Hence, we use a metric for continuous re-identification evaluation as shown below.

$$E = \frac{1}{T}\sum_{t=1}^{T} \frac{\text{number of misclassified detections at time } t}{\text{total number detections at time } t} \tag{5}$$

4.3. Results and Comparison

This section presents the results validating the efficiency of the proposed approach to match pairs of people detections as well as its performance when creating trajectories by means of the proposed inference algorithm.

4.3.1 Data Association

By means of the ROC curve, we first evaluate the performance of our approach when computing the probability of two detections belonging to the same track. The prediction on CamNeT and DukeMTMC datasets is shown in Figure 3. Table 2 shows the AUC scores of individual features evaluated on CamNeT and DukeMTMC dataset respectively. Three result groups are depicted: first, when appearance-based attribute features are extracted (best AUC on CamNeT: 0.993, best AUC on DukeMTMC: 0.999), second, when using the face features (best AUC on CamNeT: 0.965, best AUC on DukeMTMC: 0.920), and last, when using the

Feature	CamNeT: Error(%)	DukeMTMC: Error(%)
Attribute	2.9	0.01
Face	4.67	12.07
Location transition	4.49	0.5

Table 3. Inference Error Rate

Method	Cam 1	Cam 2	Cam 3	Cam 4	Cam 5	Cam 6	Cam 7	Cam 8
Baseline results [24]	366	1929	336	403	292	3370	675	365
Ours	**34**	**47**	**102**	**42**	**69**	**84**	**139**	**12**

Table 4. Single-camera fragmentation measure comparison on DukeMTMC dataset

Method	XFrag
Baseline results [32]	27
Method in [27]	24
Ours	**5**

Table 5. Crossing fragments (XFrag) measure comparison on CamNeT dataset

transition probability (best AUC on CamNeT: 0.982, best AUC on DukeMTMC: 0.987).

Among the three features, appearance features perform best, even at low FAR (False Acceptance Rate). For instance, $TAR@0.01FAR = 99.99\%$, $TAR@0.001FAR = 99.45\%$, $TAR@0.0001FAR = 83.76\%$ on DukeMTMC dataset; where TAR is the True Acceptance Rate.

4.3.2 Person Tracking

Table 3 shows the inference error rates for multi-person multi-camera tracking using the proposed inference algorithm. Since our tracking result is based on re-identification, we use the following existing performance measures for comparison:

1. Crossing fragments (XFrag): The number of true associations missed by the tracking system.
2. Fragmentation: The number of identity switches in the tracking result, when the corresponding ground-truth identity does not change.

We compare our appearance-based tracking results to two tracking methods evaluated on CamNeT dataset. The first is the baseline system [32] evaluated considering social grouping model (SGM), along with temporal and spatial constraints. The second [27] tracks multiple people by considering the long-term interdependence of features over space and time. We show the results in table 5, which indicate that the proposed method outperforms the state-of-the-art method. In table 4, we compare our DukeMTMC results with the baseline system performance. The results show that our approach of continuous entity association for single-camera tracking is better and contains far less fragmentations.

5. Conclusion and Future Work

In this paper, we present an efficient association based approach for person tracking. We address the difficulty and challenges of re-identification and association of people across cameras often in crowded environment. We model human appearance, face biometric and location transition and evaluate each feature's performance individually. We formulate a continuous evaluation metric for the problem under consideration. We highlight the efficiency of our system by comparing our results with the baseline system's performance on CamNeT and DukeMTMC datasets. To the best of our knowledge, this is the first work towards multi-person multi-camera tracking by continuous entity association.

In our future research, we intend to propagate associations in order to make a better prediction by learning from past errors (misassociations). Such an approach can deal with temporally local difficulties generated by occlusion.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant IIP #1266183.

References

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–971, 2016. 2

[2] A. Andriyenko, K. Schindler, and S. Roth. Discrete-continuous optimization for multi-target tracking. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1926–1933. IEEE, 2012. 1, 2

[3] S. Ardeshir and A. Borji. Ego2top: Matching viewers in egocentric and top-view videos. In *European Conference on Computer Vision*, pages 253–268. Springer, 2016. 2

[4] K.-W. Chen, C.-C. Lai, Y.-P. Hung, and C.-S. Chen. An adaptive learning method for target tracking across multiple cameras. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 2

[5] W. Choi. Near-online multi-target tracking with aggregated local flow descriptor. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3029–3037, 2015. 1

[6] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury. Consistent re-identification in a camera network. In *European Conference on Computer Vision*, pages 330–345. Springer, 2014. 2

[7] A. Ess, B. Leibe, K. Schindler, , and L. van Gool. A mobile vision system for robust multi-person tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*. IEEE Press, June 2008. 4

[8] J. Fan, W. Xu, Y. Wu, and Y. Gong. Human tracking using convolutional neural networks. *IEEE Transactions on Neural Networks*, 21(10):1610–1623, 2010. 2

[9] P. Fischer, A. Dosovitskiy, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. *arXiv preprint arXiv:1504.06852*, 2015. 2

[10] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer, 2008. 4

[11] C. Huang and R. Nevatia. High performance object detection by collaborative learning of joint ranking of granules features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 41–48. IEEE, 2010. 1

[12] K. Kim, D. Lee, and I. Essa. Gaussian process regression flow for analysis of motion trajectories. In *Computer vision (ICCV), 2011 IEEE international conference on*, pages 1164–1171. IEEE, 2011. 2

[13] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila. Context-based pedestrian path prediction. In *European Conference on Computer Vision*, pages 618–633. Springer, 2014. 2

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3

[15] L. Leal-Taixé, G. Pons-Moll, and B. Rosenhahn. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 120–127. IEEE, 2011. 1

[16] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. In *Computer Graphics Forum*, volume 26, pages 655–664. Wiley Online Library, 2007. 4

[17] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014. 2

[18] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1629–1642, 2015. 2

[19] W. Luo, X. Zhao, and T.-K. Kim. Multiple object tracking: A review. *arXiv preprint arXiv:1409.7618*, 1, 2014. 2

[20] R. Mazzon, S. F. Tahir, and A. Cavallaro. Person re-identification in crowd. *Pattern Recognition Letters*, 33(14):1828–1837, 2012. 2

[21] B. T. Morris and M. M. Trivedi. Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. *IEEE transactions on pattern analysis and machine intelligence*, 33(11):2287–2301, 2011. 2

[22] W. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2056–2063, 2013. 1

[23] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1846–1855, 2015. 2

[24] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer, 2016. 2, 4, 6

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 3

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[27] B. Song and A. K. Roy-Chowdhury. Robust tracking in a camera network: A multi-objective optimization framework. *IEEE Journal of Selected Topics in Signal Processing*, 2(4):582–596, 2008. 2, 6

[28] C. Stauffer. Learning to track objects through unobserved regions. In *Application of Computer Vision, 2005. WACV/MOTIONS'05 Volume 1. Seventh IEEE Workshops on*, volume 2, pages 96–102. IEEE, 2005. 2

[29] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014. 2

[30] X. Wang, T. X. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 32–39. IEEE, 2009. 1

[31] D. Yi, Z. Lei, and S. Z. Li. Deep metric learning for practical person re-identification. *arXiv preprint arXiv:1407.4979*, 2014. 2

[32] S. Zhang, E. Staudt, T. Faltemier, and A. K. Roy-Chowdhury. A camera network tracking (camnet) dataset and performance baseline. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 365–372. IEEE, 2015. 2, 4, 6

[33] W. Zheng, S. Gong, and T. Xiang. Associating groups of poeple. In *BMVC*, 2009. 2