Supplemental material: Temporal Residual Networks for Dynamic Scene Recognition

Christoph Feichtenhofer
Graz University of Technology
feichtenhofer@tugraz.at

Axel Pinz
Graz University of Technology
axel.pinz@tugraz.at

Richard P. Wildes
York University, Toronto
wildes@cse.yorku.ca

1. Baseline comparison algorithms

Slow feature analysis (SFA) approaches analyze temporal data to extract features that vary most slowly over time, taking those to be most indicative of the stable properties of the input [32]. The approach has been applied to dynamic scene recognition [28] by extracting features from filter responses that are reputed to model primate V1 cortical operations, as they result from local maxima of spatially oriented, multiscale Gabor filters [25]. The slowest varying features among those are identified by taking their temporal derivatives and subsequently are encoded via soft assignment with respect to a dictionary built with unsupervised sampling. Following encoding, the features are pooled into a feature vector via application of max-pooling to the entire video in spatial pyramid regions [18].

Bags of spacetime energies (BoSE) [11] is the penultimate version of spatiotemporal energy approaches applied to dynamic scenes [8, 10]. The approach extracted dense measurements of spatiotemporal energy across a range of scales and orientations as well as CIE-LUV colour measurements. Here, the spatiotemporal features are augmented with dense SIFT measurements [19] to more finely capture spatial orientation. The descriptors are encoded by Improved Fisher Vector (IFV) [21, 22] encoding with a visual word dictionary represented by a Gaussian Mixture Model (diagonal covariance) with 64 centres. We average the frame-level BoSE encodings over a video which simplifies the the temporal slice-based SVM prediction of the original BoSE system [11]. The simplification is employed for equality in comparison to other baselines which also train a single one-vs-rest SVMs for video classification.

Trajectory features (IDT) have been investigated with respect to a variety of video understanding tasks, e.g., [20, 23, 24, 30]. Curiously, it appears that they have not previously been applied to dynamic scene recognition. Recently, however, they have provided the basis for a number of outstanding approaches to action recognition as instantiated in improved Dense Trajectories (IDTs) [31]; therefore, it is of interest to evaluate their performance on scene recognition, as follows. Trajectory features are extracted across stabilized video sequences by concatenating a series of optical flow vectors for densely extracted interest points. Feature descriptors are aggregated across each trajectory in terms of trajectory shape [30], HOG [5], HOF [17] and MBH [6] measurements. Following extraction, the features are encoded using (improved) Fisher Vectors (FVs) [22] with dictionary represented by a Gaussian Mixture model (diagonal covariance) having 256 centres. Before training the GMM, all features are augmented with their normalized (x, y) image coordinates as an efficient way to capture location information. Details of extraction of the trajectories, their descriptors and encoding are exactly as in their original application to action recognition [31]. All DT and IDT parameters are used as in [30, 31] and their publicly available code is used to extract the descriptors.

Spatial convolutional network (S-CNN) features [3] are generated from the last convolutional layer of a VGG-16 network [27]. The model is pre-trained on ImageNet [7]. It has been shown that the features from such pretrained CNNs are transferable to many other vision domains [2, 3, 9, 13]. This approach derives its features from the last convolutional layer of a VGG-16, which uses features from the last conv-layer of VGG-16. The resulting 512-dimensional features are encoded using (improved) Fisher Vectors (FVs) [22] with dictionary represented by a Gaussian Mixture model (diagonal covariance) having 64 centres. Features from a single video are extracted with a stride of 16 frames. Before encoding the features are augmented with their normalized (x, y) image coordinates, as with the above IDT approach.

Temporal convolutional network (T-CNN) uses a stack of 10 optical flow frames as input, with optical flow extracted by a standard algorithm [1] and is first pre-trained on the UCF101 action recognition dataset [16]. The final model is a CNN-M-2048 network [2]. (In our preliminary evaluation with this implementation, a recognition accuracy of 82.6% on UCF101 (split 1) was achieved, which compares favourably to the 81.2% reported originally [26].) The same IFV encoding procedure as used for the spatial CNN above is employed, since this approach is common practice in state-of-the-art video action recognition [12] and provided slightly better performance than using the output of the last fully connected layer.

Spatiotemporal convolutional network (C3D) provides a spatiotemporal analogue to the spatial S-CNN. As a generalization of spatial convolutional neural networks, 3D spatiotemporal networks working over image spacetime, (x, y, t), have potential to more directly capture temporal aspects of the data even while maintaining spatial information. Various previous efforts have been mounted to consider this potential [14, 15, 29]. Here, C3D is considered, as it has previously been applied to dynamic scene recognition [29]. Features are extracted by applying the C3D network model, pretrained on the Sports-1M dataset [15], densely to 16-frame snippets of the input video. As in [29], the fully connected layer 6 outputs of each 16-frame clip are averaged across the video into a 4096-dimensional descriptor.

Classification is performed as in the original approaches [3, 11, 26, 28, 29, 31], with a linear SVM [4]. Before training, the descriptors are L2-normalized. All feature vectors extracted from the training set are used to train one-vs-rest linear SVM classifiers. The SVM's regularization loss trade-off parameter is set to $C = 100$. During classification, each feature type is classified by its one-vs-rest SVM to yield SVM scores for a test video and an overall classification of the video according to the maximum score.

2. Video samples of the YUP++ dataset

The videos[1], `static_camera_samples.avi` and `moving_camera_samples.avi`, show examples of the static and moving camera subsets. The codec used is H264 - MPEG-4 AVC. (High compression rates are applied in the supplemental material for the sake of constraints on submission size.) Each video shows examples for all 20 classes, ordered alphabetically from left-to-right, top-to-bottom: Beach, BuildingCollapse, Elevator, Escalator, FallingTrees, Fireworks, ForestFire, Fountain, Highway, LightningStorm, Marathon, Ocean, Railway, RushingRiver, SkyClouds, Snowing, Street, Waterfall, WavingFlags, and WindmillFarm.

References

[1] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *Proc. ECCV*, 2004. 1

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proc. BMVC*, 2014. 1

[3] M. Cimpoi, S. Maji, and A. Vedaldi. Deep filter banks for texture recognition and segmentation. In *Proc. CVPR*, 2015. 1, 2

[1] http://vision.eecs.yorku.ca/research/dynamic-scenes/

[4] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. 2

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005. 1

[6] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *Proc. ECCV*, 2006. 1

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009. 1

[8] K. Derpanis, M. Lecce, K. Daniilidis, and R. P. Wildes. Dynamic scene understanding: The role of orientation features in space and time in scene classification. In *Proc. CVPR*, 2012. 1

[9] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. Platt, et al. From captions to visual concepts and back. In *Proc. CVPR*, 2014. 1

[10] C. Feichtenhofer, A. Pinz, and R. P. Wildes. Spacetime forests with complementary features for dynamic scene recognition. In *Proc. BMVC*, 2013. 1

[11] C. Feichtenhofer, A. Pinz, and R. P. Wildes. Bags of spacetime energies for dynamic scene recognition. In *Proc. CVPR*, 2014. 1, 2

[12] A. Gorban, H. Indrees, Y. Jiang, A. R. Zamir, I. Laptev, M. Shah, and R. Sukthankar. Thumos challenge: Action recognition with a large number of classes. `http://wwwthumos.info/`, 2015. 2

[13] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik. Learning rich features from RGB-D images for object detection and segmentation. In *Proc. ECCV*. 2014. 1

[14] S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. *PAMI*, 35(1):221–231, 2013. 2

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proc. CVPR*, 2014. 2

[16] A. R. Z. Khurram Soomro and M. Shah. Ucf101: A dataset of 101 human actions calsses from videos in the wild. Technical Report CRCV-TR-12-01, UCF Center for Research in Computer Vision, 2012. 1

[17] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. CVPR*, 2008. 1

[18] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. CVPR*, 2006. 1

[19] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 1

[20] B. Moore, S. Ali, R. Mehran, and M. Shah. Visual crowd surveillance through a hydrodynamics lens. *Commun. ACM*, 54(12):64–73, 2011. 1

[21] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *Proc. CVPR*, 2007. 1

[22] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *Proc. ECCV*, 2010. 1

[23] M. Raptis and S. Soatto. Tracklet descriptors for action modeling and video analysis. In *Proc. ECCV*, 2010. 1

[24] P. Sand and S. Teller. Particle video: Long-range motion

estimation using point trajectories. In *Proc. CVPR*, 2006. 1
[25] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *PAMI*, 29(3):411–426, 2007. 1
[26] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proc. NIPS*, 2014. 1, 2
[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR*, 2014. 1
[28] C. Theriault, N. Thome, and M. Cord. Dynamic scene classification: Learning motion descriptors with slow features analysis. In *Proc. CVPR*, 2013. 1, 2
[29] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proc. ICCV*, 2015. 2
[30] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, pages 1–20, 2013. 1
[31] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proc. ICCV*, 2013. 1, 2
[32] L. Wiskott and T. J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002. 1