University of Amsterdam at THUMOS 2015

Mihir Jain[†], Jan C. van Gemert[†], Pascal Mettes[†], Cees G. M. Snoek[†⋆]

[†] ISLA, IvI, University of Amsterdam, The Netherlands
[⋆] Qualcomm Research Netherlands, The Netherlands

Abstract

This notebook paper describes our approach for the action classification task of the THUMOS 2015 benchmark challenge. We use two types of representations to capture motion and appearance. For a local motion description we employ HOG, HOF and MBH features, computed along the improved dense trajectories. The motion features are encoded into a fixed-length representation using Fisher vectors. For the appearance features, we employ a pre-trained GoogLeNet convolutional network on video frames. VLAD is used to encode the appearance features into a fixed-length representation. All actions are classified with a one-vs-rest linear SVM.

1. Classification framework

An overview of our classification pipeline is shown in Figure 1. For each video, we combine two representations, one based on local motion descriptors and the other as features from deep convolutional neural network. Motion descriptors are computed along the improved dense trajectories [7]. The CNN features are extracted per frame and aggregated to get appearance representation for the video. Finally, linear Support Vector Machine is used for classification.

1.1. Representing motion

We capture motion information by several local descriptors (HOG, HOF, and MBH) computed along the improved trajectories. Improved trajectories takes into account camera motion compensation, which is shown to be critical in action recognition [2]. To encode the local descriptors, we use Fisher vectors [5]. We first apply PCA on the motion descriptors and reduce the dimensionality by a factor of two. Then a large set of descriptors are selected at random from the UCF101 videos to estimate GMM with $K = 128$ Gaussians. This results in a $2DK$-dimensional representation, with D the number of dimensions kept after applying PCA. We perform power and ℓ_2 normalization, as advocated in [5].

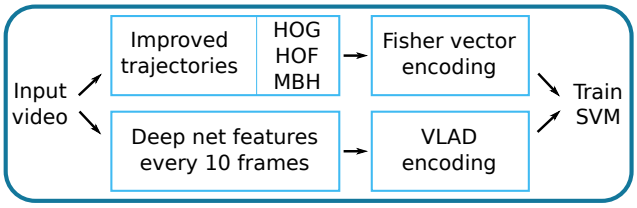

Figure 1. Overview of our classification framework, based on motion and deep net appearance representations.

1.2. Representing appearance

For the appearance representation, we employ features extracted from a deep convolutional neural network. We employ the GoogLeNet architecture [6], pre-trained on 15k ImageNet concepts [1, 4]. We compute features for each 10th frame in each video. For a given frame, we extract 1024-dimensional representation from the fully-connected layer of GoogLetNet. In addition to this, we also use 15k dimensional representation of object responses as we did in our winning approach [3] for THUMOS 2014.

Here, we extend our approach from last year by employing a state-of-the-art convolutional network for improved feature representations and by encoding the frame representations using VLAD, rather than average pooling. We apply PCA without dimensionality reduction on the frame representations, and learn a codebook on frames from UCF101.

1.3. Classifying the actions

For classifying the actions, we use a one-vs-rest linear Support Vector Machine. The C parameter is set to 1 for each of the 101 action categories. For combining the motion and appearance representations, we compute separate kernels for each representations and then sum the kernel matrices.

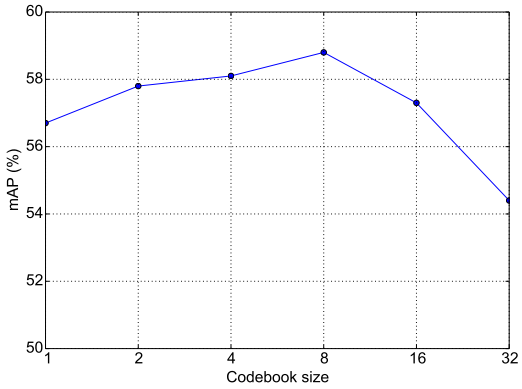

Figure 2. Performance on validation set using the deep net features with VLAD encoding for various codebook sizes.

Representation	mAP val	mAP test
Motion		
HOG+HOF+MBH (FV)	52.2%	-
Appearance		
Avg pooling	53.4%	-
VLAD	58.8%	-
Combined		
Motion + Avg pooling	63.1%	-
Motion + VLAD	**66.9%**	**68.0%**
Objects		
Avg pooling	58.2%	-
Motion + Avg pooling	**67.1%**	-

Table 1. Results (mAP %) on the THUMOS 2015 validation and test set for different representations and combinations.

2. Experiments

We report experiments performed on the THUMOS 2015 Validation set (consisting of 2,104 untrimmed videos) and Test set (5,613 untrimmed videos). For the training set, we use the 13,320 videos from UCF101.

Experimental setup. The GMM is estimated from the UCF101 dataset; the provided 'Background' set is ignored. For validation and tuning parameters, we train on UCF101 and test on validation set. For the test set, we use all the videos in both UCF101 and the Validation set to train the SVM classifiers.

Codebook size for VLAD. Figure 2 shows the mAPs on the validation set for codebook sizes (K) varying from 1 to 32. Interestingly, a small codebook yields better performance than a large codebook, which we attribute to the relatively limited variance within the actions themselves. The best mAP is obtained for $K = 8$, so we use this codebook size for the test set.

Results. The results on the validation set are shown in the middle column of Table 1. We note that appearance-based results compare favourably to the motion-based result, and combining accuracy further improves. These results can be attributed to the improved CNN features and also conform with our findings in [4].

We note that encoding 1024-dimensional CNN features (referred as 'Appearance' in Table 1) with VLAD clearly outperforms average pooling by 5.4% of mAP. Upon combining with motion also VLAD beats average pooling with mAP of 66.9% over 63.1%. We submitted only one run for the test set, i.e., 'Motion + Appearance VLAD', which resulted in mAP of 68.0%.

Although we did not use a 15k-dimensional semantic representation [4] for the submission, here we also report the results of this representation on the validation set. Results are shown at the bottom of the table, under heading 'Objects'. With average pooling object representation achieves 58.8% and when combined with motion shoots up to 67.1%. This again is in line with our findings in [4] that incorporating object appearance information helps for recognizing actions.

References

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

[2] M. Jain, H. Jégou, and P. Bouthemy. Better exploiting motion for better action recognition. In *CVPR*, 2013.

[3] M. Jain, J. C. van Gemert, and C. G. M. Snoek. University of amsterdam at thumos challenge 2014. *ECCV THUMOS Challenge*, 2014.

[4] M. Jain, J. C. van Gemert, and C. G. M. Snoek. What do 15,000 object categories tell us about classifying and localizing actions? In *CVPR*, 2015.

[5] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 105(3), 2013.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[7] H. Wang and C. Schmid. Action Recognition with Improved Trajectories. In *ICCV*, Dec. 2013.