University of Amsterdam at THUMOS Challenge 2014

Jain, M.; van Gemert, J.; Snoek, C.G.M.

Publication date
2014
Document Version
Final published version
Published in
THUMOS Challenge 2014: notebook papers

University of Amsterdam at THUMOS Challenge 2014

Mihir Jain, Jan van Gemert and Cees G. M. Snoek

ISLA, Informatics Institute, University of Amsterdam, The Netherlands

Abstract. This notebook paper describes our approach for the action classification task of the THUMOS Challenge 2014. We investigate and exploit the action-object relationship by capturing both motion and related objects. As local descriptors we use HOG, HOF and MBH computed along the improved dense trajectories. For video encoding we rely on Fisher vector. In addition, we employ deep net features learned from object attributes to capture action context. All actions are classified with a one-versus-rest linear SVM.

Keywords: Action recognition, motion trajectories, deep net features, object attributes

1 Classification framework

Our action classification framework consists of two main components: video representation and classification. The video representation is summarized in figure 1. Many of the action classes in the given dataset have related objects such as 'Billiards', 'PlayingTabla', 'RockClimbingIndoor' etc. Therefore, along with motion we also capture the appearance information of object attributes. In the following subsections, we describe these two types of representations and their classification.

1.1 Motion based representation

We capture motion information by several local descriptors (HOG, HOF and MBH) computed along the improved trajectories [4]. Improved trajectories is one of the recently proposed approaches that takes into account camera motion compensation, which is shown to be critical in action recognition [1, 4]. To encode the local descriptors, we use Fisher vector. We first apply PCA on these local descriptors and reduce the dimensionality by a factor of two. Then 256,000 descriptors are selected at random from the 'UCF101' set and the 'Background' set to estimate GMM with K (=256) Gaussians. Each video is then represented by $2DK$ dimensional Fisher vector, where D is the dimension of descriptors after PCA. Finally, we apply power and L2 normalization to the Fisher vector as done in [2].

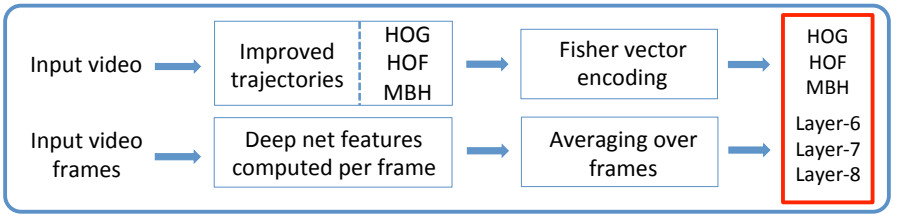

Fig. 1. Motion and appearance representations for action recognition in videos.

1.2 Appearance based representation

For appearance representation, we employ deep net features. We use the output of the last three fully connected layers of an 8-layer convolutional neural network [3]. The input is raw pixel data, the output are 15K object scores. For the final video representation, we average each of these output vectors across the frames. We refer to these three representations as: Layer-6, Layer-7 and Layer-8.

1.3 Classification and merging representations

In all the experiments, we use the SVM with linear kernel for classification. We set C=100 for the SVM and learn 101 one-versus-rest classifiers. To combine two or more of the above described six representations we simply sum the kernels.

2 Experiments

We experimented with different combinations of the six representations. In order to find best combinations, we evaluated them on the provided 'Validation' set of 1010 videos.

Setup for Validation set. Though we learn GMM using Background set also but for training we only used 13320 videos from UCF101. The classifiers were applied on Validation set and the results are reported in table 1. Among the individual representation, not surprisingly, MBH achieves the best mean average precision (mAP). But all three appearance based representations do as well as or even better than HOF is a very interesting result. Without any motion information Layer-8 is just 1.8% behind MBH.

All three motion descriptors when combined leads to 56.9%, but the gains obtained by adding appearance representations are huge. This significant improvement (mAP up to 66.8%) confirms our proposition that interdependence between action and object attributes is critically important for recognizing many actions. Further, this interdependence depends on the action category so when the best combination is used for each class the mAP boosts to 70.8%.

Setup for Test set. All videos in UCF101 and Validation sets are used as training videos. We use the best combinations obtained on Validation set for the Test set. These are the last five in Table 1, which are submitted as our five runs.

Table 1. Results as mAP on Validation and Test sets. [Motion:=HOG+HOF+MBH]

Combinations	mAP on Validation set	mAP on Test set
HOG	43.7%	–
HOF	47.4%	–
MBH	**51.5%**	–
Layer-6	47.6%	–
Layer-7	47.3%	–
Layer-8	49.7%	–
Motion	56.9%	–
Motion+Layer-6	63.3%	–
Motion+Layer-7	62.8%	–
Motion+Layer-8	**66.8%**	70.75%
Motion+Layer-6,8	66.7%	**71.00%**
Motion+Layer-7,8	66.1%	70.75%
Motion+Layer-6,7,8	65.2%	70.76%
Best per class	**70.8%**	69.32%

Acknowledgements

This research is supported by the STW STORY project, the Dutch national program COMMIT, and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20067. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

1. Jain, M., Jégou, H., Bouthemy, P.: Better exploiting motion for better action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (Jun 2013)
2. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Proceedings of the European Conference on Computer Vision (Sep 2010)
3. van de Sande, K.E.A., Fontijne, D., Snoek, C.G.M., Stokman, H., Smeulders, A.W.M.: University of Amsterdam and Euvision Technologies at ILSVRC2014. In: ILSVRC (2014)
4. Wang, H., Schmid, C.: Action Recognition with Improved Trajectories. In: Proceedings of the IEEE International Conference on Computer Vision (Dec 2013)