Extreme Learning Machine for Large-Scale Action Recognition

Gül Varol and Albert Ali Salah

Department of Computer Engineering, Boğaziçi University, Turkey

Abstract. In this paper, we describe the method we applied for the action recognition task on the THUMOS 2014 challenge dataset. We study human action recognition in RGB videos through low-level features by focusing on improved trajectory features that are densely extracted from the spatio-temporal volume. We represent each video with Fisher vector encoding and additional mid-level feautures. Finally, we use Extreme Learning Machines for classification. We achieve 62.27% mean average precision on the validation set.

Keywords: Action recognition, Extreme Learning Machine, Fisher vector

1 Introduction

Human action recognition in videos has attracted increasing interest because of the wide application area in automatic analysis of video. In spite of the great effort that computer vision scientists invested in this task, robust inference on action from image sequences remains a challenge. Human action is complex due to a variety of reasons such as viewpoint variance, background complexity, camera motion, occlusion and performance difference among different people. Besides the difficulty of action recognition, there exists the action spotting problem which is less addressed by the research community. Most of the standard human action datasets involve temporally cropped videos of a single action. THUMOS 2014 introduces challenging videos since significant part of the test videos may not include any particular action and multiple instances may occur within the same video [5]. We use UCF101 dataset [15] of 101 action categories for the training and report results on the validation set since the test labels are not revealed.

A recent action spotting method is introduced by Derpanis et al [1]. The authors propose to use a descriptor which is computed directly from raw image intensity data and to use action templates searched across video sequences. Another method which addresses spotting together with recognition is presented in [13]. Here, the authors introduce a new high-level video representation which they call action bank, inspired from object bank approach.

Spatio-temporal features are widely used to classify human actions. Researchers focus on how to treat temporal motion information and spatial gradient information to extract informative and robust features. Laptev et al. introduced

space-time interest points (STIP) which can be seen as an adaptation of scale-invariant feature transform (SIFT) [8] computed on still images to image sequences [7]. Pooling spatio-temporal descriptors with Bag of Features (BOF) technique is a widely used approach for action recognition [19] [14]. Although this technique had been previously adopted as the main paradigm for representing a video [6], a recent study showed that a better encoding technique, namely Fisher vector representation significantly increases recognition performance [10].

As the classification algorithm for the aforementioned features, most studies use the popular Support Vector Machine (SVM) including the winner of THUMOS 2013 competition [10] [17] [12] [18]. In our study, we prefer Extreme Learning Machine (ELM) which is much faster than SVM.

Our feature extraction and encoding pipeline is illustrated in Figure 1. We extract improved trajectory features densely from a subset of the training videos using the software provided by Wang et al. [17]. These features include four different aspects of video representation, namely motion boundary histogram (MBH), histogram of optical flow (HOF), histogram of oriented gradients (HOG) and trajectories. We then apply Fisher vector encoding for a global representation and concatenate some additional mid-level features for higher recognition. The final feature vector is assigned into one of the 101 action classes using ELM with linear kernel.

The rest of this paper is organized as follows: The feature extraction phase is detailed in Section 2. We briefly present ELM in Section 3 and finally report our results in Section 4.

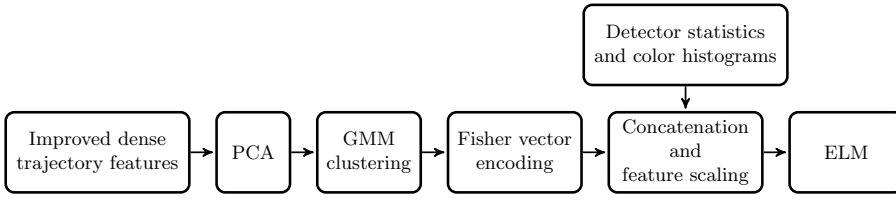

Fig. 1. Pipeline

2 Feature extraction

Improved trajectories. The improvement over the dense trajectory features [16] is achieved by compensating the camera motion effect with homography matrix computation between consecutive frames. These warped features proposed by Wang et al. [17] are shown to perform the best among various local descriptors on several benchmark action recognition datasets. We extract the features with the default parameters and obtain 96-dimensional HOG, 108-dimensional HOF and 192-dimensional MBH features for each trajectory.

Fisher vector encoding. To obtain a global descriptor for a video, we employ Fisher vector encoding proposed by Perronnin et al. [11]. This technique serves the purpose of summarizing a number of local descriptors in a feature vector. The framework is analogous to the traditional BOF representation. Instead of k-means clustering, Gaussian mixture models (GMM) are used for quantizing local descriptors. Thus, a generative dictionary is built. BOF representation only considers 0-th order statistics, i.e. word frequencies. The Fisher vector encodes both first and second order statistics by storing the difference between pooled local features and dictionary items. The final dimension of the Fisher vector becomes $2DK$, where D is the dimension of local features and K is the number of components in the GMM with diagonal covariance.

In order to decrease the size of the Fisher vector, we apply PCA prior to building the dictionary. We construct a separate GMM and Fisher vector for each modality MBH, HOG and HOF, and concatenate the vectors for the final representation. We reduce each descriptor to 64 dimensions by PCA and learn a GMM with 64 components for each aspect. We normalize the Fisher vectors as in [12] by square-rooting (power normalization) and ℓ_2 normalization so that linear classifiers can be used successfully.

Mid-level features. With the aim of supplying higher level information to the features, we extract additional descriptors. We apply seven cascade object detectors on ten equidistant frames of each video and summarized the detections with their statistics. Those detectors are the pre-trained face, upper body, eye pair, left eye, right eye, profile and people detectors. We use minimum, maximum and mean area of the detected regions throughout the video. Additionally, we compute the frequency of the detections by dividing to the number of frames. We normalize each detection area by dividing it to the frame area. Thus, we obtain 4 features per detector, which makes a total of 28-dimensional vector.

Since the class attributes indicate that some action classes involve a specific color, we make use of color histograms for additional features. For each video, we compute 48-bin normalized color histograms from ten frames on both RGB and HSV color spaces. The histograms are then averaged over the frames to get a global feature vector of size 288 ($2 \times 48 \times 3$).

The final feature vector for a video becomes the concatenation of MBH, HOG and HOF Fisher vectors; the detector statistics (DS) and two color histograms (RGBH, HSVH). We then subtract the training mean from each vector for classification.

3 Extreme Learning Machine

Extreme Learning Machine is proposed by Huang et al. [3] as an extremely fast alternative to other conventional popular learning algorithms. The proposed algorithm works for the generalized single-hidden-layer feed-forward networks, but the difference is that the hidden layer in ELM need not be tuned [2]. In the literature, Minhas et al. [9] and Iosifidis et al. [4] used ELM for action

classification from visual vocabularies, but the usage of ELM in this area is still rare.

One of the most common learning algorithms for action recognition is SVM. SVM is a powerful tool for the classification task; however, it requires iterative learning. Therefore the training might be slow depending on the feature dimensionality and the number of instances. Hyperparameter optimization, while using SVM, plays a key role in obtaining high performance. Thus, searching over a parameter grid for optimization purposes might take a long time.

In our method, we use linear kernel ELM and keep the regularization parameter as 0.85 throughout our experiments after a parameter search on the validation set.

4 Experimental results

We performed several experiments to evaluate the effect of different steps of the pipeline. First, we examine the contribution of the individual aspects of the improved trajectory features MBH, HOG and HOF. The results are presented in Table 1. The best performance of 61.02% mean average precision is obtained with the combination of the three aspects. As expected, the combination of a motion-based feature (MBH or HOF) and a spatio-based feature (HOG) yields higher performance than joining two motion-based features. Moreover, MBH is the most successful individual feature type in discriminating the action classes on this dataset.

Table 1. Mean average precision values for various feature sets

Feature set	mAP(%)
HOG	50.18
HOF	52.77
MBH	55.72
MBH+HOF	56.78
HOG+HOF	59.26
MBH+HOG	60.61
MBH+HOG+HOF	61.02

Using all the features (MBH, HOG, HOF, DS, RGBH and HSVH) concatenated, we get the best result on the validation set. The confusion matrix for the classification with ELM is presented in Figure 2.

We further compared ELM and SVM in terms of both training time and performance. The results are summarized in Table 2. The experiments are carried out on a machine with two 2.26 GHz quad-core Intel Xeon processor and 32 GB of RAM. ELM yields a mean average precision of 62.27% with 92 seconds of

training time whereas SVM performs with 43.94% performance with 14 hours of training time. ELM outperforms SVM significantly in both aspects in this task.

Table 2. ELM and SVM comparison in terms of time and performance with the best feature combination

Algorithm	mAP(%)	Training time (sec)	Testing time (sec)
SVM	43.94	51290	885
ELM	62.27	92	11

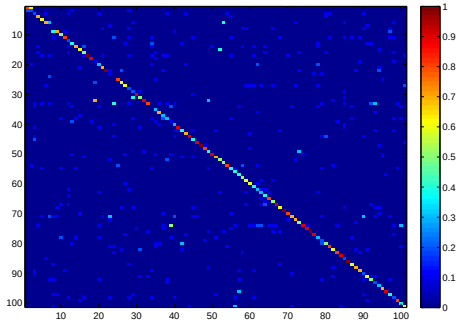

Fig. 2. Confusion matrix on the validation set

Finally, we made experiments with the BOF features and compared the results with Fisher vector representation. With the BOF encoding of 4000 visual words, a mean average precision of 40.64% is achieved with RBF kernel ELM, since the usage of linear kernel brings much worse performance with histogram-based features. Fisher vector encoding outperforms BOF representation with 20.38% increase in the performance on the validation set.

In this study, we highlight an alternative to the often used SVM for action classification task. ELM enables the opportunity to evaluate a variety of features by saving a significant time for both training and testing. We achieve 62.27% recognition performance with this method on 1010 temporally untrimmed videos of 101 different action classes.

References

1. Derpanis, K., Sizintsev, M., Cannons, K., Wildes, R.: Efficient action spotting based on a spacetime oriented structure representation. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. pp. 1990–1997 (June 2010)

2. Huang, G.B., Zhou, H., Ding, X., Zhang, R.: Extreme learning machine for regression and multiclass classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B 42(2), 513–529 (2012)
3. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: Theory and applications. Neurocomputing 70(13), 489 – 501 (2006), neural Networks Selected Papers from the 7th Brazilian Symposium on Neural Networks (SBRN '04) 7th Brazilian Symposium on Neural Networks
4. Iosifidis, A., Tefas, A., Pitas, I.: Regularized extreme learning machine for multi-view semi-supervised action recognition. Neurocomputing 145(0), 250 – 262 (2014)
5. Jiang, Y.G., Liu, J., Roshan Zamir, A., Toderici, G., Laptev, I., Shah, M., Sukthankar, R.: THUMOS challenge: Action recognition with a large number of classes. `http://crcv.ucf.edu/THUMOS14/` (2014)
6. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. pp. 1–8 (June 2008)
7. Laptev, I.: On space-time interest points. Int. J. Comput. Vision 64(2-3), 107–123 (Sep 2005)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
9. Minhas, R., Baradarani, A., Seifzadeh, S., Wu, Q.J.: Human action recognition using extreme learning machine based on visual vocabularies. Neurocomputing 73(1012), 1906 – 1917 (2010), subspace Learning / Selected papers from the European Symposium on Time Series Prediction
10. Oneata, D., Verbeek, J., Schmid, C.: Action and Event Recognition with Fisher Vectors on a Compact Feature Set. In: ICCV 2013 - IEEE Intenational Conference on Computer Vision. pp. 1817–1824. IEEE, Sydney, Australia (Dec 2013)
11. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on. pp. 1–8 (June 2007)
12. Perronnin, F., Snchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: IN: ECCV (2010)
13. Sadanand, S., Corso, J.J.: Action bank: A high-level representation of activity in video. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2012)
14. Shugao Ma, Jianming Zhang, N.I.C., Sclaroff, S.: Action recognition and localization by hierarchical space-time segments. In: Proc. of the IEEE International Conference on Computer Vision (ICCV) (2013)
15. Soomro, K., Roshan Zamir, A., Shah, M.: UCF101: A dataset of 101 human actions classes from videos in the wild. In: CRCV-TR-12-01 (2012)
16. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action Recognition by Dense Trajectories. In: IEEE Conference on Computer Vision & Pattern Recognition. pp. 3169–3176. Colorado Springs, United States (Jun 2011)
17. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: IEEE International Conference on Computer Vision. Sydney, Australia (2013)
18. Wang, H., Schmid, C.: Lear-inria submission for the thumos workshop. In: ICCV Workshop on Action Recognition with a Large Number of Classes (2013)
19. Wang, H., Ullah, M.M., Klser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: University of Central Florida, U.S.A (2009)