NII, Japan at the first THUMOS Workshop 2013

Sang Phan, Duy-Dinh Le, Shin'ichi Satoh
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan 101-8430
{plsang,ledduy,satoh}@nii.ac.jp

Abstract

We submit the results of our system to the THUMOS'13 Challenge on Recognition Task. We apply a multi-modal approach to recognize actions on UCF101 dataset. These features include Dense Trajectories (motion features), SIFT (image features) and MFCC (audio features). Moreover, we also use the Fisher vector encoding, which is a state-of-the-art feature representation on popular image classification and action recognition datasets.

1. Classification Framework

The action classification framework is shown in Figure 1. Basically, it consists of following steps: feature extraction, feature encoding and action classification.

1.1. Feature Extraction

We use feature from different modalities to model action in UCF101: motion features, still image features, and audio features. We use the Dense Trajectories [7] with Motion Boundary Histogram (MBH) descriptor for capturing moving patterns in video. This is a state of the art motion feature that is suitable for realistic videos because it can suppress camera motions. We use the standard SIFT [4] with Hessian Laplace detector for extracting still image feature. Finally, the standard MFCC [3] feature is used to capture audio features. Audio segment unit is set to 25 ms and the overlapping parameter is set to 10 ms. The 13d MFCCs along with each first and second derivatives are used for representing each audio segment.

1.2. Fisher Vector Encoding

Fisher Vector [5] representation is a newly developed feature encoding technique where the mean and variance of local descriptors that belong to each cluster are also considered. Therefore, Fisher vector encodes more information than the bag-of-words feature encoding. Following the standard implementation of Fisher vector, we use the codebook size of 256 clusters which are generated using the Gaussian Mixture Model (GMM). We further improve the expressiveness of Fisher vector by applying PCA for reducing feature dimension, i.e from 128d to 80d for SIFT and from 192d to 128d for MBH descriptor.

1.3. Action Classification

We use the popular Support Vector Machine (SVM) for classification. All the positive videos are considered as positive samples and the remaining videos are considered as negative samples. LibSVM [1] is incorporated to our system for learning steps because it is a standard implementation for SVM. We use linear kernel for training features encoded by Fisher vector. Furthermore, we also utilized the pre-computed kernel technique to reduce the training time. This technique is especially useful when the number of action classes are large.

2. Experiments

We report following experiments on the UCF101 dataset [6], based on the train/test settings provided by the workshop organizers [2]. At first, we compare the Fisher vector representation with the traditional bag-of-words representation. This experiment is reported for motion features only. In the later part, we report the results of combining multiple features, with consideration of using audio classes.

2.1. Fisher vector vs. Bag-of-words representation

For bag-of-words encoding method, we use the codebook size of 4,000 codewords trained by K-means algorithm. The soft assignment technique is also employed to reduce the quantization errors. This is the default parameters in our framework. For Fisher vector encoding, we consider applying Principal Component Analysis (PCA) to reduce the feature dimension. The comparison is shown in Figure 2. With current setting, we observed that Fisher vector representation is more robust than the bag-of-words one. Moreover, applying PCA even improves the performance more. For later experiments, we only use Fisher Vector en-

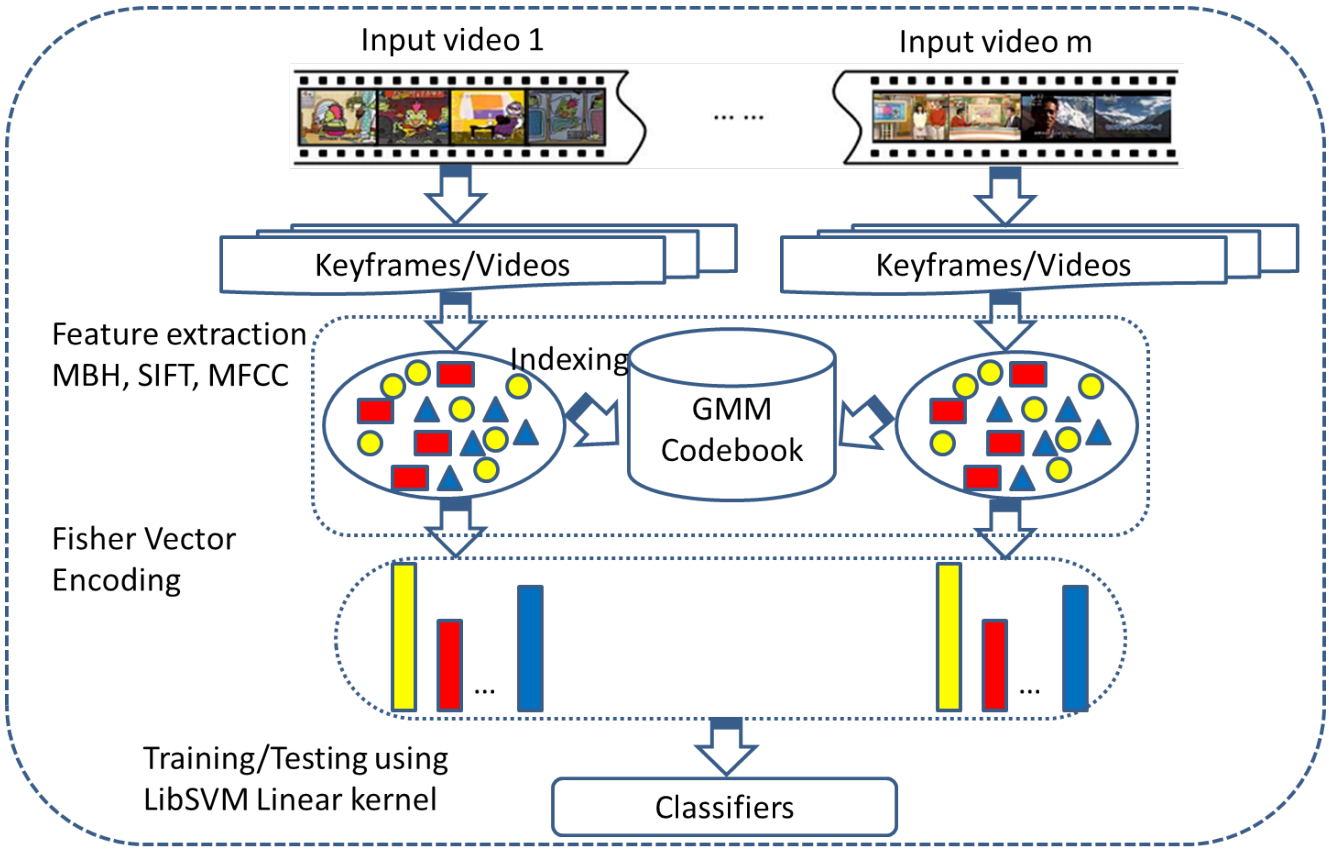


Figure 1. Action Classification Framework

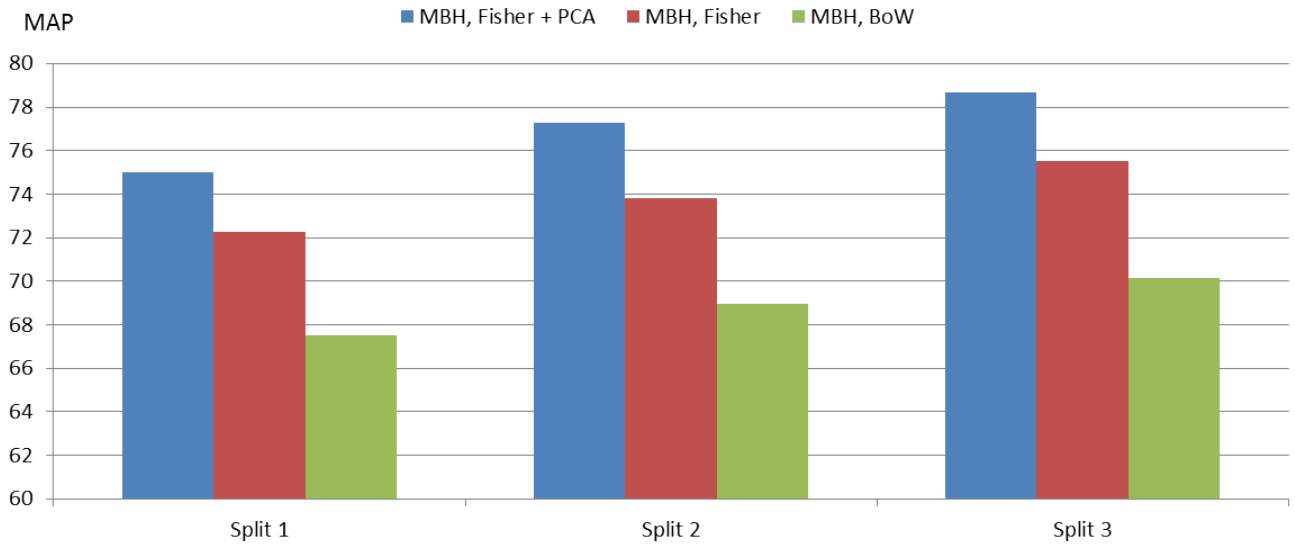


Figure 2. Fisher vector vs. Bag-of-words representation

coding with PCA. Except for audio feature, we do not apply PCA because the dimension of audio feature is rather small.

2.2. Combined runs

We investigate different combinations of motion features, image features and audio features. These features

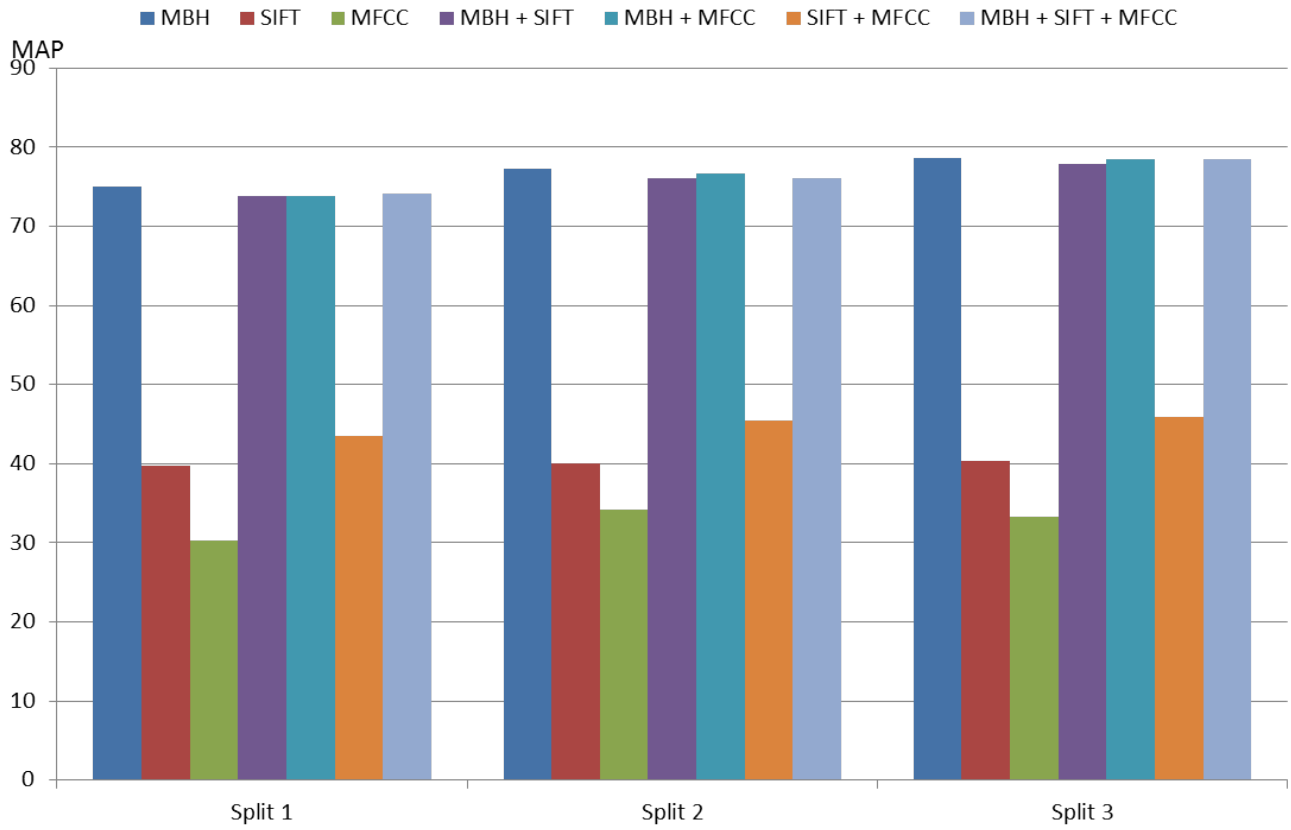


Figure 3. Performance of each single feature and their combinations

are combined using the late fusion technique with equal weights for all runs. Note that the for audio features, we only tested on 51 audio classes and their combination with other features are also done on these classes. The performance of each runs and their combinations are shown in Figure 3.

We observe that the performance of motion features are rather good. Interestingly, it is not complementary with image features and audio features. The reason can be due to the low performance of image features and audio features. However, image features and audio features are complementary to each other. These results show that the motion cues are very important to discriminate between actions, while image and audio cues are more less important. The main reason is some actions may be recorded in the same scenes and/or involved in same objects, which in turn confuse the classifiers.

References

[1] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[2] Y.-G. Jiang, J. Liu, A. Roshan Zamir, I. Laptev, M. Piccardi, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. `http://crcv.ucf.edu/ICCV13-Action-Workshop/`, 2013.

[3] C.-H. Lee, F. Soong, and B.-H. Juang. A segment model based approach to speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing, 1988. ICASSP-88.*, pages 501 –541 vol.1, 1988.

[4] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[5] F. Perronnin, J. Snchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *IN: ECCV*, 2010.

[6] K. Soomro, A. Roshan Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. In *CRCV-TR-12-01*, 2012.

[7] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action Recognition by Dense Trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States, June 2011.