ZJUDCD Submission at THUMOS Challenge 2015

Ke Ning Fei Wu
College of Computer Science, Zhejiang University, China
{ningke1688, wufei}@zju.edu.cn

Abstract

We described our system of THUMOS Challenge 2015 in this paper. In general, this paper utilized both hand-crafted temporal features(i.e., improved dense trajectories) and the learning features (via convolutional neural network) for video action detection, which shows a great improvement in experiments. Specially, in order to extract temporal features, HOG, HOF and MBH in terms of improved dense trajectories and Fisher vector encoding are employed. At the same time, CNN-based features are exploited as an auxiliary feature to boost the performance of video action recognition. In the end, the one-vs-rest linear SVM is conducted as a classifier in our system.

1. Introduction

In this paper, we describe our system for THUMOS challenge 2015[2] recognition task. Our system mainly consists of two components: 1) the extraction of hand-crafted temporal features and the extraction of learning features, 2) the fusion of hand-crafted features and the learning features. We found that both hand-crafted features and learning features have different intrinsic discriminative power to characterize video actions.

In the rest of this paper, section 2 describes our system for the recognition task. The section 3 presents experimental results.

2. System Description

2.1. Hand-crafted temporal features

We extract the hand-crafted temporal features from videos according to the improved dense trajectories[7], and encode them with Fisher vector[5]. To generate Fisher vector for motion representation, we use 256-component GMMs as a codebook. Before applying Fisher vector encoding to the local descriptors, we use principle component analysis (PCA) to reduce the dimensionality to half, and augment features as in [4] with their horizontal and vertical coordinates. We also applied T2+H3 spatial pyramids(two temporal parts and three horizontal parts), SSR (signed square root) and L2 normalization for post-processing. We run iDT+FV multiple times, and late fuse the decision scores, which can bring a little performance improvement.

2.2. Learning features

For the learning features, we extract video frame features every 5 frames by using VGG-16 network[6]. As in [8], we utilize VLAD-k with k=5 to encode them. We also use Latent Concept Descriptors(LCD) as in [8], which outperforms other CNN fully connection layer features. We use 256-component k-means centers for VLAD encoding. For all CNN features, we apply PCA-whitening to reduce their dimension into 512. In the same time, SSR, intra normalization and L2 normalization are used for post processing[1].

2.3. The fusion of classification

We use one-vs-rest linear SVM with C=100 to each of the features. To combine different features, we use late fusion by averaging decision scores.

3. Experiments

We tested our system on various datasets: THUMOS 2014 validation set, THUMOS 2014 test set[3] and THUMOS 2015 validation set[2].

3.1. Setting of validation sets

For the validation sets, we use 13320 trimmed videos from the training set (UCF101) as the training data, and test them with the validation sets. Validation 14 consists of 1010 untrimmed videos from 101 action classes. Each video may contain multiple instances of multiple actions. Validation 15 consists of 2104 untrimmed videos from validation 14 and test 14.

3.2. Setting of test sets

For the test sets, we use both the training set and the validation set as the training data.

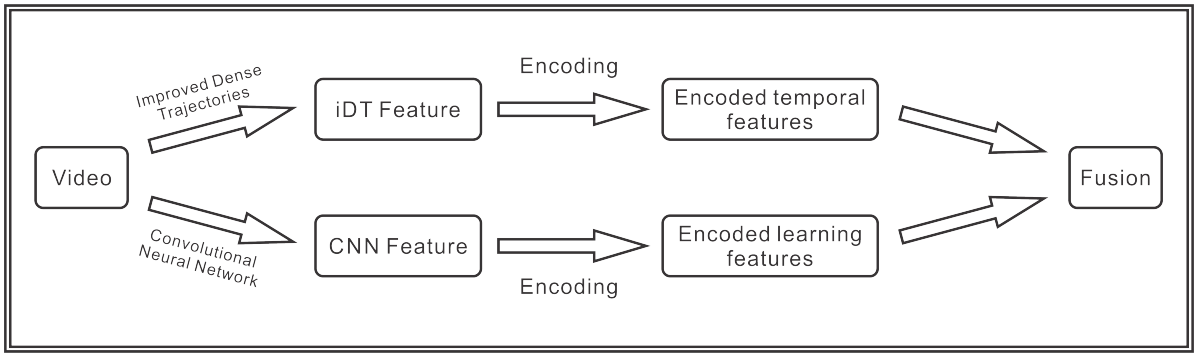

Figure 1. The pipeline of our system.

Features	Validation 2014	Test 2014	Validation 2015
$fc6_{VLAD}$	54.9609%	63.5850%	54.4367%
$fc6\_relu_{VLAD}$	52.1374%	61.2736%	50.7064%
$fc7_{VLAD}$	52.7135%	60.4173%	51.3744%
$fc7\_relu_{VLAD}$	50.2224%	60.0844%	49.1568%
LCD_{VLAD}	60.0553%	67.5938%	59.1312%
iDT_{FV}	54.9123%	66.5924%	55.2664%
iDT_{FV}+LCD_{VLAD}	66.1191%	75.1006%	63.2229%
iDT_{FV}+LCD_{VLAD} +$fc6_{VLAD}$	65.8453%	75.2212%	65.5039%

Table 1. The results on the validation and the test sets in terms of mAP

3.3. Results

For each individual feature, LCD_{VLAD} can achieve the best results over all datasets, even better than the iDT feature.

Acknowledgement

We would like to appreciate Dr. Yi Yang and Mr. Zhongwen Xu of University of Technology Sydney for the helping in this project.

References

[1] R. Arandjelovic and A. Zisserman. All about vlad. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1578–1585. IEEE, 2013.

[2] A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. `http://www.thumos.info/`, 2015.

[3] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. `http://crcv.ucf.edu/THUMOS14/`, 2014.

[4] J. Sánchez, F. Perronnin, and T. De Campos. Modeling the spatial layout of images beyond spatial pyramids. *Pattern Recognition Letters*, 33(16):2216–2223, 2012.

[5] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245, 2013.

[6] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[7] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3551–3558. IEEE, 2013.

[8] Z. Xu, Y. Yang, and A. G. Hauptmann. A discriminative cnn video representation for event detection. *arXiv preprint arXiv:1411.4006*, 2014.