Tianjin University Submission at THUMOS Challenge 2015

Yanbin Liu[†] Baixiang Fan[†] Shichao Zhao[†] Youjiang Xu[†] Yahong Han[†§]
[†] School of Computer Science and Technology, Tianjin Universtiy, China
[§] Tianjin Key Laboratory of Cognitive Computing and Application, China
{csyanbin, superfbx, zhaoshichao, yjxu, yahong}@tju.edu.cn

Abstract

In this notebook paper, we describe our method for the THUMOS 2015 Challenge in action recognition task. In recent researches, people find that action recognition benefits from motion information like improved dense trajectory (IDT) as well as appearance cue. In our approach we combine these two kinds of methods to improve the recognition of human actions. For motion, we adopt the Fisher Vector represented improved dense trajectory for its rich temporal information. For appearance cue, the latent concept descriptor (LCD) and VLAD representation is chosen to capture the static image information. All actions are classified by a one-versus-rest linear SVM with late fusion strategy. We achieve 61.67% mean average precision on the validation set.

1. Introduction

Human action recognition has gained much attention recent years due to its potential application in automatic video analysis, surveillance, sport event analysis and virtual reality. Though progressive work has been done, human action recognition still remains a problem because of intra-class variation, occlusions, view point changes and background noise. THUMOS challenge 2015 aims at the problem of human action recognition and detection on untrimmed videos, which is closer to real-world applications and more difficult. Therefore, it leads to more efficient and effective approaches in human action recognition area. In this paper, we describe our method for action recognition task in detail.

Video representation is of vital importance in recognition. Motion information and appearance cue are two popular approaches in recent years. For motion information, dense trajectory (DT) [3] and improved dense trajectory (IDT) [4] are most widely-used descriptors due to its' rich temporal information and high performance. In Thumos 2014, 10 teams of 11 used dense trajectory features (9 IDT and 1 DT). However, DT and IDT suffer from the huge storage and large computation problems except for its excellent performance.

Appearance cue is another important approach in video representation. Unlike IDT, it focuses on action performers and action scenes. Deep learning frameworks [2] [1] are widely used to capture the appearance cue. The fully connected layer along with average pooling is usually the default setting. Recently, Xu et.al. [5] find that the latent concept descriptor (LCD) with VLAD representation shows better performance than common average pooling features. In our experiments, it obtains higher MAP (59.32%) than IDT (53.91%) in Thumos 2015 validation set.

2. Framework Description

Figure 1 shows our pipeline at Thumos Challenge 2015. It is composed of two streams: motion stream and appearance stream. In motion steam, the IDT feature of each video is extracted and then PCA with whitening is applied. Final representation is represented by Fisher Vector encoding. In appearance stream, we adopt VGG net[2] 16-layer to extract deep features, then Latent Concept Descriptor (LCD) with SPP [5] is used to get descriptors. VLAD is the encoding method here. For both stream we use SVM as the action classifier and get the final prediction using late fusion strategy.

2.1. Motion stream

In motion stream, we first extract the improved dense trajectories [4] with the default parameters and obtain HOG, HOF and MBH features. After L1 normalization, we apply PCA with whitening on these local descriptors and reduce the dimensionality by a factor of two. To train the PCA transformation and Gaussian Mixture Model(GMM), we randomly sample 250,000 descriptors from the general training set. For each type of descriptor, we estimate a Gaussian Mixture Model(GMM) with $K(256)$ Gaussians. In the quantization phase, we use Fisher Vector encoding and get $2DK$ dimensional Fisher vector for each descriptor, where D is the dimension of descriptor after PCA. Then the fisher vector is normalized by power and L2 approach separately. Finally, we concatenate the normalized fisher vectors

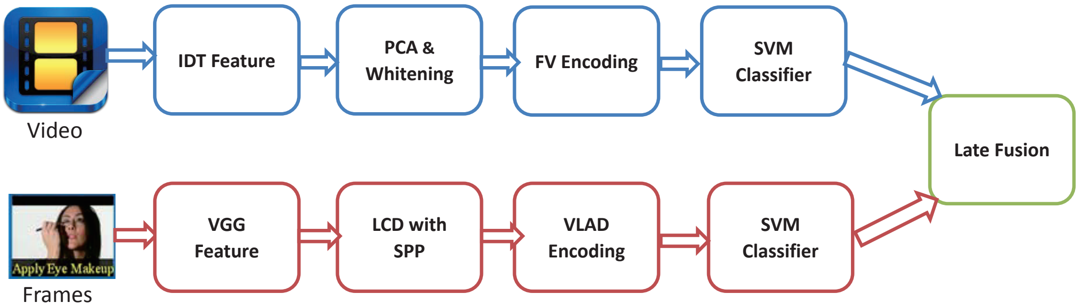

Figure 1. The pipeline used in our submission at Thumos 2015.

of all descriptor types and form the final representation of the video clip. SVM is used to train the action recognizer and predict the test labels.

2.2. Appearance stream

In appearance stream, we adopt the deep net proposed by [2]. In common setting, fully connected layer fc_6 and fc_7 along with average pooling is applied. However, average pooling is not sufficient enough to capture the appearance information. In our pipeline, the Latent Concept Descriptor (LCD) from pool_5 with Spatial Pyramid Pooling is adopted as the descriptor. As in [5], we apply four different CNN max-pooling operations and obtain (6×6), (3×3), (2×2) and (1×1) outputs for each independent convolutional filter, a total of 50 spatial locations for a single frame. Finally we get 512-D latent concept descriptor and reduce it to 256-D by PCA. As to encoding, VLAD with $K(256)$ centers is applied on the 256-D descriptors and generate KD representation for each video. The same as IDT, SVM is also the classifier here.

2.3. Classifier and fusion strategy

In this challenge, we have tried libLinear and libSVM implementation of SVM algorithm and find that libSVM performs better in most cases. So in the experiments, we use the libSVM with linear kernel for classification. We set $C = 100$ for IDT feature and learn 101 one-versus rest-classifier. As to LCD descriptor, we use default setting $C = 1$ and learn 101 one-versus-rest classifier.

IDT feature and LCD feature focus on different aspects in action recognition tasks. And they are suitable for different circumstances and action classes. To better exploit and utilize these two features, we have proposed different fusion strategy:

- Common fusion. $dec_{idt} \in \mathbf{R}^{n\times 101}$ and $dec_{lcd} \in \mathbf{R}^{n\times 101}$ denote the decision values of n tests obtained by each SVM classifier. Final decision value is computed as: $dec = (dec_{idt} + dec_{lcd})/2$.
- Weighted fusion. In validation set, we get the average precision on each of the 101 classes for IDT and LCD as: $ap_{idt} \in \mathbf{R}^{101}$ and $ap_{lcd} \in \mathbf{R}^{101}$. The final decision value is computed as: $dec(:, i) = ap_{idt}(i) * dec_{idt}(:, i) + ap_{lcd}(i) * dec_{lcd}(:, i)$.
- 01 weighted fusion. Special case of weighted fusion. If $ap_{idt}(i) > ap_{lcd}(i)$, $dec(:, i) = dec_{idt}(:, i)$. Otherwise, $dec(:, i) = dec_{lcd}(:, i)$.

3. Experiments

We performed several experiments to evaluate different pipelines and different features on the provided validation set of 2104 videos.

Table 1. mAP on validation set. Motion denotes the combination of HOG, HOF and MBH. AvgPool denotes combination of fc6 average pooling and fc7 average pooling.

Feature Combination	mAP
HOG	0.4459
HOF	0.4919
MBH	0.4874
Motion	**0.5391**
AvgPool	0.4789
LCD	**0.5932**
Motion+LCD	**0.6167**

Setup for Validation set. In the validation phase, we only use training set of 13320 videos from UCF101 to train the model. Background set is not involved to train the model. The detailed configuration of feature extraction and classifier is implemented as described above. Experimental results are reported in table 1. For motion features, HOF achieves the best mAP of 0.4919 and motion combination can be better and achieves 0.5391. For appearance features, fc6 and fc7 average pooling can achieve 0.4789. It is an interesting result that LCD with VLAD pooling can reach 0.5932, even better than motion combination and average pooled features. With motion and LCD features late fusion, we can achieve 0.6167 at the validation set.

Table 2. mAP on particular action classes.

Action Name(Class Index)	LCD result	Motion Result
BaseballPitch(7)	**0.6508**	0.0325
Billiards(12)	**0.9860**	0.4015
FieldHockeyPenalty(29)	**0.7265**	0.2499
JugglingBalls(46)	0.0872	**0.7108**
JumpRope(48)	0.1084	**0.7508**
RopeClimbing(75)	0.4542	**0.8728**
SoccerPenalty(85)	**0.5937**	0.0624
WalkingWithDog(98)	**0.6874**	0.1154

In the validation phase, we find that motion feature and LCD feature perform different for various action classes. Table 2 shows some results on particular action classes. In some scene and object dominated classes like Billiards and WalkingWithDog, LCD performs really better while in action dominated classes like RopeClimbing and JugglingBalls, Motion performs better. Motion and LCD features are good complementary to each other.

Setup for Test set. All UCF101 videos and validation videos are used to train our model. As shown in Table 2, different features perform different in various action classes, so the late fusion strategy in Section 2.3 is used. The 5 submission are listed bellow:

- Run1. Weighted fusion.
- Run2. Common fusion.
- Run3. 01 weighted fusion.
- Run4. LCD feature only as baseline.
- Run5. UCF101 train only as baseline.

4. Acknowledgements

This work was supported by the NSFC (under Grant 61202166,61472276). We would like to express our gratitude to Zhongwen Xu and Yi Yang from ITEE, The University of Queensland, Australia for providing us implementation details of LCD approach.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[2] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[3] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011.

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

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