MSR Asia MSM at THUMOS Challenge 2015

Zhaofan Qiu, Qing Li, Ting Yao, Tao Mei, and Yong Rui
Microsoft Research, Beijing, China
{v-zhqiu, v-liqing, tiyao, tmei, yongrui}@microsoft.com

Abstract

This notebook paper presents overview and comparative analysis of our systems designed for the action classification task of the THUMOS Challenge 2015. We investigate and exploit multiple spatio-temporal clues, i.e., frames, consecutive frames (optical flow) and short video clips, using 2-D or 3-D convolutional neural networks (CNNs). The choice of different CNN layers is studied as well. Furthermore, improved dense trajectory with fisher vector encoding and MFCC audio features are utilized. All actions are classified by late fusing the predictions of one-versus-rest linear SVMs learnt on each clue.

1. Introduction

Recognizing actions in videos [8] is a key ability for a variety of important applications, ranging from video surveillance, video content analysis, and video retrieval to human computer interaction. In this work, we aim at investigating multiple spatio-temporal representations, to action recognition in videos.

The remaining sections are organized as follows. Section 2 describes our action recognition system. Section 3 presents all the features, while Section 4 details the fusion strategy. In Section 5, we provide empirical evaluations, followed by the conclusions in Section 6.

2. Recognition Framework

The framework of our proposed action recognition system is shown in figure 1. The key point of our designed framework is using multiple stream features, while many actions show different characters on different time scales. After multiple stream feature extraction, all the video representations will be used to train linear SVMs and predict final action scores by applying linear fusion method.

3. Feature Extraction

To extract full useful information in video for action recognition, we choose frames, optical flow, short clips, long clips and audio as five streams on different time scales. We will describe each feature representation, and its implementation details in our experiment part.

3.1. Frame

For action recognition, individual video frames can provide useful characters as some actions are strongly associated with particular scenes and objects. To make full use of static frame appearance, VGG_19 [5], which is the recent superior CNN architecture for image classification, is picked out to extract high level visual feature from each sampled video frame. VGG_19 is a very deep convolutional networks up to 19 weight layers (16 convolutional layers and 3 fully-connected layers) for large scale image classification. With the help of pre-train process using large dataset from ImageNet challenge, the VGG_19 model can be used to extract plentiful visual concepts, e.g., scenes and objects.

For recognition task on THUMOS challenge, we fine-tune the original VGG_19 model on UCF101 dataset[6] with 13320 trimmed videos, and extract the outputs of fully-connected layers (fc6, fc7 and fc8) as deep representation of frames. Then we apply mean pooling on fully-connected layers feature of sampled frames as video representation respectively.

3.2. Optical Flow

To model the relevance of the consecutive frames, we apply another CNN to optical flow "image", which can extract motion feature between consecutive frames. As the input of this CNN, the optical flow is firstly computed between consecutive frames with [1] , and then is converted to flow "image" by centering horizontal (x) and vertical (y) flow values around 128 and multiplying by a scalar such that flow values fall between 0 and 255. After the transformation, we have got two channels for optical flow "image" while the third channel is created by calculating the flow magnitude. Furthermore, to suppress the displacement caused by camera motion, which may introduce additional noise into CNN, a global motion component is estimated by the mean vector of each flow and then subtracted from the flow.

Similar to frames, we also fine-tune the original VGG_19

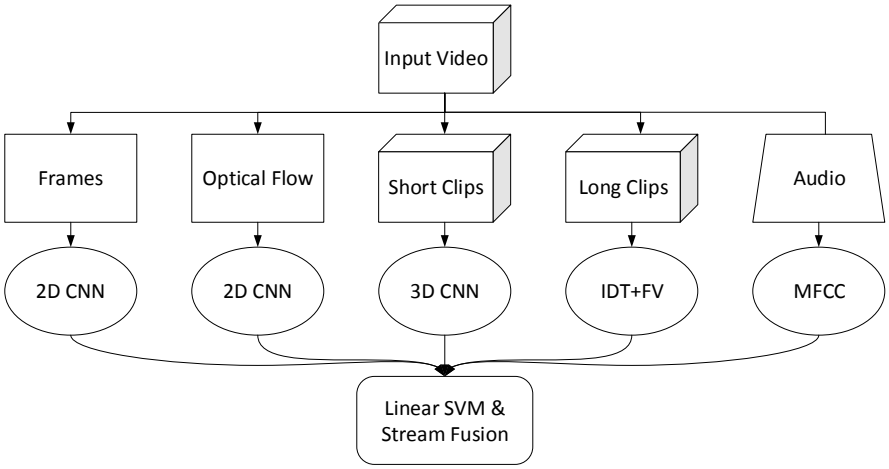

Figure 1. Framework of our proposed system.

model on optical flow "images" extracted from UCF101 dataset, and extract the outputs of fully-connected layers as visual representations of optical flows. Then, mean pooling is utilized to fuse the outputs of fully-connected layers of sampled optical flow "images" as video motion feature.

3.3. Short Clip

Besides individual frames and motions between consecutive frames, we also use 3D Convolutional Neural Networks (3D CNN) to construct video clip features from both spatial and temporal dimensions. Unlike traditional 2D CNN, 3D CNN architecture takes short video clip (multiple continuous frames) as the inputs and consists of alternating 3D convolutional and 3D pooling layers, which are further topped by a few fully-connected layers as described in [3]. For describing video clips by more powerful feature using 3D CNN, we choose the superior architecture in [7], named C3D, which aims at learning spatial-temporal features for video clips using 3D CNN trained on large-scale video dataset. As the deep network architecture is pre-trained on a large-scale video dataset [4], C3D can model general appearance and motion information simultaneously, which is important for action recognition.

Unlike using VGG_19 model, we do not fine-tune the C3D model because the original C3D model is pre-trained on Sports-1M dataset, which can extract more general feature than the fine-tuned model. As designed in C3D architecture, the input of C3D model is 16-frame clip, thus we sample continuous clips with none overlap in the video. When a video is more than one short clip, average pooling is then applied to combine clip-level to video-level feature.

3.4. Long Clip

For long-term clips, we choose the state-of-the-art hand-crafted feature - improved dense trajectory (iDT) [8] on each sampled clip. Specifically, trajectory feature, histogram of oriented gradients (HOG), histogram of flow (HOF), and motion boundary histogram (MBH) are computed for each trajectory obtained by tracking points in video clips. Because of different number of trajectory features for different clips, Fisher vector encoding is used to quantize the features and create a high dimensional representation for each clip.

Although iDT can extract feature for video of any length, for fast extraction, we convert long videos to 5-second clips with none overlap. Considering the average duration of UCF101 dataset is less than 10 seconds, we treat each video in the dataset as a long clip. But for the challenge videos, most of the videos consist of more than one long clips. Thus we predict action scores for each long clip and then apply max pooling for final prediction.

3.5. Audio

Audio feature is the most global feature (though entire video) in our system. Although audio feature itself can not get very good result for action recognition, but it can be seen as powerful additional feature, since some specific actions are highly related to audio information. For basic audio feature extraction, we choose MFCC from audio signals and then quantize them into BoWs with 4000 words as one of the video representations.

4. Classification and Stream fusion

As described in the section above, we have extracted five different video representations on different time scale. When utilizing these features for action recognition, we use the linear SVM with linear kernel [2] for classification. We set the SVM parameter C=10 for all features, and train 101 one-versus-rest classifiers with L2-norm for all the input

Table 1. Single stream results on validation set.

Stream	Layer	mAP
Frames	fc6	43.0
	fc7	41.4
	fc8	36.8
Optical flow	fc6	32.7
	fc7	30.6
	fc8	26.4
Short Clips	fc6	58.9
	fc7	58.2
	fc8	50.3
Long Clips		55.7

Table 2. Stream fusion results on validation set.

Stream	mAP
S1 all-layer	44.0
S2 all-layer	33.2
S3 all-layer	60.9
S4	55.7
S1 + S2	47.7
S1 + S3	61.7
S1 + S4	63.5
S2 + S3	62.2
S2 + S4	57.7
S3 + S4	68.2
S1+S2+S3+S4	70.0

features. Although complex fusion method can be used in our framework, we simply choose linear fusion for different streams.

5. Experiment

In order to determine the fusion parameters for different streams, we test our method on the "validation" set with 2104 untrimmed videos. Furthermore, when we test on the "validation" set, we only use the videos in UCF101 dataset as training data, and test on the validation videos shorter than 4 minutes. When we predict the scores of "test" set as we submitted, the videos in UCF101 and "validation" dataset are both used to train SVM model. Because not all the videos in UCF101 dataset have audio track, we don't consider audio stream while test on validation set, but while predict on test set, the audio feature is mixed in.

Single stream experiment. First, we test each stream in our framework on validation set, and the mAP is shown on Table 1. We can find that "short clips" stream using C3D can get the highest mAP among all single streams, which is 3.2% higher then "long clips" using IDT and Fisher Vector. And for deep representation using Neural Networks, we also get mAP on features extracted from different fully-connected layers. Not surprisingly, as presented on the previous works, fc6 layer can always get higher mAP than other layers.

Stream Fusion experiment. We apply different combination of streams in our framework on validation set. This experiment can illustrate different complementarity between streams as shown in Table 2. The four stream in the framework is marked as S1, S2, S3, S4, which represent frames, optical flow, short clips and long clips. Result of each stream is linear fusion of different layers while using representation of 2D/3D CNN. The first three rows in the table show that all-layer fusion will get a bit better performance than each layer. Then we post all combinations of two stream of our system, we can find that, the combination "S1 + S4" and "S2 + S3" get higher increasing than others, which may be caused by high complementarity of frame and iDT, optical flow and C3D. The last row show the highest mAP we can get (70.0%) which combine all the streams (except audio stream) on validation set.

6. Conclusion

In THUMOS Challenge 2015, we mainly focused on multiple visual features especially short-term and long-term temporal information which are good at describing actions. The low-level audio features can help detect some actions and slightly improve the performance. Semantic information such as ASR and OCR can be employed in future work. In addition, the fusion methods of different clues should be carefully studied.

References

[1] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, pages 25–36. Springer, 2004.

[2] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[3] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Trans. on PAMI*, 35(1):221–231, 2013.

[4] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

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

[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] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. *arXiv preprint arXiv:1412.0767*, 2014.

[8] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, pages 3551–3558. IEEE, 2013.