Action Recognition and Detection by Combining Motion and Appearance Features

Limin Wang[1,2], Yu Qiao[2], Xiaoou Tang[1,2]

[1] Department of Information Engineering, The Chinese University of Hong Kong
[2] Shenzhen Key Lab of CVPR, Shenzhen Institutes of Advanced Technology
Chinese Academy of Sciences, Shenzhen, China
07wanglimin@gmail.com, yu.qiao@siat.ac.cn, xtang@ie.cuhk.edu.hk

Abstract. We present an action recognition and detection system from temporally untrimmed videos by combining motion and appearance features. Motion and appearance provides two complementary cues for human action understanding from videos. For motion features, we adopt the Fisher vector representation with improved dense trajectories due to its rich descriptive capacity. For appearance feature, we choose the deep convolutional neural network activations due to its recent success in image based tasks. With this fused feature of iDT and CNN, we train a SVM classifier for each action class in one-vs-all scheme. We report both the recognition and detection results of our system on THUMOS 14 Challenge.

1 Introduction

Human action recognition has been one of the challenging problems in computer vision. Great research efforts have been devoted to designing effective and robust action recognition methods [1]. THUMOS challenge has become an important contest to evaluate the performance of various recognition and detection system [1]. Unlike THUMOS Action Recognition Challenge 2013 [3], THUMOS Challenge 2014 [4] focuses on action recognition and detection of temporally untrimmed videos. These temporally untrimmed videos require us to designing more effective and efficient action recognition algorithms. In this paper, we briefly describe the proposed method for the tasks of both action recognition and detection from temporal untrimmed videos.

Video representation plays an important role in the action recognition and detections. Recent study work [7] shows that the Fisher Vector representation [9] with improved Dense Trajectory (iDT) features [11] is very effective for capturing motion information, and it has obtained the state-of-the-art performance on several action recognition datasets, such as HMDB51 [6] and UCF101 [10]. However, the iDT features mainly focus on describing 3D volumes with high motion salience and may not be suitable for representing videos with static actors, such as Playing Flute and Playing Guitar. Static appearance is another

[1] http://crcv.ucf.edu/THUMOS14/home.html

important cue for understanding human action as action is usually performed in a specific pose configuration and may have interaction with specific objects. Recently image classification from static appearance has witnessed the progress by by using deep convolutional neural network [5]. Furthermore, the deep convolutional neural network (CNN) can be seen as a general feature extractor and the CNN trained on a larger dataset such as ImageNet can be effectively adapted to other datasets or other tasks [8]. The CNN features has obtained good performance on several related tasks such as scene recognition, attribute recognition, and image retrieval due to its good capacity of extracting useful static appearance information.

As motion and appearance are both important cues for understanding human action from videos, we propose a new video representation by combining both iDT features and CNN features for the tasks of action recognition and detection on temporally untrimmed videos. This new representation leverages the benefits of both traditional feature engineering (iDT+FV) and newly feature learning (CNN features), and proves that these two kinds of representation are complementary to each other for the task of human action understanding from videos.

2 Method

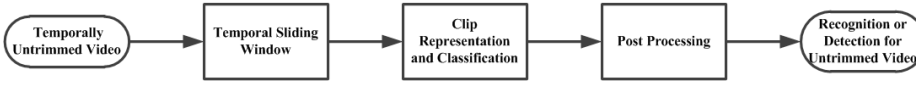

Fig. 1. The pipeline of action recognition and detection on temporally untrimmed videos.

The pipeline of our proposed action recognition and detection system is shown in Figure 2. It is mainly composed of three steps: (i) temporal sliding window, (ii) clip representation and classification, and (iii) post processing. We will describe the details of each step in the following sections.

2.1 Temporal Sliding Window

These temporal untrimmed videos contain complex and redundant visual content, and the duration of action usually occupies a small portion of the whole video stream. We firstly conduct temporal sliding window scanning to segment an continuous video stream into overlapped short video clips. Based on the statistics on the dataset of UCF101, we set the length of window as 150 frames and the sliding step as 100 frames. Then, for each short temporal window, we perform the task of action recognition independently. Eventually, the recognition results of these short window are combined to yield the final result of the whole video stream.

2.2 Cip Representation and Classification

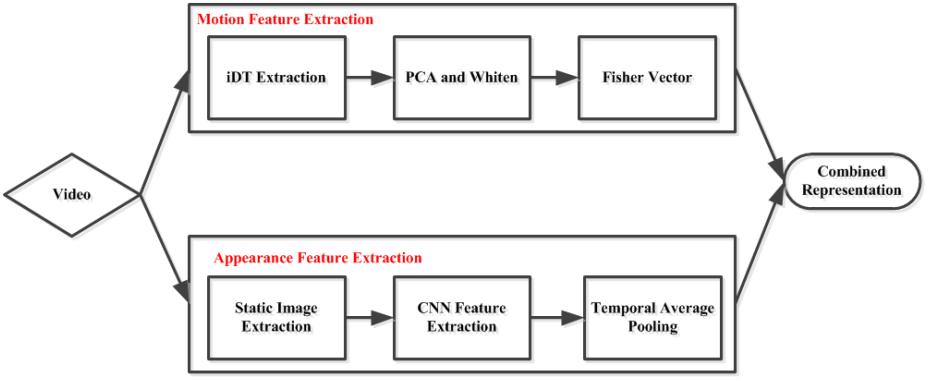

Fig. 2. The extraction of motion and appearance features for a short clip.

For each video clip, we firstly extract motion and appearance features to represent the visual content as shown in Figure 2. Then, with this combined representation, we train multiple one-vs-all SVM classifiers to perform action recognition.

Motion Feature Extraction. Motion is an important cue for human action understanding from video. According to a recent study work [7], we firstly extract improved trajectory features using the public code with default parameter settings. Totally, it extracts five kinds of descriptors: warped trajectories, HOG, warped HOF, warped MBHx, and warped MBHy. We choose the latter four kinds of descriptors due to their good performance in practice. Then, we resort to Fisher Vector representation to aggregating these local descriptors through a high-dimensional encoding scheme. To make the training of Gaussian Mixture Models (GMM), we reduce the dimension of each descriptor by a ratio of 0.5. We conduct FV encoding for each kind of descriptor independently and the resulting super vectors are normalized by intra power normalization. Finally, these normalized super vectors are concatenated to represent the motion information for this video clip.

Appearance Feature Extraction. Appearance information such as human pose, static object, and scene is able to provide complementary cues for human action recognition. We follow the recent success of deep convolutional neural network (CNN) in image classification. We firstly select 30 static frames from each short video clip and then each image is resized as 256×256. We then extract a 4096-dimensional feature vector from the cropped 227×227 region using the Caffe implementation [2] of the CNN described by Krizhevsky *et al.*

[5]. The details about the network architecture can be found in [5]. To make the CNN features more discriminative for action recognition, we fine tune the parameters of CNN using the UCF101 dataset through $50,000$ iterations. Finally, we conduct average pooling of these CNN features across the selected frames obtain a global representation for the video clip, which is further normalized by power normalization.

Classifier Training. Using the fusion feature of iDT and CNN, we train SVM classifiers for each action class using one-vs-all scheme. In order to distinguish action classes with background classes, we randomly select $4,000$ video clips from the background dataset. Thus, when we train the SVM classifier for each action class, we add these background clips as negative examples.

2.3 Post Processing

Given the classifier predictions of video clips, we need to generate the recognition and detection results for the whole video streams. To avoid the false recognition results, we use two thresholds: clip threshold (τ_1) and video threshold (τ_2). For each video clip, we assume that there are at most τ_1 action classes, and, for each video stream, there are at most τ_2 action instances. We sort all the recognition results according to their SVM scores and eliminate these results with low confidence score using the thresholds. For action detection task, we also set another threshold of τ_3 to eliminate the detection results with low confidence scores.

3 Results

In this section, we will give the implementation details of our experiments and provide the performance of our proposed method on the testing dataset.

In current implementation, we set the duration of sliding window as 150-frames and sliding step as 100-frames. We firstly divide the whole video sequence into short clips according to the temporal sliding window. Then, for each short clip, we extract improved dense trajectory features [11] using the public available code [2] with default parameter setting. We choose four types of descriptors, namely HOG, HOF, MBHx, and MBHy, and encode each of them independently. We use the PCA and whiten technique to reduce the dimension these local descriptors by a ratio of 2. For Fisher vector encoding, we choose the number of mixture as 256. Thus, the dimension for the final motion feature is $(48 + 54 + 48 + 48) \times 256 \times 2 = 101,376$.

For appearance feature, we select 15 frames from each short video clip and we then extract a 4096-dimensional CNN activation vector for each frame. Finally, we resort to an average pooling over these frames to obtain a global representation the video clip. Finally, we also conduct power and ℓ_2 normalization for this CNN appearance feature. We fuse the appearance and motion feature into a

[2] https://lear.inrialpes.fr/people/wang/improved_trajectories

Table 1. Testing results for the task of action recognition.

(τ_1,τ_2)	(1,10)	(1,20)	(5,10)	(5,20)	(101,101)
result	0.617	0.6177	0.6196	0.6174	0.6201

Table 2. Testing results for the task of action detection.

(τ_1,τ_3)	(1,0.5)	(1,0)	(1,-0.5)	(1,-1)
overlap = 0.1	0.1080	0.1373	0.1701	0.1818
overlap = 0.2	0.1042	0.1319	0.1591	0.1700
overlap = 0.3	0.0891	0.1137	0.1306	0.1405
overlap = 0.4	0.0765	0.0975	0.1090	0.1174
overlap = 0.5	0.0563	0.0695	0.0775	0.0834

single hybrid representation, whose dimension is $105,472$. We set fusion weight as 0.4 for appearance feature and 1.0 for motion motion feature.

For action recognition, we firstly estimate the predictions for each short video clips independently. We use the threshold τ_1 to keep at most τ_1 prediction for this clip. Then we use the max pooling over the clip detection results to obtain the final prediction for the whole video clips. We also use another threshold τ_2 to eliminate the low detection results from the whole video sequence. For varying threshold of τ_1 and τ_2, the recognition results are shown in following table: From these results, we observe that our proposed method is not insensitive to the parameter settings of (τ_1, τ_2).

For action detection, for each video clip, we consider there is only one action class, which means we set the threshold $\tau_1 = 1$. We add another threshold τ_3 to eliminate the detection result with low confidence score. From these results, we see that our method is sensitive to the threshold of τ_3. Larger threshold may cause the problem of deleting the true positive detection results. Smaller thresholds will lead to more detection results to kept for final evaluation.

4 Conclusion

We have presented a sliding window based method for action recognition and detection from temporally untrimmed videos by combining motion and appearance features. We evaluate the proposed methods on T 2014 action recognition and detection task. Experimental results show the effectiveness of our method on action recognition tasks, while the detection results are still low. This indicates that our single scale temporal sliding scheme may be not a good choice for temporal segmentation of video sequence. In the future, we may consider to how to design an effective temporal segmentation algorithm of dividing video sequence into short clips.

References

1. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: A review. ACM Comput. Surv. 43(3), 16 (2011)
2. Jia, Y.: Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/ (2013)
3. Jiang, Y.G., Liu, J., Roshan Zamir, A., Laptev, I., Piccardi, M., Shah, M., Sukthankar, R.: THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/ICCV13-Action-Workshop/ (2013)
4. 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)
5. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. pp. 1106–1114 (2012)
6. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: A large video database for human motion recognition. In: ICCV. pp. 2556–2563 (2011)
7. Peng, X., Wang, L., Wang, X., Qiao, Y.: Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. CoRR abs/1405.4506 (2014)
8. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. CoRR abs/1403.6382 (2014)
9. Sànchez, J., Perronnin, F., Mensink, T., Verbeek, J.J.: Image classification with the fisher vector: Theory and practice. International Journal of Computer Vision 105(3), 222–245 (2013)
10. 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)
11. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV. pp. 3551–3558 (2013)