Filomat 34:15 (2020), 4967–4974
https://doi.org/10.2298/FIL2015967W

Published by Faculty of Sciences and Mathematics,
University of Niš, Serbia
Available at: http://www.pmf.ni.ac.rs/filomat

Human Action Recognition Based on Deep Network and Feature Fusion

Dongli Wang[a], Jun Yang[a], Yan Zhou[a], Zhen Zhou[a]

[a] *School of Automation and Electronica Information, Xiangtan University, Xiangtan, China 411105*

Abstract. Feature representation is of vital importance for human action recognition. In recent few years, the application of deep learning in action recognition has become popular. However, for action recognition in videos, the advantage of single convolution feature over traditional methods is not so evident. In this paper, a novel feature representation that combines spatial and temporal feature with global motion information is proposed. Specifically, spatial and temporal feature from RGB images is extracted by convolutional neural network (CNN) and long short-term memory (LSTM) network. On the other hand, global motion information extracted from motion difference images using another separate CNN. Hereby, the motion difference images are binary video frames processed by exclusive or (XOR). Finally, support vector machine (SVM) is adopted as classifier. Experimental results on YouTube Action and UCF-50 show the superiority of the proposed method.

1. Introduction

Recognition of human actions in video is a research hotspot in the field of computer vision in recent years [1]. It has drawn a significant amount of attention from the academic community [2, 3]. It is a basic technology for many applications such as intelligent monitoring, human-computer interaction, and robotics. Human action Recognition is also a challenging task that is influenced by many factors such as different lighting conditions, perspectives, complex backgrounds, and large intra-class variations. The key way to recognize human action is to capture spatial and temporal information. In this paper, the fusion network based on convolutional neural network (CNN) and long short-term memory (LSTM) is proposed.

The traditional computer vision pipeline consists of two steps: action feature representation and action recognition. Action feature representation is the extraction of features that characterize the key information of the video. This process plays a key role in the whole recognition process. The quality of the feature directly affects the final recognition effect. The actions recognition stage takes the feature vector obtained in the previous stage as input, and learns the parameters through the algorithm then classifies. Action recognition methods can be roughly divided into two categories: hand-crafted and deep learning. For hand-crafted category, most popular descriptors are represented by HOG (Histogram of oriented gradient) [4], SIFT (Scale-invariant feature transform) [5], HOF (Histogram of Optical Flow) [6] and MBH (Motion

2010 *Mathematics Subject Classification*. Primary 93XX; Secondary 68XX
Keywords. action recognition, deep leaning, convolutional neural network, long short-term memory, feature fusion
Received: 29 September 2018; Revised: 19 November 2018; Accepted: 22 March 2019
Communicated by Shuai Li
Research supported by the National Natural Science Foundation of China (61773330, 61100140, and 61104210), the Natural Science Foundation of Hunan Province (2017JJ2253), and the Research Project of Department of Education of Hunan Province (19C1740).
Email address: yanzhou@xtu.edu.cn (Yan Zhou)

Boundary Histograms) [7]. Furthermore, there are some scholars who talk about the feature extension to 3D features, such as 3DSURF [8], HOG3D [9], etc. Due to the success of deep learning in the field of image processing, it has been considered apply to human action recognition in recent years. Thanks to the ImageNet Large Visual Recognition Challenge (ILSVRC), a number of Image classification models based on CNN have derived, such as AlexNet [10], GoogleNet [11] and VGGNet [12]. CNN not only has a good effect on image classification, but the descriptors it extracts also perform well on many other tasks, such as object detection [13], scene labeling [14] and action recognition [15].

Recently, the CNN have been used in the field of actions recognition. Some studies do not break the network's fluency using the network framework to do the end-to-end system research [2,16]; others are different, they use CNN to extract feature descriptors, then they can do feature fusion [17], or they can code in various ways and then access classifier classification [18].

In this paper, fusion network based on CNN and LSTM is proposed to extract spatial and temporal information. Different from the conventional CNN extraction feature, one network performs exclusive or (XOR) preprocessing before a video frame enters the network to extract global motion information. Converged networks can highlight local and global information, time and space information, resulting in improved accuracy. We did an evaluation on UCF-50 [19] and YouTube Action [20] datasets and achieved considerable results.

The main contributions of the paper are threefold:

1). We propose a new fusion network that takes into account both spatial and temporal information, as well as the integration of details and global motion information.
2). A new model is proposed to extract global motion information. As compensation information, its integration with spatial and temporal information makes the fusion features more representative.
3). Evaluation of proposed framework in two challenging datasets. This involves looking at how individual channel information and fusion information effect on result, and what their respective advantages and disadvantages.

2. Network Architecture

Inspired by [16, 24], this paper utilizes CNN to learn spatial features and LSTM to learn temporal features. On this basis, we simplified the motion information of sparse frames of video in the form of CNN and LSTM. Then, we designed an auxiliary network to capture the global motion information. Since XOR preprocessing used, the two networks respectively called LSTM-CNN and XOR-CNN. We devise our video recognition architecture accordingly, dividing it into two streams, as shown in Figure 1. Finally, support vector machine (SVM) is adopted as classifier. The two networks can be individually work and late fusion. In the experimental part, we evaluated the effect of this comparison.

In this task, two features are extracted from two networks. Global motion features is extracted from the motion difference image by pre-training CNN model, while spatial and temporal features is extracted by pre-training CNN model and retrained LSTM model respectively. Two networks will be described in section 2.1 and 2.1 separately.

2.1. LSTM-CNN Network

As shown in Figure 1, LSTM-CNN network accepts RGB image input. Each video is averagely extracted 40 frames as CNNs input to extract spatial information. In this part, the pre-trained model inceptionV3 [21] is considered for its good effect on extracting spatial information. Later we made a comparison of the evaluation results of the network. InceptionV3 is a network that has a small amount of parameters and can maintain high quality at low computational cost. Therefore, it is widely acclaimed in most tasks that require convolution. The proposed network wants to take advantage of its advantages to bring good results.

To capture the temporal structure information of actions, LSTM units, which was first proposed in [22], is used. LSTMs are recurrent modules which enable long-range learning. That why we introduce it to model the relationships between spatial information obtained from the CNN and temporal information.

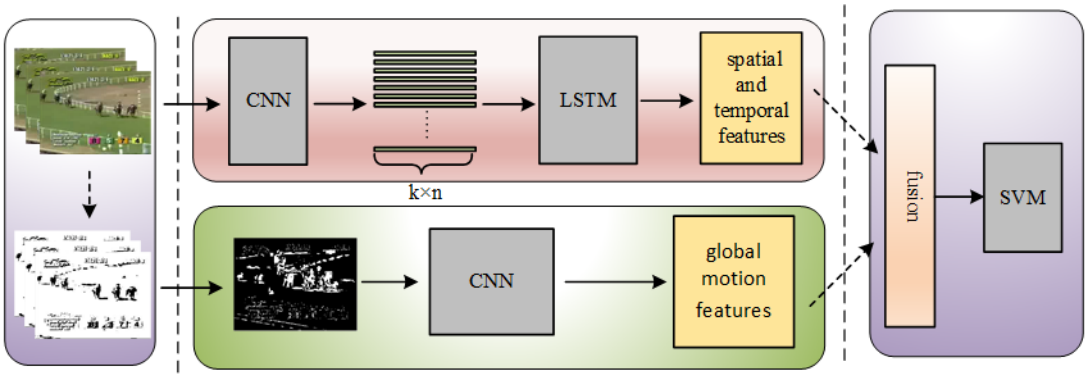

Figure 1: Converged network architecture for actions recognition.
*In the figure, k represents k frames in a video, and n represents the spatial feature dimension of a video frame.

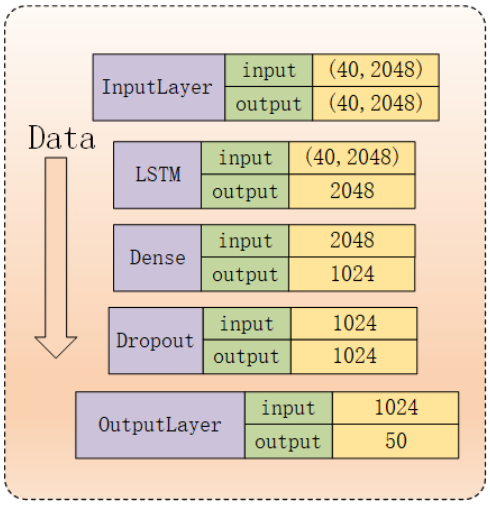

Figure 2: LSTM network structure.

The LSTM network structure on UCF-50 training work is shown in Figure 2. It has a few layers, which contain only one LSTM layer. But it is enough to extract useful temporal information.

LSTM network overcomes most of the basic problems of conventional recurrent neural networks (RNN), such as gradient disappearance and gradient explosion. The internal structure of the LSTM unit is shown in Figure 3. LSTM unit adds three gates to the standard RNN, respectively called input gate, marked as i_t, forget gate, marked as f_t, and output gate, marked as o_t.

It has two clues, one is the cell state variable C_t and the other is the output variable h_t. These two variables are updated over time. The intracellular renewal equation is shown in (1).

$$\begin{cases} f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\ i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \bar{C}_t = \sigma(W_C \cdot [h_{t-1}, x_t] + b_C) \\ C_t = f_t \cdot C_{t-1} + i_t \cdot \bar{C}_t \\ o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \end{cases} \tag{1}$$

The updated equation for h_t is shown in Equation (2).

$$h_t = o_t \cdot \tanh(C_t) \tag{2}$$

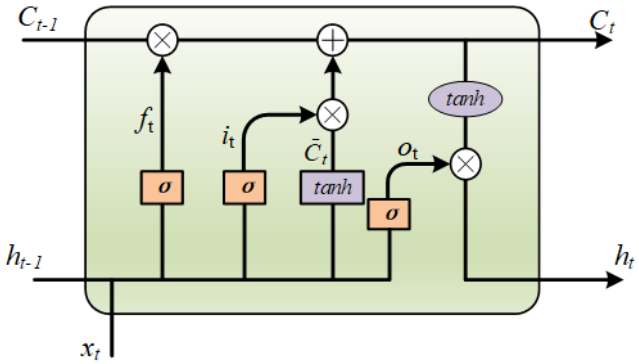

Figure 3: Internal structure of the LSTM unit.

The main process of this part is as follows: First, pre-trained inceptionV3 model is used to extract spatial features from the original video RGB frames. And features is extracted from the final average pooling layer of inceptionV3 network. Each video frame after extraction from inceptionV3 model has 2048 dimensions, so the spatial features dimension of a video is (40, 2048). The spatial feature of all videos is made into a spatial feature library. Then, LSTM network is trained by spatial feature library. Retrained LSTM model is used to extract temporal features. And the features is extracted from the dense layer, so that the final space-time feature dimension is 1024.

2.2. XOR-CNN Network

In order to reduce the calculation, we try to use the XOR method to extract image difference information without using optical flow. As shown in Figure 1, the input of this network is preprocessed by binarization and XOR. RGB video frame is extracted from video. Then the RGB images is extracted into binary image using adaptive binarization method [23], the function recorded as F(x). Finally all binary images of the same video are XOR into one image, so that a video is represented by one image.

Original video is expressed as V_o.

$$V_o = [x_1, x_2, \cdots, x_n] \tag{3}$$

the video frame extracted through the binary image is represented as

$$V_b = F(V_o) = [\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_n] \tag{4}$$

Thus, the global motion represented image can be represented as

$$V_{xor} = \bar{x}_1 \oplus \bar{x}_2 \oplus \bar{x}_3 \oplus \cdots \oplus \bar{x}_n = x_{xor} \tag{5}$$

Same as the previous network. The pre-trained inceptionV3 model is used to extract global motion features from V_xor. Similarly, on the one hand, the feature of 2048 dimension is taken out in final average pooling layer. Global motion features library also be established, and then used to fuse the space-time features at the end of the framework. On the other hand, SVM is used to classify the sample category.

3. Experiments and Result Analysis

In this section, two challenging datasets are tested several times to show the performance of the proposed network. The rest of this section is as follows. In Sect. 3.1 we describe the dataset related information. In Sect. 3.2 we introduce some experimental related settings. In Sect. 3.3, the representation of proposed framework on dataset is described. Sect. 3.4 is the evaluation of our action proposals.

3.1. Datasets

Two challenging datasets UCF-50 and YouTube are used in our experiments. All of these datasets are captured in uncontrolled environment and real world. YouTube are medium scale datasets, and UCF-50 are large scale datasets. Example frames are shown in Figure 4.

Figure 4: Examples images from video sequences on YouTube action dataset (top row) and UCF-50 (bottom).

YouTube action dataset: It contains 11 action categories: basketball shooting, biking/cycling, diving, golf swinging, horseback riding, soccer juggling, swinging, tennis swinging, trampoline jumping, and volleyball spiking, and walking with a dog. For each category, the videos are grouped into 25 groups with more than 4 action clips in it.

UCF-50: It is an action recognition dataset with 50 action categories, consisting of realistic videos taken from YouTube. For all the 50 categories, the videos are grouped into 25 groups, where each group consists of more than 4 action clips.

Both in two dataset, the video clips in the same group may share some common features, such as the same person, similar background, similar viewpoint, and so on. Our model was evaluated in the original video. The average accuracy of all categories is reported.

3.2. Settings

The extract work in this paper is divided into two blocks, in which global motion represented image extraction is performed on MATLAB, and other parts are completed on python. We execute our code on Intel(R) Core(TM) i5 system with 8G RAM and NVIDIA GTX1060 GPU. The entire experiment is based on the keras, which is a high-level neural networks API, written in python and capable of running on top of tensorflow. The base code is created on [24], on the basis of which we have merged our ideas.

3.3. Performance

The channel comparison experiment result is shown in Table 1. It can be seen that for individual global motion information, the performance achieves 65.3% on YouTube action and 62.2% on UCF-50 dataset, proving that individual global motion information is valid, but not enough to classify separately. The main reason for this may be the loss of some details and background information.

Table 1: Channel comparison experiment result

Inceptionv3(SVM)	YouTube(%)	UCF-50(%)
XOR-CNN	65.3(C=3)	62.2(C=4)
LSTM-CNN	80.4	78.2
Fusion network	88.7(c=0.001)	85.5(c=0.01)

*The parameters for SVM classification are as follows:
All kernel function is linear function, penalty parameter C is shown in the table.

For spatial and temporal information, the performance achieves 80.4% on YouTube action dataset and 78.2% on UCF-50 dataset, explaining that spatial and temporal information plays an important role in video behavior classification. It can be seen from this that spatial and temporal information is stronger than the

Table 2: Performance comparison of accuracy with other approaches on the YouTube dataset.

Algorithm	Accuracy(%)
Static + motion feature [20]	71.20
Hierarchical feature on ISA + BoF + Chi-square kernel[25]	75.80
Relative motion descriptor (RMD) + Modes [26]	81.70
Dense trajectory + HOG +HOF + MBH + BoF [27]	84.10
SIFT trajectory + HOG +HOF + MBH + BoF [27]	73.20
KLT trajectory + HOG +HOF + MBH + BoF [27]	79.50
Dense cuboids + HOG + HOF + MBH +BoF [27]	81.40
Dense trajectory + BoF [28]	84.20
CNNs+LDS [17]	86.16
XOR+LSTM+CNN+SVM	**88.70**

global motion information. When we combine two information together, the performance achieves 88.7% on YouTube action dataset and 85.5% on UCF-50 dataset. Compared to spatial and temporal information, fusion feature improves the performance by a large margin of 8.3% on the YouTube dataset and 7.3% on the UCF-50 dataset.

3.4. Evaluation of Our Action Proposals

To show the advantages of proposed approach for human action recognition, we conduct extensive comparison with other methods. The comparison results of YouTube Action are shown in Table 2 and results of UCF-50 dataset are shown in Table 3. By comparing the results, we find that the proposed method of using CNN to extract features performs better than most previous methods on YouTube and UCF-50 datasets, indicating that CNN can extract features effectively. In addition, the method in this paper is superior to [17], indicating that the integration of global motion information and temporal and spatial information can make the features more representative.

Table 3: Performance comparison of accuracy with other approaches on the UCF-50 dataset.

Algorithm	Accuracy(%)
Motion interchange patterns [29]	72.68
Orientation-based descriptor+ Gabor STIP [30]	72.90
GIST3D + STIP [32]	73.70
Motion feature [19]	76.90
Lagrangian particle trajectories [31]	81.03
Relative motion descriptor (RMD) + Modes [26]	81.80
SIFT trajectory + HOG +HOF + MBH + BoF [27]	71.80
KLT trajectory + HOG +HOF + MBH + BoF [27]	78.10
Dense cuboids + HOG + HOF + MBH +BoF [27]	80.20
Dense trajectory + HOG +HOF + MBH + BoF [27]	84.50
CNNs+LDS [17]	82.76
XOR+LSTM+CNN+SVM	**85.50**

Furthermore, we print out the recognition scores of every category on the more complex UCF-50 datasets, as shown in figure 5. It can be seen from the results that high accuracy on Drumming, Fencing, Military Parade and Playing Guitar. The main reason may be that they have a large gap between the categories and the characteristics of the action categories are outstanding. But, it also get low accuracy on Golf Swing, Javelin Throw, High Jump and Nun chucks, proving that proposed method does not work well in some cases where the speed is high and the background is confusing.

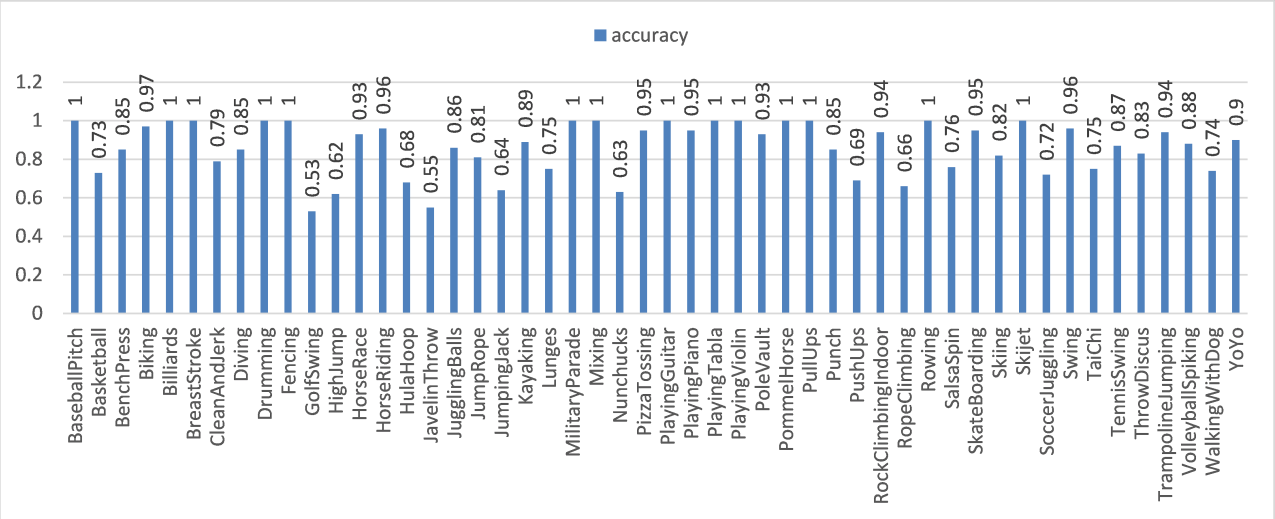

Figure 5: The recognition scores of every category on UCF-50 dataset.

4. Conclusion

In this paper, a new representative descriptor has proposed for human action recognition on realistic datasets. The advantages of CNN and LSTM have been fused to extract temporal and spatial feature, as well as global motion information. Results on the YouTube and UCF-50 datasets have demonstrated that our approach is superior to most existing descriptors. Future work will be focused on the extraction of motion difference images and the feature fusion methods.

5. Acknowledgments

This work was supported by the National Natural Science Foundation of China (61773330, 61100140, and 61104210), the Natural Science Foundation of Hunan Province (2017JJ2253), and the Research Project of Department of Education of Hunan Province (17B259).

References

[1] Poppe R. A survey on vision-based human action recognition[J]. Image and vision computing, 2010, 28(6): 976-990.
[2] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(1): 221-231.
[3] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]//Advances in neural information processing systems. 2014: 568-576.
[4] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005, 1: 886-893.
[5] Lowe D G. Object recognition from local scale-invariant features[C]//Computer vision, 1999. The proceedings of the seventh IEEE international conference on. Ieee, 1999, 2: 1150-1157.
[6] Laptev I, Marszalek M, Schmid C, et al. Learning realistic human actions from movies[C]//Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008: 1-8.
[7] Dalal N, Triggs B, Schmid C. Human detection using oriented histograms of flow and appearance[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2006: 428-441.
[8] Knopp J, Prasad M, Willems G, et al. Hough transform and 3D SURF for robust three dimensional classification[C]//European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2010: 589-602.
[9] Klaser A, Marszalek M, Schmid C. A spatio-temporal descriptor based on 3d-gradients[C]//BMVC 2008-19th British Machine Vision Conference. British Machine Vision Association, 2008: 275: 1-10.
[10] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
[11] Wu Z, Zhang Y, Yu F, et al. A gpu implementation of googlenet[J]. Tech. Rep., Technical report, 2014: 6.

[12] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
[13] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.
[14] Farabet C, Couprie C, Najman L, et al. Learning hierarchical features for scene labeling[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8): 1915-1929.
[15] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]//Advances in neural information processing systems. 2014: 568-576.
[16] Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 2625-2634.
[17] Zhang L, Feng Y, Xiang X, et al. Realistic human action recognition: When CNNS meet LDS[C]//Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017: 1622-1626.
[18] Chron G, Laptev I, Schmid C. P-cnn: Pose-based cnn features for action recognition[C]//Proceedings of the IEEE international conference on computer vision. 2015: 3218-3226.
[19] Reddy K K, Shah M. Recognizing 50 human action categories of web videos[J]. Machine Vision and Applications, 2013, 24(5): 971-981.
[20] Liu J, Luo J, Shah M. Recognizing realistic actions from videos in the wild [C]//Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on. IEEE, 2009: 1996-2003.
[21] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.
[22] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
[23] Wellner P D. Adaptive thresholding for the DigitalDesk[J]. Xerox, EPC1993-110, 1993: 1-19.
[24] Five video classification methods implemented in keras and tensorflow. [Online]. Available: https://github.com/harvitronix/five-video-classification-methods
[25] Le Q V, Zou W Y, Yeung S Y, et al. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis[C]//Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011: 3361-3368.
[26] Oshin O, Gilbert A, Bowden R. Capturing relative motion and finding modes for action recognition in the wild[J]. Computer Vision and Image Understanding, 2014, 125: 155-171.
[27] Wang H, Klaser A, Schmid C, et al. Dense trajectories and motion boundary descriptors for action recognition[J]. International journal of computer vision, 2013, 103(1): 60-79.
[28] Wang H, Klaser A, Schmid C, et al. Action recognition by dense trajectories[C]//Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011: 3169-3176.
[29] Kliper-Gross O, Gurovich Y, Hassner T, et al. Motion interchange patterns for action recognition in unconstrained videos[C]//European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2012: 256-269.
[30] Everts I, Van Gemert J C, Gevers T. Evaluation of color stips for human action recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 2850-2857.
[31] Todorovic S. Human activities as stochastic kronecker graphs[M]//Computer Vision-ECCV 2012. Springer, Berlin, Heidelberg, 2012: 130-143.
[32] Solmaz B, Assari S M, Shah M. Classifying web videos using a global video descriptor[J]. Machine vision and applications, 2013, 24(7): 1473-1485.