Action Recognition using Convolutional Neural Networks with Joint Supervision

Yupeng Li*, Yuxiao Wang*, Yongfeng Jiang† and Liang Zhang*
*Civil Aviation University of China, Tianjin, China
E-mail: yupengli666@126.com Tel: +8617320091310
†Public Security Bureau, Wenzhou, China
E-mail:13968823088@163.com Tel: +8613968823088

***Abstract*— Mapping the depth video into an optimally representation in two-dimensional space are of vital importance for depth video based human action understanding. Meanwhile, such representation will lost some useful information inevitably, a feature learning approach not only separable but also discriminative are essential for action recognition task from such representation. This paper presents a new method for action recognition base on convolutional neural networks with joint supervision which shares the merits of both representation as mentioned above and convolutional neural networks. The advantages of our method come from (*i*) The whole procedure of our method is done automatically no matter the generation of representation or deeply feature learned; (*ii*) The deeply feature using the proposed deep architectures to learned has high discriminative capacity to improve the accuracy of action recognition effectively compared with handcrafted features. We conduct experiments on two challenging datasets: MSRDailyActivity3D and SYSU 3D HOI. Experimental results show that our method outperform previous methods based on hand-crafted features. Our method also achieves superior performance to the state-of-the-art on these datasets.**

I. INTRODUCTION

Human action recognition in videos attracts increasing research interests in computer vision community due to its potential applications in video surveillance, human computer interaction, and video content analysis [1]. Video-based human action recognition is challenging because significant intra-action variations exist due to changes in viewpoint, illumination, visual appearance (such as color and texture of clothing), scale (due to different human body sizes or distances from the camera), background and speed of performing an action. Some challenges have been simplified by the use of real-time depth cameras (e.g. Kinect) that capture the texture and illumination invariant human body shape and simplify human segmentation [2]. However, action recognition from representation with two-dimensional (2D) space generated from depth video remains a major challenge and is explicitly addressed in this paper.

Most of the progress [3, 4, 7, 17, 18, 19] in the field of action recognition based depth videos over the last decade have been proposed. The performance of these approaches highly depends on the handcrafted features and is limited by the discriminative power of the handcrafted features, which are shallow high-dimensional descriptions of local or global spatio-temporal information and their performance varies from dataset to dataset. The advent of modern learnable representations such as deep convolutional neural networks (CNNs) has improved dramatically the performance of algorithms in many image-understanding tasks, offering state-of-the-art results on image recognition [9], segmentation [10], detection and retrieval [11]. However, there is performance only in color image understanding until [14] that proposed a method called Structured Images, an effective yet simple video representation for RGB-D based action recognition, which is a strategy for transforming the problem of depth videos based action recognition to image classification and making effective use of the rich information offered by the depth videos.

In spite of Structured Images dramatically outperforms existing state-of-the-art in action recognition based on RGB-D. However, it remains unclear how depth videos can be optimally represented that limit the accuracy of action recognition lies higher up. [15] proposed a discriminative feature learning approach for deep face recognition, significantly improving the previous results and setting new state-of-the-art for both face recognition and face verification tasks. Since it is not only separable but also discriminative deeply learned features which is need to provide for face recognition task.

We suppose depth videos based human action recognition is also need a discriminative deeply learned features to address the classification problems after dimension reduction by mapping the depth videos into 2D space such as Structured Images. Motivated by the above analysis, this paper proposes a depth videos based action recognition method from a novel perspective that try to exploit discriminative information quite adequately from Structured Images, as illustrated in Figure 1. To achieve this goal, we proposed a CNNs based deep architecture with joint supervision to learn discriminative feature from Structured Images. We share the merits of both representation Structured Images and CNNs. Specifically, the CNNs are trained under the joint supervision of the softmax loss and center loss, with a hyperparameter to balance the two supervision signals, to learn discriminative convolutional feature from Structured Images.

We evaluate our method on the MSRDailyActivity3D and SYSU 3D HOI datasets individually and achieve results

which are better than the state-of-the-art, and comparison with state-of-the-art show that our method achieves 1.88% and 1.66% higher accuracies respectively than the nearest competitor.

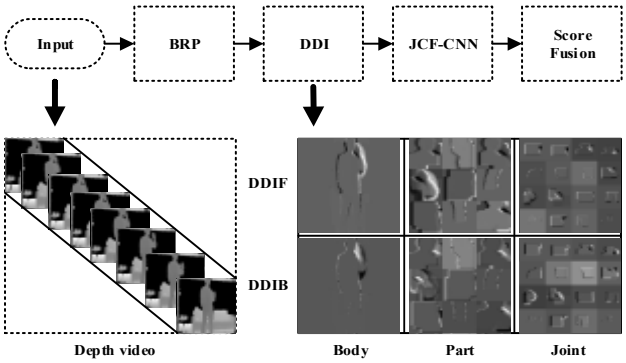

Fig. 1 The process diagram of our method.

II. RELATED WORKS

With the resurgence of neural networks invoked by Hinton and others [21], deep neural architectures have been used as an effective solution for extracting high level features from data. There are a number of attempts to apply 2D deep architectures for video recognition. In [22], spatio-temporal features are leaned unsupervised by a Convolutional Restricted Boltzmann Machine and then plugged into a CNNs for action recognition. In [23], 3D convolutional network is used to automatically learn spatio-temporal features directly from raw data. Recently, several CNNs architectures for action recognition in [24] is compared based on Sport-1M dataset, comprising 1.1 M YouTube videos of sports activities. They find that for a network, operating on individual video frames, performs similarly to the networks whose input is the stack of frames, which indicates that the learned spatiotemporal features do not capture the motion effectively. In [25], spatial and temporal streams are proposed for action recognition. Two CNNs are trained on the two streams and combined by late fusion. The spatial stream is comprised of individual frames while the temporal stream is stacked by optical flow. However, the best results of all above deep learning methods can only match the state-of-the-art results achieved by handcrafted features.

For depth videos based action recognition, many works have been reported in the past few years. Li et al. [3] sample points from silhouette of a depth image to obtain a bag of 3D points which are clustered to enable recognition. Yang et al. [26] stack differences between projected depth maps as DMM and then use HOG to extract the features on the DMM. This method transforms the problem of action recognition from 3D space to 2D space. In [5], HON4D is proposed, in which surface normal is extended to 4D space and quantized by regular polychorons. Following this method, Yang and Tian [6] cluster hypersurface normals and form the polynormal which can be used to jointly capture the local motion and geometry information. Super Normal Vector (SNV) is generated by aggregating the low-level polynormals. However, all of these methods are based on carefully hand-designed features, which are restricted to specific datasets and applications.

Our work is inspired by [14] and [15], where we propose a new method for action recognition using CNNs with joint supervision which shares the merits of both representation Structured Images and CNNs. Since we can take advantage of the rich experience of design 2D CNNs for a long term and pre-trained ImageNet models. At the same time, we can automatically complete the process of human action recognition whether in the generation of Structured Images or the proceed of CNNs training and its verification, rather than carefully constructed the hand-crafted feature.

III. STRUCTURED IMAGES REVISITED

As shown in Figure 1, our proposed method is based on 2D representation for dimension reduction and we choose Spatially Structured Dynamic Depth Images [14], referred to as Structured Images. In this section, we briefly review the map process of Structured Images. It is worth noting that our method is independent of the method of mapping depth videos into images in 2D space, and we use Structured Images due to its good performance.

Firstly, three sets of Depth Dynamic Images (DDIs) is processed hierarchically at three spatial levels guided by skeleton, namely, joint level, part level and body level, which are constructed from an image sequence through bidirectional rank pooling [12, 13]. This representation aggregates motion and structure information from global to fine-grained levels for action recognition. Each set of dynamic images is represented by two dynamic images, forward and backward(refer to as DDIF and DDIB, respectively).

A. Rank pooling

Given a sequence with k frames, which can represented as $X = < x_1, x_2, ..., x_t, ..., x_k >$ .And $\varphi(x_t) \in R^d$ be a representation or feature vector extracted from each frame x_t . Let $V_t = \frac{1}{t}\Sigma_{\tau=1}^{t}\varphi(x_t)$ be time average of these features up to time t . At each time t , a score $r_t = w^T \bullet V_t$ is assigned. In general, the later the time, the higher the score, so the score satisfies $r_i > r_j \Leftrightarrow i > j$ .The process of rank pooling is to find w^* that satisfies the following objective function:

$$\underset{w}{argmin}\ \frac{1}{2}\|w\|^2 + \lambda \sum_{i>j} \varepsilon_{ij} \qquad (1)$$
$$s.t.\ w^T \bullet (V_i - V_j) \geq 1 - \varepsilon_{ij}, \varepsilon_{ij} \geq 0$$

The parameters w^* represent the information that frame representation v_t comes before the frame representation v_{t+1} , and can be used as a descriptor of the sequence. ε_{ij} is the smallest non-negative number and λ is scalar coefficients.

In fact, accurate optimization of eq. (1) has a disadvantage: optimization is slow. We adopt rank pooling for the task of DDIs too, but make a modification. Unlike [14], we propose an approximation to rank pooling which is much faster and

works as well in practice. An alternative construction of the rank pooling does not compute the intermediate average features $V_t = \frac{1}{t}\Sigma_{\tau=1}^{t}\varphi(x_t)$, but uses directly individual video features $\varphi(x_t)$ in the definition of the ranking scores (1). In this case, the derivation above results in a weighting function which is linear in t .

B. Structured Images

Since in rank pooling the averaged feature up to time *t* is used to classify frame *t*, the pooled feature is biased towards beginning frames of the depth sequence, hence, frames at the beginning has more influence to w^* .To overcome these drawbacks, the rank pooling is applied in a bidirectional way to convert one video sequence into two dynamic images. DDIs are constructed from depth sequence. Each dynamic image is fed into a CNNs. When bidirectional rank pooling is applied to a sequence of depth maps, two DDIs, DDIF and DDIB, are generated. The resulting DDIs are also illustrated in Figure 1. As shown, DDIs effectively capture the spatiotemporal information.

IV. CNNs WITH JOINT SUPERVISION

In this section, we describe a deep architecture based on CNNs for depth videos based action recognition, which shares the benefits of both Structured Images and CNNs. We first introduce the deep architectures of CNNs with joint supervision we used. Then, we show how to adapt the model trained on large datasets ImageNet to train our networks. Finally, based on trained model and Structured Images, we describe the details of how to calculate score fusion.

A. CNNs with joint supervision

Our networks start with designing deep architecture based on CNNs for feature learning and label prediction, Map the input video to the deep feature of the last hidden output, and then map to the prediction labels from images. The networks in our method contain three separate CNNs, namely body nets, part nets and joint nets. Each net as mentioned above are designed with the same architecture used for Structured Images, in hierarchically at three spatial levels, joint level, part level and body level, respectively. We aggregate motion and structure features from global to fine-grained levels for action recognition by the three deep architectures.

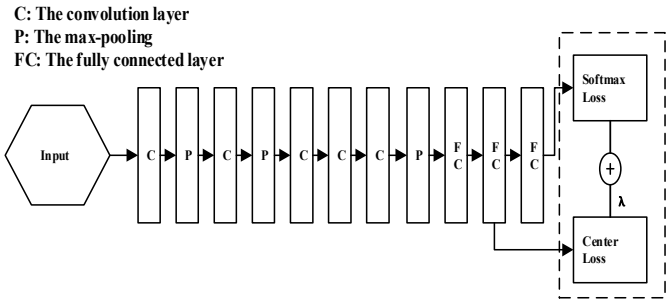

Fig. 2 Deep architecture of we proposed.

Table 1 Layer configuration of our three nets

Layer	C1	C2	C3	C4	C5	FC1	FC2	FC3
numb	96	256	384	384	256	4096	4096	1000
filter	11^2	5^2	3^2	3^2	3^2			
stride	4	1	1	1	1			
pad	0	2	1	1	1			

The details about our networks are schematically shown in Figure 2, following [8]: each net contains eight layers with weights, the first five convolutional layers and the remaining three fully-connected layers. We used this architecture due to transfer knowledge from similarly works is extremely convenient and economic, without training the networks from sketch. The training details can be found in following subsection. The layer configuration of our three nets is shown in Table 1.

As shown in Figure 2, different from other existing method only used softmax loss to guide the training process of networks. In our deep architecture, the neural network is trained under the joint monitoring of soft maximum loss and center loss, and the two monitoring signals are balanced by hyperparameters. The center loss is connected to the penultimate fully-connected layers according to our repeated trial experience. Intuitively, the softmax loss forces different classes of deep features to stay separate , center loss effectively pulls similar deep features to their center. By combining the center loss with the softmax loss to jointly supervise the learning of CNNs, the discriminative power of the deeply learned features can be highly enhanced for robust action recognition. The softmax loss function is presented as

$$L_S = -\sum_{i=1}^{m}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \qquad (2)$$

And the center loss function, formulated as

$$L_C = \frac{1}{2}\sum_{i=1}^{m}\left\|x_i - c_{y_i}\right\|_2^2 \qquad (3)$$

The $x_i \in R^d$ denotes the i th deep feature, belonging to the y_i th class. d is the feature dimension. The size of mini-batch and the number of class is m and n , respectively. $W_j \in R^d$ denotes the j th column of the weights $W \in R^{d\times n}$ in the last fully-connected layer and $b \in R^n$ is the bias term. In fact, the performance is nearly of no difference without bias term. Thus, we omit the biases for simplifying analysis. In Eq.3, The $c_{y_i} \in R^d$ denotes the y_i th class center of deep features, which can be computed by

$$\frac{\partial L_C}{\partial x_i} = x_i - c_{y_i} \qquad (4)$$

$$\Delta c_{y_i} = \frac{\alpha}{2}\sum_{i=1}^{m}(c_{y_i} - x_i) \qquad (5)$$

Where the α is a parameter restricted in $[0,1]$. In order to adopt the joint supervision of softmax loss and center loss to train the deep architecture for discriminative feature learning. The formulation can be presented as

$$L = L_S + \lambda L_C = -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} + \frac{\lambda}{2} \sum_{i=1}^{m} \left\| x_i - c_{y_i} \right\|_2^2 \quad (6)$$

Where λ is a hyperparameter used for balancing the two loss functions and affect results of the accuracy of recognition when it is changing. The conventional softmax loss can be considered as a special case of this joint supervision while λ is fit to zero.

B. Network Training

After we construct the Structured Images from depth videos and complete designs of networks, three nets are trained on the images of the three hierarchically spatial levels.

The implementation is derived by Caffe toolbox which was based on the NVIDIA Quadro P2000 card [20]. The training process is similar to [8]: mini-batch random gradient descent learning is adopted for network weights, momentum is set at 0.9, and weight attenuation is set at 0.0005. All the hidden weights use rectifier activation function; In each iteration, 256 transformed training images are sampled to construct a mini-batch of 256 samples and adjust the size of all images to 256×256; For artificially expand the training data (data increase), firstly, 224×224 patch from the selected center randomly cropped images to enhance the data by 2048 times, then random horizontal flip, but we do a small modification, no RGB jitter is initially because the Structured Image is robust noise; The network was trained with the pre-training model of ILSVRC-2012 and set the learning rate to 10^{-3}. For each network, we train 100 loops and slow down the learning rate every 20 loops. For all experimental settings, we set the dropout regularization ratio to 0.5 for reduce the complex cooperative adaptation of neurons in the network. For joint supervision, the proprietary hyperparameters α and λ are set to 0.5 and 0.003 respectively in experiential.

C. Score Fusion

During the test, we given a depth videos, after constructed the three hierarchically images, we only use Structured Images with 224 × 224 patches cropped from the center but without other data augmentation operation. We only adopt average score fusion method, the simplest fusion method compared with other two commonly used late score fusion methods (multiply and maximum score fusion), to improve the final accuracy. The average scores for each test sample are calculated from each of the three nets. The final class score for a test sample is the average of the outputs from each level of Structured Images. Thus

$$score_{test} = \frac{1}{6} \bullet \sum_{i=1}^{3} \sum_{j=1}^{2} score_j^i \quad (7)$$

Where $score_{test}$ represents the final class score for a test sample, while $score_j^i$ denotes the score of j th test sample for i th level of Structured Images.

V. EXPERIMENTS

In this section, we evaluated our proposed method on two datasets involve human-object interactions. The former is MSRDailyActivity3D Dataset [4] whereas the latter is SYSU 3D HOI Dataset [16]. We firstly generate Structured Images from all depth videos includes training and testing samples for the following procedures. CNNs and Structured Images, in all proceeding, are setting the same configuration for the two datasets provide evidence that the powerful generalization ability of our deep architectures can work. Detail of network training have describe in section 4.2, but for testing, the hyperparameteris set to zero due to we have learn the discriminative deeply features in the training process.

The MSRDailyActivity3D Dataset is a daily activity dataset which was captured by a depth camera. This dataset contains 16 classes of actions: “drink”, “eat”, “read book”, “call cellphone”, “write on paper”, “use laptop”, “use vacuum cleaner”, “cheer up”, “sit still”, “toss paper”, “play game”, “lay down on sofa”, “walking”, “play guitar”, “stand up” and “sit down”. It has 10 actors and each actor performs each activity twice, one in stand-up position and the other in sit down position. Actors in this dataset present large spatial and scaling changes. Moreover, most activities in this dataset involve human-object interactions. For this dataset, we follow the same experimental setting as [4] and obtain the final accuracy of 99.38%. The performance of our method compared to the previous approaches is shown in Table 3.

Table 2 Comparison on different training setting for MSRDailyActivity3D Dataset

Method	body	part	joint	fusion
Removed center loss	62.50%	**92.50%**	93.13%	96.88%
Normal condition	**65.63%**	90.63%	**93.75%**	**99.38%**

Table 3 Recognition accuracy comparison of our method and previous approaches on MSRDailyActivity3D

Method	Accuracy
IPM [17]	83.30%
SNV[6]	86.25%
DS+DCP+DDP+JOULE-SVM[16]	95.00%
Range Sample[7]	95.63%
MFSK+BoVW[18]	95.70%
SSDDI[14]	97.50%
Our method	**99.38%**

To highlight the ability in improving the accuracy of action recognition with joint supervision in the deep architectures, we considered another scenario for this dataset, where is set to zero while training the nets. That is to say, the joint supervision is degeneration to softmax loss alone without center loss. The results are listed in Table 2, which we can see that the accuracy of action recognition improved greatly compared with the networks without using joint supervision.

The reference experiments show that the method we proposed is effective.

Table 4 Comparison of the proposed method with previous approaches on SYSU 3D HOI Dataset

Method	Accuracy
HON4D[5]	79.22%
DS+DCP+DDP+MTDA[19]	84.21%
DS+DCP+DDP+JOULE-SVM[16]	84.89%
SSDDI[14]	95.42%
Our method	**97.08%**

The SYSU 3D HOI Dataset includes 480 depth video clips contains 12 different activities performed by 40 subjects. For each activity, each participant manipulates one of the six different objects: phone, chair, bag, wallet, mop and besom. Although each video clip corresponding RGB frames, depth sequence and skeleton data, we only use the latter two. It is challenging to our method due to the dataset was focus on human-object interactions. We follow the data protocol as [16] and report the results in Table 4, where we can see that our method obtains the final accuracy of 97.08%.

From the Table 3 and 4 we can see that our proposed method can outperform SSDDI [14] greatly and can over the state-of-the-art methods. The reasons probably are: (1) there are so many actions that are similar, such as call cellphone, drink and eat, they have similar motion shapes but have subtle motion so that the representations in 2D are very similar and confusing, which limited the higher accuracy of recognition on many existing methods; (2) the assumption we suppose in the beginning of this paper maybe right that the deeply learned features need to be not only separable but also discriminative for action recognition task; (3) training the networks with initialising the weights using the pre-trained models is profitable.

VI. CONCLUSIONS

In this paper, a method for action recognition using convolutional neural networks with joint supervision from depth videos has been proposed. The method has been evaluated on the most widely used datasets and compared with state-of-the-art methods. The proposed method achieved state-of-the-art results on individual datasets. The way of action recognition is done automatic in the whole process without careful handwork. The experimental results showed that with the joint supervision by jointly using the center loss and the softmax loss, the highly discriminative features can be obtained for robust action recognition. The experimental results have also showed that the strategies developed for applying CNNs to small datasets worked effectively. In our future work, we will combine the proposed method together with point cloud to improve the recognition accuracy.

ACKNOWLEDGMENT

Thanks to the support of the National Natural Science Foundation of China (61179045) and the Civil Aviation Security Important Projects of China (20600523).

REFERENCES

[1] Aggarwal J K, Ryoo M S. Human activity analysis: A review. ACM Computing Surveys, 2011, 43(3).
[2] Shotton J, Sharp T, Kipman A A, et al. Real-time human pose recognition in parts from single depth images, Communications of the ACM, 2013, 56(1): 116-124.
[3] Li W, Zhang Z, Liu Z. Action recognition based on a bag of 3D points. Computer Vision and Pattern Recognition Workshops. IEEE, 2010:9-14.
[4] Wu Y. Mining actionlet ensemble for action recognition with depth cameras. IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2012:1290-1297.
[5] Oreifej O, Liu Z. HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences. Computer Vision and Pattern Recognition. IEEE, 2013:716-723.
[6] Yang X, Tian Y L. Super Normal Vector for Activity Recognition Using Depth Sequences. Computer Vision and Pattern Recognition. IEEE, 2014:804-811.
[7] Lu C, Jia J, Tang C K. Range-Sample Depth Feature for Action Recognition. Computer Vision and Pattern Recognition. IEEE, 2014:772-779.
[8] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.
[9] Zhang N, Paluri M, Ranzato M, et al. PANDA: Pose Aligned Networks for Deep Attribute Modeling. Computer Vision and Pattern Recognition. IEEE, 2014:1637-1644.
[10] Oquab M, Bottou L, Laptev I, et al. Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks. IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2014:1717-1724.
[11] Girshick R B, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. Computer Vision and Pattern Recognition. IEEE, 2014: 580-587.
[12] Fernando B, Gavves E, Oramas M J, et al. Modeling video evolution for action recognition. Computer Vision and Pattern Recognition. IEEE, 2015:5378-5387.
[13] Fernando B, Anderson P, Hutter M, et al. Discriminative Hierarchical Rank Pooling for Activity Recognition. Computer Vision and Pattern Recognition. IEEE, 2016:1924-1932.
[14] Wang P, Wang S, Gao Z, et al. Structured Images for RGB-D Action Recognition. IEEE International Conference on Computer Vision Workshop. IEEE Computer Society, 2017:1005-1014.
[15] Wen Y, Zhang K, Li Z, et al. A Discriminative Feature Learning Approach for Deep Face Recognition. Computer Vision – ECCV 2016. Springer International Publishing, 2016:499-515.
[16] Hu J F, Zheng W S, Lai J, et al. Jointly learning heterogeneous features for RGB-D activity recognition. Computer Vision and Pattern Recognition. IEEE, 2015:5344-5352.
[17] Zhou Y, Ni B, Hong R, et al. Interaction part mining: A mid-level approach for fine-grained action recognition. Computer Vision and Pattern Recognition. IEEE, 2016:3323-3331.
[18] Wan J, Guo G, Li S. Explore Efficient Local Features from RGB-D Data for One-shot Learning Gesture Recognition. IEEE

Transactions on Pattern Analysis & Machine Intelligence,2016,38(8):1626-1639.
[19] Zhang Y, Yeung D Y. Multi-Task Learning in Heterogeneous Feature Spaces. AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, Usa, August. DBLP, 2011:41-75.
[20] Jia, Yangqing, Shelhamer, et al. Caffe: Convolutional Architecture for Fast Feature Embedding. Proceedings of the 22nd ACM international conference on Multimedia, 2014:675-678.
[21] Hinton G E, Osindero S, Teh Y W, et al. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18(7): 1527-1554.
[22] Taylor G W, Fergus R, Lecun Y, et al. Convolutional learning of spatio-temporal features. European conference on computer vision, 2010: 140-153.
[23] Ji S, Xu W, Yang M, et al. 3D Convolutional Neural Networks for Human Action Recognitiocn. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231.
[24] Karpathy A, Toderici G, Shetty S, et al. Large-Scale Video Classification with Convolutional Neural Networks. Computer vision and pattern recognition, 2014: 1725-1732.
[25] Simonyan K, Zisserman A. Two-Stream Convolutional Networks for Action Recognition in Videos. neural information processing systems, 2014: 568-576.
[26] Yang X, Zhang C, Tian Y, et al. Recognizing actions using depth motion maps-based histograms of oriented gradients. Acm multimedia, 2012: 1057-1060.