*I. J. Mathematical Sciences and Computing,* **2020, 6, 15-23**
Published Online December 2020 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmsc.2020.06.03

# Action Recognition Based on the Modified Two-stream CNN

**Dan zheng [a], Hang Li [a,*], and Shoulin Yin [a,*]**
[a] Software College, Shenyang Normal University, Shenyang 110034, China
Corresponding Author: lihangsoft@163.com;yslinhit@163.com

Received: 20 October 2020; Accepted: 03 November 2020; Published: 08 December 2020

**Abstract:** Human action recognition is an important research direction in computer vision areas. Its main content is to simulate human brain to analyze and recognize human action in video. It usually includes individual actions, interactions between people and the external environment. Space-time dual-channel neural network can represent the features of video from both spatial and temporal perspectives. Compared with other neural network models, it has more advantages in human action recognition. In this paper, a action recognition method based on improved space-time two-channel convolutional neural network is proposed. First, the video is divided into several equal length non-overlapping segments, and a frame image representing the static feature of the video and a stacked optical flow image representing the motion feature are sampled at random part from each segment. Then these two kinds of images are input into the spatial domain and the temporal domain convolutional neural network respectively for feature extraction, and then the segmented features of each video are fused in the two channels respectively to obtain the category prediction features of the spatial domain and the temporal domain. Finally, the video action recognition results are obtained by integrating the predictive features of the two channels. Through experiments, various data enhancement methods and transfer learning schemes are discussed to solve the over-fitting problem caused by insufficient training samples, and the effects of different segmental number, pre-training network, segmental feature fusion scheme and dual-channel integration strategy on action recognition performance are analyzed. The experiment results show that the proposed model can better learn the human action features in a complex video and better recognize the action.

**Index Terms:** Action recognition, dual-channel, convolutional neural network.

## 1. Introduction

When human beings get information from the outside world, visual information accounts for 80% of the total information obtained by various organs. This information is of great significance for understanding the nature of things. With the rapid development of mobile Internet and electronic technology, mobile phones and other video capture devices have become popular in large Numbers, and Internet short video applications have mushroomed like mushrooms, greatly reducing the cost of video shooting and sharing, which leads to the explosive growth of online video resources. These resources enrich people's life, but because of their huge amount, variety and content, how to conduct intelligent analysis, understanding and recognition of these video data has become an urgent challenge [1-5].

Human action recognition is an important research direction in the field of computer vision. The major research objectives are to simulate human brain to analyze and recognize human action in videos, which usually includes individual actions of human beings, interactions between human beings and the outside world and environment.

In the traditional action recognition methods based on artificial design features, the early features based on human body geometry or action information are only suitable for the recognition of simple human body movements in simple scenes, while the spatio-temporal interest points method is more effective in the case of relatively complex background. In this way, the interest points or dense sampling points in space-time in the video are obtained first, and local characteristics are calculated based on the space-time chunks around these points. In this way, the characteristic vector describing the video action is eventually formed by using the classic feature encoding methods such as Bag of Features (BoF), VLAD (Vector of Locally Aggregated Descriptors) or Fisher Vector [6-8].

Currently, in the local feature-based approach, the action identification method based on Dense Trajectory (DT) has obtained better identification results in many public real scene action databases. They obtain the Dense Trajectory by tracking the dense sampling points in each frame of the video, and then calculate the Trajectory characteristics to describe the action in the video.

For example, Cai [9] used multi-view super vector (MVSV) as global descriptor to code the feature of Dense Trajectory. Wang [10] improved setter Trajectory (IDT) feature using FV encoding. Peng [11] used Bag of Visual Words, (BoVW) to code space-time point of interest or features of improved dense trajectory characteristic. Based on dense trajectory

 

characteristics, Wang [12] proposed a multistage video representation model MoFAP (Motion Features, Atoms, and Phrases), which could represent the visual information in a hierarchical manner. Dense trajectories can extract actional features with wider coverage and finer granularity, but there is usually a large number of trajectory redundancy which limits the recognition effect.

Along with the deep learning successfully used in the field of speech and image recognition and so on, especially the Convolutional neural network (CNN), a variety of human action recognition methods based on deep learning framework have emerged. When the training samples are large enough, the semantic features can be learned through deep network, which is more suitable for the recognition of target and action. Karpathy [13] trained deep network DeepNet using slow fusion model to merge different image frame characteristics in the video. However, the model cannot extract the motion information of the video, so the effect is not ideal. Neural Network space-time characteristics of the depth online learning. In the network to avoid processing under the condition of light flow to obtain the motion characteristics of the video, but the time domain information extraction ability is limited, for a long time of complex human action recognition effect is not obvious. Varol used three dimensional space-time characteristics of convolution, further enhanced the action recognition effect.

Simonyan [14] first proposed to use Two data flow (Two-stream) to identify the video action of the convolutional neural network, the airspace of the network input data stream was a static frame, the time domain network input data stream was characterized the light of the interframe motion flow, each data flow depth using convolution neural network for feature extraction and motion prediction, finally the fusion of Two data flow was used to identify the final action of the results. The recognition performance of the model is similar to that of the improved dense trajectory method. Ng [15] added the long and short term memory network to the original two-channel model to strengthen the connection of time-domain information. The convolutional network layers used in the original two-channel model were relatively shallow, so Wang et al. [12] proposed to adopt a pre-training deep network model with better performance in image classification tasks, such as VGGnet, Googlenet, which enhanced the learning and modeling capabilities of video motion features. Wang used dual channel convolution characteristic figure neural network learning, and the convolution Trajectory constraints to obtain the Trajectory pooled Deep-eonvolutional Descriptors (TDD), then used FV coding to get video presentation.

Space-time dual-channel neural network can represent the features of video from both spatial and temporal perspectives. Compared with other neural network models, it has more advantages in human action recognition. Based on video segmentation, this paper uses space-time dual-channel neural network to extract frame image features of airspace and time-domain motion features, and fuses the recognition results of the airspace and time-domain of each segment to obtain the action recognition and classification of the whole video.

## 2. Methodology

### *2.1 The framework of proposed model*

In the original two-channel method, a single frame is randomly sampled from a video for action recognition. For complex actions or videos with long duration, perspective transformation and background perturbation will lead to the inability to effectively express the category information of a video using only a single frame. In order to establish an effective recognition model for long-duration complex videos, space-time dual-channel neural network is applied in this paper based on video segmentation. First, the video is divided into several equal length non-overlapping segments, and the static frame image and stacked optical flow image containing motion information are obtained by random sampling in each segment, and then input into spatial domain and temporal domain CNN respectively for feature extraction. Then the network output prediction features of each segment are fused in their respective channels. Finally, the predictive features of the two channels are integrated to obtain the final action recognition results.

Divide the video into K segments $\{S_1, S_2, \cdots, S_K\}$ of equal length according to the length. The action recognition of space-time dual-channel convolutional neural network Y based on segmentation can be expressed as:

$$Y(S_1, S_2, \cdots, S_K) = H(g(F(T_1;W), F(T_2;W), \cdots, F(T_K;W))) \quad (1)$$

Where $T_i$ represents the random sampling of the i-th segment of the video. In the spatial domain, it is RGB frame image, while in the temporal domain; it is stacked optical flow image. $F(T_i, W)$ represents the feature extraction of $T_i$ by the convolutional neural network with parameter W, whose output is the feature vector of the corresponding category number dimension. The piecewise fusion function g means to fuse K piecewise features in a certain way to obtain the spatial or temporal features. Output function H means to classify the recognition results, and Softmax function is generally used to get the probability value of each action category. In addition, the spatial network structure of each video segment is identical and the network weight is Shared. The same is true for time-domain networks.

 

### *2.2 Preprocessing of spatial domain network data*

Spatial network recognizes static RGB frame images sampled from videos. In order to test the impact of different sampling methods on action recognition performance, Splitl training/test segmentation scheme based on UCF101 data set is used to test the top-1 action recognition accuracy rate (that is, the category with the highest probability in network output is the correct identification result). Table 1 lists the identification performance of the three sampling strategies. During network training, an improved version of the GoogleNet convolutional neural network, Inception V3 model, is adopted. It can be seen that the increase in the number of sampling frames does not improve the recognition performance, but increases the data redundancy and computational complexity. Therefore, it is not advisable to conduct intensive sampling on the video. In the experiment of this paper, K video segments of equal length are sampled at random, and one frame of image is sampled for each segment.

Table 1. Comparison of UCF101 data set size and action recognition accuracy under different sampling methods of spatial domain network

| Sampling method | Sampling frames | | Top-1 recognition accuracy/% |
|---|---|---|---|
| | Training set | Testing set | |
| Completely sampling | 1791290 | 695000 | 76.25 |
| 60 frames sampling | 571430 | 226413 | 77.63 |
| 10 frames sampling | 95238 | 33735 | 77.23 |

In order to prevent the over-fitting problem in learning modeling, it usually adopts data enhancement technique, it can not only increase the size of the input data, increasing, the difference of the sample can also enhance the generalization ability of the network model. In the airspace in the network, in this paper, the video frame using the flip horizontal, Angle of rotation, translation, transformation, data increase, such as strong shear transformation method, and tested the method on InceptionV3 network model of action recognition performance. Table 2 lists the recognition accuracy of top-1 and top-5 in 5 cases. It can be seen that in the absence of any of the data enhancement technologies, the recognition accuracy decreases, which indicates the effectiveness of the data enhancement methods. Therefore, all four data enhancement technologies are adopted in the experiment of this paper.

Table 2. Comparison of action recognition accuracy under different data enhancement techniques for spatial domain/%

| Data enhancement technique | Recognition accuracy | |
|---|---|---|
| | top-1 | top-5 |
| Without flip horizontal | 75.01 | 94.17 |
| Without angle rotation | 73.61 | 91.88 |
| Without horizontal vertical migration | 73.23 | 89.72 |
| Without shear transformation | 72.34 | 90.98 |
| All operation included | 78.34 | 97.28 |

### *2.3 Time domain network data preprocessing*

The motion information in video is very important for action recognition, optical flow is a simple and practical way to express the motion information of image sequence, and it is widely used to extract the motion characteristics of action. Based on two basic assumptions derived the calculation formula of optical flow image sequence, this paper uses the method to calculate the horizontal and vertical two direction of flow. Because of the light flow value are close to zero and there is a negative, in order to be able to as input time domain network channels, need to be linear transformation, will eventually flow in two directions is saved as a two gray image, as shown in figure 1. In order to effectively extract the video motion information, this paper uses the 10 consecutive frames of horizontal and vertical flow stack form 20 intensive light flow images.

 

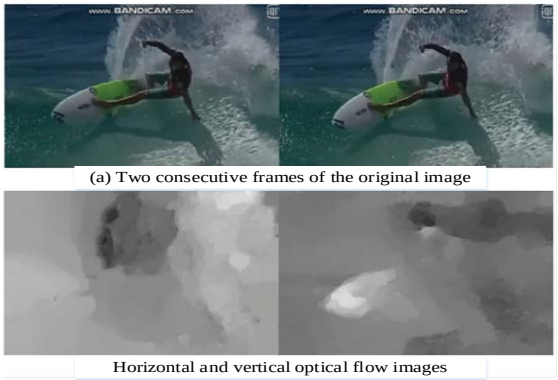


Fig. 1. Two consecutive video frames and corresponding optical flow images

Space and time domain are commonly used in advance on the ImageNet training of CNN, the network's input is RGB images, so the first channel convolution layer number is 3, 20 optical flow image but time domain network input, and the first channel convolution layer number do not match, it adopts the method of cross modal cross training in advance, will be the first three channel convolution layer has a weight average, 20 copies of it as the first convolution in the time domain network weights 20 channels. The weights of the other layers in the time-domain network are the same as those of the corresponding layers in the special network.

### *2.4 Transfer learning*

In machine learning methods, sufficient training samples [16-19] are needed to learn a good classification model. However, in practice, the existing samples for the target task are usually small in size, and manual labeling of a large number of samples is not only time-consuming and laborious, but also affected by subjective factors of the annotator. The transfer learning method can use the pre-training model to solve the problem of insufficient target task data. For the new target task, the last fully connected layer in the pre-training network model for classification needs to be replaced by a new fully connected layer for the number of target task categories. In this paper, the residual network model Resnet50/101, which is pre-trained on ImageNet, is used to conduct action recognition for UCF101 data set. During transfer learning, the last full connection layer should be set as the corresponding output of class 101.

Two transfer learning schemes are compared in the experiment. One is to update the weights of only the last classification layer of the convolutional neural network. The other is to fine-tune the entire network to update the ownership value. The recognition accuracy of the two schemes is shown in Table 3. It can be seen that the scheme using fine-tuning the entire network can achieve better identification performance, and the accuracy of top-1 and top-5 is higher than that of the scheme only fine-tuning the last layer. Therefore, this paper adopts the fine-tuning transfer learning scheme of the entire network in the experiment.

Table 3. Comparison of spatial CNN action recognition accuracy under different transfer learning schemes

| Network model | Transfer learning scheme | Recognition accuracy/% | |
|---|---|---|---|
| | | Top-1 | Top-5 |
| ResNet101 | Fine tune the last layer | 70.32 | 90.51 |
| | Fine tune the entire network | 80.97 | 95.62 |
| ResNet50 | Fine tune the last layer | 70.13 | 90.32 |
| | Fine tune the entire network | 78.98 | 94.68 |

### *2.5 Single channel segmentation feature fusion*

The spatio-temporal dual-channel model based on video segmentation contains independent spatial and temporal convolutional neural networks, and the two networks have identical parameters in other layers except that the number of input channels in the first layer is different. Single channel segmentation fusion means that the network output of each video segment is fused in a certain way in a single channel to obtain the action recognition result of the channel. In this paper, three segmentation feature fusion schemes based on maximum, mean and variance are designed.

After the transfer learning, the dimension of the eigenvector output by the last full connection layer of the network corresponds to the number of categories, and the larger the eigenvalue, the greater the possibility of the categories. Maximum segmentation feature fusion refers to taking the maximum value of the output feature value of the corresponding

 

category of all segments as the feature output of that category. This is a fusion strategy of taking the most likely mode for each category. Mean segmental feature fusion refers to taking the average of the output eigenvalues of the corresponding categories of all segments. This strategy treats the actional information in each segment equally. Segmentation based on variance fusion strategy is according to the characteristics of the block output variance to distinguish the importance of segmentation, the variance is larger, the corresponding output characteristics of discrete degree is bigger, that was to be significant to identify the action of the category, the feature recognition of contribution to the action of the video should be high, so the weight given to segment the larger. On the contrary, the small variance of piecewise output features indicates that the output features have a small degree of dispersion, low recognition degree and importance for action recognition, and low weight during fusion.

### *2.6 Double channel feature fusion*

The spatial and temporal CNNS in the dual-channel model are independent of each other, and the recognition results of the spatial and temporal domains need to be fused after the segmented features of the respective channels are fused. Based on the idea of integration learning [17], two space-time feature integration schemes, trial-and-error integration and variance integration, are discussed in this paper to further improve the recognition performance.

The design fusion method is to set the weighted coefficient $\theta_1$ and $\theta_2$ square after piecewise fusion of the spatial and temporal characteristics of the weighted sum to obtain the dual-channel output characteristics, and finally take the corresponding category of the maximum eigenvalue as the recognition result. Generally speaking, the motion information in the time domain is more important to the action recognition, so a large weight can be set. The variance integration method uses the variance of the fused spatial and temporal eigenvectors as the weighted coefficient to distinguish the importance of the two channels.

## 3. Results and analysis

The experiment in this paper is carried out in Linux system based on Pytorch0.3.0 deep learning framework. The basic parameter Settings for the two-channel network are shown in Table 4, including the initial learning rate, batch-size size, and momentum. In this paper, a pre-trained network model is used to recognize the action of UCF101 data set. A smaller learning rate will be beneficial to network training. The initial learning rate of airspace network is set as 0.0005. Since the input data of time-domain network is optical flow image, there is a certain difference between it and RGB image. The relatively large initial learning rate is conducive to the rapid convergence of the network, which is set as 0.01 in the experiment.

The learning rate in optimization is adaptive, and the learning rate is automatically updated according to the learning results. The Batch-size is set to 32 in terms of memory capacity, usage, and convergence speed. In order to effectively accelerate network convergence and momentum set to follow the traditional dual channel action recognition method is set to 0.9, airspace and time domain network training adopts cross entropy loss function as the optimization objective function, optimization method for stochastic gradient descent algorithm. UCF101 data set, the training set contains 9537 video, test set containing 3783 video. Each cycle training need 300 iterations, each iteration was randomly selected 32 video as the training sample, every sample using the foregoing data enhancement method was cut to the size of the network input after 224×224. The test set was tested after each round of training to verify the performance of the learning model, following the THUMOS13 challenge mechanism.

Table 4. Two-stream CNN parameter setting

| Channel | Initial learning rate | Batch-size | Momentum |
|---|---|---|---|
| Space domain | $5e^{-4}$ | 32 | 0.9 |
| Time domain | $1e^{-2}$ | 32 | 0.9 |

### *3.1 Action recognition performance analysis under different number of segments*

In order to model the long time video effectively, this paper divides the video into K equal length segments. When the number of segments is small, actional information extraction is insufficient and the training model is too simple. The large number of segments will lead to data redundancy and increase computation. Table 5 shows the performance of spatial channel action identification under different number of video segments when ResNET 50/101 network is used. It can be seen that when the video is divided into three segments, its action recognition performance is better. Therefore, the number of video segments is set as 3 in subsequent experiments.

 

Table 5. Comparison of spatial CNN action recognition accuracy under different number of video segments

| Network structure | Action recognition accuracy/%(Top-1) | | | | | |
|---|---|---|---|---|---|---|
| | K=1 | K=2 | K=3 | K=4 | K=5 | K=6 |
| ResNet50 | 77.42 | 78.15 | 79.26 | 78.37 | 77.94 | 77.92 |
| ResNet101 | 78.98 | 79.93 | 81.59 | 79.98 | 78.91 | 79.23 |

*3.2 Action recognition performance analysis under different network structures*

ResNet18/50/101 is presented in table 6. The improved version InceptionV3 GoogLeNet network and Inception_ResNet-v2 behavior recognition performance of 5 kinds of channel network in the airspace, given in table 7 ResNet18/50/101 three network recognition performance in time domain channel behavior. It can be seen from Table 6-7 that, compared with other network structures, ResNet101 achieves the highest behavior recognition accuracy in both the spatial channel and the temporal channel, with the accuracy of top-1 reaching 82.24% and 83.48%, respectively. In addition, it is also seen that the recognition performance of resnet18/50/101 and other three residual networks improves with the increase of network depth, which indicates the importance of depth of convolutional neural network for behavior recognition.

Table 6. Comparison of spatial CNN action recognition accuracy under different network architectures

| Network structure | Action recognition accuracy/% | |
|---|---|---|
| | Top-1 | Top-5 |
| ResNet18 | 73.94 | 93.12 |
| ResNet50 | 79.21 | 95.46 |
| ResNet101 | 83.35 | 96.71 |
| InceptionV3 | 77.53 | 93.72 |
| Inception_ResNet_v2 | 78.75 | 94.27 |

Table 7. Comparison of temporal CNN action recognition accuracy under different network architectures

| Network structure | Action recognition accuracy/% | |
|---|---|---|
| | Top-1 | Top-5 |
| ResNet18 | 76.42 | 93.41 |
| ResNet50 | 77.64 | 94.92 |
| ResNet101 | 83.56 | 96.28 |

*3.3 Behavior recognition performance analysis under different segmentation fusion schemes*

In the experiment, each video was divided into three segments of equal length, and the 101-dimensional feature vector output by the airspace channel represented the spatial behavior identification result of the input segments. As mentioned above, the behavior identification result of the entire airspace channel could be obtained after the 101-dimensional feature fusion of the three segments and Softmax function. The same is true for the time channel. Table 8 and Table 9 show the behavior recognition performance of several network structures under different piecewise fusion schemes. In the experiment, a piecewise fusion scheme based on mean value, maximum value and variance was adopted for ResNet18 residual network in both space-time channels. It can be seen that the mean-based scheme has a better recognition performance, while the maximum-based scheme has a poor overall performance, which may be because the difference of video segmentation content will lead to a large discrimination error. Therefore, the maximum-based segmentation fusion scheme is no longer used for the network structure of ResNet50 and ResNet101. It can be seen that with the increase of network depth, the recognition performance of mean-based fusion scheme is still good. Moreover, considering that the calculation of mean-based fusion scheme is simpler, the average value based on the output features of each segment is more suitable to be used as a piecewise fusion scheme.

 

Table 8. Comparison of spatial CNN action recognition accuracy under different fusion schemes based on segmentation

| Network structure | Piecewise fusion scheme | Action recognition accuracy/% | |
|---|---|---|---|
| | | Top-1 | Top-5 |
| ResNet18 | Mean value | 73.98 | 93.17 |
| | Maximum value | 69.84 | 88.74 |
| | Variance | 70.91 | 91.32 |
| ResNet50 | Mean value | 79.21 | 95.64 |
| | Variance | 78.94 | 94.51 |
| ResNet101 | Mean value | 82.35 | 95.96 |
| | Variance | 80.12 | 95.62 |

Table 9. Comparison of temporal CNN action recognition accuracy under different fusion schemes based on segmentation

| Network structure | Piecewise fusion scheme | Action recognition accuracy/% | |
|---|---|---|---|
| | | Top-1 | Top-5 |
| ResNet18 | Mean value | 76.43 | 93.41 |
| | Maximum value | 67.14 | 86.22 |
| | Variance | 65.51 | 88.56 |

### *3.4 Comparison with other methods*

Table 10 compares the performance of the proposed method with that of some methods based on traditional manual design features and deep learning on the UCF101 behavior recognition data set. In the table, the first four methods use different feature coding methods based on dense trajectory to obtain video-level representation. It can be seen that the recognition accuracy of the method based on manual features reaches the highest 88.3%. The latter 7 methods in the table are based on deep learning. The recognition accuracy of the first DeepNet network using deep learning is only 63.3%, and the accuracy of 3D convolutional neural network 3D-CNN is 85.2%, which is lower than the best manual feature method. The recognition accuracy of the original two-channel model is 88%, the recognition accuracy is 88.6% after adding ISTM cyclic neural network, and the recognition accuracy of deep convolutional neural network is 90.9%. Two-stream + LSTM combined depth characteristics and trajectory characteristics, and the recognition accuracy was 90.3%. In this paper, the space-time dual-channel model based on video segmentation is adopted to model long-time video motion information, and the recognition accuracy is 91.8%, which is 3.8 percentage points higher than the original dual-channel method. This shows that with the application of various network models and learning strategies, the deep learning-based method can achieve better recognition performance than the traditional method.

Table 10. Action recognition accuracy comparison of different methods on dataset UCF101

| Method | Accuracy/% |
|---|---|
| Two-stream | 88.1 |
| Two-stream + LSTM | 88.6 |
| Two-stream( VGGNet-16) | 90.5 |
| Proposed | 96.8 |

## 4. Conclusion

In this paper, a human behavior recognition method based on spatio-temporal two-channel convolutional neural network based on video segmentation is implemented, and the training and testing of recognition and classification are mainly carried out on UCF101 data set based on residual network model. In order to solve the problem of over-fitting caused by insufficient data set samples, the influence of various data enhancement methods on the accuracy of airspace network recognition is discussed and analyzed experimentally. At the same time, because the network needs to be adjusted when using the pre-training network model on ImageNet to classify and identify the target data set, two migration learning schemes are discussed and analyzed. The experiment shows that the global fine-tuning network can achieve greater performance improvement than the last fine-tuning layer only. Of space-time dual channel model based on section, discussed through the experiment analyzes the different number of video segmentation, pre-training network structure, section features fusion method, the space-time integration strategy, affect the performance of recognition, proved that the fusion in individual video segmentation in the dual channel convolution neural network method can

 

capture video in the behavior characteristic of the output motion characteristics, improve the identification accuracy. The behavior recognition method proposed in this paper has practical engineering value, especially for the elderly behavior prediction, it provides a great guarantee for their safety. However, due to the lack of samples, the accuracy is not particularly high. In the future, effective samples will continue to be collected.

## References

[1] Lin Teng, Hang Li, Shoulin Yin, Shahid Karim &Yang Sun (2020). An active contour model based on hybrid energy and fisher criterion for image segmentation[J]. International Journal of Image and Data Fusion. vol.11, No. 1, pp. 97-112. 2020.
[2] Jing Yu, Hang Li and Desheng Liu (2020). Modified Immune Evolutionary Algorithm for IoT Big Data Clustering and Feature Extraction Under Cloud Computing Environment [J]. Journal of Healthcare Engineering, 1, 2020.
[3] Zhang Y, Cheng L, Wu J, et al (2016). Action Recognition in Still Images With Minimum Annotation Efforts[J]. IEEE Transactions on Image Processing, 2016, 25(11):5479-5490.
[4] Yi Y, Lin M (2016). Human action recognition with graph-based multiple-instance learning[J]. Pattern Recognition, 2016, 53(C):148-162.
[5] Shoulin Yin, Hang Li, Lin Teng, Man Jiang & Shahid Karim (2020). An optimised multi-scale fusion method for airport detection in large-scale optical remote sensing images [J]. International Journal of Image and Data Fusion, vol. 11, no. 2, pp. 201-214, 2020.
[6] Kapsouras I, Nikolaidis N (2018). A Vector of Locally Aggregated Descriptors Framework for Action Recognition on Motion Capture Data[C]// 2018 26th European Signal Processing Conference (EUSIPCO). 2018.
[7] Lin B, Fang B (2018). A New Spatial-temporal Histograms of Gradients Descriptor and HOD-VLAD Encoding for Human Action Recognition[J]. International Journal of Wavelets Multiresolution and Information Processing, 2018, 17(4).
[8] Zhigang, Hongyan, Dejun, et al (2018). Action-Stage Emphasized Spatio-Temporal VLAD for Video Action Recognition.[J]. IEEE transactions on image processing: a publication of the IEEE Signal Processing Society, 2018.
[9] Cai Z, Wang L, Peng X, et al (2014). Multi-view Super Vector for Action Recognition[C]// Computer Vision & Pattern Recognition. IEEE, 2014.
[10] Wang L, Koniusz P, Huynh D Q (2019). Hallucinating Bag-of-Words and Fisher Vector IDT terms for CNN-based Action Recognition[J]. ICCV, 2019. arXiv:1906.05910
[11] Peng X, Wang L, Wang X, et al (2016). Bag of Visual Words and Fusion Methods for Action Recognition: Comprehensive Study and Good Practice[J]. Computer Vision & Image Understanding, 150(Sep.):109-125, 2016.
[12] Wang L, Xiong Y, Wang Z, et al (2016). Temporal Segment Networks: Towards Good Practices for Deep Action Recognition[C]// European Conference on Computer Vision. Springer, Cham, 2016.
[13] Karpathy A, Toderici G, Shetty S, et al (2014). Large-Scale Video Classification with Convolutional Neural Networks[C]// Computer Vision & Pattern Recognition. IEEE, 2014.
[14] Simonyan K, Zisserman A (2014). Two-Stream Convolutional Networks for Action Recognition in Videos[J]. Advances in neural information processing systems, 2014, 1.
[15] Ng Y H, Hausknecht M, Vijayanarasimhan S, et al (2015). Beyond Short Snippets: Deep Networks for Video Classification[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015.
[16] Yin, S., Li, H (2020). GSAPSO-MQC:medical image encryption based on genetic simulated annealing particle swarm optimization and modified quantum chaos system. Evolutionary Intelligence (2020). doi：10.1007/s12065-020-00440-6
[17] Xiaowei Wang, Shoulin Yin, Desheng Liu, Hang Li & Shahid Karim (2020). Accurate playground localisation based on multi-feature extraction and cascade classifier in optical remote sensing images [J]. International Journal of Image and Data Fusion, vol. 11, no. 3. pp. 233-250, 2020.
[18] S. Yin and H. Li (2020). Hot Region Selection Based on Selective Search and Modified Fuzzy C-Means in Remote Sensing Images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 5862-5871, 2020, doi: 10.1109/JSTARS.2020.3025582
[19] Jing Yu and Lulu Zhao (2021). A Novel Deep CNN Method Based on Aesthetic Rule for User Preferential Images Recommendation[J]. Journal of Applied Science and Engineering. Volume 24, Issue 1, 2021.

 

## Authors' Profiles

**Dan Zheng** graduated with a Bachelor of Engineering from Shenyang Normal University in 2017. In her college, after completing the learning task, she interests in exploring her professional knowledge. During graduate, under the guidance of his master instructor, she researches information security theory and technology.

**Hang Li** obtained his Ph.D. degree in Information Science and Engineering from Northeastern University. Hang Li is a full professor of the Software college at Shenyang Normal University. He is also a master's supervisor. He has research interests in wireless networks, mobile computing, cloud computing, social networks, network security and quantum cryptography. Prof. Li had published more than 30 international journal and international conference papers on the above research fields.

**Shoulin Yin** received the B.Eng. and M.Eng. Degree from Shenyang Normal University, Shenyang, Liaoning province, China in 2016 and 2013 respectively. Now, he is a doctor in Harbin Institute of Technology. His research interests include Multimedia Security, Network Security, Filter Algorithm, image processing and Data Mining.

**How to cite this paper:** Dan zheng, Hang Lia, and Shoulin Yin. " Action Recognition Based on the Modified Two-stream CNN ", International Journal of Mathematical Sciences and Computing (IJMSC), Vol.6, No.6, pp.15-23, 2020. DOI: 10.5815/IJMSC.2020.06.03

 