Journal on Big Data
DOI:10.32604/jbd.2020.010431
Article

Tech Science Press

Multi-Modality Video Representation for Action Recognition

Chao Zhu[1], Yike Wang[1], Dongbing Pu[1], Miao Qi[1,*], Hui Sun[2,*] and Lei Tan[3,*]

[1]College of Information Science and Technology, Northeast Normal University, Changchun, 130117, China

[2]Institute for Intelligent Elderlycare, College of Humanities and Sciences of Northeast Normal University, Changchun, 130117, China

[3]College of Humanities and information, Changchun University of Technology, Changchun, 130012, China

[*]Corresponding Author: Miao Qi. Email: qim801@nenu.edu.cn

Received: 20 May 2020; Accepted: 10 September 2020

Abstract: Nowadays, action recognition is widely applied in many fields. However, action is hard to define by single modality information. The difference between image recognition and action recognition is that action recognition needs more modality information to depict one action, such as the appearance, the motion and the dynamic information. Due to the state of action evolves with the change of time, motion information must be considered when representing an action. Most of current methods define an action by spatial information and motion information. There are two key elements of current action recognition methods: spatial information achieved by sampling sparsely on video frames' sequence and the motion content mostly represented by the optical flow which is calculated on consecutive video frames. However, the relevance between them in current methods is weak. Therefore, to strengthen the associativity, this paper presents a new architecture consisted of three streams to obtain multi-modality information. The advantages of our network are: (a) We propose a new sampling approach to sample evenly on the video sequence for acquiring the appearance information; (b) We utilize ResNet101 for gaining high-level and distinguished features; (c) We advance a three-stream architecture to capture temporal, spatial and dynamic information. Experimental results on UCF101 dataset illustrate that our method outperforms other previous methods.

Keywords: Action recognition; dynamic; appearance; spatial; motion; ResNet101; UCF101

1 Introduction

Recently, recognizing human action has a large range of applications, such as video surveillance, behavior modelling, video retrieval [1] and motion capture. Video is a sequence of still images with temporal order which human can easily memory and understand, compared with computer. Existing action recognition methods are mostly hand-crafted to extract spatial-temporal features (e.g., HOG [2], SIFT [3], LBP [4]). However, these methods exist an excessive problem of computational cost when applied to long videos.

Convolutional Neural Networks (ConvNets) [5–9] have achieved great success in image processing domain since it can automatically extract higher-order features, which include distinguishable features and semantic information. To capture higher accuracy in classification task, we usually employ ConvNets to learn more samples in the training process. Nevertheless in practice we still face some challenges. Here are the main ones.

We need to represent the content of video with sampling frames sparsely in a more efficient way. Video is a long-range sequence which is consisted of a series of static frames. Meanwhile, understanding the frames' temporal order [1,10–11] is of great significant in achieving the comprehension of video. Usually, even a short video may comprise a large majority of frames, which may incur excessive computational cost in the process of extracting features from each frame.

We need to describe the dynamics of human behavior with a better method. It is crucial for depicting the dynamicity [1] among frames for considering and relating consecutive frames comprehensively. Optical flow can describe the movement of each pixel that the value represents the magnitude of the displacement. Displacement is a vector with direction, which can easily decompose a vector into the displacement in the x and y directions [12].

Motivated by the above observations, we propose a more efficient end-to-end structure of human action recognition to gain the representation with the multi-modality video feature, as shown in Fig. 1. The first contribution of our work is to address a new fixed interval randomly method, which can gain a more random, simple video sequence to represent the appearance information of action. The second contribution is to use ResNet-101, a deep ConvNets, to extract the high-level feature of action, with less parameters to achieve distinguished image feature. The last contribution is to propose an end-to-end three-stream network for getting the corresponding behavioral scores, and achieving a final fusion for action recognition.

2 Related Work

Convolutional Neural Networks has obtained great breakthrough in classification tasks since they can get the high-level feature from each layer automatically. They calculate the difference between prediction and ground-truth for correct recognition in the stage of training. Moreover, the ConvNets can be trained by using back-propagation to modify parameters without the participation of artificial computation. ConvNets has caused the task of action recognition leap forward by its powerful ability of feature extraction. Due to the massive frames in one video, some classic methods sampling frames randomly along with temporal dimension. Besides, different from image classification, action recognition needs spatial and temporal cues. In this section, we introduce some classic methods about frame selecting strategy and temporal representation.

2.1 Frame Selecting Strategy

Selecting frames has been extensively studied in dense sequence of video. The first type is getting one frame randomly to represent the appearance of behavior, such as Two-stream Network [10]. While this method leads to three problems: (1) The frame which plays the crucial role in the sequence may not be selected, (2) As we all know, one action is consisted of several stages. For example, we could decompose running into squatting at the starting line and sprinting on the track. Thus, the whole comprehension of video can be effected by only selecting one frame, leading to part of the action would be lost easily, (3) The randomly selected frame could not represent the key object of one action, i.e. as for the action of shooting an arrow, we can recognize the behavior exactly if we have a choose of the frame that the person holding the bow and arrow, while we may have a mis-classification if we choose the last frame of the sequence of the action because it is hard to get the distinguished feature for ConvNets learning. To solve above problems, Temporal Segment Networks [11] rose a segment network on temporal dimensionality. The whole video sequence is divided into three video segments, then they randomly sample one frame from each part. However, this may lead to the problem of choosing a frame which is just on the segment's boundary. This also means the difference among selected frames is small. While our work uses the method of sampling uniformly based on the video length, which means there is fixed an interval among each proposal frame in each segment. The advantage of this method is that the difference among the proposal frames is large and the network could learn more stages' feature easily. Thus, the candidate frames selected by this method could represent the content of the whole video more sufficiently.

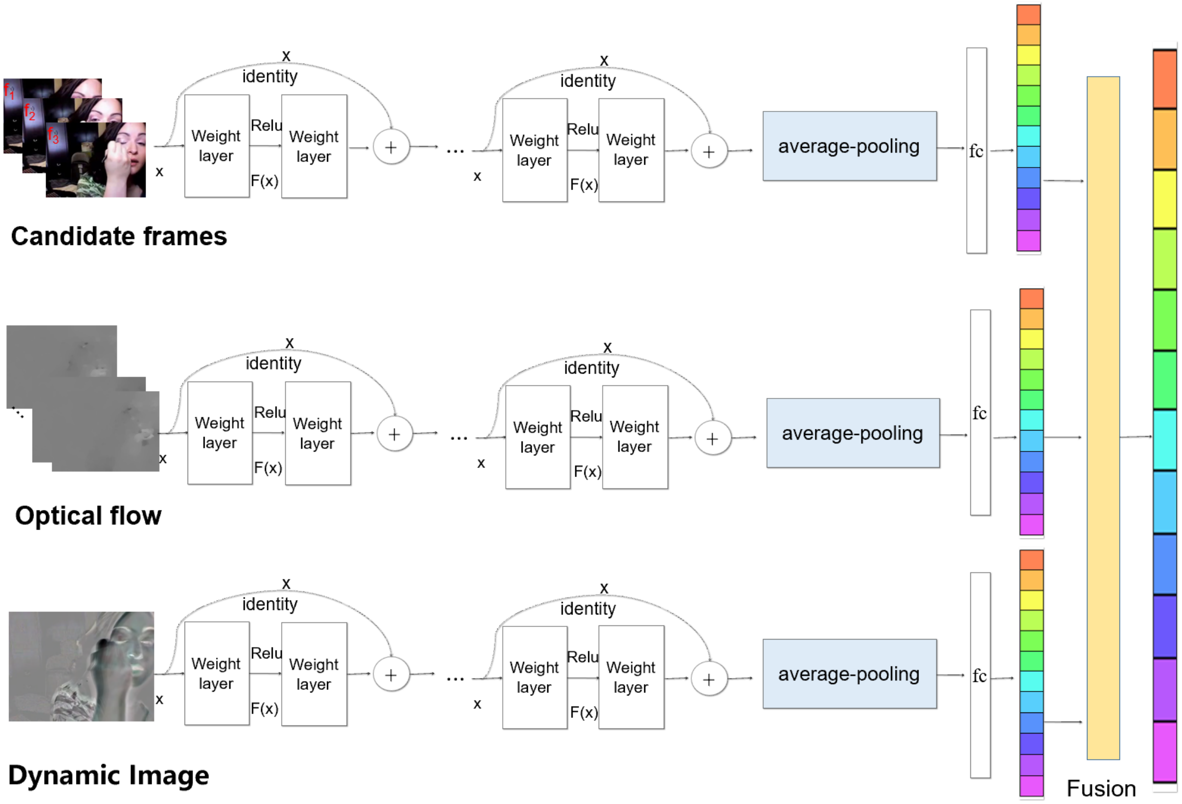

Figure 1: Illustration of proposed three-stream network takes candidate frames as the input of the spatial stream, optical flow as the input of the motion stream, a dynamic image as the input of the dynamic stream

2.2 The Temporal Representation for Video Analysis

We can find that the accuracy of ConvNets can be improved immensely by using the dynamics of a whole behavior. In practice, many approaches have adapted optical flow to describe the dynamics of action. Optical flow is the instantaneous velocity of each pixel on the imaging plane of a moving object in space. There were several works adapting optical flow to describe the dynamic of motion [1,10–11]. Literature [1] combined the spatial and motion information in a bilinear approach, which could generate a compact and sufficient feature. Nevertheless, this process cost a mass of computing resources and storage space. In general, most methods will use displacement to describe the velocity of the pixel. In literature [10], they argued that the displacement should be resolved into the shift of x and y direction, then the shifts in consecutive frames were stacked in the order of channel dimension and direction. They used this method to represent the dynamicity of the action. However, this approach is inadequate for frame sampling and lacks the information of the temporal and spatial relevance, resulting in these two modality information cannot represent the action thoroughly. Literature [12] proposed the basic algorithm of optical flow calculation by connecting two-dimensional velocity field with grayscale and introducing optical flow constraint equation. Literature [13] tested the ConvNets with the decomposition of optical flow into three components: magnitude, cosine angle and sine angle, where angle stood for the offset angle of displacement, and the magnitude was represented by the gradient. Literature [14] first proposed the concept of dynamic image, which could depict the appearance evolving with time sequence variation. They completed their work under the view that the same action category with different content should be similar but having the different behavioral dynamicity. They gave the approach of ranking pooling which used the fitted vector of function parameter and SVM to classify the parameter vector. In detail, a large range of feature representations were extracted in this method, then the rank pooling was operated based on these features, finally they used the smoothed vector of the whole video sequence to represent video. However, the problem of this method is that the feature is extracted by HOG, MBH and other methods. Additionally, it needed to train an extra SVM for classification. Literature [15] followed the idea of reference [14], they utilized the ConvNets to obtain video's feature. The feature contained not only the evolution of the appearance of the frame but also fused the temporality of the video. Though this work can describe more dynamic information, it has a deficiency of representing entire action appearance. At the same time, they argued to stack the frames to get the dynamic image for getting the temporal and spatial information. Literature [16] introduced some methods for fusing spatial and motion information in different layers using two-stream structure. However, it emerged a large amount of parameters. To solve this problem, another multi-resolution framework was presented to accelerate training [17]. They discussed a figure of ways to expand the connectivity of the ConvNets in the temporal field to utilize the partial spatial information. The disadvantage of this framework is that it could not gain sufficient motion information. Literature [18] designed a spatial-temporal decomposable ConvNets, which used 2D convolution kernel to learn the appearance feature in the bottom of the network and operated 1D convolution kernel to learn the temporal motion of the video in the top of the network. Yet, this work lacked the fully representation of the sequential information.

As the analysis above, choosing the proper frames in the video and extracting the dynamic information well play a significant role in the task of video recognition. In our work, we propose an end-to-end structure for action recognition. For describing the appearance and motion information, some frames are selected by using fixed interval randomly sampling method and the dynamicity of the video are represented with optical flow. Moreover, we utilize the network of ResNet101 to extract the appearance and motion feature in different layers. Specifically, we represent the result with the dynamic image to make a supplement of what they fuse. The comparable experimental result show that the proposed method is efficient and flexible.

3 The Proposed Approach

In our work, we propose an end-to-end three-stream structure to represent the multi-modality feature for action recognition. Section 3.1 introduces the method of fixed interval randomly sampling. Section 3.2 recommends how the spatial stream extracts the appearance information of action. Section 3.3 discusses the approach expresses the motion information of action. Section 3.4 presents a compact and effective method to fuse the appearance and temporal information with using the dynamic image to describe the appearance information evolving with time sequence variation.

3.1 Sample Method

To represent more sufficient spatial information and dig out more temporal cues, we sampled from frames more uniformly. We give a video sequence

$$\mathrm{V} = \{\mathrm{v}_{1,}\mathrm{v}_{2,}\mathrm{v}_{3,}\cdots,\mathrm{v}_{\mathrm{T}}\} \tag{1}$$

where T is the amount of frames.

To keep the homogeneity of the sampling, we divide the video into K segments on average, and there are s frames in each segment, which means

$$Segments_j = \{v_{s\times(j-1)+1,}v_{s\times(j-1)+2,}v_{s\times(j-1)+3,}\cdots,v_{s\times(j-1)+s}\}, j = 1,2,\cdots,k \tag{2}$$

First, the frame sampled randomly in the first segment as the initial state of the candidate frames. Then, we select the next proposal frame by stacking m interval, where m represents length of the segment. Finally, we can get a sub-sequence of the video:

$$Candidate = \{v_i, v_{i+m}, v_{i+2m},\cdots, v_{i+(K-1)\times m}\} \tag{3}$$

The reason of dividing the video into *K* segments is that one action is decomposed into several stages naturally. *K* frames are selected in total with Eq. (3). That is, utilizing *K* frames to represent the appearance of the action for describing the different states in different stages, shown as Fig. 2. The two advantages of this method are that: the sampling way of segmentation can make the sampling more well-distributed; Keeping fixed interval among candidate frames could make the difference among them more distinct. Due to the two consecutive frames are similar on appearance, we keep every candidate frame having the same interval to capture more distinct and variational features.

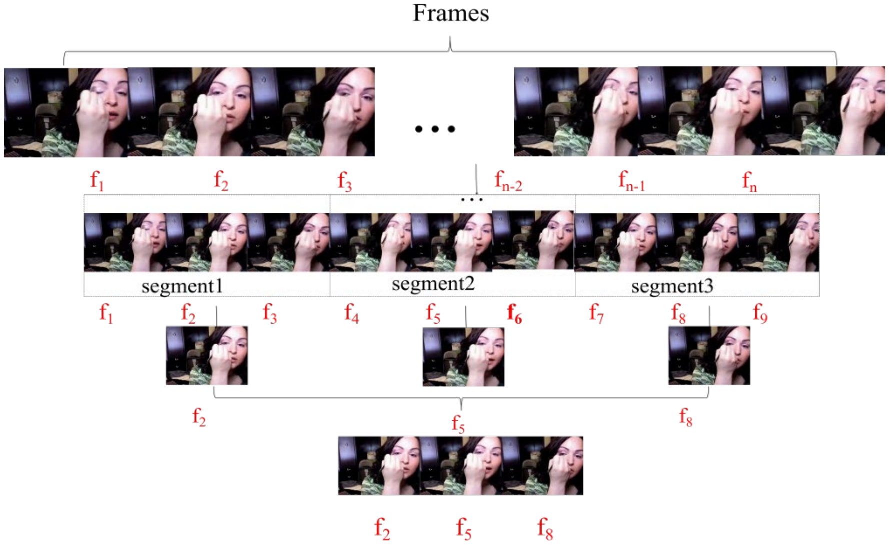

Figure 2: The process of fixed interval randomly sampling

3.2 Spatial Stream

For human being, when we have a sight of the holding a basketball, we can easily infer the category of the action. Due to the appearance of the basketball and the familiarity of the action of playing basketball, we can infer the action by only one frame. This indicates that the appearance of action is a very important clue for action recognition. Therefore, the use of RGB frames to describe the appearance of action has a great promotion effect on recognition.

The object of using spatial stream is to achieve the spatial information. Therefore, we first use the fixed interval randomly sampling method to obtain sub-sequence of the video. Secondly, the high-level features are extracted from the sub-sequence using ResNet-101 neural network. Utilizing ResNet as an extractor, the network can learn the fine-grained feature. With the deepen of the network, ResNet can extract the more higher-level feature without the vanishing gradient problem. Compared with the AlexNet [5] and VGGNet [6], there is an extra short connection added into the process of the connection in the Convolution layer to solve the vanishing gradient problem. This is used to realize the identity mapping of the input to the output. Finally, we do classify under the feature above to get the score of each frame for the classification. As the representation above, we choose several frames to do the description of the appearance. The classification results based on these frames are the scores of frame-level. For fusing these scores to get video-level score, we do statistics on the frame-level scores. We count the number of frames consistent with the predicted category and the real category to compute the top precision.

3.3 Motion Stream

Besides making the appearance information of the action available, we also need to know how the dynamicity of the action in the task of recognition. The point of utilizing the dense optical flow to describe the dynamicity of the action is proposed by [1,10–11]. The optical flow [10] is a very classic method to represent motion information in many areas such as video abstract [19] and object tracking [20]. It can be deemed as an image, which can be used to describe the situation of the movement. Each pixels' movement is regarded as a directive vector resolved into x and y direction. We can get a two-channel image to represent the dynamicity of the action. Hence, we employ this method to capture the motion information on Cartesian space.

Same with the spatial stream, we adopt a sequence of continuous t frames optical flow stacked in different directions to obtain the optical flow sequence. We use stacked optical flow as input.

$$optical\ flow\ = \{o_{q-4}^{x}, o_{q-4}^{y}, \cdots, o_{q-1}^{x}, o_{q-1}^{y}, o_{q}^{x}, o_{q}^{y}, o_{q+1}^{x}, o_{q+1}^{y}, \cdots, o_{q+5}^{x}, o_{q+5}^{y}\} \quad (4)$$

where q means the middle frame of candidate frames, x means the displacement of horizontal direction, y means the displacement of vertical direction.

Randomly sampled has a main disadvantage: The RGB frames chosen from the spatial stream randomly and the optical flow altered in the motion randomly is not corresponding exactly, which mean the optical flow and the RGB may represent the different stage of the action and the fusion of them may mislead networks and cannot classify two similar actions. Therefore, the middle frame of the candidate frames in the spatial stream is regarded as the basis of optical flow sequence. Based on this, we alter 4 serial optical flow maps in front of the basis optical flow and after 5 consecutive optical flow to make up of an optical flow sequence for depicting the motion of video.

3.4 Dynamic Stream

As mentioned above, we utilize the spatial stream to describe the appearance information and the motion stream to represent the dynamicity message of the action. While only using the two kinds of information to define an activity lacks the comprehension of the whole content with appearance and motion at the same time. To solve the problem above, we introduce the method of dynamic image as the third stream, which could encode the temporal evolution of the frames of the video.

Dynamic image is a compressed representation of the video frames by proposing a limited linear function to fit the temporal sequence. The parameters obtained from the function above are regarded as the feature vector of video. In our work, we first do rank pooling in the video achieving a dynamic image, then we send it to the ConvNets for extracting high-level features. The scheme is illustrated in Fig. 3.

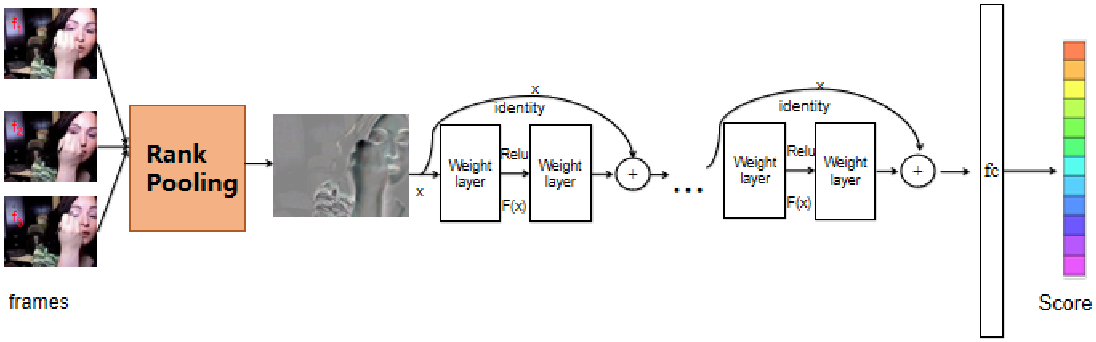

Figure 3: It shows the process of generating dynamic Image

Rank pooling puts forward that the length of the action is different, the order of the meta behavior is consistent. For the different videos with same action should have similar dynamicity. We do rank pooling in the frames, achieving the fitted function,

$$g\ =\ G(x\,|\,u), and\ u\ =\ \{u_1, u_2, u_3, \cdots, u_l\} \quad (5)$$

where has the same size with the original image.

Finally, we send it to the ConvNets to obtain high-level vectors for classification. We can also get the response of the vector to different categories.

4 Experimental Results

This chapter would introduce some experimental results to prove this paper’s viewpoint.

4.1 Dataset

UCF101 dataset is used to evaluate the performance of our method. UCF101 [21] contains 13320 video data in 101 categories of action. The action categories can be divided into five types: (1) Human-object interaction, (2) Body-motion only, (3) Human-human interaction, (4) Playing musical instruments, (5) Sports. There are 25 people do the action of each category with 4 to 7 groups. It has a large variety in capturing action, including the movement of camera and the change of appearance, posture, the objection scale, background, and the light.

4.2 Implementation Details

In our work, we propose an end-to-end structure of three-stream network, and each stream adopts ResNet101 as the basic architecture to extract feature on frame-level. Then we do pre-training in ImageNet and transfer the parameters into our task. The model of the GPU is GeForce GTX 1080. We set the learning rate in the spatial stream with 0.0005, 0.001 in the temporal stream and the dynamic stream, respectively. The parameter of dropout is set to 0.5. To avoid over-fitting, besides utilizing the dropout layer, we also operate the data augmentation on the dataset. In the spatial stream, we adopt the method of fivecrop to crop the image in the positions of top, bottom, left, right, and the middle. The operation of the horizontal rotation and normalization will be done in the same time. In the temporal stream, the image of optical flow is normalized with.

For the spatial stream, we choose the ResNet with 101 layers to extract higher semantic feature. ResNet101 neural network is consisted of 4 different scaled residual blocks, each of which block is made up of 3 convolution layers. The input image first goes through 33 residual block for getting feature maps with size of 3×3. Then, we do an average pooling by which we can keep the background information easily. Finally, the feature is input to a fully-connection layer, achieving a vector of the same length as the number of categories, which gets an activation with the help of softmax to gain the predictable score of each category.

For the temporal stream, we choose 10 consecutive optical flow with the direction of x and y, then it forms an input of the network with an image arranged by 20 channels. The same process of as spatial stream, we achieve the score of each category after the ResNet101 in the end.

4.3 Result Analysis

To validate the effective of the proposed method, extensive experiments are carried out. The method of Random select 1 RGB frame mentioned in the literature [10], which alters one frame to represent the appearance of the action. Instead, we do a segment on the video, and select one frame on each segment assuring the sampling evenly and thoroughly. In this way, we can describe the appearance information better. In the optical flow, we choose one of the frames in the proposal sequence, same with the spatial stream, as the reference flow. The input of the network is ten frames of optical flow with 20 channels which is selected from the relevance. The comparable results with literature [10] are list in Tab. 1. we can see that the fixed internal sampling randomly has a remarkable improvement compared to the traditional two-stream network.

Table 1: Several sampling methods' results on UCF101 dataset

Methods	Epoch	Accuracy
Random select 1 RGB [10]	500	73.0%
Random fixed interval sample(ours)	**100**	**82.1%**
Stack 10 optical flows [10]	500	79.4%
Select front and back 5 optical flows(ours)	**100**	**79.6%**

Action recognition can be viewed as a classification task, as previously mentioned, effective sampling method can capture multiple stage action information, which means being able to represent more complete action. Thus, we used the video classification accuracy as an index to quantify our sampling method. We calculated the average score of each category from several frames and considered the category which has the highest score as predict result. As shown in Tab. 1, with training only 100 epochs, we can achieve the accuracy of two-stream network with 500 epochs, which illustrates our sampling method can obtain more abundant features. Our sampling method work well on spatial stream and temporal stream.

Table 2: We validate different frameworks in the task of image classification in the spatial stream

Architecture	Accuracy
AlexNet	73.0%
VGGNet-16	78.4%
GoogleNet	77.1%
ResNet101	**82.1%**

We report the results of different network structures of the spatial stream in Tab. 2. Obviously, the result of ResNet101 is the best. AlexNet [5] adopts the image of as the input, then it goes through the 5 convolutional layers with two fully connected layers for classification with softmax. VGGNet-16 [6] contains 16 convolutional layers and three full connected layers, and simplify the structure of the neural network at the cost of a large amount of parameters. GoogleNet [7] proposes a network that can make good use of the computation resource, with reducing the parameters at the same time. In the research of ResNet, they found that the more layers they used in the regular plain network, the worse results got. Exactly, the vanishing gradient problem cause this phenomenon. To solve the problem above, [8] proposed the ResNet network with Identity map, which has no effect on the performance with the depth getting deeper.

Table 3: Exploration of stream in the ConvNets. Our three-stream achieves the highest accuracy

Network	Spatial	Motion	Fusion
Spatiotemporal ConvNet [18]	-	-	65.4%
Two-Stream Conv Pooling [16]	-	-	88.2%
Two-stream [10]	73.0%	79.4%	85.6%
DynamicImage [16]	86.6%	-	-
Factorized ST-ConvNet [22]	71.3%	76.0%	88.1%
Our (Spatial + Temporal + DynamicImage)	**82.1%**	**79.6%**	**88.5%**

For the action recognition, we hold the view that the key step is extracting the temporal information in a more effective way. According to this, we propose a three-stream to get the spatial and temporal information, and encode the temporal evolution of the frames of the video. These features complement each other making the proposed structure can have a better performance after fusion. We compare our method with some other structures with less representation in appearance and dynamicity, which cannot obtain a complete and effective action information. The comparable results are shown in Tab. 3. Our method used several frames as input in the spatial stream, the accuracy is higher than Two-stream method by 9.1%. Factorized ST-ConvNet used a decomposed network, the accuracy is lower than our method by 10.8%. In the motion stream, our method also achieve the best accuracy. Our accuracy is better than Two-stream method by 0.2%. Compared with Factorized ST-ConvNet method, the accuracy is higher by 3.6%. In the fusion-level, our method is better than other methods, the accuracy is higher than Two-stream by 1.4%. Above results show that the proposed method is effective and superior to other comparable methods.

5 Conclusion

Action recognition has applied to many areas as a practical research task such as intellectual robots. Many intellectual guiding robots have provided guiding service for human in the hospital. Besides, action recognition can be used to detect abnormal event in video surveillance. As a result, videos containing violent actions can be further detected and filtered. Action recognition task is of great research significance. In this paper, we present a three-stream network for action recognition demonstrated on UCF101. Our work achieves better or comparable performance for action recognition, when compared with those structures with only one stream or two streams. Experimental results have validated that fusing spatial information and motion information with the spatial appearance evolving with time sequence variation is effective. The results also have proved that the appearance and dynamicity of one action can help to represent the whole action. One action's spatial and motion information can offer distinguished feature for neural network to classify. In future work, we plan to mine the temporal relationship among different frames. For instance, we would like to utilize the method of concatenation to reason the temporal relationship among frames on high-level features.

Funding Statement: This work was supported in part by the National Natural Science Foundation of China (Nos. 61672150, 61907007), by the Fund of the Jilin Provincial Science and Technology Department Project (Nos. 20180201089GX, 20190201305JC), Provincial Department of Education Project (Nos. JJKH20190291KJ, JJKH20190294KJ, JJKH20190355KJ) and the Fundamental Research Funds for the Central Universities (No. 2412019FZ049).

Conflicts of Interest: We declare that our work has no conflicts of interest to report regarding the present study.

References

[1] D. Ali, S. Vivek and V. G. Luc, "Deep temporal linear encoding networks," in *Proc. CVPR,* Hawaii, USA, pp. 2329–2338, 2017.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR,* San Diego, USA, pp. 886–893, 2005.

[3] D. G. Lowe, "Distinctive image features from scale-Invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[4] T. Ojala, M. Pietikainen and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions." in *Proc. ICPR,* Jerusalem, Israel, vol. 1, pp. 582–585, 1994.

[5] A. Krizhevsky, I. Sutskever and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS,* Lake Tahoe, USA, vol. 1, pp. 1097–1105, 2012.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. CVPR,* Columbus, USA, 2014.

[7] C. Szegedy, W. Liu, Y. Q. Jia, P. Sermanet, S. Reed *et al.,* "Going deeper with convolutions," in *Proc. CVPR,* Boston, USA, pp. 1–9, 2015.

[8] K. M. He, X. Zhang, X. Q. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR,* Boston, USA, pp. 770–778, 2015.

[9] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in

Proc. of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[10] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. NIPS,* Montreal, Canada, vol. 1, pp. 568–576, 2014.

[11] L. M. Wang, Y. J. Xiong, Z. Wang, Y. Qiao, D. H. Lin *et al.,* "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. ECCV,* Amsterdam, Netherlands, pp. 20–36, 2016.

[12] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 2, pp. 185–203, 1981.

[13] R. H. Gao, B. Xiong and K. Grauman, "Im2Flow: Motion hallucination from static images for action recognition," in *Proc. CVPR,* Salt Lake City, pp. 5937–5947, 2018.

[14] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati and T. Tuytelaars, "Rank pooling for action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 773–787, 2015.

[15] H. Bilen, B. Fernando, E. Gavves and A. Vedaldi, "Action recognition with dynamic image networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2799–2813, 2017.

[16] C. Feichtenhofer, A. Pinz and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. CVPR,* Las Vegas, USA, pp. 1933–1941, 2016.

[17] K. Andrej, T. George, S. Sanketh, L. Thomas, S. Rahulet *et al.,* "Large-scale video classification with convolutional neural networks," in *Proc. CVPR,* Columbus, USA, pp. 1725–1732, 2014.

[18] L. Sun, K. Jia, D.Y. Yeung and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proc. ICCV,* Santiago, USA, pp. 4597–4605, 2015.

[19] S. X. Chen and Y. G. Jiang, "Motion guided spatial attention for video captioning," in *Proc. AAAI,* Hawaii, USA, 2019.

[20] B. Du, S. Ca and C. Wu, "Object tracking in satellite videos based on a multiframe optical flow tracker," *Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 99, pp. 1–13, 2019.

[21] K. Soomro, A. R. Zamir and M. Shah, "UCF101: A dataset of 101 human action classes from videos in the wild," arXiv:1212.0402, 2012.

[22] Y. H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga *et al.,* "Beyond short snippets: Deep networks for video classification," in *Proc. CVPR, Boston*, USA, pp. 4694–4702, 2015.