Video Action Recognition Based on Hybrid Convolutional Network

Yanyan Song, Li Tan(✉), Lina Zhou, Xinyue Lv, and Zihao Ma

Beijing Technology and Business University, Beijing, China
tanli@th.btbu.edu.cn

Abstract. Aiming at the problem of unbalanced distribution of spatio-temporal information in video images, this paper proposes a 2D/3D hybrid convolutional network that introduces attention mechanism, which fully captures video space information and dynamic motion information, and better reveals motion features. With the help of the dual-stream convolutional network structure, we built 2D convolution and 3D convolution parallel neural networks. In the 2D convolutional neural network, the residual structure and the LSTM network model are used to focus on the spatial feature information of the video behavior. Secondly, the 3D convolutional neural network constructed by Inception structure is used to extract the spatiotemporal feature information of video behavior. On the basis of the two high-level semantics extracted, the attention mechanism is introduced to fuse the features. Finally, the obtained significant feature vector is used for video behavior recognition. Compared with other network models on the UCF101 and HMDB51 datasets, it can be seen from the results that the proposed 2D/3D hybrid convolutional network has good recognition performance and robustness.

Keywords: Action recognition · Mixed convolution · Attention mechanism · Deep learning · Video understanding

1 Introduction

With the development of information technology, multimedia on the Internet is growing rapidly, and the number of shared videos is increasing. Video-based behavior recognition plays a key role in various fields, such as public intelligence monitoring, human-computer interaction, and video search. With the deep application of deep learning in images, more and more researchers have applied deep learning methods to video processing. Video data is a dynamic and continuous set of information compared to images. Therefore, in the data processing process, not only the information of the video space dimension but also the depth information in the video time dimension should be considered.

But video-based behavior recognition is a challenging issue. In the video, it is still difficult for the machine to recognize a person's simple actions due to problems such as posture, angle, illumination, and occlusion. To improve the effect of behavior recognition, it is necessary to extract the effective features of the video as much as possible. Based on the 2D convolutional neural network, remarkable achievements have been

X. Sun et al. (Eds.): ICAIS 2020, LNCS 12240, pp. 451–462, 2020.
https://doi.org/10.1007/978-3-030-57881-7_40

made in two-dimensional signal processing such as images. However, when the 2D convolutional neural network performs feature extraction on video data, the limitation of the 2D convolution method leads to ignoring the time information of the third dimension of the video during the convolution process. In theory, the 3D convolutional neural network can use the 3D convolution kernel to extract the spatio-temporal hybrid features of video data, but in practice, the 3D convolutional neural network does not show superior performance. Compared with the 2D convolutional neural network, the parameters of the 3D convolutional neural network are too large, and it is difficult to train the deep convolutional neural network, which increases the complexity of network optimization and the consumption of computing resources.

In order to capture the visual space information and dynamic motion information of video images, this paper proposes a 3D/2D hybrid convolutional neural network, which uses 3D convolution to extract the spatiotemporal features of video behavior, and another part uses 2D convolution to focus on extracting appearance information. Finally, the two high-level semantic information are integrated. When constructing a 3D/2D hybrid convolutional neural network, the 2D convolution is used to compensate the depth of the feature map while limiting the number of 3D convolutional layers, which makes it possible for the network to achieve better performance with less spatiotemporal fusion.

2 Related Work

Video behavior recognition is a core issue in computer vision. With the application of high-performance deep convolutional neural networks for image recognition tasks, many work has designed effective deep convolutional neural networks for behavior recognition [1–8, 14–22, 24, 26]. For example, the classical dual-stream convolution network proposed by Simonyan K et al. [1]. Video can naturally be broken down into spatial and temporal components. They designed the video recognition architecture and divided it into two streams. Each stream is implemented using ConvNet, and its softmax scores are combined through late fusion. The mainstream ConvNet framework relies on dense time sampling with predefined sampling intervals by focusing on appearance and short-term motion, thus lacking the ability to integrate remote time structures. To solve this problem, Wang L et al. proposed Time Segment Network (TSN) [2], a novel framework for motion recognition based on video, based on the idea of remote time structure modeling. It combines sparse time sampling strategies and video-level supervision to achieve efficient and effective learning using the entire action. Based on the time segment network, Zhou B et al. proposed an efficient and interpretable network model, Time Series Network (TRN) [3]. The network model can learn and infer the timing dependence of frames on multiple scales in video. The proposed time-series network sparsely samples the frames and then learns their causality to achieve efficient capture of timing relationships over multiple time scales.

Compared to 2DConvNet, 3DConvNet is able to better model temporal information through 3D convolution kernel and 3D pooling operations. Therefore, Tran D et al. proposed a simple and effective method for temporal and spatial feature learning using a deep three-dimensional convolutional network (C3D) [5] trained on large-scale surveillance video datasets, demonstrating that C3D networks can Simultaneously simulate appearance and motion information. Since the 3D convolutional neural network has more parameters than 2DConvNets, and the training video architecture requires additional large tag data sets, several variants of the proposed centralized 3D convolutional neural network fail to utilize long-range time information, thus limiting the performance of these architectures. Diba A et al. proposed a novel deep space-time feature extractor network (TTL) [6] that models the variable-time 3D convolution kernel depth over a short and long time horizon, and On this basis, the DenseNet architecture was extended, replacing the standard transition layer in the DenseNet architecture with TTL. Their proposed network architecture (T3D) [6] captures short-term, medium-term and long-term behavioral appearance and time information intensively and efficiently. Shuiwang Ji et al. developed a 3D convolutional neural network architecture based on 3D convolution feature extractor [8]. The CNN architecture generates multiple information channels from adjacent video frames and performs convolutional kernel sub-sampling in each channel to obtain a final feature representation by combining information from all channels.

In summary, it can be found that the idea of both the 2D convolutional neural network architecture and the 3D convolutional neural network architecture design is to make full use of the depth information of the video on the basis of extracting the appearance characteristics of the video behavior. Using the motion information provided by the optical flow frame or studying the timing relationship of the input frame is to some extent compensate for the video time dimension feature of the 2D convolutional architecture loss, while the 3D convolution architecture enhances the 3D extraction of spatio-temporal features under the limited convolutional layer. Therefore, we designed a 3D/2D hybrid convolutional neural network for the advantages and disadvantages of 2D convolution and 3D convolution.

3 Methods

In this section, we first introduce the built 2D convolution network and 3D convolution network. After that, we will introduce in detail the simple and effective 3D/2D mixed convolution behavior recognition model we provide for motion recognition in video.

3.1 2D Convolutional Neural Network

Res-RNN Model. In the 2D convolutional neural network part, the Res-RNN model is used to extract video behavior feature information. Firstly, the residual structure is used to build the CNN network to extract the appearance characteristics of the video behavior. On the basis of CNN, the RNN network is constructed to extract the depth features of the video behavior. The model structure is shown in the Fig. 1.

The video data contains depth information compared to the image data, and the 2D convolutional neural network focuses on extracting the appearance characteristics of a certain behavioral video frame. In the built CNN network, the residual network structure [9] is adopted to enhance the extraction capability of network features. Because residual learning is a good solution to the degradation of deep networks compared to other network structures. In the common convolutional neural network, as the number of network layers increases, the network performance does not increase proportionally,

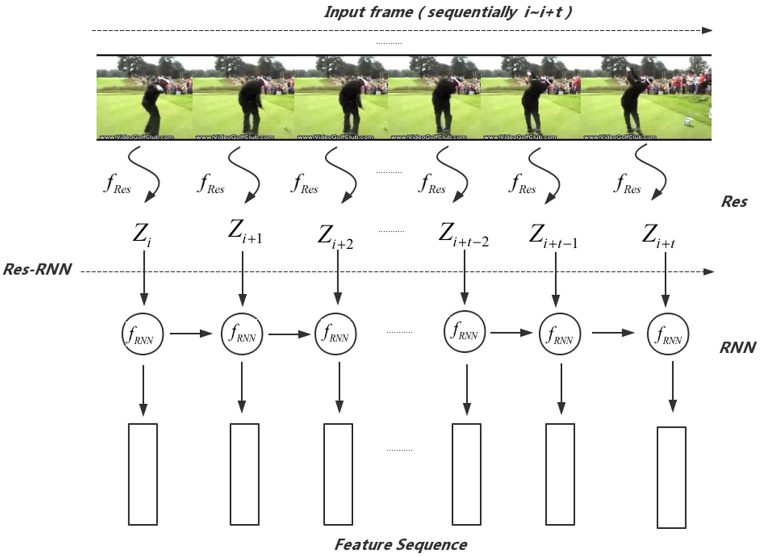

Fig. 1. Res-RNN model structure

but instead the gradient dispersion or gradient explosion problem occurs. The residual network structure utilizes the "shortcut connection" connection method to complete at least some of the redundant layer's identity mapping on the basis of the optimized network hierarchy, ensuring that the input and output of the identity layer are identical, thereby solving the problem of degraded phenomena that caused by an increase in network depth.

In order to compensate for the effect of 2D convolutional neural network on behavioral video data extraction features, and because the video input frame is time series, the long-term short-term memory model in the cyclic network [10] is used to construct the Res-RNN network. The LSTM passes the self-loop to generate a path for the gradient to continue to flow for a long time, and the weight of the self-loop is not fixed, and the accumulated time scale can be dynamically changed by the input sequence. In addition to the external RNN cycle, the LSTM also has an internal "LSTM cell" cycle (self-loop), from which it appears that the LSTM does not simply apply an element-by-element nonlinearity to the affine transformation of the input and loop elements, so LSTM is easier to learn for long-term dependencies than a simple loop architecture. It captures long-term motion information of video input frames. The structure of a certain cell network is shown in the figure below (Fig. 2).

For the input data of the behavioral video, by capturing the characteristics of the input frame at different times, the information of the previous frame of the video is used to understand the information of the current frame, and to find its dependence, thereby extracting the motion information of the behavioral video. The LSTM flows by constructing some gate control information. The important components are shown in the figure above, including the forgetting gate f, the external input gate g, the output gate q and the state unit s. Suppose that a cell unit i in the LSTM cyclic network has an input

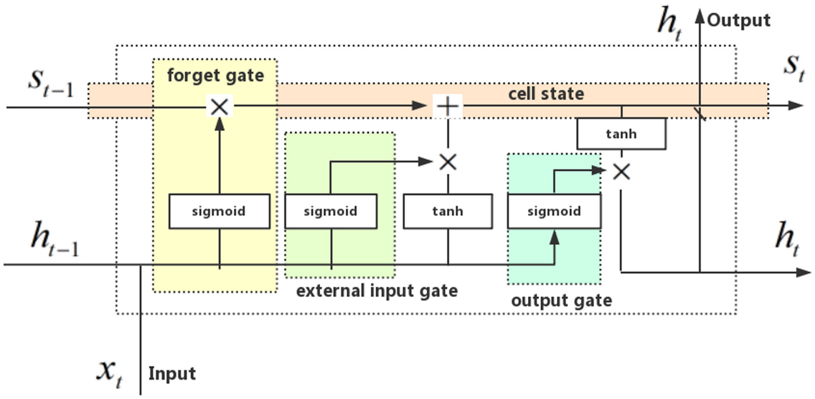

Fig. 2. LSTM cycle network "cell" block diagram

vector of $x_i^{(t)}$ at time t and an output vector of $h_i^{(t)}$. First, the weight of the self-loop is determined by the forgetting gate $f_i^{(t)}$. The weight is set by the sigmoid unit to a value between 0 and 1, where 1 represents complete reservation and 0 represents complete forgetting. The calculation formula is as follows,

$$f_i^{(t)} = sigmoid(b_i^f + \sum_j U_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f h_j^{(t-1)}) \tag{1}$$

Where b^f, U^f and W^f are the cyclic weights of the offset, input weight, and forgetting gate, respectively. Then the external input gate $g_i^{(t)}$ determines the information to be stored in the cell, which is calculated as follows,

$$g_i^{(t)} = sigmoid(b_i^g + \sum_j U_{i,j}^g x_j^{(t)} + \sum_j W_{i,j}^g h_j^{(t-1)}) \tag{2}$$

Where b^g, U^g and W^g are the cyclic weights of the offset, input weight and forgetting gate, respectively. After that, the status unit $s_i^{(t)}$ of the LSTM cell is updated and calculated as follows,

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \tanh(b_i + \sum_j U_{i,j} x_j^{(t)} + \sum_j W_{i,j} h_j^{(t-1)}) \tag{3}$$

Among them, b, U and W are the offset weights of the LSTM cells, the input weights and the forgetting gates. Finally, the output of the cell unit is calculated by the output gate $q_i^{(t)}$ and the state unit $s_i^{(t)}$, which is calculated as follows,

$$q_i^{(t)} = sigmoid(b_i^o + \sum_j U_{i,j}^o x_j^{(t)} + \sum_j W_{i,j}^o h_j^{(t-1)}) \tag{4}$$

$$h_i^{(t)} = \tanh(s_i^{(t)}) q_i^{(t)} \tag{5}$$

Where b^o, U^o and W^o are the cyclic weights of the offset, input weight and forgetting gate, respectively.

The constructed Res-RNN network extracts and encodes the video input frame into a one-dimensional feature vector through the CNN of the residual structure, and the LSTM receives the sequence one-dimensional feature vector from the Res encoder, and outputs the sequence vector containing the video depth feature for subsequent Classification forecast.

3.2 3D Convolutional Neural Network

Network Structure. In the construction of 3D convolutional neural network, the reference I3D [7] network structure is selected, which can capture the time structure of fine action and perform better model features. The network architecture diagram is as follows (Fig. 3),

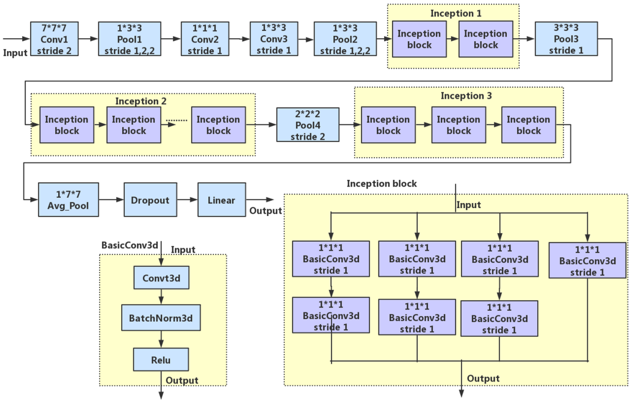

Fig. 3. 3D convolution network architecture

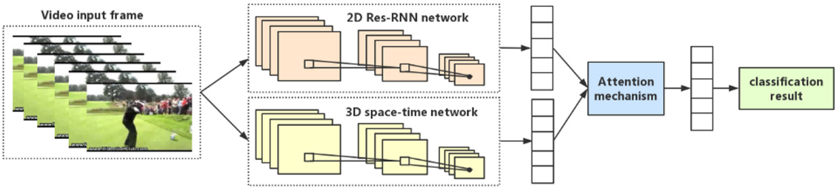

Fig. 4. Mixed convolution network behavior recognition model

3.3 Hybrid Convolutional Network

Action Recognition Model. In order to improve the accuracy of behavior recognition, the attention mechanism [11, 25] is used to fuse the features extracted by the 2D convolutional neural network and the 3D convolutional neural network. The hybrid convolutional network will learn the sequence of features with time information and convolve the input video frames layer by layer. Based on the advantages of 2D convolution and 3D convolution, the attention mechanism is used to add weights to the behavioral features extracted by the model. And the feature vectors with significant attention weights are output for the classification of video behavior. The video behavior recognition model of the 2D/3D hybrid convolutional network based on attention mechanism proposed in this paper is shown in the figure below (Fig. 4).

First, the video data is segmented into video frames using the ffmpeg tool. Using the obtained image frame data as input data of the constructed hybrid convolutional network model, feature extraction is performed separately by layer-by-layer convolution to generate two high-level semantic feature sequences. After that, the high-level semantic feature sequences obtained by convolving 2D and 3D are input into the attention mechanism. By calculating the significant attention, the corresponding weights are added to the input feature sequence, and the output obtains a new feature sequence with attention weights for the recognition of the video behavior. The attention mechanism is applied in the network, and the video frame feature vectors obtained by 2D and 3D convolution are weighted, which can enhance the saliency of the feature sequence, continuously train learning, reduce losses, and improve the accuracy of behavior recognition.

Attention Mechanism. The human brain can intentionally or unintentionally select a small amount of useful information from a large amount of input information to focus on and ignore other information. Inspired by this, the attention mechanism is introduced to select important information for the neural network to calculate. Assume that N information, $I = [x_1, \ldots\ldots., x_n]$, and question q (the object of interest, which may be itself) are input, and some task-related information is selected from input I and input to the neural network for learning. First, calculate the attention distribution α_i, that is, the probability of selecting the $i-th$ information, and the calculation formula is as follows.

$$\alpha_i = \mathrm{softmax}(score(x_i, q)) = \frac{\exp(score(x_i, q))}{\sum_{j=1}^{N} \exp(score(x_j, q))} \tag{6}$$

Where $score(x_i, q)$ is a scoring function, here we define the $score$ function as,

$$score(x_i) = sigmoid(W^T x_i + b) \tag{7}$$

On the basis of calculating the attention distribution, the soft attention mechanism is used to calculate the attention, and all the information is weighted and summed, and the input information is encoded as,

$$attn(I, q) = \sum_{i=1}^{N} \alpha_i x_i \tag{8}$$

Attention is essentially to integrate the information of the feature vector, reduce the computational complexity, obtain significant features, and improve the feature representation ability of the network model.

4 Experiment

4.1 Experiment Environment

The experimental platform is Dell server PowerEdge R430, operating system: Ubuntu 14.04, CPU: Intel (R) Core i3 3220, memory: 64 GB, GPU: NVIDIA Tesla K40 m × 2, video memory: 12 GB × 2.

4.2 Experimental Parameter Setting

Data Set. The data sets selected for the experiment were the UCF101 data set [12] and the HMDB51 data set [23]. The UCF101 data set has a total of 13320 video segments, and the number of categories is 101. It has the greatest diversity in motion. Some of its behavioral categories are shown on the left in Fig. 5. The HMDB51 dataset contains 6849 clips and is divided into 51 action categories. Each category contains at least 101 clips, and some action categories are shown on the right in Fig. 5.

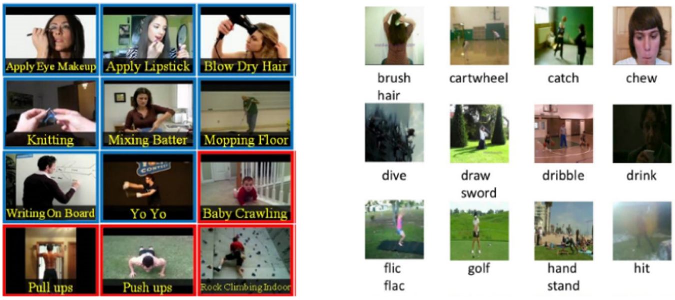

Fig. 5. UCF101 data set (left) and HMDB51 data set (right)

Experimental Parameters. The behavioral feature extraction process uses the Pytorch framework, and the model used is the 2D/3D hybrid convolutional network built above. In the training, the batch-size is set to 4, 8, 16, and 28 frames respectively for experimental comparison. At the same time, experiments are carried out on the input order of different frames to see if it affects the accuracy of behavior recognition. The BatchNormalization layer is set in the network to perform batch normalization to solve the problem of changing the distribution of data in the middle layer during the training process. In addition, consider the model without adding the BatchNormalization layer, and compare experiments with other networks to observe the performance of the built 2D/3D hybrid convolution network. The optimization method selects the Adam method and the learning rate is set to 0.001.

4.3 Experimental Result

In order to verify the effectiveness of the 2D/3D hybrid convolutional network that introduces the attention mechanism, the experimental comparison with the basic network I3D is carried out from the number of input frames, the order of input frames and whether or not the BN layer is included. The 2D/3D hybrid convolution network we built can take advantage of the network under different convolution kernels. End-to-end training is performed by taking a well-framed RGB video sequence as input. In order to verify the performance of the hybrid network, we used the UCF101 public data set for behavior recognition to test. The results are shown in Table 1 below.

Table 1. Comparison of experimental results on the UCF101 data set

Method	Dim	BN	Input	UCF101				Speed
				4f (%)	8f (%)	16f (%)	28f (%)	Average (fps)
I3D [7]	3d	Y	Order	72.46	71.35	75.11	79.39	8.00
		N	Order	46.57	37.23	42.58	58.79	7.07
		Y	Random	74.23	70.63	74.74	73.15	6.88
		N	Random	44.26	36.93	48.34	50.48	6.95
Our	3d/2d	Y	Order	**83.51**	**82.88**	**80.72**	**80.69**	**19.18**
		N	Order	73.42	72.31	72.67	71.50	22.82
		Y	Random	82.37	80.78	78.86	77.21	20.73
		N	Random	73.12	74.71	70.48	66.09	22.75

It can be seen from the results in the table that the feature extraction by inputting 4, 8, 16, 28 different frames to the convolutional neural network has a certain influence on the video behavior recognition result. It can be seen from the data in the table that for the I3D network model, the influence of the input of different frame data on the accuracy of behavior recognition is relatively large. For the proposed hybrid convolutional network, the input of different frame data has less influence on the accuracy of behavior recognition.

The partial data in the table is compared and displayed in a visual form. The result is shown in Fig. 6 and Fig. 7 below.

For the order of input frames is ordered or unordered, we can see from Fig. 6 that when the number of network input frames is large, the sequential features extracted by ordered input can improve the accuracy of behavior recognition. Secondly, as can be seen from Fig. 7, when the I3D network model removes the BN layer, the network performance is seriously degraded, and the proposed hybrid convolutional network is relatively robust. And the network model processes data at a fast speed, so that it can be seen that the hybrid convolution network model with attention mechanism has better performance advantages.

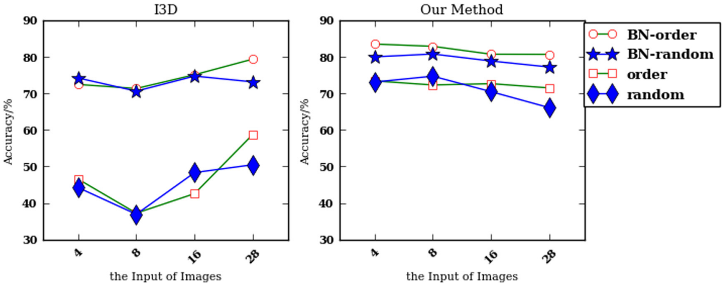

Fig. 6. Comparison of ordered or random input results

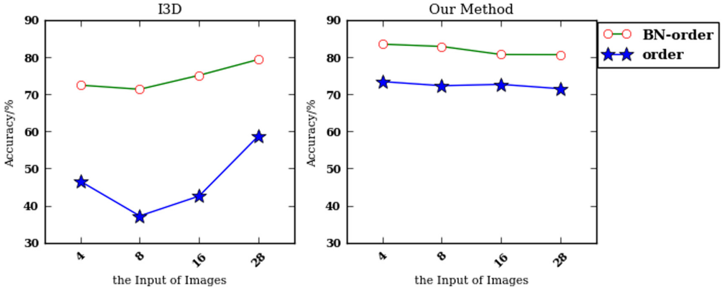

Fig. 7. Comparison of results of adding BN layers

To further understand the performance of the hybrid convolutional network model that introduces the attention mechanism, we further compare the performance of the UCF101 dataset and the HMDB51 dataset with other network models. The results are shown in Table 2.

Table 2. Comparison results with other network models on UCF101 and HMDB51

Method	Dim	UCF101	HMDB51
Two-Stream [1]	2D	78.85%	——
CRNN	2D	67.63%	34.04%
I3D [7]	3D	79.39%	28.48%
3D-ResNet [13]	3D	67.60%	25.67%
Our	3D/2D	**83.51%**	**45.71%**

5 Conclusion

A 2D/3D hybrid convolutional network with attention mechanism is introduced to extract the spatio-temporal features of video behavior through 2D/3D parallel convolutional networks, and generate two different high-level semantic information. Then, the attentional mechanism is used to apply different weights to the features to obtain a distinctive feature vector for the behavior recognition of the video. The experimental results show that the proposed methods have an accuracy of 83.51% and 45.71% on the UCF101 and HMDB101 data sets respectively. Compared with other network models, they have better recognition performance and robustness, which proves the effectiveness of the method.

Acknowledgments. This research was funded by Beijing Natural Science Foundation-Haidian Primitive Innovation Joint Fund grant number (L182007) and the National Natural Science Foundation of China grant number (61702020).

References

1. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Computer Vision and Pattern Recognition (2014)
2. Wang, L., et al.: Temporal segment networks: towards good practices for deep action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 20–36. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_2
3. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11205, pp. 831–846. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01246-5_49
4. Zhu, Y., Lan, Z., Newsam, S., Hauptmann, A.: Hidden two-stream convolutional networks for action recognition. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018. LNCS, vol. 11363, pp. 363–378. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20893-6_23
5. Tran, D., Bourdev, L., Fergus, R., et al.: Learning spatiotemporal features with 3D convolutional networks. In: IEEE International Conference on Computer Vision, Santiago, pp. 4489–4497. IEEE Computer Society (2015)
6. Diba, A., Fayyaz, M., Sharma, V., et al.: Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification. Computing Research Repository (2017)
7. Carreira, J., Zisserman, A.: Quo Vadis, action recognition? A new model and the kinetics dataset. In: IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, pp. 4724–4733. IEEE Computer Society (2017)
8. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2013)
9. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. CoRR abs/1512.03385 (2015)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
11. Vaswani, A., et al.: Attention is all you need. CoRR abs/1706.03762 (2017)
12. Soomro, K., Zamir, A.R., Shah, M.: UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. CoRR abs/1212.0402 (2012)

13. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In: IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, pp. 6546–6555. IEEE Computer Society (2018)
14. Tran, D., Wang, H., Torresani, L., et al.: A closer look at spatiotemporal convolutions for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, pp. 6450–6459. IEEE Computer Society (2018)
15. Liu, Z., Hu, H., Zhang, J.: Spatiotemporal fusion networks for video action recognition. Neural Process. Lett. **50**(2), 1877–1890 (2019). https://doi.org/10.1007/s11063-018-09972-6
16. Li, Q., Qiu, Z., Yao, T., et al.: Action recognition by learning deep multi-granular spatio-temporal video representation. In: International Conference on Multimedia Retrieval, pp. 159–166. ACM, New York (2016)
17. Zhou, Y., Sun, X., Zha, Z.-J., Zeng, W.: MiCT: mixed 3D/2D convolutional tube for human action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, pp. 449–458. IEEE Computer Society (2018)
18. Yu, W., Wei, Y., Li, L.: Human behavior recognition model based on multi-model fusion. Comput. Eng. Des. **40**(10), 3030–3036 (2019)
19. Zeng, M., Zheng, Z., Luo, S.: Dual-convolution human behavior recognition combined with LSTM. Modern Electron. Tech. **42**(19), 37–40 (2019)
20. Ma, C., Mao, Z., Cui, J., Yi, W.: Behavior recognition based on deep LSTM and dual stream convergence network. Comput. Eng. Des. **40**(09), 2631–2637 (2019)
21. Ma, L., Yu, W., Zhu, Y., Wang, C., Wang, P.: Recognition of fall behavior based on deep learning. Comput. Sci. **46**(09), 106–112 (2019)
22. Rodríguez-Moreno, I., Martínez-Otzeta, J.M., Sierra, B., Rodriguez, I., Jauregi, E.: Video activity recognition: state-of-the-art. Sensors (Basel, Switzerland) **19**(14), 3160 (2019)
23. Wishart, D.S., Tzur, D., et al.: HMDB: the human metabolome database. Nucleic Acids Res. **35**, 521–526 (2007)
24. Shah, S.M.S., Malik, T.A., Khatoon, R., Hassan, S.S., Shah, F.A.: Human behavior classification using geometrical features of skeleton and support vector machines. Comput. Mater. Continua **61**(2), 535–553 (2019)
25. Yang, K., Wang, Y., Zhang, W., Yao, J., Le, Y.: Keyphrase generation based on self-attention mechanism. Comput. Mater. Continua **61**(2), 569–581 (2019)
26. Song, W., Yu, J., Zhao, X., Wang, A.: Research on action recognition and content analysis in videos based on DNN and MLN. Comput. Mater. Continua **61**(3), 1189–1204 (2019)