Global Temporal Pyramid for Human Abnormal Action Recognition

Shengnan Chen*[a], Yuanyao Lu[b] , Pengju Zhang[c] , Yixian Fu[d]
North China University of Technology, Beijing 100144

ABSTRACT

With the development of monitoring technology and the improvement of people's security awareness, intelligent human abnormal action recognition technology in the field of action recognition is increasingly high. In most cases, abnormal human action may have little difference in appearance compared with normal behavior, so the control of visual rhythm information becomes an important factor affecting action recognition, but people often focus on the appearance information of the action and ignore the rhythm information. In this paper, we introduce the temporal pyramid module to process the visual tempos information, meanwhile, the traditional LSTM local history information transfer method is very easy to lose the context information, which is not conducive to the grasp of global information and thus will greatly affect the processing effect of the temporal pyramid. This paper introduces a non-local neural network module to enhance the network's ability to grasp global information and the model's long-range modeling capability, which is used to supplement the temporal pyramid module. Finally, this paper uses the mainstream anomaly dataset UCF-Crime to test the network performance, and the improved network model recognition accuracy AUC reaches 0.82, which is better than other state-of-the-art methods.

Keywords: abnormal action recognition, deep learning, temporal pyramid network, non-local means

1. INTRODUCTION

Video surveillance system in today's social security field occupies a fairly important position, with the continuous development of social and economic development, the expansion of the scope of human activities, and the expansion of the type of activity, monitoring almost every piece of human social activity places, and with it, the number of surveillance video is also increasing rapidly. Such a huge amount of surveillance video data only relying on manual access and analysis will obviously bring huge human costs, and limited manpower can not give full play to the security role of surveillance video so intelligent video anomaly identification technology is essential.

Computer vision technology based on deep learning can automatically learn data features and fully exploit potential correlations between massive data, so it is widely used in the era of big data. In this paper, we use the classical two-stream[1] method in the field of video action recognition processing based on deep learning to split the spatial information and temporal information, so that the information in the temporal branch can be processed more effectively while controlling the complexity of the model. The two-stream network uses RGB images as input in the spatial stream and optical stream images in the temporal stream[2,4]. At the same time, we use the inflated 3D network as the backbone network, 3D convolution is introduced into the network as an inflated convolution[3,6,9]. This inflated operation allows the model to combine the advantages of the simplicity of the dual-stream method model architecture with the excellent data processing capability of 3D convolution when dealing with large data sets. In order to enhance the ability of the I3D network to cope with deeper networks, residual connections[7] are also introduced into the model to cope with the gradient disappearance and gradient explosion that may occur in deeper networks.

A deeper and wider network model can substantially improve the overall performance of the network to better learn and process features, but abnormal action recognition pays more attention to the analysis and processing of detailed information and recognition efficiency than ordinary action recognition classification, and thinking at both levels, the way visual tempos information is processed in the network may have a very critical impact at some point. For example, in a museum, walking and running can be classified as normal and abnormal respectively, while their visual morphological differences are not significant, at this time the model can make feedback efficiently if it can quickly and accurately catch the differences in visual tempos information between the two. Traditional processing of visual tempos is mainly to build frame pyramids, using a separate network for sampling processing at each layer[8,9], and the huge amount of computation needs to consume a lot of network performance, while the introduction of temporal pyramids can directly down-sample different frame rate

inputs in one layer, which greatly saves computational costs, but such unified operation of information requires attention to the temporal context in the network. The global transfer of information, the problem of information loss in the transfer process will greatly affect the performance of the network, so this paper also makes corresponding improvements in how to better extract global information[10,11,13].

The backbone network in this paper is I3D, on which a series of improvements are made for the identification requirements of abnormal behavior, with the following main contributions:

1. Due to the large degree of influence of visual cadence information on the performance of anomalous behavior recognition systems, the introduction of a temporal pyramid structure dedicated to simulating inputs at different frame rates greatly reduces network complexity and computational complexity, improves recognition efficiency, and reduces computational costs compared to the traditional approach, which does not require the use of separate backbone networks at each layer.
2. To better assist the direct processing of temporal information by the time pyramid, a global attention mechanism is introduced and a Non-local module is inserted into the model to globally process the temporal context information to reduce the information loss problem during the transfer process.
3. In order to better handle massive surveillance data, deeper networks are needed. To address the problems of gradient disappearance and gradient explosion caused by network degradation that may be brought by training deep networks, this paper introduces residual connections in the backbone network to establish a constant mapping between layers.

2. FORMATTING OF MANUSCRIPT COMPONENTS

Video action recognition is usually a multi-category problem, where each action type can be classified into a specific result category. Feature extraction is the primary step and the cornerstone of behavior recognition, and all subsequent processing needs to be performed on the extracted features. Therefore, how to extract a robust, generalizable, and effective feature is one of the research focuses in the field of video behavior recognition. In addition to spatial scene features, the extraction of temporal features also needs to be considered for video-based action recognition.

Deep learning based convolutional neural networks have significantly advanced the development progress in the field of intelligent image recognition, and by analogy, many researchers have endeavored to migrate the already mature recognition techniques to video recognition tasks. However, video-based recognition techniques require, in contrast to image recognition, not only the acquisition of two-dimensional features of scenes in individual image video frames, but also the acquisition of temporal-series information between frames. Many researchers have given their research ideas, for example, Simonyan et al.[1] proposed Two-Stream based convolutional neural network (Two-Stream CNN), where spatial and temporal features are acquired separately using spatial and temporal stream networks, and then the features acquired from the two input streams are fused and processed, and the processed features are subsequently classified and recognized to obtain the output results[1,2,3,11,13,14]. Feichtenhofer et al.[5] explored the method of how to fuse spatio-temporal streams more effectively based on this two-stream idea, hoping to make full use of the fused features of the two for recognition and classification. And in order to solve the problem that optical stream extraction is more complicated and computationally inefficient.

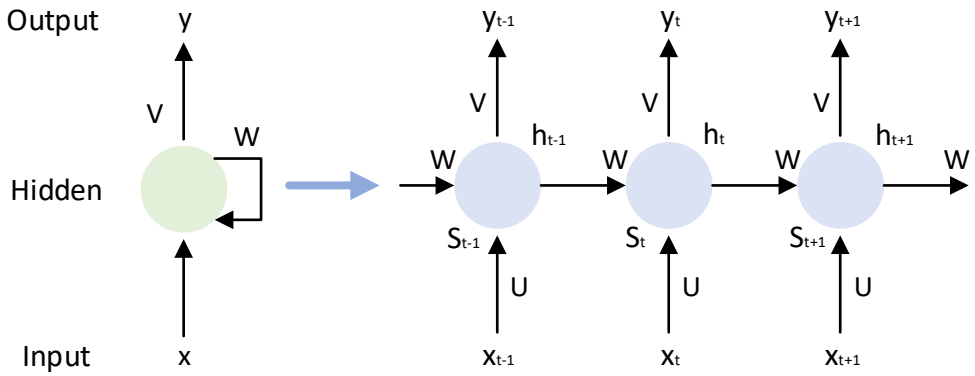

Figure 1. RNN structure diagram.

Recurrent Neural Network (RNN) has also been developed rapidly. As shown in figure 1, recurrent neural network inputs contain information from the previous moment, thus giving the network memory of the previous content, and are mainly applied to the problem of transferring historical information to videos and to the problem of capturing long-time dependencies of videos. Hochreiter et al.[20] proposed the Long Short-term Memory (LSTM) unit, which uses three different gates to achieve the preservation and forgetting of the information, compensating for the lack of gradient explosion and disappearance that existed in the initial RNN. The modeling of time series is performed using the single frame features obtained from the convolutional neural network, thus effectively improving the performance of the network for processing temporal information[4].

With the development of the era, the amount of video data is growing in huge volume, in order to better process the data, researchers try to add a temporal dimension to the two-dimensional convolutional neural network for directly processing video information. 3D convolution and 3D pooling operations can directly obtain both temporal and spatial feature information from the continuous video. Carreira et al.[6] proposed Inflated 3D ConvNet (I3D), as shown in figure 2, the number of parameters and network layers of I3D network is high. Based on the best 2D image network architecture, I3D directly inflates the coupon kernel and pooling kernel of 2D network to 3D, and follows the previous network parameters to obtain the inflated convolutional network with more layers.

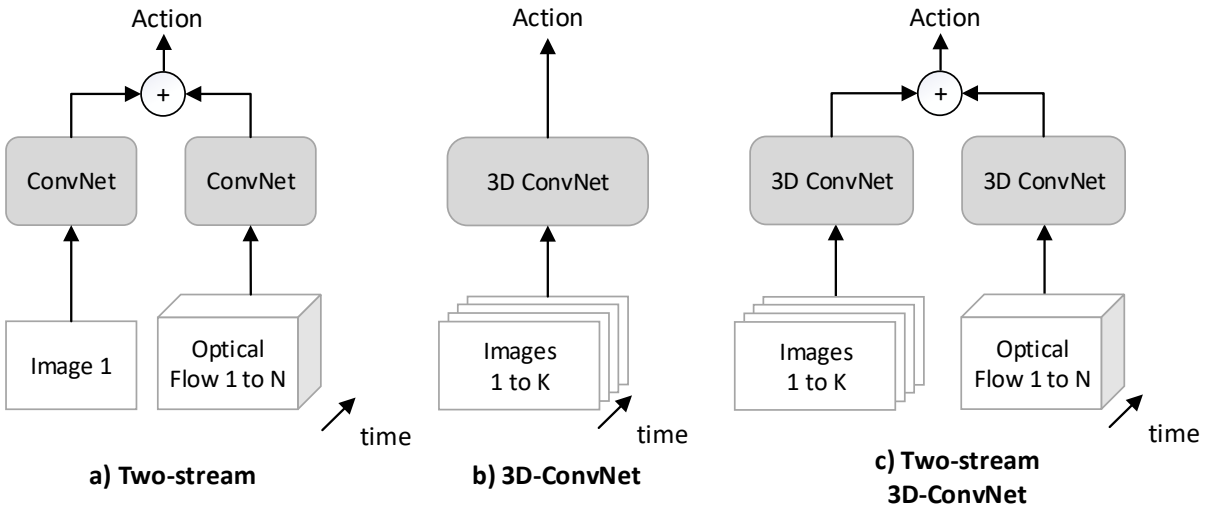

Figure 2. Structure comparison diagram between I3D and other traditional networks.

3. PROPOSED METHOD

In this paper, firstly, we adopt a two-stream based inflated 3D convolutional neural network as the main backbone network architecture, as shown in Figure 3, in which we introduce residual connections to enhance the training ability of the system for the deep network, and process spatial and temporal information in separate ways, using RGB images as input in the spatial stream processing network and optical stream images as input in the temporal stream processing network. The temporal pyramid structure is introduced in the network to process the visual tempos information, and the self-attentive non-local module is added to the main network to enhance the long time series modeling ability and the global representation ability of the information, which is used to assist the temporal pyramid module, and finally fused the obtained features effectively.

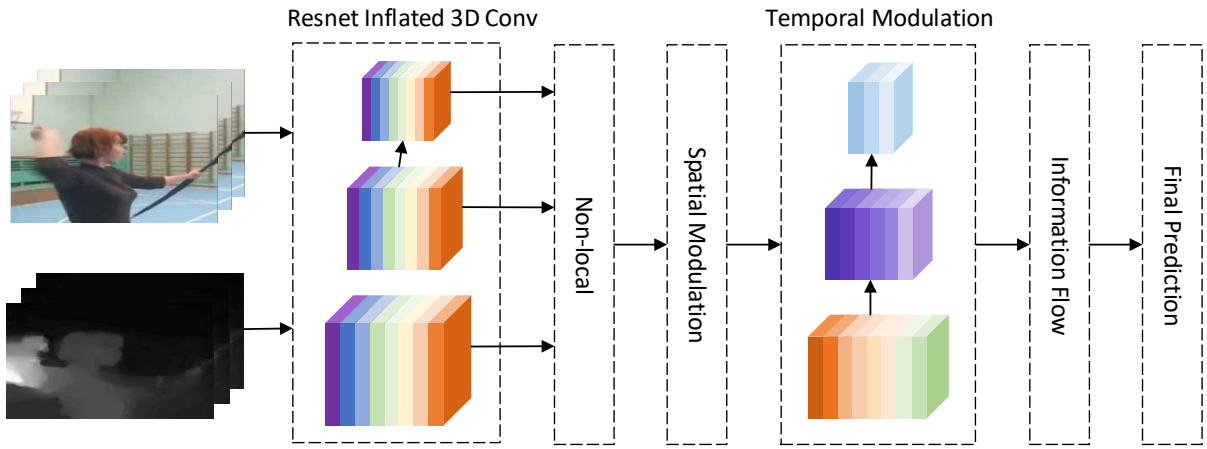

Figure 3. Our Method: Resnet I3D + TPN + Non-local.

3.1 I3D network

The Inflated Three-Dimensional Network (I3D) proposed by Joao et al[6]. skillfully uses the inflation operation in the convolutional layer, directly expanding the 2D convolutional kernel into 3D convolutional kernel and 2D pooling into 3D pooling, retaining the characteristics of two-stream convolutional networks. In addition to replacing the 2D convolutional processing with 3D convolutional processing, I3D also inflates a form of structure based on Inception matrix blocks.

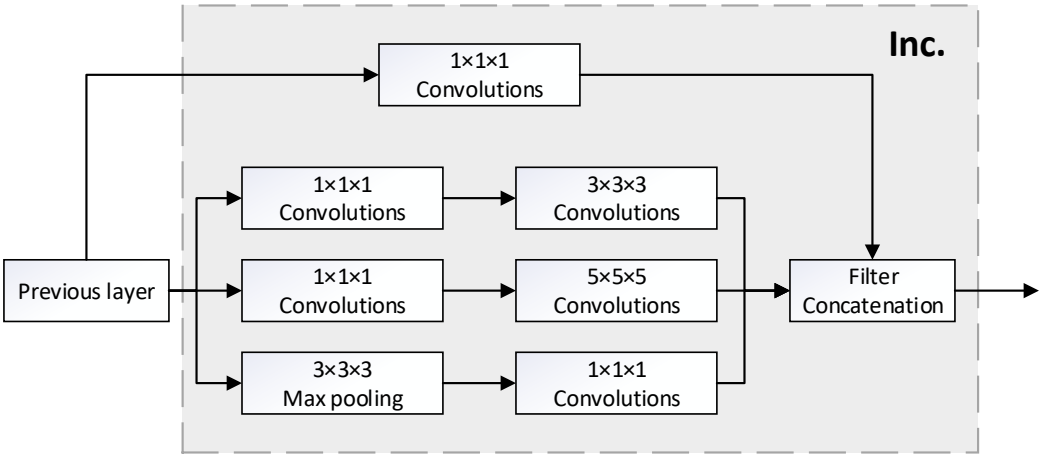

Figure 4. Inception module after expanding.

Inception emulates the repetitive stacking of neuronal cells in the human brain by clustering several sparse matrices to form a denser submatrix. The clustering approach can simultaneously take into account the sparsity of the network structure, control the number of parameters in the network, and make full use of the excellent computational performance of dense matrices. As shown in figure 4, the Inception module constructs a special neural unit structure according to this principle of forming dense with sparse clusters. The Inception module in I3D expands the two-dimensional 1 × 1 convolution, 3 × 3 convolution, 5 × 5 convolution, and 3 × 3 pooling into 1 × 1 × 1 convolution, 3 × 3 × 3 convolution, 5 × 5 × 5 convolution, and 3 × 3 × 3 pooling.

3.2 Temporal pyramid

In this paper, we plan to use the temporal pyramid structure[5,8,9,19] instead of the traditional frame pyramid and abandon the tedious practice of sampling each layer individually. When extracting features, the output features of different layers of the backbone network are directly extracted and then fed into the spatial semantic modulation network for spatial down-sampling to align feature size and shape and semantics for subsequent processing. The purpose of spatial semantic modulation is to align the receptive domain and spatial shape of features in each layer with the top layer features by using convolution with a specific step size, so that the features in the spatial dimension have the same semantics and matching shapes, and then directly simulate different frame rate inputs to temporal features by one-step down-sampling in the subsequent temporal semantic adjustment.

Since temporal pyramids unify input features in the temporal semantic adjustment layer to simulate different frame rates, while traditional frame pyramids sample each pyramid layer separately, the flexibility of temporal pyramids would be greatly limited if use the same sampling steps. We further introduced hyperparameters $\{\alpha_i\}_{i=1}^{M}$ during temporal semantic adjustment. After spatial semantic adjustment, the features are input, and the updated features at level i will be down-sampled by factor α_i. Using parametric subnetworks, the use of these hyperparameters enables the network to better handle and control the relative differences of different rhythmic information features on the time scale, thus allowing for more effective feature aggregation.

3.3 Non-local means based on self-attention

The attention mechanism is introduced to find the region with more weight and highlight the key part in the case of a large number of references, so as to obtain relevant information quickly and efficiently and establish a shortcut of communication between input and output. The self-attention mechanism differs from the attention mechanism in that instead of acquiring the weight relationship between input and output statements, self-attention acquires the relevant connections between elements within input statements or between elements within output statements[10,17,18].

The information input to the neural network is generally many vectors of different sizes, but there is a certain correlation between these different vectors, which often leads to unsatisfactory training results because of the inability to accurately obtain the correlation information between these different input vectors. For example, when dealing with lexical annotation, machine translation, semantic analysis and other textual problems, whether the contextual association information is correctly and completely obtained can greatly affect the overall recognition results of the final network.

As shown in the figure, the non-local operator is a module based on self-attentiveness, and this operator can directly capture long-range dependencies by computing the interaction between any two positions with the general formula as :

$$y_i = \frac{1}{C(X)} \sum_{\forall j} f(x_i, x_j) g(x_j) \tag{1}$$

Where j is the index of all positions, i is the index of the output position, x is the input signal, and y is the index of the output signal which has the same size as x. The binary function f calculates the scalar relationship between the input signal at index i and index j. The unary function g calculates the representation of the output signal at the index j, and C(x) is the normalization factor.

3.4 Residual connection

With the expansion of the data set and the increase of the requirements for the accuracy of the research results, the network design of deep learning is bound to go to the deeper and wider network, which brings a very fatal problem that as the number of network layers deepens, the gradient gradually disappears leading to the deeper network layers cannot be optimized and utilized, so that the theoretical deeper network will have better performance can not become a reality, and even in the This phenomenon has bottlenecked the development of deep learning until the emergence of the residual structure[7,9,16,18] , which makes people have a new research idea to deal with the possible network degradation problem of deep networks.

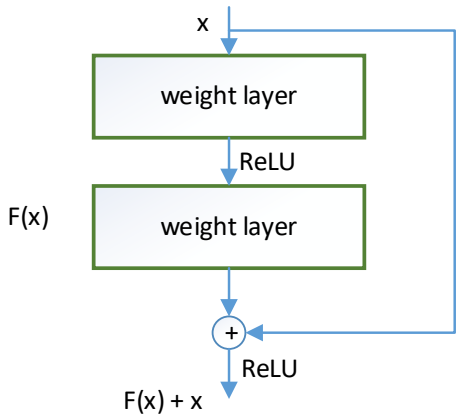

Figure 5. Diagram of the residual connection structure.

In brief, the residual structure introduces a residual connection that creates a constant mapping on the input and output, forming a shortcut as in figure 5. The objective function of the residual module is $H(x) = F(x) + x$, where $H(x)$ is the output function, and $F(x)$ represents the mapping, and$x$ represents the input. This operation, at the mathematical level, means that when deriving the gradient for each layer, a summation term is introduced in the computational equation for the derivative of the generic multiplication, avoiding that the increasingly small values of the derivatives make their product very close to zero, which leads to a network that eventually cannot be trained.

4. EXPERIMENT

4.1 Data set

This paper uses the mainstream abnormal action dataset UCF-Crime, which has a surveillance video duration of 128 hours and includes 1900 video ensembles recorded from surveillance cameras. In addition to some normal behavior accidents, the types of abnormal behavior in the dataset cover urban road traffic accidents, human behavior such as fights and brawls in public places, and public safety events such as object explosions. The average frame rate of the videos is 7274 frames.

4.2 Results

In this paper, we first try to add a layer of the temporal pyramid, and a piece of Non-local operator to the residual 3D network, and compare other existing advanced networks on the UCF-Crime dataset, in addition to comparing the effect of the network without the addition of the Non-local block and without the optical flow, as the results in table 1 show that the accuracy of our improved model has improved and the model has validity.

Table 3. Information on video and audio files that can accompany a manuscript submission.

Model	Frames	Flow	TPN	Non-local	Acc.
CSN-101	32				92.3
R(2+1)D	16	√			90.9
I3D	16	√			90.0
SlowFast-R50	32				92.6
Two-Stream I3D	64	√			92.0
	32			√	92. 6
	32×2		√		93. 3
Our Method	**32×2**	√	√	√	**94.2**

Since the non-local block and the temporal pyramid module are both flexible and plug-and-play, and can be easily interspersed in different networks, and the residual network is known for its effectiveness in training deeper networks, this

paper tries to add the non-local operator and the temporal pyramid module into different layers of the residual network to test how well they work in networks of different depths. Based on this, the results in figure 6 are obtained in this paper.

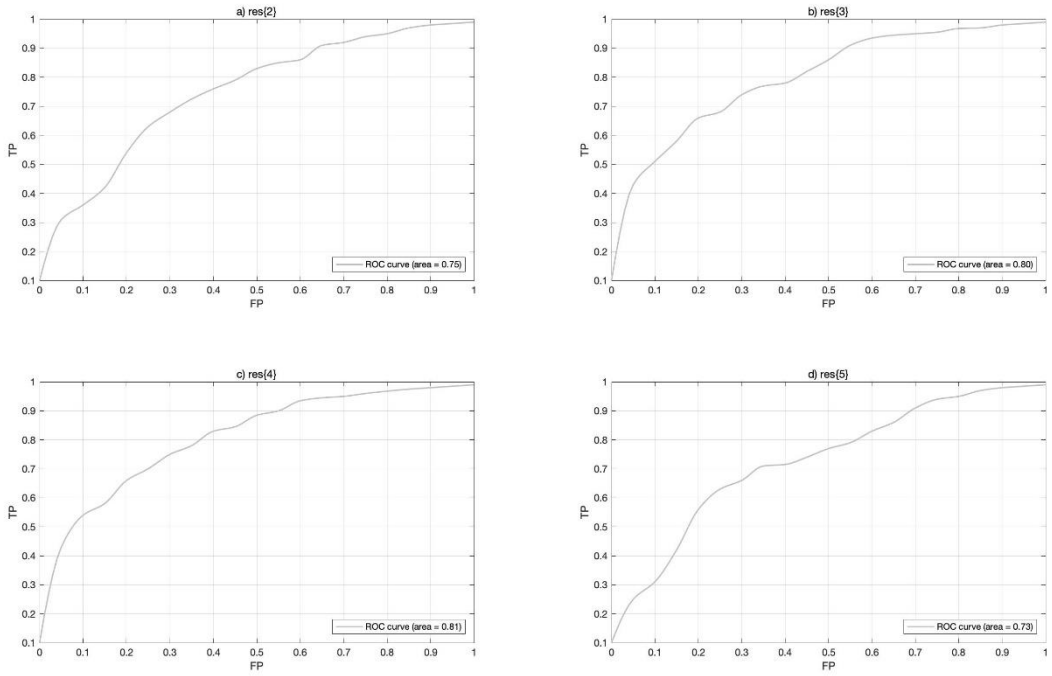

Figure 6. Comparison of ROC at different network layers.

The experimental results show that the insertion position of the module does affect the effect it plays out, and it can be seen in figure 6 that the insertion of the module in the middle layer works better than the other layers, which should be due to the fact that in deeper networks, the extracted higher-order features already contain little spatial information, so the long-range features that can be learned are very limited, and at this point, the global operation for long time series modeling can play a role is no longer very prominent.

Finally, we chose to insert a time pyramid in res{4,5} based on 5 Non-local modules, as shown in the figure 7, to be able to reflect which segment of the video has anomalous behavior by the height of the curve in a video.

Figure 7. System identification result: Breaking the public property.

5. CONCLUSION

In this paper, we focus on the processing of visual tempos information which is often overlooked by researchers, which also refers to the frequency and rapidity of the movement of the characters in the video, and this information will directly affect the overall judgment result of the behavior to some extent. Also considering that surveillance videos generally have a long time and need to focus on the extraction of temporal context information, this paper effectively combines temporal pyramids and Non-local operators, which can directly simulate the input of different frame rates while calculating global attention weights and establishing the connection between long interval pixel points in one step. Experiments show that the combination of the two can effectively improve the recognition of anomalous behavior. However, the depth of the network does not reach our ideal expected depth due to the limitation of hardware conditions such as the performance of existing computer equipment. In the future, with the follow-up of hardware conditions, this system is also intended to be trained and tested in deeper networks.

REFERENCES

[1] Simonyan, K., Zisserman, A. “Two-stream convolutional networks for action recognition in videos,” In NIPS. (2014) 568–576.

[2] A. Diba, M. Fayyaz, V. Sharma, M. M. Arzani, R. Yousefzadeh, J. Gall, Luc Van Gool. “Spatio-temporal channel correlation networks for action classification,” In Proc. ECCV, 2018.

[3] C. Feichtenhofer, A.l Pinz, A. Zisserman. “Convolutional two-stream network fusion for video action recognition,” In Proc. CVPR, 2016.

[4] Tran D, Bourdev L, Fergus R, et al. “Learning Spatiotemporal Features with 3D Convolutional Networks,” IEEE, 2015.

[5] C. Feichtenhofer, H. Fan, J. Malik, K. He. “Slowfast networks for video recognition,” In Proc. ICCV, 2019.

[6] J. Carreira, A. Zisserman, Quo vadis. “Action recognition? a new model and the kinetics dataset,” In Proc. CVPR, 2017.

[7] K. He, X. Zhang, S. Ren, J. Sun. “Deep residual learning for image recognition,” In Proc. CVPR, 2016.

[8] C. F. , H. Fan, J. M. , K. He. “Slowfast networks for video recognition,” In Proc. ICCV, 2019.

[9] C. Yang, Y. Xu, J. Shi, et al. “Temporal Pyramid Network for Action Recognition,” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.
[10] X. Wang, R. G. , A. Gupta, K. He. “Non-local neural networks,” In Proc. CVPR, 2018.
[11] A. F , P. M , M. S . “Learning optical flow from still images,” 2021.
[12] G. W. Taylor, R. Fergus, Y. LeCun, C. Bregler. “Convolutional learning of spatio-temporal features,” In Proc. ECCV, 2010.
[13] Z. Huang, S. Zhang, J. Jiang, et al. “Self-supervised Motion Learning from Static Images,” 2021.
[14] L. Wang, Z. Tong, B. Ji, et al. “TDN: Temporal Difference Networks for Efficient Action Recognition,” Computer Vision and Pattern Recognition. IEEE, 2021.
[15] Z. Li, Y. A. Farha, J. Gall, “Temporal Action Segmentation from Timestamp Supervision,” Conference on Computer Vision and Pattern Recognition (CVPR), IEEE/CVF, 2021, pp. 8361-8370.
[16] L. Peng, S. Todorovic. “Temporal deformable residual networks for action segmentation in videos,” In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pages 6742–6751.
[17] Y. Liu, J. Zhang, L. Fang, Q. Jiang, B. Zhou, “Multimodal Motion Prediction with Stacked Transformers,” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7573-7582.
[18] Z. Wang, Z. Gao, L. Wang, Z. Li, G. Wu. “Boundary-aware cascade networks for temporal action segmentation,” In European Conference on Computer Vision (ECCV), 2020.
[19] Y. Wang, M. Long, J. Wang, Philip S. Yu. “Spatiotemporal pyramid network for video action recognition,” In Proc. CVPR, 2017.
[20] Hochreiter, S, J. Schmidhuber. “Long short-term memory,” Neural Computation 9.8, 1997, pp.1735-1780.