Intelligent Human Systems Integration (IHSI 2025), Vol. 160, 2025, 103–109
https://doi.org/10.54941/ahfe1005815

An Action Recognition Method Based on 3D Feature Fusion

Yinhao Xu and Yuanyao Lu

North China University of Technology, Beijing, 100144, P. R. China

ABSTRACT

Based on the well-known two branch network SlowFast, this paper introduces a significant improvement. Specifically, we propose an enhanced SlowFast network named ESL Net. A key innovation in this network is the addition of an improved feature fusion module. This module is designed to make the most of the temporal information available in the video for effective feature fusion. It employs channel and spatial attention mechanisms to precisely identify the most significant parts of the features. By analyzing the temporal information, it can also determine the crucial elements between dual-temporal features. Extensive experiments have demonstrated that our proposed method is highly effective when applied to the UCF-101 dataset and the HMDB51 dataset, showing superior performance compared to existing methods especially SlowFast Network in terms of accuracy and robustness in human action recognition tasks.

Keywords: Human action recognition, 3D-feature fusion, Two branch network, SlowFast

INTRODUCTION

Human action recognition is a hot topic in the field of computer vision. With the development of deep learning, the research on human action recognition in videos has become more and more mature, and it is widely used in daily life such as surveillance systems, human computer interaction, and intelligent care. Human behavior recognition methods can be divided into two major categories: traditional methods and deep-learning based methods. Traditional behavior recognition methods mainly rely on human judgment to extract valuable information and design feature extraction methods. These methods developed early. Before deep learning became popular, traditional methods for behavior recognition were the main means and research objects at that time. However, the features extracted by traditional behavior recognition methods are not sufficient, and the rise of deep learning has made this machine learning method that relies on large amounts of data become the mainstream. In behavior recognition methods based on RGB data, CNN is usually used to extract the spatial features of video frames, but there are different practices in the extraction methods of motion information, which are mainly divided into three categories: two stream network, three dimensional convolutional networks, and recurrent neural networks.

To capture the spatio-temporal information in the video frame sequence, the two-stream framework was first proposed in (Feichtenhofer et al., 2019).

This framework consists of two separately operating CNNs. One of them extracts the spatial information in a single frame RGB image, and the other extracts the motion information from the video optical flow sequence. The two groups of features will complete feature fusion in the last classification layer. However, the two stream network has the problems of complex optical flow extraction operation and high computational cost, and it is not very friendly to large scale data sets or devices with limited training resources.

Using sequential models such as recurrent neural networks in conjunction with CNN to complete behavior recognition is also a common idea. Its essential idea is similar to that of the two stream structure. The difference is that a sequential model is used instead of optical flow operations to model motion information. However, RNN networks inherently have weak memory ability for long sequences, and there may be fatal problems such as gradient vanishing or gradient explosion, which make them difficult to train.

The three-dimensional convolution-based method uses three-dimensional convolution kernels to simultaneously extract spatio-temporal features and increase the correlation of spatio - temporal features. Carreira and Zisserman (2017) proposed the classic C3D network, which directly regards the video as a three-dimensional spatio-temporal structure and uses three-dimensional convolution to simultaneously extract its spatial and motion features. The feature extraction method of C3D makes the network operation more efficient. Therefore, we continue to use the 3D CNN method to uniformly model spatio-temporal information, and make improvements on the two-branch network such as SlowFast. By adding a feature fusion module, the effect of improving the network is achieved.

ROPOSED METHOD

In the retinal nerve cells of primates, 80% of the cells work at a relatively low rate. They are not sensitive to visual motion changes but can provide good spatial and color details. The remaining 20% of the cells work at a relatively high rate and can sensitively capture motion changes. Thus, He Kaiming creatively proposed a two-branch network - SlowFast. The convolution kernels in the Slow Path are used to extract appearance features, and the total number of parameters accounts for about 80% of the total parameters of the two channels. While the convolution kernels of the Fast Path are used to extract motion features, and the total number of parameters accounts for about 20%. Compared with Fast, Slow has a relatively lower frame rate but has more channels. Slow is used to capture semantic information in space, that is, Slow pathway captures the relatively static information in the video.

Slow pathway and Fast pathway do not exist independently. The information fusion between the two is one-way information fusion. The two achieve information fusion through multiple lateral connections. The direction of the lateral connection is from Fast to Slow, which means that Fast will not receive any information about Slow, but Slow can contain the information in Fast. We believe that this kind of information fusion will undoubtedly lead to information loss, that is, insufficient information utilization, which is not beneficial to improving the accuracy and robustness

of the network. Therefore, it has become an important direction of our research.

The specific network structure of the method we proposed is shown in the figure below.

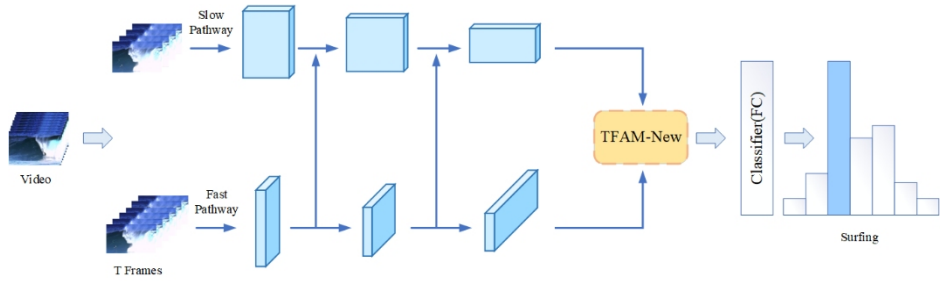

Figure 1: The model structure diagram of our method.

Currently, there are more and more ways of feature extraction and feature fusion. This is also because merging features (or feature extractors) from multiple different sources into a better feature representation can help machine learning models better understand data, thereby improving the performance and generalization ability of the model. Our feature fusion module refers to the approach of the author of TFAM and makes some improvements on this basis to propose a new structure, Temporal Fusion Attention Module (TFAM)-New, to better adapt to our human action recognition task. Specifically, TFAM-New also includes two branches: the temporal branch and the spatial branch.

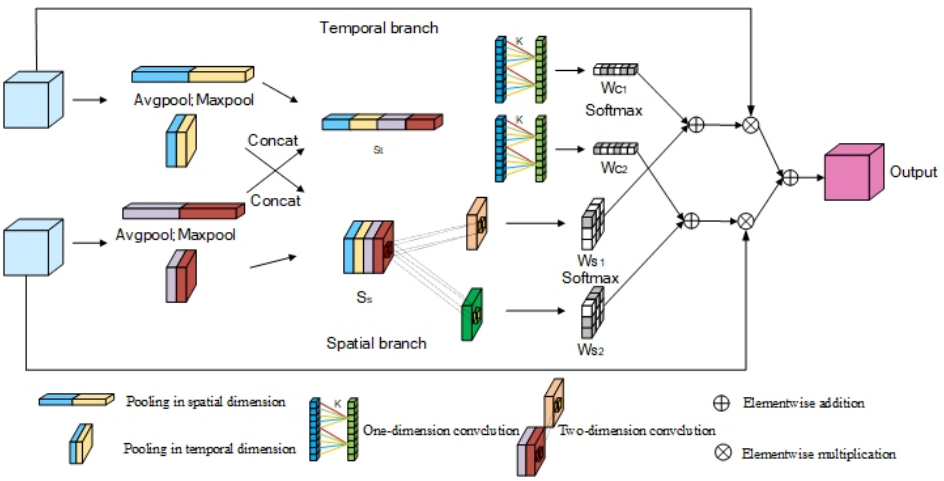

Figure 2: TFAM-new module structure diagram.

The temporal branch is used to enhance the attention to temporal information, and the spatial branch is used to enhance the attention to spatial information. Perform spatial dimension pooling and temporal

dimension pooling on two input features simultaneously, then concatenate the spatial dimension pooling and temporal dimension pooling respectively, then determine the weight, retain useful information, and finally separate. Through weight adjustment, the more valuable parts of the dual-temporal features are retained, while unimportant or misleading information is suppressed, thereby improving the accuracy and robustness of change detection.

Since the feature map sizes obtained by the slow pathway and the fast pathway after sampling at different sampling rates are different, we decided to expand the information of the slow pathway to be consistent with the fast pathway in the T dimension. There are two methods for expansion: one is similar to the padding operation in convolution calculation, and the other is to directly copy and splice the feature map. Experiments have proved that the second method is more effective. Therefore, we use the feature reuse method to convert the feature maps of the two branches into the same size for subsequent feature fusion operations. The TFAM-New module diagram is shown as Figure 3.

EVALUATION OF THE PROPOSED METHOD

To test the effectiveness of the method we proposed, we conducted tests on two publicly available datasets in the direction of human action recognition, namely the UCF-101 dataset and the HMDB51 dataset.

Figure 3: Sample of UCF-101 dataset.

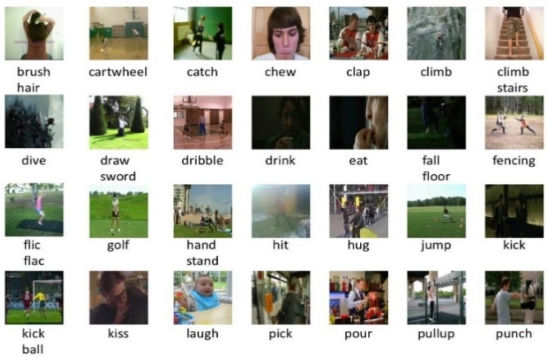

Figure 4: Sample of HMDB51 dataset.

In the process of model evaluation, we are well aware that a single evaluation index often has limitations and cannot comprehensively and accurately reflect the performance of the model. Therefore, considering that the evaluation index cannot be singularized, we carefully select two important indicators, ACC and recall rate, as the criteria for evaluating the model in the experimental part.

Accuracy can intuitively reflect the proportion of correct predictions made by the model, while recall rate focuses on measuring the model's ability to recognize positive samples. By considering these two indicators simultaneously, we can evaluate the performance of the model more comprehensively and objectively, providing a powerful basis for the optimization and improvement of the model.

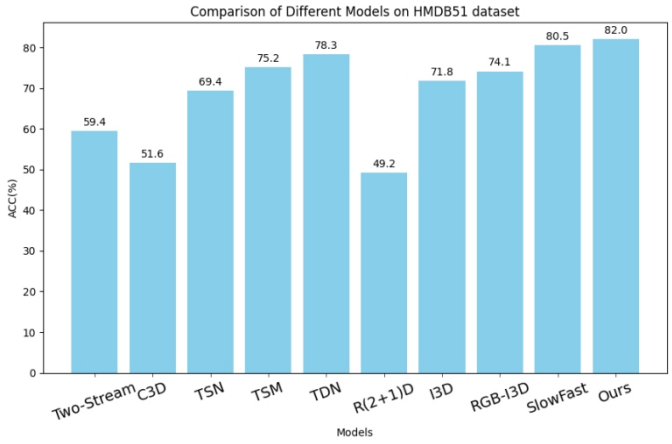

Figure 5: Comparison of different method on HMDB51 dataset.

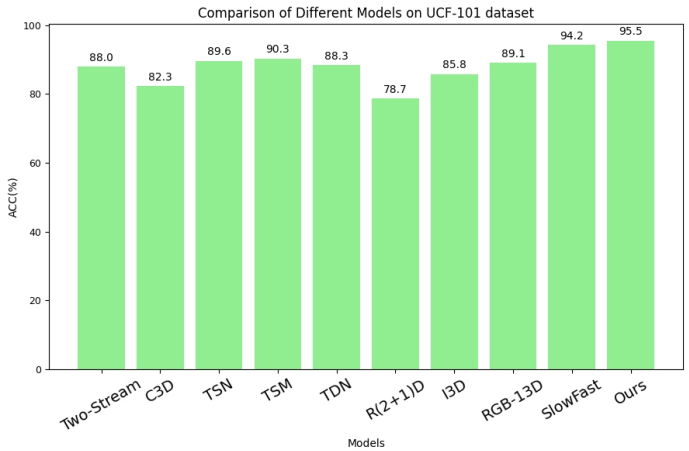

Figure 6: Comparison of different method on UCF-101 dataset.

Table 1. Comparison table of model recall rate.

Models	Two-Stream	C3D	TSN	TDN	R(2+1)D	I3D	RGB-I3D	SlowFast	Ours
UCF-101	68%	65%	77%	78%	60%	70%	78%	83%	86%
HMDB51	57%	51%	64%	66%	44%	66%	71%	75%	77%

CONCLUSION

After a series of rigorous and comprehensive experiments, we clearly saw the final experimental results. These results strongly indicate that the method we adopted is truly effective in the specific task of action recognition. Specifically, when we applied this method to test on the HDMB51 dataset, it demonstrated an accuracy rate as high as 82%. It should be noted that the HDMB51 dataset covers a large number of behavioural data of many different types and in various scenarios. Achieving such an accuracy rate in such a complex and representative dataset is sufficient to illustrate the reliability of our method. Moreover, on the UCF-101 dataset, which is widely used in action recognition research, our method performed even more outstandingly, with an astonishingly high accuracy rate of 95.5%. This fully reflects that the model we constructed has more excellent capabilities in feature extraction. With the addition of the improved module, both the accuracy rate and the recall rate have witnessed significant increases under the original SlowFast framework. The improvement in the accuracy rate means that the proportion of correct judgments made by the model for action recognition has become higher, while the increase in the recall rate indicates that the model can accurately identify more actions that should have been correctly recognized and avoid omissions. Such a significant improvement in these indicators undoubtedly proves that the improved module plays a non-negligible role and can effectively guide the entire model, enabling the model to operate and play its role in a more accurate and efficient direction in related tasks such as action recognition and video understanding.

ACKNOWLEDGMENT

This paper was supported by the National Natural Science Foundation of China (Grant Nos. 61971007 and 61571013).

REFERENCES

Carreira, João and Andrew Zisserman. (2017) 'Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset.' Available at: https://arxiv.org/abs/1705.07750.

Feichtenhofer, Christoph, Haoqi Fan, Jitendra Malik and Kaiming He. (2019) 'SlowFast Networks for Video Recognition'. Available at: https://arxiv.org/abs/1812.03982.

Simonyan, K., & Zisserman, A. (2014). 'Two-Stream Convolutional Networks for Action Recognition in Videos'. Available at: https://arxiv.org/abs/1406.2199.

Soomro, Khurram, Amir Zamir and Mubarak Shah. (2012) 'UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild'. Available at: https://arxiv.org/abs/1212.04021.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2014) 'Learning Spatiotemporal Features with 3D Convolutional Networks'. Available at: https://arxiv.org/abs/1412.0767.

Tran, Du, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun and Manohar Paluri. (2017) 'A Closer Look at Spatiotemporal Convolutions for Action Recognition'. Available at: https://arxiv.org/abs/1711.11248.

Wang, Limin, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang and Luc Van Gool. (2016) 'Temporal Segment Networks: Towards Good Practices for Deep Action Recognition'. Available at: https://arxiv.org/abs/1608.00859.

Zhao, Sijie, Xue-liang Zhang, Pengfeng Xiao and Guangjun He. (2023) 'Exchanging Dual-Encoder–Decoder: A New Strategy for Change Detection With Semantic Guidance and Spatial Localization'. Available at: https://arxiv.org/abs/2311.11302.