2024 International Wireless Communications and Mobile Computing (IWCMC) | 979-8-3503-6126-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/IWCMC61514.2024.10592362

An Action Recognition Algorithm Based on Two-Stream Deep Learning for Metaverse Applications

Jiayue Liu[1,2], Tianqi Mao[1,2], Yicheng Huang[3], Dongxuan He[4]
[1]MIIT Key Laboratory of Complex-Field Intelligent Sensing, Beijing Institute of Technology,
Beijing 100081, China
[2]Yangtze Delta Region Academy, Beijing Institute of Technology (Jiaxing), Jiaxing 314019, China
[3]Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[4]School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China
E-mails: {jiayue_liu@bit.edu.cn, maotq@bit.edu.cn, huangyic20@mails.tsinghua.edu.cn, dongxuan_he@bit.edu.cn}

***Abstract*—Action recognition algorithms have gained significant attention in recent years, which can be indispensable for a plethora of cutting-edge applications like extended reality or Metaverse. These services often pose stringent requirement on immediate sensing and cognition of the surroundings, which necessitates immediate classifications of the captured actions (e.g., video data) that classical signal processing methods can hardly attain. In this paper, we introduced a residual artificial neural network with two-stream structure to further improve the accuracy of action recognition algorithm. Specifically, two residual networks (ResNet101) are trained separately, one by spatial RGB image streams, and another by optical flow streams. The two-strem network outputs are then fed into a fusion classifier, in which information extracted by spatial network and temporal network jointly determines the classification result. Moreover, in the training process, hyper-parameters setting and optimizer selection are performed numerically to achieve optimal performance. Finally, the recognition accuracy of the proposed algorithm has been compared to other existing widely-employed counterparts, where UCF101 data set is utilized for training and testing. Simulations validates aiming that the network can achieve higher recognition accuracy than traditional algorithms, and the two-stream method shows its superiority over the single-network counterpart.**

***Index Terms*—Action recognition, deep learning, residual network, ResNet101, two-stream method.**

I. Introduction

The development of the new interactive forms such as virtual reality, and the proposal of the concept about metaverse lead to revolution of information technology. However, due to limited basic supporting technology, the interaction methods mentioned above are far from popularized. According to the concept of metaverse, applications will operate on basis of large amount of data. It is necessary to classify the raw video data that collected by interaction devices for subsequent processing works, especially those with human movements which usually contains critical information [1]. Action recognition algorithms are used to solve this type of problem, which are designed to identify the content in the video, e.g., the actions of people, extract its information, and classify the video according to the the acquired information automatically [2].

Traditional action recognition methods generally use manually designed feature templates to describe and identify actions [3]. The features of target video file are extracted, and then compared with the preset action template. If the difference is less than a threshold, the action in this video will be recognized as the specific one represented by this template. This type of method can recognize simple actions with relatively low computational complexity and high robustness. However, for complex actions or same actions at different angles, a large number of templates are required to ensure accuracy, which leads to undesirable increase in the calculation overhead and cost of template selection.

From the beginning of 21st century, with the development of machine learning and construction of the action recognition dataset, action recognition has made a breakthrough [4]. The accuracy of image recognition algorithms based on deep learning have been validated to exceed algorithms, which provide inspiration for certain areas such as action recognition and video classification. In [5], a 3-dimension (3D) conovolutional network (Conv3D, C3D) was designed for video data processing inspired by the 2D counterpart, which performed well in picture processing. Treating the video as a 3D vector, the 3D kernel structures enable C3D to capture motion features for action recognition applications. However, due to the size limitation of the input data, C3D can only recognize motion information in limited timespans, and cannot memorize information for long time. Further researches in C3D have provide enhanced solutions such as Inflated 3D ConvNet(I3D) [5] which utilizes the existing modules of ImagNet to accelerate training process and Separable 3D ConvNet S3D [6] which increases network speed and compressed network size by combining low dimensional network, the main problem of C3D remains unsolved. Long-Short Term Memory (LSTM) was introduced to solve this issue [7]. LSTM is a kind of Recurrent Neural Network, which consists of three parts: input gate, forget gate and output gate. These structures enables LSTM network to “remember” important action information for a long period and forget others in a short time. Therefore, LSTM has become a commonly used action recognition algorithm. However, LSTM is difficult in training and prone to overfitting because of its numerous parameters and complex structure.

In this paper, an action recognition algorithm based on two-

979-8-3503-6126-1/24/$31.00 ©2024 IEEE

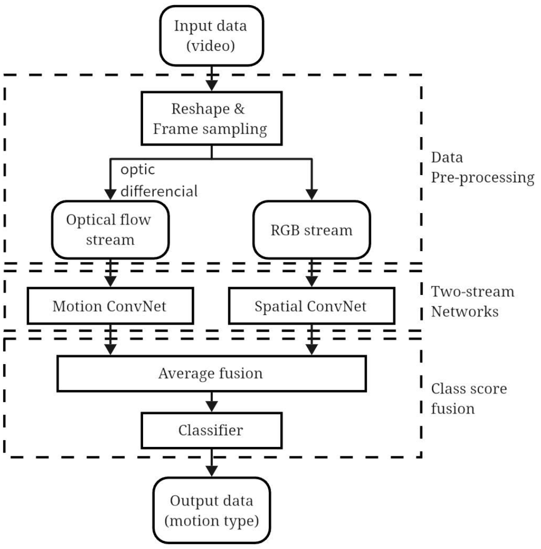


Fig. 1: Structure of the system model with two-stream structure.

stream method and deep learning neural network is introduced for accurate video classification. Specifically, the two-stream network structure is applied to process the video data in temporal and spatial modalities separately, and combine these result as output [8]. This method utilizes the complementarity of two modalities to improve classification accuracy [9]. ResNet101 is used as the core network for the two-stream network. It contains deep layers to extract video information as well as residual layers structures to reduce gradient explosion and vanishing gradient problems caused by the deep structure and ensuring the stability of the network. Therefore, it has performed well in image and video processing works [10], which is suitable for the proposed algorithm. Moreover, experimental training is introduced to discover more suitable parameters and optimizer settings for the network, which enables the network attain its maximum recognition accuracy.

II. Proposed Network

This section mainly introduces the model structure of the action recognition algorithm. The network structure starts by a pre-processing model, in which the input video is processed and divided into RGB and optical flow (OF) data stream. Then the two data streams are input separately into spatial network and motion ConvNet. After that, the output of spatial network and motion network are sent into a fusion layer, in order to form the final classification result. The model structure is shown as Fig. 1,

A. Two-stream Method

The overall process of the proposed two-stream network can be divided into four parts: data preprocessing, special ConvNet, motion ConvNet, and Class score fusion. The data preprocessing part reshapes the video data to appropriate size, and then generate RGB stream and optical flow (OF) stream based on the reshaped video $I(x, y; t)$. The RGB stream can be obtained directly by sampling the video frames, and output as a data stream with three channels $R(x, y; t)$, $G(x, y; t)$, and $B(x, y; t)$. The OF stream needs differential calculation to obtain its motion vectors. For each pixel $I(x, y; t)$ in the video, assume that the brightness remains unchanged, there is $I_x u + I_y v + I_t = 0$ in which $I_x = \frac{\partial I(x,y;t)}{\partial x}, I_y = \frac{\partial I(x,y;t)}{\partial y}, I_t = \frac{\partial I(x,y;t)}{\partial t}$, and (u, v) represents the OF vectors. The vectors are output in the form of OF stream with two channels $u(x, y; t)$ and $v(x, y; t)$. The spatial ConvNet and motion ConvNet are two separate networks with same network structure. The spatial network uses RGB stream as input, while the motion network uses OF stream as input. Both of these network output a vector composed by classification score representing the probabilities for each label $P_{RGB} = F_{spatial}(R, G, B)$ and $P_{OF} = F_{motion}(u, v)$. The class score fusion part combines the output of two networks to obtain average probability $P = E(P_{RGB}, P_{OF})$, and decides the action type according to the label that have the highest average probability.

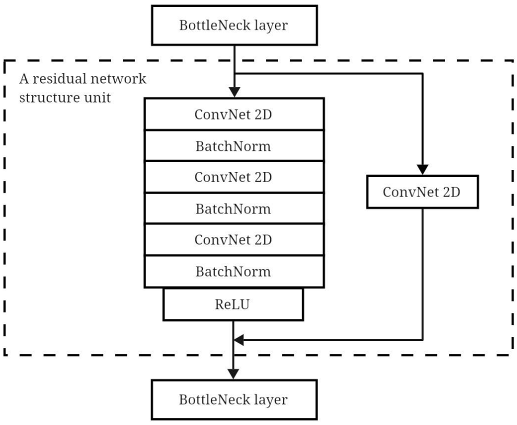


Fig. 2: The structure of one ResNet structrual unit, which is in the dashed box.

B. Residual Network

In this paper, ResNet101 is selected as the spatial and motion network. ResNet101 is a residual network (ResNet) that utilizes feedforward networks and bottleneck layers to prevent gradient explosion and vanishing, and thus enable the network to have greater depth compared to traditional neural networks. ResNet contains multiple structural units $y(x) = F(x) + x$, which consists of a group of convolutional layers $F(\cdot)$ and a short branch named residual layer. Such structural units are also in ResNet101, which structure are shown as Fig. 2. The convolutional layers and batch normalization layer are alternately connected, which forms a structure to recognize more abstract information. The residual layer is a shortcut connection that adds the input to the output of the convolutional layers to prevent gradient explosion and vanishing caused by the depth of the network. Each input and output of the structure units connect a Bottleneck layer, which can reduce the numbers of parameters in propagation, and prevent

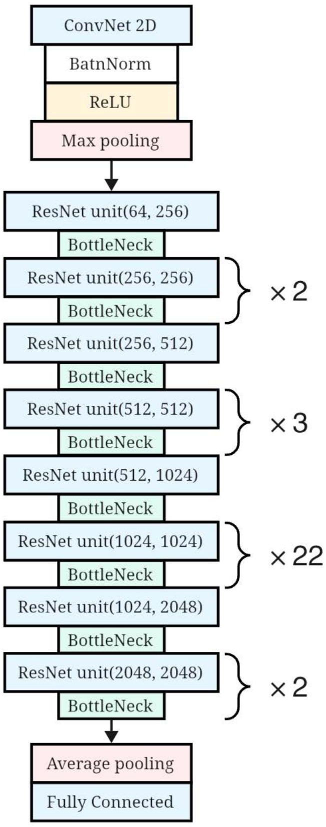


Fig. 3: The structure of ResNet101, the structure of ResNet unit is shown as Fig. 2, and the numbers represents the numbers of input and output channels.

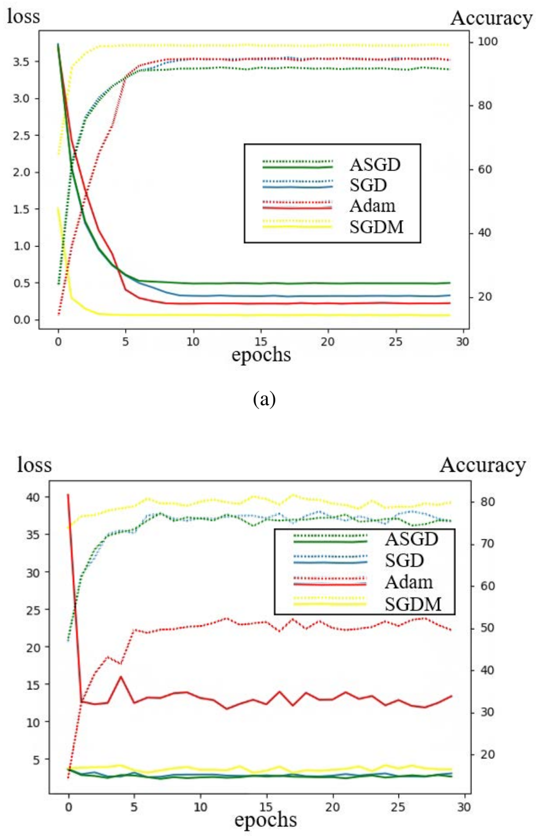
the deep network from deterioration. ResNet101 consists of three parts called input network, hidden network and output network. The input layer contains a convolutional layer with batch normalization, an activation function(ReLU) layer, and a max pooling layer. The hidden network contains certain numbers of repeated structural units. The output network contains an average pooling layer and a fully connected layer as shown in Fig. 3:

III. Network Training

This section mainly provides details about the training, including the dataset, the settings of hyper-parameters and the training process.

A. Dataset UCF101

The UCF101 dataset is a open source action recognition dataset, which contains 101 categories of videos of real human actions collected from YouTube [11]. There are significant differences in background, lighting, shooting angles, and other conditions about the videos in the UCF101 dataset, providing a rich variety of action information. In addition, this dataset is constructed based on real video which improves generalization ability of the network. There are also quite a lot preprocessed version that are accessible. In the training process, a UCF101 dataset that is sampled and preprocessed to RGB and OF data is used, so that the preprocessing work during training can be omitted. After obtaining the UCF101 dataset, it is necessary to divide it into training, validation, and testing sets in a 6:2:2 ratio. The segmentation needs to ensure uniformity, so that each motion category has video data included in each sub-dataset to ensure the accuracy and completeness of the training data.



(a)

(b)

Fig. 4: The result about the experimental training for optimizer. The four colors represent four kinds of optimizer which are Stochastic gradient descent (SGD), Averaged stochastic gradient descent (ASGD), Stochastic gradient descent with momentum (SGDM), Adaptive Moment Estimation (Adam), (a) Training loss in solid line and accuracy in dashed lines; (b) Testing loss in solid line and accuracy in dashed lines.

B. Training process and hyper-parameters setting

The spatial and motion network are trained separately. The input data for spatial network is 224×224 pixels RGB image stream, which is divided into three channels based on primary colors, while the input data for motion network is 224×224 pixels OF image stream, which is divided into two channels based on direction of vectors. The hyperparameters are setted as Table I .

In back propagation process, the labels representing the type of motions are seen as the input x, while cross entropy $E(x) = \sum P_i(x) log(Q_i(x))$ is used as the loss function. Dur-

TABLE I: Parameter setting for network training

Parameters	Spatial network	Motion network
epochs	100	100
batch size	16	12
learning rate(initial)	5e-4	5e-4
learning rate(renew)	exponential	exponential

ing this process, optimizer is used to update the parameters. We have researched four different optimizer in an experimental training with the same settings in 30 epochs. Fig. 4 shows that Stochastic Gradient Descent with Momentum (SGDM) has the lowest training loss and highest training accuracy. In addition, it also shows well in testing loss and accuracy. Therefore, SGDM is selected as the training optimizer in back propagation process.

IV. Experimental Result

In this section, the result of the action recognition system is presented. In addition, it is also compared with the existing algorithms.

The testing dataset is obtained from UCF101 dataset, which includes 20% of the video data from original dataset. The testing indicators are top1 and top5 recognition accuracy. Tested by the data that are extracted from 2664 video in dataset, the proposed network achieved a **top1 accuracy of** 85.75% **and top5 accuracy of** 95.48%.

Additionally, several different algorithms are tested in the same condition to show the superiority of the proposed scheme. In the experiment, action recognition algorithms composed of ResNet only and ResNet with OF are trained in the same hardware condition with the same hyperparameters and dataset. Meanwhile, algorithm with 3D ConvNet (C3D) and Long-Short Term Memory (LSTM) network are trained with the same epochs and appropriate parameters. Once well trained, these algorithms are tested by the same way. Recognition accuracy acts as an evaluation indicator, where the result is shown as Table II.

TABLE II: Testing result compare to other algorithms

Algorithms	Accuracy top1	Accuracy top5
ResNet101 only	78.59%	/
ResNet101 with OF	73.49%	/
C3D	77.41%	/
LSTM	81.26%	/
two-stream ResNet	**85.75%**	**95.48%**

It is obvious that the recognition accuracy of ResNet is only lower than the proposed algorithm, since it is not sensitive for temporal information. ResNet with OF performances is even worse, for it ignores the spatial information. Such result indicates that the two-stream structure is quite effective in recognizing motion information and improving the accuracy. Moreover, the result shows the proposed algorithm get better performance in accuracy over C3D and LSTM which are widely used at present.

V. Conclusion

This paper proposes an algorithm for action recognition, which adopts ResNet with two-stream method. To improve the recognition accuracy, experimental training is applied aiming to select the optimal hyperparameters. Through the experiment, the proposed algorithm has reached a high recognition accuracy, which shows an improvement over the single-network counterpart as well as other widely used algorithms.

Acknowledgment

This work was supported in part by National Key R&D Program of China under Grant No. 2022YFC3300705, in part by National Natural Science Foundation of China under Grant No. 62088101, and in part by Young Elite Scientists Sponsorship Program by CAST under Grant 2022QNRC001. *(Corresponding author: Dongxuan He, Tianqi Mao.)*

References

[1] Y. Wang, Z. Su, N. Zhang, R. Xing, D. Liu, T. H. Luan, and X. Shen, “A survey on metaverse: Fundamentals, security, and privacy,” *IEEE Communications Surveys and Tutorials*, vol. 25, no. 1, pp. 319–352, 2023.

[2] S. Herath, M. Harandi, and F. Porikli, “Going deeper into action recognition: A survey,” *Image and vision computing*, vol. 60, pp. 4–21, 2017.

[3] L. Hu, “Deep-learning-based action recognition,” Master’s thesis, Dalian University of Technology, 2021.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[5] S. Ji, W. Xu, M. Yang, and K. Yu, “3d convolutional neural networks for human action recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[6] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, “Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification,” in *Proceedings of the European conference on computer vision (ECCV)*, pp. 305–321, 2018.

[7] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2625–2634, 2015.

[8] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” *Advances in neural information processing systems*, vol. 27, 2014.

[9] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, “Modeling spatial-temporal clues in a hybrid deep learning framework for video classification,” in *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 461–470, 2015.

[10] M. Gong and Y. Shu, “Real-time detection and motion recognition of human moving objects based on deep learning and multi-scale feature fusion in video,” *IEEE Access*, vol. 8, pp. 25811–25822, 2020.

[11] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

[12] J. Liu, A. Shahroudy, D. Xu, and G. Wang, “Spatio-temporal lstm with trust gates for 3d human action recognition,” in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pp. 816–833, Springer, 2016.

[13] C. Luo and A. L. Yuille, “Grouped spatial-temporal aggregation for efficient action recognition,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5512–5521, 2019.

[14] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, 2015.

[15] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, “On the importance of initialization and momentum in deep learning,” in *International conference on machine learning*, pp. 1139–1147, PMLR, 2013.