Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 4, No. 1, 2022

Analysis of Interaction Grouping Modeling Fusion Group Behavior Recognition Algorithm

Bin Li, Fan Zhang

China Auto Information Technology (Tianjin) Co., Ltd, China

Abstract: In order to make full use of the effective information in the video, this paper proposes a multi-model interactive video behavior recognition method. In order to solve the problems of incomplete human target detection and redundant feature extraction, YOLO_V4 is used to detect the human body and remove the redundant background information. Then, it is proposed to introduce the channel attention model SE-NET into the Inception_V3 network, so as to strengthen the extraction of key features and make the network pay more attention to the details of key features. Finally, the feature information is sent to LSTM network with memory function for action recognition and classification. The multi-model mutual fusion algorithm proposed in this paper is tested and verified on an internationally published UT-Interaction data set. The experimental results show that the accuracy of interactive behavior recognition is improved, and the improved accuracy is 85.1%, which indicates that the multi-model fusion method has higher accuracy.

Keywords: Two-person interaction, Behavior recognition, Multi-model fusion, Deep learning, Feature extraction, Modelling verification.

1. Introduction

With the rapid development of the "14th Five-Year Plan", in order to maintain social stability, the construction of video surveillance system has been gradually promoted nationwide [1]. Accurately predicting and identifying behaviors in videos is of great significance to people's safety.

The behavior prediction and recognition of video interaction between two people has always been a research hotspot in the field of intelligent video surveillance. Literature [2] puts forward a method of double-person behavior recognition of key frames, but its key frames have the problem of background interference, so it is easy to mistake the background with the target. In order to fully obtain the effective information of two-person interaction behavior on the characteristics of spatio-temporal network, reference[3] proposed to construct a multi-layer convolutional neural network in three-dimensional space, using the way that two layers of convolutional networks are superimposed to form an independent sub-three-dimensional space as the model of action recognition. However, the multi-layer convolutional neural network constructed in three-dimensional space has complex structure, large parameters and high hardware requirements. Literature[4] proposes a multi-dimensional spatio-temporal two-person information system fusion method based on multi-information channels, which mainly fuses different sequence information in the skeleton network of two-person interaction system, but the multi-channel spatio-temporal fusion network has too many network parameters to be applied in practice. Literature[5] puts forward the behavior probability modeling method of two-person dynamic BOW(Bag of Word), which can effectively solve the behavior probability prediction problem of two-person collective interaction. In this method, a whole square graph is used to accurately express the whole spatio-temporal characteristics of two-person behavior, and then a probability model is built to predict the probability of two-person behavior. Literature[6] proposed a prediction framework based on the combination of local interest points and sparse representation. Each video is divided into multiple time periods, and the category of two-person activities is determined by constructing a sparse word packet in each time period. Because each video is divided into multiple time periods, the time information and space information are separated, and the key information of the video cannot be fully utilized, which leads to a low recognition rate. In reference [7], a frame of motion recognition and prediction model based on key frames is proposed. By extracting the key frames of motion video as the state nodes of the motion model, the frame of two-person behavior recognition and prediction is realized. Reference [8] proposed a multi-scale kernel model method. The local progress model and the global progress model are used to capture the relationship between the time progress and the global progress, so as to complete the action recognition of the local video.

The existing problems of double-person behavior recognition mainly include: video background interference, incomplete human target detection, false target detection, complex network, low recognition rate and so on. In order to solve the above problems, YOLO_V4 is used to detect the human target and remove the redundant background. After that, SE-SE-Inception_V3 network is proposed to recognize the human target behavior, and its advantages are as follows:

1)Inception_V3 network can extract different features by designing different convolution modules, thus effectively solving the problem of incomplete human target detection;

2) The channel attention model SE-NET can solve the problem of false detection of targets;

3) The extracted features are rich, and the network reasoning speed is fast.

2. Two-person Interaction Behavior Recognition

2.1. Human Target Detection Network-YOLO _ V4

At present, YOLO_V4[9] is selected as the target detection network. The principle of YOLO_V4 target detection is to

transform the target detection results into the regression probability of nodes, and the convolution method and neural network algorithm are used to predict the boundary box of nodes and the regression probability of the target category to which the nodes belong. YOLO_V4 is used to divide the input object image first, and then S×S grids are formed. Then, each grid detects whether an object falls into the grid. When one of the center coordinates of the detected object falls in the grid, it means that the grid can detect the object. Each grid is responsible for predicting B bounding boxes and the Confidence Score of these bounding boxes. The confidence score reflects whether the grid has detected objects and the accuracy of the detected objects. Confidence is defined as:

$$\text{Confidence} = \Pr(\text{Object}) \cdot \text{IOU}_{\text{pred}}^{\text{truth}}$$

$$\text{IOU}_{\text{pred}}^{\text{truth}} = \frac{\text{Detection} \cap \text{GroundTruth}}{\text{Detection} \cup \text{GroundTruth}}$$

$$\text{Conf} = \Pr(\text{Class}_i|\text{Object}) \cdot \Pr(\text{Object}) \cdot \text{IOU}_{\text{pred}}^{\text{truth}} = \Pr(\text{Class}_i) \cdot \text{IOU}_{\text{pred}}^{\text{truth}}$$

$$\Pr(\text{Class}_i|\text{Object}) \cdot \Pr(\text{Object}) \cdot \text{IOU}_{\text{pred}}^{\text{truth}} = \Pr(\text{Class}_i) \cdot \text{IOU}_{\text{pred}}^{\text{truth}}$$

After obtaining the confidence score, compared with the manually set threshold, if the confidence score is greater than the manually set threshold, the prediction box will be kept. When the confidence score is less than the manually set threshold, the prediction box is discarded. YOLO_V4 is used to detect the human target in the video first, and remove the interference of background redundant information.

2.2. Feature extraction network-se-incidence _ v3

In the process of feature extraction, Inception_V3 uses convolution kernels of different sizes to obtain features of different sizes, so as to extract rich features. In order to pay more attention to important features and prevent secondary features, this paper adds an SE-Module module after each sub-module of incidence _ v3, which is called SE-incidence feature extraction network [10]. Its overall structure is shown in Figure 1.

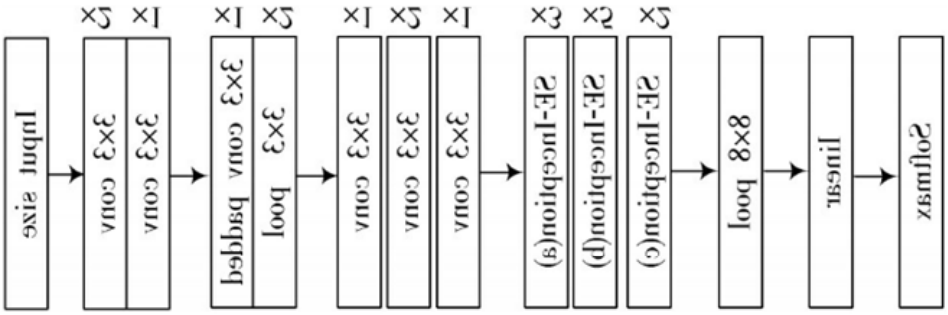

Figure 1. General structure diagram of se-incidence _ v3

Figure. 2 is a detailed structure diagram of SE-Inception. Firstly, the input network feature information X is subjected to a global pooling operation, and then the feature dimension is reduced to the original r 1, through the Full Connection Layer (FC); Then, the feature dimension is activated by ReLU function, and then transformed into the original dimension C through a global complete connection layer (FC); Finally, it is converted into a normalized weight of [0,1] by Sigmoid function.

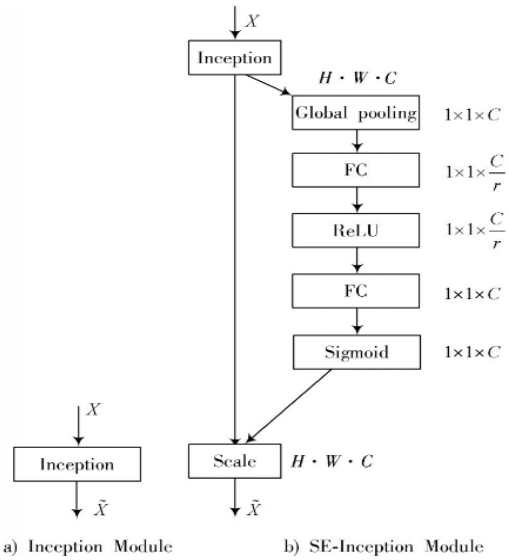

Figure 2. Structure diagram of se-induction module

Incident _ v3 submodule [11]: Incident _ v3 includes submodules A, B and C, as shown in Figure 3. There are three sub-modules A. The basic design idea is to solve the large volume integral into several small convolutions, and use two 3×3 convolution kernels instead of one 5×5 convolution kernel. One of its main features is to effectively reduce the

parameters of the model while ensuring the same receptive field, as shown in Figure 3a); There are five sub-modules B. The basic design principle of sub-module B is to decompose spatial convolution into asymmetric convolution in series. One of its main features is to decompose an N×N convolution kernel into a series combination of 1×N and N×1 convolution kernels, which reduces the computational load and greatly reduces the computational burden of hardware, as shown in Figure 3b). There are two sub-modules C. One of the main features of sub-module C is that the 3×3 convolution kernel is decomposed into 1×3 and 3×1 convolution kernels and combined in parallel, which reduces the workload of parameters and operations, as shown in Figure 3c).

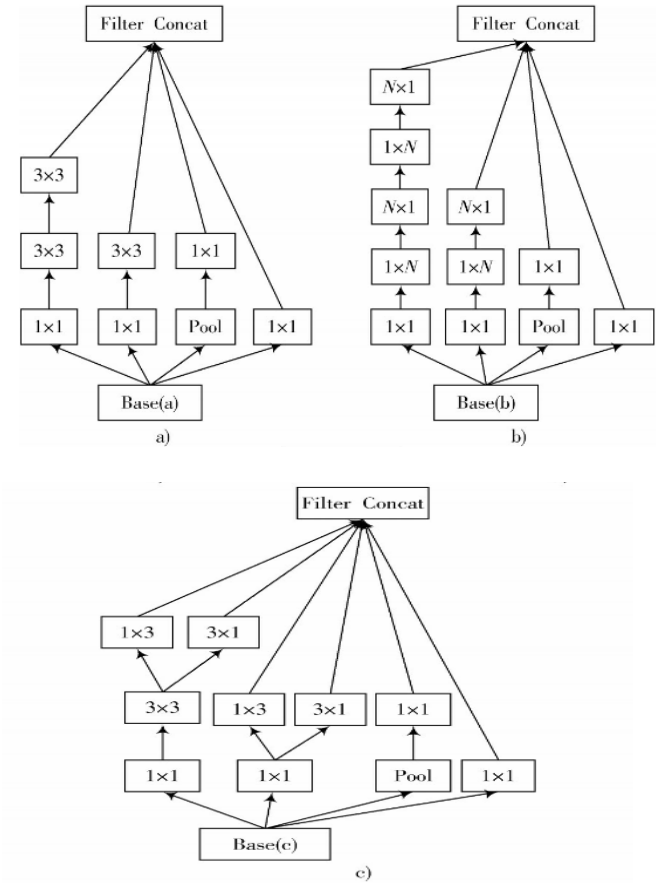

Figure 3. Structure diagram of incidence _ v3 submodule

Inception_V3 is suitable for smaller data sets because of its multi-scale features and fewer parameters. Two-person interaction UT_Interaction data set is small, so it is easy to produce over-fitting phenomenon when using large network. Inception_V3 is used as the backbone network of feature extraction, which can extract rich features and make the network reasoning faster, thus making preparations for the next behavior prediction.

2.3. Behavior prediction and classification network-lstm network

LSTM (Long Short Term Memory)[12] is a kind of Recurrent Neural Network (RNN), which has many cyclic memory functions that traditional neural networks do not have, and can find the sequence relationship between features. LSTM is effective in extracting the temporal features of images, so it has a good detection and recognition effect in the interaction between two people. The main structure of LSTM network system consists of three parts: the first part is the forgetting gate of data input, the second part is the input gate of data input, and the third part is the output gate of data output.

Intelligent building is not only a simple application of intelligent technology in architectural engineering, but also based on the deep integration of architectural design, construction, operation, service and high-tech development, embedding and use. It needs the joint efforts and cooperation of multi-disciplines, multi-types of work to finally achieve high-quality development.

The forgetting gate determines which information the model will delete from which state, and the calculation formula of forgetting gate is:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t \right] + b_f \right)$$

The input gate mainly finds the cell state to be updated and updates the information to the cell state. The calculation formula of the input gate is:

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t \right] + b_i \right)$$

The calculation formula of timely cell status pagenumber _ ebook = 185 and pagenumber _ book = 177 is:

$$\tilde{c}_t = \tanh\left(W_c \cdot \left[h_{t-1}, x_t \right] + b_c \right)$$

The information stored in the long-term memory cell state ct is:

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t$$

The output gate mainly determines which part needs to be

output, and the calculation formula of the output gate is:

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right)$$

The output value of LSTM is:

$$h_t = o_t \cdot \tanh\left(c_t\right)$$

Where: σ is Sigmoid function; Said xt t time input characteristic matrix; Wi represents the input layer to the input gate; Wf means forgetting gate; Wc represents a memory cell; Wo represents the weight matrix between output gates; Said bi input gate; Bf means forgetting gate; Bc represents a storage unit (cell); Bo indicates the offset value of the output gate.

The model sends the output result of the previous step SE-Inception_V3 to the LSTM unit of the time series module, and then averages the probability distribution of all frames, and selects the most possible label, thus completing the recognition of human motion in video.

3. Experimental Structure

3.1. Algorithm framework

The algorithm framework is shown in Figure 4.

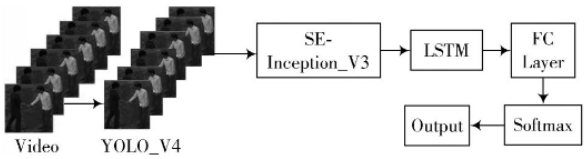

Figure 4. Algorithm Framework

The processing flow of this algorithm is as follows:

1) YOLO_V4 is used to detect the human target in the video, and the detected video is processed at frame level after the human target is detected.

2) use se-incidence _ v3 for feature extraction. Inception_V3 network can extract features of different scales, and channel attention mechanism (SE-NET) can assign weights to different convolution kernels to make them pay more attention to the main features.

3) Through LSTM for prediction and recognition, test action video frames of different types of interactive actions are sent into the trained test models respectively, and the probability value of each type of test action is obtained, so as to realize the prediction and recognition of two-person interactive behavior.

3.2. Network training

The network of this paper is built on TensorFlow. Hardware: the processor is i9-9700k；; GPU is nivda-rtx2080-ti; The GPU memory size is 11 GB. Software: the operating system is Ubuntu 18.04；; The language is Python. The YOLO_V4 network design uses Mish as an active gradient function, with the initial random learning rate of 0.1, weight coefficient of 0.005 and momentum coefficient of 0.9. Se-incidence _ v3 network adopts random gradient descent method (SGD) with high learning rate, the batch_size is set to 16, and the activation function is ReLU. In order to avoid over-fitting in training, a dropout with a threshold of 0.5 is set, the initial learning rate Lr=0.001, and the number of iterations nb_epoch=50.

3.3. Data set

A channel attention model SE-perception _ v3 based on perception _ v3 is proposed, which is used for behavior recognition algorithm of two-person interaction. In order to prove the feasibility of the algorithm, it is tested on the published UT-Interaction Datebase database [13] segmented_set2 data set. This data set includes six kinds of actions: shaking, hugging, pointing, kicking, punching and pushing. In addition, each type of action in this database is completed by different operators, and there are 60 groups of interactive actions. Because there is no periodic rule among all kinds of actions in the database, and there are many actions with high similarity among the types of actions in each database, it is more testable to recognize interactive behaviors.

The data concentration action is shown in Figure 5.

Figure 5. UT_Interaction data set

3.4. Comparison and analysis of experimental results

Comparison of accuracy results of different network models is shown in Table 1.

Table 1. Comparison of accuracy results of different network models

Network Model	Accuracy(%)
Inception_V3+LSTM	82.10
SE_Inception_V3+LSTM	84.20
YOLO_V4+SE-Inception_V3+LSTM	85.10

As can be seen from Table 1, the recognition rate has been improved by adding the channel attention module (SE-NET), and the recognition accuracy of two-person interaction behavior is 84.20%; The recognition rate of human body detected by YOLO_V4 and then sent to SE-induction _ V3 and LSTM network is higher than the original recognition rate, reaching 85.10%. It shows that the SE-Inception_V3 proposed in this paper has certain reliability in improving the accuracy of two-person interaction behavior recognition, and it also proves that it is feasible to improve the accuracy of later behavior recognition through YOLO_V4 in the early stage. Comparison of recognition rate results of different algorithms is shown in Table 2.

Table 2. Comparison of recognition rate results of different algorithms

Algorithm	Recognition Method	Recognition Rate
Algorithm in Literature[14]	STIP+Implicit Shape Model	73.30
Algorithm in Literature[15]	STIP+BP+SVM	83.30
Algorithm in Literature[2]	Description of keyframe feature library	85.00
Algorithm in this paper	YOLO+SE-Inception_V3+LSTM	85.10

Compared with literature [2], literature [14] and literature [15], the recognition rate of this algorithm framework is better. Although the recognition rate of literature [2] is the same as that of the algorithm in this paper, the extraction of key frames requires a lot of computation, requires high hardware requirements, and the description of feature library is complicated. The algorithms in literature [14] and literature [15] need to construct the STIP event sequence, and the feature extraction algorithm has high complexity and needs a lot of training. Experiments verify the feasibility of the proposed algorithm in two-person interaction.

4. Conclusion

Remarks in this paper, a multi-model fusion network is proposed. Firstly, YOLO_V4 is used to remove redundant background information and detect the target human body. Secondly, the se-incidence _ v3 proposed in this paper is used to extract features. Finally, LSTM network is used to predict and classify the two-person interaction behavior. Experimental results show that the model algorithm proposed in this paper has achieved good results on international public data sets, and the recognition accuracy of this algorithm has been significantly improved.

Although the algorithm in this paper has a good effect in the recognition of two-person interaction behavior, it relies on early human target detection and multi-feature extraction, and is not suitable for the behavior recognition of three or more people, and the recognition of multi-person interaction behavior is the key direction of future research.

References

[1] CARREIRA J,ZISSERMAN A. Quo vadis,Action recognition?a new model and the kinetics dataset [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition.Honolulu: IEEE,2017:1063-1071.

[2] Ji Xiaofei, Zuo Xinmeng. Two-person interaction behavior recognition based on statistical features of key frame feature database [J]. Computer Application, 2016,36 (8): 2287-2291.

[3] Chen Changhong, Liu Yuan. Two-person interaction behavior recognition based on improved sum-product network [J]. Computer Technology and Development, 2019,29 (10): 157-163.

[4] Pei Xiaomin, Fan Huijie, Tang Yandong. Recognition of two-person interaction behavior in multi-channel spatio-temporal fusion network [J]. Infrared and Laser Engineering, 2020,49 (5): 211-216.

[5] RYOO M S. Human activity prediction:early recognition of ongoing activities from streaming videos [C]// International Conference on Computer Vision. Barcelona, Spain:IEEE, 2011:1036-1043.

[6] XU K,QIN Z,WANG G. Human activities prediction by learning combinatorial sparse representations [C]// IEEE International Conference on Image Processing. Arizona, USA: IEEE, 2016:747-748.

[7] RAPTIS M,SIGAL L. Poselet key-framing:a model for human activity recognition [C]// IEEE Conference on Computer Vision& Pattern Recognition. Oregon, USA: IEEE, 2013: 2650-2657.

[8] KONG Y,FU Y. Max-margin action prediction machine [J].IEEE transactions on pattern analysis & machine intelligence, 2015, 38 (9):1844-1858.

[9] CHEN W,LU S,LIU B,et al. Detecting citrus in orchard environment by using improved YOLOv4 [J]. Scientific programming, 2020, 2020 (1):1-13.

[10] JIE H,LI S,ALBANIE A,et al. Squeeze-and-excitation networks [J]. IEEE transactions on pattern analysis and machine intelligence,2020,42 (8): 2011-2023.

[11] SZEGEDY C,VANHOUCKE V,IOFFE S. et al. Rethinking the inception architecture for computer vision [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas:IEEE,2016:2818-2826.

[12] Wang Yi, Ma Cuihong, Mao Zhiqiang. Research on behavior recognition based on 3D convolution and bidirectional LSTM [J]. Modern Electronic Technology, 2019,42 (14): 78-82.

[13] QIAN H,ZHOU X,ZHENG M. Abnormal behavior detection and recognition method based on improved ResNet model [J].Computers,materials and continua, 2020, 65 (3):2153-2167.

[14] ZHANG X,CUI J,TIAN L,et al. Local spatiotemporal feature based voting framework for complex human activity detection and localization [C]// Proceedings of the First Asian Conference on Pattern Recognition. Piscataway: IEEE, 2011: 12-16.

[15] YU T H, KIM T K, CIPOLLA R. Real - time action recognition by spatiotemporal semantic and structural forests[C]// Proceedings of the 21st British Machine Vision Conference. Bristol:BMVC,2010:1-12.