ISBN 978-981-14-1684-2
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
(WCSE 2019 SUMMER)
Hong Kong, 15-17 June, 2019, pp 307-311
doi: 10.18178/wcse.2019.06.046

Research on Real-time Behavior Recognition Method Based on Deep Learning

Yuanjun Ding[1], Qingqing Yang[1], Haoyang Yu[1], Hongjie Wang[1], Xiaocong Chen[1], Haibo Pu[1,2+]

[1] College of Information Engineering, Sichuan Agricultural University

[2] Key Laboratory of Agricultural Information Engineering of Sichuan Province

Abstract. With the advent of the era of big data, machine vision is growing rapidly and behavior recognition technology has a wide range of applications in our lives. As far as the current trend of behavior recognition technology is concerned, most of them have a series of problems such as slow calculation speed, low recognition accuracy and delay. In this paper, PoseNet deep neural network algorithm based on tensorflow.js is adopted to process the acquired image, train on the data set and extract the posture confidence and key point information of the human body. Through relevant algorithms, the behavior recognition of the target human body is completed, which has a broad application prospect in the future.

Keywords: Tensorflow.js, Deep Neural Networks, Key points, Behavior recognition

1. Introduction

Human behavior recognition is an emerging research direction in the field of artificial intelligence recognition, and has broad application prospects. It mainly used in video surveillance, medical diagnosis and monitoring, motion analysis, intelligent human-computer interaction and virtual display [1].In this paper, a deep neural network algorithm based on tensorflow.js is used to perform behavior recognition on single or multiple human bodies in an image. The algorithm trains on the Microsoft COCO dataset, outputs key information of the human body and returns it in json format, including the confidence of the 17 human key points and the value of the coordinate positions (x, y)[2].When processing in the browser, the posture estimation of the human body can be completed simply by processing the human body in the video and drawing out the key points and skeletons of the human body. After acquiring the key points and the skeleton of the human body, the known 17 key points are used to set the variance human body for recognition, which effectively improves the accuracy and speed of recognition.

2. Key Technologies and Framwork

2.1. Attitude detection technology

The attitude detection model mainly used in this paper is PoseNet[3]. PoseNet consists of two phases, the first phase, using the Faster R-CNN Detector[4]. In the second phase, 17 key points of the human body are estimated, which will be mentioned later. For each key point type, Full Convolutional ResNet[5] was used to predict the offset. The key point information is an important part of estimating the posture of the human body, including the position of the key point and the confidence score of the key point. This model with an accuracy of 0.685 [6] can be used to estimate a single pose or multiple poses, which gives this model a greater value for development. PoseNet currently detects 17 key points, using the basic label sample of the face and body parts of the coco dataset [7], and trains the network in a supervised manner. The key points are shown in Figure 1 below:

+ Corresponding author. Tel.: 18227589988
E-mail address: pu-hb@qq.com

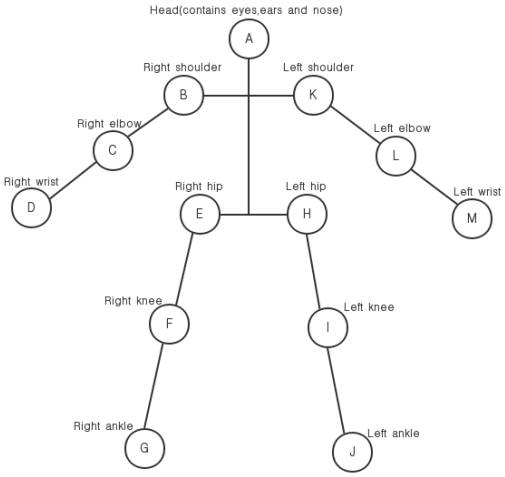

Fig. 1: Schematic diagram of key points

2.2. Convolutional netural network

The training set is the use of MoblieNet, MobileNet has been updated for many generations[8], and the version currently used by Tensorflow.js is V1, so this article simply introduces it. MobileNet V1 is a computational model primarily used on the mobile side, which adds a 1x1 convolution to traditional convolution operations. Deep separable convolution will be traditional the volume integral is solved as a depthwise convolution + a 1×1 pointwise convolution[9]. The convolution diagram is shown in Figure 2:

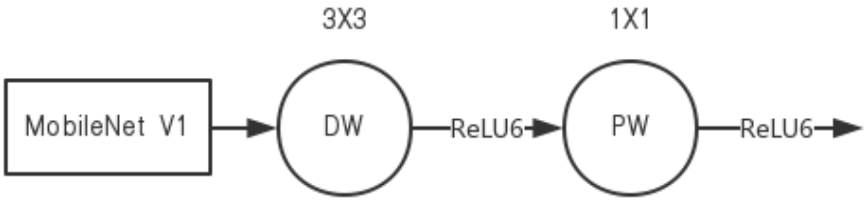

Fig. 2: Convolution diagram of MobileNet V1

2.3. Application framwork

This article mainly creates CNN (Convolutional Neural Network) on the browser and uses the GPU processing of the terminal to train these models. Therefore, it is not necessary to use a dedicated server GPU to train the neural network and support the use of WebGL [10]. When running inside the browser, Human posture estimation is mainly used through the browser, users can transfer real-time data from the camera to other devices via a mobile device. It is worth mentioning that all incoming data will remain on the client, enabling the framework to have low latency transmissions and modules with good privacy protection. Use the TensorFlow.js framework through JavaScript and the advanced layer API, where TensorFlow.js consists of two sets of APIs, the Ops API and the Layers API. The Ops API provides basic mathematical operations (such as matrix multiplication, etc.), and the Layers API provides a high-level model building block that trains neural networks [11]. The structure of the Layer API for training and use selected in this paper is shown in Figure 3:

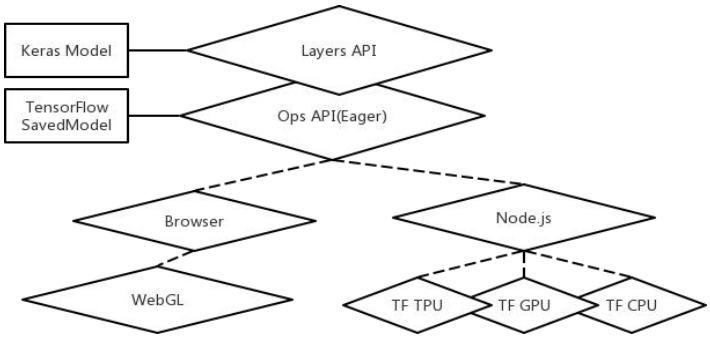

Fig. 3: Structure diagram of the layer API

3. The Process of Implemention

This paper mainly uses the Tensorflow.js framework to train on the server by calling the tensorflow-models/posenet library. And use the JavaScript syntax to call the camera (mobile phone, computer, monitoring, etc.) of the running device and observe it on the designed monitoring screen. This paper combines CNN, MobileNet V1 and PoseNet to process the video information obtained by the camera. The Architecture diagrams of processing is as follows in Figure 4:

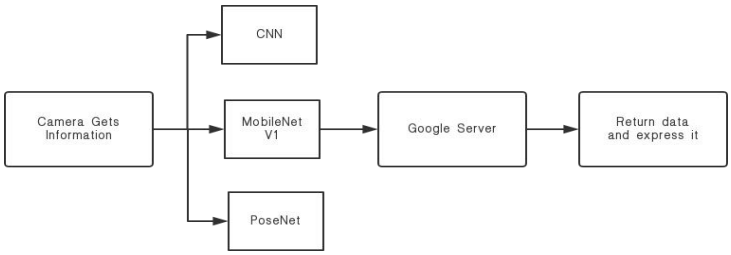

Fig. 4: Structure diagram of the Layer API

In order to reduce the algorithm complexity, improve recognition rate of training, this article selected influence on key points score calculation method of the most influential point, first of all, by will obtain the key position in the array to obtain the coordinates of the key information, and the key points out 17 had a great influence on the action points and minimal impact on subtraction, it is concluded that the biggest gap between the influence, the specific algorithm is as follows:

First, process the incoming data. The output and input are shown in formula 1:

$$G_{k,l,n} = \sum_{i,j,m} K_{i,j,m,n} \bullet F_{k+i-1,l+j-1,m} \tag{1}$$

Here, k, l are pixel values, n is the number of output channels, and m is the number of input channels. The obtained output is then compared with the trained MobileNet parameters to obtain the location of the key point, which is encapsulated into a list to calculate the range.

$$score = [keypoints[0].position \ldots keypoints[16].position]$$

$$\begin{cases} x = \max(score) \\ y = \min(score) \\ diff = x - y \end{cases} \tag{2}$$

The variance and standard deviation of the key points as shown below:

$$\begin{cases} \delta^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2 f^i}{\sum_{i=1}^{n} f_i} \\ \delta = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2 f^i}{\sum_{i=1}^{n} f_i}} \end{cases} \tag{3}$$

The variance and standard deviation of the key points' influence score can be obtained by introducing the data of key points of the human body into the algorithm, and the current behavior of the human body can be judged according to the specific score σ, as shown in Table 1 below:

Table1: Behavior score sheet

Action	σ Score
sit down	<0.34
Stand	0.34~0.52
Walk	0.52~0.63
Run	>0.63

The realization of the current behavior of the human body to identify, mainly standing, sitting, sitting, running the four most common human behavior to identify.

4. Experimental Result

The data set used in this training is the ICVL data set, which is a recorded surveillance video data consisting of 158 videos using 11 different indoor and outdoor scenes with a resolution of 1280 × 640 and an average speed of 20 fps[12].The duration of each video is from 1 minute to 6 minutes, covering the classification of human motion proposed above. This time, using different proportions of training and test data sets.15% is the verification data set, and the rest is the training data set. That is,24 videos are the verification set,134 videos are the training sets, and the recognition effect is tested and counted. The experimental results are as Table 2:

Table2: The table of behavior recognition accuracy

Category	Sit down	Stand	Walk	Run	Average
Recognition rate	88.14%	87.65%	78.40%	82.36%	84.14%

In order to test and evaluate the calculation time, the operating environment of this algorithm is CPU: Intel(R) Core(TM) i7-8565U CPU@1.80GHz 1.99GHz; memory: 8.00GB notebook. Using JavaScript+HTML to establish an experimental environment, the input video resolution is adjusted from the original 1280 × 640 to 640 × 320, and the time stamp is added to identify the image, which ensures accurate identification of the current behavior of the human body. The identification interface is shown in Figure 5 below:

Fig. 5: Interface diagram of behavior recognition

Processing time refers to the average processing time of 12 videos during testing. It can process about 27 frames in 1 second, which means that the average processing time of one frame is 37.04ms. Compare behavioral recognition rates on ICVL datasets with algorithms used in different literature, results are as Table 3:

Table3: Contrast table of recognition rates

Algorithm	Sit down	Stand	Walk	Run	Average time consuming
reference[13]	87.32%	83.67%	75.33%	71.67%	57.83
reference[14]	83.45%	81.98%	79.35%	79.96%	65.13
reference[15]	82.78%	86.78%	74.45%	73.64%	42.67
reference[16]	84.45%	85.32%	73.92%	78.76%	47.76
This article	88.14%	87.65%	78.40%	82.36%	37.04

After testing and comparing the performance of the five algorithms, the recognition rate of sitting, walking and running in this method is higher than that of the other four algorithms. Because the walking movement is small, it is easier to be confused with sitting, standing and running, so the recognition rate is the lowest. Compared with other algorithms, it can be found that the recognition rate of sitting, standing and running is improved, and the overall performance is effectively improved. The average iteration time of the influence iteration algorithm used in this paper is 37.04s, which is significantly higher than other algorithms.

In the target detection, tracking and behavior recognition, the overall calculation rate is improved. It is proved that the accuracy and efficiency of the algorithm are improved.

5. Conclusion

In order to improve the accuracy and efficiency of behavior recognition, this paper based on Tensorflow.js, the depth of the neural network algorithm to extract image in the body's key point information and training on Mobile Net dataset, the variance of data by using the method of scoring judgment arithmetic and identify the body of the current behavior, for both single-player and multiplayer identification can be relatively quick actions. According to the experimental results, the method adopted in this paper realizes the recognition of the current human behavior, and can be extended to remote monitoring, human-computer interaction, security and other production and living fields in the future, and has a broad application prospect in the future.

6. References

[1] He weihua. Research on key technologies of human behavior recognition. Chongqing university,2012. (in Chinese).

[2] Sun haifeng. Research on behavior recognition method based on complex linear dynamic system. Qingdao university of science and technology,2018. (in Chinese).

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren,Jian Sun. Deep Residual Learning for Image Recognition. In CVPR 2016.

[4] Alex Kendall, Matthew Grimes, Roberto Cipolla. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. arXiv:1505.07427,2016

[5] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In CVPR 2016.

[6] Georgia Gkioxari Ross Girshick Piotr Doll´ar Kaiming He.Detecting and Recognizing Human-Object Interactions. arXiv:1704.07333,2018.

[7] Yi Zhu, Zhenzhong Lan, Shawn Newsam, Alexander G. Hauptmann. Hidden Two-Stream Convolutional Networks for Action Recognition. arXiv:1704.00389,2017.

[8] Shuai Li, Wanqing Li , Chris Cook , Ce Zhu , Yanbo Gao. Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN. arXiv:1803.04831, 2018

[9] Shuyang Sun Zhanghui Kuang, Wanli Ouyang, Lu Sheng, and Wei Zhang.Optical Flow Guided Feature: A Fast and Robust Motion Representation for. arXiv:1711.11152, 2017.

[10] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, Kevin Murphy. Towards Accurate Multi-person Pose Estimation in the Wild.arXiv:1701.01779, 2017.

[11] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, Kevin Murphy.PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model. arXiv:1803.08225,2018.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick. Microsoft coco:Conmon objects in context. In ECCV, pages 740–755. Springer, 2014.

[13] Greff K, Srivastava R K, Koutník J, et al. LSTM: a search space odyssey. IEEE Transactions on Neural Networks and Learning Systems. 2017, 28 (10): 2222-2232.

[14] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In CVPR 2016.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren,Jian Sun. Deep Residual Learning for Image Recognition. In CVPR 2016.

[16] Alex Kendall, Matthew Grimes, Roberto Cipolla. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. arXiv:1505.07427,2016