Object Detection and Analysis of Human Body Postures Based on TensorFlow

Ling Xie
Communication University of China
School of Information and Communication Engineering
Beijing, China
e-mail: qwxieling1816@163.com

Xiao Guo
Communication University of China
School of Information and Communication Engineering
Beijing, China
e-mail: xguo@cuc.edu.cn

Abstract— **Human body posture recognition has been an important concern of research in many fields. In this field, a human pose estimation algorithm called OpenPose has been more widely used. But its efficiency is very low. We used deep learning methods based on TensorFlow to recognize human body postures. And we applied it to the judgment of the teacher's teaching states. In order to choose the algorithm that works best for our scenario, we designed eight sets of experimental schemes through combining the classification model and the detection algorithm. Then we used these groups of schemes to detect and classify the teacher's teaching states, including standing, sitting, etc. At last, we analyzed the experimental results in depth and selected the most suitable algorithm for our scenario.**

Keywords-object detection, body postures, TensorFlow, Faster R-CNN

I. INTRODUCTION

Human body posture recognition has been widely used in intelligent monitoring, human-computer interaction, video retrieval, virtual reality, etc. It has always been an active research direction in the field of computer vision. Early human body postures recognition mainly relies on multiple sensors in different parts of the human body [1]. The method of human body posture recognition based on sensors of smart phones and wearable devices (such as wristbands and watches) has been already mainstream [2, 3]. Traditional machine learning methods, such as SVM or Bayesian networks, time-domain and frequency domain analysis, etc. [4, 5]. These machine learning methods require professional knowledge in the human body posture domain for feature extraction.

In recent years, there are two major directions about the recognition of human body posture. One is to use object detection methods based on deep learning, directly use CNN to train, identify and classify human body postures. For example, scholars used neural networks and deep learning, but still requires artificial feature extraction [3,6]. Zeng M et al. used convolutional neural networks, but only used one layer for convolutional [7]. So, the learning is not enough deep to extract features of higher dimensions. Literature [8] used smart watches and deep convolutional neural networks to identify and classify human activities of different scenarios. The other is the detection of key points such as the human body, hands, face, and body posture estimation. In 2016, Zhe Cho et al. proposed an approach called OpenPose for detecting 2D poses of multiple people in an image [9]. This approach uses a bottom-up algorithm to obtain the key points and then obtain the backbone. It can achieve accurate estimation of the human body posture in the picture. Literature [10] presented a 2D human gesture grading system from monocular images based on OpenPose. Literature [11] presented an implementation of a bi-manual teleoperation system, controlled by a human through three-dimensional (3D) skeleton extraction, and achieved it based on the 3D version of the impressive OpenPose.

This paper mainly recognizes and classifies several types of human action poses in a specific scene. On the one hand, OpenPose runs very inefficiently, and its outputs are some keypoints of the human skeleton structure and their connections. These outputs still need to be input into the convolutional neural network for training. On the other hand, these types of actions are fixed, and the differences of these actions are very obvious. The object detection and recognition based on TensorFlow can easily distinguish and classify them. In short, we hope to achieve these through a simpler, more general way.

As the second generation of artificial intelligence learning system of Google, TensorFlow has a wide range of attention and use in the field of machine learning worldwide. TensorFlow has the advantages of high availability and flexibility. The object detection algorithm based on deep learning is also very convenient to implement through TensorFlow, and the hardware environment requirements are not high, which is very suitable for the research of this paper.

Using neural networks to learn the different characteristics of different actions and then build models, we can relatively accurately identify these types of action poses. So, we choose deep learning methods based on TensorFlow rather than the algorithm of OpenPose.

In this paper, we apply the human body posture recognition using deep learning method based on TensorFlow to the judgement of teacher's teaching states. In the first part, we introduced research status at home and abroad. In the second part, we will briefly introduce the object detection algorithms used in this paper and the corresponding classification models. In the third part, we present our experimental environment, data, schemes, and conduct an in-depth analysis of the experimental results. The last part is our conclusions.

978-1-7281-3488-8/19/$31.00 ©2019 IEEE
DOI 10.1109/SmartIoT.2019.00070

II. OBJECT DETECTION ALGORITHMS AND CLASSIFICATION MODELS

In this part, we will briefly introduce the object detection algorithms used in this paper (Faster R-CNN, R-FCN and SSD), and the corresponding classification models (MobileNet, InceptionNet, ResNet).

A. Classification model based on CNN

Convolutional Neural Network (CNN) is the most widely used deep learning technology, and it performs very well in the fields of image classification and object detection.

In 1998, LeCun et al. proposed LeNet to identify two-dimensional text images [12]. The LeNet has two convolutional layers, two pooling layers and a fully connected layer. The convolutional layer and the pooling layer are used to extract features and raw data to feature dimensions. While fully connected layers are used to classify and map feature dimensions to sample tags. LeNet is the prototype of modern convolutional neural networks.

Hinton and his student Alex Krizhevsky designed AlexNet in the 2012 ImageNet Large Scale Visual Recognition Challenge [13]. AlexNet deepened the learning layer of the network based on LeNet, using 5 convolution layers and 3 fully connected layers. Many improvements were made in the training. AlexNet laid the foundation for many of the better models.

In order to improve the performance of CNN on image classification, researchers at Oxford University used a 3*3 convolution kernel to extract image features in the convolutional layer, and proposed a deeper VGG model [14]. In order to better integrate the features of multi-scale models, Google proposed InceptionNet [15]. InceptionNet uses multiple convolutional kernels of different sizes to convolve the output of the previous layer in the same convolutional layer, and convolves all the convolution operations. The results are stacked together, thereby avoiding the uncertainty of manually determining the size of the convolution kernel.

Due to the existence of gradient disappearance and gradient explosion, the deeper neural network is, the more difficult it is to train, and the classification accuracy will decrease as the network depth continues to increase. In response to such problems, He Yuming et al. proposed a residual network called ResNet [16]. Each residual block of ResNet connects the activation value of the current layer to the deeper layer of the network through the shortcut on the basis of forward propagation, as shown in the figure 1.

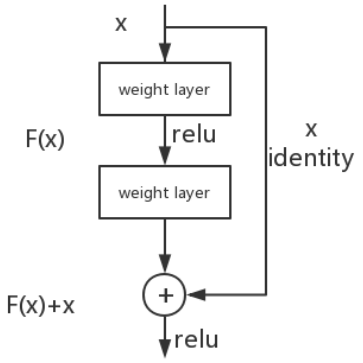


Figure 1. Schematic diagram of the shortcut of the ResNet model.

In 2016, Szegedy et al. confirmed through experiments that the combination of residual connections can significantly accelerate the training of InceptionNet, so a hybrid Inception module, Inception-ResNet network was proposed [17].

In 2017, Andrew et al. proposed an efficient network architecture, MobileNet, using deep separable convolution to construct a lightweight deep neural network [18].

In these models, ResNet and InceptionNet are widely used and have good accuracy. Inception-ResNet combines the two, can achieve higher accuracy in fewer epochs. MobileNet allows the construction of very small, low-latency models that easily meet the requirements of embedded devices with two hyperparameters.

B. Detection algorithms based on deep learning

Ross Girshick et al. proposed an object detection and recognition method based on deep learning, R-CNN [19], which divides object detection and recognition tasks into classification tasks based on candidate region extraction. Fast R-CNN proposed a region of interest (ROIs) strategy to map candidate regions to the feature layer of the CNN model, and extract the deep features of the corresponding regions directly on the feature layer, avoiding the constant input of images of different regions [20].

Ren Shaoqing, He Yuming et al. proposed the Faster R-CNN algorithm [21]. Based on Fast R-CNN, a region generation network (RPN) was introduced to replace the selective search (SS) algorithm for object candidate region extraction. This can solve the problem that the object candidate region method SS of the R-CNN and Fast R-CNN takes a relatively long time and it’s difficult to integrate into the GPU operation and the detection speed is slow. The network model of Faster R-CNN is shown as figure 2.

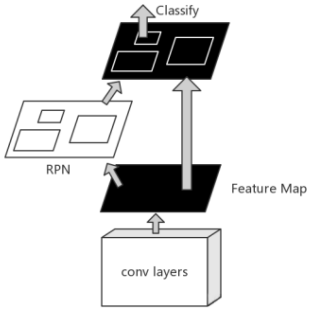


Figure 2. The Network Model of Faster R-CNN

Jifeng Dai et al. proposed a more efficient algorithm called R-FCN [22]. This method is based on the region where the detector is fully convolved and almost all calculations can be shared across the image, which can avoid repeated calculation for each ROI.

In 2015, Joseph Redmon et al. proposed a regression-based object detection and recognition algorithm, the YOLO algorithm [23]. They use regression to detect the object location and identify the object category. The regression method only needs to evaluate the whole picture with a single network to get the object bounding box and category. It belongs to an end-to-end model.

Wei Liu et al. also proposed a regression-based object detection and recognition algorithm SSD [24]. On the basis of YOLO, a target point mechanism similar to Faster R-CNN was added, and it uses only the deep features around each target to detect and recognize the object. Extract features from the feature maps of different layers of the deep neural network, so that it's possible to naturally add more scale information.

In these algorithms, comprehensive ability of Faster R-CNN is better. The efficient of R-FCN is better, but the accuracy is worse. The speed of YOLO is very fast, and the model is simple in structure. And SSD can improve the accuracy without affecting the speed on the base of YOLO.

III. EXPERIMENT AND RESULTS

A. Experimental Data and Environment

In this paper, the self-built data set is used in the experiment. We use labelVideo [25] to select and label 911 photos of the different action modes about the teacher's lectures in the classroom, including standing, sitting, writing on the blackboard, pointing to the blackboard, pointing to the lecture slides and interacting with the students. The interaction is defined as the teacher walking to the students and interacting with the students. The resolution of these pictures is 1920×1080, color channel is RGB. We also labeled the blackboards, lecture slides and desks in the photos. These pictures are at different angles, different time and different illumination angles, as shown in the figure 6. In figure 6, a) picture represents the normal angle and normal illumination, b) picture represents that the angle is oblique, c) picture represents intense illumination. We use these data as training set. In addition, 50 pictures not included in the training set were selected as test set.

The experimental environment of this paper is as the follow:

TABLE I. EXPERIMENTAL ENVIRONMENT

Classification Model	Detection Algorithms
Operating System	Ubuntu 16.04.5 LTS
CPU	Intel(R) Xeon(R) E5-2609 1.70GHz * 2
GPU	Nvidia P100*1
Memory	64GB
TensorFlow	1.12.0
OpenCV	3.4.2

B. Environmental Schemes and Results

According to the breadth of use, the classification model and the detection algorithm are combined, and the following experimental schemes are designed:

TABLE II. EXPERIMENTAL SCHEMES

Number	Classification Model	Detection Algorithms
1	MobileNet V2	SSD
2	Inception V2	SSD
3	Inception V2	Faster R-CNN
4	ResNet101	Faster R-CNN
5	ResNet101	R-FCN
6	ResNet152	Faster R-CNN
7	Inception_ResNet_v2	Faster R-CNN
8	YOLO	

According to the above eight experimental schemes, the training set was trained by using the pre-training model of each classification model, and the test set was tested by the trained model. At this stage, the experimental parameters are just default parameters. Then we compared final accuracy and efficiency of each group. The accuracy and efficiency comparisons are shown in the following figure3, 4, 5. Figure 4 shows the accuracy comparison of the six postures.

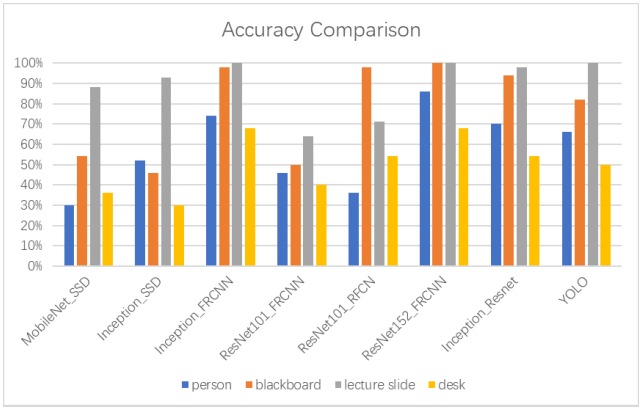


Figure 3. Accuracy Comparison of Eight Groups of Experiments.

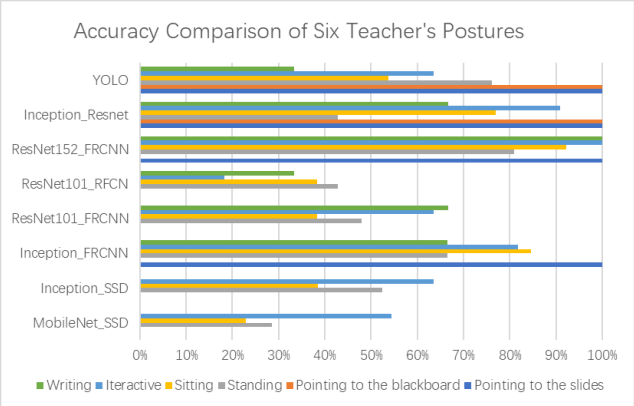


Figure 4. Accuracy Comparison of Six Teacher's Postures.

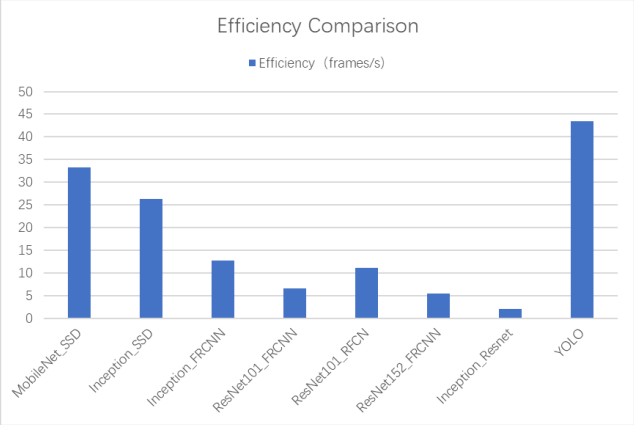


Figure 5. Efficiency Comparison of Eight Groups of Experiments.

The efficiency in the figure 5 is expressed by the number of pictures that can be inferred per second, which is frame rate.

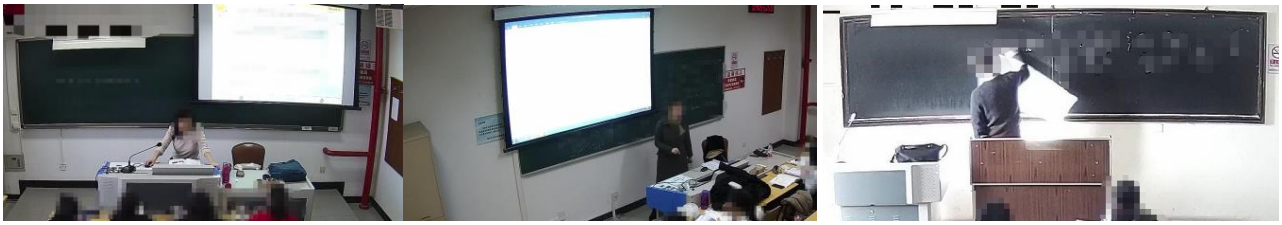
C. *Analysis of Experimental Results*

It can be clearly seen from the figures 3, 4 and 5 that the accuracy of ResNet_152 model based on the Faster R-CNN algorithm is highest, and the Inception_v2 model based on Faster R-CNN is second, but correspondingly, the inference efficiency of both is low, in particular the ResNet_152 model based on Faster R-CNN. These two sets of models have a high accuracy rate for the classification of teachers' state, and the recognition of objects (blackboard, lecture slide and desk) also performs very well.

The YOLO model has really high efficiency. And the first group and the second group (according to table 2) both based on SSD are also very efficient, but the accuracy of identifying people is lower, which is also consistent with the characteristics of the SSD algorithm. The model structure of SSD is simple, the speed is fast, and the corresponding accuracy rate is not as good as Faster R-CNN. These two groups perform better on some neatly edged objects, especially lecture slide.

We can see from the comparison between the fourth and fifth groups (according to the table 2), R-FCN improves efficiency compared to Faster R-CNN, but the accuracy rate of person decreases. It is worth noting that R-FCN can achieve better accuracy while ensuring efficiency on neatly edged objects.

From the figure 3 we can see that the accuracy of the desk in almost all groups of experiments is not high, for the following reasons:



a) normal b) angle-angled c) Strong lighting

Figure 6. Three Examples of Data Set.

- When the angle of the classroom in the picture is oblique, the desk usually only displays the desktop, just like the figure 6.b). So, it is not easy to label in the training set. And the labeled appearance is not the same when the angle is different. And the angle-angled samples occupy a small proportion in the training set, the desk in the angle-angled test picture is not easy to identify.
- In the course of the teacher's lecture, there will be movements, and the teacher may obscure the desk, so that the desk will not be recognized.

In the test set, there are 11 pictures that the angels are oblique and 3 pictures that the desks are blocked. So, the accuracy of the desk was not high.

Only one image in the test set is pointing to the lecture slide, so in the picture 4, its accuracy is 0 or 100%. Pointing to the blackboard the same. If one posture occupies more proportion in training set, its detection rate will be higher. In figure 4, the accuracy of sitting and standing are higher, because they have more samples in the training set.

The following conclusions can be drawn from several sets of experimental results with low accuracy:

- It is difficult to identify when the angle of the classroom in the picture is oblique, which is closely related to the relatively small sample of the oblique angle of the classroom in the training set.
- It is difficult to identify when the lighting conditions are strong and people are standing under strong light, just like the figure 6.c).
- It is not easy to identify when the classroom is large and the platform is far away from the camera.

IV. CONCLUSIONS

In this paper, we use several existing detection algorithms based on deep learning and classification models based on CNN to evaluate the teaching state of teachers. From the experimental results, we can see that the angle of the classroom in the picture and illumination have a certain impact on the detection. And when the classroom is large and the platform is far away from the camera, the detection is not easy too.

In eight sets of experiment, the Resnet152 model based on the Faster R-CNN detection algorithm has the highest accuracy. The Inception-v2 model based on Faster R-CNN is second. But the efficiency of the two is very low, it is difficult to meet real-time requirements. The YOLO model is very efficient, although the accuracy rate is lower than the previous two, but the efficiency comparison accuracy loss is still worth a certain degree.

In subsequent work, we will adjust the parameters according to the performance of different sets of experiments to achieve better performance. We will improve the models and algorithms to ensure high accuracy while improving efficiency. We will also try to use the OpenPose method to improve our algorithm.

ACKNOWLEDGMENT

This work is supported by “the Fundamental Research Funds for the Central Universities”.

REFERENCES

[1] Wu W, Dasgupta D, Ramirez E, et al. Classification accuracies of physical activities using smartphone motion sensors[J]. Journal of medical Internet research,2012,14(5): el30-el30.

[2] Kwapisz J R, Weiss G M, Moore S. Activity recognition using cell phone accelerometers[J]. Acm Sigkdd Explorations Newsletter, 2011, 12(2):74-82.

[3] Zhang L, Wu X, Luo D. Recognizing Human Activities from Raw Accelerometer Data Using Deep Neural Networks[C]//Proc of 2015 IEEE 14th International Conference on Machine Learning and Applications. Miami: IEEE,2015:865-870.

[4] Anjum A, Ilyas M U. Activity recognition using smartphone sensors[C]//Proc of 2013 IEEE 10th Consumer Communications and Networking Conference. Las Vegas: IEEE, 2013: 914-919.

[5] Martin H, Bernardos A M, Iglesias J, et al.Activity logging using lightweight classification techniques in mobile devices[J].Personal and Ubiquitous Computing, 2013, 17(4): 675-695.

[6] Kwon Y, Kang, Bae C. Analysis and evaluation of smartphone-based human activity recognition using a neural network approach[C]//Proc of 2015 International Joint Conference on Neural Network. Killarney: IEEE,2015:1-5.

[7] Zeng M, Nguyen L T, Yu B, et al. Convolutional neural networks for human activity recognition using mobile sensors[C]//Proc of 2014 6th International Conference on Mobile Computing, Applications and Services. Texas: IEEE,2014:197-205.

[8] Min-Cheol Kwon, Hanjong You, Jeongung Kim, Sunwoong Choi. Classification of Various Daily Activities using Convolution Neural Network and Smartwatch[C]//IEEE International Conference on Big Data (Big Data), 2018:

[9] Z. Cao, T. Simon, S. E. Wei, Y. Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields[C]//IEEE Conference on Computer Vision & Pattern Recognition, 2017: 1302-1310.

[10] Qiao S, Wang Y, Li J. Real-time human gesture grading based on OpenPose[C]// 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). IEEE, 2018.

[11] Emily-Jane Rolley-Parnell, Dimitrios Kanoulas, Arturo Laurenzi, et al. Bi-Manual Articulated Robot Teleoperation using an External RGB-D Range Sensor[C]//International Conference on Control, Automation, Robotics and Vision (ICARCV), 2018.

[12] Lécun Y, Bottou L, Bengio Y, et al. Gradient-based learning app; lied to document recognition[J]. Proceedings of the IEEE,1998,86(11):2278-2324.

[13] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//International Conference on Neural Information Processing Systems. Curran Associates Inc.2012:1097-1105.

[14] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.

[15] Szegedy C, Liu W, JiaY, et al. Going deeper with convolutions[C] //IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2015:1-9.

[16] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceeding of the IEEE conference on computer vision and pattern recognition.2016:770-778.

[17] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning[J]. 2016.

[18] Howard A G, Zhu M, Chen B, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications[J]. 2017.

[19] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Computer Vision and Pattern Recognition. IEEE, 2014:580-587.

[20] Girshick R. Fast R-CNN[J]. Computer Science, 2015.

[21] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time Object Detection with Region Proposal Networks[J]. IEEE T ransactions on Pattern Analysis&Machine Intelligence, 2015.1,2,6,7,8,9

[22] Dai J, Li Y, He K, et al. R-FCN: Object Detection via Region-based Fully Convolutional Networks[J]. 2016.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified real-time object detection[C]. Computer Vision and Pattern Recognition (CVPR),2016.

[24] W. Liu, D. Anguelov, D. Erhan, S. Christian, S. Reed, C.-Y. Fu, and A. C, Berg. SSD: single shot multibox detector[C]. European Conference on Computer Vision (ECCV),2016.1

[25] https://github.com/shawncode/labelVideo/.