2020 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

An Analysis System of Students' Attendence State in Classroom Based on Human Posture Recognition

Yuanhang Feng,Lei Zhang,Yuncheng Zhang,Xiang Li, Quanyin Zhu*
Faculty of Computer & Software Engineering, Huaiyin Institute of Technology, Huaian, China
*Corresponding author's e-mail:hyitzqy@126.com

***Abstract*—Listening state is a value reflecting the efficiency of lectures. As various human gesture recognition algorithms proposed, we could make advantages of these algorithms to construct a system to evaluate the listening state of the students. With the spring of many outstanding algorithm of machine vision, the basic theories have reached matured to complete some complex assignment. The system proposed here is a kind of integration of these technologies, which also need innovation when it's actually implemented. Also, wasted surveillance video resources universally existed in our society, which should be made advantage of but not. Especially in the scene of the classrooms, the schools hold tremendous surveillance video resources. The system is proposed to analyze the state of the students through the human posture recognition. Its basic theory is the recognition of the human body key points, including the rather important points of one's body. We could extract these information from the videos firstly, and make some analysis such as the computation of some specific angles. After these treatment of data, we could get rough results of the state of the students, whose correct rate could be higher through machine learning. There are much efforts to be made before the system could fit the requirement of the specific environment, such as semantic segmentation and super resolution. The system would also have clear visualization when finally faces customers.**

Keywords-posture recognition ;listening state; machine vision; key points; CNN ;visualization.

I. Introduction

This paper puts forward a kind of system to analyze students' state in the course. The system is based on the excellent image learning model appeared in recent two years, which could analyze the posture skeleton. This system could be applicable in video condition, also in different observational angles. It finally could measure students' movements in learning and then analyze students' learning situation.

II. Related work

A. Preprocessing and capture of human body

First, the images that have been acquired need to be preprocessed, including binarization and super resolution. To capture the people in the images accurately, a human body detection model need to be practiced in advance. Here we use the Mask-RCNN[1] algorithm to segment the whole image semantically to capture human body. The segmented areas will be identified to further analysis. In the process of noise removal, the recognition method based on gradient[2] or the feature extraction method based on lump could be used.

B. Extraction of human skeleton frame

After the images captured by the human body obtained, the human image analysis should be carried out. First, the human skeleton frame should be extracted from the images, especially to the blurred images[3]. In the process of skeletonization, we pay attention to two frames with small difference in time dimension to detect which frames are extracted incorrectly. Here we use an algorithm of human body recognition based on spatial temporal graph convolutional networks[4]. Some methods also refer to Microsoft Kinect in the process of extracting skeleton frames[5]. If there're too much difference in a very near time, it means the occurrence of actions or the mistakes which should be analyzed once more.

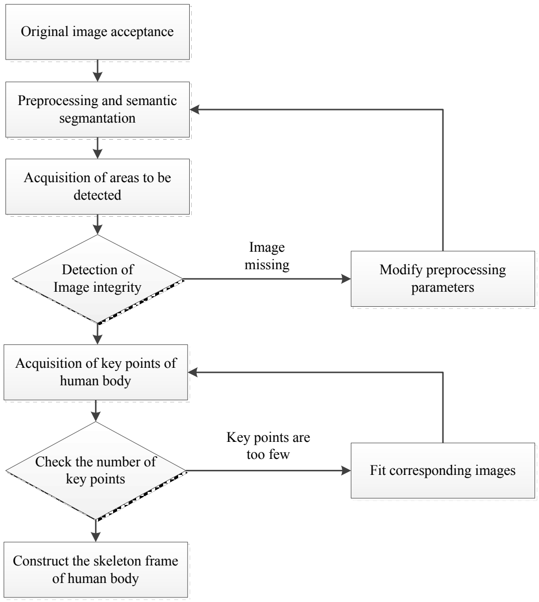
Original image acceptance
Preprocessing and semantic segmantation
Acquisition of areas to be detected
Detection of Image integrity
Image missing
Modify preprocessing parameters
Acquisition of key points of human body
Check the number of key points
Key points are too few
Fit corresponding images
Construct the skeleton frame of human body

Figure 1. Basic steps of extracting skeletal model of human body.

C. Analysis combined with key points

When the key points information has been analyzed, it's possible to judge the specific state of students with the change of human angles. For example, if the angel between the key parts such as spine and hand is too large, then the observed one could be judged to be picking up things or sleeping[6]. Similarly, the distance between the head and the desktop can also be used to judgement. It's simpler to introduce the standard students' sitting posture, which could be used to calculate the difference between the actual detection value that is the result of the comprehensive evaluation. Through this method, we could make an evaluation of students' listening state in

2473-3636/20/$31.00 ©2020 IEEE
DOI 10.1109/DCABES50732.2020.00063

classroom. It’s could also used as a method to predict the learning state, which need the visualization of the learning status of the whole class to determine the crest value of the status of the students.

D. Construction of visual graphical user interface

The system need a visual interface to display the result of the recognition. After acquiring the whole learning of a class, we should output students’ status and corresponding responsed information[7]. The final analysis result is expressed by a fitting curve, and the horizontal and vertical axes respectively represent the time and the overall students’ listening state. Similarly, the graphic output will show the skeletal frame. The overall effect will be identified by two figures, one is the direct output with skeleton lines, and the other one is the fitting curve including the analysis of students’ attendance in classroom. Visual user interface is open- source, and users can choose their own code to run the program.

III. RELATED THEORIES

A. Super resolution

As mentioned before, it is necessary to perform super-resolution operation on the images at the same time during preprocessing to achieve the purpose of capturing the human body more accurately. Technically, it takes three stages to achieve super-resolution. They are low-resolution image registration, non-uniform interpolation, blur removal and noise removal. Low-resolution frames are aligned with sub-pixel accuracy by image registering algorithm. Then these aligned low-resolution frames are placed on the high-resolution image grid. Also, very deep residual channel attention networks can also used for image super-resolution.

Deep learning can learn complex feature extraction transformation from big data, which is very suitable for image super-resolution reconstruction task. Various of deep learning network models have achieved good results in super-resolution image reconstruction. Here, we can use the multi-stage fusion network for image super-resolution reconstruction[8].Features are extracted and input into two sub-networks. Firstly, the structural feature information of low-resolution image is obtained through coding network; Secondly, the high-level features are obtained through the multi-path feedforward network composed of stage feature fusion units, in which the fusion units fuse the features of several successive layers of the network and obtain effective features in an adaptive way, and then connect different feature fusion units by multi-path connection to enhance the connection between the fusion units. This method mainly divides the network into different stages for processing, so as to make full use of features.

In the deep network, it is difficult to combine the features extracted from all layers to complete the reconstruction, so the stage feature fusion is used here.

The stage feature fusion unit includes three parts: dense connection layer, local residual layer and feature selection layer. The dense connection layer mentioned here is a concept put forward for feature reuse. But also to enhance the network's ability to express features. Now let the input feature be h, and its output can be expressed as the following formula.

$$x_i = C_{3,32}([H, x_1, \dots, x_{i-1}]) \qquad (1)$$

In the formula, x_i represents the output of the i-th convolution layer. The size and number of channels of all output feature maps are the same.[] indicates that the feature maps are merged in a cascade manner. The local residual is used to further to further sort out the information flow.

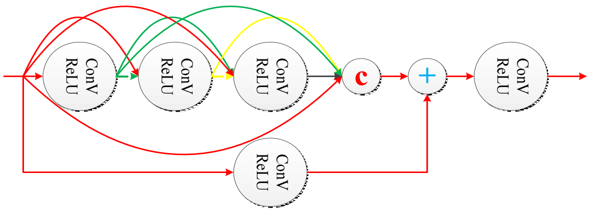

Figure 2. The schematic diagram of the feature fusion unit

In the process of jump connection in this cell, a convolution layer is introduced. Let the input feature be H, and the following formula is given.

$$\boldsymbol{H_r} = \boldsymbol{C_{1,128}}(\boldsymbol{H}) + \boldsymbol{H_d} \qquad (2)$$

At the same time, there are other depth image super-resolution reconstruction based on texture edge.

B. Semantic segmentation

This method is mainly used to detect objects in images efficiently and generate high-quality segmentation masks for each object. We have adopted the Mask-RCNN algorithm, which needs to extract more detailed spatial layout. The key lies in pixel-to-pixel alignment, and the loss function of Mask-RCNN is divided into three parts.

$$L = L_{cls} + L_{box} + L_{mask} \qquad (3)$$

L_{cls} and L_{box} are the same as those defined in fast-RCNN. L_{mask} makes the loss function on mask branch, and outputs k binary masks with the size of K*m*m and the coding resolution of m*m, that is, k categories each correspond to a binary mask. sigmoid function is used for each pixel, and L_{mask} is the average binary cross entropy loss. Mask-RCNN has no competition between classes, because other classes do not contribute losses. The mask branch has predictions for each category, and it depends on the classification layer to select the output mask.

The general semantic segmentation architecture can be regarded as an encoder-decoder network. An encoder is usually a pre-trained classification network, such as VGG, ResNet, and then a decoder network. Decoder networks constitute the difference of these architectures.

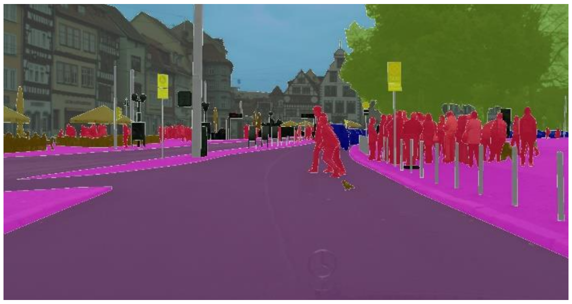

Figure 3. An example of a semantically segmented image

There are several other algorithms besides Mask-RCNN. For example, fully convolution networks for semantic segmentation could generate an output corresponding to the spatial dimension for pictures of any size. Among them, convolutional network is based on translation without deformation.

The main method is based on the success in image classification and transfer learning of deep network. Each layer of data in convolution network is a three-dimensional array of h*w*d, where h and w are spatial dimensions and d is feature or channel dimension. In this network, the first layer is an image with pixel size h*w and color channel number d. X_{ij} is recorded as the data vector at coordinates (i,j) at a specific level, and Y_{ij} is recorded at the next level. *Yij* is calculated as follows.

$$y_{ij} = f_{ks}(\{x_{si+\delta i, sj+\delta j}\}0 \leq \delta i, \delta j \leq k) \tag{4}$$

In this formula, k is the convolution kernel size, s is the step size or downsampling factor, and f_{ks} determines the type of layer.

C. Convolutional Neural Network

Convolutional neural network contains several basic structures, which are local receptive field, pooling, activation function and full connection layer. Many other networks are derived from this basic structure. This network structure is mainly used to identify displacement, scaling and other forms of 2D graphics. Now we will introduce the basic structure of neural network.

$$H_{Wb}(x) = f(w^T x) = f(\sum_{i=1}^{3} w_i x_i + b) \tag{5}$$

This unit can also be called Logistic regression model. When multiple units are combined and have a hierarchical structure, a neural network model is formed. And our convolution network can have many hidden layers, through which we can obtain features.

D. Detection of key points of human skeleton

This is the most critical technology in our whole system. It mainly detects some key points of human body, such as joints, facial features, etc., and describes human skeleton information through key points. There are many data sets that can be used to train such a model. For example, LSP and MSCOCO are excellent data sets. We are here mainly with a kind of spatial temporal graph convolutional networks to finish this work. In the form of 2D or 3D coordinates, the dynamic bone mode can be naturally represented by the time series of human joint positions. Then, human behavior recognition can be achieved by analyzing its action patterns. The main steps include the following aspects. Within each frame, a spatial map is constructed according to the natural skeleton connection relationship of human body. The same key points of two adjacent frames are connected to form a timing edge. Key points in all input frames form a node set and then all the edges in the previous steps form an edge set. That is, to form the required graph. Here we express the node set as follows. At this time, the graph is *G=(V,E)*, with n joints and a skeleton sequence of t frames.

$$V = \{v_{ti} | t = 1, \dots, T, i = 1, \dots, N\} \tag{6}$$

The feature vector on the node consists of the coordinate vector and confidence of the *i*-th joint on the *t*-th frame. The set of edges consists of two subsets. According to the connectivity of human body structure, the joints in a skeleton are connected by edges.

$$E_s = \{v_{ti} | (i, j) \in H\} \tag{7}$$

Each joint will be connected to the same joint in consecutive frames.

$$E_F = \{v_{ti} v_{(t+1)i}\} \tag{8}$$

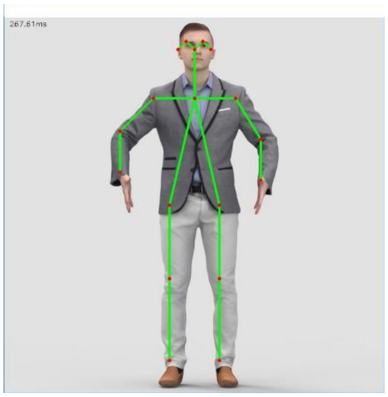

Figure 4. An example of Human Key Points Recognition.

E. Visualization technology

At the end of the system, we should design a reasonable visualization application scheme. A good visual application scheme can show our system more vividly in front of users. Here, the main task of visualization is to present our students' learning state. The front-end interface should have several elements. One is the recognition graph under the video stream, and the other is the fitting curve of students' overall state.

Since our overall plan is to analyze people's state through skeleton, it is necessary to show people the realistic results of key points of human body in the output interface.

IV. Scenarios that can be applied

Except the classroom, this system is still can be applied in other scenarios. Such a system can be used in conference rooms and other public places to supervise people's daily behavior. This system could benefit the supervision in the public areas ,which once required quantity of sources to be operated. Also ,it could also help analyze the habits of people and make the functions of prediction. Because the base of the system is the human pose estimation, we could easily transfer the application to other areas.

When talking to the topic of supervision, we have many methods to modify this system to make it quickly adapt to other application scenarios. We need to modify the angle or threshold of the corresponding event, which is always different in the different situation. But the upper analysis system is highly portable. This system has broad application prospects in monitoring, security and forecasting.

V. Conclusion

Here, we propose a system to analyze students' listening state by detecting key points of human body. The main points are the capture of key points of human body and the construction of post-analysis system. At the same time, the system has strong portability. The ultimate goal is to use such an analysis system to analyze students' listening state in classroom.The most basic idea of this project is to analyze human posture, motion trajectory and motion angle according to the key point information of human body.

However, when we actually created the system, we found other problems that needed to be solved. For example, the posture of human body under video is often blurred, which requires super-resolution. For the region to be detected, we should choose. If we use the object detection model directly to detect the human body, the accuracy of detecting the key points of the human body in the later period will decrease. At this time, we need to use semantic-based image segmentation algorithm to find out the human body more accurately. In the previous article, a semantic segmentation algorithm is mentioned, which is Mask-RCNN. When we use these algorithms to detect the human body in the image, the output results will make those noises smaller, and at the same time, the key points of the human body will be detected at the corresponding positions of the human body.

We can start our state analysis after obtaining more accurate data of key points of human body. When analyzing the state, our basic idea is to analyze the students' listening state by using the angle between different key points. Because different sitting positions can correspond to different states, we can start from some more different positions. Of course, this is aimed at those obvious postures, such as dozing off and listening to lectures. The design purpose of the system is to analyze more detailed students' postures, and analyze the available information from them. Therefore, we should let the machine learn the key points of human body, and finally get the key point model that can deal with complex situations. Finally, we need to visualize the key point information and analysis information that we get directly.

References

[1] He k, Gkioxari G ,Dollár P, et al. Mask r-cnn[C]//Proceeding of the IEEE international conference on computer vision. 2017: 2961-2969.

[2] Reddi S J, Hefny A, Sra, et al. On variance reduction in stochastic gradient descent and its asynchronous variants[C]//Advances in neural information processing systems. 2015: 2647-2655.

[3] Liao R, Yu S, An W, et al. A model-based gait recognition method with body pose and human prior knowledge[J]. Pattern Recognition , 2020, 98: 107069.

[4] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Thirty-second AAAI conference on artificial intelligence. 2018.

[5] Yi Y, Li A, Zhou X. Human action recognition based on action relevance weighted encoding[J]. Signal Processing: Image Communication, 2020, 80: 115640.

[6] Xu S, Liang L, Ji C. Gesture recognition for human–machine interaction in table tennis video based on deep semantic understanding[J] .Signal Processing: Image Communication, 2020, 81: 115688

[7] Jiang D, Wu K, Chen D, et al. A probability and integrated learning based classification algorithm for high-level human emotion recognition problems[J]. Measurement, 2020, 150: 107049.

[8] Shen M, Yu P, Wang R, Yuan J, Xue L. Image super-resolution reconstruction via deep network based on multi-staged fusion [J]. Journal of Image and Graphics, 2019,24(08):1258-1269.

[9] Li Y, Deng H, Xiang S, Wu J, Zhu L. Depth map super-resolution reconstruction based on the texture edge-guided approach[J]. Journal of Image and Graphics ,2018,23(10):1508-1517