Facial Target Detection and Keypoints Location Study Using MTCNN Model

Yuan Chai¹, Jing Liu² and Yang Li³

¹Department of Network Engineering, TJ E-Sports School of Network Science, Haikou University of Economics, Haikou, Hainan, 571132, China
²Research Center for Healthcare Data Science, Zhejiang Lab, Hangzhou, Zhejiang, 310000, China
³Foreign Languages Department, Oriental School of International Trade and Foreign Languages, Haikou University of Economics, Haikou, Hainan, 571132, China

Corresponding author’s e-mail: jingliuhnu2016@hotmail.com

Abstract: With the extremely rapid development of AI, facial recognition technology in the field of computer vision, which has the advantages of non-contact and easier to collect information, has become a hot topic in all walks of life. Facial target detection and keypoints location are two core parts of facial recognition technology. According to the analysis of the existing facial detection and keypoints location methods, by comparing their advantages and disadvantages, we find the existing problems. Based on the existing research methods, this paper proposes a high-precision facial detection and keypoints location method based on MTCNN model, which highlights the good accuracy of this method compared with the basic CNN or traditional algorithms, so as to improve the accuracy and effectiveness of facial recognition. It has more prominent advantages in this field, and is worthy of scientific research.

1. Introduction

Because of the continuous improvement of science and technology, as well as the emergence of facial recognition technology, there has been an increasing awareness of the use of identity authentication technology. Whether in daily work or in daily life, the efficiency of solving various difficult problems through facial recognition technology in various fields is extremely obvious. To realize a complete facial recognition system, it needs to be composed of several parts. facial detection and keypoints location are two important links, which directly affect the recognition accuracy of facial recognition system. Before facial detection, data will be collected, the difficulty of which will vary with the environment. The following situation will present some challenges to facial detection, such as posture or facial expression changes, light conditions, masks wearing, sunglasses and other shielding conditions. Based on the analysis of these, a cascade multi-task convolutional neural network framework is proposed in this paper. The network is divided into three stages for convolution to achieve detection and location from coarse to fine. By comparing the experimental results of the network improvement method and the original network structure, the optimal method is obtained. The two methods use the same training set and test set, which are trained by the traditional CNN Model and MTCNN Model, a facial detection and alignment algorithm based on Multitask convolutional neural network proposed by Zhang et al in 2016.[4]
2. Structural Principles

2.1. Structural Principles of CNN

CNN is short for convolutional neural network, which uses a mathematical operation called convolution on the neural network structure, so it is named convolution. Convolutional network can make CNN use of the two-dimensional structure of input data. At present, the application of this network in the field of image recognition and speech recognition can get good results. Compared with other image algorithms, convolutional neural network has less image processing steps and has certain advantages. The algorithm of facial detection and keypoints location studied in this paper belongs to this field.

2.2. Structural Principles of MTCNN

MTCNN is a kind of network architecture which is obtained by cascading multiple tasks of CNN.

It is an algorithm that combines facial detection area with keypoints location. The network architecture is divided into three layers and implemented by four steps. The first step is image preprocessing, which will resize the image into an effect with the minimum side length of 12 to serve for the subsequent P-NET. The second step is the P-NET convolutional neural network, which mainly resamples the input images to get the boxes to be tested removing the boxes that do not meet the requirements of the indicators, and then removing the parts with a high degree of convergence. This time, the main output is two parameters -- confidence and regression coefficient information. The specific network structure is shown in Figure 1.

In Fig. 1, Conv4-2 outputs the regression coefficient, while Prob1 outputs the confidence information and confidence degree of the BOX (which is the probability of facial). That is to say, each reshape image will be fed into the network, and both conv4-2 and prob1 will output. Box can be generated according to the output of P-NET. A threshold value (usually 0.6) can be set to get the candidate BOX. A point can be set in the center of the candidate BOX, and this point is the upper left corner, and expanding 12 pixels to the right and down will get a 12x12 rectangular box. Next, Confluence algorithm is applied to all 12x12 rectangular boxes detected on a frame of image, in which the confluence algorithm takes the place of traditional NMS algorithm. The main purpose of NMS algorithm in facial detection is to eliminate redundant windows and find the best facial detection location, but it depends on the setting of IOU threshold. If there are overlaps in the detected image, it is very difficult to determine the NMS threshold, which is prone to errors. If the IOU is too large, it will lead to more false detection. Conversely, it will lead to more missed detection. Based on the above situation, this paper proposes a method to replace NMS -- confluence. The difference between this method and NMS is that it does not depend on the IOU. Instead of suppressing a large number of detection results, it selects the best one among many detection results. Through Manhattan Distance, the so-called Manhattan Distance is the horizontal and vertical distance of all points. By measuring the correlation between boxes, select the box closest to other boxes in the cluster, and then remove the box with high coincidence degree. The advantage of this method is that it can keep the image which is suppressed by IOU through the original NMS mode.

Representation of Manhattan Distance between two points:

\[ MH(u,v) = |x_1 - x_2| + |y_1 - y_2| \]  \hspace{1cm} (1)

The closeness between two frames can be expressed by the distance between two diagonal points:

\[ P_{(m,n,u,v)} = MH_{(m,n)} + MH_{(u,v)} \]  \hspace{1cm} (2)

The final BOX will be placed in a four-dimensional array. The BOX given by P-NET can be corrected by regression. To change the rectangular box into a square, rec2square will be performed after modifying the BOX, and then padding will be performed, so as to turn the rectangular box into a square. Finally, these BOX boxes will be resize to 24x24x3, ready to be sent to the next R-Net network.
The third step, R-NET, is to further screen the BOX that does not meet the index requirements.

The convolution also has two parameter outputs, one is the confidence information of prob1, the other is the regression coefficient information of conv5-2. Like P-NET output, the difference lies in the network structure. The specific structure is shown in the figure. According to the two parameters of R-NET output, the generated box is filtered, the remaining box is processed by the consultation operation, then regression, rec2square, padding operation, and finally these BOXES are reset to 48x48x3, ready to be sent to O-NET. The network structure is shown in Figure 2.

O-NET is similar with R-NET. But it has three outputs, one more keypoints information of conv6-3 layer, which is also the final facial keypoints detection implemented by MTCNN. Through image preprocessing of MTCNN model and four steps of three convolution operations, a more accurate detection data will be obtained. The network structure is shown in Figure 3.
3. Keypoints
In image processing, the keypoints is essentially a feature. It is an abstract description of the physical relationship of a fixed area or space. It describes the combination or context within a certain neighborhood. It is not only a point information or a representation of a location, but also represents the combination of context and surrounding neighborhood. The goal of keypoints detection is to find out the coordinates of these points from the image by computer, so as to locate the key information of the facial. As a basic task in the field of computer vision, keypoints detection is very important for high-level tasks, such as recognition and classification.

Keypoints detection methods can be divided into two types. One type is to use coordinate regression to solve the problem, the other is to model the keypoints into thermal map, and get the keypoints position by pixel classification task and regression thermal map distribution. These two methods as means to address the problem are to find out the position and relationship of this point in the image.

4. The Choice of Loss Function
(1) Facial keypoints detection includes two tasks: classification and regression. The specific process is to classify facial and non-facial by image classification. The loss function used in facial classification is cross entropy function, as shown in formula 3.

$$L_{x} = -y_{x} \log(p_{x}) + (1 - y_{x}) \log(1 - p_{x})$$  \hspace{1cm} (3)

$p_{x}$ indicates the facial probability of the x-th position in the network test sample, while $y_{x}$ indicates the label of the sample at position X.

(2) The loss function used in facial frame regression is the mean square loss function, which is the Euclidean distance, as shown in formula 4.

$$L_{x}^{box} = \|y_{x}^{box} - y_{x}^{box}\|_{2}^{2}$$  \hspace{1cm} (4)

$y_{x}^{box}$ indicates the location of the X-person facial frame of the network test sample, while $y_{x}^{box}$ indicates the actual position of the x-th facial border.

(3) The keypoints location of facial is similar to the regression task, so the loss function is the same as that of facial frame regression.

5. Model Training
First, data sets need to be preprocessed. Second, after the feature extraction model is used to extract the features of the image, the output points correspond to the abscissa and ordinate of the keypoints, and
then the coordinates of the keypoints are returned to the processed image. Last, the image of facial keypoints detection can be printed out obtained. The specific process is as follows:

Step one is to select the appropriate data set. This experiment is based on the platform of paddle2.0 environment, and the selected public data set is Wider facial and user-defined test set. Wider facial dataset is a benchmark dataset for facial detection provided by the Multimedia Laboratory of the Chinese University of Hong Kong. There are more than 30000 images, including complex facial expression, posture, occlusion and different lighting conditions. Training data can be downloaded from this dataset and image preprocessing is carried out through transform: Step two is facial detection based on P-NET model training. Step three is R-Net model training for border regression. Step four is O-NET model training for the completion of facial keypoints localization. Each training depends on the previous model training results. After the above four steps of training, the facial box and keypoints in the picture are recognized, and the results are shown in Figure 4.

Figure 4. facial Detection and keypoints Location Results

6. Analysis of Experimental Results

This experiment is mainly to compare the three indicators, namely the accuracy rate, omission rate and detection speed. The Wider facial data set is used as training set and test set respectively. The results are compared with those of CNN model and MTCNN, as shown in Table 1.

| Detection Model | Accuracy(%) | Omission(%) | Speed(S) |
|-----------------|-------------|-------------|----------|
| CNN             | 80.269      | 18.024      | 2.463    |
| MTCNN           | 90.335      | 11.769      | 0.098    |

Through the comparative analysis of the above tables, the effectiveness of the proposed MTCNN based facial detection and keypoints location is verified. Compared with CNN Model, MTCNN Model can detect more facials in the data set with higher accuracy and detection efficiency. Thus, It is an efficient method to improve the efficiency of model detection by adding convolution task.

7. Conclusion

Taking how to carry out facial detection and keypoints location under non-special conditions as the application background, this paper studies how to improve and innovate the existing CNN technology to improve the accuracy and robustness of facial detection. Due to the influence of many adverse factors, such as illumination conditions, occlusion, facial detection and keypoints location are faciald with certain challenges, which are extremely prone to facial detection, false detection, missing detection, keypoints deviation and so on. Therefore, the research on how to improve the existing technology to improve the method of facial detection and keypoints location under non-special conditions can not only
broaden the application scope of the existing facial detection, but also lay the foundation for the facial recognition technology of abnormal driving behavior of motor vehicle drivers.

Acknowledgement
This project is supported by the Education Department of Hainan Province(Grant No. Hnky2021-54) and Haikou University of Economics(Grant No.HJKY(YB)21-01)

Reference
[1] Redmon J, Farhadi A. YOLOv3 : An incremental improvement[J]. arXiv:1804.02767, 2018.
[2] Ren S, He K, Girshick R B, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.
[3] Sun Y, Wang X, Tang X. Deep convolutional network cascade for facial point detection[C]/Computer Vision and Pattern Recognition(CVPR). Portland, OR, USA: IEEE, 2013: 3476-3483.
[4] ZHANG K, ZHANG Z, LI Z, et al. Joint facial Detection and Alignment using Multi-task Cascaded Convolutional Networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.
[5] JIA X, ZENG S, et al. Fast facial Detection Based on Improved MTCNN network[J]. Computer Engineering and Science, 2020, 42(07): 1262-1266.
[6] LAN W, ZHANG S, ZHU R, XIONG, W. Facial Detection Algorithm Based on Improved MTCNN [J]. Journal of South-Central University for Nationalities (Natural Science Edition), 2020, 39(06): 637-641.
[7] LI T. Design and Implementation of Real Time facial Recognition System Based on OpenCV[D]. Northeast Petroleum University, 2020.
[8] SUN M, LIANG L, et al. Facial keypoints Location Algorithm Based on Cascaded Convolutional Network[J]. Journal of University of Chinese Academy of Sciences, 2020, 37(04): 562-569.
[9] WU J, CHEN S. An Improved MTCNN facial Detection Algorithm[J]. Software Guide, 2019, 18(12): 78-81.