

PAPER • OPEN ACCESS

A Fast Facial Landmarks Detection and Posture Classification Algorithm

To cite this article: Shengyong Li and Zhihua Zhang 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **563** 052011

View the [article online](#) for updates and enhancements.

A Fast Facial Landmarks Detection and Posture Classification Algorithm

Shengyong LI, Zhihua ZHANG

Nantong Shipping College, Nantong, Jiangsu province, China 226000

Corresponding author's e-mail: lisy@ntsc.edu.cn

Abstract. In the field of face recognition research, the efficiency of facial landmarks detection and posture classification algorithm is a serious problem. This paper proposes a lightweight network based on convolutional neural network to quickly detect the facial landmarks of human faces. Based on this, the posture is predicted using the relative position of the obtained feature points. The model was tested by the face images captured in various scenes in real life. The accuracy and recall rate of the proposed algorithm were 96.3% and 98.2%, respectively. The test time for a single picture is 0.9ms. The proposed algorithm has less running time and its accuracy and recall rate are similar or even better than other models.

1. Introduction

Face facial landmarks detection is one of the research hotspots of machine vision. It is designed to detect the location of predefined facial signs, such as the corners of the eyes, eyebrows, nose tips. With the development of deep convolutional neural networks, facial features detection is widely used in tasks such as facial recognition [1], head posture estimation [2] and 3D facial reconstruction [3]. The latest developments in facial landmark detection mainly focus on learning characteristics through different facial shapes, different postures, and partial occlusion. Head pose estimation also plays a vital role in the field of artificial intelligence such as machine vision.

2. Related Work

A common method of facial feature detection is to learn the regression model. Many studies use deep convolutional networks to learn facial features and regressions in an end-to-end manner and predict feature points in a cascaded manner. Ren [4] and Zhu et al. [5] used convolutional neural networks or hand-crafted features to characterize facial appearance and shape information, and then used model training to predict and locate marker points. Zhu et al. [6] use a cascading method to connect the prediction module and gradually update the predicted position of the marker points. However, these studies ignore problems such as changes in image styles, so they are prone to false detections. Yu et al. [7] proposed a deep deformable network, but the slower detection speed made it unable to suffice the actual requirements. Dong et al. [8] proposed a style aggregation network for facial feature points detection, which performed better on multiple style images. However, it is less accurate when detecting faces with different angles. Head pose classification is usually based on methods such as feature points, 3D reconstruction, statistical learning, and image features. However, these methods usually cannot satisfy both accuracy and real-time.

In response to the above problems, a lightweight network has been proposed to achieve fast facial landmarks detection and pose classification without sacrificing accuracy.



3. Material and Method

The important objects of facial landmarks detection are the five key parts of the eye, eyebrow, nose and mouth. The usual method is to use a deep learning network to convolve face images to extract related features. One of the effective methods of pose classification is based on the features of the facial features, but it generally needs to be implemented step by step. The proposed method uses a single CNN to simultaneously implement facial landmarks detection and pose classification.

3.1 Lightweight convolutional neural network

The proposed network is built with reference to the VGG16 model. The original VGG16 has more layers and channels, which results in a lot of useless calculations. The proposed network consists of only 6 convolutional layers and 2 fully connected layers, which greatly reduces the number of network parameters. In addition, the size of the input image is also an important factor affecting the efficiency of the model. Therefore, the original image is scaled to 40x40 to reduce parameter calculations during back propagation. This does not affect the accuracy because the convolutional layer of the 5x5 kernel used in the original VGG16 network is replaced by two 3x3 kernel-sized convolutional layers. The change refines the features. The specific network structure is shown in figure 1.

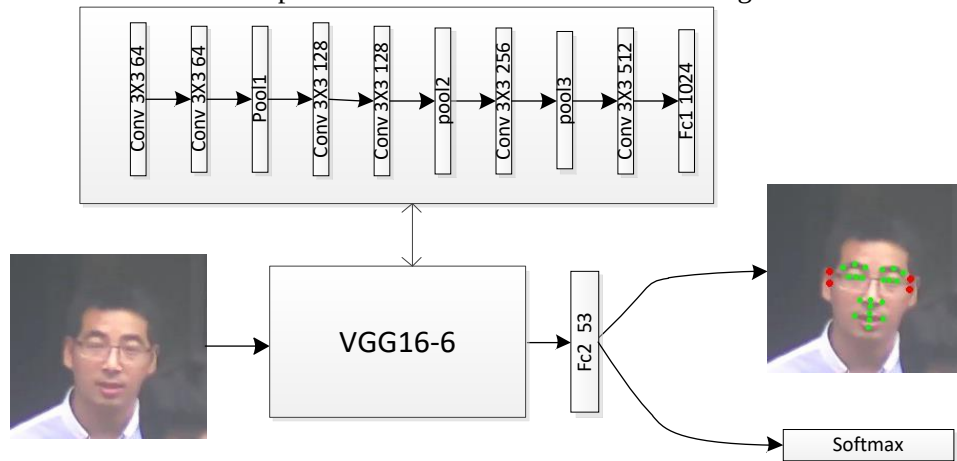


Figure 1. Lightweight network structure

In figure 1, 3x3 represents the size of the convolution kernel, and 64, 128, 256, 512, and 1024 represent the output channels of the corresponding convolutional layer or fully connected layer, respectively. In addition, the number of channels in the FC2 layer is 53, which represents the predicted coordinates of the 25 feature points output and the prediction probability of the three postures of face, side face and bow. VGG16-6 represents the network infrastructure built. The max pooling is selected in the network structure.

The original image is input into the network to learn the features, and then the two-dimensional feature is converted into a one-dimensional vector by the fully connected layer FC1 and output to the FC2 layer for facial landmarks detection. The Euclidean method (such as Equ. (1)) is used to implement loss calculation in 50 channels used for feature point location. It is used to evaluate the error between the predicted point and the real label. The Softmax function is still used in the three channels of the pose classification to implement the classification calculation.

$$L_E = \frac{1}{2N} \sum_{i=1}^N \|pred - truth\|_2^2 \quad (1)$$

Where, N is the number of samples, pred indicates the predicted point and truth represents the real label. Finally, the loss of the entire network is weighted by Euclidean and Softmax. That is 0.9x the Euclidean loss plus 0.1x the Softmax loss, as shown in Equ. (2).

$$L = 0.9L_E + 0.1L_S \quad (2)$$

Where L_S is the Softmax loss

3.2 Dataset and processing

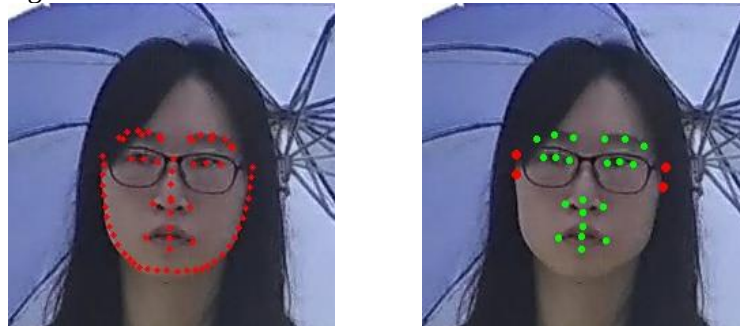
The original face image is collected in different scenes at different times. This makes the background different even if the different images belong to the same person. Therefore, the robustness of the model to each scene picture can be enhanced during the training process. This experiment uses more than 45,000 face images, and there are pictures with dense crowds, complicated background, blurred pictures, and too dark or too bright light. Of these, 40,000 images were used for training and the remaining images were used for testing. The training picture is divided into three categories: positive face, side face and low head, as shown in figure 2.



Figure 2. Positive face, side face and bow three samples

In addition, the sample needs to be pre-processed before training the network. First, the mean of the picture is subtracted from each pixel value in the image. The result is then divided by the variance of the picture. Background information may have an impact on model performance, so the image in the face frame recognized by the face detector is pre-processed and finally used to train the model.

Annotation is a huge job if the collected images are used directly to train the model. Therefore, the better model of the predecessors is used for automatic calibration of feature points. Then, the sample that was marked incorrectly was manually adjusted. In addition, 25 of the 68 feature points having the original face contour information are selected as the feature points to be learned. The purpose is to meet the requirements of rapid detection in actual engineering. This reduces the input information and causes a large number of feature points with less contribution to be discarded. Feature point selection method is shown in figure 3.



(a) Original 68 feature points. (b) The final 25 feature points.

Figure 3. Feature point selection

3.3 Evaluation method

The final purpose of the experiment is to judge the posture of the head based on the features of the facial landmarks. This is a classification task. Therefore, the precision, recall and single-frame image test time are selected as indicators for evaluation, as in Equ. (3) and Equ. (4).

$$P = \frac{TP}{TP + FP} \quad (3)$$

$$R = \frac{TP}{TP + FN} \quad (4)$$

Where, TP represents the number of samples correctly classified, FP indicates the number of negative samples that are judged to be positive samples and FN refers to the number of positive samples that are discriminated as negative samples.

4. Results and Discussion

All experiments were done by the Caffe [9] framework that exists in the Linux operating system.

4.1 Parameters Setting

During the whole training process, the momentum is set to 0.9, the weight decay coefficient is 0.0001, and the initial learning rate is assigned 0.0001. Then, the batch size is 128 and the maximum number of iterations is 200,000. In addition, the learning rate was reduced to 0.1x when the number of iterations was 80,000 and 160,000 times, respectively.

4.2 Parameters Setting

Other uncommented 600 face images were used for model testing. The advanced residual network and GoogleNet are compared to the results obtained by this method. Finally, the precision, the recall, and the detection time of each picture of the three network models were obtained. The results are shown in table 1.

Table 1. Comparison with results of advanced architecture

Network	Precision/%	Recall/%	Times of each frame/ms
ResNet	85	87.5	98
GoogleNet	97.1	98.9	3
VGG16-6	96.3	98.2	0.9

As can be seen from table 1, the test speed of the single frame picture of VGG16-6 exceeds that of GoogleNet 3x, which is 10x of the residual network. Although GoogleNet has a slightly higher precision and recall than VGG16-6, this small gap has been tested and found to have little impact on results and engineering applications. In addition, the precision and recall of VGG16-6 are more than 10% higher than the residual network. From the time index, the running speed of VGG16-6 has obvious advantages. In summary, its test results can meet the requirements of actual engineering.

5. Conclusion

Possible problems in the task of locating the facial landmarks and the corresponding head pose classification tasks. The proposed algorithm can achieve its goal quickly without sacrificing accuracy. Therefore, the method can reduce the cost of equipment operation and improve work efficiency. The lightweight network has a more important significance for the study of face recognition.

References

- [1] Liu Y, Wei F, Shao J, Sheng L, Yan J and Wang X 2018 *Int. Conf. on Computer Vision and Pattern Recognition* (Salt Lake City) (Washington: IEEE Computer Society) pp 2080-2089
- [2] Wu Y, Gou C, Ji Q 2017 *Int. Conf. on Computer Vision and Pattern Recognition* (Hawaii) (Washington: IEEE Computer Society) pp 5719-5728
- [3] Liu F, Zeng D, Zhao Q and Liu X 2016 *Int. Conf. on European Conference on Computer Vision* (Amsterdam) (London: Springer) pp 545-560
- [4] Ren S, Cao X, Wei Y and Sun J 2014 *Int. Conf. on Computer Vision and Pattern Recognition* (Columbus) (Washington: IEEE Computer Society) pp 1685-1692
- [5] Zhu S, Li C, Chen C and Tang X 2015 *Int. Conf. on Computer Vision and Pattern Recognition* (Boston) (Washington: IEEE Computer Society) pp 4998-5006
- [6] Zhu S, Li C, Chen C and Tang X 2016 *Int. Conf. on Computer Vision and Pattern Recognition* (Las Vegas) (Washington: IEEE Computer Society) pp 3409-3417
- [7] Yu X, Zhou F and Chandraker M 2016 *Int. Conf. on European Conference on Computer Vision* (Amsterdam) (London: Springer) pp 52-70

- [8] Dong X, Yan Y, Ouyang W and Yang Y 2018 *Int. Conf. on Computer Vision and Pattern Recognition* (Salt Lake City) (Washington: IEEE Computer Society) pp 379-388
- [9] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S and Darrell T 2014 *Int. Conf. on Multimedia* (Orlando) (New York: ACM) pp 675-678