Vehicle Face Detection Based on Cascaded Convolutional Neural Networks

Shen Jing
School of Automation
Nanjing University of Posts and Telecommunications
Nanjing 210021, China
601127529@qq.com

Changhui Hu*
School of Automation
Nanjing University of Posts and Telecommunications
Nanjing 210021, China
hchnjupt@126.com

Cailing Wang
School of Automation
Nanjing University of Posts and Telecommunications
Nanjing 210021, China
wangcl@njupt.edu.cn

Guangliang Zhou
School of Automation
Nanjing University of Posts and Telecommunications
Nanjing 210021, China
1466088453@qq.com

Jian Yu
School of Automation
Nanjing University of Posts and Telecommunications
Nanjing 210021, China
1256437479@qq.com

***Abstract*—Vehicle face detection is an important and challenging task in face detection. On one hand, the face image is affected by the complex environment during driving so that the facial details are easily lost. On the other hand, there is no large public vehicle face dataset. In order to solve these problems, firstly, inspired by the convolutional neural network, an improved method based on cascaded convolutional neural networks is proposed in this paper. The cascaded convolutional networks consist of three lightened convolutional neural networks. Each convolutional neural network uses multi-task learning. Feature fusion technology focusing on learning details of vehicle face image is also adopted in cascaded convolutional networks. Secondly, due to the lack of vehicle face image samples, we build a small vehicle face dataset by ourselves, named Driver FACE dataset. Experiments on Driver FACE dataset show that our method improves the performance of vehicle face detection compared with baseline.**

Keywords—cascaded convolutional neural networks, feature fusion, vehicle face detection

I. INTRODUCTION

Nowadays, face detection has been a hot research topic in the field of computer vision. Its main goal is to accurately find out the number of faces, size of faces and the location of faces from the input original image. Face detection technology has been constantly developing and improving from the traditional face detection algorithm stage to the deep learning algorithm stage [1]. It already has been widely used in many fields such as law enforcement identification, bank security certification, automatic airport screening, etc.[2]. Especially in the field of traffic monitoring, improving the face detection technology of vehicle drivers plays an important role to improve the traffic management. However, the face image acquisition condition of the vehicle driver is usually very poor, because of the influence of the complex driving environment of the vehicle [3] , such as illumination, pose, occlusion [4]. Therefore, the vehicle face detection faces a great challenge.

Early face detection methods are not enough to be put into practice. In 1995 and 1996, Chen et al. [5] and Yow et al. [6] proposed two kinds of face detectors respectively. However, it is proved that these methods lack robustness in complex environment. In 2004, Paul Viola and Michael Jones [7] designed a fast and accurate face detector, called VJ face detector. It greatly improved the speed and accuracy of face detection and was a milestone of face detection technology. In 2012, the success of AlexNet [8] made deep learning gradually become a hot research. The convolutional neural network is a very common architecture in deep learning, called CNN, which is mainly inspired by the biological natural visual cognition mechanism. In 2013, Zhang et al. [9] began to use CNN to extract features for face feature point detection and achieved good results on the FDDB [10] dataset. In 2015, Li et al. [11] proposed a cascade CNNs for face detection, which improved the performance of face detection on FDDB dataset. In 2016, Zhang et al. [12] improved cascaded CNNs, using multi-task cascaded convolutional neural networks. And that further improved the performance of face detection. In 2017, Jiang et al. [13] applied the fast R-CNN to face detection and achieved good results.

Inspired by the convolutional neural network, the main work of this paper for the task of vehicle face detection is summarized as follows: First, our method adopt the cascaded convolutional networks. It consists of three convolutional neural networks, and the number of network layers is gradually deepened, from simple to complex. We carry out the face detection task and the face feature points detection task at the same time in each convolutional network. Since there is an intrinsic correlation between the two tasks. The range of the face limits the range of the facial landmarks, and the coordinates of the facial landmarks have a strong position correlation with the output border of the face. Second, we use a new activation function, called Max-Feature-Map (MFM). It is proposed by Wu et al. [14], and its performance in the convolutional layer tends to obtain more significant and discriminative nodes. Third, our method uses feature fusion technology to fully learn the details of the vehicle face image by combining the bottom features with the high level features of the image. In order to make the network have better detection performance for vehicle face image, we build a small vehicle face dataset with a total of 1680 vehicle face images. We randomly selecting 80% of dataset for training, and the remaining 20% for testing. Since there are very few pictures in the dataset, which is not enough to support the

*Corresponding author: Changhui Hu. This work was supported by the National Natural Science Foundation of China (Grant No.61802203), Natural Science Foundation of Jiangsu Province (Grant No.BK20180761), China Postdoctoral Science Foundation (Grant No.2019M651653), Postdoctoral Research Funding Program of Jiangsu Province (Grant No.2019K124) and NUPTSF (Grant No.NY218119).

training of the whole network, thus we adopt the training method of pre-training with the public dataset and retraining with vehicle face dataset.

The remainder paper is designed as follows: Section II introduces the proposed approach. Section III introduces the experiment. Section IV summarizes this paper.

II. APPROACH

A. Overall Framework

The overall framework of vehicle face detection is shown in Fig. 1. It is mainly summarized as three stages: (1) Data processing. We resize the vehicle face image to different scales to build an image pyramid. (2) Input the processed images into the cascaded convolutional network. Each CNN detector evaluates the images, removes most of the face-free windows, and leaves the candidate windows. Meanwhile, non-maximum suppression (NMS) is used to merge highly overlapping candidate windows. The output of each CNN detector resizes the scale as the input of the next detector. (3) Output the final face windows and mark the position of the five feature points of the face (eyes, nose, mouth corners).

Fig. 1. Framework of cascaded convolutional networks

B. Each Convolutional Neural Network

1) Driver-12net: As shown in Fig. 2, Driver-12 net is the first stage of the cascaded framework. It is a fully convolutional network and uses 3*3 and 2*2 filters to obtain more detailed feature information for better performance while reducing the amount of computation. Compared with the normal convolutional neural network whose last layer uses fully connected network, the fully convolutional network can adapt to the input images of various scales. In addition, the last layer uses a 1*1 convolutional layer instead of a fully connected layer, which can widen the width of the network with minimal parameters.

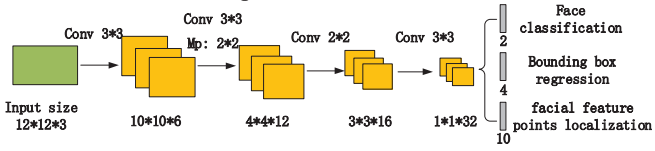

Fig. 2. Driver-12net

2) Driver-24net: The candidate windows obtained from Driver-12net will be input into the second CNN detector, as shown in Fig. 3, and we call it Driver-24net. In the first few layers of the convolutional network, we use a new activation function: Max-Feature-Map (MFM) activation function [14]. Compared with common activation functions such as Sigmoid, ReLu and PReLu, MFM is a statistical method which can not only make feature representation more compact but also realize variable selection and dimension reduction.

3) Driver-48net: Driver-48net is the last stage of the cascaded detector, so we set the layer deeper and extract the face features more carefully to ensure that the candidate windows can be screened accurately. As shown in Fig. 4, similar to Driver-24net, the first few layers of the network abandon the common activation function and use the MFM activation function. The difference is that in this stage we introduce a feature fusion method. In the deep convolutional network, the detail features are often concentrated in the bottom of the network, and fusing the bottom features and high level features in a certain proportion can make the features extracted by the network more detailed to a certain extent. In this paper, the fusion factors are 0.2 and 0.4, respectively.

C. Loss Function

As shown in the convolutional network structure, each convolutional network has three learning tasks, which can better extract the details of the face and achieve better detection results.

For face classification task, this is a simple two-class classification problem, which is divided into two categories: human face and non-face. For bounding box regression task and facial landmark localization task, we both use the Euclidean loss. Therefore, we define the final loss function as follows:

$$loss = \alpha \sum_{i=1}^{M} (-(y_i^{face} \log(p_i) + (1 - y_i^{face})(1 - \log(p_i))) + \beta \sum_{i=1}^{L} \left\| \hat{y}_i^{box} - y_i^{box} \right\|_2^2 + \gamma \sum_{i=1}^{N} \left\| \hat{y}_i^{feature} - y_i^{feature} \right\|_2^2 + L2 \quad (1)$$

where M,L,N are the number of training samples for each learning task respectively. In order to balance the importance of each task in each stage of the CNN detector, we use (α=1, β=0.5, γ=0.5) in Driver-12net and Driver-24net. In Driver-48net, the task of locating facial feature points is more important, so we use (α=1, β=0.5, γ=1). $L2$ is the regularization loss.

III. EXPERIMENTS

A. Training Dataset

Dataset is very important in deep learning. The type and number of training dataset directly affect the performance and generalization of the trained network models. This paper mainly aims at the task of vehicle face detection. However, since there is no publicly and accurately marked vehicle face dataset, and the dataset collection is difficult, the training process of this paper is divided into two phases.

Phase 1: We first pre-train the proposed cascaded convolutional networks with the WIDER FACE [15] dataset and the CelebA [16] dataset. The WIDER FACE dataset contains 32,203 images and 393,703 labeled faces, 40% of which are training sets that contain variations in scale, lighting, occlusion, and pose. The CelebA dataset contains 202,599 face images, each of which is labeled with the coordinates of the five feature points of the face. The WIDER FACE dataset is mainly used as a training set for face classification task and bounding box regression task, while the CelebA dataset is mainly used as a training set for face feature points location task.

Phase 2: After pre-training the cascaded convolutional networks, we fine-tune the networks by using a new vehicle

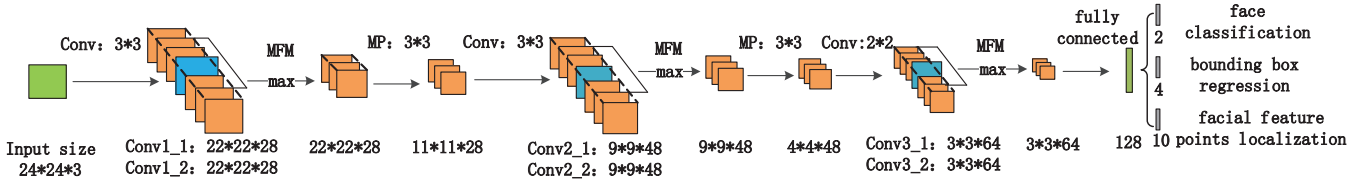

Fig. 3. Driver-24net

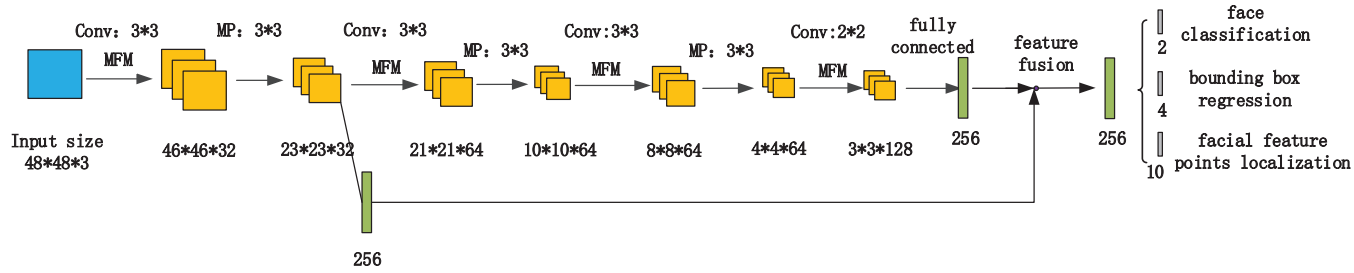

Fig. 4. Driver-48net

face dataset and LFW [17] face feature point dataset. In this paper, the vehicle face dataset we proposed is mainly collected by image acquisition tool by ourselves and collected from the Internet, which is called DRIVER FACE dataset. There are 1680 vehicle face images in this dataset, and we manually annotate the faces. Since the dataset is small and the training data is especially important for the deep neural network, we randomly select 80% of them as the training set for face classification task and bounding box regression task. The remaining 20% are used as the test set. Fig. 5 shows images and annotations of some DRIVER FACE dataset. In addition, the LFW face feature point dataset contains 5590 faces, each of which is labeled with the coordinates of the five feature points of the face. We use it as the training set for the face feature point location task.

Fig. 5. Part of images and annotations in DRIVER FACE dataset

B. Training Process

Since the training process of the two phases is basically the same, in this paper we mainly describe the training process of the second phase. At this time, the training of the cascaded convolutional networks is based on the model that has been trained in the first phase.

For three learning tasks, we mainly divide the training dataset into four categories: positives, part faces, negatives, and feature point faces. The selection criteria for positives, part faces and negatives is determined by the Intersection-over-Union (IoU) rate between the random clipping window and any ground-truth face in the image. The IoU rate of positives is greater than 0.65. The IoU rate of part faces is between 0.4 and 0.65. The IoU rate of negatives is less than 0.3. The positives and negatives are for face classification task, and the positives and part faces are for bounding box regression task.

For Driver-12net training, we randomly crop the positives, part faces and negatives that we need from the DRIVER FACE dataset, and crop faces from the LFW dataset as feature point faces. And we resize all samples to 12*12.

For Driver-24net training, we use the trained Driver-12net to detect faces from DRIVER FACE dataset to get the positives, part faces and negatives. We crop faces from the LFW dataset as feature point faces. All samples are resized to 24*24.

For Driver-48net training, similar to Driver-24net, we use the trained Driver-12net and the trained Driver-24net to detect faces from DRIVER FACE dataset to get the samples we need. And we resize all samples to 48*48.

C. Evaluation on Face Detection

1) Evaluation on FDDB: FDDB dataset contains 2,845 images with 5,171 faces. As shown in Fig. 6, we compare our method after the first phase of training with other methods [7, 11, 18, 19, 20,]. Our method has a good performance.

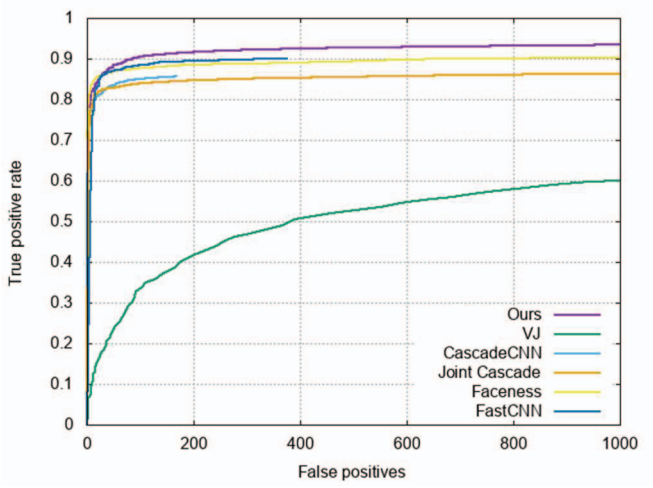

Fig. 6. Evalution on FDDB

2) Evaluation on DRIVER FACE: On DRIVER FACE test set, we test our method after each phase of training, respectively. And we evaluate the results using the evaluation criteria in PASCAL VOC [21]. In order to prove

the improvement of our method on vehicle face by using cascaded networks, we also use the same training method and the same dataset to train MTCNN [12]. We compare these two methods on DRIVER FACE test set. As shown in Table I, after the first phase of training, our method is slightly lower than the MTCNN on precision, but the recall is slightly higher. In general, the F-Measure value is slightly higher than MTCNN.

TABLE I. EVALUATION AFTER THE FIRST PHASE OF TRAINING

Method	*recall(%)*	*precision(%)*	*F-Measure(%)*
MTCNN	59.006	96.284	73.171
Ours	60.455	95.425	74.017

However, the performance of the cascaded networks without training with vehicle face images is not very well. Thus, the second phase of training that using DRIVER FACE training set is necessary. Fig. 7 shows that the faces which has not been detected in the first phase has been correctly detected after the second phase of training. It is obviously that the detection performance has improved.

Fig. 7. Detections of some images on DRIVER FACE test set

Table II confirms above. Compared with the first phase, the performance after retraining with the vehicle face images is significantly improved. Moreover, in precision and recall, our method is higher than MTCNN. Our method can get a good result on vehicle face image.

TABLE II. EVALUATION AFTER THE SECOND PHASE OF TRAINING

Method	*recall(%)*	*precision(%)*	*F-Measure(%)*
MTCNN	85.921	92.428	89.056
Ours	86.957	94.382	90.517

IV. CONCLUSION

In this paper, a method based on cascaded convolutional networks are proposed to solve the vehicle face detection task. The cascaded convolutional networks consist of a fully convolutional network and two convolutional networks using MFM activation function, and each convolutional network uses multi-task learning to improve the accuracy of face detection. Our method detects the vehicle face and feature points at the same time, which also facilitates the face alignment work required for vehicle face recognition. Our method also uses the feature fusion to learn more details of vehicle face image. In addition, the proposed method trains the network in stages, and improves performance on vehicle face detection with limited resources.

REFERENCES

[1] Hu C, Lu X, Liu P, Jing X, and Yue D, "Single Sample Face Recognition Under Varying Illumination via QRCP Decomposition", IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 2624-2638, May 2019.

[2] Zhang Cha and Zhang Zhengyou, "A Survey of Recent Advances in Face Detection", Technical Report, MSRTR-2010-66, June. 2010.

[3] Hu C, Lu X, Ye M, Zeng W. Singular value decomposition and local near neighbors for face recognition under varying illumination[J]. Pattern Recognition, 2017, 64: 60-83.

[4] Hu C, Ye M, Ji S, Zeng W, Lu X. A new face recognition method based on image decomposition for single sample per person problem [J]. Neurocomputing, 2015, 160: 287-299.

[5] Yachida M, Wu H, Chen Q. Face detection by fuzzy pattern matching[C]. International Conference on Computer Vision, 1995, 591-596.

[6] Taagepera R, Yow K C, Cipolla R. Feature-based human face detection[J]. Image and Vision Computing, 1997, 15(9): 713-735.

[7] Viola P, Jones M J, "Robust real-time face detection. International journal of computer vision", vol.57, no. 2, pp. 137-154, 2004.

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

[9] Zhang C, Zhang Z. Improving multiview face detection with multi-task deep convolutional neural networks[C]. Applications of Computer Vision, 2014: 1036-1041.

[10] V. Jain, and E. G. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings", Technical Report UMCS-2010-009, University of Massachusetts, Amherst, 2010.

[11] Li H, Lin Z, Shen X, et al, "A convolutional neural network cascade for face detection", in IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5325-5334.

[12] Zhang K, Zhang Z, Li Z, et al. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks[J]. IEEE Signal Processing Letters, 2016, 23(10):1499-1503.

[13] Jiang H, Learned-Miller E. Face Detection with the Faster R-CNN[J]. 2016.

[14] Wu X, He R, Sun Z. A Lightened CNN for Deep Face Representation[J]. Computer Science, 2015.

[15] Yang S, Luo P, Loy C C, Tang X, "WIDER FACE: A Face Detection Benchmark". arXiv preprint arXiv:1511.06523.

[16] Liu Z, Luo P, Wang X, Tang X, "Deep learning face attributes in the wild", in IEEE International Conference on Computer Vision, 2015, pp. 3730-3738.

[17] Sun Y, Wang X, Tang X. Deep Convolutional Network Cascade for Facial Point Detection. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[18] Chen D, Ren S, Wei Y, Cao X, Sun J. Joint cascade face detection and alignment. In ECCV, pages 109–122. Springer, 2014.

[19] Yang S, Luo P, Loy C C, Tang X. From facial parts responses to face detection: A deep learning approach. In ICCV, pages 3676–3684, 2015.

[20] Danai Triantafyllidou, Anastasios Tefas. A Fast Deep Convolutional Neural Network for Face Detection in Big Visual Data. Advances in Big Data, 2016.

[21] Everingham M. The PASCAL Visual Object Classes Challenge, (VOC2007) Results[J]. Lecture Notes in Computer Science, 2007, 111(1):98-136.