A Deep Learning Approach for Face Detection and Location on Highway

Yang Zhang 1,2, Peihua Lv1,2 and Xiaobo Lu1,2,*

1School of Automation, Southeast University, Nanjing 210096, China
2Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University, Nanjing 210096, China

*E-mail: xblu2013@126.com

Abstract. Face detection and location technique is a hot research direction during recent years. Especially, driver face detection on highway is still a challenging problem in social safety deserving research. This paper proposes a novel algorithm based on the improved Multi-task Cascaded Convolutional Networks (MTCNN) and Support Vector Machine (SVM) to realize accurate face region detection and feature location of driver's face on highway, predicting face and feature location via a coarse-to-fine pattern. The proposed algorithm is verified under various complex highway conditions. Experimental results show that the proposed model shows satisfied performance compared to other state-of-the-art techniques used in driver face detection and alignment, keeping robust to the occlusions, varying pose and extreme illumination on highway.

Keywords. Driver face detection; Face alignment; Convolutional Networks; Support Vector Machine

1. Introduction

Face detection and alignment technology has been widely used in various practical fields, especially in driver face detection and alignment area, which involves public security and traffic order. With the fast development of digital image processing technology, various detection and alignment techniques have been proposed [1]. However, in real applications, there exit varying illumination, occasion as well pose affect the detection and alignment performance.

Face detection is a hot research direction in these years. In 2004, Viola [2] put forward cascade detection method based on AdaBoost with Haar-Like features to perform cascaded classifiers. Unfortunately, some later researches [3, 4] indicate that it cannot keep continued competitiveness in real applications which affect the visual consistency of faces. Afterwards, deep CNNs are utilized in face detection. Yang et al. [5] proposed the deep neural networks for facial feature recognition. However, this algorithm is time consumely in real condition. Later, Kaipeng Zhang designed the MTCNN [6] model, which is consisted of three layers of precise deep convolutional networks. Nevertheless, due to its fixed training samples and limited network structure, when it comes to real condition, such as driver face detection, it show unsatisfied performance.

At the same time, face alignment also attracts wide study interests. The research fruits in this area can be approximately divided into two parts[7-12], regression-based techniques and template fitting methods. Nevertheless, previous methods about face detection and alignment methods overlooked the inherent relationship among these two issues.

The proposed algorithm is verified under various complex highway conditions. Experimental results show that the proposed model shows satisfied performance compared to other state-of-the-art techniques used in driver face detection and alignment, keeping robust to the occlusions, varying pose and extreme illumination on highway.
In this paper, we propose a driver face detection and alignment model based on improved MTCNN and SVM[13], called IMT-SVM, forming adaptive detection and alignment model. We train this model on our own traffic driver face database, which is constructed by Public Security Department of Jiangsu Province, verifying the performance on the testing dataset. According to the experimental result, the IMT-SVM model shows high detection rate as well as low error detection rate compared to other state-of-the-art method.

2. Basic Theory

In this paper, we do research on the structural improvement of MTCNN and realize the fusion of SVM, improving the performance of the original algorithm. In this section, the basic theory about face detection and alignment which are utilized in our model are introduced.

2.1. MTCNN

The main purpose of MTCNN is to construct the image pyramid of the corresponding face image. The overall stages of MTCNN which is constituted of three convolutional networks are illustrated below.

2.1.1. Proposal Network (P-Net). The fully convolutional network is utilized to get the rough facial windows as well the corresponding vectors contain bounding box regression information. This can be concluded as a two-class classification issue which can be solved by the cross-entropy loss.

\[ L_{\text{det}} = -\left( y_{\text{det}} \log(p_i) + \left(1 - y_{\text{det}}\right) \log(1 - p_i) \right) \]  

\[ y_{\text{det}} \in \{0, 1\} \]  

In Eq(1), \( xi \) is the input image, \( pi \) is the probability represents \( xi \) being the face. Eq(2) indicates the label of ground-truth.

2.1.2. Refine Network (R-Net). This layer of network play the role of calibration based on bounding box regression and NMS, aiming at rejects major false rough facial windows. This objective can be summed as a regression problem, as well overcomed by Euclidean loss.

\[ L_{\text{box}} = \left\| \hat{y}_{\text{box}} - y_{\text{box}} \right\|^2 \]  

In Eq(3), \( \hat{y}_{\text{box}} \) and \( y_{\text{box}} \) represent the regression target calculated by the network and the corresponding real coordinate, respectively.

2.1.3. Output Network (O-Net). This stage proposes more supervisions to mark face region. Most important of all, this stage marks out five facial features’ coordinates. Facial features detection belongs to the regression issue through minimizing the defined Euclidean loss:

\[ L_{\text{landmark}} = \left\| \hat{y}_{\text{landmark}} - y_{\text{landmark}} \right\|^2 \]  

In Eq(4), \( \hat{y}_{\text{landmark}} \) and \( y_{\text{landmark}} \) are the coordinates of facial features correspond to trained network and real condition for the i-th input image, respectively. The facial features are consisted of five feature points, including left eye, right eye, nose, left mouth as well as right mouth.

2.2. SVM

SVM is a classical classification method applied in pattern recognition field. SVM maps the pixels’ data into the space consisted of higher dimensional which is contribute to constructing the optimal separating hyperplane, aiming at solving the quadratic programming and local minima issue. In our model, we propose SVM for classification issues of two classes, whether it is face region or not,
realizing multiple classification judgment. The combination of SVM and MTCNN can be utilized in complex conditions.

3. The proposed method

3.1. The Whole Procedure of IMT-SVM

Fig. 1 performs the procedure chart of the IMT-SVM algorithm. Firstly, the P-Net can get the rough facial windows in the input driver face image. Secondly, the accurate face region is labeled through R-Net. Thirdly, SVM model is utilized to judge whether it is driver face region. If not, the false sample will be deleted. Finally, the five facial features’ coordinates are labeled through O-Net.

![Figure 1. The Procedure of IMT-SVM algorithm](image)

3.2. Architecture of Proposed Convolutional network

In IMT-SVM model, we make improvement on the architectures of P-Net, R-Net as well O-Net, which share similar architecture (Fig. 2). The size of input image is 28*28. C1 and C2 are the first and second convolutional layer, respectively, which consist of six feature maps. They share the same convolutional kernel of 5*5.

![Figure 2. The Architecture of Proposed Convolutional network](image)

3.3. Manual Hard Sample Mining

We use the training set of our own traffic driver face database to train this model. The testing set is utilized to verify the performance. What should be noticed is that, even though the online hard sample
mining is conducted in MTCNN. However, in real application, especially in highway condition, as the result of complex environment, it is necessary to manual generate representative negative sample, strengthening the detector in training procedure. Experimental results indicate that this strategy performs better performance with adding manual hard sample, showing in Section IV.

4. Experimental Results

In this section, We use our own traffic driver face database, which is constructed by Public Security Department of Jiangsu Province, as dataset, containing approximately 1500 driver face images in different traffic conditions. We randomly choose 1000 images for training IMT-SVM model, remaining 500 images for verifying the model's performance. Double kernel CPU are used to train the networks. The performances of our method on face detection and alignment are shown below.

4.1. The Performances of Face Detection

Fig.3 and Fig.4 show the testing sample of IMT-SVM model under different environment. The detection and alignment results of IMT-SVM model in the dark environment and day time are performed. We can get the conclusion that the proposed model in this paper can realize good performance in complex environment on highway. Table 1 is the comparing table, indicating the detecting performances of IMT-SVM technique with MTCNN as well as Cascade CNN [14]. From the experimental result, the proposed technique in this paper shows high detection rate as well as low error detection rate compared to other state-of-the-art method [15]. This fully proves the IMT-SVM algorithm own excellent performance on improving the accuracy of drive face detection.

![Figure 3. The detection and alignment result of IMT-SVM model under the dark environment](image1)

(a) Original image  (b) Result picture  (c) Magnified labelled face

![Figure 4. The detection and alignment result of IMT-SVM model during the day time](image2)

(a) Original image  (b) Result picture  (c) Magnified labelled face
### Table 1. Detecting Performances Between our Method and Other Comparison Techniques

| Detection Method | IMT-SVM (OUR) | MTCNN [6] | Cascade CNN [19] |
|------------------|---------------|-----------|------------------|
| Detection Rate   | 83%           | 75.6%     | 68.9%            |
| Error Detection Rate | 1.35%         | 11.28%    | 18.7%            |

4.2. The Performances of Face Alignment

Then, we compare the alignment performances of our method with MTCNN as well as Cascade CNN. Average alignment errors on left eye (LE), right eye (RE), nose (N), left mouth (LM) as well as right mouth (RM) of the above methods are shown in Fig.5. For nose tip, which is the most difficult in detection, the average error of IMT-SVM method is approximately 2.3% lower than MTCNN and about 4.2% lower than Cascade CNN. When it comes to other face features, including LE, RE, LM and RM, IMT-SVM algorithm’s error rates still rank the lowest in the contrast methods. According to the result, we can prove that the IMT-SVM model shows outstanding performance on improving the precision of driver face alignment.

![Figure 5. The alignment results of our model and other comparison techniques](image)

5. Conclusions

This paper proposes a driver face detection and alignment model based on improved MTCNN and SVM. As face detection through convolutional network in MTCNN show unsatisfied error detection rate, it is desirable to use SVM for multiple classification judgment. In addition, improved MTCNN is proposed to realize precise facial point detection under complex highway environment. Experimental results show that the IMT-SVM model in this paper is effective, keeping robust to the occlusions, varying pose and extreme illumination on highway.

However, our work has some limitations under complex illumination environment and shadowed face images. Thus, in future work, denoising method [16][17] is a direction worth studying and making improvements[18][19].

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.61871123), Key Research and Development Program in Jiangsu Province (No.BE2016739) and a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.
References
[1] B. Yang, J. Yan, Z. Lei, and S. Z. Li, “Aggregate channel features formulti-view face detection,” in IEEE International Joint Conference on Biometrics, 2014, pp. 1-8.
[2] P. Viola and M. J. Jones, “Robust real-time face detection.International journal of computer vision,” vol. 57, no. 2, pp. 137-154, 2004
[3] M. T. Pham, Y. Gao, V. D. D. Hoang, and T. J. Cham, “Fast polygonalintegration and its application in extending haar-like features to improve object detection,” in IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 942-949.
[4] Q. Zhu, M. C. Yeh, K. T. Cheng, and S. Avidan, “Fast human detection using a cascade of histograms of oriented gradients,” in IEEE Computer Conference on Computer Vision and Pattern Recognition, 2006, pp. 1491-1498.
[5] S. Yang, P. Luo, C. C. Loy, and X. Tang, “From facial parts responses to face detection: A deep learning approach,” in IEEE International Conference on Computer Vision, 2015, pp. 3676-3684.
[6] Zhang K, Zhang Z, Li Z, et al. “Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks,” IEEE Signal Processing Letters, 2016, 23(10):1499-1503.
[7] Y. Sun, X. Wang, and X. Tang. Deep Convolutional Network Cascade for Facial Point Detection. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013
[8] B. Amberg, T. Vetter. Optimal landmark detection using shape models and branch and bound. In Proc. ICCV, 2011.
[9] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In Proc. CVPR, 2011.
[10] J. Zhang, S. Shan, M. Kan, and X. Chen, “Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment,” in European Conference on Computer Vision, 2014, pp. 1-16.
[11] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In Proc. CVPR, 2012.
[12] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In Proc. CVPR, 2012.
[13] Suykens J A K, Vandewalle J. “Least Squares Support Vector Machine Classifiers,” Neural Processing Letters, 1999, 9(3):293-300.
[14] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, “A convolutional neural network cascade for face detection,” in IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5325-5334.
[15] K. He, X. Zhang, S. Ren, J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in IEEE International Conference on Computer Vision, 2015, pp. 1026-1034.
[16] D. Shi, “Face image information processing and recognition,” Electronic Industry Press, 2010.
[17] Hu Changhui, Lu Xiaobo, Ye Mengjun, Zeng Weili. “Singular value decomposition and local near neighbors for face recognition under varying illumination,” Pattern Recognition, 2017, 64: 60-83.
[18] W. Zeng and X. Lu, “Region-based nonlocal means algorithm for noise removal,” Electronics Letters, vol. 47, pp. 1125-1127, 2011.
[19] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, “A convolutional neural network cascade for face detection,” in IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5325-5334.