Mathematical Biosciences and Engineering

MBE, 18(5): 6638–6651.
DOI: 10.3934/mbe.2021329
Received: 07 May 2021
Accepted: 26 July 2021
Published: 05 August 2021

http://www.aimspress.com/journal/MBE

Research article

Robust face recognition based on multi-task convolutional neural network

Huilin Ge, Yuewei Dai*, Zhiyu Zhu and Biao Wang

School of Electronic Information, Jiangsu University of Science and Technology, Zhenjiang 212003, China

* **Correspondence:** Email: dywjust@163.com; Tel: +8618914568855.

Abstract: *Purpose:* Due to the lack of prior knowledge of face images, large illumination changes, and complex backgrounds, the accuracy of face recognition is low. To address this issue, we propose a face detection and recognition algorithm based on multi-task convolutional neural network (MTCNN). *Methods:* In our paper, MTCNN mainly uses three cascaded networks, and adopts the idea of candidate box plus classifier to perform fast and efficient face recognition. The model is trained on a database of 50 faces we have collected, and Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measurement (SSIM), and receiver operating characteristic (ROC) curve are used to analyse MTCNN, Region-CNN (R-CNN) and Faster R-CNN. *Results:* The average PSNR of this technique is 1.24 dB higher than that of R-CNN and 0.94 dB higher than that of Faster R-CNN. The average SSIM value of MTCNN is 10.3% higher than R-CNN and 8.7% higher than Faster R-CNN. The Area Under Curve (AUC) of MTCNN is 97.56%, the AUC of R-CNN is 91.24%, and the AUC of Faster R-CNN is 92.01%. MTCNN has the best comprehensive performance in face recognition. For the face images with defective features, MTCNN still has the best effect. *Conclusions:* This algorithm can effectively improve face recognition to a certain extent. The accuracy rate and the reduction of the false detection rate of face detection can not only be better used in key places, ensure the safety of property and security of the people, improve safety, but also better reduce the waste of human resources and improve efficiency.

Keywords: multi-task CNN; image recognition; peak signal-to-noise ratio; structural similarity index measurement

1. Introduction

In recent years, artificial intelligence technique has advanced rapidly [1–3]. Biometric recognition [4–6], which includes face recognition, voice recognition, fingerprint recognition, iris recognition, eye pattern recognition, etc. will occupy a very important position in the field of

artificial intelligence in the future. At present, technologies such as smart cards based on radio frequency identification, second-generation ID cards, and user passwords are mostly used in identification, and biometrics will gradually occupy an important market share.

Due to the superiority of biometrics, it is widely used in bank payment, securities, transportation, e-commerce, airport subway entrance access control, attendance, and criminal investigation by public security and judicial departments [7–9]. Major enterprises, institutions, companies, and government agencies have established their own biometric-based access control systems and attendance systems to improve the informatization and intelligence of management, greatly improve management efficiency, and effectively liberate the labor force [10–13].

Face recognition technology started relatively late, but the technology related to face recognition [37] has developed rapidly, and has achieved remarkable results in recognition accuracy, etc., and related technical achievements have attracted worldwide attention. Due to the lack of prior knowledge of face images, large illumination changes, complex backgrounds, and variable face angles, the demand for face images is large, expression changes are large, and face occlusion leads to low accuracy of face recognition.

Deep learning applications often use convolutional neural networks to achieve image processing and recognition with high efficiency and accuracy [14–17]. Facial images are highly structured images. Combining with prior facial knowledge is a very popular method in face recognition.

Dong et al. proposed an image super-resolution method based on deep convolutional neural network [18], which realizes the mapping from the low-resolution end to the high-resolution end of the image, and extends the traditional super-resolution method based on coding coefficients. Since then, the research of neural network combined with image super-resolution has continued to deepen. Kim et al. [19] proposed the use of very deep convolutional neural network on the basis of VGG network to improve the super-resolution accuracy by using multiple filters in the neural network, and realize the use of image context information.

Yu et al. proposed a transforming and distinguishing neural network [20] for the serious problems of multi-posture and degradation, and solving the problem of multi-posture and image misalignment. As the depth of the network increases, the features gradually disappear during the transmission process. In response to this problem, the multi-scale residual network image super-resolution (MSRN) algorithm [21] uses the combination of local multi-scale features and global features to maximize utilizing the features of low-resolution images; the problem of feature disappearance during transmission is solved. There are various existing algorithms and deep learning models [38–40] in which not only can they facilitate face recognition, but also perform human activity recognition and motion prediction.

In this paper, we propose a face detection and recognition model based on multi-task convolutional neural network (MTCNN) [22–24]. The recognition model combined with deep learning has a high accuracy rate, and has a shorter recognition time, which can reduce the waste of human resources.

2. Methodology

2.1. CNN

2.1.1. MTCNN

MTCNN implements the face area detection and face key point detection together, and its subject

framework is similar to cascade. The whole can be divided into a three-layer network structure of Proposal Network (P-Net), Refine Network (R-Net), and Output Network (O-Net) [25,26]. It is a multi-task neural network model for face detection tasks which mainly uses three cascaded networks and the idea of candidate boxes plus classifiers.

The three cascaded networks are P-Net for quickly generating candidate windows, R-Net for filtering and selecting high-precision candidate windows, and O-Net for generating final bounding boxes and face key points. And many convolutional neural network models that deal with image problems; the model also uses image pyramids, border regression, non-maximum suppression and other technologies. The network structure of P-Net, R-Net and O-Net is shown in Figure 1.

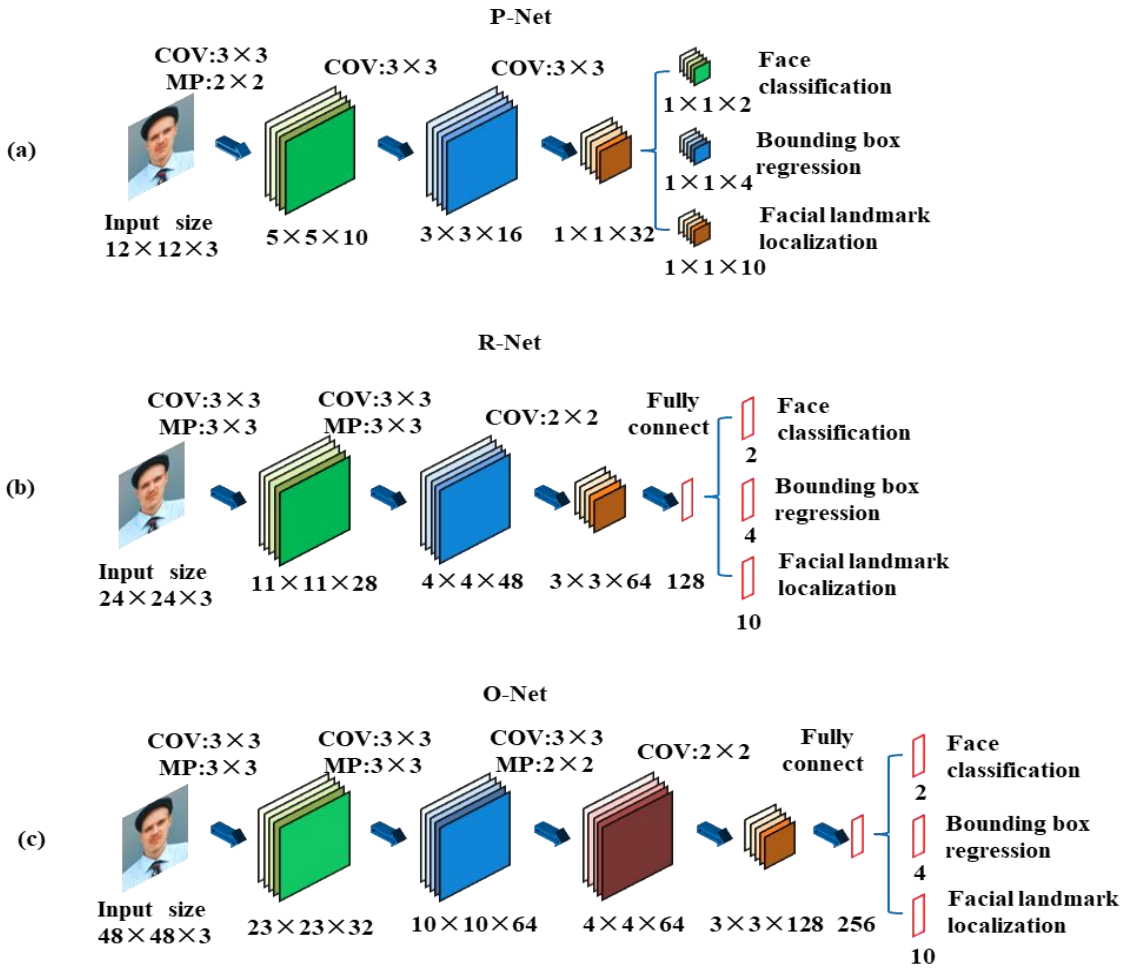

Figure 1. Network structure of P-Net, R-Net and O-Net (a) P-Net; (b) R-Net; (c) O-Net.

To balance computational expenses and performance, MTCNN avoids the huge performance consumption caused by traditional ideas such as sliding windows and classifiers, we first use a small model to generate a candidate frame for the target area with a certain possibility, and then use a more complex model for fine classification. Then we return the higher-precision area box, and let this step be executed recursively. The network structure of MTCNN is presented by Figure 2.

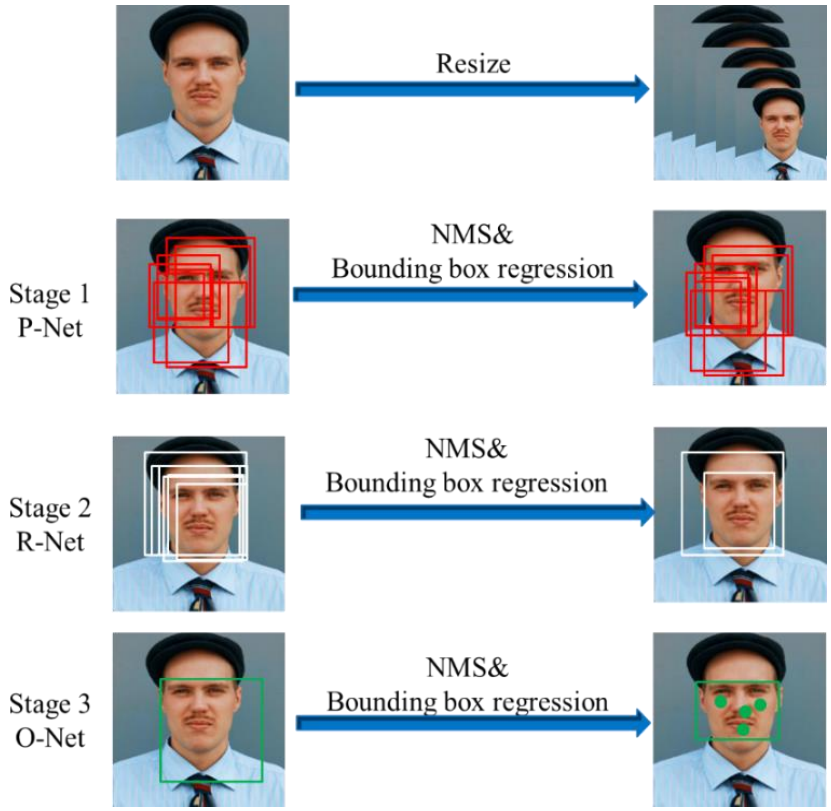

Figure 2. Our network pertaining to MTCNN.

2.1.2. R-CNN

R-CNN [27] draws on the idea of sliding window and adopts the scheme of area recognition. The specific identification scheme is the first step. Given an input image, extract 2000 independent candidate regions from the image; the second step is to use CNN to extract a fixed-length feature vector for each region; the third step, we use SVM to classify each area. Figure 3 shows the process of R-CNN in the recognition of faces.

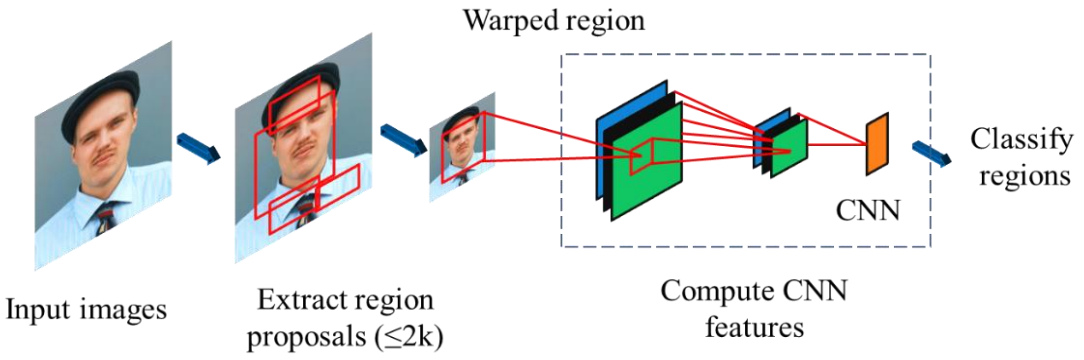

Figure 3. The process of R-CNN in the recognition of faces.

2.1.3. Faster R-CNN

Faster R-CNN [28,29] creatively uses the convolutional network to generate the suggestion frame by itself, and shares the convolutional network with the target detection network, so that the number of suggestion frames is reduced from about 2000 to 300. This framework can even be combined with nonlinear anisotropic diffusion filtering and other morphological methods to perform denoising of image. To ensure excellent face recognition, the noise and the other unrelated background areas can be removed. The process of Faster R-CNN to recognize faces is shown in Figure 4.

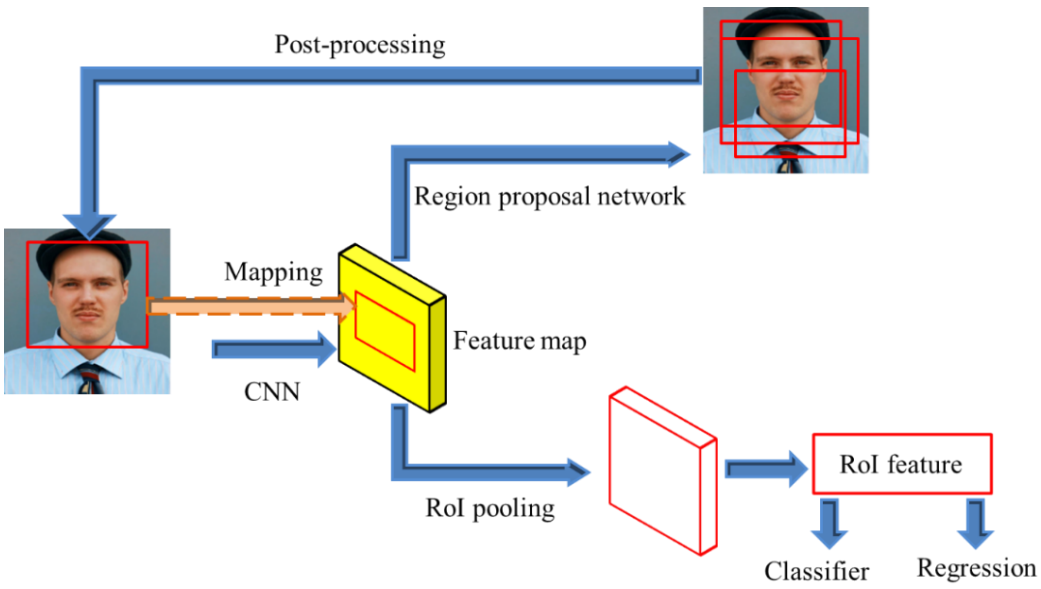

Figure 4. The process of Faster R-CNN in the recognition of faces.

2.2. Mapping function

The commonly used kernel functions [30] include linear kernel functions [31,32], Gaussian kernel functions [33,34], and other similar algorithms. One of the most commonly used is the Gaussian kernel function, which can map feature data to infinite dimensions.

2.2.1 Linear kernel function

The linear kernel function is mainly used in the case of linear separability, which can achieve a good classification effect. The mathematical expression of the linear kernel function is shown in Eq (1). There are few parameters in the function, so the computation rate is very fast, and the dimension of its input space is the same as the dimension of the feature space, which is suitable for the first attempt in the classification task.

$$K(X,Y) = (X^T Y + C) \tag{1}$$

In Eq (1), X and Y represent eigenvectors, and C represents a constant.

2.2.2 Gaussian kernel function

The Gaussian kernel function can map the sample from the input space to the higher-dimensional feature space, and it can achieve good results regardless of the sample size. When the classification task cannot determine which kernel function to use, the Gaussian kernel function is the most widely used one of the kernel functions. The mathematical expression of the Gaussian kernel function is shown in Eq (2).

$$K(X,Y) = exp\left(-\frac{\|X-Y\|^2}{2\sigma^2}\right) \tag{2}$$

In particular, σ in Eq (2) is the width parameter of the function, which controls the radial range of the function.

2.3. Database

We collect 2500 face images for our face recognition research in this paper. During the test, the data set is divided into the original image data set and the test data set according to the parity position. The two images at the corresponding positions of the two data sets are the comparison objects, as shown in Figure 5. Then, we apply the algorithm of this paper to these two data sets, and extract the images out in turn. Table 1 show the gender ratio and picture size of the two data sets.

Table 1. Gender ratio and image size of the two data sets.

Characteristics	Original image data set	Test data set
Male: Female	10: 15	13: 12
Image size	1897×1897	1897×1897

2.4. Evaluation Index

The high-resolution face image reconstructed after the algorithm processing requires a certain measurement standard to examine the performance of the algorithm. The earlier evaluation standard was subjective evaluation through naked eye observation. This method is simple and direct. The objective evaluation calculates the similarity between the synthesized image and the original image, and has a specific value to measure the reconstruction result of the image. Compared with the subjective evaluation, its advantage is that the comparison result is more concise and accurate. At present, the commonly used objective evaluation methods include Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measurement (SSIM) [35,36].

2.4.1 PSNR

PSNR is an image quality evaluation method based on the error between corresponding pixels. It is the most common image quality evaluation index, and its expression is as Eq (3).

$$PSNR = 10\log_{10}\left(\frac{(2^b-1)^2}{MSE}\right) \tag{3}$$

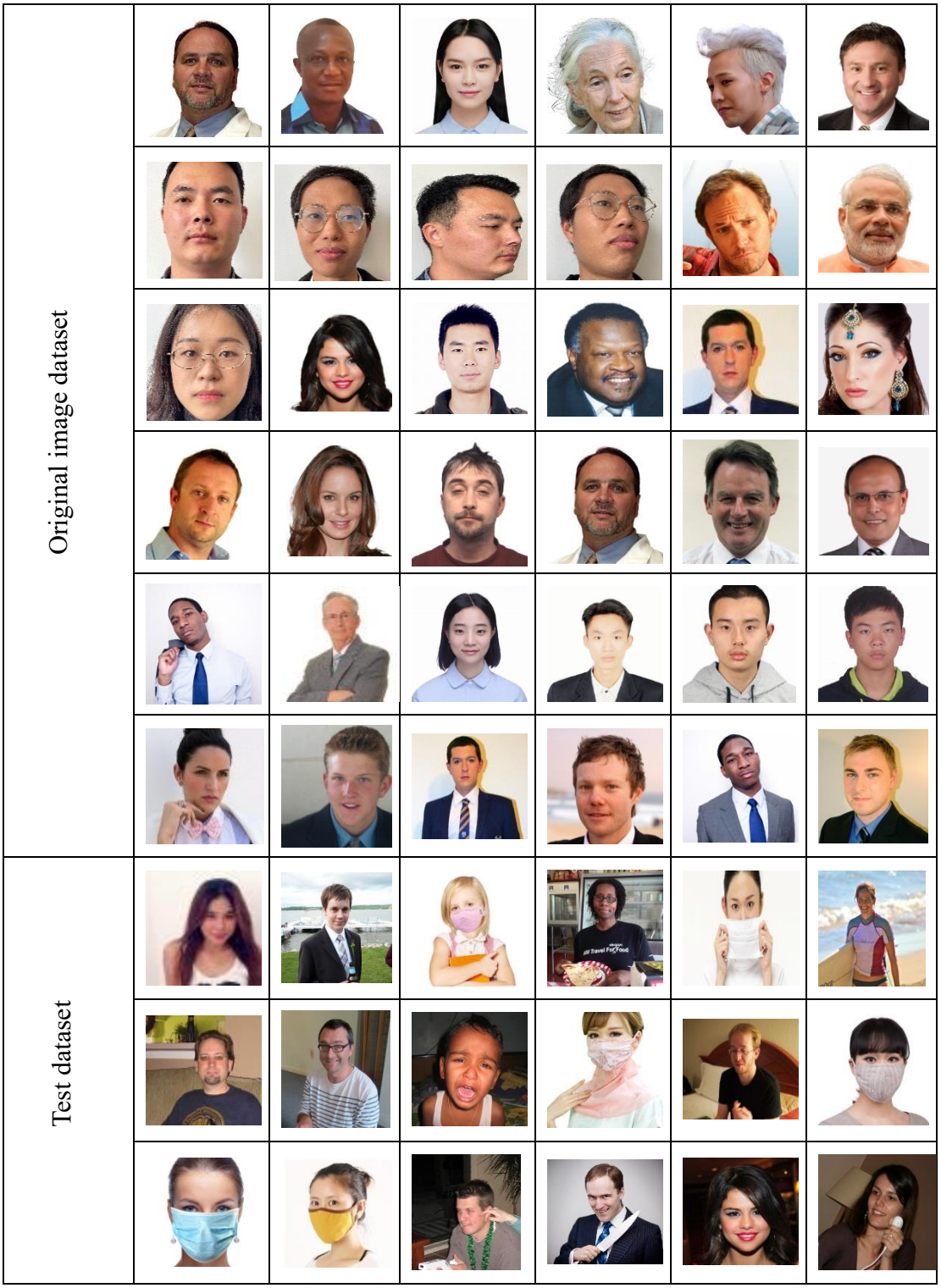

Figure 5. Sample images of the original image data set and the test data set.

In Eq (3), b is the number of bits of the pixel, which is usually 8. The unit of PSNR is decibel (dB). The larger the value means higher quality of the reconstructed image. Mean Square Error (MSE) represents the mean square error between images. The symbol h represents the height of the image, and w represents the width of the image. The expression is shown in Eq (4).

$$MSE = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} [X_1(i,j) - X_2(i,j)]^2 \quad (4)$$

However, because the human perception of an area will be affected by the surrounding area and other reasons, this method does not take into account the human visual characteristics, but only calculates the difference between pixels, so the result of the peak signal-to-noise ratio evaluation method will appear to be different from that of humans.

2.4.2 SSIM

The value of SSIM is a value between 0 and 1. The larger the value of SSIM, explant the smaller the difference between both photos of the faces. Therefore, the larger the value of SSIM replies the better the image reconstruction quality. The SSIM method measures image similarity from three aspects: brightness (L), contrast (C), and structure (S). The expression of SSIM is shown in Eq (5).

$$SSIM(X,Y) = L(X,Y) \times C(X,Y) \times S(X,Y) \quad (5)$$

SSIM simulates the human perception of changes in image information, uses the image average to model the image brightness, the image standard deviation to model the image contrast, and the image covariance to model the image structure, which makes up for the shortcomings of the PSNR method.

2.4.3 ROC

Receiver operating characteristic (ROC) curve is a comprehensive indicator that reflects the sensitivity and specificity of continuous variables. Each point on the ROC curve reflects the susceptibility to the same signal stimulus.

The abscissa is false positive rate (FPR), that is, the proportion of all negative samples that are predicted to be positive but actually negative. The larger the FPR, the positive prediction the more negative classes in the class. The computational FPR is as Eq (6).

$$FPR = \frac{FP}{TN+FN} \quad (6)$$

The ordinate is true positive rate (TPR), the proportion of all positive samples that are predicted to be positive and actually positive. The larger the TPR stands the more actual positive classes in the predicted positive class. The computational TPR is as Eq (7).

$$\text{TPR} = \frac{TP}{TP+FP} \quad (7)$$

In Eqs (6) and (7), True Positive (TP), the prediction is a positive sample and the actual number of features is also a positive sample. False Positive (FP), the number of features predicted to be a positive sample and actually a negative sample. True Negative (TN) is predicted to be a negative sample and is actually the number of features of a negative sample. False Negative (FN), the number of features predicted to be negative samples and actually positive samples.

3. Results

3.1. Optimizing parameters

In order to better recognize the face image, this paper implements the Gaussian kernel function as the mapping function in the model. In order to achieve the best results, we need to adjust three parameters, which are the local constraint parameter λ, the kernel function similarity parameter σ, and the high resolution layer constraint parameter k.

3.1.1 Local constraint term parameter λ

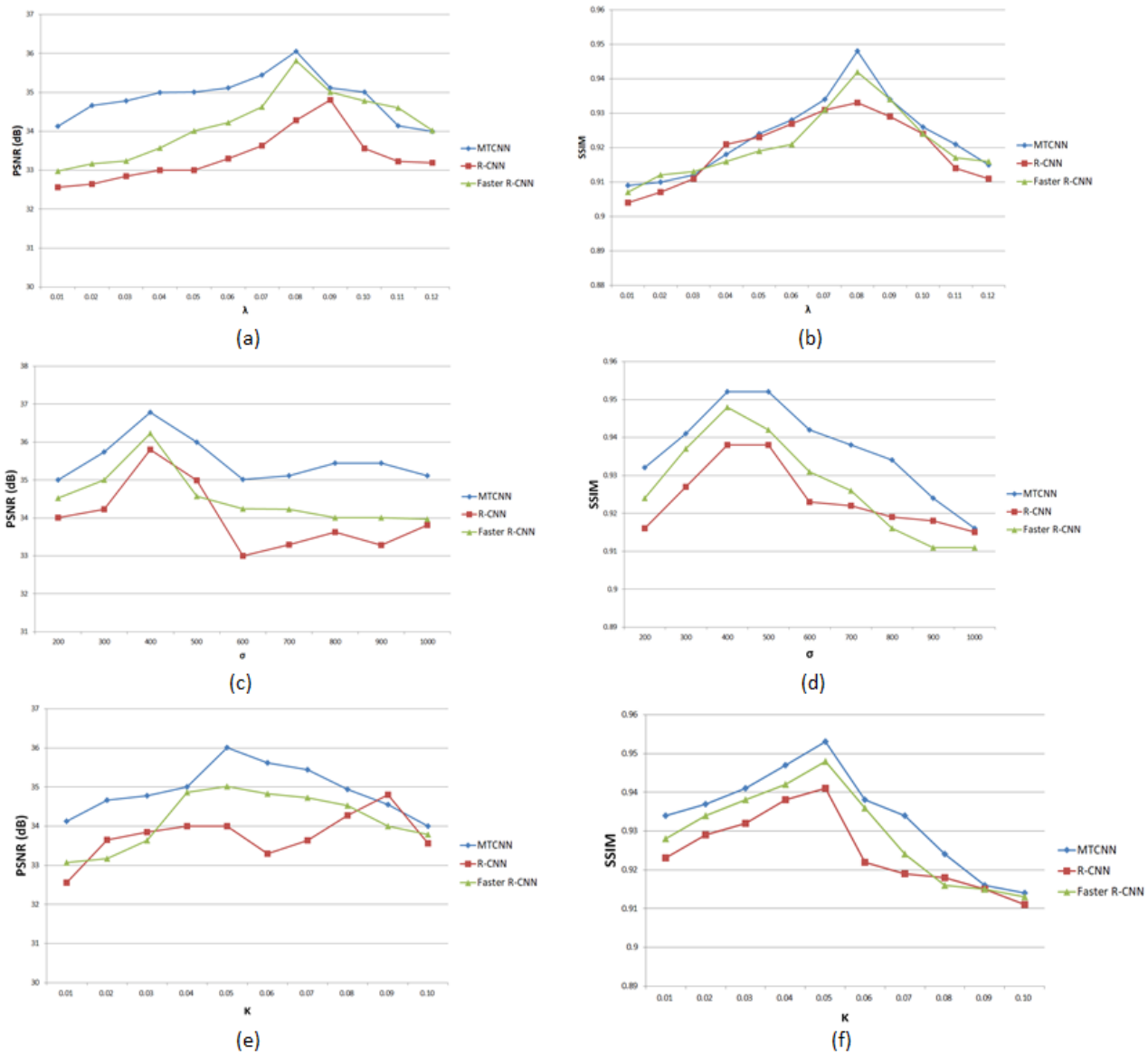

Figure 6. Parameter optimization of λ, σ and k in the face recognition framework based on MTCNN, R-XNN, and Faster R-CNN.

We fix the value of k and the kernel function similarity parameter σ, and set the value range of λ to be 0.01 to 0.12. The mean value of PSNR and mean value of SSIM as the evaluation criteria change with the value of λ as shown in Figure 6(a),(b). As shown in the figure, when λ = 0.08, the average PSNR of MTCNN and Faster R-CNN is the largest, and when λ = 0.08, the average PSNR of R-CNN is the largest. The average value of SSIM reaches its peak. Since the difference between the two values is very small, the value of λ is selected here as 0.08.

3.1.2 Kernel function similarity parameter σ

We fix the values of k and λ, and set the value of the kernel function similarity parameter σ from 200 to 1000. The average value of PSNR and the average value of SSIM as the evaluation criteria change with the value of σ as shown in Figure 6(c),(d). We find that when the σ value is 400, the average PSNR of the three models reaches the maximum, while the average SSIM reaches the peak at σ = 400 and 500. Combining the two objective evaluation criteria, we choose σ = 400.

3.1.3 Constraint parameter k of the high-resolution layer

We set the values of λ and the kernel function similarity parameter σ, and set the value of the high-resolution layer error term parameter k to 0.01 to 0.10. As the evaluation criteria, the mean value of PSNR and mean value of SSIM change with the value of k as shown in Figure 6(e),(f). It can be clearly seen from the following line chart that both the mean PSNR and the mean SSIM peak when k takes 0.05, so we determine the optimal value of k to be 0.05.

3.2. PSNR and SSIM mean results

From the optimization parameters, we know that when our parameters are set to λ as 0.08, σ as 400 and k as 0.05, the three models are in the optimal state. Table 2 shows the average values of PSNR and SSIM of the three models.

Table 2. Results of mean PSNR and mean SSIM of the three models.

Model	PSNR (dB)	SSIM
MTCNN	36.245	0.954
R-CNN	35.005	0.927
Faster R-CNN	35.305	0.938

From Table 2, the mean PSNR and SSIM pertaining to the MTCNN are better than R-CNN and Faster R-CNN. The average PSNR of our method is 1.24 dB higher than that of R-CNN and 0.94 dB higher than that of Faster R-CNN. The average SSIM of MTCNN is 10.3% higher than R-CNN and 8.7% higher than Faster R-CNN.

3.3. ROC curve

Figure 7 shows the ROC curves of MTCNN, R-CNN and Faster R-CNN. The Area Under Curve (AUC) of MTCNN is 97.56%, the AUC of R-CNN is 91.24%, and the AUC of Faster R-

CNN is 92.01%. MTCNN has the best overall face recognition performance. For detection of faces, MTCNN still has the best effect.

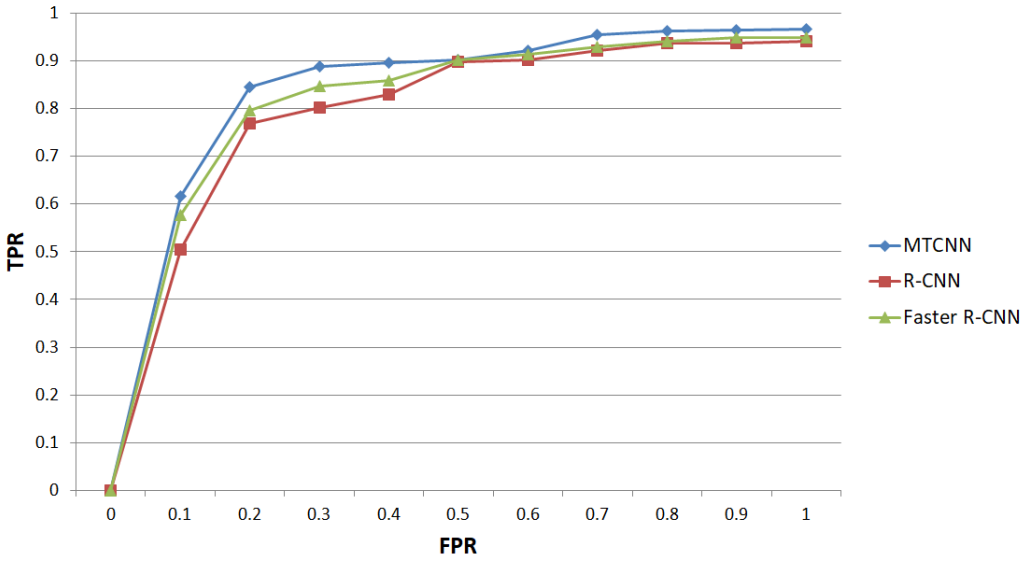

Figure 7. ROC curves of MTCNN, R-CNN and Faster R-CNN.

4. Discussion

In actual scenes, the acquired face images are usually of poor resolution and low-quality, which is caused by a variety of reasons: first, the location of the surveillance camera is high, the shooting range is large, and the target face image is small; second, the monitoring The device is limited by storage space and highly compresses video images, so the image loses detailed information; third, external environments such as rainy weather and poor lighting will further reduce the quality of the captured images. In response to these problems, the face recognition technology combined with convolutional neural network realizes its practical application value.

Although this paper has conducted considerable research worthy of investigation, and performed exploration on learning-based image classification methods, there are still significant limitations that need to be addressed. Despite the sound conclusion of this research in this field, there are still some issues worthy of attention and future implementation of other relevant networks is required for a thorough analysis.

With the continuous deepening of research on convolutional neural networks, vector-based convolution and pooling processing have been fully studied. In fact, in a network, we can apply the Riemannian manifold geometry to the data in the middle layer for processing. This pooling and iterative process in the form of a matrix can have a positive effect on the final output of the network.

5. Conclusions

MTCNN mainly uses three cascaded networks, and uses the idea of candidate box plus classifier to execute fast and efficient face recognition. Among the three evaluation indicators, MTCNN has the best overall face recognition performance, and for defective faces, MTCNN still has the best effect as

well as performance.

The technology based on MTCNN has a good development prospect, which greatly improves the accuracy of face recognition. While improving accuracy, it also improves the security of the image recognition system.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgment

This research is funded by the National Natural Science Foundation of China (No. 62006102).

References

1. C Fernández, M. A. Vicente, M. O. Martínez-Rach, Implementation of a face recognition system as experimental practices in an artificial intelligence and pattern recognition course, *Comput. Appl. Eng. Educ.*, **12** (2020), 48–50.
2. J. X. Zeng, P. Chen, J. Q. Tian, X. Fu, Fuzzy kernel two-dimensional principal component analysis for face recognition, *AICA*, 2015.
3. D. C. Wise, Face recognition under expressions and lighting variations using artificial intelligence and image synthesizing, *JCER*, **2** (2012), 186–190.
4. H. Chakraborty, V. Balasubramanian, S. Panchanathan, Generalized batch mode active learning for face-based biometric recognition, *Pattern Recognit.*, **23** (2013), 134–140.
5. P. M. Shende, M. V. Sarode, M. M. Ghonge, A survey based on fingerprint, face and iris biometric recognition system, image quality assessment and fake biometric, *Int. J. Comput. Sci. Eng.*, **25** (2014), 221–225.
6. D. Sharma, A. Kumar, An empirical analysis over the four different feature-based face and iris biometric recognition tech, *Int. Jrnl. Adv. Comput. Sci. Appl.*, **3** (2012), 1–13.
7. F. W. Wheeler, X. Liu, P. H. Tu, R. T. Hoctor, Multi-frame image restoration for face recognition, in *IEEE Workshop on Signal Processing Applications for Public Security & Forensics*, 2007.
8. Z. Liu, H. Zhang, S. Wang, W. Hong, J. Ma, Y. He, Reliability evaluation of public security face recognition system based on continuous bayesian network, *Math. Probl. Eng.*, **3** (2020), 1–9.
9. D. Li, X. Zhang, L. S. Yi, X. Zhao, Multiple-step model training for face recognition, *Int. Conf. Appl. Tech. Cyber. Sec. Int.*, 2017.
10. B. S. Satari, N. Rahman, Z. Abidin, Face recognition for security efficiency in managing and monitoring visitors of an organization, in *2014 ISBAST*, 2014.
11. Q. Liang, W. Fang, College student attendance system based on face recognition, *Iop. Conf.*, 2018.
12. S. Kasaei, S. A. Monadjemi, K. Jamshidi, Application of the face recognition procedure in citizenship and immigration management system, *Int. J. Electron. Comput. Eng. Syst.*, (2014), 611–614.
13. Z. Zhang, Design plan of dormitory name management system based on face recognition, *China Commun.*, **12** (2018), 220–230.
14. K. K. L. Wong, G. Fortino, D. Abbott, Deep learning-based cardiovascular image diagnosis: A promising challenge, *Future Gener Comput. Syst.*, **110** (2020), 802–811.

15. W. Wang, J. Yang, J. Xiao, S. Li, D. Zhou, Face recognition based on deep learning, *Int. Conf. Pervas. Comput. Appl.*, 2014.
16. K. Grm, V. Štruc, A. Artiges, M. Caron, H. K. Ekenel., Strengths and weaknesses of deep learning models for face recognition against image degradations, *Iet. Biom.*, **7** (2018), 81–89.
17. J. Zhao, Y. Lv, Z. Zhou, F. Cao, A novel deep learning algorithm for incomplete face recognition: Low-rank-recovery network, *Neural Netw.*, **13**(2017), 94.
18. C. Dong, C. C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, *IEEE Trans Pattern Anal. Mach. Int.*, **38** (2016), 295–307.
19. J. Kim, J. K. Lee, K. M. Lee, Accurate image super-resolution using very deep convolutional networks, *IEEE Conf. Comp. Vis. Pat. Recognit. IEEE*, 2016.
20. X. Yu, F. Porikli, Face hallucination with tiny unaligned images by transformative discriminative neural networks, *31st AAAI Conf. Art. Int.*, 2017.
21. J. Li, F. Fang, K. Mei, G. Zhang, Multi-scale residual network for image super-resolution, *ECCV*, 2018.
22. J. Xiang, G, Zhu, Joint face detection and facial expression recognition with Mtcnn, *Int. Conf. Infor. Sci. Control Eng. IEEE Comp. Soc.*, (2017), 424–427.
23. E. Jose, M. Greeshma, M. T. Haridas, M. H. Supriya, Face recognition based surveillance system using faceNet and Mtcnn on jetson TX2, in *2019 5th ICACCS, IEEE*, 2019.
24. C. Wu, Y. Zhang, Mtcnn and facenet based access control system for face detection and recognition, *Autom. Control. Comput. Sci.*, **55** (2021), 102–112.
25. L. H. Ma, H. Y. Fan, Z. M. Lu, D. Tian, Acceleration of multi-task cascaded convolutional networks, *IET Image Process.*, **14** (2020), 2435–2441.
26. X. Zhao, S. Lin, X. Chen, C. Ou, C. Liao, Application of face image detection based on deep learning in privacy security of intelligent cloud platform, *Multimed. Tools Appl.*, **79** (2020), 205–210.
27. H. Chen, Y. Chen, X. Tian, R. Jiang, A cascade face spoofing detector based on face anti-spoofing R-CNN and improved retinex LBP, *IEEE Access*, **7** (2019), 170116–170133.
28. X. Qin, Y. Zhou, Z. He, Y. Wang, Z. Tang, A faster R-CNN based method for comic characters face detection, in *2017 14th ICDAR IEEE Computer Society*, 2017.
29. L. J. Halawa, A. Wibowo, F. Ernawan, Face recognition using faster R-CNN with inception-V2 architecture for CCTV camera, in *2019 3rd ICICoS*, 2019.
30. T. Hofmann, B. Schölkopf, A. J. Smola, Kernel methods in machine learning, *Ann. Stat.*, (2008), 1171–1220.
31. Y. Q. Bai, M. E. Ghami, C. Roos, A comparative study of kernel functions for primal-dual interior-point algorithms in linear optimization, *SIAM J. Optim.*, **15** (2014), 101–128.
32. A. R. Webb, An approach to non-linear principal components analysis using radially symmetric kernel functions, *Stat. Comput.*, **6** (1996), 159–168.
33. E. H. Lieb, Gaussian kernels have only gaussian maximizers, *Inv. Math.*, **102** (1990), 9–208.
34. L. Jiang, B. Zhu, X. Rao, G. Berney, Y. Tao, Discrimination of black walnut shell and pulp in hyperspectral fluorescence imagery using Gaussian kernel function approach, *J. Food Eng.*, **81** (2007), 108–117.
35. D. S. Turaga, Y. Chen, J. Caviedes, No reference PSNR estimation for compressed pictures, *Signal Process. Image Commun.*, **19** (2004), 173–184.
36. H. Alain, D. Ziou, Image quality metrics: PSNR vs. SSIM, *20th Int. Conf. Pat. Recognit. IEEE*, 2010.

37. Z. Tang, G. Zhao, T. Ouyang, Two-phase deep learning model for short-term wind direction forecasting, *Renew. Energy*, **173** (2021), 1005–1016.
38. K. Lan, S. Fong, L. S. Liu, R. K. Wong, N. Dey, R. C. Millham, et al, A clustering based variable sub-window approach using particle swarm optimisation for biomedical sensor data monitoring, *Enterp. Inf. Syst.*, **15** (2021), 15–35.
39. T. Li, S. Fong, K. K. L. Wong, Y. Wu, X. Yang, X. Li, Fusing wearable and remote sensing data streams by fast incremental learning with swarm decision table for human activity recognition, *Inf. Fusion*, **60** (2020), 41–64.

AIMS Press