Received: 21 April 2021 | Revised: 24 August 2021 | Accepted: 4 September 2021

IET Image Processing

DOI: 10.1049/ipr2.12344

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

Heterogeneous face detection based on multi-task cascaded convolutional neural network

XianBen Yang | **Wei Zhang**

College of Computer Science and Technology, Beihua University, Jilin City, China

Correspondence
Wei Zhang, College of Computer Science and Technology, Beihua University, Jilin City 13200, China.
Email: yangxianben@stu.cpu.edu.cn

Abstract

Facial target detection is an important task in computer vision. Because heterogeneous face detection shows broad prospects, it has attracted extensive attention from the academic community. In recent years, with the rise of deep learning and its applications in computer vision, face detection technology has made great strides. This paper uses multi-task cascaded convolutional neural network (MTCNN) for heterogeneous face feature detection. This algorithm makes full use of the advantages of image pyramid, boundary regression, fully convolutional attention networks and non-maximum suppression. The main idea of this paper is to use candidate frame plus classifier for fast and efficient face detection. Specifically, the candidate window is generated by the proposal network (P-Net), and the high-precision candidate window is filtered and selected by the reduced network (R-Net), and the final bounding box and facial key points are generated by the output network (O-Net). In order to prove the effectiveness of this method in visible light, near-infrared and sketch face recognition scenes, it was verified in the datasets of CUFS, CUFSF and CASIA NIR-VIS 2.0. Experiments show that this method is effective for face images in heterogeneous face and is better than the latest algorithms.

1 | INTRODUCTION

Artificial intelligence (AI) technology is widely studied, which not only exists in news reports, but also appears in our daily lives, for example, AI-based biometrics technology is applied for public safety. It involves a large number of applications of biometrics, access control systems, and social media tagging to character retrieval in multimedia. The AI-based biometric technology is an important application of artificial intelligence technology, which receives lots of research interests. The research of biometric technology is an important part of artificial intelligence. In biometric dataset, everyone has several modalities such as near-infrared face images, portrait sketches, thermal infrared face images and three-dimensional face model images. Face detection as an emerging biometric recognition application has become one of the most active research topics in the field of pattern recognition and machine learning. Compared with other biometric recognition technologies such as fingerprints and iris, face detection is more and more popular in the field of computer vision and modal recognition due to its natural, non-contact, non-mandatory, and concurrent characteristics. The facial images with different modalities are also called heterogeneous facial images. Heterogeneous facial images use different imaging mechanisms to make the facial images of the same person exist in different ways, and the performance difference between them is huge, which greatly reduces the recognition accuracy of traditional face detection facial data sets. Specifically, the research of heterogeneous face detection algorithms faces many different severe challenges. Taking the near-infrared image and the RGB face image as examples, the image information between the two images is very different, which will seriously affect the detection between the two images. The coordinate system between heterogeneous facial images is different. Take the two-dimensional face image and the three-dimensional structure face model as examples. The coordinate systems of the two are different, and they cannot be directly compared and recognized. And when the low-resolution face is extracted, when the image and the high-resolution heterogeneous face image

have obvious detail information, there is a problem of unequal image information due to the lack of information. It greatly affects the recognition accuracy of such images, but such images have application scenarios in security and medical identification. Therefore, accurate recognition of such images is of great help in real life.

Now, public safety systems can effectively detect the face area in captured image by employing image analysis technology to eliminate the interference that have nothing to do with face information. However, traditional face detection technology still has certain defects in the face detection of heterogeneous images, such as near-infrared images, thermal infrared images, low-resolution images etc., because traditional face detection methods are limited to the visible light modalities for face images. Therefore, for the heterogeneous face images composed of different imaging mechanisms, it is impossible for the traditional methods to effectively detect the face area in the heterogeneous image, which eventually leads to considerable performance gap among the different modalities face images of the same person. In order to solve this problem, there is an urgent need to conduct research on face detection technology on the heterogeneous face images.

This paper proposes a multi-task cascaded convolutional neural network [2, 9, 12, 13, 15, 16], which aims to effectively detect the faces in heterogeneous images, while ensuring high-precision performance in the invisible light mode and reducing the time consumption. At the same time, the internal connections between the position of the face landmark and bounding box are also considered. The main idea of this paper is to use multi-task cascaded convolutional neural network to effectively eliminate the modal difference between heterogeneous face images, thereby improving the recognition accuracy of different modal faces.

The rest of this article is organized as follows. Section 2 introduces the current mainstream face detection methods, and then describes the related theoretical techniques in detail. Section 3 discusses the multi-task cascaded convolutional neural network. Section 4 illustrates the experimental procedure for evaluating the performance of the algorithm. Finally, in Section 5, we give our conclusions and directions for future improvements.

2 | RELATED WORKS

2.1 | Face detection

Face detection is to detect the face and locate its region in the input image. The most popular face detection method is the Viola-Jones algorithm proposed by Viola and Jones [1]. It is based on the Harr-Like function and the AdaBoost algorithm [10] to perform real-time and high-performance face detection on the original image. Among them, the idea of cascading AdaBoost classifiers for target detection is to use multiple AdaBoost classifiers. Cooperate to complete the classification of candidate frames. These classifiers form a pipeline to determine the candidate frame image in the sliding window and determine whether it is a human face or a non-human face. Although this method can quickly exclude a large number of windows that are not human faces, some images that are not human faces may also be judged as human faces. Because the essence of this idea is to use a simple strong classifier in the early window of the face to quickly eliminate a large number of non-human images, while ensuring a high recall rate, so that the number of samples at all levels can eventually be strong, and the classifier is small .The basis for this is that most of the images to be detected are not human faces but backgrounds, that is, human faces are a sparse event. In practical applications, there will be multiple targets in the image, which will reduce the accuracy performance [10, 11, 14, 15, 16]. In addition, there is also a commonly used deformable part model algorithm (DPM) [8]. The DPM method uses FHOG for feature extraction, which can artificially design activation features and regard the face as a whole formed by the organic combination of multiple parts. Since the face has multiple parts, such as eyes, ears, mouth, nose, etc., and each part also has different characteristics. It has a very good detection effect for distorted, gender, multi-posture, multi-angle face, etc. As a result, the workload will increase significantly, causing the model to be too large, and the complex calculations in the judgment process are difficult to meet real-time requirements. Convolutional neural networks were soon used in face detection problems after their success in image classification problems. They greatly surpassed the previous AdaBoost framework in accuracy. There are currently some high-precision and efficient algorithms. The scheme of directly using sliding window and convolutional network to classify window images is too computationally expensive to achieve real-time. The method of using convolutional network for face detection uses various methods to solve or avoid this problem. Recently, convolutional neural networks (CNNs) achieve remarkable progresses in a variety of computer vision tasks [25, 26, 27]. Taking image classification [17, 23] and face recognition [18] as examples, some studies have carried out face detection experiments based on convolutional neural networks. Yang et al. performed facial attribute recognition based on convolutional neural networks [2], and Li et al. conducted face detection based on cascaded CNN [3]. Mukherjee et al.[29] have discussed the formulation for both the methods, i.e., using hand-crafted features followed by training a simple classifier and an entirely modern approach of learning features from data using neural networks. Ren et al. [30] have presented a method for real time detection and tracking of the human face. The proposed method combines the Convolution Neural Network detection and the Kalman filter tracking. Convolution Neural Network is used to detect the face in the video, which is more accurate than traditional detection method. When the face is largely deflected or severely occluded, Kalman filter tracking is utilized to predict the face position. They try to increase the face detection rate, while meeting the real time requirements. Luo et al. [31] have suggested deep cascaded detection method that iteratively exploits bounding-box regression, a localization technique, to approach the detection of potential faces in images. Although these methods have achieved exciting results in the detection of visible light modal face images, they have defects in processing heterogeneous face

images, or computation exhausting, or ignoring the inherent relationship between the position of the face landmark, and finally resulting in out of practical application. Therefore, the existing CNN-based methods are not suitable for face detection in heterogeneous images. Moreover, Ricardo F proposed a template matching method for face detection in heterogeneous images [4, 19, 22, 24]. Although this method can achieve excellent recognition rate, but it is calculation exhausted. In this context, the proposed face detection method can achieve high-precision performance and reduce time consumption in the invisible light mode.

2.2 | Structure similarity

Structural similarity (SSIM) is an index used to measure the similarity of pictures, and it can also be used to judge the quality of pictures after compression. Since the structural information of the image is the structural feature of the object in the field of view, it is independent of the brightness and contrast of the image. Therefore, SSIM tests the similarity of two images through brightness, contrast and structure. Compared with traditional image quality measurement indicators, structural similarity is more in line with the human eye's judgment on image quality in the measurement of image quality. Therefore, we use SSIM, which can reflect the difficulty of adapting to heterogeneous images due to the limited spectrum of the visible spectrum-based method.

2.3 | Image pyramid

Image pyramid, mainly used for image segmentation, is a kind of multi-scale expression in an image. It is an effective but simple-concept structure that interprets images with multiple resolutions. The pyramids of an image are a series of images arranged in a pyramid shape with gradually reduced resolution and derived from the same original image. It is obtained by down-sampling in steps, and the sampling is stopped until a certain termination condition is reached. The bottom of the pyramid is a high-resolution representation of the image, while the top is a low-resolution approximation. That is why the layers of images are considered as pyramids. The higher the level, the smaller the image and the lower the resolution. The application of image pyramid in face detection is to construct an image pyramid by transforming the original image at different scales to adapt to the detection of faces of different sizes. In order to improve the adaptability of face detection algorithm, we use image pyramid, which can help us to detect the smallest detected face.

2.4 | Parametric rectified linear unit

PRelu is a special type of Relu that equipped with parameters. Compared with Relu's method of filtering negative values [28], PRule adds parameters to negative values instead of filtering them directly.

For PReLU, the coefficient of the negative part is not constant and is adaptively learned. This approach will bring more calculations and more over-simulation to the algorithm. The possibility of integration, because PReLU only adds a very small number of parameters, so the calculation amount of the network and the risk of overfitting are only a little increased. But it can retain more data information, and the complexity of the entire model is not greatly improved, which can effectively avoid the problem of overfitting, so it also helps the model's training results have better fitting performance.

3 | METHODS

3.1 | Multi-task cascaded convolutional network (MTCNN)

This paper attempts to employ the convolutional neural networks to improve the detection efficiency of face images. In the process of image classification and recognition effectively processed by convolutional neural networks, the amount of calculation brought is too large to achieve real-time and other types of problems. Specifically, this paper adopts a face detection algorithm based on MTCNN. MTCNN is a cascaded network consisting of Proposal Network (P-Net), Refine Network (R-Net) and Output Network (O-Net). It adopts the idea of cascading, setting different confidence thresholds and IOU thresholds for each layer of the network, and each layer provides corresponding functions according to the network structure settings to effectively process the original picture information. The algorithm first uses Proposal Network (P-Net) to obtain the window of the face area and the boundary box regression, and the face area window will be corrected by the result of the boundary box regression, and then uses the non-maximum suppression (NMS) to merge the overlapping windows. Second, Refine Network (R-Net) is used to refine the input selection through the more powerful CNN network, filter out most non-face candidate windows, and then use border regression and facial key point locator to carry out the border of the face area regression and keypoint positioning are used to correct the results of bounding box regression, and R-Net uses NMS to merge overlapping windows. Finally, Output Network (O-Net) is employed to process the output of R-Net for further extraction, and at the same time perform face discrimination, face area border regression and face feature positioning, and finally output the upper left corner coordinates and lower right corner of the face area coordinates and five feature points of the face area. To some extent, the features of this model are similar with the HAAR cascade detection. They both use the cascade method, which rejects most of the image areas at the initial stage, effectively reducing the amount of calculation of later CNN. The specific procedure is shown in Figure 1.

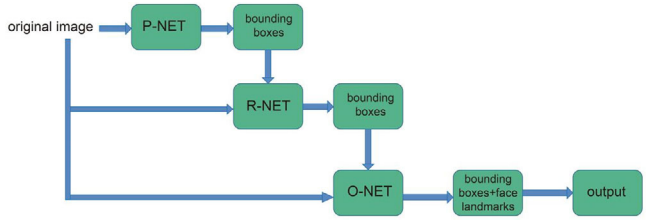

FIGURE 1 Face detection flow chart

3.2 | Proposal network (P-Net)

We use a fully convolutional attention network, called as proposal network (P-Net), in which the convolution kernel of each network layer is 3 × 3, and the input image size is 12 × 12 × 3 (12 × 12 means the length and width of input image, and 3 represents the colour channel) as shown in Figure 2a. P-Net obtains the candidate face window and its bounding box regression vector. The candidates are then calibrated based on the estimated bounding box regression vector. After that, we use non-maximum suppression (NMS) to merge highly overlapping candidates. The learning process is formulated as a two-class classification problem, and the cross-entropy loss is used for face classification. Cross entropy is an important concept in Shannon's information theory, used to measure the difference information between two probability distributions. In machine learning, cross entropy is used as the loss function, by assuming p is the true distribution of the sample, q is the distribution predicted by the model, and measuring the similarity between p and q to achieve the face classification task. Here, p_i represents the probability that the i^{th} candidate form is a face; y_i^{det} represents the true mark corresponding to the i^{th} candidate form. The cross-entropy formula for multimedia tools and applications is as follows:

$$L_i^{\text{det}} = -\left(y_i^{\text{det}} \times \log(p_i) + \left(1 - y_i^{\text{det}}\right) \times (1 - \log(p_i))\right), \quad y_i^{\text{det}} \in \{01\} \tag{1}$$

3.3 | Refine network (R-Net)

All candidates are fed to another CNN, which is called as Refined Network (R-Net) and the size of the convolution kernel of the first and second layers is 3 × 3 while that of the third layer is 2 × 2. The final fully connected layer has 128 neurons, and the number of neurons in the input layer is 24 × 24 × 3 (24 represents the length and width of the input image, and 3 represents the colour channel), as shown in Figure 2b. It

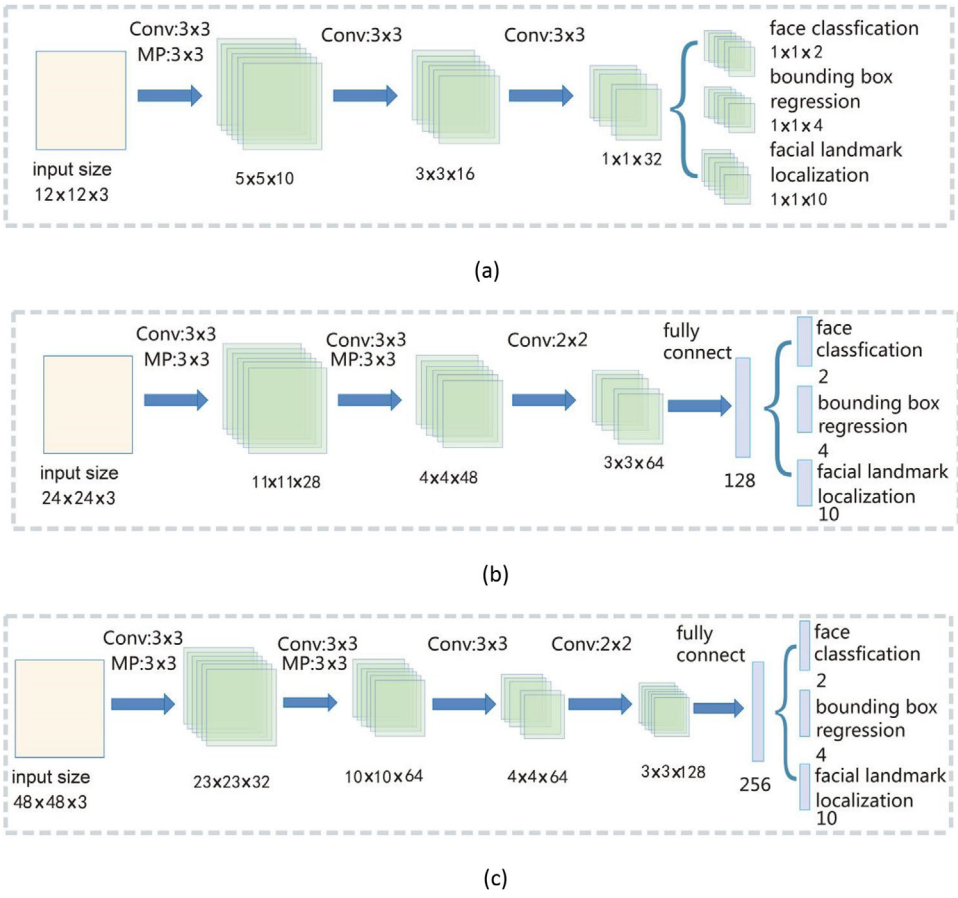

FIGURE 2 The architecture of MTCNN. (a) P-Net (Proposal Network); (b) R-Net (Refine Network); (c) O-Net (Output Network). MP, max pooling; Conv, convolution

further rejects a large number of wrong candidates, filters out most non-face candidate windows, and then uses border regression and facial key point locator to perform border regression and key point positioning of the face area to correct the results of bounding box regression, and Use NMS to deal with the problem of overlapping windows caused by different dimensional features and different windows. Among them, for the bounding box regression, the offset between it and the nearest ground fact (that is, the left, top, height and width of the bounding box) are predicted for each candidate window. The learning process is formulated as a regression problem, and the Euclidean loss is calculated for each sample x_i.

$$L_i^{box} = \left\| \hat{L}_i^{box} - L_i^{box} \right\|_2^2 \tag{2}$$

3.4 | Output network (O-Net)

This module, called as Output Network (O-Net), is similar to the previous module, but its goal is to identify facial regions with more supervision information. O-NET has four CNN layers and a fully connected layer. The convolution kernels of the first three layers are 3 × 3, why that of the last layer is 2 × 2. The final fully connected layer has 256 neurons, and the input size is 48 × 48 × 3 (48 × 48 represents the length and width of the input image, and 3 represents the colour channel), as shown in Figure 2c. The network outputs the positions of five facial landmarks to find five landmark points (two eyes, nose and two corners of the mouth) on the output surface. It retains more image features, and at the same time performs face discrimination, face area border regression and face feature positioning, and finally outputs the upper left and lower right coordinates of the face area and five feature points of the face area. O-Net has more characteristic input and more complex network structure, and also has better performance. The output of this layer is used as the final network model output. The landmark points are calculated as follows:

$$L_i^{landmark} = \left(\hat{y}_i^{landmark} - y_i^{landmark}\right)^2 \quad y_i^{landmark} \in R \tag{3}$$

3.5 | Overall flow

The overall flow of our method is shown in Figure 3. Given an image, we first adjust it to different scales to build an image pyramid, and then develop a fully convolutional attention network (P-Net) according to us to obtain the candidate face window and its bounding box regression direction, and then the candidates are calibrated based on the estimated bounding box regression vector and highly overlapping candidates are merged based on non-maximum suppression (NMS). Then all candidates are sent to another CNN (R-Net), which further rejects a large number of wrong candidates, uses bounding box regression for calibration, and performs NMS. Finally, the facial area is recognized through O-Net. In particular, the network will output the positions of five facial landmarks.

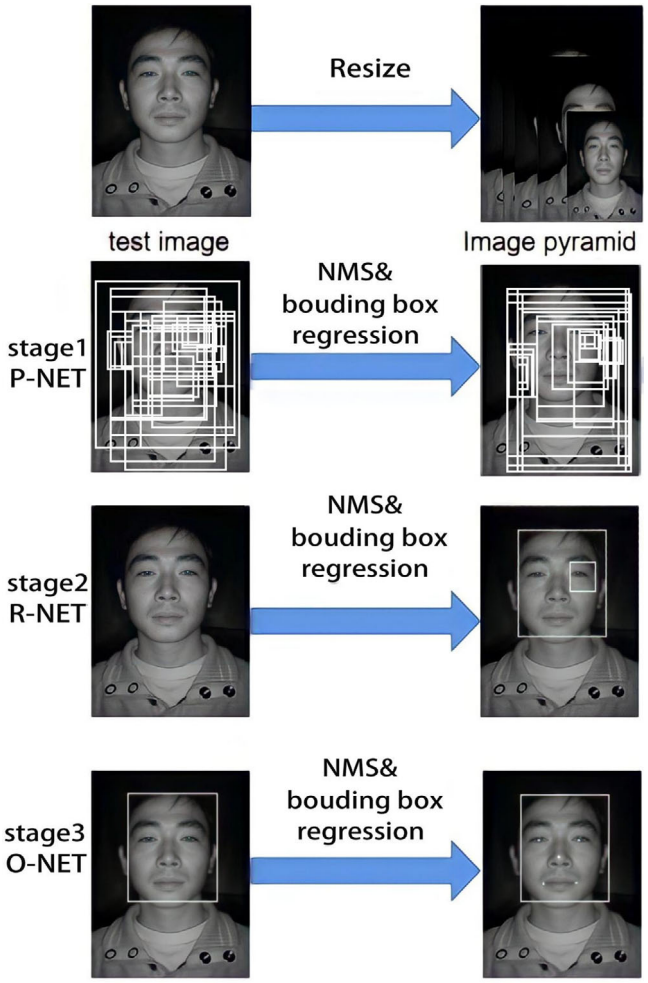

FIGURE 3 Pipeline of cascaded framework

4 | EXPERIMENTS

4.1 | Dataset

In order to evaluate the performance of the method, a dataset from Sketch-Visible light (VIS) and Near-infrared (NIR)-VIS is collected for experiment. Both NIR-VIS and Sketch-VIS are usually used in security inspection, but their spectral components are different, so there is a matching issue between them [19, 20, 22, 23, 24]. Sketch-VIS samples mainly come from the CUHK Face Sketch Database (CUFS) data set [5]and the CUHK Face Sketch FERET (CUFSF) data set [7]. The NIR-VIS samples mainly come from the CASIA NIR-VIS 2.0 data set [6]. These data sets are detailed as follows:

1. CUFS dataset: This facial dataset includes facial images of 188 students from the Chinese University of Hong Kong

TABLE 1 Face detection sample distribution data set

Purpose \ Dataset	CUFS dataset	CUFSF dataset	CASIA NIR-VIS 2.0 dataset	Total
Training set	606 (50%)	1872 (78.40%)	14064 (80%)	16564
Validation set	0	0	1758 (10%)	1750
Test set	606 (50%)	516 (21.60%)	1758 (10%)	2866
Total	1212	2388	17,580	21,180

and 123 individuals from the AR dataset, and the XM2VTS dataset . The avatars of 295 people. Therefore, it includes a total of 606 people, each of whom has a face photo, the light changes, and exaggerated sketches drawn by the artist when viewing the photo. There are a total of 1212 facial images. Since Sketch facial images are drawn by the artist based on the verbal description of witnesses, the hand-drawn drawings are very different from the photos of real criminals, that is, the modalities between images of different modalities. The difference is huge. At the same time, the data set only includes a few hundred pictures, and there are very few pictures that can be used for training, so the model is very easy to overfit, so it has a certain difficulty in recognition. An example is shown in Figure 4a

2. CUFSF data set: This face data set contains 1194 people from FERET (Phillips et al., 1997). For everyone, there is a photo of a face that changes as the light changes, and a photo drawn by the artist while looking at the photo to zoom in on the sketch. The data set has a total of 2388 facial images with different modalities. Compared with CUFS, the amount of data in CUFSF is more, and the data changes are greater, including changes in lighting, expressions, etc., so it is more challenging. One of them is shown in Figure 4b.
3. CASIA NIR-VIS 2.0 data set: This face data set is collected by the Chinese Academy of Sciences and is the largest cross-modal face recognition data set. It contains 17,580 NIR-VIS images of 725 volunteers. These images contain internal changes of the class, such as: changes in posture, age, resolution, and expressions, etc. By detecting near-infrared face images and visible light images, we can simulate the detection efficiency of algorithms in dark environments and visible light scenes. This data A sample of the set is shown in Figure 4c.

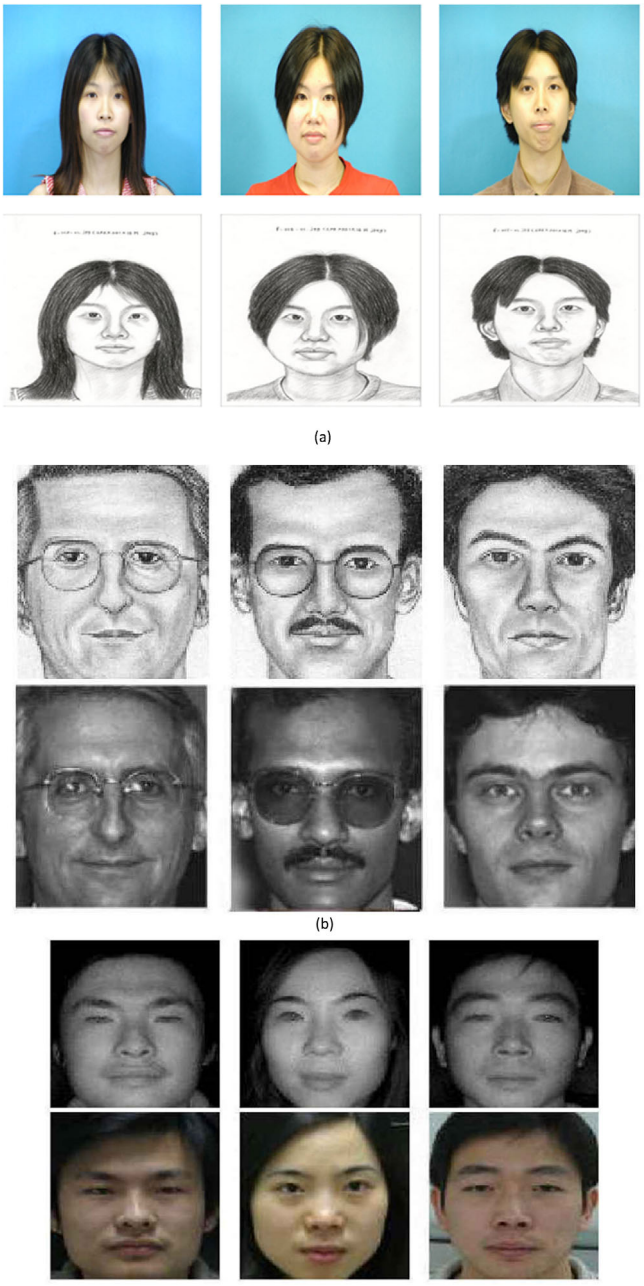

FIGURE 4 The style diagram of each data set. (a) CUFS data set; (b) CUFSF data set; (c) CASIA NIR-VIS 2.0 data set

This research divides the above data sets into training set, validation set and test set. The distribution of CUFS data set, CUFSF data set, CASIA NIR-VIS 2.0 data set is shown in Table 1.

Due to the sizes of the CUFS and CUFSF data sets are relatively small, there are few samples can be assigned to the verification set, which will result in the inability to fully estimate the model, and will also lead to the test results not being robust and significant. Therefore, the validation set of CUFS and CUFSF is set to 0.

Because of the modal gap between different modalities, it is difficult to adapt to heterogeneous images based on the visible spectrum. We use each mode (Visible-Visible, Visible-NIR, Visible-Sktech and Sktech-NIR) to calculate the SSIM (Structure and Similarity Score) on the same surface to reflect the spectral limit. The ssim for each mode is shown in Table 2.

4.2 | Experimental training process of MTCNN

In the training stage, the image is randomly cropped before input into the network, and horizontal flipping is used as data enhancement strategy. The model was trained and iterated on

TABLE 2 SSIM between each mode

Reference image	Visible	Visible	Visible	Sktech
Compare image	Visible	NIR	Sktech	NIR
SSIM	1.00	0.63	0.41	0.36

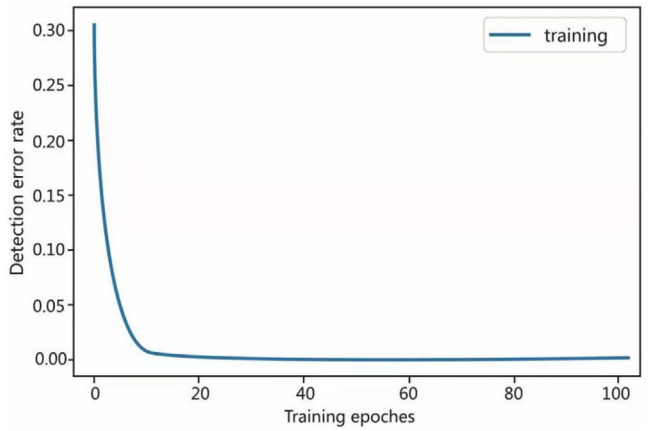

FIGURE 5 Training error curve

the wider face dataset with 100 epochs. The data input size of p-net is 12 × 12 × 3, that of R-Net is 24 × 24 × 3, and that of o-net is 48 × 48 × 3. Data set annotation is divided into four categories, IOU < 0.3 is negative sample, IOU > 0.65 is positive sample, IOU is 0.4–0.65 is part of the face, IOU is 0.3–0.4 is an unclear area. Two different learning rates, 0.001 and 0.0001, were used to train the iterative model. The learning rate is 0.005, and the learning rate is reduced once every 100 backward propagation. The experimental results show that when the larger learning rate of 0.001 is used, the model iteration is faster and the training time is shorter, but the final detection effect of the model is not ideal; when using 0.0001 as the basic learning rate, the iteration time of the model is delayed, but the final effect of the model is ideal. Large batch can let the model see more samples in the same iteration, which can make learning more stable and achieve better results. But at the same time, with the increase of batches, the mean value of the overall sample noise remains unchanged, but the variance decreases, and the sample noise helps the optimizer to avoid the local optimum and improve the overall generalization ability; in this experiment, the batch size is set to 32, that is, 32 samples are propagated backward each time. Training error curve. As shown in Figure 5.

4.3 | Experimental results and analysis of MTCNN

In our experiment, the network structure model was implemented on TensorFlow platform. In the training process, the batch is set as 100, and the initial learning rate is 0.01. When the evaluation index no longer improves, the learning rate is reduced by 10 times, and the minimal learning rate is set as 0.00001. An epoch means that all training sets are separately trained once.

TABLE 3 Face detection accuracy of each data set

Purpose \ Dataset	CUFS dataset	CUFSF dataset	CASIA NIR-VIS 2.0 dataset	Average accuracy
Validation set	0	0	96.63%	96.63%
Test set	98.32%	97.12%	96.41%	97.28%
Average accuracy	98.32%	97.12%	96.52%	96.95%

TABLE 4 Ablation experiments of the proposed methods on algorithms

Method	Average precision		
	CUFS (%)	CUFSF (%)	CASLA NIR-VIS 2.0 (%)
12-net + 24-net + 48-net	83.23	82.80	79.89
P-net + 24-net + 48-net	89.42	86.32	83.24
P-net + R-net + 48-net	92.05	90.88	89.29
P-net + 24-net + O-net	90.27	89.12	88.51
12-net + R-net + 48-net	89.35	88.12	87.12
12-net + R-net + O-net	87.62	85.62	84.38
12-net + 24-net + O-net	90.03	88.96	86.69
P-net + R-net + O-net	**98.32**	**97.12**	**96.52**

In this experiment, 100 epochs are conducted. The verification data set and the test data set are used to test the obtained network structure model, and the detection accuracy on each data set is shown in Table 3.

Since the number of images in the CUFS and CUFSF datasets is relatively small, so there is overfit risk. The average accuracy of the CUFS, CUFSF, and CASIA NIR-VIS 2.0 training data sets is 97.83%, and the average accuracy of the test data sets is 96.95%, the accuracy gap between the test data set and the training result is only 0.88%, demonstrating that the training network without over-fitting. In order to better understand the proposed MTCNN, we will train the detection networks 12-net, 24-net and 48-net according to the cascade structure. Then conduct a large number of ablation experiments to check the improvement of different network substructures. Table 4 shows that we performed ablation experiments on the proposed model. It can be seen from the table that the improved subnet will improve the overall network to a certain extent, but the overall network performance will also be restricted by other cascaded networks.

In order to verify the performance of the proposed algorithm, we use the datasets of CUFS, CUFSF and CASIA NIR-VIS 2.0 for comparison experiments. The detection accuracies of various methods for these data sets are shown in Tables 5–7 respectively.

At the same time, the speed comparison of MTCNN and other face detection methods is shown in Table 8

The calculation cost comparison between MTCNN and other methods is shown in Table 9

TABLE 5 Performance comparison of different face detection algorithms for CUFS dataset

Method	Detection accuracy (%)
MTCNN	98.32
SSH	89.13
R-CNN	97.96
HR	96.23

TABLE 6 Performance comparison of different face detection algorithms for CUFSF data set

Method	Detection accuracy (%)
MTCNN	97.12
SSH	86.23
R-CNN	96.48
HR	95.93

There is also a speed comparison between our CNNS and the previous CNNS, see Table 10.

To sum up, the proposed MTCNN achieves the best detection accuracy performance both in the dataset of CUFS and CUFSF, as well as the second fastest computation speed (only 5 FPS slow than the SSH), which means the proposed MTCNN is a satisfied solution for practical application.

5 | CONCLUSION

The study of heterogeneous face recognition is helpful to the application of face detection under unnatural conditions in real scenario, such as security inspection and law enforcement. In this paper, a multi-task cascaded convolutional neural network is proposed for face recognition in the scenes of visible light, near-infrared and sketch. The experimental results show that MTCNN has higher detection accuracy and excellent processing speed, indicating that the proposed method is suitable for face detection in different forms of images.

Although excellent results are achieved in the existing datasets, currently only the images with Sketch modalities, Visible light and Near-infrared are covered in this paper, therefore we plan to extend this method to large-scale heterogeneous face datasets in the near future.

TABLE 7 Performance comparison of different face detection algorithms for CASIA NIR-VIS 2.0 data set

Method	Detection accuracy (%)
MTCNN	96.52
SSH	82.23
R-CNN	94.78
HR	93.83

TABLE 8 MTCNN speed comparison with other methods

MTCNN	Nvidia Titan Black	95 FPS
SSH	Nvidia Titan Black	100 FPS
R-CNN	Nvidia Titan Black	23 FPS
HR	Nvidia Titan Black	5 FPS

TABLE 9 Comparison of computational cost between MTCNN and other methods

Method	Computational cost
MTCNN	10.5 ms
SSH	10 ms
R-CNN	43 ms
HR	200 ms

TABLE 10 Comparison of the speed of our CNNS and the previous CNNS [3]

Group	CNN	300 × forward propagation (s)
Group 1	12-Net[3]	0.038
	P-Net	0.031
Group 2	24-Net[3]	0.738
	R-Net	0.458
Group 3	48-Net[3]	3.577
	O-Net	1.347

CONFLICT OF INTEREST

We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in CUFS at http://mmlab.ie.cuhk.edu.hk/archive/facesketch.html. The data that support the findings of this study are openly available in CUFSF at http://mmlab.ie.cuhk.edu.hk/archive/cufsf/ The data that support the findings of this study are available from Institute of Automation, Chinese Academy of Sciences (CASIA).Restrictions apply to the availability of these data, which were used under license for this study. Data are available at https://mega.nz/folder/77oE3YoB#m-8_NSxfxhkb_EInzHfJuA with the permission of Institute of Automation, Chinese Academy of Sciences (CASIA).

ORCID

Wei Zhang https://orcid.org/0000-0002-5959-3037

REFERENCES

1. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vision 57(2), 137–154 (2004)
2. Yang, S., Luo, P., Loy, C.C., Tang, X.: From facial parts responses to face detection: A deep learning approach. In: IEEE International Conference on Computer Vision, pp. 3676–3684. IEEE, Piscataway (2015)
3. Li, H., Lin, Z., Shen, X., Brandt, J., Hua, G.: A convolutional neural network cascade for face detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5325-5334. IEEE, Piscataway (2015)
4. Neves, A., Ribeiro, R.: Algorithms for face detection on infrared thermal images. Int. J. Adv. Softw. 10, 499–512 (2018)
5. Wang, X., Tang, X.: Face photo-sketch synthesis and recognition. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 31, 1955–1967 (2009)
6. Li, S.Z., Yi, D., Lei, Z., Liao, S.: The CASIA NIR-VIS 2.0 face database. In: 9th IEEE Workshop on Perception Beyond the Visible Spectrum (PBVS, in conjunction with CVPR 2013). IEEE, Piscataway (2013)
7. Zhang, W., Wang, X., Tang, X.: Coupled information-theoretic encoding for face photo-sketch recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Piscataway (2011)
8. Ranjan, R., Patel, V.M., Chellappa, R.: A deep pyramid deformable part model for face detection. In: IEEE International Conference on Biometrics Theory, Applications and Systems, pp. 1–8. IEEE, Piscataway (2015)
9. Zhang, K., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. arXiv preprint, arxiv: 1604.02878 (2016)
10. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Computer Society Conference on Computer Vision & Pattern Recognition. IEEE, Piscataway (2003)
11. Wang, Y.: An analysis of the Viola-Jones face detection algorithm. Image Process. Line 4, 128–148 (2014)
12. Chen, D., Ren, S., Wei, Y., et al.: Joint cascade face detection and alignment. In: European Conference on Computer Vision, pp 109–122. Springer, Cham (2014)
13. Zou, Y., Liu, X.H.: Compressed deep convolution neural network for face recognition. In: 4th International Conference on Machinery, Materials and Information Technology Applications. Springer, London (2016)
14. Yang, B., Yan, J., Lei, Z., Li, S.Z.: Aggregate channel features for multi-view face detection. In: IEEE International Joint Conference on Biometrics, pp. 1–8. IEEE, Piscataway (2014)
15. Pham, M.T., Gao, Y., Hoang, V.D.D., Cham, T.J.: Fast polygonal integration and its application in extending haar-like features to improve object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 942–949. IEEE, Piscataway (2010)
16. Zhu, Q., Yeh, M.C., Cheng, K.T., Avidan, S.: Fast human detection using a cascade of histograms of oriented gradients. In: IEEE Computer Conference on Computer Vision and Pattern Recognition, pp. 1491–1498. IEEE, Piscataway (2006)
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105. MIT Press, Cambridge (2012)
18. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in Neural Information Processing Systems, pp. 1988–1996. MIT Press, Cambridge (2014)
19. Liu, S., Yi, D., Lei, Z., et al.: Heterogeneous face image matching using multi-scale features. In: 2012 5th IAPR International Conference on Biometrics (ICB), pp. 79–84. IEEE, Piscataway (2012)
20. Tang, X., Wang, X.: Face photo recognition using sketch. In: Proceedings of International Conference on Image Processing. IEEE, Piscataway (2002)
21. Huang, X., Lei, Z., Fan, M., et al.: Regularized discriminative spectral regression method for heterogeneous face matching. IEEE Trans. Image Process. 22(1), 353–362 (2013)
22. He, R., Wu, X., Sun, Z., et al.: Learning invariant deep representation for nir-vis face recognition. In: Thirty-First AAAI Conference on Artificial Intelligence. AAAI Press, Palo Alto (2017)
23. Chan, T.-H., Jia, K., Gao, S., Lu, J., Zeng, Z., Ma, Y.: Pcanet: A simple deep learning baseline for image classification. IEEE Trans. Image Process. 24(12), 5017–5032 (2015)
24. Ouyang, S., Hospedales, T., Song, Y.Z., et al.: A survey on heterogeneous face recognition: Sketch, infra-red, 3D and low-resolution. Image Vision Comput. 56, 28–48 (2016)
25. McCurrie, M., et al.: Convolutional neural networks for subjective face attributes. Image Vision Comput. 78, 14–25 (2018)
26. Altameem, T., Altameem, A.: Facial expression recognition using human machine interaction and multi-modal visualization analysis for healthcare applications. Image Vision Comput. 103, 104044 (2020)
27. Zhang, Y., et al.: Deep multimodal fusion for semantic image segmentation: A survey. Image Vision Comput. 105, 104042 (2020)
28. He, K., et al.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: IEEE International Conference on Computer Vision, pp. 1026–1034. IEEE, Piscataway (2015)
29. Mukherjee, S., Saha, S., Lahiri, S., Das, A., Bhunia, A.K., Konwer, A., Chakraborty, A.: Convolutional neural network based face detection. In: Proceeding of 1st International Conference on Electronics, Materials Engineering and Nano-Technology, pp. 1–5. IEEE, Piscataway (2017)
30. Ren, Z., Yang, S., Zou, F., Yang, F., Luan, C., Li, K.: A face tracking framework based on convolutional neural networks and Kalman fifilter. In: Proceeding of 8th IEEE International Conference on Software Engineering and Service Science, pp. 410–413. IEEE, Piscataway (2017)
31. Luo, D., Wen, G., Li, D., Hu, Y., Huan, E.: Deep-learning-based face detection using iterative bounding-box regression. Multimed Tools Appl. 77, 24663–24680 (2018)

How to cite this article: Yang, X.B., Zhang, W.: Heterogeneous face detection based on multi-task cascaded convolutional neural network. IET Image Process. 16, 207–215 (2022).
https://doi.org/10.1049/ipr2.12344