Hindawi
Scientific Programming
Volume 2020, Article ID 8861987, 10 pages
https://doi.org/10.1155/2020/8861987

Research Article

Face Detection and Recognition Based on Visual Attention Mechanism Guidance Model in Unrestricted Posture

Zhenguo Yuan

School of Mechanical and Electrical Engineering, Guangdong Industry Polytechnic, Guangzhou, China

Correspondence should be addressed to Zhenguo Yuan; 2016001053@gdip.edu.cn

Received 14 April 2020; Revised 28 April 2020; Accepted 6 May 2020; Published 20 May 2020

Academic Editor: Chenxi Huang

Performance of face detection and recognition is affected and damaged because occlusion often leads to missed detection. To reduce the recognition accuracy caused by facial occlusion and enhance the accuracy of face detection, a visual attention mechanism guidance model is proposed in this paper, which uses the visual attention mechanism to guide the model highlight the visible area of the occluded face; the face detection problem is simplified into the high-level semantic feature detection problem through the improved analytical network, and the location and scale of the face are predicted by the activation map to avoid additional parameter settings. A large number of simulation experiment results show that our proposed method is superior to other comparison algorithms for the accuracy of occlusion face detection and recognition on the face database. In addition, our proposed method achieves a better balance between detection accuracy and speed, which can be used in the field of security surveillance.

1. Introduction

There are still some challenging problems in face detection and recognition technology mainly due to the nonrigid features and the influence of complex background [1–3]. Traditional face detection algorithms mostly use a semi-supervised learning method. Since the traditional method needs to design different artificial features for different tasks, such as grayscale features, contour features, and HOG features, these features are easily affected by the imaging angle, and the generalization ability is poor. At the same time, the object occlusion will also lead to the missed detection, thereby reducing the accuracy of the detector. Therefore, it is of great practical significance to study the occlusion problem for face detection and recognition task [4].

The face detection and recognition model based on machine learning is a popular research direction in the field of computer vision [5]. By directly extracting features from the detection area and then using machine learning algorithms to classify and recognize, the accuracy of the model classification can be improved to a certain extent, but the characterization ability of features directly affects the recognition accuracy of the system [6]. Compared with detection and recognition algorithms for shallow learning models such as boosting, decision trees, and neural networks, deep learning represented by convolutional neural networks implements the deep nonlinear network structures through operations such as local receptive fields and weigh sharing. The hierarchical strategy can learn the most essential feature representation in the data set [7]. At present, mainstream deep learning-based face detectors usually adopt a two-stage network structure, which is divided into face detection and face recognition.

Most convolutional neural networks use a classification loss function to measure the difference between the predicted value and the actual value and then complete the classification of the image through the training process to expand the distance between different types of images. Wang et al. [8] used 3-dimensional face information as a feature where the robustness and accuracy of the algorithm are improved through a large amount of data training. Corrow et al. [9] used DeepID for face recognition by partitioning different parts of the face, extracting features separately, and

then using the Bayes algorithm to perform complex operations on the features, and finally obtaining face feature information, effectively improving the accuracy of recognition. However, none of the above algorithms solves the recognition problem under nonlimiting conditions. Therefore, how to increase the distance between classes while reducing the distance of intraclasses in the recognition process is the important topic of the face recognition task. Abbad et al. [10] realized the feedback of the loss function during the training process by adding a loss verification method and used the positive samples to reduce the distance between the classes, but this method is more dependent on the samples. Madhavan and Kumar [11] proposed a ternary loss algorithm that unifies the training data into triple elements; each triple contains positive value, negative value, and sample anchor point, which can effectively reduce the intraclass distance.

Although the above method can solve some nonlimiting problems, it has poor performance in convergence speed, especially when the number of network layers is too large, and vanishing gradient phenomenon will occur. In order to solve this kind of problem, Su et al. [12] proposed a multi-inception structure-based convolutional network neural algorithm for face recognition. By transforming the traditional Softmax loss method and combining Softmax and TripletLoss, a larger interclass distance and a smaller intraclass distance can be obtained. Experiment result proves that the algorithm increases the depth and width of the network, and intraclass spacing can be effectively reduced during the training process. Based on the above description, it can be seen that the convolutional neural network faces different problems when processing different data and application scenarios. Some scenes pay more attention to calculation speed, and some pay more attention to detection accuracy. More scholars strive to find a universal model with high performance in all aspects. This is also the ultimate goal of this study.

Aiming at the problem that occlusion affects the accuracy of face detection and recognition, this paper proposes a deep network with multilevel feature fusion. This network uses a visual attention mechanism to guide the model to highlight the visible area of the occlusion face; the detection recognition problem can be simplified to a high-level semantic feature detection problem, and the position and scale of the face are predicted by means of activation maps, avoiding additional parameter settings. A large number of simulation experiment results show that the proposed method is better than the existing mainstream method in the detection and recognition of the occlusion face on the public data set and has achieved a faster detection speed, which can be used in the field of security surveillance.

The innovations of this study are summarized as follows:

(1) In view of the detection omission caused by occlusion in face detection, a solution is proposed, that is, a visual attention mechanism guidance model is proposed, which uses the visual attention mechanism guidance model to highlight the visible area of the blocked face, thereby improving face detection and recognition accuracy.

(2) The new model parameters have been simplified. Through the improved analysis network, the face detection problem is simplified to the advanced semantic feature detection problem, and the position and scale of the face are predicted by the activation map to avoid additional parameter settings.

2. Face Detection Network

The YOLO-V3 network is a better deep learning model in the field of object recognition, the network evolved from the YOLO and YOLO-V2 networks [13]. Compared with the deep learning network based on region proposal, the YOLO network transforms the detection problem into a regression problem. The network does not need to adopt exhaustive candidate regions but directly generates the confidence and bounding box coordinates of the object through regression. Compared with the Faster-RCNN network, the detection speed is greatly improved [14].

The YOLO detection model is shown in Figure 1. The network divides each image in the training set into an $S \times S$ ($S = 13$) grid. If the center of the real object falls into the grid, the grid is responsible for detecting the category of the object. Multiple bounding boxes are predicted in each grid, and each predicted bounding box is scored to demonstrate that the bounding box completely contains the confidence of the object, which is defined as follows:

$$C = P_r(\text{object}) \times \text{IoU}_{\text{pred}}^{\text{truth}}, \quad P_r(\text{object} \in \{0, 1\}), \tag{1}$$

where $P_r(\text{object})$ indicates the probability of the object contained in the bounding box. If there is an object in the bounding box, we have $P_r(\text{object}) = 1$; otherwise $P_r(\text{object}) = 0$; $\text{IoU}_{\text{pred}}^{\text{truth}}$ indicates the Intersection over Union (IoU) between the prediction result and the benchmark frame. The confidence reflects whether the grid contains objects and the accuracy of prediction boundary box. When multiple bounding boxes detect the same object, YOLO uses a nonmaximum suppression method to select the best bounding box.

Although YOLO has obtained a faster detection speed, its detection accuracy is not as good as Faster R-CNN. In order to solve this problem, YOLO-V2 introduces the idea of the anchor mechanism in the Faster R-CNN network and uses the k-means clustering method or fuzzy c-means method [15] to generate a suitable prior bounding box. Therefore, the number of anchor boxes required by the YOLO-V2 algorithm to achieve the same IoU is reduced. YOLO-V2 improves the network structure and replaces the fully connected layer in the YOLO output layer with a convolutional layer [16]. In addition, YOLO-V2 also introduces batch normalization, dimensional clustering, fine-grained features, multiscale training, and other strategies; compared with YOLO, YOLO-V2 greatly improves the detection accuracy. YOLO-V3 is an improved model based on YOLO-V2. By using multiscale prediction to detect the

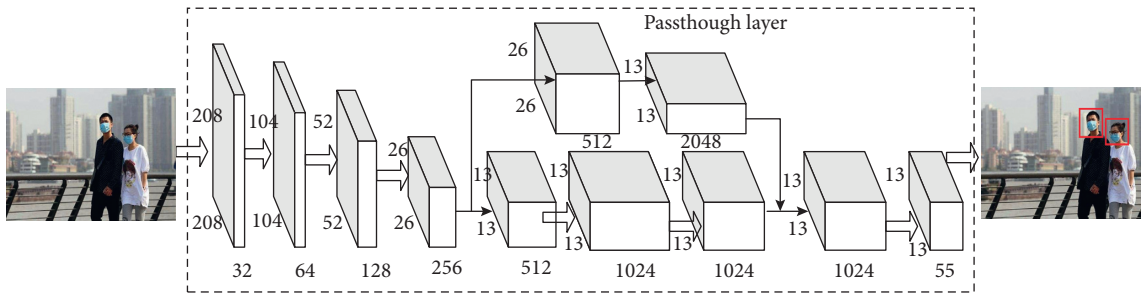


Figure 1: Face detection and recognition process.

final object, its network structure is more complicated than YOLO-V2. YOLO-V3 can predict the bounding boxes of different scales, which can detect small objects more effectively than YOLO-V2, but there are still missing detections for partially occluded face objects [17].

3. Occlusion Face Detection and Recognition Algorithm Combined with the Visual Attention Mechanism

Figure 2 is our proposed face detection and recognition model in this paper, which consists of two parts: feature extraction network and face analysis network. The input image is extracted by the feature extraction network to extract high-level semantic features, and the feature-guided attention module is used for feature fusion; the face analysis network predicts the face position, height, and offset heat map on the basis of the obtained high-level semantic feature and obtains the face boundary box.

3.1. Feature Extraction Network. The feature extraction network is designed on the basis of feature pyramid networks (FPN) [18–21], including basic networks and visual attention networks. ResNet50 has excellent performance in visual tasks such as image classification, so it is used as the basic network. ResNet50 can be divided into 5 levels, and the downsampling rate of each level relative to the input image is i; {1, 2, 3, 4, 5} represents the number of levels. In order to make full use of the location information of the shallow feature map and the semantic information of the deep feature, the shallow and deep feature maps are used to guide the attention network for feature fusion. The process can be described as follows: firstly, the number of C3 and C4 feature channels can be reduced to 256 by the convolutional layer with the size of the 1×1 convolution kernel, which can reduce the amount of calculation; then, the backbone network feature maps (namely, P3 and P4) after bilinear interpolation and upsampling 2 times are, respectively, input into the guided attention module to feature fusion [22–24].

3.2. Visual Attention Network. Different feature channels of the convolutional network will have different responses to a specific area of the face, which is to say that the occlusion form of the object can be described through different feature channels, and the occlusion form $O(n)$ is defined as the following equation:

$$o(n) = [v_0, p_0, v_1, p_1, \ldots, v_k, p_k], \tag{2}$$

where p_i denote different areas of face objects and $v_i \in \{0, 1\}, i \in [0, k]$ is used to indicate whether the partial area i of the face is visible.

The weights of traditional CNN channels are usually fixed and the same, which limits the network's ability to express different occlusion forms [19, 21, 23–26]. Patil et al. [27] recalibrated the weight of each channel so that the feature channel expressing the visible area of the occlusion object has a greater contribution to the final convolution feature, which can highlight the occlusion object in the background. The channel weighting process can be expressed as the following equation:

$$F_o(n) = \mathbf{\Omega}_n \mathbf{F}_c, \tag{3}$$

where $\mathbf{F}_c$ is the channel feature and $\mathbf{\Omega}_n$ is the channel weighting vector corresponding to the occlusion form n. The visual attention module is to get the attention vector $\mathbf{\Omega}_n$ through learning and finally achieve the reweighting of the feature channel for $\mathbf{\Omega}_n$ so that the network can adaptively express different occlusion forms. However, the existing models only consider the relationship between channels and ignore the importance of spatial information for the feature map. Because the spatial information of the feature map is helpful for the network to locate the region of interest, the feature channel attention mechanism and the spatial attention mechanism are used in the feature description task in literature [28]. Similarly, spatial attention mechanism is applied to object detection tasks in literature [29], so as to guide the network to highlight the useful features for current tasks. On the basis of the above description and analysis, the feature space information is used in the face detection and recognition task to highlight the occlusion of the face object area, and the spatial attention module is constructed to achieve the face detection and recognition task. The spatial attention module obtains the spatial attention map from the spatial information of the statistical feature map, which is used to reactivate the input features, so as to guide the network to focus on the occluded face and suppress the background interference.

As shown in Figure 3, the visual attention network consists of two submodules: channel and spatial attention. The input of the visual attention network is two feature maps

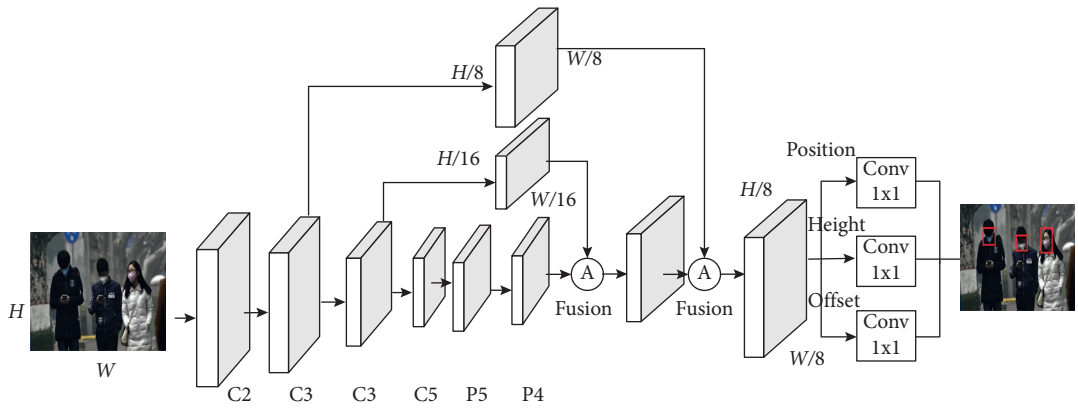


Figure 2: Model framework.

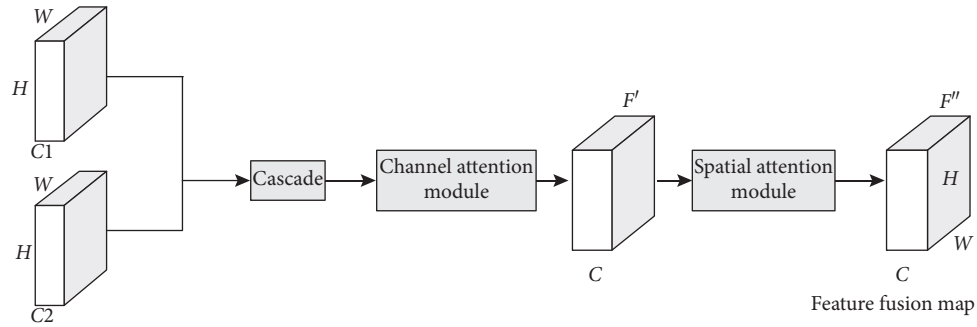


Figure 3: The overall structure of the visual attention module.

(such as C4 and P4) from the shallow convolution layer and the deep convolution layer, respectively. Firstly, the input features are connected in the channel dimension to get $F \in R^{H\times W\times C}$, and then F is input to channel attention module. After a series of operations, the spatial attention module is to achieve the feature fusion fusion. Therefore, by using the attention module to model the correlation between feature channels and the spatial information of the feature map, the network can not only enhance the feature representation of the relevant areas but also obtain the location information of the area of interest [29]. While making full use of the useful features to deal with the problem of face occlusion, it also suppresses the useless clutter information, which is conducive to improving the accuracy of face and recognition detection.

Figure 4 is the feature channel attention module proposed in this paper. For the input feature map F, the global information of each feature channel is obtained by global average pooling and maximum pooling operations to form the channel descriptor z_c^{avg} and $z_c^{\max}$, and then the feature channel attention vector $\Omega_c \in R^{1\times 1\times C}$ is obtained through the two fully connected layers FC1 and FC2. Finally, the deep learning method makes the network to automatically characterize the occlusion form of different samples. The specific steps are shown in equation (3).

$$\Omega_c = \sigma\left(W_2\left(\delta\left(W_1 z_c^{\text{avg}}\right)\right) + W_2\left(\delta\left(W_1 z_c^{\max}\right)\right)\right), \tag{4}$$

where σ and δ are sigmoid function and ReLU function, respectively. $W_1 \in R^{\text{Clr}\times C}$ and $W_2 \in R^{C\times \text{Clr}}$ represent two fully connected layer parameters, where r is the ratio of downsampling dimensionality reduction. $\Omega_c \in R^{1\times 1\times C}$ is used to weight the input feature F channel by channel to obtain F'. The process can be written as the following equation:

$$F' = \Omega_s \otimes F', \tag{5}$$

where $\otimes$ represents the dot-product channel by channel.

Since the useful information for partial-occluding face objects is usually obscured by the background, the network also needs to determine the spatial location of the useful information while enhancing the feature expression of the occlusion object through the channel attention module. Unlike the channel attention mechanism, the spatial attention mechanism is mainly used to highlight the areas in the feature map that are related to the current task, which is to guide the network to focus on the visible area of the occlusion object [30].

In the spatial attention module, the maximum pooling operation is firstly performed on the input feature map F' in the channel dimension to obtain the feature map $F_{\max}{}' \in R^{H\times W\times 1}$, which is used to count the spatial information of the feature map; then, the feature map is input into a 3×3 convolution layer f_c and output by the sigmoid function to obtain the spatial attention map $M_s \in R^{H\times W\times 1}$:

$$M_s = \sigma\left(f_c\left(F'_{\max}\right)\right), \tag{6}$$

where σ is the sigmoid function. Finally, the spatial attention map M_s is used to reactivate the input F' to obtain the final feature map F'':

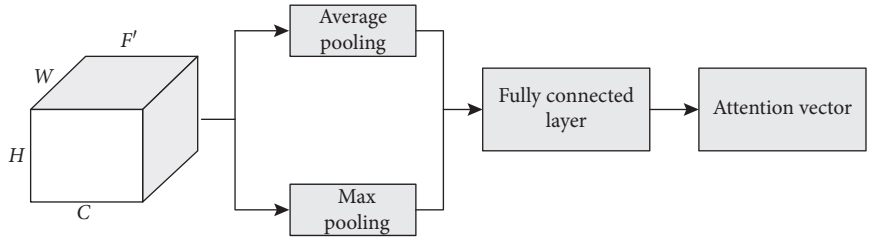


Figure 4: Feature channel attention module structure.

$$F'' = M_s \otimes F', \tag{7}$$

where $\otimes$ represents the dot product between feature maps.

Face detection and recognition task is regarded as a high-level semantic feature detection problem. On the basis of obtaining semantic features, the final prediction bounding box is obtained through the face analysis network. In this paper, the position, height, and position offset of the face are firstly predicted, and the size of the bounding box is obtained by simple geometric transformation, and then, the simple recognition network can get the high-precision recognition effect [31]. Specifically, after the predicted height h of the face is obtained, the width $w = h \cdot \alpha$ of the bounding box can be calculated by the length-width ratio of the bounding box. If the output feature map of the feature extraction network is $F_{\text{final}} \in R^{H/s \times W/s}$, three heat maps are predicted by three parallel 1×1 convolutional layers and correspond to center position $H_c \in R^{H/s \times W/s}$, height $H_h \in R^{H/s \times W/s}$ and position offset $H_{\text{offset}} \in R^{H/s \times W/s}$, respectively. s is the sampling rate of the output activation map relative to the input image. By predicting the heat map, the limitation of the prior frame adopted by the traditional method is avoided, and a more flexible face detection and recognition is realized in the same network.

3.2.1. Position Prediction. Face location prediction is achieved through the location heat map H_c. In this paper, the position prediction problem is simplified as a binary classification problem.

The position of the object center on the feature map F_{final} is (x_c, y_c). And the object center pixel is selected as a positive sample and the other positions as negative samples. The cross-entropy loss function is used to optimize the training position-prediction branch. The training true value H_c^{gt} is generated by a 2D Gaussian function, and the truth value at any position can be obtained by calculating equation (8):

$$H_c^{gt}(i, j) = \max\left(G(i, j; x_c, y_c, \sigma_w, \sigma_h)\right), \tag{8}$$

$$G(i, j, x_c, y_c, \sigma_w, \sigma_h) = e^{-\left(\left((i-x_c)^2/\sigma_w^2\right)+\left((j-y_c)^2/\sigma_h^2\right)\right)}, \tag{9}$$

where (x_c, y_c) is the central location of the object and σ_w and σ_h are variances of the width and height of the object, respectively. In order to alleviate the imbalance of positive and negative samples in the training process, focal loss was defined as the predicted loss function of the center position:

$$\begin{cases} L_c = \dfrac{-1}{N} \displaystyle\sum_{i=1}^{W/s} \sum_{j=1}^{H/s} BW, & \\ \left(1 - p_{i,j}\right)^{\alpha} \log\left(p_{i,j}\right), & H_c^{gt}(i, j) = 1, \\ \left(1 - H_c^{gt}(i, j)\right)^{\beta} p_{i,j}^{\alpha} \log\left(1 - p_{i,j}\right), & \text{others}, \end{cases} \tag{10}$$

where $p(i, j)$ indicates the prediction score of the object center at (i, j) in the prediction heat map, N is the number of objects in the picture, and α and β are the balance factors, generally set to 2 and 4.

3.2.2. Height Prediction. Given the position of the face k in the height heat map which is (x_k, y_k), its corresponding true value is $H_h^{gt}(x_k, y_k) = \log(h_k)$, where h_k denotes the height of the object k. In this paper, the true value within the radius r of (x_k, y_k) is set to $\log(h_k)$, and the radius r is set according to the width of the object, which is generally set as $r = 0.5w_k$. Our proposed model in this paper uses the L1 loss function for training. The loss function is denoted as follows:

$$L_h = \frac{1}{N} \sum_{k=1}^{N} L1\left(h_k, H_h^{gt}(x_k, y_k)\right), \tag{11}$$

where h_k is the predicted height of the object k in the heat map and N is the number of objects in the image.

3.2.3. Deviation Prediction. Since the convolutional network is usually a downsampling process, the position (x, y) on the input image is mapped into the heat map, whose position can be expressed as $(x/s, t\,y/s)$, where s is the downsampling rate of the network. When the position on the activation map is remapped back to the input image, an error will be generated, especially affecting the detection and recognition result of the dim-small face. To alleviate this problem, the position prediction of the object is corrected by predicting the deviation/offset of the center position [32], and the corresponding true value can be rewritten as the following equation:

$$H_{\text{offset}}^{gt}(x_k, y_k) = \left(\frac{x_k}{s} - \left\lfloor \frac{x_k}{s} \right\rfloor, \frac{y_k}{s} - \left\lfloor \frac{y_k}{s} \right\rfloor\right). \tag{12}$$

Finally, the multitask loss function weighted optimization can be adopted to train our proposed network, where the weighted loss function can be denoted as follows:

$$L = \lambda_c L_c + \lambda_h L_h + \lambda_o L_{\text{offset}}, \tag{13}$$

where λ_c, λ_h, and λ_o are weighting factors, which are set to 0.01, 1, and 0.12, respectively。

4. Experimental Results and Analysis

4.1. Experimental Data Set and Parameter Settings. In order to evaluate the performance of the face detection and recognition algorithm based on the visual attention-guided mechanism proposed in this paper, LFW (labeled faces in the wild database), CMUFD database (CMU face detection database) [23], and UCFI database (UCD color face image) [25] are used as face detection and recognition data sets. It consists of 500 images with a resolution of 2048×1024. Since CMUFD contains a large number of partial occlusion face images, it is selected as the verification and comparison test of the proposed method; UCFI contains about 350,000 face samples, where the standard testing set consists of 4,024 images with a resolution of 640×680 in a simple scenario. In order to verify the generalization of our proposed method, some testing experiments are performed on the UCFI data set. Face objects of all training data have been accurately marked. Except for the object area, the rest is marked as background, which means that the labeled data set can be used for training and testing of face detection and recognition models.

The network proposed in this paper selects pretrained ResNet50 as the backbone network. Its parameters are set as follows: depth = 40, growth_rate = 12, bottleneck = True, reduction = 0.5, minibatch is set to 16; learning rate is set to 0.001, dropout parameter is set to 0.8, and the maximum number of iterations is set to 10,000; in order to improve the optimization efficiency, this paper uses the Adam optimization algorithm. The Adam optimization algorithm is an extension of the stochastic gradient descent algorithm, which can iteratively update the neural network weights based on the training data; the initialization of learning rate is set to 0.25; then, when training to the 30th epoch, the learning rate is changed to 0.025. The nonmaximum suppression algorithm (NMS) is used to filter out the redundant face results. The threshold of Intersection over Union (IoU) is set to 0.5, and only the face results with the object confidence score greater than 0.1 are retained [33].

Our proposed network module is based on the PyTorch deep learning framework. The experimental environment is Xeon (Xeon) E7-8890 v2 @ 2.80 GHz (X4), 128 GB (DDR3 1600 MHz), Nvidia GeForce GTX 1080 Ti, Ubuntu 6.04, 64-bit operating system.

4.2. Evaluation Index. At present, the video surveillance intelligent analysis system has been able to detect and recognize the face with different scales. However, existing algorithms have a large number of false detection for face objects under partial occlusion mainly due to incomplete occlusion of face objects and the similarity of the face and background gray. Therefore, the standard evaluation indicators are selected as performance evaluation, which is the false positive per image (FPPI) of each image, focusing on the frequency of occurrence of false positive, as shown in Table 1.

In the detection stage, the evaluation criteria are the detection rate (DR) and the false detection rate (false positive per image, FPPI):

$$\begin{aligned} \text{DR} &= \frac{TP}{(\text{TP} + \text{FN})}, \\ \text{FPPI} &= \frac{FP}{(\text{FP} + \text{TN})}, \end{aligned} \tag{14}$$

where TP represents the number of positive samples detected correctly, $TP + FN$ represents the number of positive samples included the image, and FP represents the number of false positive samples. In addition, we also use the log-average miss rate (MR) to characterize the performance of the detector for the face. This paper mainly focuses on the occlusion situation for face detection and recognition. Therefore, we define the visible range to characterize the occlusion situation of the face. Given that the proportion of the object visible area to the total area is λ, if $\lambda > 0.7$, it means that the object is in a normal state and is denoted as N; if $0.2 < \lambda < 0.7$, it means that the object is in a serious occlusion state and is denoted as H. For ease of analysis, we will also divide the data set into four types of subsets, which are, respectively, recorded as mixed face data set (mixed), bare occlusion face data set (bare), partially occlusion data set (partial), and severely occlusion data set (heavy) [34].

4.3. Performance Analysis for Face Detection. In order to verify the effectiveness of the feature-guided attention network, the detector with the attention network removed is used as a test baseline (baseline), and Face++ proposed in [17] is used as a comparison method. Baseline adopts the feature fusion method consistent with FPN to build the model, where CA means the channel attention module and SA means the space attention module, and the performance of the detector after adding each module is compared in the experiment, and the test results are shown in Table 2.

Compared with the baseline model, the face detector's missed detection (MR) of the occlusion image has a significant decrease after adding the attention module, indicating that our proposed attention mechanism can effectively guide the detector to focus on the occlusion object. Compared with Face++, the proposed method has a significant reduction in MR under all evaluation criteria, especially under the severe occlusion evaluation criteria, and its missed detection (MR) decreased by 18.2%, indicating that the proposed method has good effectiveness for face detection in the complex scenario.

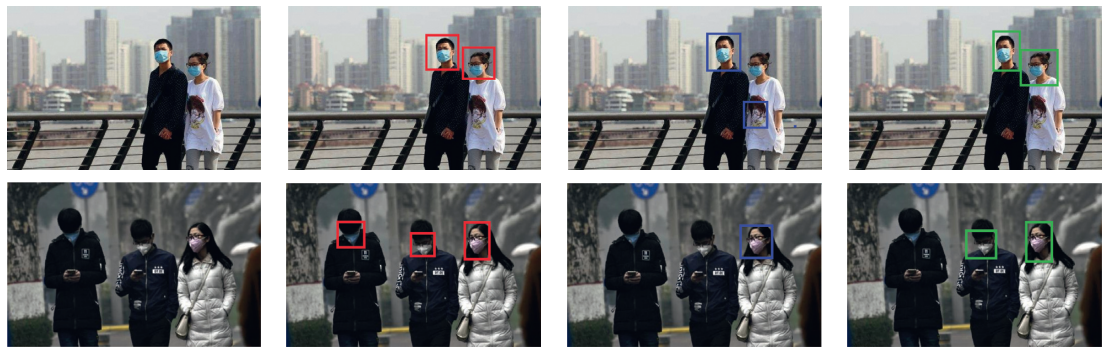
In order to more intuitively understand the attention module on the performance of the face detector, Figure 5 gives visualized face prediction results. Figure 5(a) is a face image, and input it to Baseline, Baseline + CA, and Baseline + CA + SA, respectively, to obtain the position prediction. Through observation, it can be found that the feature response in Figure 5(d) is closer to the face visible

Table 1: Statistics of face detection results.

Results	Benchmark	
	Face (positive)	Background (negative)
Face (positive)	(True positive) TP	(False positive) FP
Background (negative)	(False negative) FN	(True negative) TN

Table 2: Comparative analysis of visual attention network verification results in the face detection module.

Methods	N (%)	H (%)	$N+H$ (%)
Face++	16.11	56.74	38.12
Baseline	12.16	41.05	37.92
Baseline + CA	11.79	39.31	37.83
Baseline + CA + SA	11.53	38.73	37.26



(a) (b) (c) (d)

Figure 5: Performance comparison for face detection. (a) Face image with occlusion. (b) Baseline + CA + SA. (c) Baseline + SA. (d) Baseline + CA.

area, while there is still background interference in Figures 5(b) and 5(c), but the interference of Figure 5(c) is significantly less than that of Figure 5(b). This also proves that the attention module can guide the network to highlight to the visible part of the occlusion face, while also reducing the impact of background noise on detection performance.

4.4. Performance Analysis for Face Recognition. Since the recognition algorithm based on deep learning has achieved great results in the field of natural image, some classical deep learning algorithms will be used to make a comparison with our proposed algorithm. In order to qualitatively and quantitatively analyze the accuracy of the proposed algorithm for occluded face detection, this paper selects the comparison algorithm as FACEILD [35], Faster-FCC [36], KSDD [37], DNET [14], ResNet [7], and ConvNet [38] to further verify the performance of the proposed method.

From the experiment results in Table 3, it can be seen that the face detection results proposed in this paper are better than the DNET mainly due to the improvement of the face detection accuracy of the attention perception fusion module and the use of the multiscale pyramid pooling layer to capture high-level semantic features. The complementary features can effectively preserve the clear boundary of the face, while the combination of the multiple side output and pyramid pooling layer output can extract rich global context information and adapt to the two classification problems of face recognition. The heavy data set contains the most complex face image in the whole testing data. The face is seriously occluded, especially the image contrast is small and the face is fuzzy, which directly affects the detection and recognition effect of the network. From the precision comparison results in Table 3, it can be seen that the detection rate on the heavy data is the lowest mainly because the occlusion greatly reduces the perception ability of the deep network. However, the face detection based on the visual attention-guided mechanism proposed in this paper is also better than other deep networks. It can be seen that our proposed algorithm achieves 59.78% detection accuracy under the heavy evaluation standard, which is better than the comparison methods. In terms of detection efficiency, the detection and recognition speed of the input image with a resolution of 1024×2048 is 0.22 s, achieving a good balance between speed and accuracy. If the input image with a smaller resolution is detected, the detection speed of this method will be further improved.

In the selected detection image, the shape and scale of the face are quite different, especially the gray level of the face and the adjacent background is similar. According to the maximum analysis of the response map, these appearance changes cannot get the accurate boundary, which leads to ConvNet, KSDD, and FACEILD cannot get the accurate face. However, it is proved that the face with serious

Table 3: Recognition rate under different data sets.

Data set	Models						
	KSDD (%)	DNET (%)	ResNet (%)	ConvNet (%)	Faster-FCC (%)	FACEILD (%)	Proposed (%)
Bare	74.11	85.47	87.22	86.28	83.23	90.15	**90.06**
Mixed	69.12	78.22	81.25	87.24	79.39	87.09	**87.28**
Partial	78.28	85.28	88.20	85.69	78.29	88.36	**89.21**
Heavy	51.91	52.90	59.75	59.11	52.44	58.69	**59.51**

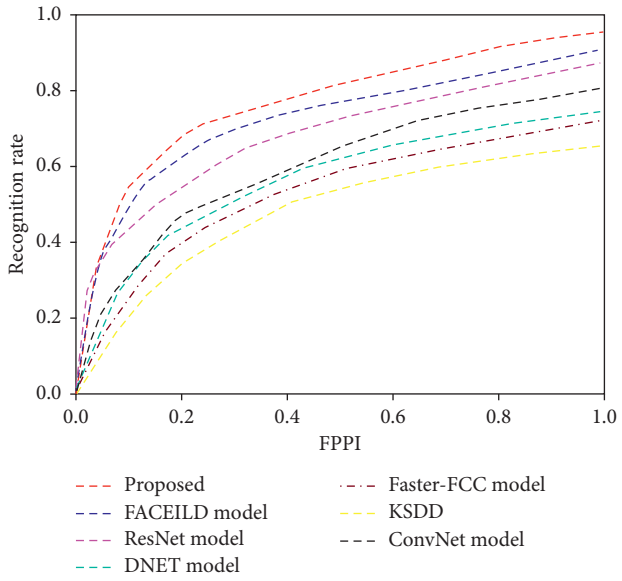


Figure 6: Relationship curve between FPPI and recognition rate on the LFW database.

occlusion such as deformation and low contrast can be accurately detected and recognized. KSDD is a lightweight network structure based on the VGG network. Although it can balance the contradiction between robustness and speed, it is still easy to be disturbed by occlusion, resulting in deviation of the detection center. From the detection results, it can be seen that the ConvNet detection and recognition has deviated from the face center. Our proposed model in this paper uses the attention-guided mechanism to highlight the visual area of occluded faces so that our proposed algorithm in this paper can better adapt to the influence of occluded interference in face detection.

4.5. Generalization Analysis. In order to verify the generalization performance of the proposed method, the proposed method was trained on the LFW training set, and cross-data set experiments were performed on the CMUFD database. The heavy subset consists of face objects with a height greater than 50 pixels and a visible range of [0.20, 0.65]. As shown in Figure 6, FPPI represents the statistical results of face detection and recognition algorithms at different detection rates. In order to facilitate comparison in different deep networks, the experiment mainly discusses the detection results of each algorithm when FPPI = 1 for analysis. The recognition rate of the ResNet algorithm is 91.88%, the recognition rate of the KSSD algorithm is 51.91%, and the detection rate of DenseNet algorithm is only 58.69%. The reason is that most deep detection methods only use the side-output feature and ignore the importance of global structural features. Our proposed paper uses a visual attention mechanism to guide the model to highlight the occlusion object visible area and simplify the face detection and recognition problem to a high-level semantic feature detection problem through an improved analytical network and uses the activation map to predict the location and scale of the face, which can avoid additional parameter settings and further reduce the false detection rate of each image. It can be clearly observed from Figure 6 that the performance of the proposed algorithm is obviously another algorithm.

5. Conclusions

Performance of face detection and recognition is affected and damaged because occlusion often leads to missed detection. In order to improve the accuracy of face detection and recognition, a visual attention mechanism guidance model is proposed in this paper, which uses the visual attention mechanism to guide the model highlight the visible area of the occluded face. The face detection problem is simplified into the high-level semantic feature detection problem through the improved analytical network, and the location and scale of the face are predicted by the activation map to avoid additional parameter settings. A large number of simulation experiment results show that our proposed method is superior to other comparison algorithms for the accuracy of occlusion face detection and recognition on the face database. In addition, our proposed method achieves a better balance between accuracy and speed, which can be used in the field of security surveillance. However, the performance of the proposed algorithm is sensitive to parameters, and its generalization is not high. How to improve this problem will be more conducive to the model applied to other scenarios or data.

Data Availability

All the data used to support the findings of this study are available within the article.

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

This work was financially supported by the key project of Education Bureau of Guangdong Province (Exploring the

Training and Practice of BeiDou + Intelligent Logistics Innovative Talents (2018CSLKT3-107) and Application of BeiDou + Blockchain in road transportation (2018GkQNCX072)).

References

[1] A. J. Colmenarez and T. S. Huang, "Face detection and recognition," *NATO ASI Series F Computer and Systems Sciences*, vol. 11, no. 2, pp. 208–218, 1998.

[2] T. Kondo and H. Yan, "Automatic human face detection and recognition under non-uniform illumination," *Pattern Recognition*, vol. 32, no. 10, pp. 1707–1718, 1999.

[3] L. H. Koh, S. Ranganath, and Y. V. Venkatesh, "An integrated automatic face detection and recognition system," *Pattern Recognition the Journal of the Pattern Recognition Society*, vol. 35, no. 6, pp. 1259–1273, 2002.

[4] S. Chaudhry and R. Chandra, "Face detection and recognition in an unconstrained environment for mobile visual assistive system," *Applied Soft Computing*, vol. 53, pp. 168–180, 2017.

[5] M. H. Siddiqi, R. Ali, A. M. Khan, E. S. Kim, G. J. Kim, and S. Lee, "Facial expression recognition using active contour-based face detection, facial movement-based feature extraction, and non-linear feature selection," *Multimedia Systems*, vol. 21, no. 6, pp. 541–555, 2015.

[6] S. Zhang, X. Zhu, Z. Lei, X. Wang, H. Shi, and S. Z. Li, "Detecting face with densely connected face proposal network," *Neurocomputing*, vol. 284, pp. 119–127, 2018.

[7] H. Ling, J. Wu, J. Huang, J. Chen, and P. Li, "Attention-based convolutional neural network for deep face recognition," *Multimedia Tools and Applications*, vol. 79, no. 9-10, pp. 5595–5616, 2020.

[8] H. Wang, D. S. Zhang, and Z. H. Miao, "Face recognition with single sample per person using HOG–LDB and SVDL," *Signal Image & Video Processing*, vol. 13, no. 19, 2019.

[9] S. L. Corrow, A. Albonico, and J. J. S. Barton, "Diagnosing prosopagnosia: the utility of visual noise in the cambridge face recognition test," *Perception*, vol. 47, no. 3, pp. 330–343, 2018.

[10] A. Abbad, O. Elharrouss, K. Abbad, and H. Tairi, "Application of meemd in post-processing of dimensionality reduction methods for face recognition," *Iet Biometrics*, vol. 8, no. 1, pp. 59–68, 2019.

[11] S. Madhavan and N. Kumar, "Incremental methods in face recognition: a survey," *Artificial Intelligence Review*, vol. 284, no. 5, pp. 119–127, 2019.

[12] Y. Su, Z. Liu, and M. Wang, "Sparse representation-based face recognition against expression and illumination," *IET Image Processing*, vol. 12, no. 5, pp. 826–832, 2018.

[13] L. I. Yan, S. Shan, R. Wang, Z. Cui, and X. Chen, "Fusing magnitude and phase features with multiple face models for robust face recognition," *Frontiers of Computer Science*, vol. 12, no. 6, 2018.

[14] Y. Zhang, K. Shang, J. Wang, N. Li, and M. M. Y. Zhang, "Patch strategy for deep face recognition," *IET Image Processing*, vol. 12, no. 5, pp. 819–825, 2018.

[15] Y. Jiang, F.-L. Chung, S. Wang, Z. Deng, J. Wang, and P. Qian, "Collaborative fuzzy clustering from multiple weighted views," *IEEE Transactions on Cybernetics*, vol. 45, no. 4, pp. 688–701, 2015.

[16] Y. Zhang, Y. Huang, S. Yu, and L. Wang, "Cross-view gait recognition by discriminative feature learning," *IEEE Transactions on Image Processing*, vol. 99, pp. 1001–1015, 2019.

[17] L. Liu, P. Fieguth, G. Zhao, M. Pietikäinen, and D. Hu, "Extended local binary patterns for face recognition," *Information Sciences*, vol. 24, no. 5, pp. 25–37, 2016.

[18] H. Shao, S. Chen, J. Zhao, W. Cui, and Y. U. Tianshu, "Face recognition based on subset selection via metric learning on manifold," *Frontiers of Information Technology & Electronic Engineering*, vol. 16, no. 12, pp. 102–118, 2015.

[19] A. K. Bobak, A. J. Dowsett, and S. Bate, "Solving the border control problem: evidence of enhanced face matching in individuals with extraordinary face recognition skills," *PLoS One*, vol. 11, no. 2, Article ID e0148148, 2016.

[20] X. Liu, M. Kan, W. Wu, S. Shan, and X. Chen, "Viplfacenet: an open source deep face recognition sdk," *Frontiers of Computer Science*, vol. 11, no. 2, pp. 208–218, 2017.

[21] A. Rikhtegar, M. Pooyan, and M. T. Manzuri-Shalmani, "Genetic algorithm-optimised structure of convolutional neural network for face recognition applications," *IET Computer Vision*, vol. 10, no. 6, pp. 559–566, 2016.

[22] J. Zhao, Y. Lv, Z. Zhou, and F. Cao, "A novel deep learning algorithm for incomplete face recognition: low-rank-recovery network," *Neural Networks: The Official Journal of the International Neural Network Society*, vol. 94, pp. 115–124, 2017.

[23] Y. Cheng, Z. Li, L. Jiao, H. Lu, and X. Cao, "Enhanced retinal modeling for face recognition and facial feature point detection under complex illumination conditions," *Journal of Electronic Imaging*, vol. 25, no. 4, Article ID 043028, 2016.

[24] M. S. Sarfraz and R. Stiefelhagen, "Deep perceptual mapping for cross-modal face recognition," *International Journal of Computer Vision*, vol. 96, no. 8, p. 125058, 2016.

[25] X. Dong and H. Zhang, "Weighted neighbor sparse subspace based collaborative representation for face recognition," *Journal of Computational and Theoretical Nanoscience*, vol. 14, no. 4, pp. 1906–1913, 2017.

[26] S. Lou, X. Zhao, Y. Chuang, H. Yu, and S. Zhang, "Graph regularized sparsity discriminant analysis for face recognition," *Neurocomputing*, vol. 173, no. P2, pp. 290–297, 2015.

[27] H. Patil, A. Kothari, and K. Bhurchandi, "Expression invariant face recognition using semidecimated dwt, patch-ldsmt, feature and score level fusion," *Applied Intelligence*, vol. 44, no. 4, pp. 913–930, 2015.

[28] H. Shi, X. Wang, D. Yi, Z. Lei, X. Zhu, and S. Z. Li, "Cross-modality face recognition via heterogeneous joint bayesian," *IEEE Signal Processing Letters*, vol. 24, no. 1, pp. 81–85, 2017.

[29] G. F. Lu, Y. Wang, and J. Zou, "Graph maximum margin criterion for face recognition," *Neural Processing Letters*, vol. 44, no. 2, pp. 1258–1268, 2015.

[30] H. Ran, W. Xiang, S. Zhenan, and T. Tieniu, "Wasserstein CNN: learning invariant features for NIR-VIS face recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 41, no. 7, pp. 1761–1773, 2018.

[31] S. Karahan, M. K. Yildirim, K. Kirtac, F. S. Rende, and H. K. Ekenel, "How image degradations affect deep CNN-based face recognition?" in *Proceedings of the 2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, vol. 12, no. 14, pp. 6–23, Darmstadt, Germany, September 2016.

[32] J. Wang and Z. Li, "Research on face recognition based on CNN," in *Proceedings of the IOP Conference*, pp. 170–177, Melbourne, Australia, September 2018.

[33] M. Matsugu, K. Mori, and T. Suzuki, "Face recognition using svm combined with CNN for face detection," *Lecture Notes in Computer Science*, vol. 253, no. 251, p. 1, 2004.

[34] L. Jing, Q. Tao, W. Chang, X. Kai, and W. Fang-Qing, "Robust face recognition using the deep C2D-CNN model based on decision-level fusion," *Sensors*, vol. 18, no. 7, pp. 2080–2093, 2018.

[35] Y. X. Yang, C. Wen, K. Xie, F. Q. Wen, G. Q. Sheng, and X. G. Tang, "Face recognition using the SR-CNN model," *Sensors*, vol. 18, no. 12, p. 1, 2018.
[36] A. Rikhtegar, M. Pooyan, and M. T. Manzuri-Shalmani, "GA-optimized structure of cnn for face recognition applications," *IET Computer Vision*, vol. 10, no. 6, pp. 559–566, 2016.
[37] B. Samik and D. Sukhendu, "Mutual variation of information on transfer-CNN for face recognition with degraded probe samples," *Neurocomputing*, vol. 310, pp. 299–315, 2018.
[38] Z. Lu, X. Jiang, and A. Kot, "Feature fusion with covariance matrix regularization in face recognition," *Signal Processing*, vol. 144, pp. 296–305, 2018.