Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 5591020, 11 pages
https://doi.org/10.1155/2021/5591020

Research Article

Masked Face Recognition Algorithm for a Contactless Distribution Cabinet

GuiLing Wu

Xinyang Agriculture and Forestry University, Xinyang 464000, China

Correspondence should be addressed to GuiLing Wu; wgl@xyafu.edu.cn

Received 2 March 2021; Accepted 10 May 2021; Published 22 May 2021

Academic Editor: Yandong He

A contactless delivery cabinet is an important courier self-pickup device, for the reason that COVID-19 can be transmitted by human contact. During the pandemic period of COVID-19, wearing a mask to take delivery is a common application scenario, which makes the study of masked face recognition algorithm greatly significant. A masked face recognition algorithm based on attention mechanism is proposed in this paper in order to improve the recognition rate of masked face images. First, the masked face image is separated by the local constrained dictionary learning method, and the face image part is separated. Then, the dilated convolution is used to reduce the resolution reduction in the subsampling process. Finally, according to the important feature information of the face image, the attention mechanism neural network is used to reduce the information loss in the subsampling process and improve the face recognition rate. In the experimental part, the RMFRD and SMFRD databases of Wuhan University were selected to compare the recognition rate. The experimental results show that the proposed algorithm has a better recognition rate.

1. Introduction

The contactless delivery scene is becoming increasingly normal, since COVID-19 can be transmitted by contact. In the scene of contactless express delivery, a self-pickup cabinet is an important piece of terminal express delivery equipment, and face recognition is one of the effective way to achieve contactless express delivery. During the pandemic of COVID-19, nearly everyone wears masks, which makes the traditional face recognition technology almost ineffective on face recognition self-pickup cabinets. Therefore, it is greatly urgent to improve the performance of existing face recognition technology for face mask recognition.

In recent years, face recognition technology has made great breakthroughs in both theoretical progress and practical applications. It has become a frontier research direction in the field of pattern recognition. However, the recognition problem of occluded face images, such as mask, hairstyle, sunglasses, and hat occlusions, often appears in the process of face processing. These occlusions greatly interfere with the correct recognition of human faces. The low-rank representation [1], however, can quickly solve the occlusion problem. A new iterative method is used to effectively improve the recognition rate and the robustness of large-area image occlusion recognition. Literature [2] proposed a method combining structured occlusion coding and sparse representation-based classifier [2], which cleverly used structured sparse coding to deal with image occlusion problems. In addition, the long and short-term memory network autoencoder [3] was also commonly used to solve the problem of facial occlusion, which better improved the robustness of image noise reduction. However, the occlusion problem has not been completely solved.

Face recognition with occlusion has attracted extensive attention in academic circles. Occlusion processing methods are generally divided into video occlusion processing and image occlusion processing methods. Object tracking is usually used to deal with the occlusion problem in dynamic video. For example, a tracking method based on video monitoring is proposed in literature [4], which can automatically detect and process occlusion objects. Literature [5] proposed a new object tracking technology, which can track

people's dynamic behaviors and actions and maintain pixel tracking allocation even under large shielding. It has been used in different experiments of indoor human behavior supervision due to its robustness. The methods for processing occluded images can be divided into five categories: low-rank representation, image restoration, fuzzy analysis, robust principal component analysis, and structural occlusion coding. Literature [6] proposed a robust low-rank representation method to solve the problem of face recognition with occlusion. This method mainly combines robust representation and low error estimation. At present, the multiscale fractal coding and reconstruction method of image restoration has been proposed in literature [7], which has a good effect on texture images and images with large holes. Literature [8] proposed a method based on fuzzy principal component analysis to detect the occlusion area and restore the face area. However, the fuzzy principal component analysis has a large amount of computation and is not ideal for large-area occlusion processing.

Low-rank robust principal component analysis is a mainstream occlusion face feature extraction method, which combines structural occlusion coding with sparse representation classification. Literature [9] proposed a new nonnegative sparse representation method for robust face recognition in large-scale databases. However, this algorithm has a large amount of computation and a complex structure. Literature [10] proposed an occlusion dictionary method, which plays an increasingly important role in face recognition and can effectively deal with various occlusion objects. It can distinguish the features of nonoccluded and occluded regions and encode the corresponding parts of the dictionary, respectively. The occlusion problem is resolved by using detection/mask scores in literature [11]. And it introduced a plug-and-play occlusion handling algorithm to deal with the occlusion between different object instances. Occlusion face recognition problem is the most critical step towards the practical face recognition technology.

Face recognition systems must solve the problem of occlusion. Objects like hats, scarves, and sunglasses are very common. Sometimes, a large area of occlusion will seriously destroy the information of the original image, resulting in the failure of large image recognition. Literature [12] proposed the method of sparse error and graphical model to continuously overlap and finally display the mask. The Markov random field model is transformed into the calculation of the sparse representation of the training image, so as to find the occlusion area accurately. Therefore, great research significance lies in how to separate the blocked face images from blocking images.

A masked face recognition algorithm based on attention mechanism is proposed in the paper, in order to improve the face recognition rate of masks. First, the masked image is separated by local constrained dictionary learning method; that is, the mask and the face image are separated. Then, dilated convolution is adopted to reduce the problem of resolution reduction in the process of subsampling, and an attention mechanism is designed to reduce information loss in the process of subsampling. Finally, according to the important feature information of the face image, the attention mechanism neural network algorithm is used for face recognition.

2. Materials and Methods

2.1. Relevant Work. In recent years, researchers have been devoted to the study of effective face recognition with occlusion [10] and proposed many effective algorithms. It mainly includes three methods: the generation model of the occlusion problem, the discriminant model, and robust feature extraction. These three methods will be introduced in the following paragraphs.

2.1.1. Generation Model. In the robust learning process of occluded face image data, dictionary atom, low-rank structure of face image, and occluded structure are used to represent the occluded part of the image to be recognized. The local feature loss in face image is processed by reconstruction of noise such as occlusion to correct its influence on recognition performance. By fully understanding the content of the occlusion part, the recognition efficiency is improved. Generating models mainly include robust subspace regression and robust structured error coding.

Robust subspace regression is established by projecting the high-dimensional feature data of different categories of face images into the low-dimensional subspace in a linear or nonlinear way. Then, an independent subspace is established for the occlusion part, and the existing dictionary atoms are used to represent the occlusion in the face image, which can achieve robust recognition effect for the occlusion face. At present, robust subspace solutions to occlusion face recognition mainly include sparse representation, collaborative representation, and occlusion dictionary learning.

(1) *Sparse Representation.* The study in [12] first applies the sparse representation to the field of face recognition and proposes a method based on sparse representation classification (SRC). The sparsity nonzero principle is used to select the most appropriate sparse matrix to represent the image to be recognized more flexibly and comprehensively. Thus, face image classification can be achieved, and the error caused by occlusion and damage can be handled uniformly.

(2) *Collaborative Representation Classification.* The essence is to use training samples from all categories to jointly represent the image to be recognized. The study in [13] through the analysis of the SRC method can effectively enhance the capacity of classification difference, on the basis of classification based on collaborative said method for face recognition.

(3) *Occlusion Dictionary Learning.* The purpose is to learn a new set of dictionaries from the original training samples, which can well represent the ability of the original training samples. Then, it is used for image processing and classification. The study in [14] uses low-rank matrix restoration to train relatively clean face images from face images to be recognized

as a new feature dictionary. Then, the feature dictionary is learned using Fisher's criterion dictionary learning method. Ensure that subdictionaries of the pending class in the new dictionary are well represented for samples in the same category, but not for samples in other categories. This method can effectively reduce the reconstruction error and improve the performance of face recognition with occlusion.

Because of the spatial continuity and locality of occlusion, the error caused by occlusion region has its specific spatial structure, and occlusion will destroy the low-rank structure of face image. Therefore, it is crucial to improve the performance of occluded face recognition to effectively and accurately reconstruct the low-rank structure of the face image from the data damaged by occlusion. And robust structured error coding is to use the low-rank structure of face image to recognize face. Face images in natural environments are affected by different kinds of occlusion noises, leading to great differences between actual low-rank structures of face images and low-rank structures processed by PCA. In order to improve the robustness of occlusion face recognition, the study in [15] proposed robust principal component analysis (RPCA). After decomposition of all training sample matrix Y by low-rank matrix, low-rank content matrix Z and sparse content matrix E were obtained. Thus, the recovery of the low-rank subspace of the training sample is realized. The method takes into account how to recover the low-rank structure from the training samples with large errors but sparse structure, so the effect of sparse noise is effectively suppressed and has strong robustness. To increase the interclass information between low-rank matrices of different categories of faces, literature [16] expressed all training samples as an observation matrix D. After matrix D is decomposed, the low-rank matrix A without occlusion and the sparse error matrix E are obtained. RPCA is applied to the low-rank matrix A, and the subspace obtained is used as the occlusion dictionary of face images. Then, the image reconstruction was identified and the error size was classified according to the sparse representation classification and occlusion dictionary.

2.1.2. Discriminant Model. There are two error indexes that are used as the main discriminant model to estimate the occlusion location. One is the local similarity error between the occlusion image and the original image, and the other is the spatial local error caused by the occlusion. The occluded face image is regarded as the Mosaic of occluded area and unoccluded area, and the unoccluded area is given a larger weight to code. In the process of recognition, the occlusion area may be directly discarded or the image reconstruction may be carried out according to the occlusion area. The key consideration is how to accurately detect the occlusion position, without understanding the content of the occlusion area, so as to eliminate or suppress the impact of occlusion on face recognition. Compared with the generation model, the discriminant model can save a lot of reconstruction time and avoid the introduction of new noise and other problems during reconstruction. The discriminant models mainly include error weight measurement based on local similarity and occlusion error support estimation.

The error weight measurement based on local similarity is to obtain the error information of the occluded image and its original image by comparing the local similarity. The error is given different weights to measure the occlusion position. When human eyes judge the similarity of two images, they only judge their similar areas and directly ignore the contents of nonsimilar areas. For the recognition of occluded face image, the local similarity between the occluded image and its original image is compared first. Then, the error information obtained is weighted with different weights to code the error, and the occlusion location is estimated. Thus, the occluded area can be detected and only the unoccluded area can be identified. The error weight measurement model based on robust sparse coding and the error weight measurement model based on correlation entropy are used to measure the occlusion region through the local similar errors of the two images. Then, the weight of occlusion features is allocated adaptively to suppress or eliminate the influence of occlusion on the performance of face recognition. However, both of these algorithms have some shortcomings. For example, the two algorithms give different weights adaptively according to the error caused by occlusion, but neither of them provides any technology to guarantee that the error caused by occlusion must be large. Moreover, a new noise is introduced in the iterative weighting process, which reduces the efficiency of occlusion face recognition.

Occlusion error support estimation is used to estimate the occlusion position for the spatial local error structure caused by occlusion. In practical face recognition applications, the original images with occluded images may not always exist, so most methods need to reconstruct the original unoccluded images for the occluded face images to be recognized. This will lead to uncertain reconstruction errors in the reconstructed image, resulting in poor detection effect of occluded areas.

Therefore, in order to accurately and directly detect the occlusion region, it is necessary to make full use of the spatial local continuity of occlusion error.

2.1.3. Robust Feature Extraction. Robust face image feature extraction is low-order function, such as color, texture, brightness and expression, age, and gender. Such multiscale decomposition features more orientation to avoid interference between various features. To extract image features, suppress or eliminate the impact of obstacle features on recognition performance and achieve robust recognition effect. Existing robust feature extraction methods for occlusion face recognition mainly include "shallow" feature extraction and "deep" feature extraction.

Shallow robust feature extraction is the traditional robust feature extraction method applied to the occlusion case. The most effective and representative features unrelated to occlusion are extracted, so as to achieve robust feature extraction from occlusion face images. Based on image

gradient direction (IGD), feature extraction shaded face images can be measured between the relationships. In order to solve the actual occlusion face recognition problem, literature [17] proposed a low-order model based on adaptive sparse gradient direction. The generalized gradient direction is used to extract effective features and enhance the continuity of the model to maintain the robustness of face recognition. Literature [18] proposed the subspace learning framework of image gradient direction. The occluded test samples and training samples are mapped into the gradient face space. The image gradient feature is robust to the image noise in the gradient face space. It can extract robust face features without the effect of occlusion to a large extent. Mapping the test samples into the image gradient PCA subspace will result in a reconstructed image with almost no occlusion. Thus, robust feature extraction can be realized for occluded face images, and the extracted features can be used for recognition and classification. This method is implemented under the condition that the difference of the occlusion region of two completely different images approximately obeying the uniform distribution. In practice, the difference between the occluded image and its original image is not uniformly distributed. This method is not suitable for face image recognition with arbitrary occlusion because of its poor performance in the real environment.

Deep learning has the characteristics of automatic learning features, and the extracted "deep" features are more expressive and stable than those designed manually. Therefore, deep learning is used to acquire more abstract and expressive deep features. It is expected that the abstract semantic information of data can be represented by multilevel high-order features to obtain robust features. It was considered a good way to overcome the limitations of artificially designed "shallow" features. A robust "depth" feature should both minimize intraclass differences and maximize interclass differences in the image. In order to identify whether multiple face images belong to the same person, the DeepID2 neural network structure was proposed in literature [19]. The convolutional neural network is used for feature learning. Different face features are extracted through face recognition signals, and the interclass differences between different face images are increased. And through the face verification signal to extract the features of the same face, reduce the intraclass differences, so as to learn the strong ability to distinguish features. However, literature [20] adopted forward and backward greedy algorithms to select some effective complementary DeepID2 vectors due to the large number of features to be learned. The features after dimensionality reduction are input into the joint Bayesian model for face classification and recognition. Although the network structure is not designed to distinguish between occluded and nonoccluded faces, the features of deep learning are adaptively robust to occluding. In order to further extract better facial features, literature [21] improved the DeepID2 network and proposed the DeepID2+ network structure. Supervised signals are added in each convolution layer, and the 512-dimension features of the final output are binarized by using threshold values. Therefore, it not only ensures the accuracy of recognition but also improves the speed of face retrieval. The results show that DeepID2+ is robust to mask human face from bottom to top and black blocks of different sizes. When the occlusion is less than 20% and the block is less than 30×30, the validation accuracy of the output of DeepID2+ is almost unchanged. Furthermore, it provides a new way to deal with the occlusion face recognition.

2.2. The Proposed Algorithm

2.2.1. Mask Separation with Locality Constraint Dictionary Learning Method. Local constraint dictionary (LCD) means that, given the dictionary Q_H, x_i can be approximately represented by a linear combination of the atoms q_j in the dictionary; that is, $x_i = \sum_{j=1}^{k} c_{ji} q_j$. The embedded $y_i = p(x_i)$ in the d-dimensional space can be represented by the insertion of q_j in a low-dimensional space of $p(q_i)$ nearly linear; that is, $x_j = \sum_{j=1}^{k} c_{ji} p(q_j)$. According to the l_2 distance, in order to minimize the errors of the above two linear representations, the following two expressions should be minimized simultaneously with respect to Q_H and $C = [c_1, c_2, \ldots, c_N]$.

$$\sum_{j=1}^{N} \left\| p(x_i) - \sum_{j=1}^{K} c_{ji} p\left(q_j\right) \right\|^2, \quad \sum_{j=1}^{N} \left\| x_i - \sum_{j=1}^{K} c_{ji} q_j \right\|^2. \tag{1}$$

To ensure choice invariance, constraints $\forall_i, \sum_{j=1}^{K} c_{ji} = 1$, $\|c_i\|_0 = \tau$. Furthermore, if q_j is not in the τ neighborhood of $x_i (i = 1, 2, \ldots, N)$, c_{ij} is equal to 0. If q_j is in the neighborhood of x_i, then $c_{ji} \neq 0$. The corresponding function is shown as

$$\min_{C, Q_N} \left\| X - Q_H C \right\|_F^2 + \lambda \sum_{i=1}^{N} \sum_{j=1}^{K} c_{ji}^2 \left\| x_i - q_j \right\|^2 + \mu \|C\|_F^2, \tag{2}$$

where $I^T c_i = 0$, $c_{ji} = 0$, $\forall_i = 1, 2, \ldots, N$, and $q_j \notin \Omega_\tau(x_i)$. And i is the column vector with 1 entry. $\Omega_\tau(x_i) \in R^{n \times \tau}$ is the τ neighborhood of x_i, which contains τ nearest neighbors of x_i. λ is the parameter, $\lambda < 0$.

The form of a face image y blocked by an occlusion object v is $u = y + v$, which violates the low-dimensional linear lighting model and leads to SOC classification error. Occlusion categories in real scenes are predictable and can be collected in advance. Inspired by literature [22], the algorithm in this paper constructs a subdictionary of occlusion object $B = [B_1, B_2, \ldots, B_S]$, and different subscripts represent different categories. Then, add it to the original dictionary as $R = [Q, B]$. The mask v belonging to class t is then represented by the corresponding subdictionary V_t.

Specifically, if a face of the r-th object is occluded by the occlusion of the t-th category, then u is represented linearly by Q_r and B_t as $u = y + v = Q_r X_r + B_t C_t$. By looking for appropriate sparse solutions, a series of coefficients, whose nonzero terms represent faces and occluded objects, can be obtained. The specific form is as follows:

$$W = \left[x^T, c^T\right]^T = \left[0, \ldots, 0, x_{r,1}, x_{r,2}, \ldots, x_{r,n}, 0, \ldots, 0, c_{r,1}, c_{r,2}, \ldots, c_{r,n_i}, 0, \ldots, 0\right]^T. \tag{3}$$

It can be seen that a well-constructed dictionary can deal with all kinds of occlusion effectively and is highly robust in the actual scene. The equation for occlusion identification problem is as follows:

$$\begin{aligned} &\widehat{\omega}_0 = \arg\min_w \|w\|_0 = \arg\min_{x,c} \|x; c\|_0, \\ &\text{s.t.} \quad \left\| u - [Q, B] \begin{bmatrix} x \\ c \end{bmatrix} \right\|_2 = \|u - Rw\|_2 \le \varepsilon. \end{aligned} \tag{4}$$

When the data are gradually increasing, it is not feasible to solve problem (4), so l_1 normal form is used to replace Al_0 normal form in problem (4), so problem (4) becomes the following problem:

$$\begin{aligned} &\widehat{\omega}_1 = \arg\min_w \|w\|_1 = \arg\min_{x,c} \|x; c\|_1, \\ &\text{s.t.} \quad \left\| u - [Q, B] \begin{bmatrix} x \\ c \end{bmatrix} \right\|_2 = \|u - Rw\|_2 \le \varepsilon. \end{aligned} \tag{5}$$

Calculate the error between $\widehat{y}$ and $\widehat{y}_i$, where $\widehat{y}_i = Q\delta_i(\widehat{x})$. Even with occlusion, the minimum residual difference can be allocated to sample u. It makes up the occlusion dictionary and the clean face dictionary. The two dictionaries obtained are of considerable use for subsequent occlusion work, as Figure 1 shows. This separates the mask from the face. Compared with the original face, the pixel value of the separated face does not decrease, and the face dictionary composed is more conducive to the subsequent classification. In this paper, the recognition results are presented in a subsequent experiment.

2.2.2. Face Recognition Algorithm Using Attention Mechanism. The overall network structure proposed in this paper uses two paths for secondary sampling, rather than the traditional U-shaped structure. Dilated convolution is used to obtain more detailed information. ResNet is used to extract features from feature context information, so as to obtain a larger sensory field of vision. Meanwhile, attention modules are designed to improve accuracy by imitating human visual mechanism [23]. Finally, a feature fusion module is designed to integrate the collected features of different sensory fields to obtain better results.

(1) Dilated Convolution. In order to improve the resolution while maintaining a fixed field of view, the algorithm in this paper uses dilated convolution in the spatial information path. The expression of ordinary convolution is as follows:

$$P(x, y) \cdot M(x, y) = \sum_{i=0}^{\omega} \sum_{j=0}^{m} K(i, j) \cdot P(x - i, y - j). \tag{6}$$

In the equation, $P(x, y)$ is the pixel value of the original image at the point (x, y) and $M(x, y)$ is the convolution kernel multiplied by it with the size of $\omega \times m$.

The dilated convolution is calculated as follows:

$$P(x, y) \cdot M'(x, y) = \sum_{i=0}^{\omega} \sum_{j=0}^{m} M'(i, j) \cdot P(x - l \times i, y - l \times j), \tag{7}$$

where l is the expansion factor and $M'(x, y)$ is the dilated convolution kernel. It can be seen from equations (6) and (7) that the dilated convolution is essentially 0 filling of the convolution kernel. It can increase the perception field of the convolution kernel, while retaining the original pixel information, thus increasing the resolution. If the size of the convolution kernel is k and the expansion rate is l, then the actual effective size of the dilated convolution is $k + (k - 1) \times (l - 1)$. Compared with ordinary convolution of the same size, expansive convolution not only expands the perception field but also maintains the same resolution as ordinary convolution.

(2) Attention Mechanism. An attention mechanism is adopted in this paper in order to make up for the loss of details caused by subsampling and better guide model training. It can enhance the target features and suppress the background through the weighted processing of feature map [24]. The target features here are the contour and texture information of the eyes, eyebrows, and face due to the features of the face of the mask. The attention mechanism is mainly composed of spatial attention mechanism, channel attention mechanism, and pyramid attention mechanism, as shown in Figure 2.

(3) Spatial Attention Mechanism. The mechanism of spatial attention is mainly put forward by imitating human visual mechanism. When the human eye sees an image, it automatically gives greater attention to key locations. When you see a rabbit, for example, you pay more attention to the rabbit's ears. Therefore, different parts of the image feature map should have different weights [25]. The spatial attention mechanism proposed in this paper is shown in Figure 3.

The resulting feature graph uses maximum and mean global pooling on the channel dimensions. The feature information of different positions on the feature map is compared and extracted, and the feature weight of each position is obtained.

The maximum global pooling expression is

$$P(i, j) = \max\left(I_t(i, j)\right), \quad t \in (1, c), \tag{8}$$

where $I_t(i, j)$ is the eigenvalue of the feature graph on channel t-th at position (i, j), $P(i, j)$ is the feature graph after maximum pooled, and c is the number of channels.

The global pooling of the mean value is expressed as

$$Q(i, j) = \frac{1}{d} \sum_{n=1}^{d} T_n(i, j). \tag{9}$$

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

Figure 1: The example of mask separation. (a) Mask face image. (b) Reference image. (c) Separated mask. (d) Separated face.

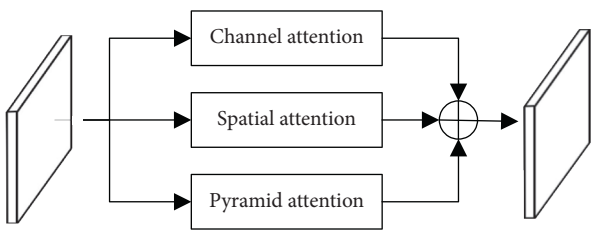

Figure 2: Structure of attention mechanism.

In the equation, $T_n(i, j)$ is the eigenvalue of the nth channel feature graph at the position (i, j). $Q_n(i, j)$ is the feature graph after averaging pooling, and d is the number of channels.

The two feature maps are connected in channel dimension by means of channel connection. In order to better integrate the information extracted by the two methods, the size of 1×1 convolution kernel is used for learning. The final weight of attention of the obtained feature map was calculated by sigmoid function [26], so as to avoid errors caused by excessive weight coefficient.

The sigmoid function is expressed as follows:

$$\mathrm{SF}(x) = \frac{1}{1 + e^{-x}}, \tag{10}$$

where SF (x) is the response of the output and x is the input.

Spatial attention mechanism can effectively extract salient information of each position in the feature map. Based on this, a weight is assigned to the eigenvalue of each position to extract the main feature of the target effectively.

(4) Channel Attention Mechanism. Each channel of the feature map extracted by convolutional neural network (CNN) represents an image feature, such as texture and shape. The target features here are the contour and texture information of the eyes, eyebrows, and face due to the features of the face of the mask. In the image, each feature contains different information, and its contribution to image segmentation is also different. Therefore, different attention should be paid to each different feature [27] and different weights should be assigned. Channel attention mechanism is designed to assign weight to features so that the network can focus on important features, as shown in Figure 4.

By using global maximum pooling and global average pooling, the channel information is modeled in spatial dimension and the characteristic information of each channel is obtained.

The calculation process of maximum pooling is

$$Q_c = \max\left(I_c(i, j)\right). \tag{11}$$

In the equation, $i \in (1, h)$, $j \in (1, \omega)$, c represents c-th feature graph, and Q_c represents the output of the feature graph after maximum pooling.

The calculation process of average pooling is

$$Q_t = \frac{1}{\omega}\frac{1}{h}\sum_{i=1}^{t}\sum_{j=1}^{\omega} I_t(i, j). \tag{12}$$

In the equation, $i \in (1, h)$, $j \in (1, \omega)$, t represents the m-th feature graph, and Q_t represents the output of the feature graph after average pooling.

In order to use less computation to integrate the feature graphs obtained by global pooling, the two feature graphs are, respectively, passed through a convolution kernel of size 1×1. Then, nonlinear components are added through BN layer and ReLU layer to make the model fit better. This method can prevent the occurrence of overfitting phenomenon to a certain extent. Finally, the two feature graphs obtained are fused and the final weight is obtained through the sigmoid function.

(5) Pyramid Attention Mechanism. Human vision tends to integrate a variety of information when discriminating objects. For example, distinguishing a rabbit from a cat pays more attention to the shape of its ears, and distinguishing a panda from a bear pays more attention to its color. It can be seen that different features in different positions on the feature map should receive different attention. In order to obtain different information of different position of image better, a pyramid type attention model is proposed. By extracting the feature map of different perceptual field, the image information under different perceptual field is obtained. This information was fused to obtain the final weight coefficient, as shown in Figure 5.

The feature graphs are, respectively, passed through 3×3, 1×1, and 5×5 convolution kernels. Because these convolution kernels are used in low-resolution high-level feature maps, there is not much computational burden. Then the feature images are collected by convolution kernels of different sizes. In this way, contextual information can be better integrated and feature information can be obtained at different scales. Through these convolution kernels of different sizes, the characteristic information under different perceptual fields is obtained. Based on the above feature

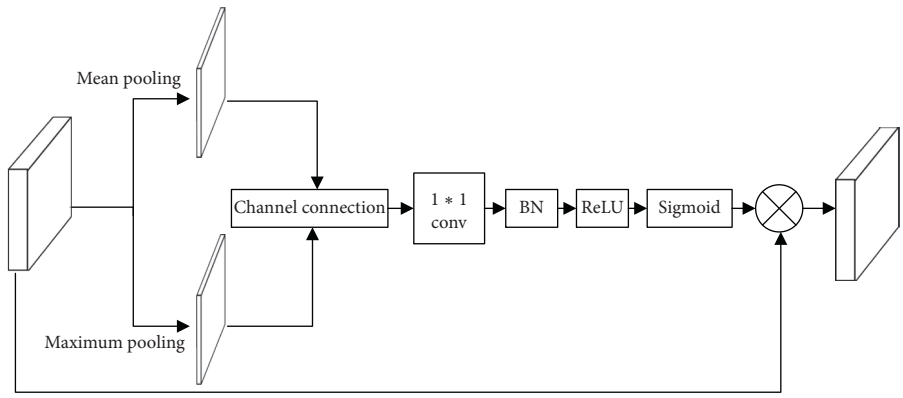

Figure 3: Spatial attention mechanism.

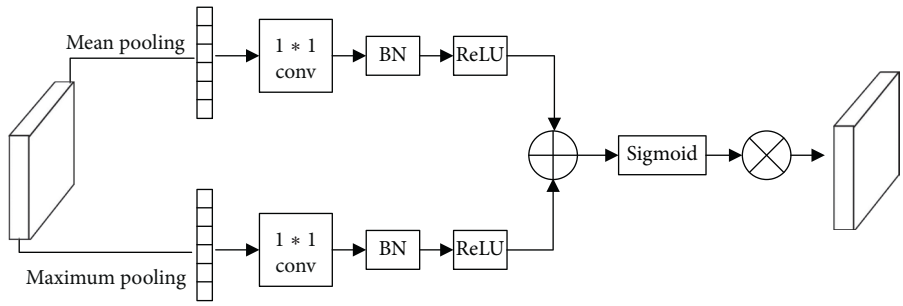

Figure 4: Channel attention mechanism.

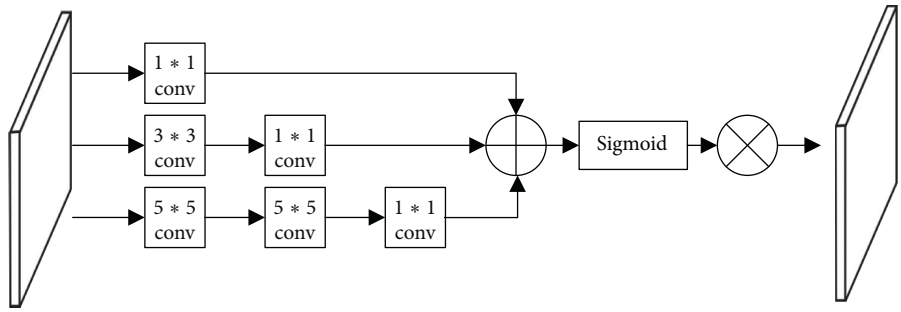

Figure 5: Pyramid attention mechanism.

information, the feature weights of different positions are obtained through a 1×1 convolution kernel. Finally, the obtained feature graphs are added for fusion, and the final weight is obtained through the sigmoid function.

(6) Feature Fusion Module. For different types of objects, the importance of vision perception is different. For larger objects, the features acquired from the larger visual field are more important. For smaller objects, the features of larger visual field will collect too much peripheral information and lead to errors. The traditional feature fusion methods are generally cascade or addition, which does not take into account the different sensory field of different feature graphs and ignores the specificity of features. In view of this, the feature fusion module designed in this paper assigns different weights to feature images with different perception fields to achieve better feature fusion, as shown in Figure 6.

First, the two input feature maps are linked at the channel dimension level. Secondly, the cascaded feature image is fused by a 3×3 convolution kernel to realize the preliminary fusion of feature image information. Global pooling operation is carried out to extract the information of each feature map. Then, the obtained feature image is passed through the convolution kernel of size of 1×1, so that the network learns the weight according to the overall information of each feature image. Finally, the sigmoid function is used to get the final weight, which is multiplied by the original feature graph. Through the feature fusion module,

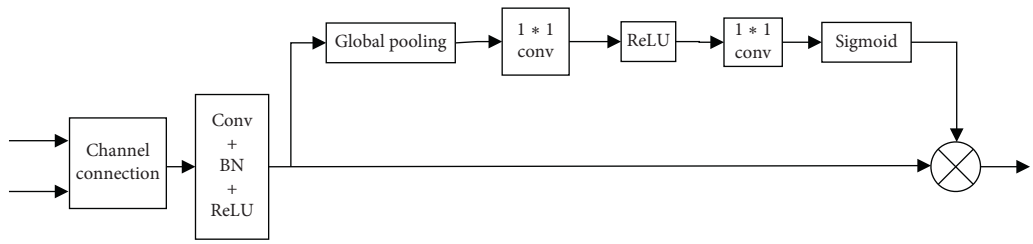

Figure 6: Feature fusion module.

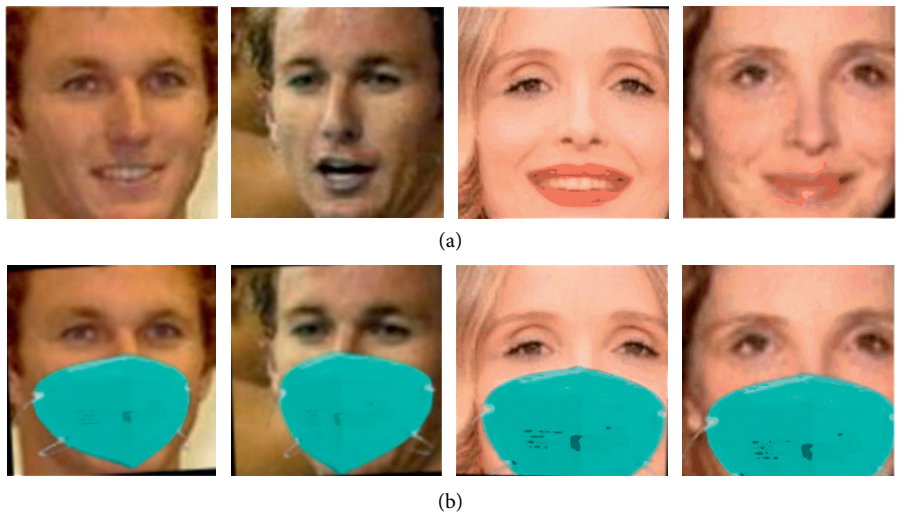

Figure 7: The face images of SMFRD. (a) Face image without mask. (b) Masked face images.

weight is assigned to the feature images under different perceptual fields, so that the feature specificity under different perceptual fields can be reflected, and the features can be fused better.

3. Results and Discussion

In this paper, the RMFRD (real-world face recognition dataset) and SMFRD (simulated face recognition dataset) [28] opened by Wuhan University are selected as the experimental databases. RMFRD and SMFRD are the first real face mask dataset in the world, and the simulated masked image data in SMFRD are based on LFW [29] and Webface [30] datasets. The experiment in this paper is carried out on a 16 GB memory NVIDIA GeForce GTX1080TI GPU workstation.

3.1. Experiment on SMFRD Dataset. The SMFRD database simulates wearing a mask on the existing public large-scale face dataset to obtain a large number of face images. Among them, public large-scale face datasets include commonly used LFW and Webface datasets. In this way, a simulated masked face dataset was constructed, covering 500,000 face images from 10,000 subjects. Figure 7 shows multiple pairs of face images.

In the training process, 5000 people were randomly selected from 10000 people as the training set. Among them, a normal image without a mask was selected as the baseline image for recognition. During the training process, other normal images without masks were first selected for training. Then, all face images wearing masks were selected for training.

In the test phase, the remaining 5000 people were selected as the test set. During the test, a normal image without a mask was selected as the baseline image for recognition. During the test, other normal images without masks were selected for identification tests. Then, all faces wearing masks were selected for recognition tests.

In this paper, the PCA + SVM [31], SRC [32], CNN, and DCGAN + CNN [33] are selected for comparison. In the DCGAN + CNN method, DCGAN is used to fill the occlusion face image, and CNN is also a fine-tuned VGGFACE model for face recognition. All methods were tested on the RMFRD dataset, and the results are shown in Table 1. It can be seen that the expression recognition accuracy of the proposed method in this paper is high, regardless of whether the face image is covered by a mask or not. In the deep learning method, although DCGAN + CNN method also fills the face image, the consistency of the image obtained is poor, which affects the accuracy of expression recognition. The proposed algorithm in this paper has already separated the mask part before face recognition and then extracted the important features such as eyes and eyebrows by using the attention mechanism, so the algorithm in this paper has higher stability.

Table 1: Comparison of recognition rates with different algorithms.

Methods	Image without mask	Image with mask
PCA + SVM [31]	91.22	68.63
SRC [32]	92.80	72.42
CNN [33]	97.52	69.63
DCGAN + CNN [33]	97.59	75.85
Proposed	98.39	95.31

(a) (b) (c)

Figure 8: The images of RMFRD. (a) Masked face images. (b) and (c) Face images without mask.

3.2. Experiment on RMFRD Dataset. The face images of RMFRD dataset come from Internet resources, in which the front image of public figures and their corresponding face mask images are captured by the crawler tool. Then, the unreasonable face images were manually removed and a tagging tool was experimented with to crop the exact face areas. The dataset consisted of 5,000 photos of 525 people wearing masks and 90,000 photos of 525 people not wearing masks. Figure 8 shows multiple pairs of face images.

In the training process, 300 people were randomly selected from 525 people as the training set. Among them, a normal image without a mask was selected as the baseline image for recognition. During the training process, other normal images without masks were first selected for training. Then, all face images wearing masks were selected for training.

In the test phase, the remaining 325 people were selected as the test set. During the test, a normal image without a

Table 2: Comparison of recognition rates with different algorithms.

Methods	Image without mask	Image with mask
PCA + SVM [31]	90.14	67.82
SRC [32]	91.92	72.37
CNN [33]	97.21	68.69
DCGAN + CNN [33]	97.36	75.21
Proposed	98.10	95.22

Table 3: Comparison of time consuming with different algorithms.

Methods	Image without mask (%)	Image with mask (%)
PCA + SVM [31]	0.92	0.93
SRC [32]	1.20	1.20
CNN [33]	0.87	0.89
DCGAN + CNN [33]	1.23	1.25
Proposed	0.91	0.98

mask was selected as the baseline image for recognition. During the test, other normal images without masks were selected for identification tests. Then, all faces wearing masks were selected for recognition tests.

PCA + SVM, SRC, CNN, and DCGAN + CNN algorithms were selected, and experiments were carried out on RMFRD dataset at the same time as the proposed algorithm. The recognition results are shown in Table 2. It can be seen that the expression recognition accuracy of the method presented in this paper is higher than other methods.

3.3. The Algorithm Efficiency. In order to compare the efficiency of the algorithm, the time consumption of the recognition process is compared. The time consumption of different face recognition methods is compared, which is as shown in Table 3. It can be seen from the table that the recognition time of the algorithm in this paper is less than 1 second, which can meet the experiment of actual use scenarios.

4. Conclusions

Since COVID-19 can be spread by contact, contactless delivery cabinets are becoming increasingly normal. The self-delivery cabinet is an important piece of terminal express delivery equipment in the contactless express delivery scenario, and the masked face recognition is one of the effective way to achieve the contactless express delivery. In this paper, the local constrained dictionary learning method firstly is used to separate the masked face image, and the parts of the mask are separated to reduce the impact on the recognition algorithm. Then, the dilated convolution method is used to reduce the impact of resolution reduction in the sampling process. According to the important features of face images, such as eyes and eyebrows, multiple features are extracted by using attention machine neural network. Finally, the algorithm comparison experiment is carried out through the simulation face masked image database and the real face masked image database. Experimental results show that the proposed algorithm has a better recognition rate and also has important application value in contactless express delivery scenarios.

Data Availability

The RMFRD (real-world face recognition dataset) and SMFRD (simulated face recognition dataset) opened by Wuhan University are selected as the experimental databases. The data can be downloaded from https://github.com/X-zhangyang/Real-World-Masked-Face-Dataset. Because of the programmer's reason, the source code of this algorithm is not convenient to provide directly. The readers can program if interested, according to the idea of the paper.

Conflicts of Interest

The author declares no conflicts of interest.

Acknowledgments

This work was supported by the Henan Province Science and Technology Tackling Plan Project (182102210533).

References

[1] J. Qian, J. Yang, F. Zhang, and Z. Lin, "Robust low-rank regularized regression for face recognition with occlusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Columbus, OH, USA, June 2014.

[2] Y. Ouyang, N. Sang, and R. Huang, "Accurate and robust facial expressions recognition by fusing multiple sparse representation based classifiers," *Neurocomputing*, vol. 149, pp. 71–78, 2015.

[3] F. Zhao, J. Feng, J. Zhao, W. Yang, and S. Yan, "Robust LSTM-autoencoders for face de-occlusion in the wild," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 778–790, 2017.

[4] H. Possegger, T. Mauthner, P. M. Roth, and H. Bischof, "Occlusion geodesics for online multi-object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1306–1313, Columbus, OH, USA, June 2014.

[5] L. A. Camuñas-Mesa, T. Serrano-Gotarredona, S. H. Ieng, R. Benosman, and B. Linares-Barranco, "Event-driven stereo visual tracking algorithm to solve object occlusion," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 9, pp. 4223–4237, 2017.

[6] M. Iliadis, H. Wang, R. Molina, and A. K. Katsaggelos, "Robust and low-rank representation for fast face identification with occlusions," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2203–2218, 2017.

[7] Z. Ying, G. Jun, C. Guo, and F. Wengang, "Multi-scale and multi-orientation texture feature extraction method based on fractal theory," *Chinese Journal of Scientific Instrument*, vol. 29, no. 4, p. 787, 2008.

[8] N. Naik, P. Jenkins, and N. Savage, "A ransomware detection method using fuzzy hashing for mitigating the risk of occlusion of information systems," in *Proceedings of the 2019 International Symposium on Systems Engineering (ISSE)*, pp. 1–6, IEEE, Edinburgh, UK, October 2019.

[9] S. Zeng, J. Gou, and L. Deng, "An antinoise sparse representation method for robust face recognition via joint l_1 and l_2 regularization," *Expert Systems with Applications*, vol. 82, pp. 1–9, 2017.

[10] W. Ou, X. You, D. Tao, P. Zhang, Y. Tang, and Z. Zhu, "Robust face recognition via occlusion dictionary learning," *Pattern Recognition*, vol. 47, no. 4, pp. 1559–1572, 2014.

[11] Y. Chen, G. Lin, S. Li et al., "BANet: bidirectional aggregation network with occlusion handling for panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3793–3802, Seattle, WA, USA, August 2020.

[12] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2008.

[13] X. Song, Y. Chen, Z.-H. Feng, G. Hu, T. Zhang, and X.-J. Wu, "Collaborative representation based face classification exploiting block weighted LBP and analysis dictionary learning," *Pattern Recognition*, vol. 88, pp. 127–138, 2019.

[14] H. Zhang, W. He, L. Zhang, H. Shen, and Q. Yuan, "Hyperspectral image restoration using low-rank matrix recovery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 8, pp. 4729–4743, 2013.

[15] Z. Gao, L.-F. Cheong, and Y.-X. Wang, "Block-sparse RPCA for salient motion detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 10, pp. 1975–1987, 2014.

[16] C.-P. Wei, C.-F. Chen, and Y.-C. F. Wang, "Robust face recognition with structurally incoherent low-rank matrix decomposition," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3294–3307, 2014.

[17] Q. Xu and Y. Xu, "Extremely low order time-fractional differential equation and application in combustion process," *Communications in Nonlinear Science and Numerical Simulation*, vol. 64, pp. 135–148, 2018.

[18] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "Subspace learning from image gradient orientations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 12, pp. 2454–2466, 2012.

[19] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: face recognition with very deep neural networks," 2015, https://arxiv.org/abs/1502.00873.

[20] M. Fachrurrozi, A. Wijaya, and M. N. Rachmatullah, "New optimization technique to extract facial features," *IAENG International Journal of Computer Science*, vol. 45, no. 4, 2018.

[21] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2892–2900, Boston, MA, USA, June 2015.

[22] Y. Wen, W. Liu, M. Yang, Y. Fu, Y. Xiang, and R. Hu, "Structured occlusion coding for robust face recognition," *Neurocomputing*, vol. 178, pp. 11–24, 2016.

[23] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in *Proceedings of Advances in Neural Information Processing Systems*, pp. 5998–6008, Long Beach, CA, USA, December 2017.

[24] A. Gilra and W. Gerstner, "Non-linear motor control by local learning in spiking neural networks," in *Proceedings of the International Conference on Machine Learning*, Stockholm, Sweden, July 2018.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, IEEE, Las Vegas, NV, USA, June 2016.

[26] X. Yin, J. A. N. Goudriaan, E. A. Lantinga, J. Vos, and H. J. Spiertz, "A flexible sigmoid function of determinate growth," *Annals of Botany*, vol. 91, no. 3, pp. 361–371, 2003.

[27] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, Munich, Germany, July 2018.

[28] Z. Wang, G. Wang, B. Huang et al., "Masked face recognition dataset and application," 2020, https://arxiv.org/abs/2003.09093.

[29] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: a dataset and benchmark for large-scale face recognition," in *Proceedings of the European Conference on Computer Vision*, pp. 87–102, Springer, Amsterdam, Netherlands, July 2016.

[30] S. Pepin and C. Körner, "Web-FACE: a new canopy free-air CO_2 enrichment system for tall trees in mature forests," *Oecologia*, vol. 133, no. 1, pp. 1–9, 2002.

[31] Y. Luo, C. Wu, and Y. Zhang, "Facial expression recognition based on principal component analysis and support vector machine applied in intelligent wheelchair," *Application Research of Computers*, vol. 29, no. 8, pp. 3166–3168, 2012.

[32] M. Zhu, S. Li, and H. Ye, "An occluded facial expression recognition method based on sparse representation," *Public Relations & Artificial Intelligence*, vol. 27, no. 8, pp. 708–712, 2014.

[33] R. Yeh, C. Chen, T. Y. Lim, and M. Hasegawa-Johnson, "Semantic image inpainting with perceptual and contextual losses," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6882–6890, Honolulu, HI, USA, July 2017.