A Convolutional Neural Network based on TensorFlow for Face Recognition

Liping Yuan[1] ,Zhiyi Qu[1] ,Yufeng Zhao[1] ,Hongshuai Zhang[1] ,Qing Nian[2]
1. School of Information Science & Engineering, Lanzhou University
2. Technical Reconnaissance Detachment of Public Security Bureau Lanzhou
Lanzhou, China
yuanlp15@lzu.edu.cn, quzy@lzu.edu.cn, zhaoyf10@lzu.edu.cn, zhanghsh15@lzu.edu.cn,554506103@qq.com

***Abstract*—Face recognition is a hot research field in computer vision, and it has a high practical value for the detection and recognition of specific sensitive characters. Research found that in traditional hand-crafted features, there are uncontrolled environments such as pose, facial expression, illumination and occlusion influencing the accuracy of recognition and it has poor performance, so the deep learning method is adopted. On the basis of face detection, a Convolutional Neural Network (CNN) based on TensorFlow, an open source deep learning framework, is proposed for face recognition. Experimental results show that the proposed method has better recognition accuracy and higher robustness in complex environment.**

Keywords—face recognition, deep learning, CNN, Tensorflow

I. INTRODUCTION

With the rapid development of information technology, the network has brought great convenience to people, where people are willing to get lots of information. However, the security of information has become increasingly serious. There are many applications including online banking, e-commerce, as well as national security and so on, which are in urgent need of effective identification technology to protect the user's security. In addition, some terrorists spread a lot of bad information in the network, which has brought a very significant harm to the stability and harmony of the country. Compared with the traditional identity authentication technology, the identity authentication technology based on biometrics is more efficient, convenient and safe, which can not only meet the basic security needs of citizens, but also play a vital role in the national information security.

At present, the recognition technology of biometric characteristics mainly uses computer vision, graphics and image processing, pattern recognition and other techniques to extract human characteristics, including physiological characteristics and behavior characteristics. The physiological characteristics that can be detected and verified are fingerprint, iris, retina, facial features, DNA, etc., and behavioral characteristics of voice, gait, handwriting, etc. [1]. Compared with fingerprints and iris, face recognition is a non-contact recognition technology with a higher nature, acceptability and intuition.

When face recognition technology is used to obtain face images, there are no contact, therefore which can be used for covert operation, and widely used in sensitive personnel monitoring. It is more convenient and easily accepted by users when using, and it can record the faces recognized, and used for afterwards such as tracking criminals, personnel attendance and so on, which biometric identification (for example, fingerprint, iris) cannot carry. In the face recognition there is not high requirement for hardware facilities, like digital cameras, cameras and other equipments whose cost is gradually decrease, so the practical space of face recognition is very large.

Face recognition also has some shortcomings. First of all, the face will change with time and expression changes, and some minor changes such as makeup, glasses and other decorations will bring recognition errors causing some uncertainty. Secondly, although everyone's face is a unique, there are still two or more similar faces, which will bring some trouble to identify. Finally, information acquisition is also an insurmountable technical difficulties in the process of recognition, and illumination condition, acquisition angle and the distance between the face and acquisition device will affect the effect of the acquisition, thus affecting the accuracy of identification [2].

In the face of a large number of sensitive images on the Internet, we use the technology of image processing to detect and analyze them. Compared with the existing algorithms of face detection and face recognition, a sensitive person identification system is proposed. Using the latest advanced learning framework, the design of the neural network model is designed to realize the effective face recognition. For sensitive persons, they can use the system to select and identify in a large number of unknown information in the images. And then assist the public security departments and effectively block the spread of bad information, in order to ensure the safety and stability of network information.

II. RELATED WORKS

In the face detection technology, the common face detection methods are divided into feature-based detection methods and statistical model-based detection methods. The former is more mature, but limited to light, expression, pose and other conditions. In addition, the structural feature-based detection algorithm is relatively large, generally cannot meet people's requirements for real-time detection. The statistical model-based detection algorithm is the mainstream of the research direction, in the statistical process using a large number of samples for training, making the results have a certain reliability. Among them, Adaboost algorithm [3] is the

fastest detection method in speed. The clearest description of state information are the algorithms based on Hidden Markov Model (HMM) and texture feature. In our work, we apply Adaboost algorithm to train the cascade classifier to detect and track human faces accurately and in real time, which is benefit for back-end face recognition.

Face recognition technology can be traced back to 18th century, Galton [4] in the "Nature" published an article based on facial information for identification. In general, according to different algorithms of classification, face recognition can be divided into the following directions. Template-based methods, which utilize the correlation between the image gray level and the computer template to carry on face recognition. The methods based on geometric features, use geometric vector to represent the local characteristics of the face and clustering method to design classifier [5]. In the methods based on elastic graph, the grid is used as the template to compare the images to the grid. Although recognition performance is better, the calculation is large and the recognition speed is slow. Based on the HMM, probabilistic estimation and the singular value feature of facial image are transformed into vectors, and the HMM is used for recognition. The Eigenface method based on Principal Component Analysis (PCA), using PCA transform to reduce the original image processing, and then to classify and identify [6]. In the Fisherface method based on Linear Discriminant Analysis (LDA), the features of the image are reduced and then LDA is used to transform and extract the features of the principal components after the dimension reduction. It is expected that large inter-class divergence and small intra-class divergence. The Local Binary Pattern (LBP) method, which is based on local feature extraction, compares the pixel points in the image and the pixels around the image, and then sums up the results according to the comparison result. At the local level of the image, the description of the feature is the purpose of describing the image. Based on the Support Vector Machine (SVM), the generalization ability and the high classification ability are obtained by finding the optimal classification hyperplane. However, the selection of kernel function and its parameters in SVM is more flexible and cannot be controlled. For the method based on neural network, this method makes use of the learning and classification ability of the neural network, and directly takes the image pixels as the input of the network. The output of the network can distinguish which class belongs to avoid the complex feature extraction. Experiments show that the method has strong adaptability and robustness. The recognition performance can be further improved by appropriately increasing in the number of network neurons and the number of training samples.

III. FACE DETECTION

It needs face detection before face recognition, and the methods of face detection include statistical models, Support Vector Machine (SVM) method, as well as elastic graph matching method. In statistical feature extraction methods based on parameters, the core idea of feature extraction and feature point positioning is to regard the part of the image features as a kind of model information, and the feature of a large number of samples and no feature samples are trained (and to train a large number of feature class sample and non-feature class sample respectively), finally different classifiers are constructed for feature extraction and subsequent recognition. In this paper, the learning algorithm based on Adaboost is adopted, simply and efficiently.

The Haar eigenvalue, that is, the rectangle feature is the difference between the sum of the black pixels in the feature rectangle and the sum of the white pixels. Using the traditional calculation method, it is necessary to scan all the pixels in the rectangle. The integral graph, an improved algorithm, is adopted.

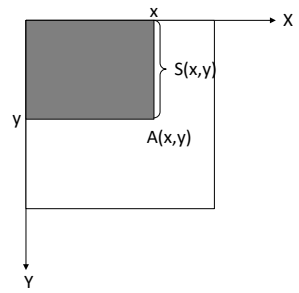

Fig. 1 The integral of the point (x, y).

As shown in Fig. 1, *A (x, y)* represents the integral of the point (x, y), and *S (x, y)* represents the sum of all the original images in the *y* direction of the point *(x, y)*.

$$A(x,y) = \sum_{x1<x,y1<y} image(x1,y1) \quad (1)$$

$$A(x,y) = A(x-1,y) + S(x,y) \quad (2)$$

Where, *image (x1, y1)* represents the original image, and the color image is the color value of the point, and for the gray scale image, the range of its gray value is 0 ~ 255.

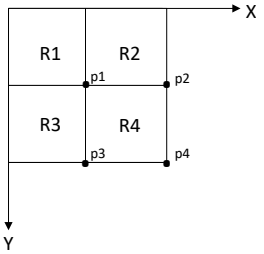

Fig. 2 The pixel value of a region.

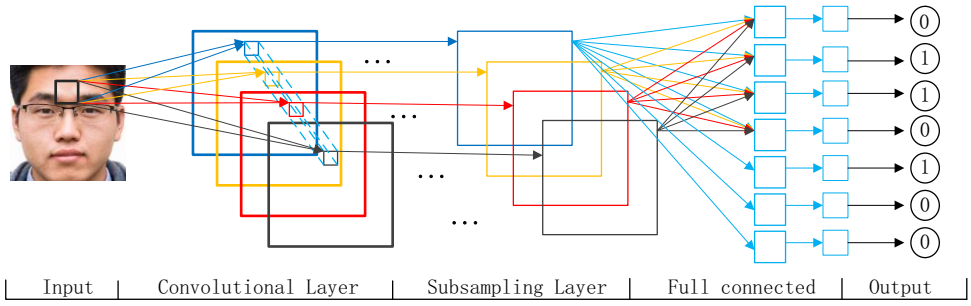

Fig.3 The model of CNN

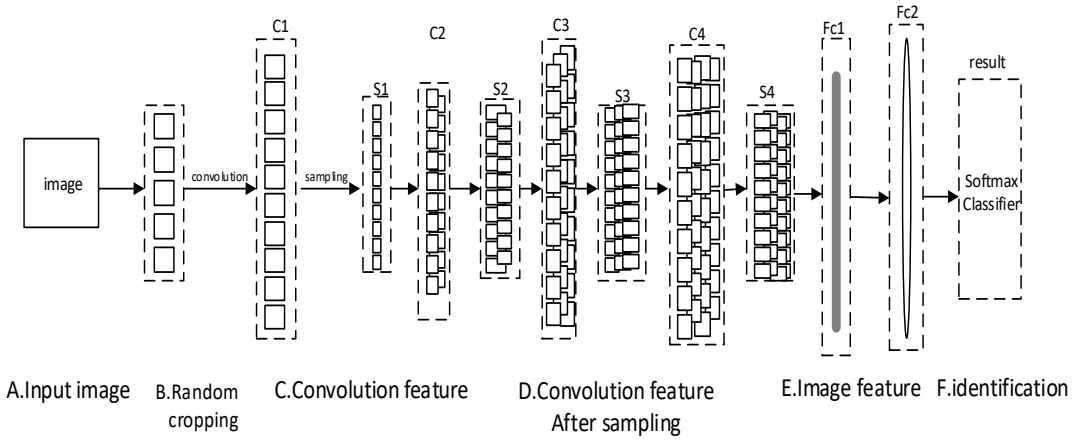

Fig.4 Network structure of our CNN

As shown in the Fig. 2, there are four areas, namely R1, R2, R3, R4, coordinates points p1, p2, p3, p4. The pixel value of the region R4 is A (p4) + A (p1) -A (p2) -A (p3).

The pixel value of a region can be calculated using the integral plot of the endpoints of the region. Let *(x, y)* be the origin of the rectangle, w be the width, h be the height. Let *SumOfRect (x, y, w, h)* be the region and the value.

$$\begin{aligned} Sumof\ \mathrm{Re}ct(x,y,w,h) &= Sum(x-1,y-1) \\ &+Sum(x+w-1,y+h-1)-Sum(x-1,y+h-1) \\ &-Sum(x+w-1,y-1) \end{aligned} \quad (3)$$

Therefore, the eigenvalues of the rectangle feature are only related to the integral graph of the endpoint of the feature rectangle, instead of the coordinates of the image. The eigenvalues can be obtained by calculating the integral graph of the endpoints of the feature rectangles and then simply adding and subtracting them. Therefore, the calculation speed of the feature is greatly improved, and the detection speed of the target is also improved.

IV. CONVOLUTIONAL NEURAL NETWORK

In 1960s, Hubel and Wiesel found that the unique network structure which can effectively reduce the complexity of the feedback neural network in the study of neurons with local sensitivity and direction selection in the cortex of cats. And then a convolution neural network (Convolutional Neural Networks -CNN) is proposed. Now, CNN has become a hot research topic in many scientific fields, especially in the field of pattern classification, because it avoids the complex preprocessing of the image and can directly input the original image, so it has been widely used [9].

A. Convolution neural network model

Convolution neural network is a non-fully connected multi-layer neural network, which contains convolutional layers (Convolutional Layer, C) and sampling layer (Subsampling Layer S), and the hidden layer of the whole network is composed of convoluted layer and sub-sampling Layer alternately, the whole model is shown in Fig. 3.

In general, the basic structure of CNN consists of two layers, one is the convolution layer, the feature extraction layer, and the input of each neuron is connected with the local acceptance domain of the previous layer, and extracts the local features. Once the local feature is extracted, and its positional relationship with other features is determined, each layer has a plurality of feature maps (Feature Map), which is extracted by a convolution filter and then each feature map has multiple neurons.

The other is the sampling layer, which is a feature map layer. Each computing layer of the network is composed of a plurality of feature maps, and each feature map is a plane, where the weights of all neurons are equal. Additionally, the number of free parameters of the network is reduced due to the shared weights of the neurons on the map. Each convolution layer in the convolutional neural network is followed by a

computational layer for local averaging and quadratic extraction. The unique structure with feature extraction twice reduces feature resolution.

In this paper, we conduct our experiments on the Linux system, training CNN model based on TensorFlow (0.10.0 CPU only). The network structure used is shown in Fig. 4. The design of the network consists of 11 layers, including 4 convolution layers, 4 sampling layers, 2 fully connected layers and 1 output layer. Except for input, each layer contains the training parameters (connection weights). The size of the input image is 112*112, normalized by the process of face detection.

TensorFlow is an open source software library for numerical computation using data flow graphs [8]. It is a system that transfers complex data structures to artificial neural networks for analysis and processing. It can be used in many areas of deep learning such as speech recognition and image recognition.

V. EXPERIMENTAL RESULTS

In this paper, the partial parameters of the convolutions are set to be same. For example, the size of the convolution kernel is 3*3, the stride is 1 and the padding is SAME which means the image size is unchanged before and after the convolution. And the weight is initialized from the standard deviation of 0.1 Gaussian sampling. The last two layers are the fully connected layers or the feature map, the number of neurons is 2048 and 224 respectively. The weights are initialized from a Gaussian distribution with a standard deviation of 0.1, the threshold is initialized to 0.1, and the dropout is performed with a probability of 50%. Table 1 details our CNN architectures.

TABLE I. OUR CNN STRUCTURE

Convolution Layer	**Partial Parameters**			
	Pool	***Stride***	***Num_Channels***	***Threshold Initial***
Conv1	2*2	2	64	0
Conv2	2*2	2	64	0.1
Conv3	2*2	2	128	0
Conv4	4*4	4	128	0.1
Fully Connected Layer				
Fc1	2048			
Fc2	224			
softmax				

The activation function we used is Rectification Linear Unit (RELU) [7]. In the result layer, there are two neurons, corresponding to 0 and 1 respectively. Following [9], softmax function is used in the last layer for predicting a single class. Softmax regression is a supervised learning algorithm, and it has a good effect on the calculation method and the recognition effect. The loss function of the softmax function is the Sigmoid Cross Entropy Loss (SCEL) which is described in (4).

$$J(\theta) = -\frac{1}{m}\left[\sum_{y}^{m} y^{(i)} \log_{h_\theta}(x^{(i)}) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right] \quad (4)$$

The process of training the model is a process of constantly optimizing the parameters of the model. Our training goal is to find the optimal model parameter θ, which can make the loss function $J(\theta)$ to the minimum. In this paper, a gradient descent algorithm is used to obtain the optimal solution by iteration. After deduction, the gradient formula is given as follows.

$$\nabla_{\theta_j} J_{(\theta)} = -\frac{1}{m}\sum_{i=1}^{m}\left[x^{(i)}(I\left\{y^{i}=j\right\} - p(y^{(i)}=j \mid x^{(i)};\theta))\right] \quad (5)$$

In practical applications, a parameter will not be arbitrarily set to 0. It will be added to the loss function of a weight attenuation term (weight decay), the weight attenuation term can retain all the parameters ($\theta_1, \theta_2, \ldots, \theta_k$). At the same time, the parameter redundancy problem of softmax regression is solved. In this paper, we use L2 regularity, the weight decay coefficient is 5*10e-4 which is the L2 regular coefficient.

In the CNN structure training, we collected from the Google for 778 pictures containing specific target, which selected 712 target face images. And then cut, according to different classification. Since the performance of CNN depends on the training data, and the amount of data we collect is small, we adopt the threshold clipping method, as shown in Fig. 5.

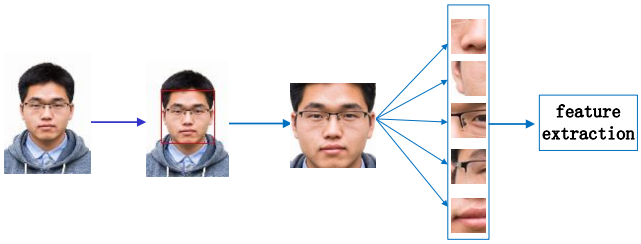

Fig. 5 Threshold clipping.

Each image is cut out 5, the threshold is set to 40*40 pixels. If the original image size is less than the threshold value. It will not be cropped. All of the original images will be taken as the positive samples together with the cropped pictures. The clipping data is shown in Table 2.

TABLE II. CLIPPING DATA

State	**Cut Result**			
	Front	***Left***	***Right***	***Total***
Before cropping	384	179	149	712
After cropping	1904	859	639	3402

Thus, we obtained 1904 positive face samples, 859 left face samples and 639 right face samples, as a total of 3402. In the same way, we take 729 different face images from the Face Detection Data Set and Benchmark (FDDB) excluding the target, and get 3636 negative samples after threshold cutting.

Finally, we use the 7038 sample images to train the CNN structure model. The validation data is set as 700, batch size 128 and the pixel depth is 255. Total training time is 1 hour. The input data is a single channel picture, and the data is preprocessed as follows.

$$x' = \frac{x - \mu_x}{\mu} \tag{6}$$

Where $x' \in (0,255)$ and $x \in (0,255)$. $\mu = 255$, $\mu_x = 255/2.0$. Thus, the data is normalized to (-0.5, 0.5).

We selected 100 target faces and 98 non-target faces as the test samples, using the traditional feature recognition method (here is LBPH) as a contrast, as shown in Table 3.

TABLE III. RECOGNITION RESULTS COMPARISON

Algorithms	Test Results		
	True positive	*False positive*	*Accuracy*
LBPH	76	34	77.55%
Our CNN	87	22	87.00%

The result of the test is that the method with our CNN has an accuracy of 87.0%. False positive is 22, that is, non-target is recognized as the target. And 13 missed, that is, the target identified as non-target. The traditional method of false positive is 34, missing 27. The CNN model is significantly better.

VI. CONCLUSIONS

The performance of face recognition is often very poor in a constrained environment, such as illumination changes, shooting angle, distance, pose, and facial expressions and so on which will affect the accuracy of recognition. Compared with traditional hand-crafted features, CNN learning features have better robustness to face recognition in complex environments.

CNN is mainly used to identify the displacement, scaling and other forms of distortion invariant two-dimensional graphics. Since CNN's feature detection layer learns by training data, it avoids the feature extraction of the display and implicitly learns from the training data when CNN is used. Moreover, because the weights of the neurons on the same feature map are the same, the network can parallel learning, which is a big advantage compared with the convolutional network of neurons connected network.

TensorFlow is the latest second-generation of Google artificial intelligence learning system based on deep learning DistBelief framework, which has been improved in all aspects, better performance, fully open source and can be run on more devices. In the existing face recognition research, the introduction and use of the framework is less, and the experiment proves that we have achieved a better face recognition effect in TensorFlow framework.

In the training of CNN, we classify the training samples according to the pose, and pretreat the training data. Experiments show that normalization can reduce the data dimension, effectively reduce the training time and improve the recognition rate. Motivated by [20], although the color image contains more information, there is no significant improvement in the recognition accuracy, so we choose a single channel image. Convolutional neural network has a unique advantage in speech recognition and image processing with its special structure of local weight sharing. Its layout is closer to the actual biological neural network, and the weight sharing reduces the complexity of the network, especially the image of multidimensional input-vector can be a direct input to the network to avoid the complexity of data reconstruction in the process of feature extraction and classification. In the future work, in order to improve the accuracy of face detection, we will use CNN cascade [11] to improve.

REFERENCES

[1] Sun Dongmei, Qiu Zhengding. A Survey of the Emerging Biometric Technology. ACTA ELECTRONICA SINCA. Vol. 29. 2001.

[2] P.J.Phillips, H.Moon. The FERET Evaluation Methodology for Face-Recognition Algorithms [J]. IEEE Transactions on PAMI, 22(10), 2000, 1090-1104.

[3] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C], proceedings of the 2011 IEEE Computer Society Conference, Computer Vision and Pattern Recognition, 2001, 1:I-511~518.

[4] Galton S. F. Personal identification and description [J]. Nature, 1888, 21:173-177.

[5] Zhang Y, Zhou Z H. Cost-sensitive face recognition [J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2010, 32(10): 1758-1769.

[6] Turk, M. Over Twenty Years of Eigenfaces. ACM Trans. Multimedia Comput. Commum. Appl. 9, 2013.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenent classification with deep convolutional neural networks. In Advances in neural information processing systems, p1097-1105, 2012.

[8] https://www.tensorflow.org/.

[9] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. arXiv preprint arXiv: 1412.1265, 2014.

[10] G. Hu et al. "When face recognition meets with deep learning: An evaluation of convolutional neural networks for face recognition." [Online]. Available: http://arxiv.org/abs/1504.02351.

[11] H. Li, Z. Lin, X. Shen, J. Brandt and G. Hua. A Convolutional Neural Network Cascade for Face Detection. Computer Vision and Pattern Recognition (CVPR), 2015 Conference on.