ISSN: 2664-9640

MobileNet-based face recognition algorithm for nursing Application on Robot

Jian Yang, Shuai Jiang *, Kai Chen, Xiaozhi Yang

School of Mechanical and Electrical Engineering, Chengdu university of technology, Cheng Du 610059, China.

* Corresponding Author

Abstract

In this paper, based on an in-depth study of the theories related to deep learning and convolutional neural networks, we propose an improved face recognition algorithm based on MobileNet-v2 and use the improved ArcFace-w as the loss function in the network to design a face recognition system with an improved network model, which is applied to a nursing robot experimental platform and accomplishes a light and efficient identity recognition task and can accurately complete the work of matching drug services.

Keywords

MobileNet-v2; ArcFace-w ;in-depth study; nursing robot.

1. Introduction

The human face has been one of the most direct and simple markers of human identity in human civilization. It has also been one of the most active research topics in computer vision and machine learning research for nearly 50 years. The class of face recognition belongs to biometric identification. The human face contains information such as identity, age, gender, and race, as well as facial expressions that reflect emotional and psychological states. The analysis of human faces and expressions involves multiple disciplines, and its research areas cover psychology, engineering, and neuroscience. Unlike several other biometric traits, face recognition does not require human cooperation and can be performed in an unobtrusive way. It can be recognized by static faces or by capturing dynamic faces in video or camera equipment, and is suitable for behavioral biometric identification. Nowadays, with the rapid development of artificial intelligence and deep learning, face recognition technology has now been applied in a large number of fields including security and law enforcement, health, education, marketing, finance, entertainment and human-computer interaction with good results.

Thus the problem of patient identification in nursing robots can be solved based on the face recognition approach. Although face recognition already has some significant breakthrough technologies, it still faces many technical difficulties in real-life applications. In particular, the accuracy and speed of face recognition in hospitals, homes and special extreme environments in crisis, where there are more different head postures, facial expressions and obscured conditions, still needs to be studied in depth. Therefore, it is of great academic value to develop a lightweight and efficient face recognition system based on deep learning technology for the face recognition scenario of mobile nursing robots, and to solve the key problems of face recognition in the application of mobile nursing robots, which can not only improve the functional level of China's medical service system, but also alleviate the conflicts between doctors and patients under the shortage of medical facilities, large number of patients, and critical and special extreme conditions. It is of great social significance.

ISSN: 2664-9640

1.1. Face Recognition Process

The complete process of face recognition is described in Figure 1. Usually a complete face recognition process first acquires faces from cameras, videos or pictures for face detection, then aligns and crops the detected faces, eliminates invalid information and performs image pre-processing, and then extracts feature vectors from the faces by face recognition algorithms to perform feature comparison calculations with the faces to be recognized to produce the final output results.

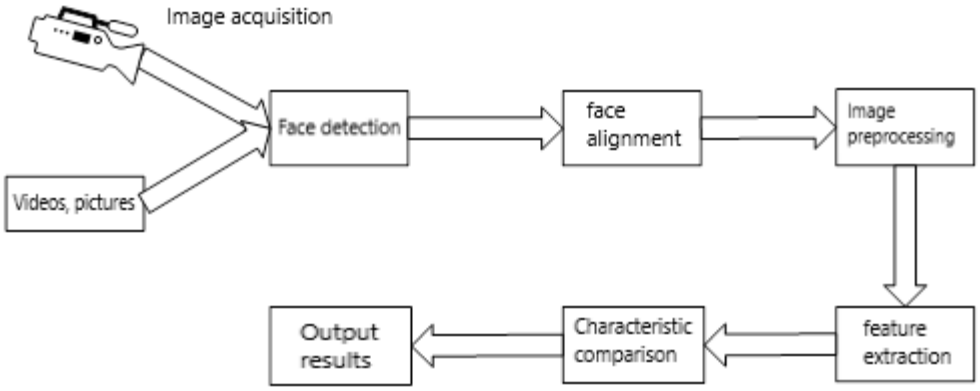

Figure 1: The complete process of face recognition

1.2. Deep learning based approach

From the early days of ANN to today's deep learning with higher accuracy and deeper networks, the field of face recognition is tired of success with the help of neural networks, which has made deep learning a hot topic in machine vision.

It has revolutionized machine learning in the last few years, with researchers in all fields, including but not limited to social sciences, engineering, and life sciences, using deeper frameworks, mixing their existing models, and getting radical results. Many researchers, especially in the field of face recognition, believe it has significant computational power as well as higher recognition accuracy. Most researchers are currently focusing on introducing improved techniques to better recognize faces by fusing existing models with deep networks as a research direction, and it has been recognized that deep neural networks and their related techniques are well suited to achieve high performance in terms of accuracy and robustness. Its ability to classify a large number of unlabeled face images in a robust and accurate manner gives it an edge over traditional face recognition methods.

2. Self-made face dataset and pre-processing

Deep learning training relies on large-scale data, and excellent data can effectively improve the performance of the network Face recognition based on deep learning requires high-quality, multi-featured face datasets. For face recognition in some specific scenarios and populations, excellent face datasets made according to the scenarios and populations should also be required.

2.1. Public data sets

Face recognition is a complex task that requires not only a good network architecture, but also a very large and correctly labeled training dataset. In this paper, the following datasets are used in the actual training and testing process: CASIA-WebFace, LFW, IMDB-WIK and self-built Asian middle-aged and elderly faces datasets. The initial test sets used for model evaluation are the self-built Asian middle-aged and elderly faces test set, LFW test set and IMDB-WIKI test set.

ISSN: 2664-9640

Table 1: Parameters of parts of spider-like robot

Dataset	Number of people	Number of pictures	Description
CASIA-WebFace	10K	500K	Identity Tags
VGGFace2	9K	3310K	Face area and identity tagging
AgeDB	0.5K	16K	Identity Tags
CAF	4.7K	314K	No noise, identity label
IJB-A	24K	49K	Face area and identity tagging
CelebA	10K	200K	5 feature points, face attributes
GANFaces-500k	10K	500K	Synthetic data
MegaFace	690K	1M	Identity Tags
IMDB-WIKI	26K	523K	Identity tag (age)
LFW	5K	10K	Identity Tags
RMFRD	0.5K	95K	Wearing a mask and not wearing a mask
SMFRD	10K	500K	Dataset generation for faces wearing masks

2.2. Public dataset framework

The framework for acquiring the datasets in this chapter is shown in Figure 2 below

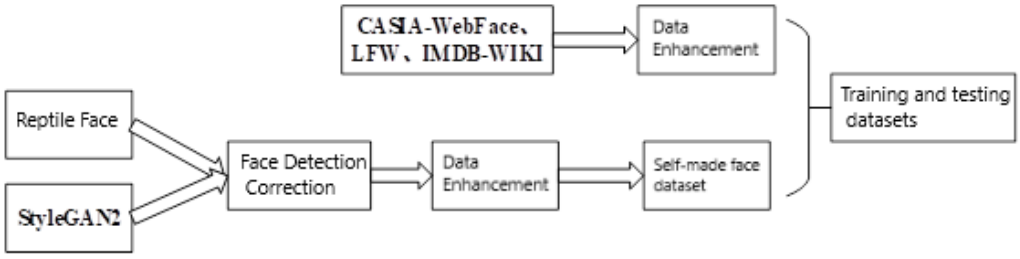

Figure 2: The framework for acquiring the datasets

In the experiments of this paper, the datasets include CASIA-WebFace, LFW, IMDB-WIKI and homemade Asian middle-aged and elderly datasets, and they are used as training datasets as well as testing datasets.

2.3. Self-made Asian middle-aged and elderly dataset

Since only a small portion of the above three communal datasets are middle-aged and elderly, they are not well suited for the applicable scenarios and populations of care robots. Therefore, in this paper, we make our own dataset of Asian middle-aged and elderly people, and after image pre-processing, the dataset has 3223 people and 131 301 images. One part is based on the face images of Asian middle-aged and elderly people crawled from the Internet, with Asian yellow middle-aged and elderly people as the main collection objects, and covering some patients' face data, part of which is shown in Figure 3. The other part is based on the face generation model of StyleGAN2 network to generate the Asian middle-aged and elderly faces needed in this paper, and its generated faces are shown in Figure 4 below.

ISSN: 2664-9640

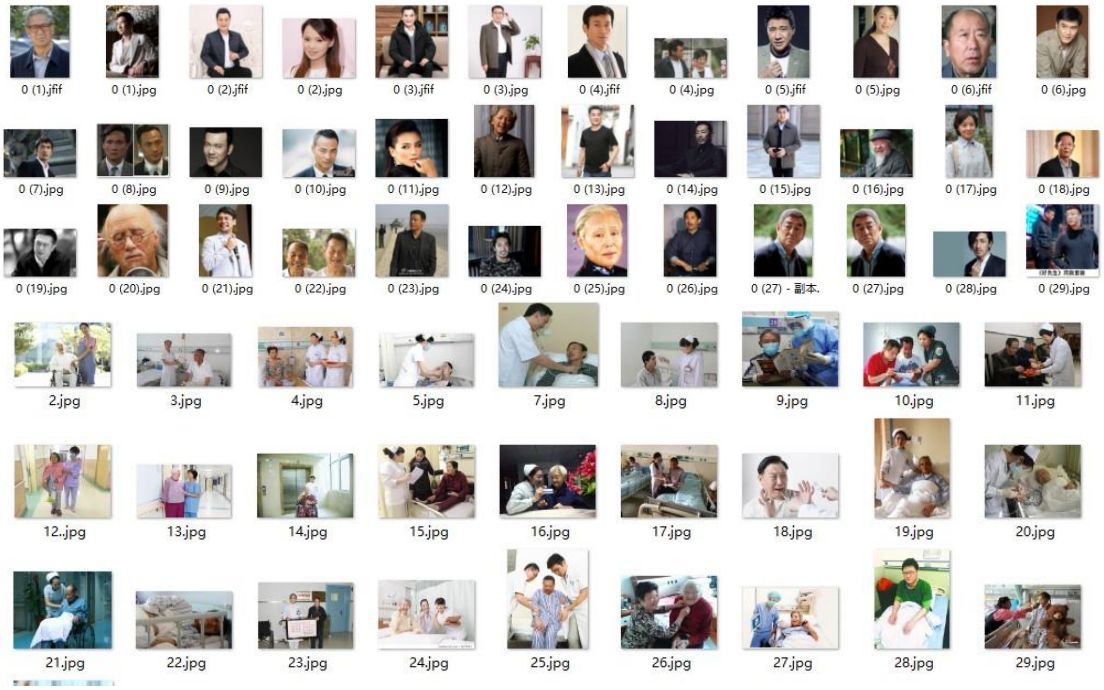

Figure 3: Crawler pictures

Figure 4: StyleGAN2 network generates images

2.4. Data pre-processing

2.4.1 Image Enlargement

At present, image augmentation generally has operations such as mirroring, panning and cropping. Image augmentation can also be interpreted as increasing the amount of the same type of data by randomly changing the training samples, reducing the dependence of the model on some features, making the model obtain higher-level feature information, improving the generalization ability of the model to prevent overfitting, and thus improving the accuracy of the model.

2.4.2 Image normalization

Since the image format of both face datasets CASIA-WebFace and LFW is 250×250, it is necessary to scale the size of the three datasets to a uniform format before inputting them into the face recognition algorithm. Also, because the image sizes required for subsequent models are different, face detection and cropping is performed using the MTCNN face detection network to unify the three dataset image sizes to 112×112 for subsequent operations

ISSN: 2664-9640

2.4.3 Data Cleaning

In the process of building the training set there will be some faces and non-faces, for this situation, the face detection module composed of AdaBoost+Haar-like is used to eliminate the part of faces, so as to achieve data cleaning of the application dataset.

3. MobileNet-based face recognition algorithm design

Based on the MobileNet-v2 network structure with deep separable convolution, we streamline the network structure according to the design rules to make the network structure lightweight and efficient in using the device resources without degrading the training accuracy and face recognition rate. The ArcFace loss function is studied and improved to improve its robustness for the adaptation scenario and population in this paper.

3.1. MobileNet

In 2017, Google proposed MobileNet, a lightweight network mainly for embedded devices, in response to the fact that most existing convolutional neural networks are complex and can only be used on computers with good performance. its biggest innovation is that it proposes Depthwise Separable Convolution, which is a form of decomposed convolution. It decomposes a standard convolution into a depth convolution and a 1×1 convolution called a point convolution. Its structure is shown in Figure 5.

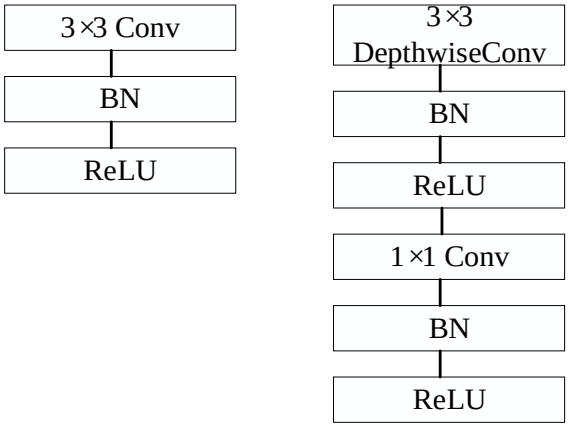

Figure 5: StyleGAN2 network generates images

3.2. Network Structure Design

The current mainstream lightweight networks are mainly MobileNet and ShuffleNet series. This design combines these two networks, using MobileNet as the main backbone network, plus ShuffleNet advantages to design a more lightweight and efficient face recognition network.

In deep networks, the higher the number of layers the clearer the features are, but a large number of convolutional layers stacked in the higher layers make the features in the higher layers become unclear and the noise increases. And the number of convolutional kernels in the high level is large, and it is inevitable that a large number of highly similar convolutional kernels will appear. Therefore, MobileNet-v2 is selected as the backbone network of the face recognition model and pruned according to the network design principles, and the network structure after pruning is shown in Table 2 below.

Comparing Table 1 and Table 2, we can see that the input size of the image is changed to 112×112 in order to reduce its computation, and a large number of Block blocks and duplicate blocks in the original network structure are deleted in order to reduce the number of parameters; the number of channels c and the expansion coefficient t of the residual network are reduced, which can effectively prevent overfitting and reduce the convolution caused by the

ISSN: 2664-9640

deep separable convolution. This can effectively prevent overfitting, reduce the amount of convolution caused by deep separable convolution, and reduce the MAC value. Most of the incomplete network training and insufficient feature learning are due to the abnormal expansion rate of the network. To avoid such problems, the expansion form of the original network is changed from spindle shape to hourglass shape in this paper.

Table 2: Network structure of MobileNet-v2

Input	*Operator*	*t*	*c*	*n*	*s*
$224^2\times3$	Conv2d	-	32	1	2
$112^2\times32$	Bottleneck	1	16	1	1
$112^2\times16$	Bottleneck	6	24	2	2
$56^2\times24$	Bottleneck	6	32	3	2
$28^2\times32$	Bottleneck	6	64	4	2
$14^2\times64$	Bottleneck	6	96	3	1
$14^2\times96$	Bottleneck	6	160	3	2
$7^2\times160$	Bottleneck	6	320	1	1
$7^2\times320$	Conv2d 1×1	-	1280	1	1
$7^2\times1280$	Avgpool 7×7	-	-	1	-
1×1×1280	Conv2d 1×1	-	k	-	-

Table 3:Improved network structure of MobileNet-v2

Input	*Operator*	*t*	*c*	*n*	*s*
$112^2\times3$	Conv2d	-	32	1	2
$56^2\times32$	Bottleneck	1	8	1	1
$56^2\times8$	Bottleneck	6	12	2	2
$28^2\times12$	Bottleneck	5	16	3	2
$14^2\times16$	Bottleneck	4	32	4	2
$7^2\times32$	Bottleneck	3	64	3	2
$3^2\times64$	Bottleneck	2	128	1	1
$3^2\times128$	Conv2d 1×1	-	1280	1	1
$3^2\times1280$	Avgpool 3×3	-	-	1	-
1×1×1280	Conv2d 1×1	-	k	-	-

In the design of the residual structure of the network, this paper borrows the residual network design principle of ShuffleNet network, that is, the difference in the number and width of channels between layers will increase the MAC value of the network, and designs two residual structures as in Figure 6(b), where DW is Depth Wise and SW is Separable Wise. adding DW and SW to the ordinary residual structure , using parallel expansion, the residual structure of the higher layers of the network mainly adopts the structure in (b) in Fig. and the residual structure of the lower layers adopts the structure in (a) in Fig. In the improved MobileNet-v2 based network framework structure, some bottleneck layers are replaced purposefully. And in order to improve the feature differentiability and accuracy of the network, the ArcFace loss function with the best performance is selected as the loss function of the network in this paper.

ISSN: 2664-9640

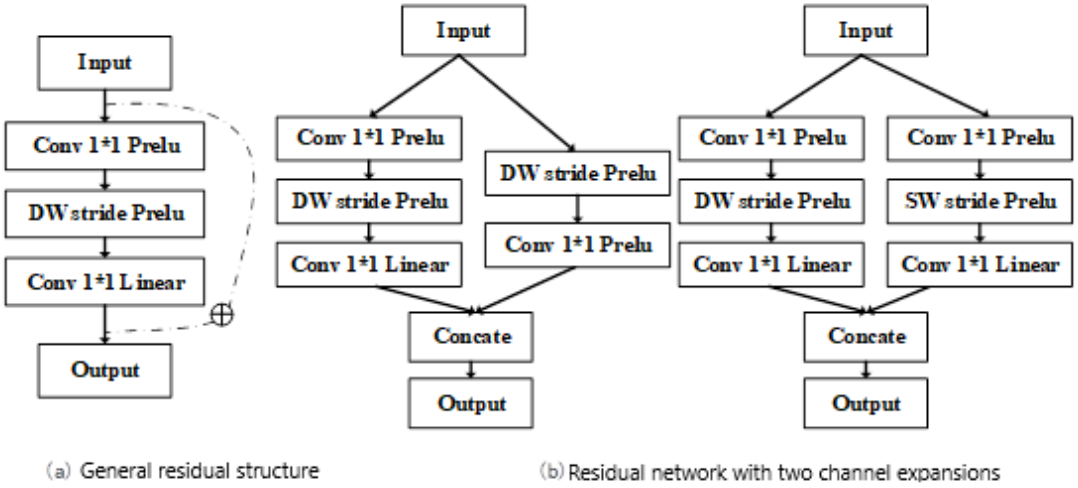

Figure 6: Residual structure

3.3. Design of loss function

Softmax loss is the most common classification function, widely used in deep learning and neural networks, often used in multilevel classification, mapping multiple neuron outputs into the (0, 1) interval and normalized to ensure that the sum is 1, and the sum of the probabilities of multiple classifications is just 1. Suppose there is an array V, Vi is the ith element in the array, and its Softmax value is Equation (1) . Suppose the input is 3, 1, -3, then it is mapped to the interval (0, 1) by Softmax, and the output is cumulative to 1. The process is shown in Figure 7.

$$S_i = \frac{e^i}{\sum_j e^{V_j}} \tag{1}$$

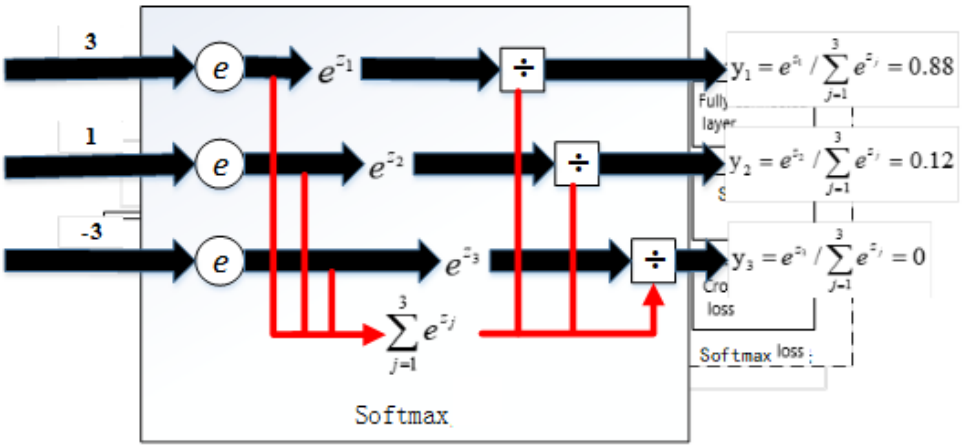

Figure 7: Softmax example

Softmax loss is based on the cross-entropy loss of the Softmax function, whose formula is shown in equation (2), and it can have a large optimization space in reducing or increasing the inter-class distance because it only deals with the classification problem between samples. Where F(x_i) is the feature of the ith sample; N is the number of batch training samples (batch_size), K

ISSN: 2664-9640

is the number of output neurons of the last fully connected layer, which can also be said to be the number of categories, and W_j and b_j are the training weights and bias of the jth category.

$$L_s = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T F(x_i)+b_i}}{\sum_{j=1}^{K} e^{W_j^T F(x_i)+b_j}} \tag{2}$$

The workflow of Softmax loss in image classification task can be understood by Figure 8. After the input training data is convolved, the output feature map F is then multiplied by the feature extraction layer and the kind weight matrix W_K in the last layer of the classification layer to get the scores of various classes, and then the kind probabilities are obtained by the Softmax function, and finally the cross entropy loss is obtained. The kind weights can be regarded as the representatives of all samples of a kind; the dot product of sample feature map matrix and kind weights can be regarded as the similarity of samples and kinds. However, in face recognition tasks, it is usually desired that the face features learned by the deep learning network are strongly differentiable, like Softmax loss as the loss function of the network, which only constrains the similarity between feature classes that can be amplified, which will lead to the network's ability to focus too much on classification. When the data gap within a class is large or the similarity between classes is large, it will have the situation of misclassification within classes and indistinguishability between classes. Therefore, given the limitations of Softmax, it is necessary to optimize for it by constraining the intra-class distance and expanding the inter-class distance. So DENG et al. proposed ArcFace (Additive Angular Margin Loss) loss function in order to make the features of the network have better differentiability (Deng et al., 2019)。

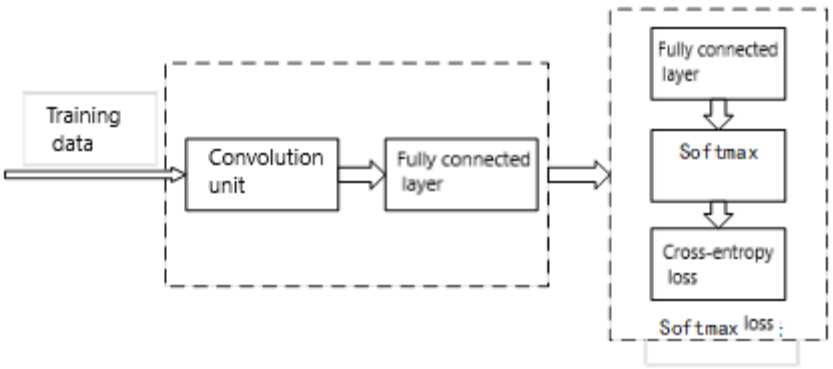

Figure 8: Softmax classification process

The ArcFace implementation is shown in Figure 9, where the ArcFace loss function performs an L2 regularization of both the feature vectors and the weights， and multiplying the two matrices $\frac{F_{x_i}}{\|F_{x_i}\|}$ and $\frac{W_j}{\|W_j\|}$ by equation (3) yields $\cos(\theta_j)$， The corresponding true label in this value is subjected to the inverse cosine operation to obtain the θ_j-value, making the network only affected by the angle of its features and weights. And then, in order to enhance the differentiability, the angular edge m is first added to the interior of $\cos(\theta_j)$ so that the Target Logit value $\cos(\theta_j + m)$ is less than $\cos(\theta_j)$ in the range of $\theta_j \in [0, \pi - m]$. Target logit refers to the fully connected layer output of the network. The constraint strength of the cosine angle is enhanced, and then a hyperparameter s is added to the obtained eigenvalues to amplify the output.

ISSN: 2664-9640

$$W_j^T F(x_i) = \left\|W_j^T\right\| \left\|F(x_i)\right\| \cos\theta_j \quad j \in [1,\cdots,n] \tag{3}$$

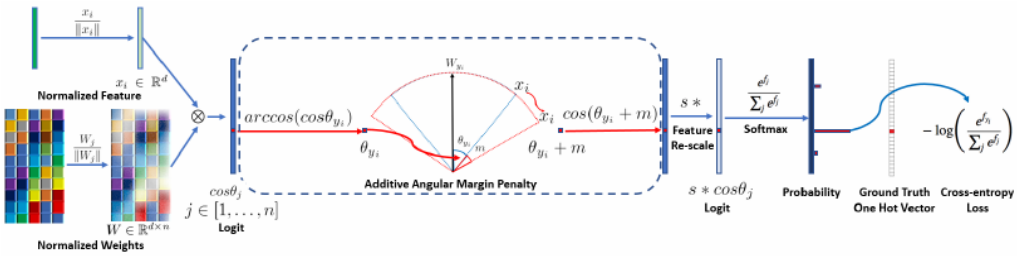

Figure 9: ArcFace Flow Chart

Finally the obtained result $s*\cos(\theta_j+m)$ is put into Softmax to output the predicted value，as shown in Equation (4). In other words Target logit is the output of the predicted category as the true category in the output matrix of the fully connected layer. With the change of hyperparameters its constraint ability is stronger and the feature change is shown in Figure 10 below.

Arcface（s=1，m=0） Arcface（s=64，m=0.5） Arcface（s=128，m=0.5）

Figure 10: Effect of ArcFace hyperparameter changes on features

$$L_{Arc} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s*\cos(\theta_j+m)}}{e^{s*\cos(\theta_{y_i}+m)} + \sum_{j=1,j\neq y_i}^{K} e^{s*\cos(\theta_j)}} \tag{4}$$

In this paper, we propose to introduce Taylor series in the inverse cosine operation in ArcFace in order to make the network more capable of classification constraints in face recognition tasks. After matrix multiplication of features and weights by the original authors to obtain the cosine of the label, the label θ_j is obtained by the inverse cosine, which is expressed by equation (5). Then from the trigonometric equation (6), while the Taylor series expansion is performed on the right side of the equation to obtain the equation (7).

$$\theta_j = arccos(\cos\theta_j) \tag{5}$$

$$arccos\, x = \frac{\pi}{2} - arcsin\, x \tag{6}$$

$$\theta_j = \frac{\pi}{2} - \left[\cos\theta_j + \frac{1}{2}\times\frac{\cos^2\theta_j}{3} + \frac{1}{2}\times\frac{3}{4}\frac{\cos^5\theta_j}{5} + \frac{1}{2}\times\frac{3}{4}\times\frac{5}{6}\times\frac{\cos^7\theta_j}{7} + \cdots\right] \tag{7}$$

In feature classification, if too many features are retained does not necessarily lead to an increase in the correct rate, but rather a decrease. In equation (7)， It can be seen that $\arcsin x$ is in the Taylor expansion of $x = \cos\theta_j$,Beacause of $0 < \cos\theta_j < 1$ and $\theta_j \in [0, \frac{\pi}{2} - m]$， The value of its later terms will become smaller and smaller, and although the weights are increasing, the

ISSN: 2664-9640

later the weights will account for a smaller percentage, so if the subsequent unnecessary features are rounded off, it may be possible to improve the correct rate. At the same time the value of θ_j increases as the number of terms of the Taylor expansion decreases. where $\cos(\theta_j + m)$ becomes smaller than the original, further strengthening the constraint and making the network classification task more demanding. To facilitate subsequent discrimination, the improved ArcFace is called ArcFace-w

4. Conclusion

In this paper, based on an in-depth study of the theories related to deep learning and convolutional neural networks, we propose an improved face recognition algorithm based on MobileNet-v2 and use the improved ArcFace-w as the loss function in the network to design a face recognition system with an improved network model, which is applied to a nursing robot experimental platform and accomplishes a light and efficient identity recognition task and can accurately complete the work of matching drug services.

References

[1] Jian Y, David Z, Frangi A F, et al. 2004. Two-dimensional PCA: a new approach to appearance-based face representation and recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 26: 131-137.

[2] Bartlett M S, Movellan J R, Sejnowski T J. 2002. Face recognition by independent component analysis[J]. IEEE Transactions on Neural Networks, 13(6): 1450-1464.

[3] Guo G, Fu Y, Dyer C R, et al. 2008. Image-Based Human Age Estimation by Manifold Learning and Locally Adjusted Robust Regression[J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 17(7): 1178-88.

[4] Kawulok M, Jing W, Hancock E R. 2011. Supervised relevance maps for increasing the distinctiveness of facial images[J]. Pattern Recognition, 44(4): 929-939.

[5] Xue Y. 2007. Non-negative matrix factorization for face recognition[J]. Dissertation Abstracts International, Volume: 69-02, Section: B, page: 1124: Adviser: Chong-Sze Tong.

[6] He X, Yan S, Hu Y, et al. 2005. Face Recognition Using Laplacian Faces[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(3): 328-340.

[7] Kasturi Rangachar, Goldgof Dmitry, Soundararajan Padmanabhan, Manohar Vasant, Garofolo John, Bowers Rachel, Boonstra Matthew, Korzhova Valentina, Zhang Jing. 2009. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: data, metrics, and protocol.[J]. IEEE transactions on pattern analysis and machine intelligence, 31(2).

[8] Yuille A L, Hallinan P W, Cohen D S. 1992. Feature extraction from faces using deformable templates[J]. International Journal of Computer Vision, 8(2): 99-111.

[9] Brunelli R, Poggio T. 1993. Face Recognition: Features Versus Templates[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 15(10): 1042-1052.

[10] Lades, M, Vorbruggen, et al. 1993. Distortion invariant object recognition in the dynamic link architecture[J]. Computers IEEE Transactions on.

[11] Conde C, Serrano A, Rodriguezaragon L J, et al. 2007. An automatic 2D, 2.5D & 3D score-based fusion face verification system[C]//International Workshop on Computer Architecture for Machine Perception & Sensing. IEEE.

[12] Wiskott L, Fellous J M, Kuiger N, et al. 1997. Face recognition by elastic bunch graph matching[C]//International Conference on Computer Analysis of Images and Patterns. Springer, Berlin, Heidelberg.