LOW-RESOLUTION FACE RECOGNITION WITH LARGER IMAGINARY MARGIN

Jiacheng Yang, Jiawei Li, Zhihao Ouyang, Qianggang Ding, Shu-Tao Xia

Department of Computer Science and Technology, Tsinghua University, Beijing, China

ABSTRACT

Large-margin based methods have made significant progress in traditional face recognition problem. However, existing large-margin loss of deep convolutional neural networks(CNN) is easy to fail or have not played a big role when the face images has a low resolution. In this paper, we propose a two-step framework to address this problem. First, We construct a super-network which can be an ordinary CNN, called SRnet. Then a large-margin based loss is used to learn the generated faces from SRnet. we have found that SRnet and large-margin softmax loss function can promote each other in low resolution face recognition field for the first time. Compared to directly train large-margin models on low-resolution faces, the proposed framework aims to learn a larger maginary margin, which is better for the discriminant classification with the help of SRnet. Our proposed method has been validated on the CASIA-WebFace[1] and LFW[2] datasets which are training dataset and testing dataset, respectively. The extensive experiments show that our method can actually learn a larger margin.

Index Terms— low resolution, face recognition, large margin, SRnet

1. INTRODUCTION

In the past few decades, face recognition[3, 4, 5, 6, 7] developed rapidly and performed very well in many application scenarios. Still, there are existing lots of challenging conditions such as occlusion, various poses, lighting and various expression. People usually use high-quality face images datasets in these experiments and applications. Few experiments have been conducted to identify low-resolution face images in the real world, such as surveillance[8] systems. The challenge of face recognition in the surveillance system is that the face is far away from the camera so we cannot obtain high resolution image. Therefore, it is of great social significance to study low-resolution face recognition[9, 8, 10, 11, 12] to improve the recognition accuracy under this scene.

Recently, many low-resolution face recognition studies [13], [14] and [15] have adopted a two-step network structure framework. They firstly use low-resolution face images to obtain high-resolution images by bicubic unsampling, [13] maps different low resolutions to the same high resolution, and then identifies them through an MRCNN network, it can adapt to any low-resolution face recognition. [14] proposed a FECNN network to perform nonlinear mapping on the high-resolution images, which are obtained by upsampling in order to obtain a reconstructed image that is more conducive to recognition, and then fine-tuned it with the VGG[16] network trained by high-resolution images. [15] proposed the backbone network and branch network to identify different low-resolution face images. The backbone network is mainly used to extract features, and the branch network is used to map high-resolution features and low-resolution features to the same hidden space. It trains different branch networks and shares the same backbone network for different resolution recognition. A common problem with the above networks is that bicubic upsampling improves resolution but does not increase the information contained in low resolution images. Therefore, [17] and [18] directly use the matrix mapping method to map the low-resolution face features and the corresponding high-resolution face features to the same space for matching training. Moreover, the accuracy of low-resolution face recognition is improved by reducing the feature distance of the same type of faces and increasing the feature distance of different types of faces. [19] proposed a better angular loss to reduce the distance between the same class and increase the distance between different classes, and achieved higher accuracy in face recognition.

In order to solve the above problems, we propose a new two-step low-resolution face recognition network structure, using SRnet to super-resolute low-resolution face images, and then combined with resNet, using large margin softmax loss function. The test accuracy achieves considerable improvement on the CASIA-WebFace and LFW datasets.

2. LOW RESOLUTION FACE RECOGNITION

In low-resolution face recognition scenarios, tasks can be done by measuring the distance between the features of the LR and HR face images,as formula 1.

$$d_{ij} = dist(l_i, h_j) \tag{1}$$

Where $l_i \in R^n, i = 1, 2, ..., N$ and $h_j \in R^m, i = 1, 2, ..., N$ $(n < m)$ represents the n-dimensional feature vector of the LR image and the m-dimensional vector of the

HR image, respectively. Since the feature dimensions of LR and HR do not match, the general distance such as Euclidean distance is obviously not directly applicable. In order to solve the above problem, the conventional method is to map the features of the LR image to the HR space by SR, and then calculate the distance in the HR space,as formula 2.

$$d_{ij} = dist(f_{SR}(l_i), h_j) \tag{2}$$

In this paper, we use SRnet to map both LR and HR images to a common new space with a feature dimension of M, called coupled mappings method. The distance can be calculated by Equation 3 below.

$$d_{ij} = dist(f_L(l_i), f_H(h_j)) \tag{3}$$

For an ideal new space, it should satisfy that the feature distances of LR and HR pictures of the same category are closer. Let $f_L(l) = W_L^T l$ and $f_H(h) = W_H^T h$ be the linear mapping of the LR and HR images, respectively. The loss function is expressed as the formula 4.

$$J(W_L, W_H) = \sum_{i=1}^{N} \parallel W_L^T l_i - W_H^T h_i \parallel^2 \tag{4}$$

Where N represents the number of training images, W_L and W_H are two mapping matrices, the sizes are $n \times M$ and $m \times M$, respectively.

3. LARGE MARGIN

[14] further proposed a method to increase the feature distance between different classes, called LMCP. They created two graphs, namely the graph Gw of the same category and the graph G_b of different categories. P_w and P_b respectively represent the relationship matrices between images in G_w and G_b space, see Equations 5, respectively.

$$P_{x,ij} = \begin{cases} exp(-\frac{\|h_j - h_i\|_2}{\sigma}), & if\, h_i, h_j\, connected\, in\, G_x \\ 0, & otherwise \end{cases} \tag{5}$$

x stands for w and b. The new common space where LR and HR are mapped should have this characteristic: the feature distances in G_w is smaller than the feature distances in G_b. In order to achieve this, we use Equations 6 and 7 to constrain the mapping.

$$\min_{W_L^T, W_H^T} \sum_{i,j} \parallel W_L^T l_i - W_H^T h_j \parallel_2^2 P_{w,ij} + \parallel W_L^T l_i - W_L^T l_j \parallel_2^2 P_{w,ij} + \parallel W_H^T h_i - W_H^T h_j \parallel_2^2 P_{w,ij} \tag{6}$$

$$\max_{W_L^T, W_H^T} \sum_{i,j} \parallel W_L^T l_i - W_H^T h_j \parallel_2^2 P_{b,ij} + \parallel W_L^T l_i - W_L^T l_j \parallel_2^2 P_{b,ij} + \parallel W_H^T h_i - W_H^T h_j \parallel_2^2 P_{b,ij} \tag{7}$$

Where Pw and Pb represent the weight matrix in the Gw and Gb spaces, respectively. After using the above equations, the feature distance of the images in the Gw space is minimized and the feature distance of the image in the Gb space is maximized.

[19] proposed a large margin method to improve the softmax loss function. Assuming that the i-th input has the feature x_i and the label y_i, the original softmax function can be written as equation 8.

$$L = \frac{1}{N} \sum_i L_i = \frac{1}{N} \sum_i -log(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}) \tag{8}$$

Where f_j denotes the score of the i-th input divided into the j-th class, and N is the total number of samples of the training data. f_{y_i} can be written as $f_{y_i} = W_{y_i}^T x_i$, where W_{y_i} is the Y_i column of W(offset b is omitted). So

$$f_j = \parallel W_j \parallel \parallel x_i \parallel cos(\theta_j) \quad 0 \leq \theta_j \leq \pi \tag{9}$$

At this point, the loss becomes equation 10.

$$L_i = -log(\frac{e^{\|W_{y_i}\| \|x_i\| cos(\theta_{y_i})}}{\sum_j e^{\|W_j\| \|x_i\| cos(\theta_j)}}) \tag{10}$$

After the softmax function is written in the form of Equation 10, for the input x, the modulo $\parallel W \parallel$ of the parameter W is normalized, then the class to which the sample i belongs is determined only by the angle between the sample x and W_j. if the angle of x and W_j is the smallest, then the sample i belongs to the c-class.

Consider the problem of the two classifications, suppose we have a sample x belonging to class 1, and for the original softmax then we make $W_1^T x > W_2^T x$, ie $\parallel W_1 \parallel \parallel x \parallel cos(\theta_1) > \parallel W_2 \parallel \parallel x \parallel cos(\theta_2)$, so that the sample x can be classified correctly. However, if we want to make the classification more discriminative, we can set a decision margin, requiring $\parallel W_1 \parallel \parallel x \parallel cos(m\theta_1) > \parallel W_2 \parallel \parallel x \parallel cos(\theta_2) \;\; 0 \leq \theta_1 \leq \frac{\pi}{m}$ m, where m is a positive integer, we can see equation 11.

$$\parallel W_1 \parallel \parallel x \parallel cos(\theta_1) > \parallel W_1 \parallel \parallel x \parallel cos(m\theta_1) > \parallel W_2 \parallel \parallel x \parallel cos(\theta_2) \tag{11}$$

So this new taxonomy is a stronger rule for correctly classifying sample x, resulting in a stronger decision boundary for class 1, our proposed architecture use this large margin method for low resolution face recognition.

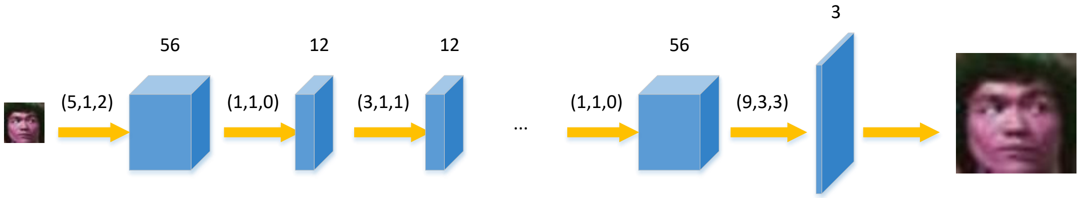

Fig. 1. SRnet. (a,b,c) stand for kernel size, stride and padding respectively. The numbers above the blue rectangle represent channels. The input is a low resolution image, and the output is a super-resolution image[20] whose visual effect is not so good because it is used for classification and verification.

4. PROPOSED ARCHITECTURE

4.1. SRnet

Our SRnet is built with reference to FSRCNN[21] which is good at super-resoluting. But FSRCNN is just for rebuilding better visual effects, not for face recognition. Our SRnet can better extract the facial features needed based on the task of low-resolution face recognition, but at the expense of visual effects. So the network parameters and function between our SRnet and FSRCNN are completely different. SRnet can be broken down into five parts, too: feature extraction, shrinking, mapping, expansion, and deconvolution. The first four parts are the convolutional layer and the last one is the deconvolution[22] layer. The feature extraction layer is followed by a shrink layer. In order to reduce the extracted feature dimension, the filter size of this layer is 1x1, but the number is smaller than the feature extraction layer, so this layer can reduce the number of parameters. The nonlinear mapping layer is the most critical network layer that affects the super-resolution effect. The key factors are the width and depth of the filter. The number of layers in the nonlinear mapping layer is also a sensitive variable, which determines the accuracy and complexity of the mapping. The expansion layer is the inverse of the shrink layer which saves computational overhead. But if the high-resolution image is reconstructed directly from the low-dimensional feature vector, the effect is poor. So the extension layer increases the learned high-resolution features dimensions to improve the reconstruction effect, the filter size of this layer is the same as the shrink layer, ie 1x1. The number is the same as the feature extraction layer, too. The last layer is the deconvolution layer. Upsampling the high-resolution feature vectors learned by the previous extension layer can be seen as the reverse operation of the convolutional layer.

Past low-resolution face recognition first uses interpolation to increase resolution and then passes through a CNN network such as SRCNN[23]. The interpolation such as bicubic can not increase information for image, their reconstruction effect mainly depends on the following CNN, the kernel size is often 1×1 in order to keep the image dimensions. We use SRnet to make the entire training process possible in the neural network. We have done a comparative experiment between SRnet and bicubic interpolation plus ordinary CNN network. We have found that their role in low-resolution face recognition is quite similar. Therefore, large-margin promotes SRnet extracting features with universality.

The network structure of SRnet is shown in Figure 1. Each convolution layer is followed by a layer of PReLu activation function.

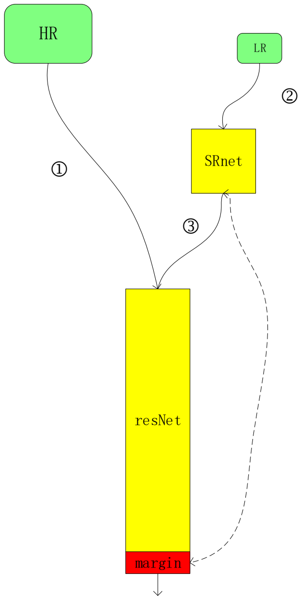

Fig. 2. overall network.First, use HR training the resNet network with SL-margin in the left part. Then, train the SRnet with LR as input and corresponding HR as supervision. Finally, connect SRnet and resNet with SL-margin, fitune the overall network.

Table 1. The testset accuracy rate of LR and HR face verification in different conditions(unit:%)

Fold	1	2	3	4	5	6	7	8	9	10	AVE
LR	93.67	93.67	94.5	94	91.17	93.17	94.33	93.83	94.33	92.5	93.52
LR+SRnet	92.17	93	92	92.67	89.33	91	93.83	91.33	93.83	92.33	92.15
LR+SL-margin	95.83	95.83	95.5	95.5	94.5	96.33	96.33	95.5	97	96.83	95.92
HR	93.67	94.5	93.67	94.33	93.33	93.5	96.67	92.83	95.33	96	94.38
HR+SL-margin	98.67	99	99	98.83	98.5	98.83	99	98.83	98.83	99.33	98.98
LR+SRnet+SL-margin	97.5	97.67	98	96.5	97.5	98.67	97.17	98.17	98.5	98.17	97.78
LR+srcnn*+SL-margin	97	97.83	98.17	96.17	97.33	98.5	97.5	98.17	98.5	98.5	97.77

4.2. overall architecture

The overall network framework we propose is shown in Figure 2. We trained our CNN model using the open network-collected training dataset CASIA-WebFace (after excluding the identity images that appeared in the testset). In our experiments, the high resolution of images in CASIA-WebFace and LFW datasets is 96×96, and we downsample them to 32×32 resolution as the low resolution images. The resNet network is built with reference to Sphereface network[19]. But our resNet connect to the SRnet, and perform better in low resolution face recognition field compared to Sphereface network.

The large margin method in our architecture is equation 10 and 11, we call it sphereface large margin(SL- margin). Compared with another large margin method of equation 7 which we called WB-margin, we find that SL-margin has lower complexity. Our training dataset CASIA WebFace has 494,414 face images. For WB-margin, every two picture needs to calculate a P as equation 7, which is a NP problem. So our architecture uses SL-margin. What's more, we find that the SL-margin can benefit SRnet to get better face features in our low resolution face recognition tasks.

The training process is as follows:

1) Use the high-resolution face images training the resNet network with SL-margin in the left part in figure 2.

2) Train the SRnet in figure 2 with low resolution face images as input and corresponding high resolution face images as supervision. The loss function is Euclidean distance loss here.

3) Remove the loss layer of SRnet and the input layer of the resNet network framework with SL-Margin loss, and use the output of SRnet as the input of resNet network. We call the connection of SRnet, resNet network and SL-margin as overall network.

4) Fine-tuning the overall network using parameters trained in 1) and 2).

5. EXPERIMENT

We think that our biggest innovation is that SRnet and SL-margin promote each other to achieve better accuracy in the field of low-resolution face recognition, and this property can be proved through controlled experiments only, so we only have done several experiments base on CASIA-WebFace and LFW datasets. We think these experiments are enough to prove this property.

The accuracy of the testset obtained using the Sphereface network of this paper is shown in Table 1. LR means 32X32 low resolution images and HR means 96X96 high resolution images. We used a ten-fold cross-validation method during the test, and the final accuracy was their average. From the table we can see that when there is no SRnet and SL-margin, the accuracy of LR is 93.25% and HR is 94.38%, only a difference of 1.13%. When adding SL-margin, the accuracy of LR and HR has become 95.92% and 98.98%, respectively. They have increased 2.67% and 4.6% respectively, but the accuracy of LR and HR has a gap of 3.06%. We want to make the accuracy of LR as close as possible to HR. When adding SRnet, the accuracy of LR has become 92.15%, it decreases because what we need is classification accuracy rather than image reconstruction. We need something to guide SRnet to extract facial features that are good for classification, SL-margin does it. When we use SRnet and SL-margin to guide the network, the accuracy of LR increases to 97.78%, only 1.2% lower than HR. We can also see that when only SL-margin used, the accuracy of LR is only 95.92%, so SRnet promotes SL-margin to separate different types of facial features.

We have also done a comparative experiment between SRnet and bicubic interpolation plus ordinary CNN network called srcnn* in the table, which is a five layers CNN network. The accuracy is 97.78% and 97.77%, respectively. So arbitrary SR network and large margin softmax loss function can promote each other.

6. CONCLUSION

In this study, we have proposed an overall architecture for low resolution face recognition. It has been validated on the CASIA-WebFace and LFW datasets which are training dataset and testing dataset, respectively. we increase the accuracy of low resolution images from 93.25% to 97.78% on the above datasets. We have also done a lot of comparison experiments and found that SRnet and large margin softmax loss function can promote each other for the first time.

7. REFERENCES

[1] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[2] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[3] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld, "Face recognition: A literature survey," *ACM computing surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[4] Changxing Ding and Dacheng Tao, "Robust face recognition via multimodal deep face representation," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2049–2058, 2015.

[5] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al., "Deep face recognition.," in *BMVC*, 2015, vol. 1, p. 6.

[6] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Sparsifying neural network connections for face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4856–4864.

[7] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[8] Wilman WW Zou and Pong C Yuen, "Very low resolution face recognition problem," *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 327–340, 2012.

[9] Zhifei Wang, Zhenjiang Miao, QM Jonathan Wu, Yanli Wan, and Zhen Tang, "Low-resolution face recognition: a review," *The Visual Computer*, vol. 30, no. 4, pp. 359–386, 2014.

[10] Pablo H Hennings-Yeomans, Simon Baker, and BVK Vijaya Kumar, "Simultaneous super-resolution and feature extraction for recognition of low-resolution faces," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[11] Bo Li, Hong Chang, Shiguang Shan, and Xilin Chen, "Low-resolution face recognition via coupled locality preserving mappings," *IEEE Signal processing letters*, vol. 17, no. 1, pp. 20–23, 2010.

[12] Sivaram Prasad Mudunuri and Soma Biswas, "Low resolution face recognition across variations in pose and illumination," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 5, pp. 1034–1040, 2016.

[13] Chunhui Ding, Tianlong Bao, Saleem Karmoshi, and Ming Zhu, "Low-resolution face recognition via convolutional neural network," in *Communication Software and Networks (ICCSN), 2017 IEEE 9th International Conference on*. IEEE, 2017, pp. 1157–1161.

[14] Erfan Zangeneh, Mohammad Rahmati, and Yalda Mohsenzadeh, "Low resolution face recognition using a two-branch deep convolutional neural network architecture," *arXiv preprint arXiv:1706.06247*, 2017.

[15] Ze Lu, Xudong Jiang, and Alex ChiChung Kot, "Deep coupled resnet for low-resolution face recognition," *IEEE Signal Processing Letters*, 2018.

[16] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[17] Fuwei Yang, Wenming Yang, Riqiang Gao, and Qingmin Liao, "Discriminative multidimensional scaling for low-resolution face recognition," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 388–392, 2018.

[18] Jiaqi Zhang, Zhenhua Guo, Xiu Li, and Youbin Chen, "Large margin coupled mapping for low resolution face recognition," in *Pacific Rim International Conference on Artificial Intelligence*. Springer, 2016, pp. 661–672.

[19] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song, "Sphereface: Deep hypersphere embedding for face recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 1, p. 1.

[20] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[21] Chao Dong, Chen Change Loy, and Xiaoou Tang, "Accelerating the super-resolution convolutional neural network," in *European Conference on Computer Vision*. Springer, 2016, pp. 391–407.

[22] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus, "Deconvolutional networks," 2010.

[23] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*. Springer, 2014, pp. 184–199.