LOW RESOLUTION FACE RECOGNITION AND RECONSTRUCTION VIA DEEP CANONICAL CORRELATION ANALYSIS

Zhao Zhang[1] *Yun-Hao Yuan*[1,*] *Xiao-Bo Shen*[2] *Yun Li*[1]

[1]School of Information Engineering, Yangzhou University, 225127, China
[2] School of Computer Science and Engineering, Nanyang Technological University, 639798, Singapoe
yhyuan@yzu.edu.cn

ABSTRACT

Low-resolution (LR) face identification is always a challenge in computer vision. In this paper, we propose a new LR face recognition and reconstruction method using deep canonical correlation analysis (DCCA). Unlike linear CCA-based methods, our proposed method can learn flexible nonlinear representations by passing LR and high-resolution (HR) image principal component features through multiple stacked layers of nonlinear transformation. As the nonlinear transformation in deep neural networks is implicit, we apply radial basis function based neural network to learn an explicit mapping between principal components and correlational features. In addition, we also design two residual compensation methods for identification and vision enhancement, respectively. The proposed approach is compared with existing LR face recognition and reconstruction algorithms. A number of experimental results on benchmark datasets have demonstrated the effectiveness and robustness of our method.

Index Terms— Low resolution, face recognition, image reconstruction, deep learning, canonical correlation

1. INTRODUCTION

In video surveillance, human face is essential information for robust recognition and authentication. However, owing to environment influence and imaging equipment limitation, face images often present very low resolution (LR), which makes subsequent face recognition or authentication tasks challenging. A simple example is shown in Fig. 1, where a female face image is blurry due to the long distance. To overcome this issue, many face super-resolution (FSR) techniques have been proposed in recent decades. FSR aims at inferring high-resolution (HR) facial images or recognition features from LR ones. Roughly, FSR methods can be divided into vision-oriented and recognition-oriented methods.

Fig. 1. A typical low-resolution face image.

Vision-oriented methods focus on obtaining good visual effects by image reconstruction. Typical methods include manifold-based [1, 2], dictionary-based [3, 4] and regression-based methods [5, 6, 7, 8]. Differently, recognition-oriented techniques aim at achieving high recognition accuracy on LR face images. Li et al. [9] first reconstructed HR features instead of HR images for face recognition. After that, sparse-representation-based methods [10] and deep-learning-based [11] methods are springing up with encouraging performances.

Recently, Huang et al. [12] used canonical correlation analysis (CCA) to learn the linear correlations between HR and LR facial features, which achieves satisfactory reconstruction quality. Nevertheless, CCA is a linear learning approach in essence, thus difficult to measure the nonlinear relationships between HR and LR facial images. To solve this issue, Zhang et al. [13] proposed a kernel CCA (KCCA) based LR face recognition method, where the nonlinear correlation between HR and LR face features can be well depicted by two kernel mappings. However, an obvious drawback of KCCA is that the learned nonlinear representation is limited by the fixed kernel. More importantly, projecting original features into kernel spaces is opaque, which makes the mapping process irreversible, therefore becoming difficult to reconstruct HR face images from LR ones.

Deep neural network (DNN) has been widely used in computer vision, which is a powerful tool for uncovering the nonlinear information hidden in the data and thus obtains the

*Corresponding author.

This work is supported by the National Natural Science Foundation of China under Grant Nos. 61402203, 61703362, 61472344, 61611540347, the Natural Science Foundation of Jiangsu Province of China under Grant Nos. BK20161338, BK20170513, and Yangzhou Science Project Fund under Grant No. YZ2017292. Moreover, it is also sponsored by Excellent Young Backbone Teacher (Qing Lan) Project and Scientific Innovation Project Fund of Yangzhou University under Grant No. 2017CXJ033.

great success. Motivated by advanced ideas from DNN and CCA, we propose a new LR face recognition and reconstruction approach based on deep CCA (DCCA). The proposed method can learn flexible nonlinear representations by passing HR and LR facial features via multiple stacked layers of nonlinear transformation. Specifically, we first extract global facial structure both in HR and LR training images, and then employ DCCA to learn the nonlinear consistency between HR and LR facial features. At last, given a LR test image, we use the neighborhood-based reconstruction approach [1] to generate HR facial feature from LR one in a coherent subspace. For recognition purpose, residual compensation is implemented in HR feature space and the nearest neighbor classifier is used for identification. For reconstruction purpose, we apply a radial basis function (RBF) based neural network to build a regression model from correlational features to principal component features. Many experimental results show the effectiveness of our proposed method.

2. CCA

Given two centered sets $X = [x_1, x_2, \ldots, x_m] \in \mathbb{R}^{p\times m}$ and $Y = [y_1, y_2, \ldots, y_m] \in \mathbb{R}^{q\times m}$, CCA aims to seek $n \leq \min(p, q)$ pairs of basis vectors $W_x \in \mathbb{R}^{p\times m}$ and $W_y \in \mathbb{R}^{q\times m}$, making the correlation coefficient of canonical projections $W_x^\top X$ and $W_y^\top Y$ maximized. There are many expressions for the optimization objective, one of which is

$$\begin{aligned} &\max_{W_x, W_y} \operatorname{tr}\left(W_x^\top X Y^\top W_y\right) \\ &s.t.\ W_x^\top X X^\top W_x = W_y^\top Y Y^\top W_y = I, \end{aligned} \tag{1}$$

where $\operatorname{tr}(A)$ denotes the trace of matrix A, and I is the identity matrix. The optimization problem (1) can be solved by the generalized eigenvalue problem.

3. PROPOSED APPROACH

Our approach employs a two-step framework: the first step is to carry out facial features/images reconstruction. In the second step, different residual compensation methods are adopted according to the purposes of identification and vision.

3.1. FSR for vision enhancement

Assume the HR face set is $I^h = [i_1^h, i_2^h, \ldots, i_m^h] \in \mathbb{R}^{p\times m}$ and the corresponding LR set is $I^l = [i_1^l, i_2^l, \ldots, i_m^l] \in \mathbb{R}^{q\times m}$, First, we center the LR and HR training images by $\hat{I}^l = \{I_j^l - \mu^l\}_{j=1}^m$ and $\hat{I}^h = \{I_j^h - \mu^h\}_{j=1}^m$, where μ^l and μ^h denote the mean faces of the LR and HR images. Then, to improve the computational efficiency and reduce the noise, we use principal component analysis (PCA) to extract the global facial features of LR and HR training sets by the following:

$$X^l = P_l^T \hat{I}^l \text{ and } X^h = P_h^T \hat{I}^h, \tag{2}$$

where P_l and P_h are the PCA projection matrices.

Considering that X^l and X^h come from the same faces differing in resolution, it is natural that they have the intrinsic consistency. Therefore, we use DCCA to learn flexible nonlinear representations to enhance consistency, as follows:

$$\begin{aligned} \left(\theta_f^*, \theta_g^*, W_f^*, W_g^*\right) = &\underset{\theta_f, \theta_g, W_f, W_g}{\arg\max} \operatorname{tr}\left(W_f^\top \Sigma_{fg} W_g\right) \\ &s.t.\ W_f^\top \Sigma_{ff} W_f = W_g^\top \Sigma_{gg} W_g = I, \end{aligned} \tag{3}$$

where

$$\begin{aligned} \Sigma_{ff} &= F(X^l; \theta_f) F(X^l; \theta_f)^\top + r_f I, \\ \Sigma_{gg} &= G(X^h; \theta_g) G(X^h; \theta_g)^\top + r_g I, \\ \Sigma_{fg} &= F(X^l; \theta_f) G(X^h; \theta_g)^\top, \end{aligned}$$

$F\left(X^l; \theta_f\right)$ and $G\left(X^h; \theta_g\right)$ are the centered outputs of two DNNs, θ_f and θ_g are the vectors containing all parameters of two DNNs, r_f and r_g are two small positive numbers. We employ mini-batch gradient descent (MBGD) to solve the model (3), as used in [14]. Once the optimal projection matrices and the parameters of DNNs are obtained, we can get the correlational features $C^l = W_f^{*\top} F(X^l; \theta_f^*)$ and $C^h = W_g^{*\top} G(X^h; \theta_g^*)$.

3.1.1. Facial reconstruction

If a new LR face image i_t is given, we first compute its principal component feature by $x_t^l = P_l^\top (i_t - \mu_l)$, then transform it to the coherent subspace by $c_t^l = W_f^{*\top} F(x_t^l; \theta_f^*)$.

Now, we reconstruct the corresponding HR correlational features $\tilde{c}_t^h$ using the idea of neighborhood reconstruction. For c_t^l, we find its nearest k neighbors $\{C_{t_j}^l\}_{j=1}^k$ in C^l measured by Euclidean distance, and the weight coefficients $A = \{\alpha_{t_j}\}_{j=1}^k$ are obtained via minimizing the reconstruction error:

$$\varepsilon = \left\| c_t^l - \sum\nolimits_{j=1}^k \alpha_{t_j} C_{t_j}^l \right\| \ \ s.t. \sum\nolimits_{j=1}^k \alpha_{t_j} = 1, \tag{4}$$

where $\|\cdot\|$ denotes the 2-norm of a vector. The corresponding HR feature $\tilde{c}_t^h$ can be reconstructed by applying weight A to $\{C_{t_j}^h\}_{j=1}^k$ in C^h:

$$\tilde{c}_t^h = \sum\nolimits_{j=1}^k \alpha_{t_j} C_{t_j}^h. \tag{5}$$

In order to reconstruct the HR image, we need to get the corresponding principal component feature x_t^h from c_t^h. Considering the non-linear mapping implemented by DNN is implicit, we reestablish the relationship between C^h and X^h using the following mapping:

$$X^h = W_{RBF} \Phi. \tag{6}$$

The (i, j)th element in the matrix $\Phi \in \mathbb{R}^{m\times m}$ is calculated by $(\Phi)_{ij} = \exp(-||c_i^h - c_j^h||/2\sigma^2)$ with c_i^h as the ith column in

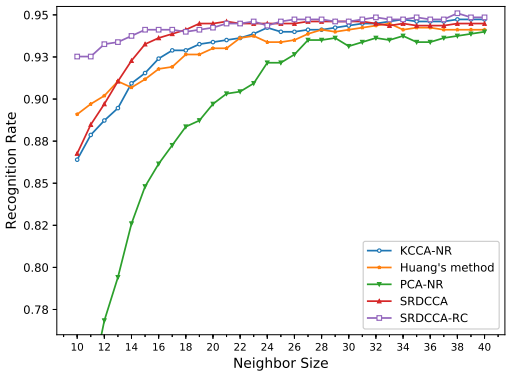

Fig. 2. Recognition rate vs neighborhood size.

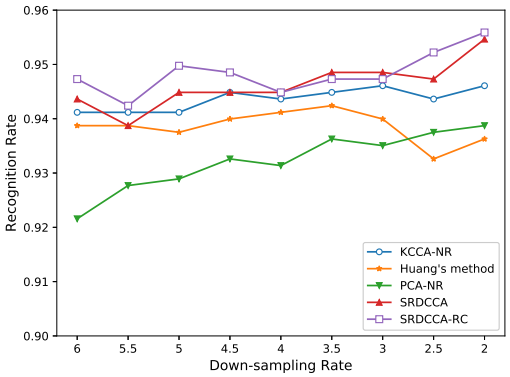

Fig. 3. Recognition rate vs down-sampling rate.

C^h and σ as the parameter of RBF. Accordingly, the weighting coefficient matrix can be obtained by $W_{RBF} = X^h\Phi^{-1}$. It is easily to calculate x_t^h with W_{RBF}. Then, the preliminary reconstruction image can be expressed as:

$$\tilde{i}_t^h = P_h x_t^h + \mu^h. \tag{7}$$

3.1.2. Residual compensation

During face reconstruction, many details are inevitably lost in space transformation and neighbor reconstruction processes. To improve the final reconstruction outcome, we add the processing of residual compensation.

First we use the method in section 3.1.1 to generate HR face set $\tilde{I}^h$ according to LR face training set I^l. Then we get HR and LR residual face sets:

$$R^h = I^h - \tilde{I}^h \text{ and } R^l = I^l - \text{downsample}(\tilde{I}^h). \tag{8}$$

Same form as (4) and (5), the HR residual r_t^h can be calculated by keeping the neighborhood relationship from R^l to R^h, then the high-quality image we eventually produce is

$$i_t^h = \tilde{i}_t^h + r_t^h. \tag{9}$$

3.2. FSR for face recognition

In most surveillance scenarios, we prefer to identify people rather than get better visual effect of human faces. It is regrettable that great majority of current algorithms are to reconstruct high-quality faces instead of enhancing the recognition accuracy.

For recognition purpose, we propose a new approach in this section. Let us start from (7) again. The feature of the new inputted LR image has been calculated as $\tilde{c}_t^h$ in the correlation space, as well as reconstructed HR images $\tilde{i}_t^h$. In the section 3.1.2, the residual compensation method performs well in the vision enhancement, but it also brings some noise, which is unprofitable to recognition tasks. Therefore, in the second step, we design a method to compensate the residuals on recognition feature directly.

First we reconstruct feature set $\tilde{C}^h$ according to the LR feature set C^l with method in section 3.1.1, and the corresponding HR set $\tilde{I}^h$ can be generated. Following the down sampling, we get $\tilde{I}^l$. Then we calculate correlational feature of $\tilde{I}^l$ by

$$\tilde{C}^l = W_f^{*\top} F\left(P_l^\top\left(\tilde{I}^l - \mu^l\right); \theta_f^*\right). \tag{10}$$

Then the feature residual sets are expressed as

$$E^h = C^h - \tilde{C}^h \text{ and } E^l = C^l - \tilde{C}^l. \tag{11}$$

We down-sample the reconstructed test image and extract the correlational feature c_t^l. Then, we minimize

$$\varepsilon = \left\| r_c^l - \sum\nolimits_{j=1}^k \beta_{t_j} E_{t_j}^l \right\| \ \ s.t. \sum\nolimits_{j=1}^k \beta_{t_j} = 1, \tag{12}$$

where feature residual $r_c^l = c_t^l - \tilde{c}_t^l$ and $\{E^l{}_{t_j}\}_{j=1}^k$ are nearest k neighbors of r_c^l. We apply the obtained weight $B = \{\beta_{t_j}\}_{j=1}^k$ on $\{E_{t_j}^h\}_{j=1}^k$ in E^h to reconfigure the corresponding HR feature residual:

$$r_c^h = \sum\nolimits_{j=1}^k \beta_{t_j} E_{t_j}^h. \tag{13}$$

Finally, we obtain the recognition feature

$$c^h = \tilde{c}_t^h + r_c^h \tag{14}$$

that can be fed to a classifier, such as the nearest neighbor classifier.

4. EXPERIMENT

In this section, we test the effectiveness of the proposed method in recognition and reconstruction two scenarios separately. Explanatorily, we denote the proposed method containing residual compensation as SRDCCA-RC, and the method without residual compensation as SRDCCA.

4.1. Recognition experiment

We carry out our SRDCCA and SRDCCA-RC methods on the CMU PIE and Yale-B databases to test recognition performance, and compare them with Bicubic-PCA, PCA-NR, Huang's method [12], and KCCA-NR [13]. Bicubic-PCA applies bicubic interpolation on LR face image, then extracts the principal component feature for recognition. PCA-NR carries out the neighborhood-based reconstruction to obtain the corresponding HR principal component feature from the LR one. The reconstructed feature is then used to recognize. In addition, the nearest neighbor (NN) classifier is used in all the experiments. In our method, we retain the 99% of spectral energy for HR and LR face images in the PCA transformation. In all hidden layers of DNNs, We select the rectified linear units as activation function defined as $h(x) = \max(x, 0)$.

4.1.1. Experiment on the CMU PIE database

We run two tests on the CMU PIE that is a face database rich in light, posture and angle variety. We select all 24 frontal face images for each person, and use all even numbered images as the training set, and leave the rest as the test set. In our method, the DNNs contain two hidden layers, each layer with 1000 cells. We set the learning rate as 0.088 and the max epoch as 150 in MBGD.

Fig. 2. demonstrates the accuracy of various approaches in the presence of different neighborhood sizes. In this task, the size of HR and LR face images is set to 64×64 and 16×16, and the final recognition features are fixed in 30 dimensions.

Fig. 3. illustrates the effectiveness of our approach under various down-sampling rates. In this task, we set the size of HR images to 64×64, the neighborhood size and the recognition feature dimension to 30, and let down-sampling rate change from 2 to 6.

In the above two tests, our approaches (SRDCCA, SRDCCA-RC) outperform other methods, showing a satisfactory robustness.

4.1.2. Experiment on the Yale-B database

The Yale-B database contains 5760 images of 10 subjects under 576 viewing conditions. We use subset-1 of the database to test the effectiveness of our approach for extremely low-resolution image recognizing. In the seven images of each person, the first four are used as training sets and the rest for testing. The HR face image size is still set to 64×64. Differently, LR images present very low resolution: 4×4. For parameter settings, the DNNs still consist of two hidden layers, each layer with 800 cells. In MBGD, we set the learning rate to 0.001 and the max epoch to 15. The experimental results are recorded in Fig. 4.

From the results, our approach shows a better recognition effect at exceedingly low resolution, and therefore has higher application value.

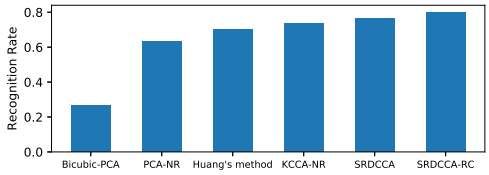

Fig. 4. Recognition rate on very low-resolution images.

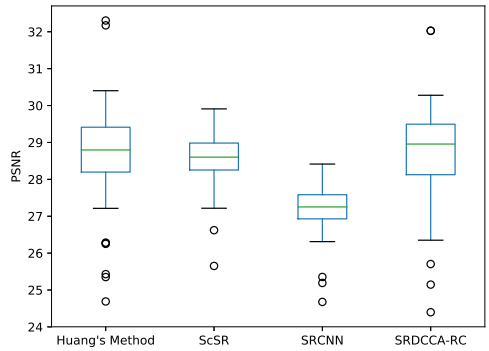

Fig. 5. PSNR on the CAS-PEAL database.

4.2. Reconstruction experiment

We evaluate the effect of our method by reconstructing the image on the CAS-PEAL database compared with Huang's method [12], ScSR [10] and SRCNN [5]. We use 1040 frontal face images for reconstruction task, in which 1000 images are used as the training set, and the rest as the test set. The size of HR and LR face images is set to 100×100 and 25×25. In our method, the structure of DNNs is the same as that in section 4.1.1. In MBGD, we set the learning rate to 0.088 and the max epoch to 150. The peak signal to noise ratio (PSNR) result is illustrated in Fig. 5. PSNR is defined as:

$$PSNR = 10 \log_{10} \left(\frac{peakval^2}{MSE} \right), \tag{15}$$

where $peakval$ is the maximum possible pixel value of the image. MSE is the mean square error between the reconstructed image and the original image. The higher the PSNR value is, the lower the picture distortion becomes.

5. CONCLUSION AND FUTURE WORK

In this paper, we propose a DCCA-based approach to address the trouble of the LR face recognition and reconstruction. Many experimental results have shown that our proposed method is promising both in recognition and reconstruction. Considering the architecture of DNNs is designed simply and parameters selection is based on experience, solving these problems is what we are going to do in the future.

6. REFERENCES

[1] H. Chang, D. Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2004, pp. 275–282.

[2] X. Lu, Y. Yuan, and P. Yan, "Image super-resolution via double sparsity regularized manifold learning," *IEEE Transactions on Circuits & Systems for Video Technology*, vol. 23, no. 12, pp. 2022–2033, 2013.

[3] J. Yang, Z. Wang, Z. Lin, S. D. Cohen, and T. S. Huang, "Coupled dictionary training for image super-resolution," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3467–78, 2012.

[4] Q. Pan, Y. Liang, L. Zhang, and S. Wang, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2216–2223.

[5] C. Dong, C. L. Chen, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 38, no. 2, pp. 295, 2016.

[6] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, and et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 4681–4690.

[7] Y. Song, J. Zhang, S. He, L. Bao, and Q. Yang, "Learning to hallucinate face images via component generation and enhancement," in *International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, 2017, pp. 4537–4543.

[8] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Learning face hallucination in the wild," in *AAAI Conference on Artificial Intelligence*. AAAI Press, 2015, pp. 3871–3877.

[9] B. Li, H. Chang, S. Shan, X. Chen, and W. Gao, "Hallucinating facial images and features," in *International Conference on Pattern Recognition*. IEEE, 2008, pp. 1–4.

[10] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861, 2010.

[11] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 4792–4800.

[12] H. Huang, H. He, X. Fan, and J. Zhang, "Super-resolution of human face image using canonical correlation analysis," *Pattern Recognition*, vol. 43, no. 7, pp. 2532–2543, 2010.

[13] Z. Zhang, Y. Yuan, Y. Li, B. Li, and J. Qiang, "Face hallucination and recognition using kernel canonical correlation analysis," in *International Conference on Neural Information Processing*. Springer, 2017, pp. 633–641.

[14] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, "Unsupervised learning of acoustic features via deep canonical correlation analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2015, pp. 4590–4594.