Two-stage Patch-based Multi-View Face Super-resolution

Zhuo Hui and Kin-Man Lam
Centre for Signal Processing, Department of Electronic and Information Engineering,
The Hong Kong Polytechnic University, Kowloon, Hong Kong
E-mail: {06824233d, enkmlam}@polyu.edu.hk Tel: +852-27666207

***Abstract*— In this paper, we propose a learning-based method to generate a high-resolution (HR) face in frontal view from a low-resolution (LR) face in an arbitrary pose. This HR virtual face (HRVF) method is based on two stages of pixel-structure learning. In the first stage of our algorithm, initially estimated HR frontal-view images are generated from non-frontal-view LR input images, based on a patch-based learning method. In the second stage, the estimated frontal-view image will be used to search for similar faces from the interpolated LR frontal-view face database. The targeted HR frontal-view face image is then constructed based on the local patches of the HR faces of the corresponding LR face images in the database. Experiments show that the proposed algorithm can produce a better performance than existing methods.**

I. INTRODUCTION

Face recognition can be highly accurate when state-of-the-art technologies are used under restricted settings, such as faces in frontal view, under even lighting conditions, etc., as described in [1], [2]. However, variations in facial appearance and poses cause great difficulty in face recognition, significantly degrading the recognition accuracy. In most real situations, images are captured in uncontrolled conditions, which may be of low resolution and non-frontal view. However, most of the current facial-image super-resolution (SR) methods focus on single-view or frontal-view learning, e.g. [3]-[7] proposed recently. Therefore, an accurate and efficient facial-image SR algorithm is necessary to generate a HR frontal-view facial image from its corresponding LR facial images in an arbitrary pose.

The key issue in the HRVF problem is to establish the outline of a frontal view or to estimate the low-frequency content of a face, which is different from the objective of reconstructing the high-frequency content in frontal-view face SR methods. Therefore, only a few methods to solve the problem have been proposed so far. Jia et al. [8] proposed a patch-based method to reconstruct HR images in two steps with the use of a tensor and a Bayesian field probability model. Both the tensor and the probability model require high computational complexity. Vetter [9] extended the scope of the HRVF problem to the 3-D space, and solved it with a linear-object class method. Inspired by the ideas proposed in linear-object classes, [10] and [11] proposed a two-step patch-based local linear regression (LLR) method to reconstruct HR images. Unlike Vetter's method, the LLR method attempts to predict the HR frontal-view facial images in the 2-D domain. Therefore, linear-object classes are established in the 2-D space. The two methods consider patches as the basic units of linear classes, and apply LLR to estimate the correspondence existing in the patches between frontal and non-frontal face images.

In our algorithm, we first use patch-based learning to reconstruct an initial frontal-view face image, and then local-structure learning is used to refine and improve the correspondence relationship learnt in the first phase. Experimental results show that our proposed method performs well in reconstructing local facial features and sharp edges.

The remainder of the paper is organized as follows. In Section II, a mathematical relation between face images of frontal view and non-frontal view is derived, and the advantages of our proposed algorithm are described. In Section III, our algorithm is presented in detail. Finally, experimental results are shown in Section IV, and a conclusion is given at the end.

II. MATHEMATICAL RELATIONSHIP BETWEEN FRONTAL-VIEW AND NON-FRONTAL-VIEW FACES

In this section, we will derive the linear relationship existing between a non-frontal-view face and its corresponding frontal-view image. We assume that the 3-D face surface is Lambertian, and a 3-D face is modeled using a cylinder [10]. Although the face geometries of two different persons are not exactly the same, they should be similar as the spatial configuration of facial features is similar on all faces. Hence, as in [9] and [10], an approximate linear relationship can be established between the frontal and non-frontal views by means of linear regression.

For a given lighting source $\vec{s}$, the intensity at each point in the 3-D surface can be represented as follows:

$$\Gamma(x, y, z) = \rho(x, y, z)\cos\alpha\,, \tag{1}$$

where $\Gamma(x, y, z)$ represents the intensity function at the point (x, y, z) in the 3-D space, $\rho(x, y, z)$ denotes the albedo of the point in the 3-D surface, and α represents the angle between the normal at the point and the lighting source.

Let $\mathbf{I}_0$ denote a frontal-view facial image under a frontal light source (in this case, α is equal to zero), and $\mathbf{P}_0$ the corresponding projection matrix, which projects points in the 3-D space onto the 2-D image plane. Then, we have

$$\mathbf{I}_0 = \mathbf{P}_0\Gamma. \quad (2)$$

Similarly, a non-frontal view of the face image, denoted as $\mathbf{I}_n$, can be generated with the projection matrix $\mathbf{P}_n$, as follows:

$$\mathbf{I}_n = \mathbf{P}_n\Gamma. \quad (3)$$

The projection matrix $\mathbf{P}_n$, which projects a 3-D face in pose n onto the 2-D image plane, is determined by both the viewpoint taken and the 3-D geometry of the face under consideration. $\mathbf{P}_n$ is a $m \times k$ matrix, where m and k is the number of pixels in $\mathbf{I}_n$ and the number of points on the 3-D face surface Γ, respectively. The elements in the projection matrices are either 1 or 0. If a pixel in the 3-D space is occluded in the 2-D domain, the corresponding position in the projection matrix is filled with 0. Otherwise, it is 1.

The intensity function Γ in the 3-D space is reconstructed using (3) so as to estimate $\mathbf{I_0}$ in (2). However, Γ cannot be predicted; this is an ill-posed problem, as the number of pixels in $\mathbf{I}_n$ is less than that in Γ. We consider those visible points in the 2-D images, and define the assembly of the visible points in the 3-D space Γ as $\widetilde{\Gamma}$, which contains the same number of pixels as $\mathbf{I}_n$ (i.e. $m = k$). Hence, the projection matrix $\mathbf{P}_n$ can be considered unitary; the number of pixels in the 2-D space is equal to that in the 3-D space. Consequently, we have

$$\widetilde{\Gamma} = \mathbf{P}_n^T\mathbf{I}_n = \mathbf{P}_n^T\mathbf{P}_n\Gamma. \quad (4)$$

Using the assumption [11] that the missing points in the 3-D space can be interpolated or extrapolated from the existing points in $\mathbf{I}_n$, the relations between $\widetilde{\Gamma}$ and Γ can be shown as:

$$\Gamma \cong \tilde{\Gamma} + \varepsilon\mathbf{I}_n = \mathbf{P}_n^T\mathbf{I}_n + \varepsilon\mathbf{I}_n, \quad (5)$$

where ε is a matrix representing the neighborhood relationship between $\mathbf{I}_n$ and Γ. Therefore, substitute the results in (5) into (2), we have

$$\begin{aligned}\mathbf{I_0} &= \mathbf{P_0}\Gamma \\ &\approx \mathbf{P}_0(\mathbf{P}_n^T\mathbf{I}_n + \varepsilon\mathbf{I}_n) = (\mathbf{P}_0\mathbf{P}_n^T + \mathbf{P}_0\varepsilon)\mathbf{I}_n. \quad (6)\end{aligned}$$

Denote $$\mathbf{M}_n = \mathbf{P}_0\mathbf{P}_n^T + \mathbf{P}_0\varepsilon, \quad (7)$$

then $$\mathbf{I}_0 \approx \widetilde{\mathbf{I}}_0 = \mathbf{M}_n\mathbf{I}_n. \quad (8)$$

As mentioned previously, $\mathbf{M}_n$ depends on the 3-D geometry of the person under consideration. Therefore, $\mathbf{M}_n$ varies for different people, unless we establish a 3-D model, as did Vetter [9]. In [10][11], linear regression is applied to estimate the relationships. If $\mathbf{I_0}$ and $\mathbf{I}_n$ are considered to form linear classes with a correspondence relationship and $\mathbf{M}_n$ approximates the linear relation between these two classes, then some information about local features may be lost or distorted, as the corresponding pixels in two classes may not be linearly related. In other words, a uniform representation of the correspondence relationship for a certain class may not work accurately for local structures and features. Therefore, patched-based estimation is employed to improve the accuracy. The size of the patches to be used greatly affects the accuracy of the estimation of $\mathbf{M}_n$, in particular when the number of training samples in a certain class is small. However, with patch-based learning, the size of a class is usually small, which may cause degradation in the estimation. For all of these reasons, we aim to take more pixels in patches into consideration, and learn the relation of these pixels with respect to the corresponding pixels in the target HR frontal-view images. Suppose that $\mathbf{I_0}$ and $\mathbf{I}_n$ in (8) are of low resolution, and we reconstruct an approximate frontal-view HR face image in the first phase. This initially estimated frontal-view image can be considered as an interpolation of its LR frontal view. Thus, these initially estimated images can be projected back to the interpolated LR space formed from the training samples and refine the initially predicated structures. Our method employs eigentransformation [16], in which the target HR images are constructed using PCA and the linear relationship existing between the HR and the LR space. In our case, eigentransformation is applied to patches so as to retain the local structures. The major contribution of this paper is to propose a novel method to generate an initial estimation of frontal-view face images, and to apply back projection and eigentransformation to patches to refine local structures.

III. Two Stages of Patch-based Learning

In our paper, we propose to use two stages of patch-based learning with PCA to reconstruct a HR frontal-view image from a LR face in an arbitrary pose. In the first stage, the frontal view of a face is estimated based on the learned relationship between patches of LR non-frontal-view images and the corresponding HR frontal-view images in the training dataset. The input non-frontal-view LR image is partitioned into overlapped patches, and the weights of similar patches in the non-frontal LR database that contribute to the input LR patches are learned using PCA. These weights are projected on to the space containing the corresponding HR frontal-view patches, and the initial frontal-view images can be estimated. In the next stage, these estimated images are partitioned into patches. Based on the weights learned for each patch in the LR frontal-view images in the training set that contribute to the estimated images, eigentransformation is applied to generate the corresponding weights of the HR frontal-view images for constructing the target HR frontal images.

A. First Stage of PCA-based Learning

Hu et al. in [12] proposed a local-structure learning method based on the assumption that two similar face images should have similar local pixel structures, and thus these local structures can be learnt for SR reconstruction. Hence, to improve the estimation accuracy, we first reconstruct the frontal-view templates of an input face using patch-based learning. This stage aims to generate the outline or appearance of the target HR frontal-view facial images. Suppose that each facial image is divided into N overlapped patches. In the training set, the patches of the i^{th} LR non-frontal image $\mathbf{L}_n^i$ and the corresponding patches of its HR frontal-view facial image $\mathbf{H}_0^i$ are denoted as $\mathbf{L}_n^i = \{l_{i,n}^1, l_{i,n}^2, \ldots, l_{i,n}^N\}$ and

$\mathbf{H}_0^i = \{h_{i,0}^1, h_{i,0}^2, \ldots, h_{i,0}^N\}$, respectively. Similarly, the patches of an input LR non-frontal-view image are denoted as $\mathbf{L}_n^t = \{l_{t,n}^1, l_{t,n}^2, \ldots, l_{t,n}^N\}$. The reconstruction error ξ can be expressed as follows:

$$\xi^u = \left\| l_{t,n}^u - \sum_{v \le N} c_{uv} l_{i,n}^v \right\|, \text{ where } 1 \le u \le N. \quad (9)$$

c_{uv} represents the weights that each patch, indexed by v, in a different LR training image contributes to the reconstruction of the u^{th} patch. Denote V_u and Λ_u as the matrices of the eigenvectors and the eigenvalues, respectively, of the matrix $(l_{t,n}^u - L_u)^T (l_{t,n}^u - L_u)$, where L_u is the mean patch matrix of the u^{th} patch. Then, the eigen matrix of $(l_{t,n}^u - L_u)(l_{t,n}^u - L_u)^T$ and the corresponding weight matrix W_u can be computed as follows:

$$E_u = (l_{t,n}^u - L_u) V_u \Lambda_u^{-1/2} \text{ and } W_u = E_u^t (l_{t,n}^u - L_u). \quad (10)$$

Hence, the input u^{th} patch can be expressed as follows:

$$l_{t,n}^u = E_u W_u + L_u. \quad (11)$$

Substitute (10) into (11), we have

$$l_{t,n}^u = (l_{t,n}^u - L_u) V_u \Lambda_u^{-1/2} W_u + L_u. \quad (12)$$

The weights that each patch in the training set contributes to the input can be expressed as $c_{uv} = V_u \Lambda_u^{-1/2} W_u$. With these weights, the corresponding HR patches of the HR frontal-view facial image can be reconstructed as follows:

$$h_{t,0}^u = \sum_{u,v \le N} c_{uv} \left(h_{i,0}^u - H_u \right) + H_u, \quad (13)$$

With all the estimated patches, an initial frontal-view HR facial image can be reconstructed.

B. Second Stage of Patch-based PCA Back-projection Learning

Using the estimated frontal-view images generated in the first stage, we aim to refine their local structures and textures in this stage. Patch-based eigentransformation is performed by extracting the features of the patches, so as to refine the correspondence between linear-object classes. Compared to the local-structure learning [12] and the neighboring-embedding method [13], [14], eigentransformation can reduce the negative effect of the structure of the initially estimated images on the target HR frontal-view images. This is due to the fact that [12-14] are based on the assumption that input LR images possess either a similar appearance outline to or mostly similar low-frequency content to the target HR images. However, this is not true in cases of multi-view super-resolution. As the initially estimated images are predicated on non-frontal-view LR samples, the reconstructed structure needs to be further refined. In other words, the initially estimated images cannot be considered to contain most of the low-frequency content of the target HR ones. However, as LR frontal-view face images and the corresponding HR images possess similar local structures, we can project the patches of the initially estimated images back to the LR frontal-view patch space so as to derive the weights that each LR frontal-view sample contributes to the patches of the initially estimated images. As illustrated in [12], the center pixel in each patch can be approximately represented by its neighbors as follows:

$$I_{re}(x,y) \approx \sum_{u,v} w_{u,v}(x,y) \times I_{re}(x+u, y+v), \quad (14)$$

where $I_{re}(x, y)$ refers to the frontal-view patch generated by LR frontal-view images in the training set, $w_{u,v}(x, y)$ represents the weights of the neighboring pixels contributed to the reconstruction of the center pixel, and can be determined by the following relations:

$$w_{u,v}(x,y) = c_{x,y} \exp\left\{ -\frac{\left(I_{re}(x,y) - I_{re}(x+u, y+v)\right)^2}{\sigma^2} \right\}, \quad (15)$$

where $c_{x,y} = \dfrac{I(x,y)}{\sum_{u,v} w'_{u,v}(x,y) \times I(x+u, y+v) + \varepsilon}$.

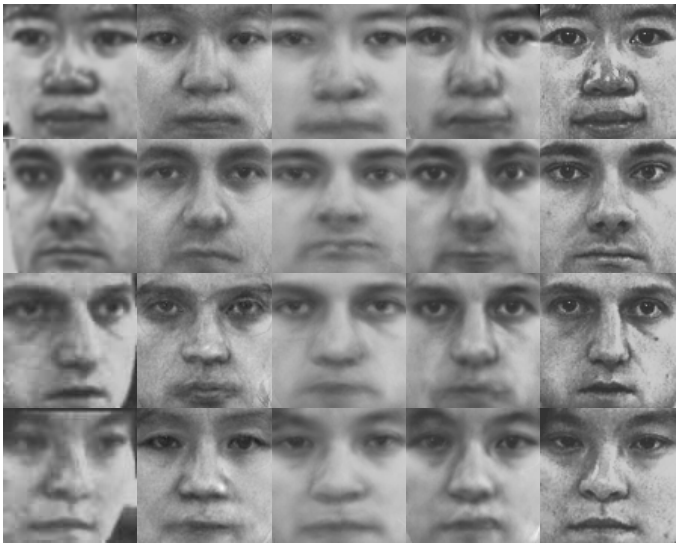
ε is a small positive value that prevents the denominator from being zero and $I(x, y)$ denotes the corresponding input patches of initial estimated images. The weights $w_{u,v}(x, y)$ that we have learned for the patches can be viewed as the priors of the local structures. Similar to the first stage, PCA is used to search for reference patches from the interpolated LR face image from the training set. By means of eigentransformation [16], we can directly apply the same weights in the space of HR frontal-view images to reconstruct the target HR images, as we did in the first phase. That is, the weights used to express a certain patch of an initially estimated image can be directly applied in the reconstruction of the target HR image.

IV. EXPERIMENT RESULTS

In the experiments, five pose subsets of the CMU PIE database [15] were used. The poses covered included ±45°, ±22.5°, and frontal view. The numbers of images selected for training and testing were both 68. Leave-one-out was used for testing. All the faces were coarsely aligned based on the positions of the two eyes, and were cropped to a size of 90×90. The LR images are generated by applying to the HR images a Gaussian blurring kernel of size 7×7, and down-sampled by a factor of 2 in both the horizontal and vertical directions. The size of the patches is set at 7×7, and the overlapping size is 6 in each direction. The performance of our proposed method is compared with Jia's method [8] and the locally-linear regression method [10], [11] in terms of the Mean Squared Error (MSE), the Edge Stability Mean Squared Error (ESMSE), and the Structural Similarity Index (SSIM). Table I

tabulates the performances of these methods. Fig. 1 demonstrates the visual qualities of the final reconstructed images using the three methods. In view of the ESMSE values, our proposed method can perform well in producing sharp edges and in retrieving detail information. Moreover, our method can improve the accuracy of local feature reconstruction. As illustrated in Fig. 1 and based on the SSIM values in Table I, our proposed method is robust in terms of facial-feature reconstruction. For example, our proposed algorithm can better reconstruct the beard around the mouth of the second image in Fig. 1. Moreover, our proposed method can reconstruct details that are more similar to the target HR image.

Our method was simulated on a computer of 1.67GHz CPU with 1GByte DRAM, using MatLab 7.2. The entire runtime required by our algorithm is about 18.9 seconds, while Jia's method and the locally-linear regression method require about 15.4s and 21.7s, respectively.



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

Fig. 1 The HR frontal-view images reconstructed using different methods: (a) the input LR non-frontal-view image, (b) Jia's method [8], (c) the LLR reconstruction method [10],[11], (d) our local-structure learning method, and (e) the ground-truth images.

TABLE I THE DIFFERENT CRITERIA (MSE, ESMSE, SSIM) MEASURED BASED ON JIA'S METHOD, THE LLR RECONSTRUCTION METHOD, AND OUR PROPOSED LOCAL-STRUCTURE LEARNING.

	Jia's Method	LLR	Our method
MSE	485.45	423.42	365.51
ESMSE	4.39	4.67	4.43
SSIM	0.48	0.56	0.61

V. CONCLUSION

In this paper, we have proposed a two stage learning based method to estimate HR frontal-view facial images when LR non-frontal-view facial images are given. With the help of a two-phase approach, our algorithm employs a PCA patch based reconstruction method to outline the structure of a whole face as a global approach. In the second phase, locally structure and texture has been refined through weights back projection and eigentransformation implemented in patch level. Experiment results have shown that our proposed algorithm can perform better than Jia's method and locally-linear regression in terms of both visual quality and PSNR.

ACKNOWLEDGEMENT

This project was supported by an internal grant from The Hong Kong Polytechnic University, Hong Kong (G-U932).

REFERENCES

[1] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey", Proc. IEEE, vol. 83, no.5, pp. 705-740, 1995.

[2] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature surveys", ACM Comput. Surv. vol. 35, no. 4, pp. 399-458, 2003.

[3] W. T. Freeman, T. R. Jones, E. C. Pasztor, "Example-based super-resolution", IEEE Computer Graphics and Applications. vol.22, no.5, pp. 56-65, 2000.

[4] W. T. Freeman, E. C. Pasztor, "Learning low-level vision", International Journal of Computer Vision, vol. 40, no. 1. pp. 25-47, 2000.

[5] Z. Hui, K. M. Lam, "Wavelet-based eigentransformation for face super-resolution" Proc. Pacific-Rim conference on multimedia (PCM) vol. 2, pp. 226-234, 2010.

[6] Y. He, K. H. Yap, L. Chen, L. P. Chau, "A Nonlinear Least Square Technique for Simultaneous Image Registration and Super-Resolution" IEEE Tran. Image Processing, vol. 16, no. 11, pp. 2830-2841, 2007.

[7] J. Sun, N. N. Zheng, H.Tao, H. Y. Shum, "Image hallucination with primal sketch priors", Proc. IEEE conference on Computer Vision and Patteren Recoginition (CVPR), vol. 2, 2003.

[8] K. Jia, S.G. Gong, "Generalized Face Super-Resolution", IEEE Trans. Image Processing, vol. 17, no. 6, pp. 873-886, 2008.

[9] T. Vetter, T. Poggio, "Linear Object classes and image synthesis from a single example image". IEEE Trans. Pattern Analysis and Machine Intellig., vol. 19, no. 7, pp. 733-742, 1997.

[10] K.M. Lam and H. Yan, "An Analytical-to-holistic Approach for Face Recognition Based on a Single Frontal View", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 20, no. 7, pp. 673-86, 1998.

[11] X. Ma, H. Huang, S. Wang, C. Qi, "A simple approach to multiview face hallucination", IEEE Signal Processing Letters, vol. 17, no. 6, pp. 579-582, 2010.

[12] Y. Hu, K. M. Lam, G. Qiu, T. Shen, "From local pixel structure to global image super-resolution: a new face hallucination framework", IEEE Trans. on Image Processing, vol. 20, no. 2, pp. 433-445, 2011.

[13] W. Fan, D. Y. Yeung, "Image Hallucination using neighbor embedding over visual primitive manifolds", In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1-7, 2007.

[14] J. Tenebaum, V. Silva, J. Langford, "A global geometric framework for nonlinear dimensionality reduction", Science, 290 (5500), pp. 2319-2323, 2000.

[15] T. Sim, S. Baker, M. Bsat, "The CMU pose, illumination, and expression (PIE) database", in Proc 5th Int Conf Auto. Face and Gesture Recognition, Washington, DC, pp. 46-51, 2002.

[16] X. Wang, X. Tang, "Hallucinating face by eigentransformation" IEEE Trans. Systems Man Cybernet. 35, pp.425-434, 2006.