NEIGHBOR COMBINATION AND TRANSFORMATION FOR HALLUCINATING FACES

Wei Liu, Dahua Lin, and Xiaoou Tang

Department of Information Engineering
The Chinese University of Hong Kong, Shatin, Hong Kong
{wliu, dhlin4, xtang}@ie.cuhk.edu.hk

ABSTRACT

In this paper, we propose a novel face hallucination framework based on image patches, which exploits local geometry structures of overlapping patches to hallucinate different components associated with one facial image. To achieve local fidelity while preserving smoothness in the target high-resolution image, we develop a neighbor combination super-resolution model for high-resolution patch synthesis. For further enhancing the detailed information, we propose another model, which effectively learns neighbor transformations between low- and high-resolution image patch residuals to compensate modeling errors caused by the first model. Experiments demonstrate that our approach can hallucinate high quality super-resolution faces.

1. INTRODUCTION

In recent years, the face recognition technique has shown increasing significance in multimedia applications. To raise the performance of a face recognition system, it is often useful to render a high-resolution face image from the low-resolution one, which is called face hallucination or face super-resolution. A large quantity of super-resolution techniques have been proposed in the past decade [1][2][4][5]. Most of them are based on synthesis of local texture, and are usually applied to generic images without special consideration on the special characterisitics of face images.

Baker and Kanade [1][2] develop a hallucination method targeted at face image. It infers the high frequency components from a parent structure by recognizing the local features from the training set. Liu et al.[5] develop a two-step statistical approach integrating a global parametric model and a local model with Markov Random Field. Both of the two methods are based on an known down-resolution function, which is sometimes difficult to obtain in practice.

Instead of using a probabilistic model, Chang et al. [3] enforce the linear combination in a local neighborhood (i.e. k-NN) of each image patch based on the assumption that the the low-resolution and the high-resolution patch space share similar local manifold structure, which is inspired by the recent manifold learning methods, locally linear embedding (LLE).

In this paper, we propose a novel framework based on image patches for solving single facial image super-resolution problems. We develop neighbor combination and neighbor transformation models for face hallucination.

2. SUPER-RESOLUTION THROUGH NEIGHBOR COMBINATION

As in Locally Linear Embedding(LLE), our hallucination algorithm is based on the assumption that small patches in low resolution space and high resolution space form manifolds with the same local structure, which is characterized by the weights of neighboring patches. Therefore we can approximately assume that corresponding patches in low- and high-resolution share the same weight vectors and constituent samples. Then we can synthesize the high-resolution patch employing the weights analyzed according to input low-resolution patches.

We denote the available low-resolution face images as $\{I_L^{(i)}\}_{i=1}^n$ and their high-resolution counterparts as $\{I_H^{(i)}\}_{i=1}^n$, where n is sample number. Each image in the training set is divided into a set of m small overlapping image patch sets $\{\mathbf{p}_{L,j}^{(i)} | \mathbf{p}_{L,j}^{(i)} \in I_L^{(i)}, 1 \le j \le m, 1 \le i \le n\}$ and $\{\mathbf{p}_{H,j}^{(i)} | \mathbf{p}_{H,j}^{(i)} \in I_H^{(i)}, 1 \le j \le m, 1 \le i \le n\}$. The same division configuration is adopted for low resolution images and high resolution images in order to establish the correspondence between patches.

In testing stage, given a low-resolution image I_L as input, we firstly decompose it into overlapping patches $\{\mathbf{p}_{L,j}\}_{j=1}^m$ as a same way. Then, we use a linear combination over k-Nearest Neighboring samples of each low-resolution patch to reconstruct the patch. The optimal weight vector is achieved

0-7803-9332-5/05/$20.00 ©2005 IEEE

by minimizing the following reconstruction error

$$\varepsilon(\mathbf{w}_j) = \|\mathbf{p}_{L,j} - \sum_{i=1}^{k} \omega_{ij} \mathbf{p}_{L,j}^{(N(j,i))}\|^2$$
$$s.t. \quad \sum_{i=1}^{k} \omega_{ij} = 1, \quad j = 1, 2, \cdots, m \tag{1}$$

where $\mathbf{w}_j$ is the local weight vector associated with the patch $\mathbf{p}_{L,j}$, $N(j,i)$ denotes the index of face identity with rank-i similarity at the jth patch. Solving the least squares problem with constraint (eq.(1)) gives

$$\omega_{ij} = \frac{\sum_{t=1}^{k} (\mathbf{R}_j)_{it}}{\sum_{t=1}^{k} \sum_{s=1}^{k} (\mathbf{R}_j)_{st}}, \ \mathbf{R}_j = \mathbf{Q}_j^{-1},$$
$$(\mathbf{Q}_j)_{st} = (\mathbf{p}_{L,j} - \mathbf{p}_{L,j}^{(N(j,s))})^{\mathrm{T}} (\mathbf{p}_{L,j} - \mathbf{p}_{L,j}^{(N(j,t))}) \tag{2}$$

where $\mathbf{R}_j$ and $\mathbf{Q}_j$ are both $k \times k$ matrices. To avoid the singularity problem of $\mathbf{Q}_j$, regularization technique with parameter r is introduced to $\mathbf{Q}_j$ before inversion; i.e. $\mathbf{R}_j = (\mathbf{Q}_j + r\mathbf{I})^{-1}$. After $\mathbf{w}_j$ is solved, we synthesize the high-resolution patch $\mathbf{p}_{H,j}$ one-to-one corresponding to $\mathbf{p}_{L,j}$ as below

$$\hat{\mathbf{p}}_{H,j} = \sum_{i=1}^{k} \omega_{ij} \mathbf{p}_{H,j}^{(N(j,i))} \tag{3}$$

Based on the formulation and derivation shown above, we develop an algorithm for hallucinating faces through neighbor combination. The key of our algorithm is to preserve linear combination weights.

To assure local compatibility and smoothness between the hallucinated patches, we superpose patches in adjacent regions of one image and blend pixels in the overlapping area to form one whole facial image.

- **Step 1.** For each patch $\mathbf{p}_{L,j} (j = 1, 2, \cdots, m)$ in an input low-resolution face I_L:

  (**1.1**) Find k_1-NN(Nearest Neighbors) of different people in the training patch ensemble $\{\mathbf{p}_{L,j}^{(i)}\}_{i=1}^{n}$, denote them as $N(j,i)(i = 1, \cdots, k_1)$.

  (**1.2**) Compute the reconstruction weight $\mathbf{w}_j$ based on the selected neighbors using (2), which just minimizes error of reconstructing $\mathbf{p}_{L,j}$.

  (**1.3**) Synthesize the high-resolution patch $\hat{\mathbf{p}}_{H,j}$ applying $\mathbf{w}_j$ to perform linearly weighted combination over the neighbor patches as (3).

- **Step 2.** Concatenate and integrate the hallucinated high-resolution patches $\{\hat{\mathbf{p}}_{H,j}\}_{j=1}^{m}$ to form one facial image $\hat{I}_H$, which is the target high-resolution facial image.

3. RESIDUE SUPER-RESOLUTION THROUGH NEIGHBOR TRANSFORMATION

3.1. Generalized Singular Value Decomposition (GSVD)

Generalized Singular Value Decomposition (GSVD) is a powerful mathematical method which can be employed to model the correlation between two spaces with different dimensions. When two related data sets $\mathbf{X} \in \Re^{d_1 \times n}$ and $\mathbf{Y} \in \Re^{d_2 \times n}$ with an equal number of observations are given, a natural question arising is how to exploit the correlations between them to estimate one set from the other. In general, the dimensionality, complexity, and energy of the data sets is different. These issues in conjunction with high dimensionality of data present a number of challenges.

In particular, we only discuss the case that the number of samples is small with respect to dimensions of data, i.e. $n < d_1 < d_2$. Performing GSVD on the matrices $\mathbf{X}$ and $\mathbf{Y}$ results in simultaneously decomposition on the two matrices as follows

$$\mathbf{X} = \mathbf{U}\mathbf{C}\mathbf{H}^{\mathrm{T}} \tag{4}$$
$$\mathbf{Y} = \mathbf{V}\mathbf{S}\mathbf{H}^{\mathrm{T}} \tag{5}$$

where $\mathbf{U}(d_1 \times d_1)$ and $\mathbf{V}(d_2 \times d_2)$ are both unitary matrices, the square matrix $H(n \times n)$ is shared in two decompositions, and nonnegative diagonal matrices $\mathbf{C}(d_1 \times n)$ and $\mathbf{S}(d_2 \times n)$ satisfying $\mathbf{C}^{\mathrm{T}}\mathbf{C} + \mathbf{S}^{\mathrm{T}}\mathbf{S} = \mathbf{I}$. Realizing that the nonzero elements of $\mathbf{C}$ are on its main diagonal, we only utilize the first n rows (saved in $n \times n$ square matrix $\dot{\mathbf{C}}$) of $\mathbf{C}$ and the first n columns (saved in $d_1 \times n$ matrix $\dot{\mathbf{U}}$) of $\mathbf{U}$ and have the following relationship

$$\mathbf{U}\mathbf{C} = [\dot{\mathbf{U}} \ \ \bar{\mathbf{U}}] \begin{bmatrix} \dot{\mathbf{C}} \\ \mathbf{0} \end{bmatrix} = \dot{\mathbf{U}}\dot{\mathbf{C}} \tag{6}$$

Combine (4) and (6), we derive $\mathbf{X} = \dot{\mathbf{U}}\dot{\mathbf{C}}\mathbf{H}^{\mathrm{T}}$. Furthermore, if the diagonal matrix $\dot{\mathbf{C}}$ is invertible, with consideration that $\dot{\mathbf{U}}$ is orthogonal, the share matrix $\mathbf{H}^{\mathrm{T}}$ is solved as $\dot{\mathbf{C}}^{-1}\dot{\mathbf{U}}^{\mathrm{T}}\mathbf{X}$ and is substituted into (5), thus we derive below

$$\mathbf{Y} = \mathbf{V}\mathbf{S}\mathbf{H}^{\mathrm{T}} = \mathbf{V}\mathbf{S}\dot{\mathbf{C}}^{-1}\dot{\mathbf{U}}^{\mathrm{T}}\mathbf{X} = \mathbf{T}\mathbf{X} \tag{7}$$

where $\mathbf{T} = \mathbf{V}\mathbf{S}\dot{\mathbf{C}}^{-1}\dot{\mathbf{U}}^{\mathrm{T}}$ is the resultant transformation matrix which can transform $\mathbf{X}$ to $\mathbf{Y}$.

3.2. Residue Compensation Using Neighbor Transformation

Realizing that the neighbor combination super-resolution learning model will lose detailed information of face images inevitably, single super-resolution learning model is not enough to achieve satisfactory results. Another learning model is applied to establish the relation between low-resolution image residual and high-resolution image residual caused by the first model, which well compensate the hallucinated result acquired in the first model.

Because GSVD provides the capability of establish the relation between two spaces of different dimensions, we achieve above goal through explicitly learning a GSVD-based transformation between low- and high-resolution image residual. Due to the fact that GSVD is not capable of dealing with high dimensional data such as images, we still divide whole images as overlapping patches as the same mode to the first model, so GSVD is performed on each pairwise patch residue.

Specifically, we split the training sets into two disjoint halves. The first half is for neighbor combination super-resolution learning. Take the low-resolution images of remain half training data as test images for the first model, hallucinate faces directly. Then two residual images for each sample can be constructed as below: one is obtained by subtracting the low-resolution image by a down-sampled version of the hallucinated image, the other is obtained by subtracting the actual high-resolution image by the hallucinated image. Thus, we acquire pairwise residue image sets $\{R_L^{(i)}\}_{i=1}^{l}$ and $\{R_H^{(i)}\}_{i=1}^{l}(l = n/2)$. Afterwards, we gain available residue image patches as $\{\mathbf{r}_{L,j}^{(i)}|\mathbf{r}_{L,j}^{(i)} \in R_L^{(i)}, 1 \leq j \leq m, 1 \leq i \leq l\}$ and $\{\mathbf{r}_{H,j}^{(i)}|\mathbf{r}_{H,j}^{(i)} \in R_H^{(i)}, 1 \leq j \leq m, 1 \leq i \leq l\}$.

With a low-resolution image residual R_L given as input, we divide it into overlapping patches $\{\mathbf{r}_{L,j}\}_{j=1}^{m}$ the same to the first model. This time we abandon the neighbor combination trick to estimate a high-resolution patch residual since most of those residual vectors are sparse, which will deteriorate linear reconstruction quality. Instead we apply neighbor transformations to implement residue super-resolution with help of GSVD. The detailed algorithm is as below

- **Step 1.** For each residue patch $\mathbf{r}_{L,j}(j = 1, 2, \cdots, m)$ in an input low-resolution residue face image R_L:

  (**1.1**) Find k_2-NNs of different people in the residue patch ensemble $\{\mathbf{r}_{L,j}^{(i)}\}_{i=1}^{l}$, denote them as $N(j, i)(i = 1, \cdots, k_2)$.

  (**1.2**) Calculate the transformation matrix $\mathbf{T}_j$ as (7), through performing GSVD on matrices $\mathbf{X}_j = [\mathbf{r}_{L,j}^{(N(j,1))}, \cdots, \mathbf{r}_{L,j}^{(N(j,k_2))}]$ and $\mathbf{Y}_j = [\mathbf{r}_{H,j}^{(N(j,1))}, \cdots, \mathbf{r}_{H,j}^{(N(j,k_2))}]$. (Make sure that k_2 is not larger than the dimension of low-resolution patch; if $\dot{\mathbf{C}}$ is singular, $\mathbf{T}_j$ is set to 0.)

  (**1.3**) Synthesize the high-resolution reside patch $\hat{\mathbf{r}}_{H,j}$ using the neighbor transformation $\mathbf{T}_j$, i.e. $\hat{\mathbf{r}}_{H,j} = \mathbf{T}_j\mathbf{r}_{L,j}$.

- **Step 2.** Concatenate and integrate the hallucinated high-resolution residue patches $\{\hat{\mathbf{r}}_{H,j}\}_{j=1}^{m}$ to form one residue facial image $\hat{R}_H$.

4. OUR FACE HALLUCINATION FRAMEWORK

Benefiting from the proposed first algorithm, we can recover the global face structure and main local features of the target high-resolution face in the super-resolution image. We denote the hallucinated result as I_H^g. Employing the second algorithm, we can infer the high-resolution residue, called by I_H^r, from low-resolution residue, and further enhance the quality of hallucination. We cascade the two super-resolution models as a integrated framework, which is described as below

1. For an input low-resolution image I_L, the super-resolution image I_H^g is hallucinated by the neighbor combination algorithm performed on the first half training data.

2. Construct the low-resolution residual image I_L^r by subtracting the input image I_L with down-sampled version of hallucinated image I_H^g.

3. Infer the super-resolution residue I_H^r by the neighbor transformation algorithm running on the remain half of training data.

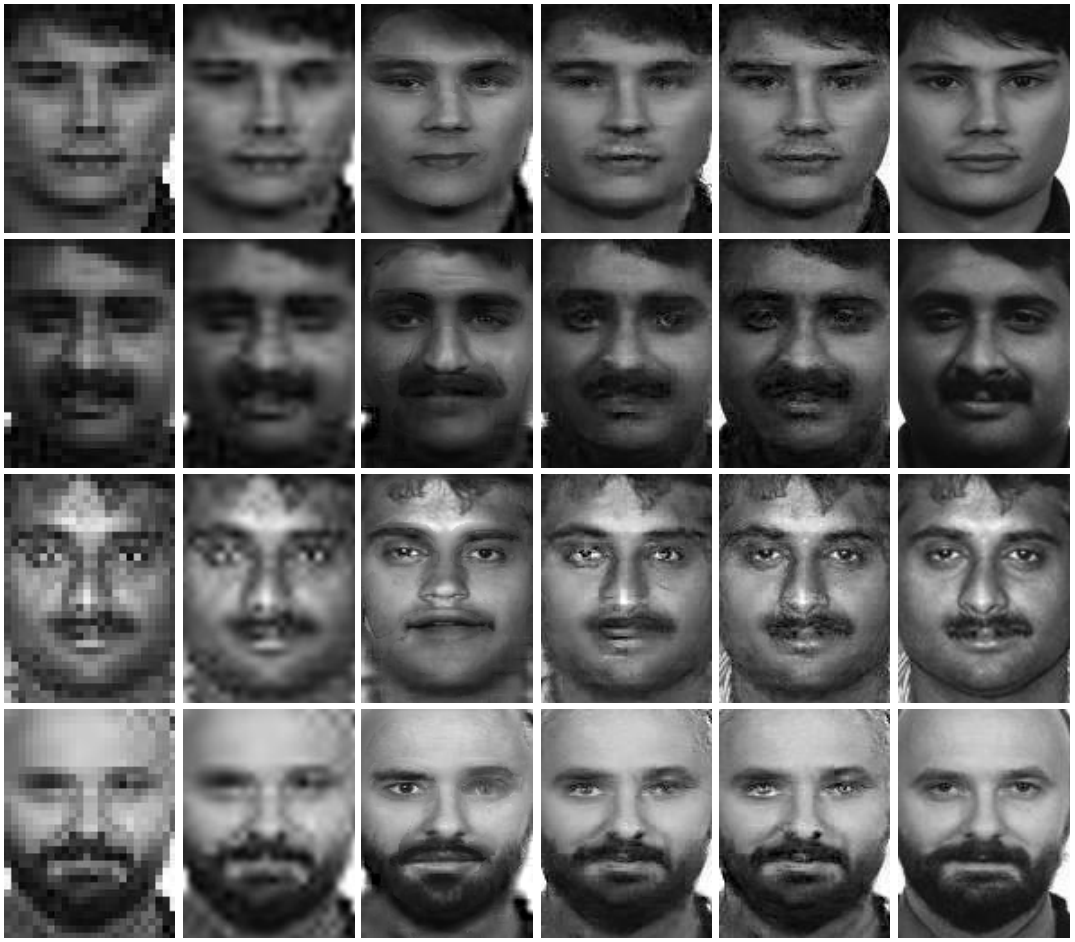
4. Add the inferred residue image I_H^r to the super-resolution image I_H^g to obtain the final result $I_H^* = I_H^g + I_H^r$.

5. EXPERIMENTAL RESULTS

Our experiments were conducted with a mixed database with samples collected from two databases **XM2VTS** [6] and **FERET** [7]. Our training data set consists of about 1400 images. Among all these samples, we select a half part samples for training the global model and the remain part for training residue compensation. Other samples and some outside samples are for testing. As a necessary preamble steps, we performs geometric normalization by an affine transform based on coordinates of eyes and mouth. After the transform, each image is cropped to a 96×128 grayscale image as the high-resolution one. The corresponding low-resolution images are obtained by smoothing and down-sampling, which are 24×32 images.

In our experiments, for each low-resolution image, $m = 682$ overlapped patches are extracted by sliding a small (3×3) window pixel by pixel. For high-resolution images, 682 (12×12) patches are extracted as well. The patches in low-resolution image and high-resolution image are in one-to-one correspondence. The regularization parameter r is set to 10^{-5}, and the numbers k_1 and k_2 of neighbors are set to 5 and 3 respectively. Our experiments show that such configuration on parameters yields the most satisfactory results.

The resultant images are shown in Figure 1. We can see that neighbor combination super-resolution already produces good hallucinated results, and neighbor transformation super-resolution further enhances hallucination quality, which makes the final resultant image approximate the groundtruth fairly well.



(a) Input 24*32 (b) Cubic B-Spline (c) Baker et al. (d) Neighbor Combination (e) Neighbor Combination +Transformation (f) Original96*128

Fig. 1. Comparison between our method and others.

We compare our algorithm with other existing methods, including Cubic B-Spline, Baker's algorithms. From Figure 1, We can clearly see the limitation of other methods. It is conspicuous that our method have significant superiority over others in terms of preserving both global structure and subtle details.

Acknowledgement

The work described in this paper was fully supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region. The work was done while all the authors are with the Chinese University of Hong Kong.

6. REFERENCES

[1] S. Baker and T. Kanade, "Limits on Super-Resolution and How to Break them," *IEEE Trans. on PAMI*, Vol. 24, No. 9, pp. 1167-1183, 2002.

[2] S. Baker and T. Kanade, "Hallucinating Faces," in *Proc. of Inter. Conf. on Automatic Face and Gesture Recognition*, pp. 83-88, 2000.

[3] H. Chang, D.Y. Yeung, and Y. Xiong, "Super-Resolution Through Neighbor Embedding," in *Proc. of CVPR*, Vol. 1, pp. 275-282, 2004.

[4] W.T. Freeman and E.C. Pasztor, "Learning Low-Level Vision," in *Proc. of ICCV*, Vol. 2, pp. 1182-1189, 1999.

[5] C. Liu, H. Shum, and C. Zhang, "A Two-Step Approach to Hallucinating Faces: Global Parametric Model and Local Nonparametric Model," in *Proc. of CVPR*, Vol. 1, pp. 192-198, 2001.

[6] K.Messer, J.Matas, J.Kittler, J.Luettin, and G.Matitre, "XM2VTSDB: The extended M2VTS database," in *Proc. of the Second International Conference on AVBPA*, 1999.

[7] P. Philips, H. Moon, P. Pauss, and S.Rivzvi, "The FERET Evaluation Methodology for Face-Recognition Algorithms," in *Proc. of CVPR*, pp. 137-143, 1997.