Generalized Face Super-Resolution

Kui Jia and Shaogang Gong

***Abstract*—Existing learning-based face super-resolution (hallucination) techniques generate high-resolution images of a single facial modality (i.e., at a fixed expression, pose and illumination) given one or set of low-resolution face images as probe. Here, we present a generalized approach based on a hierarchical tensor (multilinear) space representation for hallucinating high-resolution face images across multiple modalities, achieving generalization to variations in expression and pose. In particular, we formulate a unified tensor which can be reduced to two parts: a global image-based tensor for modeling the mappings among different facial modalities, and a local patch-based multiresolution tensor for incorporating high-resolution image details. For realistic hallucination of unregistered low-resolution faces contained in raw images, we develop an automatic face alignment algorithm capable of pixel-wise alignment by iteratively warping the probing face to its projection in the space of training face images. Our experiments show not only performance superiority over existing benchmark face super-resolution techniques on single modal face hallucination, but also novelty of our approach in coping with multimodal hallucination and its robustness in automatic alignment under practical imaging conditions.**

***Index Terms*—Face hallucination, super-resolution, tensor.**

I. Problem Statement

DUE to the intrinsic nonrigidness and extrinsic uncontrollable imaging conditions, face images of noncooperative human subjects captured by live surveillance cameras from a distance often consist of nonlinear variations caused by changes in expression, viewpoint (pose), or illumination. The difficulties in analyzing face images are further compounded when the resolution of face images becomes low, which is typical in CCTV videos. The missing high-resolution details of facial features and in appearance can deteriorate effective face image analysis and recognition. This raises two important questions need be addressed. 1) Given low-resolution face images, how do we recover or synthesize the lost high-frequency details at a predefined pose and expression but with unknown identity? 2) How do we build a unified model capable of coping with variations from different and multiple modalities where each modality defines one source of variation such as change in expression, pose, or illumination?

There have been some considerations in addressing the second problem [9]–[11], [15], [16], [18]–[20], [24]. However, these techniques can only be applied to high-resolution face images. To address the first problem, super-resolution techniques [25], [26] need to be exploited in order to generate higher resolution images given a single or a set of low-resolution input images.

The computation of super-resolution requires the recovery or synthesis of high-frequency information that was lost during the image formation process, which can be performed using two different approaches: reconstruction-based [1]–[7], and learning-based [22], [23], [27], [29]–[31], [36], [43]–[45]. Reconstruction-based approach inherits limitations when magnification factor increases [27]. In our approach, we focus on learning-based super-resolution, which when applied to human face images, is also known as "face hallucination" [28].

In this paper, we are motivated by the desire to develop a generalized model capable of hallucinating face images across multiple modalities such as expression or pose variations, given any low-resolution face image input of a single modality. To this end, we formulate a unified tensor space representation incorporating both global and local tensors.

Specifically, we model both the high- and low-resolution training face images of multiple modalities using a unified tensor space representation. We reduce this unified tensor to two components: a global image-based tensor that models the mappings among different facial modalities, and a local patch-based multiresolution tensor that incorporates high-resolution face image details into the cross-modality mapping process. Given any low-resolution face image input of a single modality, we first synthesize multiple low-resolution face images of different modalities using the trained global tensor. Based on these synthesized low-resolution face images, we then use the trained local tensor to construct the corresponding high-resolution image for each facial modality. Due to that the high-resolution face image constructed by the local tensor lacks the highest frequency visual information, we further add a high-frequency component residue using nonparametric patch learning from high-resolution training face images. We integrate this sequential statistical modeling process into a Bayesian framework, so that given any low-resolution face image of a single modality, we are able to obtain hallucinated high-resolution face images of multiple modalities.

Another important underlying issue for face hallucination is the requirement for accurate pixel-wise face alignment (registration): even a small amount of misalignment can dramatically degrade the hallucination result. Most existing techniques require that both training and test face images have been manually registered to a predefined template [28], [31]. Consequently, they are of limited use in many practical scenarios, where the faces contained in raw images are normally of nonfrontal views at low resolution. Recently, Liu *et al.* [37] proposed automatically registering low-resolution face images using an algorithm derived from the model of Lucas–Kanade [33]. However, this algorithm only considers a *mean* training face (although variance

Manuscript received December 20, 2006; revised February 4, 2008. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hassan Foroosh.

K. Jia is with the Shenzhen Institute of Advanced Integration Technology, Chinese Academy of Sciences/Chinese University of Hong Kong, Shenzhen, China (e-mail: chrisjia@dcs.qmul.ac.uk; kui.jia@sub.siat.ac.cn).

S. Gong is with the Department of Computer Science, Queen Mary, University of London, London E1 4NS, U.K. (e-mail: sgg@dcs.qmul.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2008.922421

is taken into account) as registration template; thus, its registration process is rather ad hoc and sensitive to initialization. To overcome the problem, we develop a different automatic face alignment algorithm in which automatic pixel-wise face registration is realized by iteratively optimizing geometric transformation parameters, which warp any probing face contained in raw images to its projection in a principal component analysis (PCA) subspace constructed from the low-resolution training face images. We demonstrate the effectiveness of our algorithm on accurate registration for generalized face hallucination with many examples.

II. Related Work

Traditional learning-based face super-resolution (hallucination) techniques generate high-resolution upright frontal view face images without nonlinear variations caused by expression, pose or illumination changes, by modeling the mapping prior between low- and high-resolution face image spaces [22], [27], [28], [31], [36]. These techniques build training model prior either globally using holistic face images, or locally using patches or pixels. They can be grouped into two categories.

a) **Global face space parameter estimation**: Capel and Zisserman [31] computed eigenfaces from a training face database as a model prior to constrain and super-resolve low-resolution face images. Combined with a MAP estimator, they recovered super-resolution images from a high-resolution eigenface space. Wang and Tang [21] considered the face hallucination problem as a transformation between different face styles. They used PCA to fit an input low-resolution face image to a linear combination of low-resolution face images in the training set. A high-resolution image was then rendered by replacing the low-resolution training images with their high-resolution correspondences, while retaining the same combination coefficients.

b) **Local image primitive intensity restoration**: Rather than globally using the whole face, Baker and Kanade [27] locally established a prior based on a set of training face images pixel by pixel using Gaussian, Laplacian and feature pyramids. Freeman *et al.* [29], [30] used a homogeneous Markov random field (MRF) model which builds the relation between low-resolution "observation" patches and underlying high-resolution "scene" patches, and the relation between neighboring "scene" patches. They used this model for generic image super-resolution. Liu and Shum [36] combined the PCA model-based approach with Freeman *et al.*'s image primitive technique. They developed a mixture model combining a global parametric model called the "global face image" carrying common facial properties, and a local nonparametric model called the "local feature image" recording local individualities. The high-resolution image is a composition of both.

Learning-based techniques have also been extended to video by taking into consideration of temporal constraints. In [42], a direct application of Freeman *et al.*'s model to video sequences was attempted, but severe video artifacts were encountered. As a remedy, an ad-hoc solution was proposed, consisting of re-using high-resolution solutions for achieving more coherent videos. Alternatively, Dedeoglu *et al.* [43] extended the work of [27] to super-resolve a single face video sequence, using different videos of the same face as training data. By exploiting a Bayesian framework and spatio-temporal constraints, they reported an extremely high 16×16 face magnification factor.

However, none of these techniques addressed the problem of generalization of face image super-resolution to multimodalities of nonlinear variations in expression, pose or illumination. To that end, we recently proposed a multimodal tensor model for face super-resolution [16], [17]. Part of this paper is an extension of our earlier work in [17]. Despite its ability for multiple expression hallucination, the earlier model did not in general cope well with nonlinear deformations. We extend the model here by building a unified tensor space representation incorporating both mapping relations amongst different modalities and also between high- and low-resolutions. We choose a global image-based tensor to perform synthesis across different facial modalities, and a local patch-based multiresolution tensor for hallucination. Another new contribution of this paper is an automatic face alignment algorithm. For applications in practical scenarios where faces captured in raw images are normally of nonfrontal views at low resolution, we develop an automatic face alignment algorithm in order to register the low-resolution faces so as to obtain good generalized face hallucination results.

III. Tensor Space Face Representation

A. Multilinear Analysis: Tensor SVD

Multilinear analysis [10]–[14] is a general extension of traditional linear methods such as PCA or matrix SVD. Instead of modeling relations within vectors or matrices, multilinear analysis provides a means to investigate the mappings between multiple factor spaces.[1] Given an N^{th}-order tensor $\mathcal{A} \in R^{I_1 \times I_2 \cdots \times I_N}$, an element of $\mathcal{A}$ is denoted as $\mathcal{A}_{i_1 \cdots i_n \cdots i_N}$ or $a_{i_1 \cdots i_n \cdots i_N}$, where $1 \leq i_n \leq I_n$. If we refer to I_n rank in tensor terminology, we generalize the matrix definition and refer to column vectors of matrices as mode-1 vectors and row vectors of matrices as mode-2 vectors. The mode-n vectors of the N^{th} order tensor are the I_n-dimensional vectors obtained from $\mathcal{A}$ by varying index i_n while keeping the other indices fixed. We can unfold or flatten tensor $\mathcal{A}$ by taking the mode-n vectors as the column vectors of matrix $\mathbf{A}^{(n)} \in R^{I_n \times (I_1 I_2 \cdots I_{n-1} I_{n+1} \cdots I_N)}$. This provides easy manipulation in tensor algebra. We can also reconstruct a tensor by the inverse process of mode-n unfolding.

We can generalize the product of two matrices to the product of a tensor and a matrix. The mode-n product of a tensor $\mathcal{A} \in R^{I_1 \times I_2 \times \cdots \times I_n \times \cdots \times I_N}$ by a matrix $\mathbf{M} \in R^{J_n \times I_n}$, denoted by $\mathcal{A} \times_n \mathbf{M}$, is a tensor $\mathcal{B} \in R^{I_1 \times \cdots \times I_{n-1} \times J_n \times I_{n+1} \times \cdots \times I_N}$ whose entries are computed by

$$(\mathcal{A} \times_n \mathbf{M})_{i_1 \cdots i_{n-1} j_n i_{n+1} \cdots i_N} = \sum_{i_n} a_{i_1 \cdots i_{n-1} i_n i_{n+1} \cdots i_N} m_{j_n i_n}.$$

[1]We denote scalars by lower-case letters $(a, b, \cdots; \alpha, \beta, \cdots)$, vectors by upper-case $(A, B, \cdots)$, matrices by bold upper-case $(\mathbf{A}, \mathbf{B}, \cdots)$, and tensors by calligraphic letters $(\mathcal{A}, \mathcal{B}, \cdots)$.

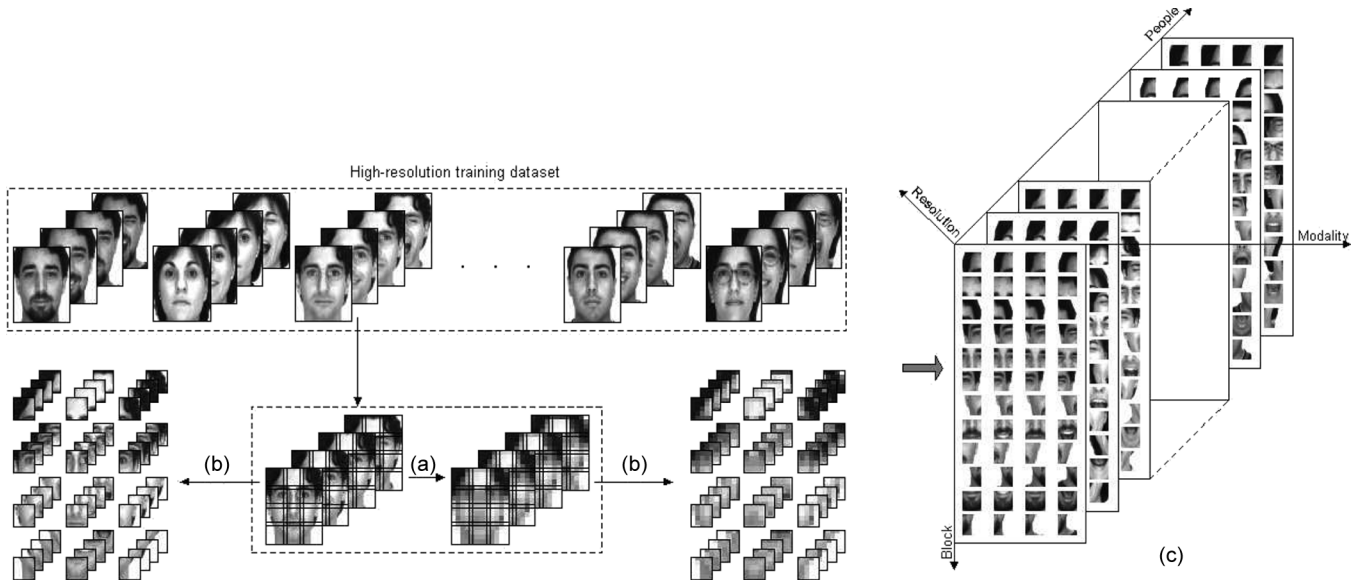

Fig. 1. Tensor construction illustration using block-wise images of multiple facial expressions at low and high resolution. (a) Face images of multiple modalities in the high-resolution training dataset are blurred and subsampled to get their corresponding low-resolution versions (here, multiple facial expression images from one training individual are shown as an example of such a process). (b) Low- and high-resolution training face images are uniformly decomposed into overlapped image blocks. (c) Fifth-order tensor $\mathcal{D}$ for the obtained block-wise image ensemble; only high-resolution images are shown.

This mode-n product of tensor and matrix can be expressed in terms of unfolding matrices for ease of usage

$$\mathbf{B}^{(n)} = \mathbf{M}\mathbf{A}^{(n)}. \tag{1}$$

In singular value decomposition of matrices, a matrix $\mathbf{D}$ is decomposed as $\mathbf{U}_1\mathbf{\Sigma}\mathbf{U}_2^T$, the product of an orthogonal column space represented by the left matrix $\mathbf{U}_1 \in R^{I_1\times J_1}$, a diagonal singular value matrix $\mathbf{\Sigma} \in R^{J_1\times J_2}$, and an orthogonal row space represented by the right matrix $\mathbf{U}_2 \in R^{I_2\times J_2}$. This matrix product can also be written as a mode-n product $\mathbf{D} = \mathbf{\Sigma} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2$. We can thus generalize SVD of matrices to multilinear Higher-Order SVD (HOSVD). An N^{th}-order tensor $\mathcal{A} \in R^{I_1\times I_2\times\cdots\times I_N}$ can be written as the product

$$\mathcal{A} = \mathcal{Z} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times \cdots \times_N \mathbf{U}_N \tag{2}$$

where $\mathbf{U}_n$ is a unitary matrix, and $\mathcal{Z}$ is the core tensor having the property of all-orthogonality, that is, two subtensors $\mathcal{Z}_{i_n=\alpha}$ and $\mathcal{Z}_{i_n=\beta}$ are orthogonal for all possible values of n, α and β subject to $\alpha \neq \beta$. The HOSVD of a given tensor $\mathcal{A}$ can be computed as follows. The mode-n singular matrix $\mathbf{U}_n$ can directly be found as the left singular matrix of the mode-n matrix unfolding of $\mathcal{A}$, afterwards, based on the product of tensor and matrix as in (1), the core tensor $\mathcal{Z}$ can be computed by $\mathcal{Z} = \mathcal{A} \times_1 \mathbf{U}_1^T \times_2 \mathbf{U}_2^T \cdots \times_N \mathbf{U}_N^T$. Equation (2) gives the basic representation of a multilinear model. With mode-n unfolding and folding, and by rearranging (2), we have

$$\mathcal{S} = \mathcal{B} \times_n V_n^T \tag{3}$$

where $\mathcal{S}$ is a subtensor of $\mathcal{A}$ corresponding to a fixed row vector V_n^T of singular matrix $\mathbf{U}_n$, and

$$\mathcal{B} = \mathcal{Z} \times_1 \mathbf{U}_1 \cdots \times_{n-1} \mathbf{U}_{n-1} \times_{n+1} \mathbf{U}_{n+1} \cdots \times_N \mathbf{U}_N.$$

This expression is the basis for recovering original data from a tensor structure. If we index into basis tensor $\mathcal{B}$ for particular V_n^T, we can recover different modal sample vector data.

B. Modeling Multimodalities at Different Resolutions

Face images share some common properties in pixel distribution, even though they appear differently under variations in expression, viewpoint or illumination. A tensor structure provides a powerful mechanism to incorporate information and interaction of these image ensembles of multiple modalities at different resolutions. More precisely, given a training dataset of high-resolution face images, we blur and subsample them with different Gaussian filters and subsampling factors, while keeping the image size unchanged, so to generate a set of low-resolution training face images. To further improve the modeling accuracy, we uniformly decompose these face images into overlapped image blocks, and then obtain a hierarchical ensemble containing block-wise face images of multiple modalities at low- and high-resolution. An illustration of how to obtain such a hierarchical ensemble is shown in Fig. 1(a) and (b). With these training data in place, we can construct a fifth-order tensor $\mathcal{D}$ [see Fig. 1(c)]. We use HOSVD to decompose $\mathcal{D}$ into

$$\mathcal{D} = \mathcal{Z} \times_1 \mathbf{U}_{\text{idens}} \times_2 \mathbf{U}_{\text{modes}} \times_3 \mathbf{U}_{\text{resos}} \times_4 \mathbf{U}_{\text{blocks}} \times_5 \mathbf{U}_{\text{pixels}} \tag{4}$$

where tensor $\mathcal{D}$ groups these block-wise training images into a tensor structure, and the core tensor $\mathcal{Z}$ governs the interactions between the five mode factors. Whilst mode matrix $\mathbf{U}_{\text{idens}}$ spans the parameter space of identity, mode matrices $\mathbf{U}_{\text{modes}}$, $\mathbf{U}_{\text{resos}}$, $\mathbf{U}_{\text{blocks}}$, and $\mathbf{U}_{\text{pixels}}$ span the spaces of modality, resolution, block-position, and pixel-value, respectively.

By utilizing the mappings among multiple factor spaces inherently embedded in the tensor structure, given a face image of a single modality we can synthesize multiple images of different modalities. Reversely, given a low-resolution face image

as a probe, we can also reconstruct/hallucinate its high-resolution correspondence.

C. Global Multimodal Tensor for Face Transformation

Equation (4) builds a tensor structure on decomposed image blocks, and models the relations amongst face images of multiple modalities at different resolutions. For face image synthesis across different expressions or viewpoints, we take the whole face image as blocks, and reduce (4) at a specific image resolution to a global image-based multimodal tensor

$$\mathcal{G} = \mathcal{Z}_{\mathcal{G}} \times_1 \mathbf{U}_{\text{idens}} \times_2 \mathbf{U}_{\text{modes}} \times_3 \mathbf{U}_{\text{pixels}}. \quad (5)$$

With the global tensor $\mathcal{G}$ containing face images of multiple modalities, we can perform face transformation in a tensor parameter vector space. Based on (3), images of different modalities can be synthesized given their identity parameter vector in tensor space. This identity parameter vector can be computed by projecting test modal face images onto the multimodal tensor $\mathcal{G}$. More precisely, suppose the basis tensor of $\mathcal{G}$ is $\mathcal{B}_{\mathcal{G}} = \mathcal{Z}_{\mathcal{G}} \times_2 \mathbf{U}_{\text{modes}} \times_3 \mathbf{U}_{\text{pixels}}$, we can index into $\mathcal{B}_{\mathcal{G}}$ at a particular modality m to yield a basis subtensor $\mathcal{B}_{\mathcal{G}_m} = \mathcal{Z}_{\mathcal{G}} \times_3 \mathbf{U}_{\text{pixels}} \times_2 V_m^T$. Then the subtensor containing individual image data can be expressed as

$$\mathcal{G}_m = \mathcal{B}_{\mathcal{G}_m} \times_1 V^T + \mathcal{E}_m \quad (6)$$

where V^T represents the identity parameter row vector and $\mathcal{E}_m$ stands for the tensor modeling error for modality m. To simplify notation, we use mode-1 unfolding matrices to represent tensors. The matrix representation of (6) becomes

$$\mathbf{G}_m^{(1)} = V^T \mathbf{B}_{\mathbf{G}_m}^{(1)} + e_m. \quad (7)$$

Equation (7) provides a possible solution for the identity parameter vector V^T. Applying it to a different facial modality m', the corresponding image data can be computed as

$$\mathbf{G}_{m'}^{(1)} = V^T \mathbf{B}_{\mathbf{G}_{m'}}^{(1)} + e_{m'}. \quad (8)$$

As noted in (7) and (8), the modeling errors e_m and $e_{m'}$ may degrade the recovered image quality of different facial modalities. To overcome this problem, we introduce a local patch-based multiresolution tensor to hallucinate high-resolution face images for each modality.

D. Local Multiresolution Tensor for Face Hallucination

To model high-resolution details for the purpose of face hallucination, we uniformly decompose the low- and high-resolution face images into small overlapped patches, and perform tensor modeling at patch level. We take these patches as blocks and reduce (4) at a specific modality to a local patch-based multiresolution tensor

$$\mathcal{L} = \mathcal{Z}_{\mathcal{L}} \times_1 \mathbf{U}_{\text{idens}} \times_2 \mathbf{U}_{\text{resos}} \times_3 \mathbf{U}_{\text{patches}} \times_4 \mathbf{U}_{\text{pixels}}. \quad (9)$$

Similar to Section III-C, from the unified identity parameter vector for low- and high-resolution, we can reconstruct high-resolution image data with all the decomposed patches. The final high-resolution face images for each modality are compositions of their corresponding overlapped small patches.

IV. Hallucinating Face Images Across Modalities

Tensor space modeling of either the whole face image or big patches can incorporate more information about different face modal variations, which is advantageous in synthesizing holistic facial image structures across modalities. However, this approach is also poor for recovering high-resolution details. To compensate for this disadvantage, based on synthesized low-resolution face images of multiple modalities, we perform super-resolution for each facial modality using local patch-based multiresolution tensor. Furthermore, for the highest frequency visual information which cannot be recovered by the local tensor, we add nonparametric local patch updating by learning from high-resolution training data.

Suppose that $H_1, H_2, \ldots, H_M$ are the high-resolution images to be hallucinated for different facial modalities, $S_1, S_2, \ldots, S_M$ are their low-resolution correspondences to be synthesized, and L_1 is any low-resolution face input of single modality. Our problem of multimodal face hallucination can be formulated into a Bayesian framework. The task comes as finding the maximum *a posteriori* (MAP) estimation of $H_1, H_2, \ldots, H_M$ given L_1. We consider the case of two modal face hallucination as an example, which can be formulated as

$$\{H_{1\text{MAP}}, H_{2\text{MAP}}\} = \arg \max_{H_1, S_1, H_2, S_2} \log P(H_1, H_2, S_1, S_2 | L_1). \quad (10)$$

By applying Bayes rule, we have

$$P(H_1, H_2, S_1, S_2 | L_1) = P(H_1, H_2 | S_1, S_2, L_1) P(S_1, S_2 | L_1). \quad (11)$$

During the sequential processes of our face hallucination across modalities, the high-resolution face image is independently reconstructed, based on the synthesized low-resolution image for each modality. The above expression then yields

$$\begin{aligned} &P(H_1, H_2, S_1, S_2 | L_1) \\ &\quad = P(H_1, H_2 | S_1, S_2) P(S_1, S_2 | L_1) \\ &\quad = P(H_1 | S_1) P(H_2 | S_2) P(S_1, S_2 | L_1) \\ &\quad = P(S_1 | H_1) P(S_2 | H_2) P(H_1) P(H_2) P(S_1, S_2 | L_1). \end{aligned} \quad (12)$$

Assuming H^{lm} represents face images containing low- and middle-frequency information, and H^h contains high-frequency part, the high-resolution image is naturally composed from the two

$$H = H^{lm} + H^h. \quad (13)$$

Since H^{lm} contributes the main part of S after blurring and subsampling, then the probability $P(S|H)$ can be approximated as $P(S|H^{lm})$. Based on (13), we also have $P(H) = P(H^h|H^{lm})P(H^{lm})$, and the estimation of H given H^{lm} is equivalent to the estimation of H^h given H^{lm}, we then reformulate probability $P(S_1|H_1)P(S_2|H_2)P(H_1)P(H_2)$ as

$$\begin{aligned} &P(S_1|H_1)P(S_2|H_2)P(H_1)P(H_2) \\ &\quad = P\left(S_1|H_1^{lm}\right)P\left(S_2|H_2^{lm}\right)P\left(H_1^h|H_1^{lm}\right) \\ &P\left(H_1^{lm}\right)P\left(H_2^h|H_2^{lm}\right)P\left(H_2^{lm}\right) \\ &\quad = P\left(H_1^{lm}|S_1\right)P\left(H_2^{lm}|S_2\right)P\left(H_1|H_1^{lm}\right)P\left(H_2|H_2^{lm}\right). \end{aligned} \tag{14}$$

The synthesis of S_1 and S_2 are independent, so we can rewrite probability $P(S_1, S_2|L_1)$ as

$$\begin{aligned} P(S_1, S_2|L_1) &= P(S_1|L_1)P(S_2|L_1) \\ &= P(S_1|L_1)P(L_1|S_2)P(S_2) \\ &= P(L_1|S_1)P(S_1)P(S_2). \end{aligned} \tag{15}$$

Based on (14) and (15), the MAP inference problem of (10) can be finally formulated as (16), shown at the bottom of the page. Probabilities $P(L_1|S_1)P(S_1)P(S_2)$, $P(H_1^{lm}|S_1)P(H_2^{lm}|S_2)$, and $P(H_1|H_1^{lm})P(H_2|H_2^{lm})$ sequentially constrain S_1, S_2, H_1^{lm}, H_2^{lm}, and H_1, H_2 in (16). This leads to a three-step sequential solution. In the first step, by using a global image-based multimodal tensor, we can synthesize the low-resolution S_1, S_2 for different facial modalities. After obtaining S_1, S_2, the H_1^{lm}, H_2^{lm} containing low- and middle-frequency image information can be computed using the local patch-based multiresolution tensor. The final high-resolution H_1, H_2 are computed by maximizing $P(H_1|H_1^{lm})P(H_2|H_2^{lm})$ in the third step.

A. Global Multimodal Low-Resolution Face Image Synthesis

The synthesis of S_1 and S_2 are computed by maximizing probability $P(L_1|S_1)P(S_1)P(S_2)$. Since L_1 and S_1 are the low-resolution given and synthesized face images with the same modality, we regard their relationship as Gaussian

$$P(L_1|S_1) = \frac{1}{f}\exp\left\{-||S_1 - L_1||^2/\lambda\right\} \tag{17}$$

where f is a normalization constant and λ scales the variance. Prior constraints $P(S_1)P(S_2)$ are assumed to be Gaussian as well, so $P(S_1) = (1/F')\exp(-(S_1 - \mu_{S_1})^T\mathbf{\Lambda}^{-1}(S_1 - \mu_{S_1}))$, where $\mathbf{\Lambda}$ is the covariance matrix of all training face images with one single modality. However, due to the orthogonality of tensor decomposition, the above prior $P(S_1)$ simply leads the optimum to the mean value μ_{S_1}. The same condition applies to $P(S_2)$. In this sense, (15) degenerates to the maximum likelihood (ML) estimation.

Equation (5) shows that the global multimodal tensor incorporates the image data of multiple facial modalities. If we index into its basis subtensor at a particular modality m^1, then the subtensor containing the individual image data as in (6) can be approximated by $\mathcal{G}_{m^1} = \mathcal{B}_{\mathcal{G}_{m^1}} \times_1 V^T$. We unfold it into matrix representation and it becomes $\mathbf{G}_{m^1}^{(1)T} = \mathbf{B}_{\mathbf{G}_{m^1}^{(1)T}}V$. Similarly, we can obtain a subtensor for modality m', which is $\mathbf{G}_{m'}^{(1)T} = \mathbf{B}_{\mathbf{G}_{m'}^{(1)T}}\widetilde{V}$. Suppose $\mathbf{G}_{m^1}^{(1)T}$ and $\mathbf{G}_{m'}^{(1)T}$ correspond to face images S_1 and S_2 respectively, then we substitute for S_1 in (17) resulting in

$$P(L_1|S_1) = \frac{1}{f}\exp\left\{-\left\|\mathbf{B}_{\mathbf{G}_{m^1}^{(1)T}}V - L_1\right\|^2/\lambda\right\}. \tag{18}$$

In reality the given low-resolution L_1 and synthesized S_1 have the same modality. By setting $\mathbf{B}_{\mathbf{G}_{m^1}^{(1)T}}V = L_1$, we maximize (18) and approximately compute

$$V = \left(\mathbf{B}_{\mathbf{G}_{m^1}^{(1)}}\mathbf{B}_{\mathbf{G}_{m^1}^{(1)T}}\right)^{-1}\mathbf{B}_{\mathbf{G}_{m^1}^{(1)T}}L_1 \tag{19}$$

where $(\mathbf{B}_{\mathbf{G}_{m^1}^{(1)}}\mathbf{B}_{\mathbf{G}_{m^1}^{(1)T}})^{-1}\mathbf{B}_{\mathbf{G}_{m^1}^{(1)T}}$ is the pseudoinverse of $\mathbf{B}_{\mathbf{G}_{m^1}^{(1)T}}$. Because of the uniqueness of the identity parameter vector V for each individual person in a tensor space, we choose $\widetilde{V} = V$, and the synthesis of face images across different modalities such as S_2 is then computed as

$$S_2 = \mathbf{G}_{m'}^{(1)T}\left(\mathbf{B}_{\mathbf{G}_{m^1}^{(1)}}\mathbf{B}_{\mathbf{G}_{m^1}^{(1)T}}\right)^{-1}\mathbf{B}_{\mathbf{G}_{m^1}^{(1)T}}L_1. \tag{20}$$

Equation (20) provides the computation method for face image synthesis across different modalities given any low-resolution input of single modality.

B. Local Patch-Based Face Image Hallucination

Face images of multiple modalities synthesized by the global image-based tensor are at a low resolution. To obtain their hallucinated high-resolution correspondences of each modality containing low- and middle-frequency visual information, we maximize $P(H_1^{lm}|S_1)P(H_2^{lm}|S_2)$ using the local patch-based multiresolution tensor. The inference of H_1^{lm}, H_2^{lm} from S_1, S_2 is independent. In the following, we take H_1^{lm} as an example to illustrate this second process.

Since the training local multiresolution tensor is constructed from small overlapped patches, we decompose the synthesized S_1 uniformly in the same way as decomposing training data, and factorize the likelihood $P(H_1^{lm}|S_1)$ at patch level as $P(H_1^{lm}|S_1) = \prod_{p=1}^{N}P(H_{1p}^{lm}|S_{1p})$. Assuming $\mathbf{A}$ is the blurring and subsampling operator connecting H_{1p}^{lm} and S_{1p} in

$$\begin{aligned} &\{H_{1\mathrm{MAP}}, H_{2\mathrm{MAP}}\} \\ &\quad = \arg\max_{H_1, H_1^{lm}, S_1, H_2, H_2^{lm}, S_2} \log\left(P(L_1|S_1)P(S_1)P(S_2)P\left(H_1^{lm}|S_1\right)P\left(H_2^{lm}|S_2\right)P\left(H_1|H_1^{lm}\right)P\left(H_2|H_2^{lm}\right)\right). \end{aligned} \tag{16}$$

an imaging observation model, we regard these processes as Gaussian, therefore

$$P\left(H_1^{lm}|S_1\right)=\prod_{p=1}^{N}\frac{1}{w}\exp\left\{-\left\|\mathbf{A}H_{1p}^{lm}-S_{1p}\right\|^2/\beta\right\} \quad (21)$$

where w is a normalization constant and β scales the variance.

Suppose the local multiresolution tensor in (9) has a basis tensor $\mathcal{B}_{\mathcal{L}}=\mathcal{Z}_{\mathcal{L}}\times_2\mathbf{U}_{\text{resos}}\times_3\mathbf{U}_{\text{patches}}\times_4\mathbf{U}_{\text{pixels}}$. We index into this basis tensor at a particular resolution r and patch position p, yielding a basis subtensor $\mathcal{B}_{\mathcal{L}r,p}=\mathcal{Z}_{\mathcal{L}}\times_4\mathbf{U}_{\text{pixels}}\times_2V_r^T\times_3V_p^T$. Then as described in Section IV-A, the subtensor containing the pixel data for that particular patch can be approximated as $\mathcal{L}_{r,p}=\mathcal{B}_{\mathcal{L}r,p}\times_1V^T$, and its unfolded representation is $\mathbf{L}_{r,p}^{(1)T}=\mathbf{B}_{\mathbf{L}r,p}^{(1)T}V$. Similarly, we can obtain a subtensor for resolution r' of the same patch position, which is $\mathbf{L}_{r',p}^{(1)T}=\mathbf{B}_{\mathbf{L}r',p}^{(1)T}\widetilde{V}$. Suppose $\mathbf{L}_{r,p}^{(1)T}$ and $\mathbf{L}_{r',p}^{(1)T}$ correspond to S_{1p} and H_{1p}^{lm}, respectively; we substitute them in (21) as

$$P\left(H_1^{lm}|S_1\right)=\prod_{p=1}^{N}\frac{1}{w}\exp\left\{-\left\|\mathbf{A}\mathbf{B}_{\mathbf{L}r',p}^{(1)T}\widetilde{V}-\mathbf{B}_{\mathbf{L}r,p}^{(1)T}V\right\|^2/\beta\right\}. \quad (22)$$

We optimize the parameter $\widetilde{V}$ based on the construction properties of the local multiresolution patch tensor, which suggests that the relation between $\mathbf{B}_{\mathbf{L}r',p}^{(1)T}$ and $\mathbf{B}_{\mathbf{L}r,p}^{(1)T}$ observes a basic imaging observation model through the blurring and subsampling operator $\mathbf{A}$. This is consistent with the uniqueness of the identity parameter vector in a tensor space as well. By setting $\widetilde{V}=V$, we can approximately compute H_{1p}^{lm} as $H_{1p}^{lm}=\mathbf{B}_{\mathbf{L}r',p}^{(1)T}\mathbf{\Psi}S_{1p}$ where $\mathbf{\Psi}$ is the pseudoinverse of $\mathbf{B}_{\mathbf{L}r,p}^{(1)T}$ and is equal to $(\mathbf{B}_{\mathbf{L}r,p}^{(1)}\mathbf{B}_{\mathbf{L}r,p}^{(1)T})^{-1}\mathbf{B}_{\mathbf{L}r,p}^{(1)T}$. After reconstructing all the patches at different positions, the final hallucinated face image H_1^{lm} is simply a composition of the corresponding hallucinated small patches.

C. High-Frequency Residue Recovery

Face images of multiple modalities recovered by global and local tensors contain only low- and middle-frequency information. We recover the highest frequency part by patch learning from the high-resolution training data. The inference of H_1, H_2 from H_1^{lm}, H_2^{lm} is independent. In the following, we take H_1 as an example to illustrate how to hallucinate the final high-resolution face images.

We use a MRF to model the H_1 to be inferred. By decomposing H_1^{lm} into square patches

$$P\left(H_1|H_1^{lm}\right)=P\left(H_1^{lm}|H_1\right)P(H_1)$$
$$=\prod_{q=1}^{Q}P\left(H_{1q}^{lm}|H_{1q}\right)P(H_1). \quad (23)$$

The difference between H_1 and H_1^{lm} is the high-frequency band information. Since the high-frequency information depends on the lower-frequency band, we use the Laplacian image $L_{H_1^{lm}}$ of H_1^{lm} to represent the middle-frequency band. To infer H_1, we use the sum of squared differences of Laplacian images as metrics to model $\prod_{q=1}^{Q}P(H_{1q}^{lm}|H_{1q})$ as

$$\prod_{q=1}^{Q}P\left(H_{1q}^{lm}|H_{1q}\right)\propto\prod_{q=1}^{Q}\exp\left\{-\left\|L_{H_{1q}^{lm}}-L_{H_{1q}^{(t)}}\right\|^2\right\} \quad (24)$$

where $L_{H_{1q}^{(t)}}$ are the Laplacian images from high-resolution training face images. Comparing the Laplacian images $L_{H_{1q}^{lm}}$ with $\{L_{H_{1q}^{(t)}}\}_{t=1}^{\mathrm{T}}$ from the training dataset, the patch $H_{1q}^{(t)}$ with $L_{H_{1q}^{(t)}}$ closest to $L_{H_{1q}^{lm}}$ is the most probable to be chosen as H_{1q}. An illustration of this process is shown in Fig. 2(a). Since we model the high-resolution image as a MRF, based on the Hammersley–Clifford theorem, $P(H_1)$ is a product $\prod_{H_{1q},H_{1\overline{q}}}\Phi(H_{1q},H_{1\overline{q}})$ of compatibility functions $\Phi(H_{1q},H_{1\overline{q}})$ over all neighboring pairs, where $H_{1q},H_{1\overline{q}}$ are one of the neighboring patch pairs in a 4-neighbor system. The compatibility function $\Phi(H_{1q},H_{1\overline{q}})$ is defined using the similarity of pixel values on the overlapping area of the neighboring patches $\Phi(H_{1q},H_{1\overline{q}})\propto\exp\{-\|O_{H_{1q}}-O_{H_{1\overline{q}}}\|^2\}$, where $O_{H_{1q}}$ denotes the pixels of patch H_{1q} overlapping with neighboring patch $H_{1\overline{q}}$, and vice versa for $O_{H_{1\overline{q}}}$. We illustrate this 4-neighbor system and the corresponding patch overlapping relations in Fig. 2(b). H_1 is then estimated as

$$\arg\max_{H_1}\prod_{q=1}^{Q}P\left(H_{1q}^{lm}|H_{1q}\right)\prod_{(q,\overline{q})}\Phi(H_{1q},H_{1\overline{q}}). \quad (25)$$

Similar procedures can be independently repeated for estimations of face images of other modalities.

Solving probabilistic (25) to obtain H_1 is not a trivial task. We use the iterated conditional modes (ICM) algorithm [41]. More specifically, we maximize $P(H_{1q}^{lm}|H_{1q})$ for all patch positions $q\in\{1,\ldots,Q\}$ to yield the initial maximum likelihood estimate of H_1. Based on this initial estimate, we then pick a random patch position q and update the estimate of H_{1q} using the current estimates of its neighbors $H_{1\overline{q}}$ by maximizing $P(H_{1q}^{lm}|H_{1q})\prod_{(q,\overline{q})}\Phi(H_{1q},H_{1\overline{q}})$. We repeat this random patch selection and updating process until converging to the final high-resolution image H_1. The pseudo code for our generalized face super-resolution algorithm is as follows.

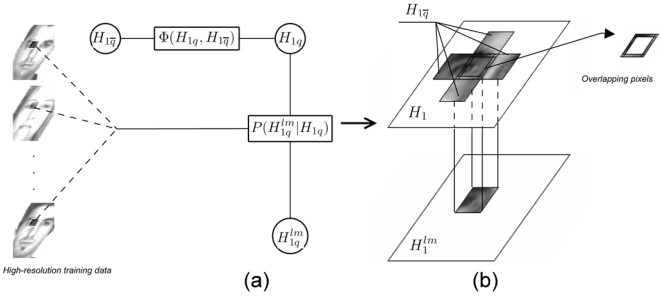

Fig. 2. Illustration of patch-based high-frequency residue recovery.

Algorithm 1: Generalized face super-resolution.

input: single face image L_{m^1} with modality m^1 at low-resolution r; $m^1 \in \{1, \ldots, M\}$

output: multiple face images H_m $(m = 1, \ldots, M)$ of different modalities at high-resolution r'

Step I:

for *each different modality* $m' \neq m^1; m' \in \{1, \ldots, M\}$

do

$$S_{m'} = \mathbf{G}_{m'}^{(1)T}(\mathbf{B}_{\mathbf{G}_{m^1}^{(1)}}\mathbf{B}_{\mathbf{G}_{m^1}^{(1)T}})^{-1}\mathbf{B}_{\mathbf{G}_{m^1}^{(1)T}}L_{m^1}$$

end

Step II:

for *each patch position* $p \in \{1, \ldots, N\}$ *on any modality* $m \in \{1, \ldots, M\}$

do

$H_m^{lm} \longleftarrow H_{m,p}^{lm} = (\mathbf{B}_{\mathbf{L}_{r',p}^{(1)T}})_m \mathbf{\Psi}_m S_{m,p}; (\mathbf{B}_{\mathbf{L}_{r',p}^{(1)T}})_m,$ $\mathbf{\Psi}_m$ for different modality m

end

Step III:

for *each modality* $m \in \{1, \ldots, M\}$ **do**

for *all patch positions* $q \in \{1, \ldots, Q\}$ **do**

$H_{m,q} \longleftarrow \arg\max_{H_{m,q}} P(H_{m,q}^{lm}|H_{m,q})$

end

repeat

pick a random patch location q;

$H_{m,q} \longleftarrow \arg\max_{H_{m,q}} P(H_{m,q}^{lm}|H_{m,q}) \prod_{(q,\overline{q})} \Phi(H_{m,q}, H_{m,\overline{q}})$

until $H_m converges$;

end

V. Simulation With Manual Alignment

A. Multiple Facial Expression Hallucination

We chose the benchmark AR face database for a set of simulated experiments. The original AR dataset consists of 126 people, and for each individual, it includes images of different facial expressions, illumination conditions and occlusions. We chose expression images of neutral, smile, anger, and scream for experiments on multiple facial expression hallucination. To establish a standard training dataset, we used a face image size of 64×48 and aligned the data manually by marking the location of three points: the centers of mouth and two eyes. These three points define an affine warp, which was used to warp the images into a canonical form.

For all the 504 (126×4) high-resolution facial expression images in the training dataset, we blurred and subsampled them to obtain their low-resolution (16×12) samples. We used the "leave-one-out" methodology to perform the multiple facial expression hallucination experiments. That is, from the 126 individuals, we used all four facial expression images of 125 of them as training data, and one expression image of the remaining person as the test input (the test expression is known in this case). In the step of image synthesis across different facial expressions, we used the four low-resolution expression images of these 125 training individuals to build our global image-based tensor. After obtaining the synthesized low-resolution image of each facial expression, we interpolated those low-resolution training images to the size of 64×48, and decomposed each of the 125×4 pairs of low- and high-resolution training face images into 768 small 3×3 patches which overlapped horizontally and vertically with each other by 1 pixel (the patch size and overlapping size were experimentally determined). For face hallucination of any expression, we chose the decomposed low- and high-resolution face image pairs of that expression from the training 125 individuals to build our local patch-based multiresolution tensor. The obtained high-resolution (64×48) images with facial expressions of neutral, smile, anger, and scream did not contain the highest frequency visual information. So we additionally used the nonparametric patch learning method (Section IV-C) to compensate for this. The nonparametric patch size was chosen as 6×6 with three pixels overlapping both horizontally and vertically. Finally we obtained the four facial expression hallucinations for each test expression input.

Some example results are shown in Fig. 3. These results suggest that any hallucination of low-resolution input with the same expression [given in Fig. 3(a)] is always better than those with other expressions, which is naturally an expense of generating nonlinear variations across different facial expressions. Also in Fig. 3, comparative investigations between Fig. 3(k) and (o) and Fig. 3(m) and (q) suggest that the hallucinated smiling and screaming images have no identical muscle changes compared with their ground truth. However, the muscle change intensities of these expressions have been successfully synthesized and hallucinated.

We also quantified our performance by evaluating the peak signal-to-noise ratio (PSNR) between the ground truth face images and the hallucinated images of multiple facial expressions, which is commonly used as a quality measure in the domain of image compression. We plot the PSNR values of the hallucinated results of four different facial expressions from each test expression of 126 individuals in Fig. 4 (better shown in electronic version), where the black lines are for neutral expression, the red lines are for smiling expression, the green lines for angry expression and the blue lines for screaming expression. Consistent with Fig. 3, the black line in Fig. 4(a), the red line in Fig. 4(b), the green line in Fig. 4(c) and the blue line in Fig. 4(d) correspond to the hallucinated results with the same facial expressions as their respective low-resolution inputs, and they all have the highest PSNR values compared with other expressions.

In the above experiments, we used facial expression images of 125 people for training. What if we use training data from fewer people? We performed further experiments using facial expression images of fewer training people which were randomly selected from the 126 people in the AR face database.

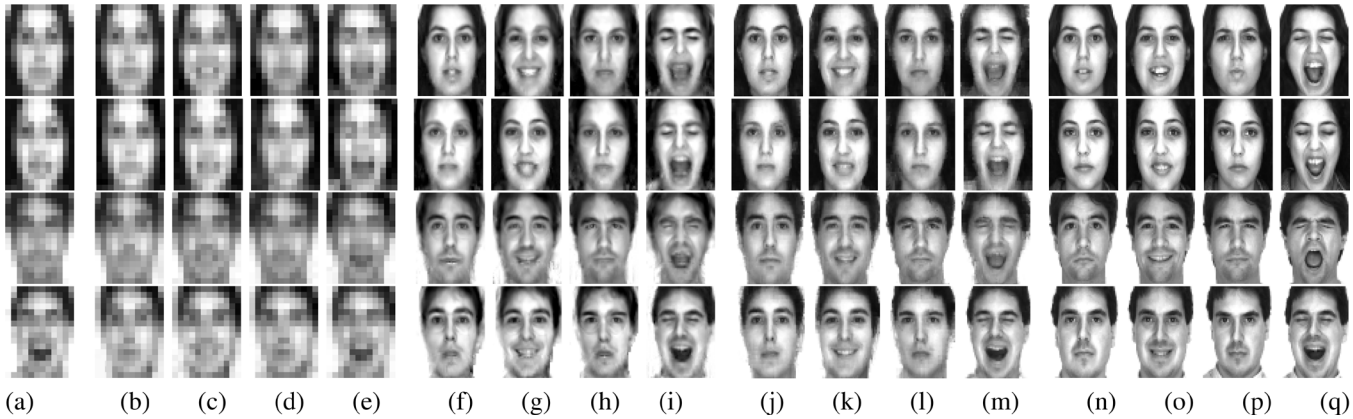

Fig. 3. Simulated experiments on multiple facial expression hallucination. (a) Low-resolution input images (16 × 12) of different facial expressions (obtained by downsampling original test input images). (b)–(e) Synthesized low-resolution (16 × 12) images with expressions of neutral, smile, anger, and scream, respectively, using the global image-based tensor. (f)–(i) Hallucinated high-resolution (64 × 48) face images with the four expressions, using the local patch-based multiresolution tensor. (j)–(m) Final hallucination results after adding the high-frequency component residue using nonparametric patch learning. (n)–(q) Ground truth face images (64 × 48) of corresponding expressions.

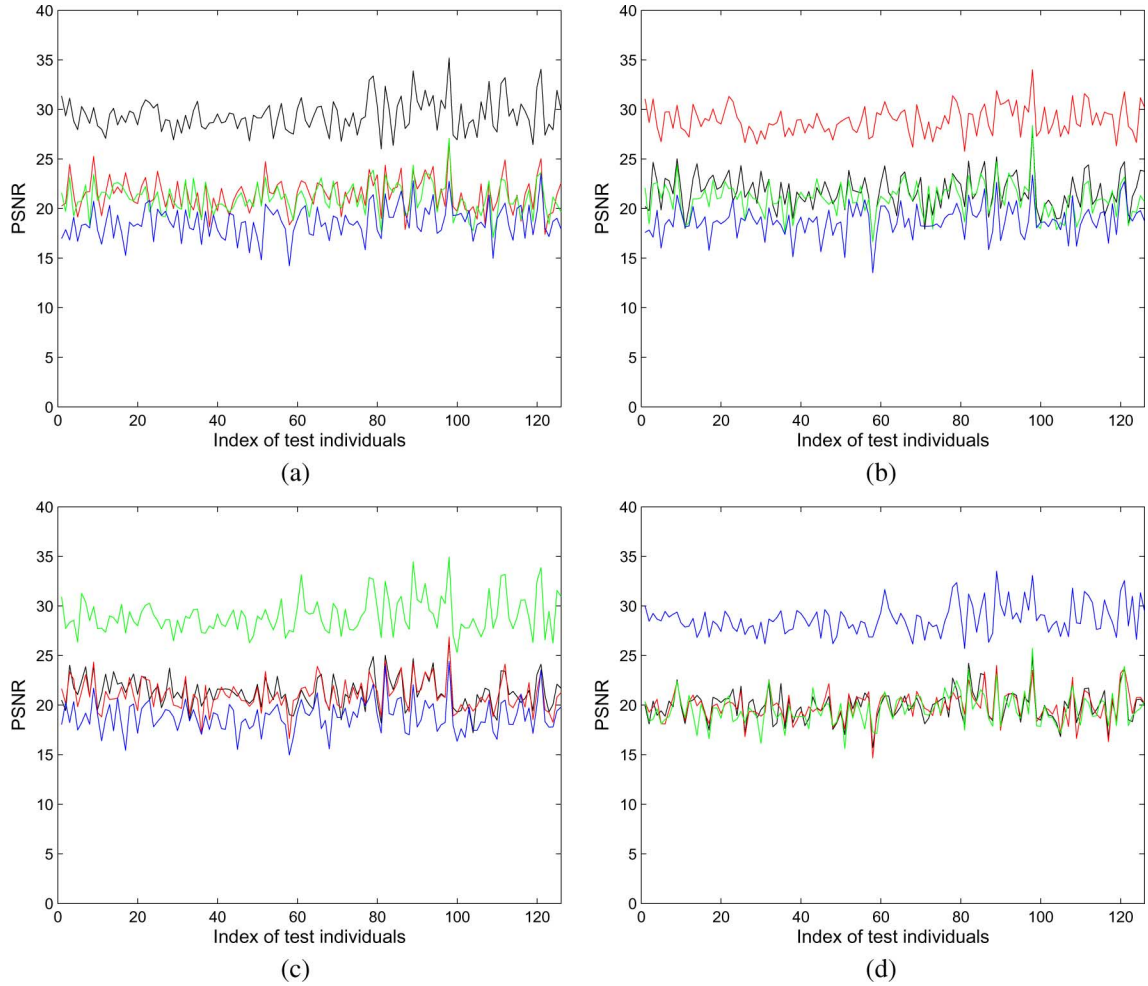

Fig. 4. PSNR values of the hallucinated results of four different facial expressions from each test expression of 126 individuals (better shown in electronic version). The black lines stand for the PSNR values of the hallucinated neutral expression face images, the red lines stand for the PSNR values of the hallucinated smiling expression face images, the green lines for angry expression, and the blue lines for screaming expression. (a) Low-resolution test inputs are with neutral expression. (b) Low-resolution test inputs are with smiling expression. (c) Low-resolution test inputs are with angry expression. (d) Low-resolution test inputs are with screaming expression.

Some example results are given in Fig. 5, which show that hallucination of all four facial expressions are acceptable when the numbers of training people are more than 65. When the numbers of training people become less than 65, some or all four facial expression hallucinations become worse, and can be extremely different from the ground truth face images. This is because when the size of training samples decreases, the learning process of face hallucination may converge to a local minimum,

Fig. 5. Examples of facial expression hallucination using different numbers of training people, given one low-resolution probe of neutral expression.

Fig. 6. Comparative experiments of our sequential tensor space approach with other benchmark face super-resolution techniques on single facial expression. (a) Low-resolution input images (16 × 12) with expressions of neutral, smile, anger, and scream, respectively. (b) Super-resolution results (64 × 48) using Capel and Zisserman's PCA-based Bayesian approach. (c) Super-resolution results (64 × 48) using Freeman *et al.*'s technique adapted to be based on an inhomogeneous Markov network. (d) Hallucination results (64 × 48) using Baker and Kanade's approach. (e) Hallucinated results (64 × 48) using first global then local tensors of our sequential approach. (f) Our final hallucination results (64 × 48) after adding the high-frequency component residue to (e). (g) Ground truth face images (64 × 48).

which can typically be a face of some training person in a small training set.

B. Comparisons With Other Face Super-Resolution Techniques

We compared our models with other benchmark techniques for face image super-resolution, including Capel and Zisserman [31], Freeman *et al.* [29], [30], and Baker and Kanade [27], [28]. These traditional techniques can only perform face super-resolution under a single modality; therefore, in this comparative experiment, given any low-resolution probe of a single facial expression, we only chose our approach's hallucination result with the same expression as output (from our multiple results of different facial expressions).

Some example results are presented in Fig. 6. Compared with the results in Fig. 6(b) using Capel and Zisserman's technique, the hallucinated results in Fig. 6(e) using our first global then local tensors recovered more facial details, and are closer to the ground truth images in Fig. 6(g). This suggests that super-resolution through local modeling has the advantage of recovering high-resolution facial details. After adding the high-frequency component residue our final hallucination results are shown in Fig. 6(f).

TABLE I
PSNR VALUES BETWEEN GROUND TRUTH FACE IMAGES AND HALLUCINATION RESULTS IN FIG. 6 USING DIFFERENT TECHNIQUES

PSNR	Capel	Freeman	Baker	Our tensors	Our final
Neutral	23.1717	26.9647	29.3221	25.5105	29.6813
Smile	22.5980	25.9577	28.4390	24.6619	28.6154
Anger	21.7997	25.4694	27.1970	23.1744	27.7260
Scream	20.7043	25.1834	27.0243	22.7340	27.5355

In the original work of Freeman *et al.* [29], [30], they assumed a homogeneous Markov network which is an appropriate model for generic image super-resolution. But this assumption is not very suitable for face super-resolution since human faces have strong structural patterns which is essentially inhomogeneous. The hallucination results in column (c) using Freeman *et al.*'s technique are based on an inhomogeneous MRF. This is then similar to the third step of the high-frequency residue compensation in our sequential approach. The results in column (c) shows that Freeman *et al.*'s technique is good at hallucinating and reproducing details of local face regions, but poor at retaining the global symmetric facial structures. For example, the face appearance lighting distribution of the result in the second row in column (c) is not natural, and the mouth of the face in the third row in column (c) is not symmetric. On the contrary, the results in column (f) using our sequential approach are cleaner and appear closer to natural human faces. This is essentially benefiting from our method's advantages of sequential global and local face modeling and super-resolution. Baker and Kanade's approach does not incorporate global constraint either, consequently their results in column (d) expose some unsmoothed noisy artifacts, which compare poorly against the results in column (f) using our sequential approach. We also give in Table I the PSNR values between the ground truth face images and those face hallucination results in Fig. 6 using different techniques. Table I shows that our sequential approach outperforms all the other face super-resolution techniques in terms of PSNR. This supports our claims and analysis above. In summary, our sequential tensor space face hallucination approach is superior to other benchmark learning-based face super-resolution techniques for both single modal face hallucination and hallucination across multiple facial modalities.

C. Comparisons With Our Previous Work

In the above experiments, the sequential steps of the global image-based multimodal tensor and the local patch-based multiresolution tensor are necessary for generalized face super-resolution, which is also one of the technical differences between the super-resolution model in this paper and that in our earlier work [17]. Our previous model constructs a training tensor $\mathcal{C}$ decomposable as $\mathcal{C} = \mathcal{Z}_{\mathcal{C}} \times_1 \mathbf{U}_{\text{idens}} \times_2 \mathbf{U}_{\text{modes}} \times_3 \mathbf{U}_{\text{resos}} \times_4 \mathbf{U}_{\text{patches}} \times_5 \mathbf{U}_{\text{pixels}}$. Based on $\mathcal{C}$, generalized face super-resolution is patch-wisely performed in one single step. That is, face hallucination is a composition of its corresponding hallucinated local patches. However, this single step solution is limited because the nonlinear appearance variation relations between different facial modalities cannot be accurately modeled

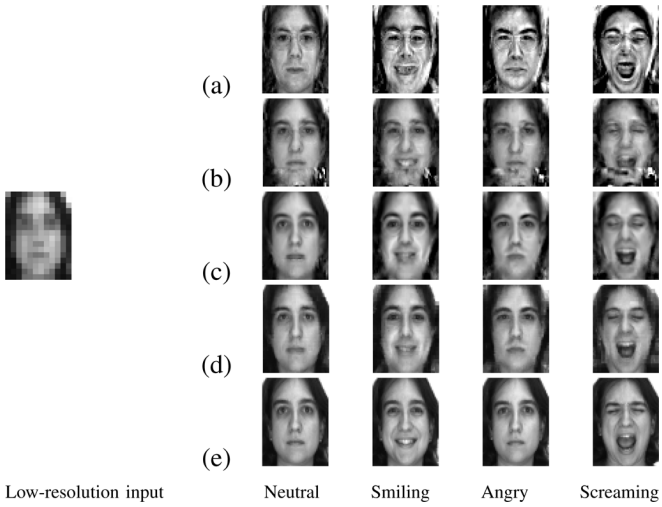

Fig. 7. Comparisons of our sequential approach with single step approaches on multi-expression hallucination. Given one low-resolution neutral face image (16 × 12). (a) Hallucination results (64 × 48) of neutral, smiling, angry, and screaming expression images, using the tensor model $\mathcal{C}'$. (b) Hallucination results (64 × 48) of the four expression face images, using our previous tensor model $\mathcal{C}$ in [17]. (c) Hallucinated high-resolution (64 × 48) images using first global then local tensors of our sequential approach. (d) Our final results (64 × 48) after adding the high-frequency component residue to (c). (e) Ground truth face images (64 × 48) of corresponding expressions.

by local patches alone. As a result, super-resolution results in [17] are poor around the mouth region where large appearance variations normally appear when facial expressions change. Our new model here consists of an extra global image-based multimodal tensor $\mathcal{G} = \mathcal{Z}_{\mathcal{G}} \times_1 \mathbf{U}_{\text{idens}} \times_2 \mathbf{U}_{\text{modes}} \times_3 \mathbf{U}_{\text{pixels}}$ that captures intermodal facial structural variation information so is advantageous in synthesizing holistic facial structures across modalities.

To demonstrate the advantage of our sequential step solution, we designed the following comparative experiments on multi-expression hallucination in a single step. We used our previous tensor model $\mathcal{C} = \mathcal{Z}_{\mathcal{C}} \times_1 \mathbf{U}_{\text{idens}} \times_2 \mathbf{U}_{\text{modes}} \times_3 \mathbf{U}_{\text{resos}} \times_4 \mathbf{U}_{\text{patches}} \times_5 \mathbf{U}_{\text{pixels}}$ in [17] (patch size is experimentally decided as 4 × 4 with one pixel overlapping) and a replacement tensor model, similar to that described in Section IV but without a sequential process, $\mathcal{C}' = \mathcal{Z}_{\mathcal{C}'} \times_1 \mathbf{U}_{\text{idens}} \times_2 \mathbf{U}_{\text{modes}} \times_3 \mathbf{U}_{\text{resos}} \times_4 \mathbf{U}_{\text{pixels}}$ to perform, respectively, face hallucination across multi-expressions in one single step.

Example results are shown in Fig. 7, which demonstrate that although the tensor model $\mathcal{C}'$ can generate high-resolution holistic structure of nonlinear facial variations such as different expressions, it is poor at recovering smoothly high-resolution details [Fig. 7(a)]. Our previous tensor model $\mathcal{C}$ is also capable of hallucinating high-resolution details. But the block-type artifacts around the mouth region [Fig. 7(b)] show its weak generalization ability based on local patches alone.

Another technical difference between this paper and our previous work in [17] is on face alignment. In [17], all face images are required to be manually aligned before performing super-resolution. This is not practical for automatic applications in many practical scenarios. We describe a new automatic face alignment algorithm in the following Section VI.

D. Multiview Face Hallucination

We also applied our approach to multiview face hallucination. We used face images from a subset of AR, FERET and Yale databases to form an experimental dataset consisting of 1475 face images of 295 different individuals, in which each individual has five different face views. We manually aligned these face images and established a standard training dataset similar to that used for our multi-expression hallucination experiments above. Some example results are shown in Fig. 8.

VI. Automatic Face Alignment

Accurate pixel-wise face alignment is necessary for successful face hallucination. Traditional learning-based techniques require both training and test face images to be manually aligned. We develop an automatic face alignment algorithm to register low-resolution faces in raw images. The registration process is initialized by an Adaboost [35] face detector, which provides the rough coordinate position of any face in a raw image. Based on this initial position, we start face registration by iteratively optimizing affine warping parameters.

Specifically, assume we have a low-resolution training face dataset $\mathbf{D}$ (which can be obtained by subsampling the high-resolution training face images for tensor construction as in Section III), each face image from which can be registered to a predefined face template. Let $\mathbf{z} = \{z_i\} = \{(x_i, y_i)\}$ be the spatial coordinates of the template. We want to estimate a warping function $\mathbf{W}(z, \mathbf{p})$ so that $I(\mathbf{W}(\mathbf{z}, \mathbf{p}))$ is close to the low-resolution face template, where I is the raw image. Affine transformation is chosen as the warping function

$$\mathbf{W}(z, \mathbf{p}) = \begin{bmatrix} p_1 & p_3 & p_5 \\ p_2 & p_4 & p_6 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{26}$$

where $\mathbf{p} = (p_1, \cdots, p_6)^T$. Let the mean face image of the training set $\mathbf{D}$ be μ, we use intuitively the objective expression

$$\sum_z [I(\mathbf{W}(z, \mathbf{p})) - \mu(z)]^2 \tag{27}$$

to find the optimal affine warp parameter $\mathbf{p}^*$ by minimizing the designated sum of squared error. However, due to the intrinsic nonrigidness of human faces, using the mean face μ as the registration template leads to a biased estimation of $\mathbf{p}^*$, and as a consequence, the warped face image does not retain the facial geometric properties of the original face in the raw image. As a remedy, rather than using the mean, our registration is based on an iterative projected estimation.

More precisely, we apply PCA on our low-resolution training face dataset $\mathbf{D}$, and obtain the eigenvectors $\{K_k\}_{k=1}^R$, eigenvalues $\{\sigma_k^2\}_{k=1}^R$ and the mean face μ. The orthogonal eigenvectors construct the subspace $\Omega = \text{span}(K_1, \ldots, K_R) \sim \mathbb{R}^R$. Thus, the reconstructed image of the warped face in subspace Ω can be expressed as $(\mathbf{K}X + \mu)$, where $\mathbf{K} = [K_1, \ldots, K_R]$, and X is the specified coefficient vector of the warped face image projected in Ω. Instead of using the mean μ, we then use the expression

$$\sum_z [I(\mathbf{W}(z, \mathbf{p})) - (\mathbf{K}X + \mu)_z]^2 \tag{28}$$

Fig. 8. Simulated experiments on hallucinating multiple viewpoint face images. (a) Low-resolution input images (14 × 9) at different single views (obtained by downsampling original test input images). (b)–(f) Synthesized low-resolution (14 × 9) images at frontal, yaw $-/+45°$, and tilt $-/+45°$ views respectively, using the global image-based tensor. (g)–(k) Hallucinated high-resolution (56 × 36) images at these five views, using the local patch-based multiresolution tensor. (l)–(p) Final hallucination results (56 × 36) after adding the high-frequency component residue, using nonparametric patch learning. (q)–(u) Ground truth face images (56 × 36) at these five views.

to simultaneously find the optimal affine warp parameter $\mathbf{p}^*$ and the coefficient vector X^* in the low-resolution PCA subspace. Using this iteratively reconstructed face in PCA subspace as the registration template is geometrically more consistent than the use of the mean face μ, which is equivalent to setting the coefficient vector $X = \mathbf{0}$.

In general, the quadratic expression in (28) is nonconvex, and global optimization over both the warping parameter $\mathbf{p}$ and the corresponding PCA subspace coefficient vector X can result in local minima. We address this problem by iterating over these two sets of parameters: first optimize the warping parameter $\mathbf{p}$ by nonlinear gradient descent since $I(\mathbf{W}(\cdot))$ is nonlinear. We then update the coefficient vector X based on the newly warped face image, and finally iterate. While this method is not guaranteed to converge to a global minimum, it is effective when registration is initialized at the rough face position provided by an automatic face detector.

The minimization of the quadratic expression in (28) is performed by fixing the n^{th} update of the coefficient vector X^n. Similar to the Lucas–Kanade approach [33], based on the current $\mathbf{p}^n$, we wish to compute an increment $\Delta\mathbf{p}^n$

$$\mathbf{p}^{(n+1)} \longleftarrow \mathbf{p}^n + \Delta\mathbf{p}^n \tag{29}$$

then (28) becomes

$$\sum_z \left[I\left(\mathbf{W}(z, \mathbf{p}^n + \Delta\mathbf{p}^n)\right) - (\mathbf{K}X^n + \mu)_z\right]^2. \tag{30}$$

The nonlinear expression in (30) can be linearized by a first order Taylor expansion resulting in the objective expression as

$$\sum_z \left[I\left(\mathbf{W}(z, \mathbf{p}^n)\right) + \nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}^n}\Delta\mathbf{p}^n - (\mathbf{K}X^n + \mu)_z\right]^2 \tag{31}$$

where $\nabla I = (\partial I/\partial x, \partial I/\partial y)$ is the gradient of raw image I evaluated at $\mathbf{W}(\mathbf{z}, \mathbf{p}^n)$. The term $\partial\mathbf{W}/\partial\mathbf{p}^n$ is the Jacobian of the warp. For assumed affine transformation, we have

$$\frac{\partial \mathbf{W}}{\partial \mathbf{p}^n} = \begin{bmatrix} x & 0 & y & 0 & 1 & 0 \\ 0 & x & 0 & y & 0 & 1 \end{bmatrix}. \tag{32}$$

As the minimization of (31) is a least squares problem, a closed form solution is given as

$$\Delta\mathbf{p}^n = \mathbf{H}^{-1}\sum_z \left[\nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}^n}\right]^T \left((\mathbf{K}X^n + \mu)_z - I\left(\mathbf{W}(z, \mathbf{p}^n)\right)\right) \tag{33}$$

where $\mathbf{H}$ is the Hessian matrix and $\mathbf{H} = \sum_z [\nabla I(\partial\mathbf{W}/\partial\mathbf{p}^n)]^T[\nabla I(\partial\mathbf{W}/\partial\mathbf{p}^n)]$. After obtaining the $(n+1)^{th}$ iteration warping parameter $\mathbf{p}^{(n+1)}$, we update the corresponding PCA subspace coefficient vector as

$$X^{(n+1)} = \mathbf{K}^{-1}\left[I\left(\mathbf{W}\left(\mathbf{z}, \mathbf{p}^{(n+1)}\right)\right) - \mu\right]. \tag{34}$$

The global optimization over these two sets of parameters $\mathbf{p}$ and X is realized by iterating the above processes until the sum of squared difference in (30) becomes stable or a specified iteration limit is reached. The coefficient vector X^0 can be initialized by setting $X^0 = \mathbf{0}$, or in other words, the iteration process can start by taking the mean of the low-resolution training face images as a template.

VII. Experimental Results

We performed automatic face alignment and then generalized face super-resolution experiments on the MIT+CMU dataset [34]. We used an AdaBoost face detector [35] to initialize the process. As some of the original test images in the MIT+CMU dataset contain faces of higher resolution, we subsampled them to ensure the sizes of the faces contained in them ranged between 32 × 24 and 16 × 12 (we only considered situations where the Adaboost face detector provided initial face positions). Similarly to our simulated experiments on multiple facial expression hallucination, we chose expression images of neutral, smile, anger, and scream in the AR face dataset for training.

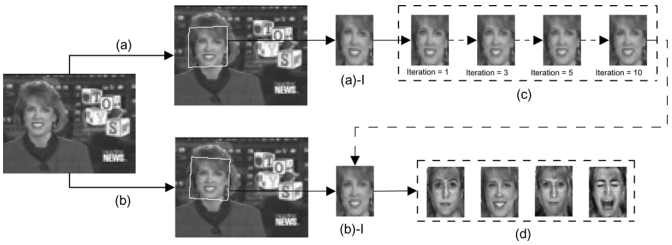

Fig. 9. Comparative experiment on automatic face alignment. (a) Process using the mean of the low-resolution training face images as the registration template; (a)-I is its aligned (warped) low-resolution face. (b) Our proposed automatic face alignment process, with iteratively reconstructed faces in the PCA subspace of the low-resolution training face images as the registration templates; (b)-I is our aligned (warped) low-resolution face. (c) Intermediate face warping results in our iterative optimization of face alignment process [the optimization starts from (a)-I, and converges at (b)-I after ten iterations]. (d) Super-resolved multi-expression images.

These training images were manually aligned to the standard high-resolution size of 64×48.

We performed a comparative experiment as shown in Fig. 9. In Fig. 9(a), the mean of the low-resolution training face images (which can be obtained by subsampling the high-resolution training face images) was used as the registration template, consequently its warped low-resolution face (a)-I is misaligned and does not retain the facial geometric property of the face object contained in the test raw image. Fig. 9(b) used our proposed automatic face alignment algorithm, in which the iteratively reconstructed faces in the PCA subspace of the low-resolution training face images were taken as the registration templates. Clearly, our warped low-resolution face (b)-I is geometrically more consistent with the face in the test raw image. The intermediate face warping results in Fig. 9(c) demonstrate the robustness of our iterative optimization of face alignment algorithm, which was realized by iteratively minimizing the quadratic expression in (28). The optimization process started from the misaligned face warping (a)-I, and converged at our face warping (b)-I after ten iterations. We show more experimental results in Fig. 10. The low-resolution faces and their corresponding hallucinated multiple facial expression images are displayed aside the raw test images, from which the low-resolution faces were automatically detected, aligned and extracted. The results are shown at the resolution of 64×48. Fig. 10 demonstrates that our generalized face super-resolution approach is able to produce reasonable results although the quality of many raw test images is degraded by different kinds of imaging noise. However, these hallucinated results appear noisy, and are relatively poor compared with the results from our simulated experiments. This shows a weakness of learning-based super-resolution techniques which require a certain similarity between training and test images.

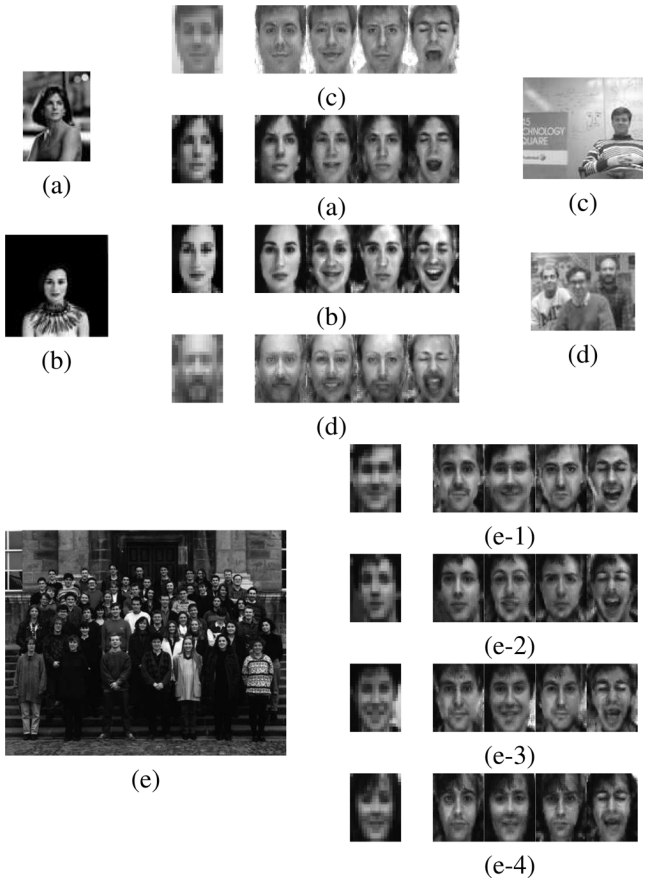

Fig. 10. Experimental results on multiple facial expression super-resolution based on automatic face alignment: the automatically aligned and extracted low-resolution faces in rows (e-1), (e-3), and (e-4) are with a smiling expression, and others are with a neutral expression.

VIII. Discussion

A. Global Versus Local

Global modeling of face images is advantageous in capturing the holistic structure and variation of facial appearance, while local modeling on patches or pixels incorporates more high-frequency information. Consequently, global super-resolution approaches are more robust in ensuring reconstruction lying in face space, and local approaches can recover higher resolution image details. In this paper, we formulate a unified tensor space based model for face hallucination across modalities: a global tensor for synthesizing multimodal holistic facial appearances, and a local tensor for hallucinating high-resolution faces for each modality.

B. Manual Alignment Versus Automatic Alignment

Pixel-wise accuracy in alignment between test and training face images is essential in learning-based face super-resolution. In our simulated experiments on public face databases, the test inputs were manually aligned by three predefined feature points: the mouth center and centers of the left and right eyes. This is unrealistic in application to many practical scenarios, where the faces captured in raw images are normally of nonfrontal views at low resolution. One may adopt an automatic face detector to find faces. However, accuracy in the position and scale of face images detected by current state-of-the-art face detectors remains poor.

To cope with this problem, we develop an automatic face alignment algorithm to register low-resolution faces in raw images. The registration process is initialized by an automatic face detector, which provides a rough coordinate position of any face

found in the raw image. Based on this initial position, we initiate pixel-wise face registration by iteratively optimizing affine transformation parameters, which warp any probe face in raw images to its projection in a PCA subspace constructed from low-resolution training face images.

C. Synthesis Versus Recognition

Face hallucination is generally a high-resolution face synthesis problem. However, face images captured in practical scenarios, such as by surveillance cameras, usually are at low resolution, which significantly limit the performance of face recognition systems. We can apply hallucination techniques to low-resolution face images, to effectively enhance the original image quality, and hence improve the accuracy of recognition.

Rather than separating the high-resolution face synthesis and recognition into two independent steps, for instance face recognition is only performed after yielding synthesized high-resolution face images, the two processes can be performed simultaneously. To this end, one can directly use the recovered test face coefficient vector in the high-resolution face subspace for both high-resolution synthesis and recognition [32]. Furthermore, given the intrinsic coupling power of the tensor space for mapping different scales (i.e., variations in image resolution) with different modalities, we can use the unique tensor space identity parameter vector for both multimodal high-resolution face synthesis, and face recognition across different expressions, poses or illuminations, as demonstrated in our earlier work [16]. In this way, face synthesis and recognition are unified.

IX. Summary

In this paper, we proposed a generalized approach to hallucinate face images across multiple modalities (generalization to variations such as facial expression or pose) based on a unified global and a local tensor space representation. Specifically, we modeled the high- and low-resolution training face images of multiple modalities. We can reduce this unified model to a global image-based tensor for modeling the mappings among different facial modalities, and a local patch-based multiresolution tensor for incorporating high-resolution face image details. Given any low-resolution face input of a single modality, we first synthesize multiple low-resolution face images of different modalities using a trained global tensor. Based on these synthesized low-resolution face images, we then use the trained local tensor to construct corresponding high-resolution image for each facial modality. For the highest frequency visual information, we further add a high-frequency component residue using nonparametric patch learning from high-resolution training face images. For applications in practical scenarios where faces captured in raw images are normally nonfrontal views at low resolution, we developed an automatic pixel-wise face registration algorithm. The registration was realized by iteratively optimizing affine transformation parameters, which warp any probing face in raw images to its projection in a PCA subspace constructed from the low-resolution training face images. We performed simulated and practical experiments on multi-expression and multiview face hallucination. Compared with other benchmark face super-resolution techniques, our experimental results demonstrate performance superiority and novelty in terms of both single modal face hallucination, and hallucination across multiple facial modalities.

References

[1] M. Elad and A. Feuer, "Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images," *IEEE Trans. Image Process.*, vol. 6, no. 12, pp. 1646–1658, Dec. 1997.

[2] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graph. Models Image Process.*, vol. 53, pp. 231–239, 1991.

[3] R. R. Schulz and R. L. Stevenson, "Extraction of high-resolution frames from video sequences," *IEEE Trans. Image Process.*, vol. 5, no. 6, pp. 996–1011, Jun. 1996.

[4] R. C. Hardie, K. J. Barnard, and E. E. Armstrong, "Joint MAP registration and high-resolution image estimation using a sequence of undersampled images," *IEEE Trans. Image Process.*, vol. 6, no. 12, pp. 1621–1633, Dec. 1997.

[5] G. Rochefort, F. Champagnat, G. Le Besnerais, and J. F. Giovannelli, "An improved observation model for super-resolution under affine motion," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3325–3337, Nov. 2006.

[6] H. He and L. P. Kondi, "An image super-resolution algorithm for different error levels per frame," *IEEE Trans. Image Process.*, vol. 15, no. 3, pp. 592–603, Mar. 2006.

[7] S. Farsiu, M. Elad, and P. Milanfar, "Multiframe demosaicing and super-resolution of color images," *IEEE Trans. Image Process.*, vol. 15, no. 1, pp. 141–159, Jan. 2006.

[8] D. Robinson and P. Milanfar, "Statistical performance analysis of super-resolution," *IEEE Trans. Image Process.*, vol. 15, no. 6, pp. 1413–1428, Jun. 2006.

[9] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[10] M. A. O. Vasilesescu and D. Terzopoulos, "Multilinear image analysis for facial recognition," in *Proc. Int. Conf. Pattern Recognition*, 2002, pp. 511–514.

[11] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear analysis of image ensembles: TensorFaces," in *Proc. Eur. Conf. Computer Vision*, 2002, pp. 447–460.

[12] D. Vlasic, M. Brand, H. Pfister, and J. Popovic, "Face transfer with multilinear models," *ACM Trans. Graphics*, vol. 24, no. 3, pp. 426–433, 2005.

[13] T. G. Kolda, "Orthogonal tensor decompositions," *SIAM J. Matrix Anal. Appl.*, vol. 23, pp. 243–255, 2001.

[14] L. D. Lathauwer, B. D. Moor, and J. Vandewalle, "Multilinear singular value tensor decompositions," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1253–1278, 2000.

[15] H. Wang and N. Ahuja, "Facial expression decomposition," in *Proc. IEEE Int. Conf. Computer Vision*, 2003, pp. 958–965.

[16] K. Jia and S. Gong, "Multi-modal tensor face for simultaneous super-resolution and recognition," in *Proc. IEEE Int. Conf. Computer Vision*, 2005, pp. 1683–1690.

[17] K. Jia and S. Gong, "Multi-resolution tensor for facial expression hallucination," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006, pp. 395–402.

[18] L. Zhang and D. Samaras, "Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 351–363, Mar. 2006.

[19] X. He, S. Yan, Y. Hu, P. Niyogi, and H. J. Zhang, "Face recognition using laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[20] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1063–1074, Sep. 2003.

[21] X. Wang and X. Tang, "Hallucinating face by eigentransformation," *IEEE Trans. Syst., Man, Cybern. C, Cybern.*, vol. 35, no. 3, pp. 425–434, Mar. 2005.

[22] W. Liu, D. Lin, and X. Tang, "Hallucinating faces: TensorPatch super-resolution and coupled residue compensation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005, pp. 478–484.

[23] H. Chang, D. Yeung, and Y. Xiong, "Super-resolution through neighbour embedding," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004, pp. 275–282.

[24] X. Lu, A. K. Jain, and D. Colbry, "Matching 2.5D face scans to 3D models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 31–43, Jan. 2006.
[25] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Advances and challenges in super-resolution," *Int. J. Imag. Syst. Technol.*, vol. 14, no. 2, pp. 47–57, 2004.
[26] S. Park, M. Park, and M. G. Kang, "Super-resolution image reconstruction: A technical overview," *IEEE Signal Process. Mag.*, vol. 20, no. 3, pp. 21–36, Mar. 2003.
[27] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1167–1183, Sep. 2002.
[28] S. Baker and T. Kanade, "Hallucinnating faces," in *Proc. IEEE Automatic Face and Gesture Recognition*, 2000, pp. 83–90.
[29] W. T. Freeman and E. C. Pasztor, "Learning low-level vision," in *Proc. IEEE Int. Conf. Computer Vision*, 1999, pp. 1182–1189.
[30] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning low-level vision," *Int. J. Comput. Vis.*, vol. 40, no. 1, pp. 25–47, 2000.
[31] D. P. Capel and A. Zisserman, "Super-resolution from multiple views using learnt image models," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001, pp. 627–634.
[32] B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes, III, and R. M. Mersereau, "Eigenface-domain super-resolution for face recognition," *IEEE Trans. Image Process.*, vol. 12, no. 5, pp. 597–606, May 2003.
[33] S. Baker and I. Matthews, "Lucas–Kanade 20 years on: A unifying framework," *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 221–255, Mar. 2004.
[34] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, pp. 23–38, Jan. 1998.
[35] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.
[36] C. Liu, H. Shum, and C. Zhang, "A two-step approach to hallucinating faces: Global parametric model and local nonparametric model," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001, pp. 192–198.
[37] C. Liu, H. Shum, and W. T. Freeman, "Face hallucination: Theory and practice," *Int. J. Comput. Vis.*, vol. 75, no. 1, pp. 115–134, 2007.
[38] P. J. Burt, "Fast filter transforms for image processing," *Comput. Graph. Image Process.*, vol. 16, pp. 20–51, 1980.
[39] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. COM-31, no. 4, pp. 532–540, Apr. 1983.
[40] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2001, pp. 511–518.
[41] J. E. Besag, "On the statistical analysis of dirty pictures (with discussion)," *J. Roy. Statist. Soc. B*, vol. 48, no. 3, pp. 259–302, 1986.
[42] C. M. Bishop, A. Blake, and B. Marthi, "Super-resolution enhancement of video," presented at the Artificial Intelligence and Statistics, Society for Artificial Intelligence and Statistics, 2003.
[43] G. Dedeoglu, T. Kanade, and J. August, "High-zoom video hallucination by exploiting spatio-temporal regularities," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004, pp. 151–158.
[44] J. S. Park and S. W. Lee, "Resolution enhancement of facial image based on top-down learning," in *Proc. Int. Multimedia Conf.: 1st ACM SIGMM Int. Workshop on Video Surveillance*, 2003, pp. 59–64.
[45] J. Sun, N. Zhang, H. Tao, and H. Shum, "Image hallucination with primal sketch priors," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2003, pp. 729–736.

Kui Jia received the B.Eng. degree in electrical engineering from Northwestern Polytechnic University, China, in 2001, the M.Eng. degree in electrical and computer engineering from the National University of Singapore in 2003, and the Ph.D. degree in computer science from Queen Mary, University of London, London, U.K., in 2007.

He is currently with the Shenzhen Institute of Advanced Integration Technology, Chinese Academy of Sciences/Chinese University of Hong Kong, as a Research Assistant Professor. His research focuses on learning for vision, image enhancement, super-resolution, human activity analysis, face analysis, and object recognition.

Shaogang Gong received the D.Phil. degree in computer vision from Oxford University, Oxford, U.K., in 1989.

He is Professor of visual computation at Queen Mary, University of London, London, U.K. He has published over 170 papers and the book *Dynamic Vision: From Images to Face Recognition*. His work focuses on motion and video analysis; object detection, tracking, and recognition; face and expression recognition; gesture and action recognition; human behavior recognition and anomaly detection.

He is a Fellow of the Institution of Electrical Engineers and a Member of the U.K. Computing Research Committee.