In Proc. IEEE International Conference on Computer Vision (ICCV'05), October 2005

Multi-Modal Tensor Face for Simultaneous Super-Resolution and Recognition

Kui Jia and Shaogang Gong
Department of Computer Science
Queen Mary, University of London, London E1 4NS, UK
{chrisjia,sgg}@dcs.qmul.ac.uk

Abstract

Face images of non-frontal views under poor illumination with low resolution reduce dramatically face recognition accuracy. This is evident most compellingly by the very low recognition rate of all existing face recognition systems when applied to live CCTV camera input. In this paper, we present a Bayesian framework to perform multi-modal (such as variations in viewpoint and illumination) face image super-resolution for recognition in tensor space. Given a single modal low-resolution face image, we benefit from the multiple factor interactions of training tensor, and super-resolve its high-resolution reconstructions across different modalities for face recognition. Instead of performing pixel-domain super-resolution and recognition independently as two separate sequential processes, we integrate the tasks of super-resolution and recognition by directly computing a maximum likelihood identity parameter vector in high-resolution tensor space for recognition. We show results from multi-modal super-resolution and face recognition experiments across different imaging modalities, using low-resolution images as testing inputs and demonstrate improved recognition rates over standard tensorface and eigenface representations.

1. Introduction

Many representations and models have been proposed for face recognition in recent years, mostly based on linear models such as PCA [1], ICA [3] and LDA [2]. Most of them cope poorly with nonlinear variations in viewing conditions away from the training data. More recently TensorFace [5, 4] has been proposed for a multi-linear analysis to model explicitly the multiple modes of variations in facial shape, expression, pose and illumination and their inter-relationships. Reported experiments suggested improved recognition performance over traditional approach [1]. However, the recognition rates based on these algorithms decrease dramatically with low-resolution inputs. To overcome this problem, super-resolution techniques [14, 16, 18, 17] can be exploited to generate a high-resolution image given a single or set of low-resolution input images. The computation of super-resolution requires the recovering of lost high-frequency information occurring during the image formation process. Super-resolution can be performed using either reconstruction-based [8, 9, 10, 11] or learning-based [15, 13, 14, 16, 18, 19] approaches. In this work, we focus on learning-based approaches.

Capel and Zisserman [16] used eigenface from a training face database as model prior to constrain and super-resolve low-resolution face images. To further improve the performance, they divided human face into six unrelated parts and apply PCA on them separately. Combined with MAP estimator, they can recover the result from a high-resolution eigenface space. A similar method was proposed by Baker and Kanade [13]. Rather than using the whole or parts of a face, they established the prior based on a set of training face images pixel by pixel using Gaussian, Laplacian and feature pyramids. Freeman and Pasztor [15] took a different approach for learning-based super-resolution. Specifically, they tried to recover the lost high-frequency information from low-level image primitives, which were learnt from several general training images. They broke the images and scenes into a Markov network, and learned the parameters of the network from the training data. To find the best scene explanation given new image data, they applied belief propagation in the Markov network. A very similar image hallucination approach was also introduced in [19]. They used the primal sketch as the prior to recover the smoothed high-frequency information. Liu and Shum [18] combined the PCA model-based approach and Freeman's image primitive technique. They developed a mixture model combing a global parametric model called "global face image" carrying common facial properties, and a local nonparametric model called "local feature image" recording local individualities. The high-resolution face image was naturally a composition of both.

To go beyond the current super-resolution techniques which only consider face images under fixed imaging conditions in terms of pose, expression and illumination, we present in this work a Bayesian model to perform simultaneously multi-modal face image super-resolution and recognition in tensor space. Given a single modal low-resolution face image, we benefit from the multiple factor interactions of training tensor, and super-resolve its high-resolution reconstructions across different modalities for face recognition. Instead of performing pixel-domain super-resolution and recognition independently as two separate sequential processes, we integrate the tasks of super-resolution and recognition by directly computing a maximum likelihood identity parameter vector in high-resolution tensor space for recognition.

The paper is organized as follows. Section 2 introduces multilinear analysis and tensor singular value decomposition (SVD). In section 3, we derive a Bayesian framework to perform multi-modal super-resolution, and present an algorithm optimizing the high-resolution identity parameter vector in tensor space. Section 4 discusses experimental results before conclusions are drawn in section 5.

2. Multilinear Analysis: Tensor SVD

Multilinear analysis [5, 7, 6] is a general extension of the traditional linear methods such as PCA or matrix SVD. Instead of modelling the relations within vectors or matrices, multilinear analysis provides a means to investigate the mappings between multiple factor spaces. In this context, the multilinear equivalents of vectors (first order) and matrices (second order) are called tensors, multidimensional matrices or multiway arrays. Tensor singular value decomposition or higher-order singular value decomposition (HOSVD) [7] is a multilinear generalization of the concept of matrix SVD. In the following, we denote scalars by lower-case letters ($a, b, \dots; \alpha, \beta, \dots$), vectors by upper-case ($A, B, \dots$), matrices by bold upper-case ($\mathbf{A}, \mathbf{B}, \dots$), and tensors by calligraphic letters ($\mathcal{A}, \mathcal{B}, \dots$).

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 $\mathcal{A}_{i_1 \dots i_n \dots i_N}$ or $a_{i_1 \dots i_n \dots i_N}$, where $1 \leq i_n \leq I_n$. If we refer to I_n rank in tensor terminology, we generalize the matrix definition and call 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 the tensor $\mathcal{A}$ by taking the mode-n vectors as the column vectors of matrix $\mathbf{A}_{(n)} \in R^{I_n \times (I_1 I_2 \dots I_{n-1} I_{n+1} \dots I_N)}$. These tensor unfoldings provide an easy manipulation in tensor algebra and if necessary, we can reconstruct the tensor by an 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

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

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

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

Given the tensor $\mathcal{A} \in R^{I_1 \times I_2 \cdots \times I_N}$ and the matrices $\mathbf{F} \in R^{J_n \times I_n}$ and $\mathbf{G} \in R^{J_m \times I_m}$, the following property holds true in tensor algebra [6, 7]:

$$(\mathcal{A} \times_n \mathbf{F}) \times_m \mathbf{G} = (\mathcal{A} \times_m \mathbf{G}) \times_n \mathbf{F} = \mathcal{A} \times_n \mathbf{F} \times_m \mathbf{G}.$$

In singular value decompositions 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 in terms of mode-n product as $\mathbf{D} = \mathbf{\Sigma} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2$. We can generalize the 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 Eq.(1), the core tensor $\mathcal{Z}$ can be computed by

$$\mathcal{Z} = \mathcal{A} \times_1 \mathbf{U}_1^T \times_2 \mathbf{U}_2^T \cdots \times_N \mathbf{U}_N^T.$$

Eq.(2) gives the basic representation of multilinear model. If we investigate the mode-n unfolding and folding, and rearrange Eq.(2), we can have

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

where $\mathcal{S}$ is a subtensor of $\mathcal{A}$ corresponding to a fixed row vector V_n^T of the 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 tensor structure. If we index into basis tensor $\mathcal{B}$ for more particular V_n^T, we can get different modal sample vector data.

3. Multi-Modal Super-Resolution in Tensor Space

In this section, we first build a tensor structure for face images of different modalities including varying illumination, viewpoint (head pose) and people identity. We then derive an algorithm for super-resolution in tensor parameter vector space.

3.1. Modelling Face Images in Tensor Space

We construct a tensor structure from multi-modal face images and use HOSVD to decompose them. The decomposed model can be expressed as

$$\mathcal{D} = \mathcal{Z} \times_1 \mathbf{U}_{idens} \times_2 \mathbf{U}_{views} \times_3 \mathbf{U}_{illums} \times_4 \mathbf{U}_{pixels},$$

where tensor $\mathcal{D}$ groups the multi-modal face images into a tensor structure, and the core tensor $\mathcal{Z}$ governs the interactions between the 4 mode factors. The mode matrix $\mathbf{U}_{idens}$ spans the parameter space of different people identities, the mode matrix $\mathbf{U}_{views}$ spans the parameter space of changing head poses, and the mode matrix $\mathbf{U}_{illums}$ spanning the space of varying illumination parameters, the mode matrix $\mathbf{U}_{pixels}$ spanning space of face images.

With decomposed tensor of multi-modal face images, we can perform super-resolution in tensor parameter vector space. In such a formulation, the observation is an identity parameter vector computed by projecting testing low-resolution face images onto a tensor constructed from low-resolution training images, and proposed algorithm super-resolve the true identity parameter vector in a tensor constructed from high-resolution training images. We start with the pixel-domain image observation model. Assuming D_L is a vectorized observed low-resolution image, D_H is the unknown true scene, and $\mathbf{A}$ is a linear operator that incorporates the motion, blurring and downsampling processes, the observation model can be expressed as

$$D_L = \mathbf{A}D_H + n, \tag{3}$$

where n represents the noise in these processes.

The unknown high-resolution image D_H and observed image D_L have identity parameter vectors that lie in the respective tensor spaces. These parameter vectors provide a unique representation for any people identity independent of the potentially varying modalities such as viewpoint and illumination. Rather than performing super-resolution on pixel-domain modal by modal, we derive a model for the reconstruction of identity parameter vectors in the high-resolution tensor space.

Based on the tensor algebra introduced in section 2, suppose we have a basis tensor

$$\mathcal{B} = \mathcal{Z} \times_2 \mathbf{U}_{views} \times_3 \mathbf{U}_{illums} \times_4 \mathbf{U}_{pixels}, \tag{4}$$

we can index into this basis tensor for a particular viewpoint v and illumination l to yield a basis subtensor

$$\mathcal{B}_{v,l} = \mathcal{Z} \times_4 \mathbf{U}_{pixels} \times_2 V_v^T \times_3 V_l^T,$$

for each of the face imaging modalities. Then the subtensor containing the individual image data can be expressed as

$$\mathcal{D}_{v,l} = \mathcal{B}_{v,l} \times_1 V^T + \mathcal{E}_{v,l}, \tag{5}$$

where V^T represents the identity parameter row vector and $\mathcal{E}_{v,l}$ stands for the tensor modelling error for modalities of viewpoints v and illumination l. For ease of notation and readability, we will use the mode-1 unfolding matrix to represent tensors. Then the matrix representation of Eq.(5) becomes

$$\mathbf{D}_{v,l}^{(1)} = V^T \mathbf{B}_{v,l}^{(1)} + e_{v,l}. \tag{6}$$

The counterpart of pixel-domain image observation model (3) is then given as

$$\hat{\mathbf{B}}_{v,l}^{T(1)} \hat{V} + \hat{e}_{v,l} = \mathbf{A}\mathbf{B}_{v,l}^{T(1)} V + \mathbf{A}e_{v,l} + n, \tag{7}$$

where $\hat{\mathbf{B}}_{v,l}^{T(1)}$ and $\mathbf{B}_{v,l}^{T(1)}$ are the low-resolution and high-resolution unfolded basis subtensor, $\hat{V}$ and V are the identity parameter vectors for the low-resolution testing face image and unknown high-resolution image.

Independent of changing viewpoints v and illuminations l, the low- and high-resolution parameter vectors $\hat{V}$ and V are the unique representations of the low-resolution input and its corresponding high-resolution image to be estimated. Without loss of generality we can rewrite Eq.(7) as

$$\hat{\mathbf{B}}^{T(1)} \hat{V} + \hat{E} = \mathbf{A}\mathbf{B}^{T(1)} V + \mathbf{A}E + N, \tag{8}$$

where $\hat{\mathbf{B}}^{T(1)}$ and $\mathbf{B}^{T(1)}$ are the unfolded basis tensors, and $\hat{E}$ and E are the combined tensor modelling error over all modal face images.

Low-resolution observation images contain very little high-frequency information after the processes of downsampling and blurring, so we can safely neglect the error $\hat{E}$ and multiply both sides of Eq.(8) by $\mathbf{\Psi} = (\hat{\mathbf{B}}^{(1)} \hat{\mathbf{B}}^{T(1)})^{-1} \hat{\mathbf{B}}^{(1)})$ on the left, we obtain

$$\hat{V} = \mathbf{\Psi}\mathbf{A}\mathbf{B}^{T(1)} V + \mathbf{\Psi}\mathbf{A}E + \mathbf{\Psi}N, \tag{9}$$

where $\mathbf{\Psi}$ is the pseudoinverse of $\hat{\mathbf{B}}^{T(1)}$. Eq.(9) gives the relation between the unknown "true" identity parameter vector V and the observed low-resolution counterpart $\hat{V}$. In Fig.(1), we use the multi-view example to illustrate the whole process of our multi-modal super-resolution and recognition in tensor space.

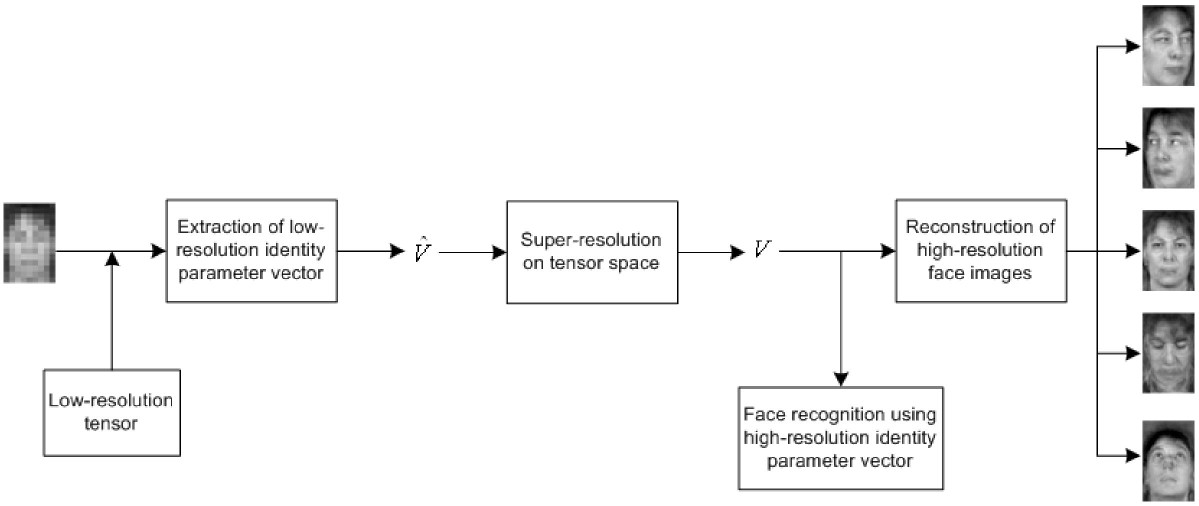
Extraction of low-resolution identity parameter vector
$\hat{V}$
Super-resolution on tensor space
V
Reconstruction of high-resolution face images
Low-resolution tensor
Face recognition using high-resolution identity parameter vector

Figure 1: An illustration of our multi-modal super-resolution and recognition process in tensor space using a multi-view super-resolution example.

3.2. A Bayesian Formulation

We use the Bayesian estimation algorithm to solve Eq.(9). The maximum *a posteriori* probability (MAP) estimation of the high-resolution identity parameter vector V can be expressed as

$$\widetilde{V} = \arg\max_{V}\{p(\hat{V}|V)p(V)\}, \tag{10}$$

where $p(\hat{V}|V)$ is the conditional probability modelling the relations between $\hat{V}$ and V, and $p(V)$ is a prior probability. We can assume the prior probability as Gaussian

$$p(V) = \frac{1}{Z}\exp(-(V-\mu_V)^T\mathbf{\Lambda}^{-1}(V-\mu_V)),$$

where $\mathbf{\Lambda}$ is the covariance matrix for all the training parameter vectors V_i. In our tensor structure, the indentity parameter vectors V_i comes from the row vectors of *orthogonal* matrix $\mathbf{U}_{idens}$. In this sense, the prior $p(V)$ just simply leads the optimum $\widetilde{V}$ in Eq.(10) to the mean value μ_V. So Eq.(10) degenerates to the maximum likelihood (ML) estimator

$$\widetilde{V} = \arg\max_{V} p(\hat{V}|V). \tag{11}$$

To solve the above equation, we define a total noise F that consists of the tensor representation error E and the pixel-domain observation noise N, and rewrite Eq.(9) as

$$\hat{V} = \mathbf{\Psi}\mathbf{A}\mathbf{B}^{T(1)}V + \mathbf{\Psi}F. \tag{12}$$

Now we need derive the distribution of the projected noise $p(\mathbf{\Psi}F)$. Before that, we can write the probability distribution of F as

$$p(F) = \frac{1}{Z}\exp\left(-(F-\mu_F)^T\mathbf{K}^{-1}(F-\mu_F)\right),$$

where $\mathbf{K}$ is a defined diagonal covariance matrix and Z is a normalization constant. Since $\hat{\mathbf{B}}^{(1)}\hat{\mathbf{B}}^{T(1)}$ is nonsingular, $p(\mathbf{\Psi}F)$ can also be modeled as jointly Gaussian, then we have

$$p(\mathbf{\Psi}F) = \frac{1}{Z}\exp\left(-(\mathbf{\Psi}F-\mathbf{\Psi}\mu_F)^T\mathbf{Q}^{-1}(\mathbf{\Psi}F-\mathbf{\Psi}\mu_F)\right), \tag{13}$$

where $\mathbf{\Psi}\mu_F$ is the projected mean error and $\mathbf{Q}$ is the new covariance matrix computed by

$$\mathbf{Q} = \mathbf{\Psi}\mathbf{K}\hat{\mathbf{B}}^{T(1)}. \tag{14}$$

Based on Eq.(12) and Eq.(13), we find the conditional probability $p(\hat{V}|V)$ as

$$\begin{aligned} p(\hat{V}|V) &= \frac{1}{Z}\exp\Big(-(\hat{V}-\mathbf{\Psi}\mathbf{A}\hat{\mathbf{B}}^{T(1)}V-\mathbf{\Psi}\mu_F)^T \\ &\qquad \mathbf{Q}^{-1}(\hat{V}-\mathbf{\Psi}\mathbf{A}\hat{\mathbf{B}}^{T(1)}V-\mathbf{\Psi}\mu_F)\Big). \end{aligned}$$

Then finally we obtain the ML estimator $\widetilde{V}$ as

$$\begin{aligned} \widetilde{V} &= \arg\min_{V}\Big((\hat{V}-\mathbf{\Psi}\mathbf{A}\hat{\mathbf{B}}^{T(1)}V-\mathbf{\Psi}\mu_F)^T \\ &\qquad \mathbf{Q}^{-1}(\hat{V}-\mathbf{\Psi}\mathbf{A}\hat{\mathbf{B}}^{T(1)}V-\mathbf{\Psi}\mu_F)\Big). \end{aligned} \tag{15}$$

In the above expression of ML estimation, the statistics of mean μ_F and covariance matrix $\mathbf{K}$ can be computed based on the training images. Assuming we have I training people, and for each of them we have M training images of different modalities, then we estimate the mean and covariance matrix as follows

$$\mu_F \cong \frac{1}{IM}\sum_{i=1}^{I}\sum_{m=1}^{M}(\hat{\mathbf{D}}_{i,m}^{T(1)} - \mathbf{A}\mathbf{B}_m^{T(1)}V_i),$$

and

$$\mathbf{K} \cong \frac{1}{IM}\sum_{i=1}^{I}\sum_{m=1}^{M}(\hat{\mathbf{D}}_{i,m}^{T(1)} - \mathbf{A}\mathbf{B}_{m}^{T(1)}V_i - \mu_F) \cdot(\hat{\mathbf{D}}_{i,m}^{T(1)} - \mathbf{A}\mathbf{B}_{m}^{T(1)}V_i - \mu_F)^T,$$

where $\hat{\mathbf{D}}_{i,m}^{T(1)}$ represents every low-resolution training image and V_i is the high-resolution identity parameter vector for each training people. We set off-diagonals of $\mathbf{K}$ to zero and use Eq.(14) to obtain $\mathbf{Q}$.

We use the iterative steepest descent method for ML estimation of $\widetilde{V}$. Defining $C(V)$ as the cost function to be minimized, V can be updated in the direction of the negative gradient of $C(V)$. The updating equation can be expressed as

$$V_{n+1} = V_n - \alpha\nabla C(V_n), \qquad (16)$$

where α is the step size. We choose the cost function according to Eq.(15) as

$$\begin{aligned} C(V) \quad = \quad & (\hat{V} - \mathbf{\Psi}\mathbf{A}\hat{\mathbf{B}}^{T(1)}V - \mathbf{\Psi}\mu_F)^T \\ & \mathbf{Q}^{-1}(\hat{V} - \mathbf{\Psi}\mathbf{A}\hat{\mathbf{B}}^{T(1)}V - \mathbf{\Psi}\mu_F), \end{aligned}$$

and take the derivative of $C(V)$ with respect to V, the gradient can be computed as

$$\nabla C(V) = -\hat{\mathbf{B}}^{(1)}\mathbf{A}^T\mathbf{\Psi}^T\mathbf{Q}^{-1}(\hat{V} - \mathbf{\Psi}\mathbf{A}\hat{\mathbf{B}}^{T(1)}V - \mathbf{\Psi}\mu_F).$$

In summary, everything but $\hat{V}$ and V are known (In our experiments, the low-resolution images are blurred and downsampled manually, so we keep the the image observation model parameter $\mathbf{A}$ in the data preparation processes). The identity parameter vector $\hat{V}$ on low-resolution tensor space is obtained by projecting the testing face image $\hat{D}$ onto basis subtensors of all modalities, and then reconstruct them by projecting back, the parameter vector that gives the minimum reconstruction error is chosen as $\hat{V}$, which is essentially a modal estimation process. Based on Eq.(6), the expression can be written as

$$\hat{V} = \arg\min_{\hat{V}_{v,l}} \|\hat{D} - \hat{\mathbf{B}}_{v,l}^{T(1)}\hat{V}_{v,l}\|, \qquad (17)$$

for all the combinations of viewpoints v and illumination l, where $\hat{V}_{v,l}$ can be computed as $\hat{V}_{v,l} = \mathbf{\Psi}_{v,l}\hat{D}$ and $\mathbf{\Psi}_{v,l}$ is the pseudoinverse of $\hat{\mathbf{B}}_{v,l}^{T(1)}$. To summarize, the complete algorithm is as follows.

- Compute the initial estimate of V_0 by bilinearly interpolating the given low-resolution testing face image to the same size of the high-resolution training images, and projecting it onto the training tensor space.
- Obtain the identity parameter vector $\hat{V}$ using Eq.(17).
- Repeat the process of optimizing V_n in Eq.(16).
- Obtain the ML estimation $\widetilde{V}$.

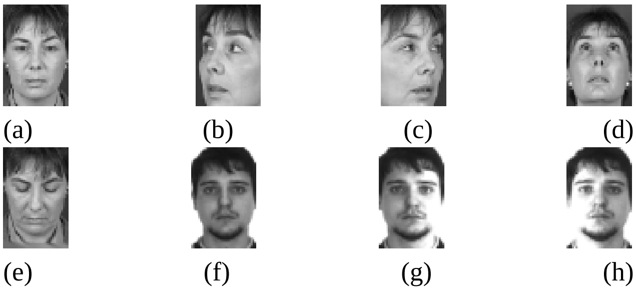

Figure 2: Example images in our dataset: (a), (b), (c), (d) and (e) are 56×36 face images at frontal, yaw -/+45 degrees and tilt -/+ 45 degrees views; (f), (g) and (h) are 56×36 face images under three different illumination conditions of Illum-I, Illum-II and Illum-III.

4. Experiments

In this section, we present first results on super-resolving face images in multiple views given a single view low-resolution testing image. We then show results on super-resolving face images under different illumination conditions given a single illumination low-resolution testing image. We further present results on face recognition across different 3D pose and illumination conditions, based on super-resolved identity parameter vectors in a high-resolution tensor space.

For our experiments, we used face images from a subset of AR, FERET and Yale databases to form two datasets for multi-view and multi-illumination experiments respectively. The multi-view dataset has two sets of face images of 295 different individuals captured at two different occasions, and each set consists of 1475 images of these 295 individuals, in which each individual has 5 different view face images. For multi-illumination dataset, we has one subset of 399 images of 133 person, each of them have 3 face images with 3 different illuminations (Illum-I, Illum-II an Illum-III), and another subset of 133 images of the same 133 persons, but with a different expression under condition of illum-I. Originally face images from AR, FERET and YALE databases have different sizes, and also the area of the image occupied by face varies considerably. To establish a standard training dataset, we aligned these face images manually by hand marking the location of 3 points: the centers of the eyeballs and the lower tip of the nose. These 3 points define an affine warp, which was used to warp the images into a canonical form. Examples of our dataset are shown in Fig.2.

4.1. Multi-Modal Super-Resolutions

We performed two sets of experiments on multi-modal super-resolution using our model derived in section 3. In the first experiment, we used one set of 1475 face im-

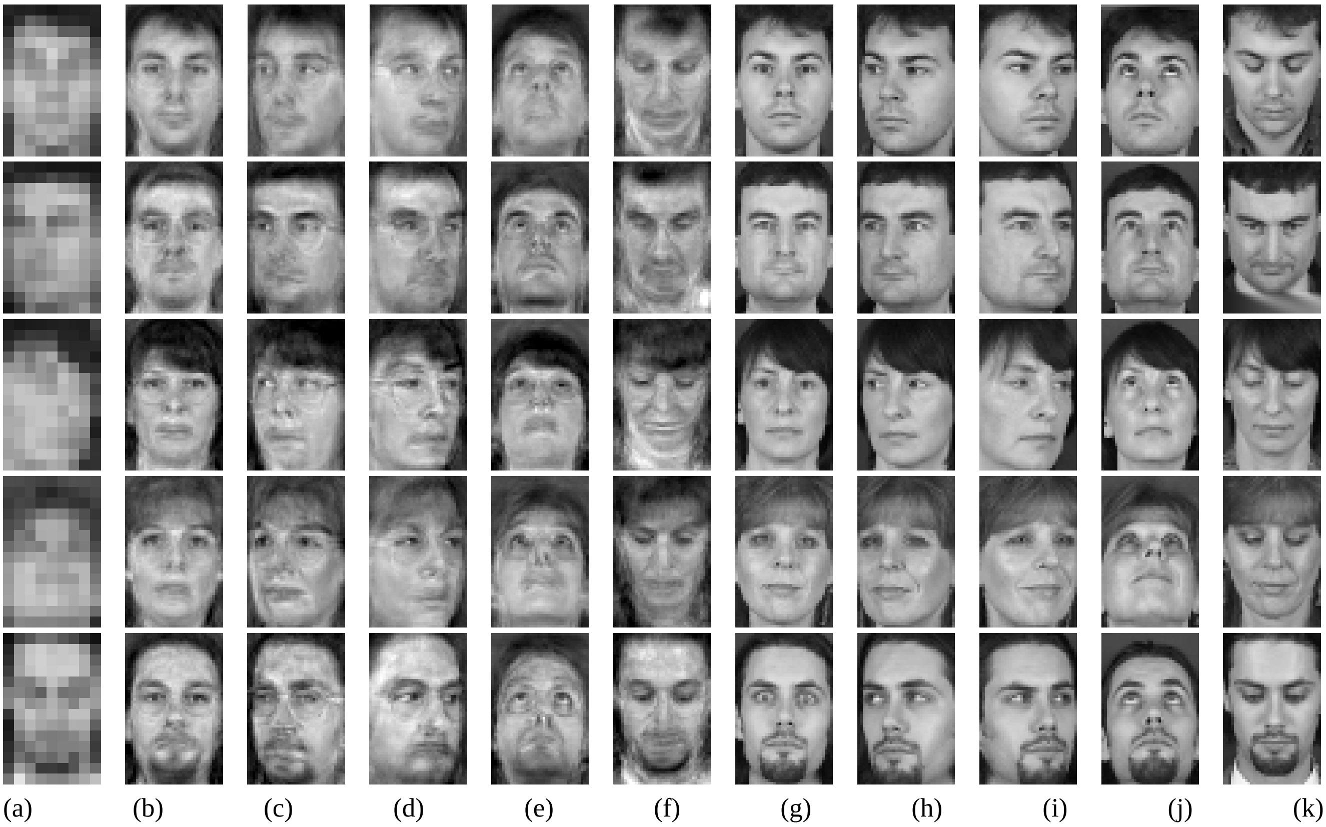

Figure 3: Experiments on super-resolving multi-view face images given a single view low-resolution input: (a) are low-resolution input images (14×9) at different single views (obtained by downsampling original testing input images); (b) - (f) are high-resolution reconstruction results (56×36) at frontal, yaw -/+45 degrees, and tilt -/+45 degrees views respectively; and (g) - (k) are ground truth face images at these 5 views.

ages of 295 individuals in our multi-view dataset. Given a low-resolution single view face image, we super-resolved 5 high-resolution outputs at 5 different views covering the frontal, yaw -/+45 degrees, and tilt -/+45 degrees. Some example results from this experiment is shown in Figure 3. In the second experiment, we used the first subset of 399 images of 133 persons in our multi-illumination dataset, to perform super-resolution and yield three high-resolution outputs under three different illumination conditions (Illum-I, Illum-II and Illum-III) given only one single illumination low-resolution input. Some example results are shown in Figure 4. In both of these two experiments, we used the "leave-one-out" methodology. That is in each of the dataset, those images which were not selected as the testing image were used to construct the model tensors.

The high-resolution reconstruction results shown in Fig.3 and Fig.4 are clearly promising and go beyond what existing methods are capable of in terms of generalizing into significantly different views in super-resolution. Although not perfect, it does not seem to affect the recognition performance using super-resolved identity parameter vector in the high-resolution tensor space. In next section, we show results on recognition experiments using our model.

4.2. Recognition Experiments

Our multi-view dataset has two sets of face images captured at two different occasions. For multi-view face recognition experiment, we used the first set as training dataset and the second as testing dataset. We set up three comparative face recognition experiments, which are our Multi-Model TensorSuperResolution, TensorFace and EigenFace. In the first one using our Multi-Model TensorSuperResolution, we used the yaw -/+45 degrees and tilt -/+45 degrees view high-resolution training face images to build our high-resolution tensor, and used all 5 view low-resolution training images (obtained by downsampling the high-resolution training images) to build the low-resolution tensor. We used the frontal view low-resolution face images in the testing dataset as the testing images. For each of these testing images, we projected it to the low-resolution training tensor to get its low-resolution identity parameter vector $\hat{V}$ as defined in Eq.(7) and computed in Eq.(17), and then per-

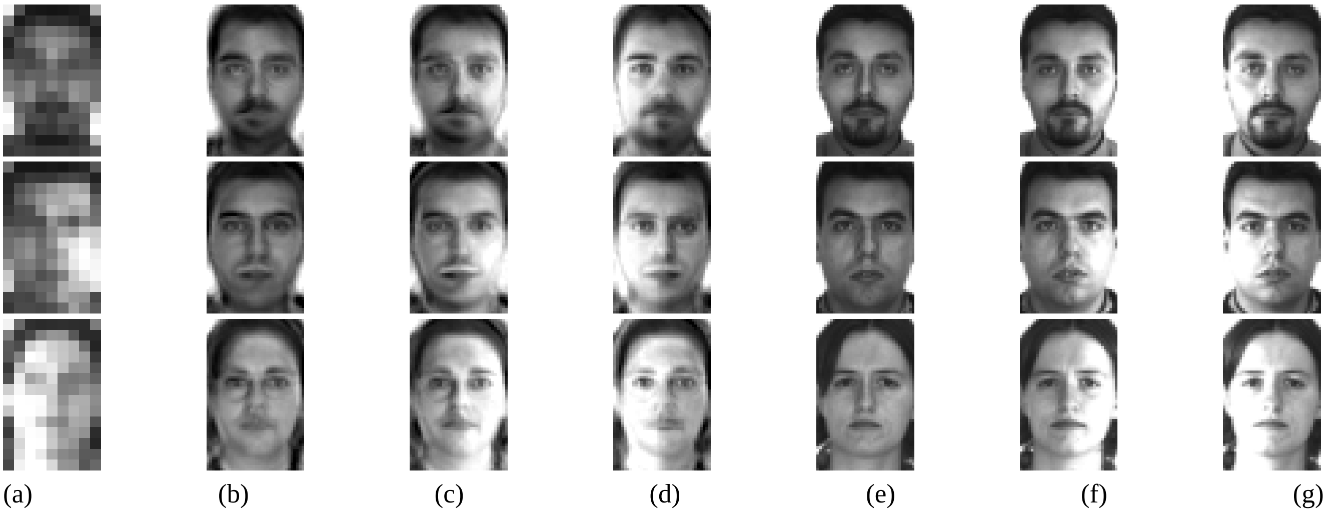

Figure 4: Experiments on super-resolving face images under multiple illumination conditions given a single illumination low-resolution input: (a) are low-resolution input images (14×9) under 3 different illumination conditions (obtained by downsampling original testing input images); (b) - (d) are high-resolution reconstruction results (56×36) at Illum-I, Illum-II and Illum-III repectively; and (e) - (g) are ground truth face images under these 3 illumination conditions.

formed super-resolution using our high-resolution training tensor and the corresponding low-resolution training subtensor obtained by removing frontal view information. After getting the estimated identity parameter vector $\widetilde{V}$ as in Eq.(15), we employed nearest neighbour based recognition by computing its $L2$ norm to every identity parameter vectors V_i in high-resolution training tensor. In the second TensorFace experiment, we also used the yaw -/+45 degrees and tilt -/+45 degrees view high-resolution training face images to build the high-resolution tensor, and used the frontal view low-resolution face images in the testing dataset as the testing images. We bilinearly interpolated testing images to the same size of high-resolution images. We projected these interpolated images onto subtensors of yaw -/+45 degrees and tilt -/+45 degrees to get $V_{v=2,3,4,5}$, the identity parameter vector in training tensor that yields the smallest $L2$ norms among $v = 2$, $v = 3$, $v = 4$ and $v = 5$ identifies the testing frontal image. In the last EigenFace experiment, we performed PCA using all the yaw -/+45 and tilt -/+45 degrees view high-resolution training face images, and used the frontal high-resolution face images in the testing dataset as testing images, recognition can be done in eigenspace. We tabulate the results as below:

Recognition experiments	Recognition rates
Experiment I: Face recognition across views using our Multi-Model TensorSuperResolution	74.6%
Experiment II: Face recognition across views using low-resolution TensorFace	51.4%
Experiment III: Face recognition across views using high-resolution EigenFace	39.7%

Table 1: Face recognition comparison across multiple views.

For face recognition under different illumination conditions, we have two subsets in our multi-illumination datasets, we used the first subset as training dataset and the second one as testing dataset. Similar to the multi-view face recognition, we also performed three experiments for comparison and the results are tabulated as below:

Recognition experiments	Recognition rates
Experiment I: Face recognition under changing illuminations using our Multi-Model TensorSuperResolution	86.2%
Experiment II: Face recognition under changing illuminations using low-resolution TensorFace	66.2%
Experiment III: Face recognition under changing illuminations using high-resolution EigenFace	45.9%

Table 2: Face recognition comparison under changing illumination conditions.

5. Conclusion

In summary, we present a multi-modal face image super-resolution and recognition system in tensor space. By introducing the tensor structure that models multiple factor interactions into a Bayesian framework, we can super-resolve the high-resolution tensor identity parameter vector, given a single modal low-resolution face image. Based on the super-resolved identity parameter vector, we can directly perform face recognition across different views and under changing illumination conditions, we can also reconstruct multiple high-resolution face images of different modalities. Experimental results verify our declaration.

References

[1] M.A.Turk and A.P.Pentland, “Face recognition using eigenfaces”, *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 586-591, 1991.

[2] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, “Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, *European Conf. Computer Vision*, pp. 45-58, 1996.

[3] M. S. Bartlett, J.R.Movellan, and T. J. Sejnowski, “Face recognition by Independent Component Analysis”, *IEEE Trans. on Neural Networks*, Vol.13, No.6, pp. 1450-1464, 2002.

[4] M.A.O. Vasilesescu, D. Terzopoulos, “ Multilinear image analysis for facial recognition”, *Proc. of International Conf. on Pattern Recognition*, 2002.

[5] M. A. O. Vasilescu, D. Terzopoulos, “Multilinear analysis of image ensembles: TensorFaces”, *Proc. 7th European Conference on Computer Vision*, 2002.

[6] T.G.Kolda, “Orthogonal tensor decompositions”, *SIAM Journal on Matrix Analysis and Applications*, Vol.23, pp. 243-255, 2001.

[7] L.D.Lathauwer, B.D.Moor, and J.Vandewalle, “Multilinear Singular Value Tensor Decompositions”, *SIAM Journal on Matrix Analysis and Applications*, Vol.21, No.4, pp.1253-1278, 2000.

[8] M. Elad and A. Feuer, “Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images”, *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1646-1658, Dec. 1997.

[9] M. Irani and S. Peleg, “Improving resolution by image registration”, *CVGIP: Graphical Models and Image Proc.*, vol. 53, pp. 231-239, May 1991.

[10] R. R. Schulz and R. L. Stevenson, “Extraction of high-resolution frames from video sequences”, *IEEE Transactions on Image Processing*, vol. 5, pp. 996-1011, June 1996.

[11] 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 Transactions on Image Processing.*, vol. 6, pp. 1621-1633, Dec. 1997.

[12] J. S. DeBonet and P. A. Viola, “A non-parametric multi-scale statistical model for natural images”, *Advances in Neural Information Processing Systems (NIPS)*, vol. 10, 1998.

[13] S. Baker and T. Kanade, “Limits on super-resolution and how to break them”, *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*,June 2000.

[14] S. Baker and T. Kanade, “Hallucinnating Faces”, *Proc. of IEEE Automatic Face and Gesture Recognition*, pp.83-90, March 2000

[15] W. Freeman and E. Pasztor, “Learning low-level vision”, *7th International Conference on Computer Vision*, pp. 1182-1189,1999.

[16] D. P. Capel and A. Zisserman, “Super-resolution from multiple views using learnt image models”, *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, 2001.

[17] B.K.Gunturk and A.U.Batur, “Eigenface-Domain Super-Resolution for Face Recognition”, *IEEE Tran. on Image Processing*, Vol.12, No.5, pp. 597-606, 2003.

[18] C. Liu, H. Shum and C. Zhang, “A Two-Step Approach to Hallucinating Faces: Global Parametric Model and Local Nonparametric Model”, *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, pp 192-198, 2001.

[19] J. Sun, N. Zhang, H. Tao and H. Shum, “Image Hallucination with Primal Sketch Priors”, *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, 2003.