Multi-Resolution Patch Tensor for Facial Expression Hallucination

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

Abstract

In this paper, we propose a sequential approach to hallucinate/synthesize high-resolution images of multiple facial expressions. We propose an idea of multi-resolution tensor for super-resolution, and decompose facial expression images into small local patches. We build a multi-resolution patch tensor across different facial expressions. By unifying the identity parameters and learning the subspace mappings across different resolutions and expressions, we simplify the facial expression hallucination as a problem of parameter recovery in a patch tensor space. We further add a high-frequency component residue using nonparametric patch learning from high-resolution training data. We integrate the sequential statistical modelling into a Bayesian framework, so that given any low-resolution facial image of a single expression, we are able to synthesize multiple facial expression images in high-resolution. We show promising experimental results from both facial expression database and live video sequences.

1. Introduction

Automatic facial expression analysis requires effective and robust image representation that can capture sufficiently discriminative details about facial muscle changes. Many modelling approaches have been introduced which no matter holistically or locally, rely on regions of interest conveying rich deformation information, and perform feature extraction around salient eyelids, eyebrows, mouth, and their appearance. In particular, Facial Action Coding System (FACS) defines 44 muscular action units to decribe the facial motion and deformation with regard to location and intensity [10, 9], whilst Active Appearance Models (AAM) select and label manually salient local landmark points on facial regions [12, 13]. However, the accuracy and robustness of an expression model suffer dramatically when the resolutions of facial expression image becomes low. This becomes a particular problem when facial expression images are captured at median to long distance away from the camera or when a subject face is not the sole focus of the camera view .

Super-resolution is a technique [15, 17, 19, 18] to generate high-resolution images given a single or set of low-resolution input images. Super-resolution can be performed using either reconstruction-based [6, 7, 8, 11] or learning-based [16, 14, 15, 17, 19, 21] approaches. In particular, Capel and Zisserman [17] divided human face into six unrelated parts and applied PCA on them separately. Combined with MAP estimator, they can recover the result from a high-resolution eigenface space. Baker and Kanade [14] attempted to establish the prior based on a set of training face images pixel by pixel using Gaussian, Laplacian and feature pyramids. Freeman and Pasztor [16] tried to recover the lost high-frequency information from low-level image primitives by representing images using Markov network parameters obtained from a training data set. Liu and Shum [19] combined the PCA model-based approach and Freeman's image primitive technique to form a mixture model of "global face image" carrying common facial properties and "local feature image" recording local individualities. Jia and Gong [5] developed a multi-modal face image super-resolution and recognition system across different views and illuminations. They constructed two training tensors in high- and low-resolution separately, and performed multiple face image super-resolutions by the inferrence of high-resolution tensor identity parameter vectors.

However, none of the existing approaches addressed the problem of super-resolving and generalising facial images undergoing non-linear deformation, such as across different facial expressions. To this end, we propose in this paper a sequential approach to hallucinate/synthesize high-resolution images of multiple facial expressions. We propose an idea of multi-resolution tensor for super-resolution, and decompose facial expression images into small local patches. We build a multi-resolution patch tensor across different facial expressions. By unifying the identity parameters and learning the subspace mappings across different resolutions and expressions, we simplify the facial expression hallucination as a problem of parameter recovery in a

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

patch tensor space. We further add a high-frequency component residue using nonparametric patch learning from high-resolution training data. We integrate a sequential statistical modelling into a Bayesian framework, so that given any low-resolution facial image of a single expression, we are able to synthesize multiple facial expression images in high-resolution. We show promising experimental results from both facial expression databases and live video sequences.

The paper is organized as follows. Section 2 introduces an idea of multi-resolution tensor for super-resolution, and formulates a multi-resolution patch tensor model of multiple facial expressions. In section 3 we derive a Bayesian framework which integrates our sequential process of facial expression hallucination. Section 4 presents experimental results before conclusions are drawn in Section 5.

2. Modelling Facial Expression Images Using Multi-Resolution Patch Tensor

Super-resolution requires a suitable model for generating high-resolution images given a single or set of low-resolution images. Multilinear (tensor) analysis provides an effective approach to model the multiple factor interactions of an image ensemble. Motivated by the combination of these two ideas, in this section we introduce a concept of multi-resolution tensor, and its effective usage when applied on super-resolution. We further extend this idea to facial expression hallucination. Overall, we uniformly decompose the multi-resolution facial expression images into small overlapped patches, and then group these patch pixels of different positions, expressions and resolutions as an emsemble. We construct a multi-resolution patch tensor which later is used for multiple facial expression hallucination. Before presenting how to model facial expression images using multi-resolution patch tensor, let us first briefly introduce some properties of tensor algebra.

2.1. Multilinear Analysis: Tensor SVD

Multilinear analysis [2, 4, 3, 1] is a general extension of the traditional linear methods such as PCA or matrix SVD. Instead of modelling relations within vectors or matrices, multilinear analysis provides a means to investigate the mappings between multiple factor spaces. In the following, we denote scalars by lower-case letters $(a, b, \ldots; \alpha, \beta, \ldots)$, vectors by upper-case $(A, B, \ldots)$, matrices by bold upper-case $(\mathbf{A}, \mathbf{B}, \ldots)$, and tensors by calligraphic letters $(\mathcal{A}, \mathcal{B}, \ldots)$.

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 $\mathcal{A}_{i_1 \ldots i_n \ldots i_N}$ or $a_{i_1 \ldots i_n \ldots i_N}$, 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 $\mathbf{A}_{(n)} \in R^{I_n \times (I_1 I_2 \ldots I_{n-1} I_{n+1} \ldots I_N)}$. 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

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

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 [3, 4]:

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

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

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

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.

2.2. Multi-Resolution Tensor for Super-Resolution

A tensor structure provides a powerful mechanism to incorporate the information and interaction of image ensembles with different resolutions. Benefiting from the mapping relations of multiple factor spaces inherently embedded in the tensor structure, we can recover higher resolution images given any corresponding lower resolution images. More precisely, given a training dataset of high-resolution images, of which they bear some common properties of pixel distributions, as shared by all human faces. We blur and sub-sample these high-resoluiton images with different Gaussian filters and sub-sampling factors, while keeping the image size unchanged. We then obtain a hierarchical ensemble containing images of multiple resolutions. With these training images in place, we construct a tensor structure and use HOSVD to decompose them. The decomposed model can be expressed as

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

where tensor $\mathcal{D}$ groups these training images of multiple resolutions into a tensor structure, and the core tensor $\mathcal{Z}$ governs the interactions between the 3 mode factors. The mode matrix $\mathbf{U}_{idens}$ spans the parameter space of identities for these training images, the mode matrix $\mathbf{U}_{resos}$ spans the parameter space of different resolutions, and the mode matrix $\mathbf{U}_{pixels}$ spanning space of image pixels.

With the constructed tensor of these training images of multiple resolutions, we can perform super-resolution in a tensor parameter vector space. Based on the tensor theories in section 2.1, specifically as suggested in Eq.(3), the image data of different resolutions can be recovered given their single identity parameter vector in tensor space. In such a formulation of super-resolution, this single identity parameter vector can be computed by projecting testing resolution images onto the multi-resolution tensor.

More precisely, suppose we have a basis tensor

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

we can index into this basis tensor at a particular resolution r to yield a basis subtensor

$$\mathcal{B}_r = \mathcal{Z} \times_3 \mathbf{U}_{pixels} \times_2 V_r^T.$$

Then the subtensor containing the individual image data can be expressed as

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

where V^T represents the single identity parameter row vector and $\mathcal{E}_r$ stands for the tensor modelling error for resolution r. 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

$$\mathbf{D}_r^{(1)} = V^T \mathbf{B}_r^{(1)} + e_r. \tag{6}$$

Eq.(6) provides a possible solution for the single identity parameter vector V^T. Applied it on other resolution r', the corresponding resolution image data can be computed as

$$\mathbf{D}_{r'}^{(1)} = V^T \mathbf{B}_{r'}^{(1)} + e_{r'}. \tag{7}$$

We have to emphasize that as noted in Eq.(6) and Eq.(7), the modelling errors e_r and $e_{r'}$ may seriously deteriorate the recovered image quality of different resolutions. To overcome this problem, we build this multi-resolution tensor based on decomposed small patches as follows.

2.3. Modelling Facial Expression Images

Traditional facial expression modelling approaches segment and rely on regions of interest conveying rich deformation information, and perform feature extraction around salient areas. These heuristic segmentation of facial regions restricts the accuracy and stability of facial expression analysis. In this paper, we decompose the facial expression images into small overlapped patches uniformly, and perform analysis and modelling on patch scale without consideration of heuristic locating and segmenting.

After patch decomposition, we apply the above proposed multi-resolution tensor on patch level, which provides a model for capturing the variations of facial expression images. Then a new tensor structure can be given as

$$\mathcal{D} = \mathcal{Z} \times_1 \mathbf{U}_{idens} \times_2 \mathbf{U}_{exps} \times_3 \mathbf{U}_{resos} \times_4 \mathbf{U}_{patches} \times_5 \mathbf{U}_{pixels}, \tag{8}$$

where mode matrix $\mathbf{U}_{exps}$ spans the parameter space of different facial expressions, and the mode matrix $\mathbf{U}_{patches}$ spans the parameter space of overlapped patches. An illustration of this construction process is given in Fig. 1. Similarly as deduced in section 2.2, by unified identity parameter vector for different expressions and patch positions, we can recover the image pixel data on all the decomposed patches, and on each of them for multiple facial expressions.

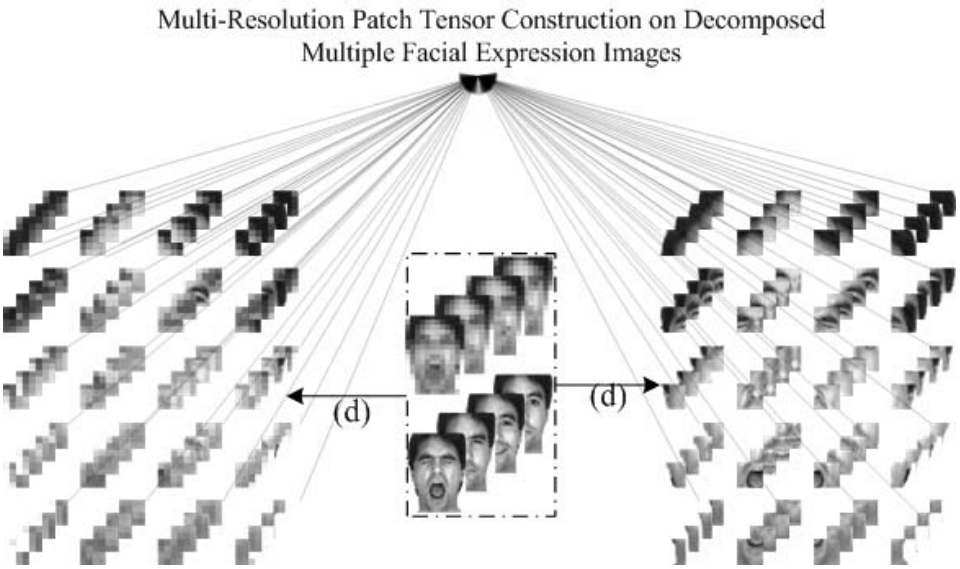

Figure 1. Multiple facial expression images of different resolutions are decomposed into small overlapped patches in (d), on which multi-resolution patch tensor can be constructed.

3. A Sequential Process for Facial Expression Hallucination

The tensor recovery is optimal in minimizing the global mean square error. However, this will lose the highest frequency information of image data. The tensor modelling error in Eq.(7) may also make the recovered image deviating from the original truth. To compensate for this deviation, based on tensor recovered images, we add nonparametric local patch updating by learning from high-resolution training data, and present a sequential process to obtain the middle- and high-frequency information in two steps.

Suppose that $H_1, H_2, ..., H_S$ are the high-resolution images to be recovered for different facial expressions, and L_1 is the low-resolution input of one facial expression. Our probem of multiple facial expression hallucination can be formulated into a Bayesian framework. The task comes as finding the Maximum A Posterior (MAP) estimation of $H_1, H_2, ..., H_S$ given L_1. We take the case of two expression hallucination as an example, which can be formulated as

$$\{H_{1MAP}, H_{2MAP}\} = \arg \max_{H_1,H_2} \log P(H_1, H_2|L_1). \quad (9)$$

By applying the Bayes rule, the probability $P(H_1, H_2|L_1)$ becomes

$$P(H_1, H_2|L_1) = P(H_1|L_1, H_2)P(H_2|L_1). \quad (10)$$

The given low-resolution input L_1 observes the basic imaging observation model, and the estimation of its corresponding high-resolution image H_1 is independent of other expressions, we then rewrite the above as

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

Assuming H^m represents facial expression images containing low- and middle-frequency part information, and H^h containing information of high-frequency part, the high-resolution image is naturally a composition of them,

$$H = H^m + H^h. \quad (12)$$

Since H^m contributes the main part of L after blurring and sub-sampling, and $P(H)$ is equal to $P(H^h|H^m)P(H^m)$, we reformulate the MAP problem of Eq.(9) resulting in

$$\begin{aligned} P(L_1|H_1)P(H_1)P(H_2) \quad = \quad & P(L_1|H_1^m)P(H_1^h|H_1^m) \\ & P(H_2^h|H_2^m)P(H_1^m)P(H_2^m). \end{aligned} \quad (13)$$

Based on Eq.(12) and (13), the MAP inference problem of Eq.(9) can be finally formulated as

$$\begin{aligned} &\{H_{1MAP}, H_{2MAP}\} \\ &\quad = \arg \max_{H_1,H_1^m,H_2,H_2^m} \log \Big(P(L_1|H_1^m)P(H_1^m)P(H_2^m) \\ &\qquad P(H_1|H_1^m)P(H_2|H_2^m) \Big) \end{aligned} \quad (14)$$

The probabilities $P(L_1|H_1^m)P(H_1^m)P(H_2^m)$ and $P(H_1|H_1^m)P(H_2|H_2^m)$ sequentially constrain H_1^m, H_2^m and H_1, H_2 in Eq.(14). This leads to a two-step sequential solution. In the first step, by using a multi-resolution patch tensor, we can recover the H^m for different facial expressions. After obtaining H_1^m, H_2^m, the final high-resolution

H_1, H_2 can be computed in a second step by maximizing $P(H_1|H_1^m)P(H_2|H_2^m)$.

3.1. Middle-Frequency Component Recovery Using Multi-Resolution Patch Tensor

Due to the orthogonality of tensor decompostion [5], the prior constraint $P(H_1^m)P(H_2^m)$ can be ignored in Eq.(14). We decompose the facial expression images into small overlapped patches, and the inference of images H^m containing middle-frenquency part is carried on a patch level. We factorize the likelihood $P(L_1|H_1^m)$ onto patch level and it becomes

$$P(L_1|H_1^m) = \prod_{p=1}^{N} P(L_{1p}|H_{1p}^m)$$

Assuming $\mathbf{A}$ is the blurring and sub-sampling operator connecting L_{1p} and H_{1p}^m in a imaging observation model, we regard these processes as Gaussian therefore

$$P(L_1|H_1^m) = \prod_{p=1}^{N} \frac{1}{Z} \exp\{-\|\mathbf{A}H_{1p}^m - L_{1p}\|^2/\lambda\}, \quad (15)$$

where Z is a normalization constant and λ scales the variance.

Eq.(8) shows our multi-resolution patch tensor structure incorporarting multiple facial expressions, of which suppose we have a basis tensor

$$\mathcal{B} = \mathcal{Z} \times_2 \mathbf{U}_{exps} \times_3 \mathbf{U}_{resos} \times_4 \mathbf{U}_{patches} \times_5 \mathbf{U}_{pixels},$$

we index into this basis tensor at particular expression e, resolution r and patch position p, yielding a basis subtensor

$$\mathcal{B}_{e,r,p} = \mathcal{Z} \times_5 \mathbf{U}_{pixels} \times_2 V_e^T \times_3 V_r^T \times_4 V_p^T.$$

Then the subtensor containing the pixel data for that particular patch can be approximated as $\mathcal{D}_{e,r,p} = \mathcal{B}_{e,r,p} \times_1 V^T$. We unfold it into matrix representation and it becomes $\mathbf{D}_{e,r,p}^{(1)T} = \mathbf{B}_{e,r,p}^{(1)T} \cdot V$. Similarly we can obtain a subtensor for resolution r' of the same facial expression and patch postion, which is $\mathbf{D}_{e,r',p}^{(1)T} = \mathbf{B}_{e,r',p}^{(1)T} \cdot \widetilde{V}$. Suppose $\mathbf{D}_{e,r,p}^{(1)T}$ and $\mathbf{D}_{e,r',p}^{(1)T}$ correspond to the L_{1p} and H_{1p}^m respectively, we replace them in Eq.(15) resulting in

$$P(L_1|H_1^m) = \prod_{p=1}^{N} \frac{1}{Z} \exp\{-\|\mathbf{A}\mathbf{B}_{e,r',p}^{(1)T} \cdot \widetilde{V} - \mathbf{B}_{e,r,p}^{(1)T} \cdot V\|^2/\lambda\}. \quad (16)$$

We optimize the paramter $\widetilde{V}$ based on the constructing properties of our multi-resolution patch tensor, which suggest that the relation between $\mathbf{B}_{e,r',p}^{(1)T}$ and $\mathbf{B}_{e,r,p}^{(1)T}$ observes a basic imaging observation model. In reality, this is consistent with the uniqueness of the identity parameter vector in a tensor space. By setting $\widetilde{V} = V$, we can approximately compute H_{1p}^m as $H_{1p}^m = \mathbf{B}_{e,r',p}^{(1)T} \mathbf{\Psi} L_{1p}$ where $\mathbf{\Psi}$ is the pseudoinverse of $\mathbf{B}_{e,r,p}^{(1)T}$ and is equal to $(\mathbf{B}_{e,r,p}^{(1)} \mathbf{B}_{e,r,p}^{(1)T})^{-1} \mathbf{B}_{e,r,p}^{(1)T}$. By choosing different $e = s$ and $p = n$, we have the general equation for patch recovering of multiple facial expressions at different patch positions, which is formulated as

$$H_{sn}^m = \mathbf{B}_{s,r',n}^{(1)T} \mathbf{\Psi} L_{1n}. \quad (17)$$

After reconstructing all the patches at different positions for multiple facial expressions, the final higher resolution facial expression images are simply composition of their corresponding overlapped small patches.

3.2. High-Frequency Residue Recovery Using Non-parametric Patch Learning

Facial expression images recovered by multi-resolution patch tensor contain the low- and middle-frequency information, we then compensate for the highest frequency part by patch learning from the high-resolution training data. The inference of H_1, H_2 from H_1^m, H_2^m is independent. In the following we take H_1 as example to illustrate how to hallucinate the final high-resolution facial expression images.

We use Markov Random Field (MRF) to model the H_1 to be inferred. By decomposing H_1^m into square patches, we have

$$P(H_1|H_1^m) = P(H_1^m|H_1)P(H_1) = \prod_{q=1}^{M} P(H_{1q}^m|H_{1q})P(H_1).$$

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

$$\prod_{q=1}^{M} P(H_{1q}^m|H_{1q}) \propto -\prod_{q=1}^{M} \|L_{H_{1q}^m} - L_{H_{1q}^{(i)}}\|^2,$$

where $L_{H_{1q}^{(i)}}$ are the Laplacian images from high-resolution training facial expression images. Compare the Laplacian images $L_{H_{1q}^m}$ with $\{L_{H_{1q}^{(i)}}\}_{i=1}^k$ from the training dataset, the patch $H_{1q}^{(i)}$ with $L_{H_{1q}^{(i)}}$ closest to $L_{H_{1q}^m}$ is most probable to be chosen as H_{1q}. Since we model the high-resolution images as MRF, based on the Hammersley-Clifford theorem, $P(H_1)$ is a product $\prod_{H_{1q}, H_{1\bar{q}}} \Phi(H_{1q}, H_{1\bar{q}})$ of compatibility functions $\Phi(H_{1q}, H_{1\bar{q}})$ over all neighboring pairs, where

$H_{1q}, H_{1\bar{q}}$ are the neighboring patch pair, and the compatibility functions are measured using differences of pixel values. Then finally H_1 is estimated as

$$\arg\max_{H_1} \prod_{q=1}^{M} P(H_{1q}^m | H_{1q}) \prod_{(q,\bar{q})} \Phi(H_{1q}, H_{1\bar{q}}). \quad (18)$$

Similiar procedures can be independently repeated for estimations of other high-resolution facial expression images H_s.

Solving probabilistic Eq.(18) is not a trivial task to obtain H_1. We use the Iterated Conditional Modes (ICM) algorithm [20] for this purpose. The pseudo code for our algorithm of multiple facial expression hallucination is as follows.

Algorithm 1: Algorithm for multiple facial expression hallucination

input : single low-resolution image L_1
output: multiple facial expression images H_s, $s = 1, \ldots, S$
Step I:
for *different expression* $e = s$ *and patch position* $p = n$ **do**
  $H_s^m \longleftarrow H_{sn}^m = \mathbf{B}_{s,r',n}^{(1)T} \mathbf{\Psi} L_{1n}$
end
Step II:
repeat
  for any expression s, pick a random patch location q;
  $H_s \longleftarrow \arg\max_{H_s} \prod_{q=1}^{M} P(H_{sq}^m | H_{sq}) \prod_{(q,\bar{q})} \Phi(H_{sq}, H_{s\bar{q}})$
until $H_s converges$;

4. Experiments

We tested our facial expression hallucination approach on both a benchmark facial expression database and live surveillant video sequences. For simulated experiments, we chose the benchmark AR face database. Original AR dataset has 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 testing our approach of multiple facial expression hallucination. To establish a standard training dataset, we used facial image size of 64×48, and aligned them 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.

For all the 504 (126×4) high-resolution facial expression images in the training dataset, we blurred and subsampled them to obtain low-resolution versions of these images. For each experiment, we decomposed every facial expression image into 336 small 4×4 patches which overlapped horizontally and vertically with each other by 1 pixel (the patch size and overlapping size were experimentally decided). We chose the 8 high- and low-resolution facial expression images of 125 individuals to build up a training multi-resolution patch tensor, and one low-resolution expression image of the remaining person as the test input. After obtaining the hallucinated results using the multi-resolution patch tensor (Step I), for resulting images of different facial expressions, we performed the nonparametric patch learning from the corresponding high-resolution expression images of the 125 training individuals (Step II). We experimentally chose the patch size of 6×6 to iteratively update them. Some of the example results are shown as in Fig.2.

Fig.2 shows that any hallucinated result with the same facial expression as the low-resolution input in column (a) is always better than those with other expressions, which is natually an expense of generating nonlinear variations across different facial expressions. Also in Fig.2, comparative investigations of column (g) with column (k), and column(i) with column(m) suggest that, the hallucinated smile and scream images have no identical muscle changes compared to their ground truth expressions. However, the muscle change intensities of these facial expressions have been successfully synthesized.

For testing the robustness of our approach when applied on real data, we captured video sequences from a corridor surveillant camera, and selected frames where low-resolution human faces existed. We cropped out the faces as testing input for hallucinating multiple facial expression images. The high-resolution training facial expression images from the AR database, and the multi-resolution patch tensor constructed from them were used in this process. Fig.3 demonstrates the illustration examples.

In the live sequence experiments, we have no ground truth images which could be collected, to verify the likeness of hallucinated results. But the synthesized multiple high-resolution facial expression images do demonstrate the robustness of our approach. As suggested in Fig.3, the quality of these synthesized images is as good if not better than as those in simulated experiments.

5. Conclusion

In summary, we propose a novel idea of multi-resolution tensor for super-resolution, and present a sequential approach to hallucinate/synthesize high-resolution images of multiple facial expressions. We decompose facial expression images into small patches, and build a multi-resolution patch tensor across different facial expressions. By unifying the identity parameters and learning the subspace mappings across different resolutions and expressions, we simplify the facial expression hallucination as a problem of pa-

(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l) (m)

Figure 2. Examples from simulated experiments on multiple facial expression hallucination: (a) Low-resolution input images (16×12) of different expressions (obtained by downsampling original testing input images). (b)-(e) The first step hallucination results (Eq.(17)) (64×48) with expressions of neutral, smile, anger and scream respectively, using multi-resolution patch tensor. (f)-(i) The second step hallucination results (Eq.(18)) with the 4 expressions, using nonparametric patch learning. (j)-(m) Ground truth facial images of corresponding expresssions.

rameter recovery in a patch tensor space. We further add a high-frequency component residue using nonparametric patch learning from high-resolution training data. We integrate the sequential statistical modelling into a Bayesian framework. Given any low-resolution facial image of single expression, we are able to super-resolve multiple facial expression images in high-resolution. Experiments on both simulated database and live video data verify our method.

References

[1] M.A.O. Vasilesescu, D. Terzopoulos, " Multilinear image analysis for facial recognition", *Proc. of International Conf. on Pattern Recognition*, 2002. 2

[2] M. A. O. Vasilescu, D. Terzopoulos, "Multilinear analysis of image ensembles: TensorFaces", *Proc. 7th European Conference on Computer Vision*, 2002. 2

[3] T.G.Kolda, "Orthogonal tensor decompositions", *SIAM Journal on Matrix Analysis and Applications*, Vol.23, pp. 243-255, 2001. 2

[4] 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. 2

[5] K.Jia and S.Gong, "Multi-modal tensor face for simultaneous super-resolution and recognition", *Proc. IEEE International Conference on Computer Vision*, Oct, 2005. 1, 5

[6] 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. 1

[7] M. Irani and S. Peleg, "Improving resolution by image registration", *CVGIP: Graphical Models and Image Proc.*, vol. 53, pp. 231-239, May 1991. 1

[8] 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. 1

[9] P. Ekman and W.V. Friesen, "Facial Action Coding System (FACS): Manual",*Palo Alto: Consulting Psychologists Press*, 1978. 1

[10] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski, "Classifying Facial Actions", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974-989, Oct. 1999. 1

[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. 1

[12] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active Appearance Models",*Proc. European Conf. Computer Vision*, vol. 2, pp. 484-498, 1998. 1

[13] G.J. Edwards, T.F. Cootes, and C.J. Taylor, "Face Recognition Using Active Appearance Models", *Proc. European Conf. Computer Vision*, vol. 2, pp. 581-695, 1998. 1

[14] 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. 1

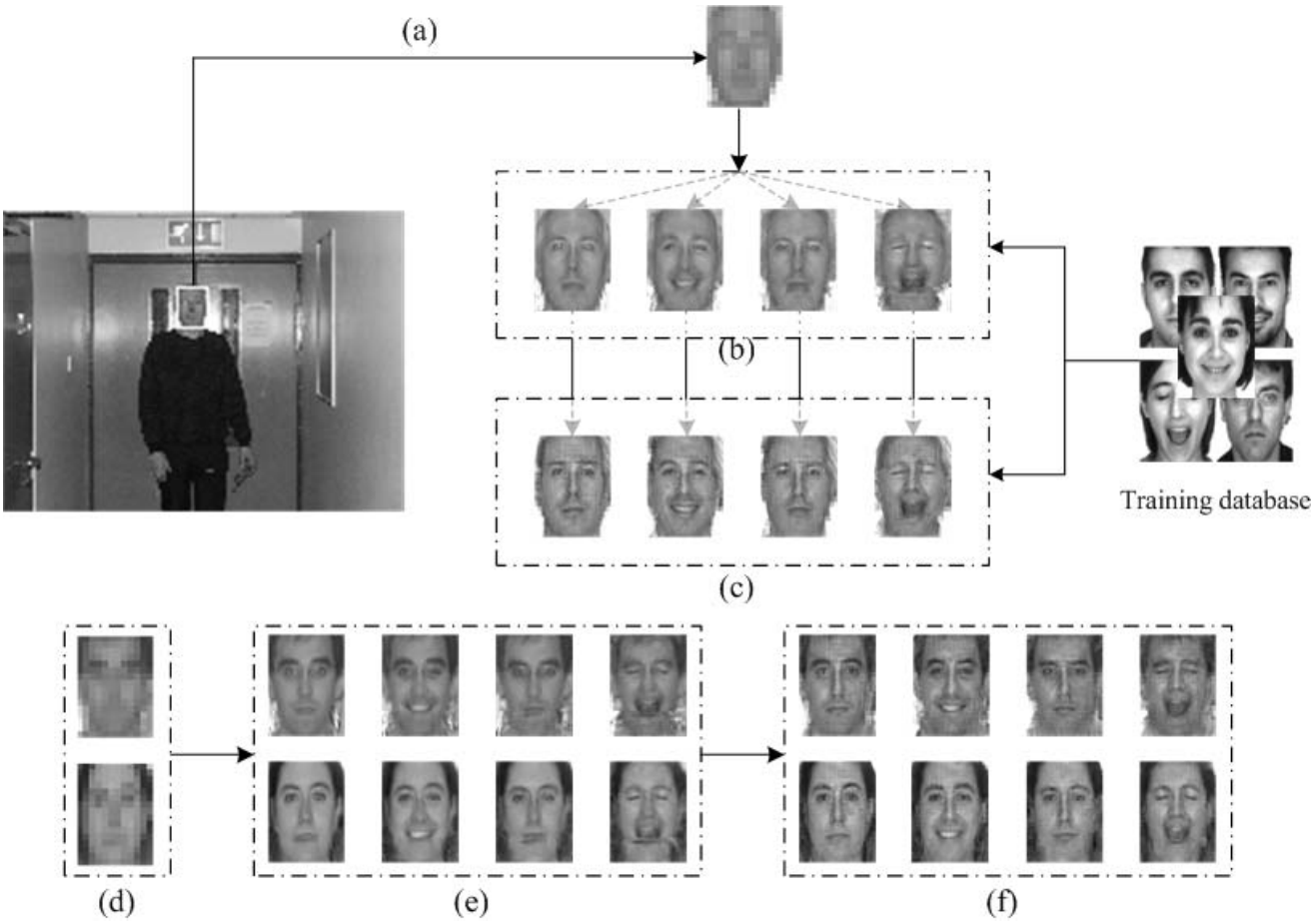

Figure 3. Examples of experiments on multiple facial expression hallucination on live video data: (a) Low resolution facial images in live video were located and segmented first. (b) Hallucinated high-resolution images after the first step (Eq.(17)) using multi-resolution patch tensor. (c) Final high-resolution multiple facial expression images after the second step process of nonparametric patch updating (Eq.(18)). More segmented low-resolution inputs and their two-step hallucinated results are shown in (d)-(f).

[15] S. Baker and T. Kanade, "Hallucinnating Faces", *Proc. of IEEE Automatic Face and Gesture Recognition*, pp.83-90, March 2000 1

[16] W. Freeman and E. Pasztor, "Learning low-level vision", *7th International Conference on Computer Vision*, pp. 1182-1189,1999. 1

[17] 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. 1

[18] 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. 1

[19] 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. 1

[20] J. E. Besag, "On the statistical analysis of dirty pictures (with discussion)", *Journal of the Royal Statistical Society B*, 48(3):259-302,1986. 6

[21] 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. 1