Accepted by *International Journal of Computer Vision*

Face Hallucination: Theory and Practice

Ce Liu* Heung-Yeung Shum† William T. Freeman*

*Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology
celiu@mit.edu, billf@mit.edu

†Microsoft Research Asia
hshum@microsoft.com

Abstract

In this paper, we study face hallucination, or synthesizing a high-resolution face image from an input low-resolution image, with the help of a large collection of other high-resolution face images. Our theoretical contribution is a two-step statistical modeling approach that integrates both a global parametric model and a local nonparametric model. At the first step, we derive a global linear model to learn the relationship between the high-resolution face images and their smoothed and down-sampled lower resolution ones. At the second step, we model the residue between an original high-resolution image and the reconstructed high-resolution image after applying the learned linear model by a patch-based non-parametric Markov network, to capture the high-frequency content. By integrating both global and local models, we can generate photorealistic face images. A practical contribution is a robust warping algorithm to align the low-resolution face images to obtain good hallucination results. The effectiveness of our approach is demonstrated by extensive experiments generating high-quality hallucinated face images from low-resolution input with no manual alignment.

1. Introduction

Many computer vision tasks require inferring a missing high-resolution image from the low-resolution input. Of particular interest is to infer high-resolution (abbr. *high-res*) face images from low-resolution (abbr. *low-res*) ones. This problem was introduced by Baker and Kanade [1] as *face hallucination*. This technique has many applications in image enhancement, image compression and face recognition. It can be especially useful in a surveillance system where the resolution a of face image is normally low in video, but the details of facial features which can be found in a potential high-res image may be crucial for identification and further analysis.

However, hallucinating faces is challenging because people are so familiar with the face. A small error, e.g. an asymmetry of the eyes, might be significant to human perception, whereas for super resolution of generic images the errors in textured regions, e.g. leaves, are often overlooked. This specialized perception of faces requires that a face synthesis system be accurate at representing facial features. A similar problem was encountered with a face cartoon system [7].

We propose that a successful face hallucination algorithm should meet the following three constraints:

(a) **Data constraint.** The result must be close to the input image when smoothed and down-sampled.

(b) **Global constraint.** The result must have common characteristics of a human face, *e.g.* eyes, mouth and nose, symmetry, etc. The facial features should be coherent.

(c) **Local constraint.** The result must have specific characteristics of this face image, with photorealistic local features.

The first requirement can easily be met. For example, it can be simply formulated as a linear constraint on the result. The second and third constraints are more difficult to formulate, but it is important to satisfy all the three requirements in order to hallucinate faces well. Without constraints on specific face features, the result can be too smooth. Without the global face similarity constraint, the result could be noisy or not in agreement with ordinary facial features.

Such global and local constraints motivate us to design a hybrid approach in this paper. We combine a global parametric model which generalizes well for common faces, with a local nonparametric model which learns local textures from example faces. This approach can be applied to modeling visual patterns other than faces, in particular for structured objects with both global coherence such as contour, symmetry, or illumination effects, and precise local textures or patterns, analogous to skin and hair, such as spokes or leaves.

(a) Low-res input (b) Hallucinated by our system (c) Original high-res

Figure 1. Illustration of face hallucination (from Figure 9 (e)). Note that the detailed facial features such as eyes, eyebrows, nose, mouth and teeth of the hallucinated face (b) are different from the ground truth (c), but perceptually we see it as a valid face image. The processing from (a) to (b) is entirely automatic.

We incorporate all the constraints in a statistical face model and find the *maximum a posteriori* (MAP) solution for the hallucinated face. The data constraint is modeled as a Gaussian distribution (a soft constraint), or simply as an equality constraint (a hard constraint). The global constraint assumes a Gaussian distribution learned by principal component analysis (PCA). The local constraint utilizes a patch-based nonparametric Markov network to learn the statistical relationship between the global face image and the local features. A two-step approach is then used in hallucinating faces. First, an optimal global face image is pursued in the eigen-space when constraints (a) and (b) are satisfied. Second, an optimal local feature image is inferred from the optimal global image by minimizing the energy of the Markov network with constraint (c) applied. The sum of the global and local image forms the final result. An example of hallucinated image from an input low-resolution image is shown in Figure 1. Although the facial feature details of the hallucinated face are different from those in the original, we may perceive it as a valid human face taken by a camera.

At a practical matter, the other challenge in face hallucination is the difficulty of aligning faces at low-res images. Many learning-based image synthesis models require alignment between the test sample and the training examples, e.g. [7]. Even a small amount of misalignment can dramatically degrade the synthesized result. However, the facial features may contain very few pixels; in real images the faces are normally not upright; the scale and position must be estimated at sub-pixel level. Therefore, alignment at low-res requires that very accurate measurements be made from very little data.

To address the alignment challenge, we design a face alignment algorithm to align faces at low-res. The alignment algorithm finds an affine transform to warp the input image to a tem-

plate to maximize the probability of low-res face image, determined from an eigenspace representation. To make that alignment step robust, multiple candidate starting points are explored through a stochastic algorithm from which the best alignment result is selected automatically. We demonstrate through many examples that our system is able to find and hallucinate high-res face image with vivid details from low-res pictures, without manual intervention.

Our work is built upon Takeo Kanade's pioneering work on image registration [26] and face hallucination [1]. We acknowledge his contributions to computer vision that inspired our work in this paper.

This paper is organized as follows. After reviewing related work in Section 2, we introduce the details of our global and local face modeling in Section 3. Many examples of a toy experiment where the low-res input is well registered at high-res are shown in Section 4. Face alignment on the low-res image is introduced in Section 5, and the hallucination results using the aligned low-res images are shown in Section 6. Other applications such as random face synthesis are also explored in Section 6. Discussion is given in Section 7 and Section 8 concludes the paper.

2. Related Work

Finding a high resolution image, given a low-resolution input, is an under-constrained problem: many images can yield the input after being smoothed and down-sampled. We find it natural to divide super resolution work into two categories, based on which additional constraints are used to infer the high-res image.

(a) The constraints come from a direct, temporal correspondence over multiple frames. The hallucinated high-res information should be consistent from frame to frame.

(b) The constraints come from an indirect, spatial correspondence with other samples. This can be described in a parametric form, or else by examples, learning the statistical correlation between low-res and high-res information from a database of training images.

Obviously these two categories may be applied at very different scenarios. But both of them can address face hallucination problem. They may also be combined as in [10].

2.1 Direct, Temporal Correspondence

We may have multiple (noisy) observations for a still or temporally slowly changing scene. Through motion analysis, the observations are registered from frame to frame and a high resolution image can be inferred from matching each frame. Early work on video-based super

resolution includes [19, 28]. In [4] motion blur was taken into account. Multiple sequences are used in [35] for both spatial and temporal super resolution. Utilizing a video sequence can significantly enhance the resolution of an image, and should be exploited where possible. In this paper, however, we focus on the application scenario where only a single input image is provided.

2.2 Indirect Spatial Correspondence

Where can we find the high-frequency information for a single input low-res image? It can be obtained from either parametric or nonparametric methods. Parametric approaches to super-resolution, also known as image interpolation, have had some success [18, 34, 29, 16]. However, using parametric methods it is often difficult to interpolate details well within texture and corner-like local regions of intensities. In estimating such details, example-based approaches often perform better.

The general idea of example-based approach is to collect a database, learn the statistical correlation between the low-res and high-res, and apply it to the input image. The problem under this category can be *generic*, where the input can be any image, or *object-specific*, where we assume that only images of a certain object category are input, such as face.

2.2.1 Generic image super-resolution

Most learning-based super-resolution algorithms such as [15, 17, 14, 13] assume homogeneous Markov random field (MRF) for images. Let $\mathcal{L}$ denote an image lattice, and $\mathbf{v}$ a certain position on the lattice with $I_{\mathbf{v}}$ as the pixel value. $I_{\mathbf{v}}^{-}$ represents all pixels on $\mathcal{L}$ other than $I_{\mathbf{v}}$. I is a Markov random field if

$$p(I_{\mathbf{v}}|I_{\mathbf{v}}^{-}) = p(I_{\mathbf{v}}|N_{\mathbf{v}}), \tag{1}$$

where $N_{\mathbf{v}}$ is the neighborhood of $\mathbf{v}$. This definition indicates what a pixel is only relies on the pixels in its neighborhood. Further I is a *homogeneous* MRF if the conditional density function is independent of the position $\mathbf{v}$.

Although originally proposed for texture synthesis, the multi-resolution nonparametric sampling method developed by De Bonet [5] infers the high-frequency texture features from the low-frequency features named *parent structure*. His texture synthesis results indicate that in homogeneous MRF, the high-frequency component locally depends on the low-frequency part. Freeman *et al.* [15] proposed a parametric Markov network to learn the statistics between the "scene" and "image", as a framework to handle various low-level vision tasks, including super-resolution. In their work the conditional density function of each image patch given its scene

patch is also homogeneous. If the scene is the high-frequency part and image the low-frequency input, Markov network can be applied in super resolution work. They elaborated this application in [14]. Hertzmann *et al.* [17] and Efros & Freeman [12] generalize local feature transform methods. When given a pair of training images, an analogous image is inferred from the input by the local similarity between the training pair. "Image analogies" [17] can fulfill super resolution work if the training pairs are low-resolution and high-resolution images respectively.

All of above methods do local feature transfer/inference on low-level vision. People also tried to approach generic image super resolution at a higher level. In [37], a primal sketch is estimated from the low-res image to guide finding the edge in high-res image. In [38] a graphical model based on multiple local regressors is proposed to make the inference problem tractable.

2.2.2 Face hallucination

The generic super resolution algorithms perform well in hallucinating images provided (a) the training image details generalize to the test image, and (b) the synthesized image details are primarily textures, not semantically important structures. They often fail in hallucinating structural visual patterns which break the homogeneous assumptions, such as the human face. To specialize to face hallucination, the homogeneous MRF assumption has to be abandoned, leading to the work by Baker and Kanade[1]. They only follow that the size of each pixel's neighborhood is equal. The statistics between the low-res and high-res images at each position is learnt in a nonparametric way by a number of training pairs. Similar to [5], the features on high-frequency image are inferred from the *parent structure* by nearest neighbor searching. The final gray level image is then obtained by gradient descent to fit the constraints by the inferred local features. They also discuss the limits on super resolution and how to break them in their method [2]. The images hallucinated by [1] appear to be noisy at places. In their model, the global constraint is not incorporated. The global properties of face, such as explicit contour, coherent illumination and symmetry are somewhat missing.

It is interesting to note that all previous models use local feature inference in MRF without global correspondence being taken into account. Such global modeling is, however, essential to pursue good performance in face hallucination. Principal component analysis (PCA) can be used to model the global variance of facial appearance in an eigen-space: it has been successfully used for face recognition [40] and generative face modeling by ASM and AAM [8]. Encouraged by recent success of patch-based nonparametric sampling for texture synthesis [22, 12], we built a non-parametric patch-based Markov network as in [14] to model the statistics between the local feature image and the global face image in eigen-space.

The early version of this global and local modeling appeared in [24]. A subspace-based super resolution approach similar to our global face model was proposed at the same time in [6]. A number of papers on face hallucination appeared subsequently. In [10] both temporal correspondence and a prior model are used to hallucinate faces. In [42], a mask is designed to do face hallucination on the inner part of the face only to avoid artifacts on hair and background, though these artifacts can be properly handled by local modeling and appropriate smoothing in this paper. In [21, 20] the task was generalized to handle different poses.

In this paper we focus on facial appearance modeling as in [24], but address additional practical considerations. We elaborate more on the global face modeling, in particular on hard constraint and show that it may generate results more faithful to the low-res. Importantly, we also show how to apply face hallucination to unregistered images, resulting in high quality high-res face images synthesized from low-res face input, using no manual intervention.

3. Theory and Algorithms

3.1 A Bayesian Formulation to Face Hallucination

Let I_H and I_L denote the high-resolution and low-resolution face images respectively. If I_L is reduced from I_H by a factor of s, following [1], we compute I_L by

$$I_L(m,n) = \frac{1}{s^2}\sum_{i=0}^{s-1}\sum_{j=0}^{s-1} I_H(sm+i, sn+j) \tag{2}$$

where s is always an integer. We take $s = 4$ unless otherwise specified. Eq. (2) combines a smoothing step and a down-sampling step, more consistent with image formation as integration over the pixel [1]. To simplify the notation, if I_H and I_L are N-D and M-D long vectors respectively ($M = N/s^2$), Eq. (2) can be rewritten as

$$I_L = \mathbf{A} I_H \tag{3}$$

where $\mathbf{A} = [a_1, a_2, \cdots, a_M]^T$ is a $M \times N$ matrix. Each row vector a_i^T in $\mathbf{A}$ smooths a $s \times s$ block in I_H to one pixel in I_L.

To compute I_H from I_L is straightforward in (3), but the inverse process is full of uncertainty. It is clear that many I_H satisfy the constraint of Eq. (3). Thus we should find the optimal one to maximize the posterior probability $p(I_H|I_L)$, based on the *maximum a posteriori* (MAP) criterion. Bayes' rule for this estimation problem is:

$$p(I_H|I_L) = \frac{p(I_L|I_H)p(I_H)}{p(I_L)}. \tag{4}$$

Since $p(I_L)$ is the evidence remaining constant, MAP actually maximizes the product of the likelihood $p(I_L|I_H)$ and prior $p(I_H)$. The MAP estimate of optimal solution is under the prior $p(I_H)$

$$I_H^* = \arg\max_{I_H} p(I_L|I_H)p(I_H). \tag{5}$$

3.2 Global and Local Face Modeling

Note that equation (5) contains the prior distribution of a face image $p(I_H)$. Looking for a sophisticated face prior model has been a long term research goal in computer vision. Current face prior models either capture the common features of faces in a parametric way, for example through eigenfaces [40] and AAM [8], or represent the individual characteristics such as local features [1] in a nonparametric way. But both the common features and the individual characteristics of faces are required in face hallucination. Therefore we develop a mixture model combining a global parametric model called the *global face image* I_H^g which carries the common features of face, and a local nonparametric one called the *local feature image* I_H^l which records the local individualities. The full-resolution face image is their sum,

$$I_H = I_H^l + I_H^g. \tag{6}$$

Since I_L is the low-frequency part of I_H, the global face I_H^g contributes the main part of AI_H and the local features I_H^l lie on the high-frequency band. Mathematically,

$$\mathbf{A}I_H^g = \mathbf{A}I_H, \quad \mathbf{A}I_H^l = 0. \tag{7}$$

To ensure that $\mathbf{A}I_H^l = 0$, I_H^l can be defined in terms of wavelets, but we find it unnecessary to do that in practice. We decompose the prior model of the face as

$$p(I_H) = p(I_H^l, I_H^g) = p(I_H^l|I_H^g)p(I_H^g). \tag{8}$$

Now we shall reformulate the MAP problem (5) under this mixture model for faces. The likelihood $p(I_L|I_H)$ can be simply regarded as a soft constraint on I_H. If each pixel on I_L is identically treated, the distribution has a Gaussian form

$$p(I_L|I_H) = \frac{1}{Z}\exp\{-\frac{1}{2\sigma^2}(\mathbf{A}I_H - I_L)^T(\mathbf{A}I_H - I_L)\}, \tag{9}$$

where Z is a normalization constant and σ^2 evaluates the variance of the assumed additive Gaussian noise. Using Eq. (7), Eq. (9) can be rewritten as

$$\begin{aligned} p(I_L|I_H) =& \frac{1}{Z}\exp\{-\frac{1}{2\sigma^2}(\mathbf{A}I_H^g - I_L)^T(\mathbf{A}I_H^g - I_L)\} \\ =& p(I_L|I_H^g) \end{aligned} \tag{10}$$

In the limit of no observation noise the likelihood function can alternatively be formulated as a delta function

$$p(I_L|I_H) = \delta(I_L - \mathbf{A}I_H) = \delta(I_L - \mathbf{A}I_H^g) \tag{11}$$

From Equations (8), (10) and (11), the MAP inference problem, Eq. (5), can be rewritten as

$$I_H^* = \arg\max_{I_H^g, I_H^l} p(I_L|I_H^g)p(I_H^g)p(I_H^l|I_H^g) \tag{12}$$

It is clear that $p(I_L|I_H^g)p(I_H^g)$ and $p(I_H^l|I_H^g)$ constrain I_H^g and I_H^l respectively. The optimization strategy is naturally divided into two steps. At the first step we find the optimal global face I_H^{g*} by maximizing $p(I_L|I_H^g)p(I_H^g)$. At the second stage the optimal local feature image I_H^{l*} is computed by maximizing $p(I_H^l|I_H^{g*})$. Finally $I_H^* = I_H^{g*} + I_H^{l*}$ is our desired result.

3.3 Global modeling: a linear parametric model

We apply PCA to modeling the global face image I_H^g. Given a set of training face images $\{I_H^{(i)}\}_{i=1}^k$, we can compute the eigenvectors $\{b_i\}_{i=1}^r (b_i \in \mathbb{R}^N, i = 1, \cdots, r)$, eigenvalues $\{\sigma_i^2\}_{i=1}^r$ and mean face μ by standard singular value decomposition (SVD) [32]. The orthogonal eigenvectors construct the eigen-subspace $\Omega = \mathrm{span}(b_1, \cdots, b_r) \sim \mathbb{R}^r$. Thus I_H^g is in fact the reconstructed image of I_H in Ω

$$I_H^g = \mathbf{B}X + \mu, \;\; X = \mathbf{B}^T(I_H - \mu), \tag{13}$$

where $\mathbf{B} = [b_1, \cdots, b_r]_{N\times l}$, and $X = (x_1, \cdots, x_r)^T$ is a vector in Ω. Intuitively, I_H^g is linearly controlled by the coefficients x_i with corresponding eigenvectors $\mathbf{b}_i$. Since the eigenvectors are analyzed from the training data, they represent the irrelevant common properties of the face, such as lighting, scale and pose etc. Thus I_H^g retains the common features of I_H with individual characteristics lost.

Since the random variable I_H^g is determined by X in (13), its distribution can be replaced by X. Maximizing $p(I_L|I_H^g)p(I_H^g)$ in (12) is equivalent to maximizing $p(I_L|X)p(X)$. We approximate the prior $p(X)$ by a simple Gaussian:

$$p(X) = \frac{1}{Z'} \exp\{-\frac{1}{2} X^T \mathbf{\Lambda}^{-1} X\}, \tag{14}$$

where $\mathbf{\Lambda} = \text{diag}(\sigma_1^2, \cdots, \sigma_l^2)$ and Z' is a normalization constant. For the likelihood $p(I_L|X)$ we have two choices, corresponding to a choice of hard or soft constraints.

3.3.1 Hard constraint

When $r > N$, i.e. the number of eigenvectors is greater than the dimension of the I_L, then X is under constrained, or there is enough freedom to precisely formulate the constraint. The hard constraint for the eigenspace representation of the high-res image rendering to the observed low-res image is

$$\mathbf{A}(\mathbf{B}X + \mu) = I_L \tag{15}$$

Let $\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{N \times r}$. Note that it is not necessary to explicitly write down $\mathbf{A}$, a huge sparse matrix in order to compute $\mathbf{C}$. The i, jth entry $\mathbf{C}_{ij}$ is the average of the ith block (in scan line order) of eigenvector b_j. In other words, each column vector of $\mathbf{C}$ is a smoothed and downsampled eigenvector. The above equation can be rewritten as

$$\mathbf{C}X = I_L - \mathbf{A}\mu \tag{16}$$

Let QR decomposition [36] of $\mathbf{C}^T$ be

$$\mathbf{C}^T = [\mathbf{Q}_1 \; \mathbf{Q}_2] \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{bmatrix} \tag{17}$$

where $[\mathbf{Q}_1 \; \mathbf{Q}_2] \in \mathbb{R}^{r \times r}$ is a unitary matrix, forming a set of bases in the space of X. The span of the column vectors of $\mathbf{Q}_2$ forms the null space of $\mathbf{C}$. Let

$$X = [\mathbf{Q}_1 \; \mathbf{Q}_2] \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \mathbf{Q}_1 u_1 + \mathbf{Q}_2 u_2, \tag{18}$$

and we have

$$\mathbf{C}X = [\mathbf{R}_1^T \; \mathbf{0}] \begin{bmatrix} \mathbf{Q}_1^T \\ \mathbf{Q}_2^T \end{bmatrix} [\mathbf{Q}_1 \; \mathbf{Q}_2] \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \mathbf{R}_1^T u_1 = I_L - \mathbf{A}\mu \tag{19}$$

In this model $\mathbf{R}_1$ is an invertable square matrix. So we have

$$u_1 = (\mathbf{R}_1^T)^{-1}(I_L - \mathbf{A}\mu) \tag{20}$$

Now we combine Eq. (18) to (20) to maximize $p(I_L|X)p(X)$, or equivalently minimize the following function:

$$\begin{aligned} F(X) &= X^T \mathbf{\Lambda}^{-1} X \\ &= (u_1^T \mathbf{Q}_1^T + u_2^T \mathbf{Q}_2^T)\mathbf{\Lambda}^{-1}(\mathbf{Q}_1 u_1 + \mathbf{Q}_2 u_2) \\ &= u_1^T \mathbf{Q}_1^T \mathbf{\Lambda}^{-1} \mathbf{Q}_1 u_1 + 2u_2^T \mathbf{Q}_2^T \mathbf{\Lambda}^{-1} \mathbf{Q}_1 u_1 + u_2^T \mathbf{Q}_2^T \mathbf{\Lambda}^{-1} \mathbf{Q}_2 u_2 \end{aligned} \tag{21}$$

As u_1 is determined by the low-res image I_L through Eq.(20), the optimal u_2 is given as

$$u_2^* = -(\mathbf{Q}_2^T \mathbf{\Lambda}^{-1} \mathbf{Q}_2)^{-1} \mathbf{Q}_2^T \mathbf{\Lambda}^{-1} \mathbf{Q}_1 u_1 \tag{22}$$

Combining Eq.(18), (20) and (22), we obtain

$$X^* = (\mathbf{I} - \mathbf{Q}_2(\mathbf{Q}_2^T \mathbf{\Lambda}^{-1} \mathbf{Q}_2)^{-1} \mathbf{Q}_2^T \mathbf{\Lambda}^{-1}) \mathbf{Q}_1 (\mathbf{R}_1^T)^{-1}(I_L - \mathbf{A}\mu) \tag{23}$$

In practice, we first solve u_1 based on Eq. (20) and u_2^* based on Eq. (22), and then combine them to find X based on Eq. (18). In this way we can avoid computing an inverse matrix. All the matrices are computed off-line in the training step.

3.3.2 Soft constraint

When $r < N$, i.e. the number of eigenvectors is smaller than the dimension of I_L, X would be over-constrained in Eq. (15). We shall formulate the likelihood as a soft constraint. The likelihood, Eq. (10), becomes

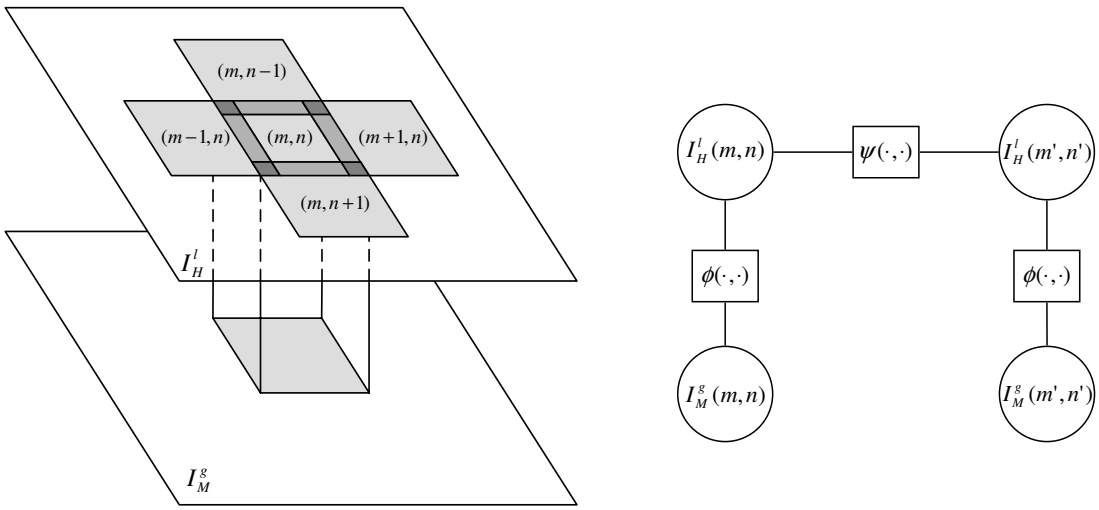
$$p(I_L|X) = \frac{1}{Z} \exp\{-\frac{1}{\sigma^2}[\mathbf{A}(\mathbf{B}X + \mu) - I_L]^T [\mathbf{A}(\mathbf{B}X + \mu) - I_L]\}. \tag{24}$$

The optimal X^* maximizing $p(I_L|X)p(X)$ is

$$X^* = \arg\min_X \sigma^2 X^T \mathbf{\Lambda}^{-1} X + [\mathbf{A}(\mathbf{B}X + \mu) - I_L]^T [\mathbf{A}(\mathbf{B}X + \mu) - I_L]. \tag{25}$$

Since the objective function is a quadratic form, the solution is straightforward:

$$\begin{aligned} X^* &= (\mathbf{B}^T \mathbf{A}^T \mathbf{A} \mathbf{B} + \sigma^2 \mathbf{\Lambda}^{-1})^{-1} \mathbf{B}^T \mathbf{A}^T (I_L - \mathbf{A}\mu) \\ &= (\mathbf{C}^T \mathbf{C} + \sigma^2 \mathbf{\Lambda}^{-1})^{-1} \mathbf{C}^T (I_L - \mathbf{A}\mu) \end{aligned} \tag{26}$$



(a) Markov network (b) Factor graph

Figure 2. (a) Illustration of the patch-based nonparametric Markov network. The compatibility function $\psi(\cdot)$ is defined on the similarity of the two neighboring patches on the overlapping area. (b) The corresponding factor graph.

To ensure numerical stability, the inverse of $\mathbf{B}^T\mathbf{A}^T\mathbf{A}\mathbf{B} + \sigma^2\mathbf{\Lambda}^{-1}$ is computed by SVD. The optimal global face image $I_H^{g*} = \mathbf{B}X^* + \mu$. Since matrix $\mathbf{B}$, $\mathbf{\Lambda}$ and μ are learnt by PCA, and $\mathbf{A}$ is constant as a smoothing and down-sampling function, all matrices on the right side of (26) can be computed offline. Furthermore, we want to allow the "softness" parameter σ^2 to be as small as possible. When $\sigma \to 0$, Eq.(26) becomes

$$X^* = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T(I_L - \mathbf{A}\mu) \tag{27}$$

Will the soft constraint approach the hard constraint when $\sigma \to 0$? When $r < N$, it is impossible to apply hard constraint. When $r > N$, the inverse of $\mathbf{C}^T\mathbf{C}$ does not exist. Therefore, the soft constraint is not equivalent to the hard constraint in any circumstances.

Once given I_L as input, I_H^{g*} can be computed very quickly through solving linear systems. I_H^{g*} is a smoothed version of a human face, which will be improved by the local model in next subsection.

3.4 Local modeling: patch-based nonparametric Markov Network

In most cases when PCA is used, the random variable is regarded as a composition of two parts: the principal components and an unmodeled residue which is always assumed independent of the former. But in our mixture modeling, the residue $I_H^l = I_H - I_H^g$ is the highest

frequency component, dependent on the lower frequency part [15], *i.e.*, I_H^g. That independence assumption fails in our model. To carefully model $p(I_H^l|I_H^g)$, we use patch-based nonparametric Markov network [15, 14] and do inference using max-product belief propagation [11]. Such a patch-based nonparametric approach has been used in texture synthesis [22, 12] as well. An early version of this Markov network optimized by simulated annealing is in [24].

Following [15], we assume the high-frequency band to be conditionally independent of the low-frequency band given the middle-frequency band. Mathematically

$$p(I_H^l|I_H^g) = p(I_H^l|I_M^g) \tag{28}$$

where

$$I_M^g = I_H^g - f * I_H^g. \tag{29}$$

f is a Gaussian filter.

The likelihood function in Eq. (28) can be written as

$$p(I_H^l|I_M^g) = \frac{1}{Z} \prod_{(m,n)} \phi(I_H^l(m,n), I_M^g(m,n)) \prod_{(m,n),(m',n')\in\varepsilon} \psi(I_H^l(m,n), I_H^l(m',n')) \tag{30}$$

where $I_H^l(m,n)$ and $I_M^g(m,n)$ denote the patches centered at $(ms+s/2,\ ns+s/2)$ with patch size $s+2$. We choose $s=4$, though other choices give similar results. ε denotes the set of neighbors. We choose a 4-neighbor system. Neighboring patches overlap at a 2 pixel width strip where compatibility function ψ is computed.

Following [15], the compatibility functions $\phi(\cdot,\cdot)$ are computed in a nonparametric manner. From the collected face database (see Section 4 for the details) we have training pairs $\{(Y_H^{l(i)}, Y_M^{g(i)})\}_{i=1}^K$ where Y has the same dimension as I. For an input patch $I_M^g(m,n)$ we may obtain a set of training samples that match this patch within a small tolerance ϵ

$$\Omega(m,n) = \{Y_H^{l(i)}(m,n) \mid \|Y_M^{g(i)} - I_M^g(m,n)\| < \epsilon\ i = 1, \cdots, K\}. \tag{31}$$

We rewrite the set as $\Omega(m,n) = \{Y_H^{*l(j)}(m,n)\}_{j=1}^{r_{mn}}$, where r_{mn} is the number of elements. Intuitively, this set contains all the local face patches at the location (m,n) whose corresponding global face patches (middle frequency component) match the given global face patch. The compatibility function is defined as

$$\phi(I_H^l(m,n) = Y_H^{*l(j)}(m,n), I_M^g(m,n)) = \exp\{-\frac{1}{2\sigma_\phi^2}\|Y_M^{g(i)} - I_M^g(m,n)\|^2\} \tag{32}$$

This function is indeed defined on a discrete set with r_{mn} states. This method is *nonparametric* because $\phi(\cdot)$ is defined on observed samples.

Compatibility function ψ is simply defined such that neighboring patches agree with each other on the overlapping area. Without losing generalization, the ψ for two horizontally neighboring patches are defined as

$$\psi(I_H^l(m,n), I_H^l(m+1,n)) = \exp\{-\frac{1}{2\sigma_\psi^2}\|\mathcal{R}_H^l(m,n) - \mathcal{L}_H^l(m+1,n)\|^2\} \tag{33}$$

where $\mathcal{R}$ and $\mathcal{L}$ denote the right most 2 columns and left most 2 columns of a patch, respectively. Function ψ for vertically neighboring patches are defined similarly.

Once the Markov network is set, we use max-product belief propagation to minimize the energy. Please refer to [15, 11] for the details of belief propagation.

3.5 Post-Processing

When PCA is applied to reconstruct an image we may see a ghosting effect, similar to the Gibbs effect when a signal is reconstructed by Fourier bases [30]. Inevitably this artifact is propagated to the final reconstructed high-res face images through the Markov network. This is partially caused by the misalignment of face images. In the training database we try to align facial feature points, but other image features, such as hair strings and clothes are not necessarily well aligned. To avoid this problem some other researchers tried to only do face hallucination in the interior region of face [42, 9]. However, we found that this artifact can be easily removed by a post-processing step.

The artifacts as shown in Figure 7 row (c) can be regarded as noise, which can be removed by bilateral filtering [39] by appropriately setting the spatial and intensity variance. But the artifacts or noise mainly distribute around the image boundary. Inspired by the adaptive bilateral filtering work [23], we design the parameters of the bilateral filter to be dependent on image coordinates. The rule of thumb is to smooth less in the center but more around the boundary.

We could have encoded local image statistics in the face modeling, e.g. modeling the marginal of the band-pass filtering responses [45]. Leaving this as our future work, we find that current modeling is sufficient to generate good results.

4. Experiment on a Simplified Scenario

In this section, we study the effectiveness of our face hallucination algorithm by assuming that the face images are well aligned in both training and test as in [1, 24]. The practical

issue of face alignment in low-res images will be discussed in the next section, resulting in a fully automatic algorithm. Only for investigation in this current section do we use manual intervention to register the low-res images.

In our experiments, the high-res faces are collected from public face databases such as AR [27] and FERET [31], and MSRA Cartoon face database [7]. There are a total of 4,476 samples, including Caucasian, Asian and Black, both male and female adults, frontal faces. The lighting of the images is mostly from an indoor environment. We use face detection [43] and alignment [44] algorithms to register face images. We choose the 87 feature point system as proposed in [7], and allow the user to modify any misalignments.

After registration we compute the mean shape of facial feature points, and warp each face image to the mean shape by affine transform. This affine transform is estimated to minimize the sum of matching errors. Even though an affine transform may distort the face if the pose is not strictly upright or the facial shape is different from the mean, in the real application we shall use affine transformation to extract low-res face images. After affine warping, the facial features are almost registered, eye to eye and mouth to mouth, but not exactly (for exact registration more sophisticated warping techniques are needed), and we do not need exact registration. From the total 4,476 high-res samples after warping and cropping we extract 46 images and downsample them for testing, using the remaining 4430 for training.

The mean face and the top ten eigen-faces corresponding to the ten largest eigenvalues computed by SVD are displayed in Figure 3. To better visualize the eigenfaces in Figure 3(a), we multiply each eigenface by $\pm 3\sigma$, add to the mean face (d) and display the results in (b) and (c), respectively. Clearly the facial properties such as lighting, pose, race and gender are modified by the different eigenfaces [40]. For instance, side lighting is controlled by the 5th eigenface, race appears to be modified by the 1st, 2nd and 3rd eigenfaces, gender is affected by many, e.g. 1st to 4th, 6th and 10th, pose is changed by 8th, and background lighting is controlled by 1st, 2nd and 6th. Interestingly, each eigenface normally changes a mixture of facial properties, e.g. the 1st and 2nd eigenfaces appear to change gender, race and lighting simultaneously. The ability of the eigenfaces to model these various facial properties makes them a useful model for the global face.

The results of reconstructing the global face from a low-res input using the soft constraint are shown in Figure 5. We have chosen 8 typical samples out of 40 for illustration. The number of eigenvectors r varies from 20, 100, 500 to $1,000$, and the corresponding results are shown from (b) to (e). Not surprisingly, the fewer eigenvectors, the smoother and closer to the mean face the reconstruction is. The reconstruction with insufficient eigenvectors lacks the individual facial features such as the correct lighting effects. The change of the reconstruction from $r = 500$

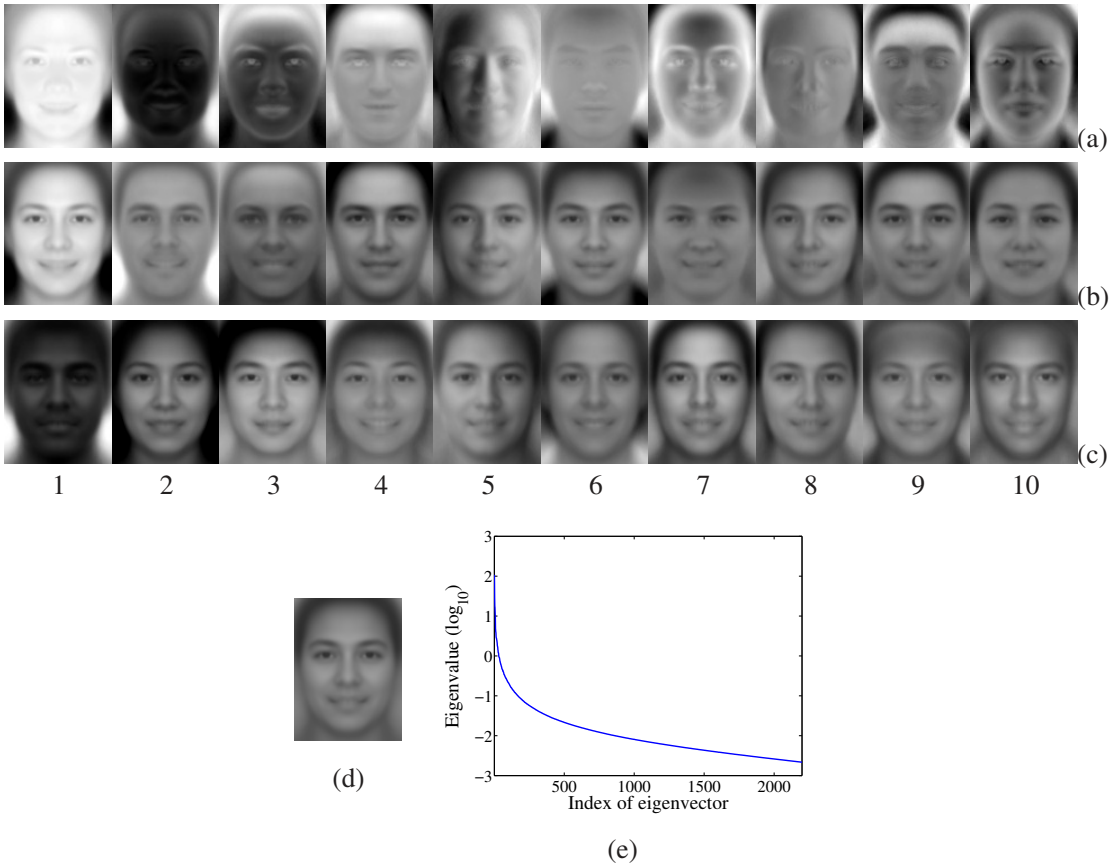


Figure 3. Eigenface [40]. (a) Top 10 eigenvectors corresponding to the 10 largest eigenvalues. (b) Eigenvectors (eigenfaces) are multiplied by 3σ where σ is the square root of eigenvalue and added to the mean face. (c) Eigenvectors are multiplied by -3σ and added to the mean face. (d) Mean face. (e) The logarithm of eigenvalues.

to $r = 1,000$ is too subtle to perceive. Therefore, we choose $r = 500$ for the soft constraint reconstruction.

The global face reconstructions from hard constraint are shown in Figure 6. From Eq. (17) the number of eigenvectors r should be larger than the dimension of the low-res N: $r > N$ ($N = 32 \times 24 = 768$). We gradually increase r from $1,000$ to $2,500$ in steps of 500, and display the results from (b) to (e). When $r = 1,000$, namely the number of eigenvectors are just above the low-res dimension, the reconstruction is poor because the main function of the eigenvectors is to satisfy the hard constraint, i.e. Eq. (15), and there is little freedom to maximize the posterior. As r increases, the eigenvectors have more freedom to maximize the posterior while satisfying the hard constraint, and therefore the reconstruction has fewer artifacts, as shown in (c) and (d). There is little visual difference between the reconstruction

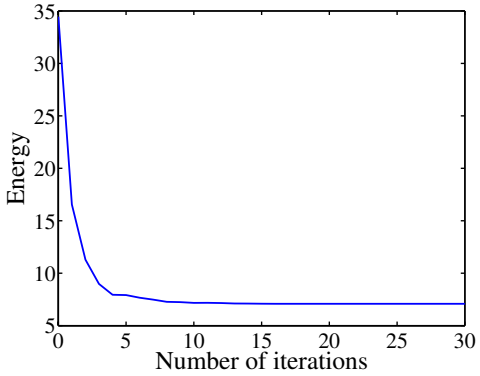


Figure 4. The energy converges quickly in max-product belief propagation.

from $r = 2,000$ and $r = 2,500$, so we choose $r = 2,000$ for the hard constraint reconstruction.

After the global face is reconstructed, we use the patch-based nonparametric Markov network model to infer the optimal local face, i.e. adding local details. For each patch we search in the database for the top 20 closest candidates, and use max-product belief propagation to minimize the energy. This algorithm converges in about 15 steps, as shown in Figure 4. The global faces (soft constraint) and the inferred high-res faces (global+local) are listed in Figure 7 (b) and (c). The ghosting effects are noticeably removed after the adaptive bilateral filtering is applied, as shown in (d). The results of the hard constraint (global+local+bilateral) are shown in (e). Comparing (d) and (e), we may observe that the soft constraint generates cleaner and sharper features with strong facial features such as eyeballs and teeth, but the results are close to the mean face. A typical example is the hair of the 6th sample from left to right, whose hair gets blurred by the soft constraint-based hallucination. The hard constraint generates images that well preserve the distinguished features of the low-res images, e.g. the hair details of the 6th sample are hallucinated, even though they are different from the original image. Nevertheless, the results generated using the hard constraint lack the crisp features of those generated using the soft constraint. In summary, soft constraint *beautifies* face in hallucination, whereas hard constraint faithfully *reproduces* facial details.

A significant advantage of the soft constraint over the hard constraint is the low memory load: the hard constraint requires 224MB memory whereas the soft constraint requires 50MB. The load of the Markov Network is about 800MB, but it can be significantly reduced when a clustering-based technique is used. In this paper we do not address the engineering work of reducing memory requirement.

We also compare our results with other approaches, e.g. bicubic interpolation and the inhomogeneous Markov Network [15] in Figure 7 (f) and (g). Note that this Markov Network implementation is the same as the local face part in our modeling, except that I_H^g is replaced

Figure 5. Experimental results on reconstructing the global face I_H^g using the soft constraint. (a) Input 24×32 low-res images. From (b) to (f) are global faces inferred using the soft constraint with different eigenspace dimensions. (b) $r=20$, (c) $r=100$, (d) $r=500$ and (e) $r=1000$. (f) Original 96×128 high-res images. With fewer eigenvectors the reconstruction is smooth, close to the mean face, but lacks the distinguishing facial feature of the input low-res face. With more eigenvectors the reconstruction is closer to the individual face image, but we observe ghosting effects at edges, similar to the Gibbs effect in reconstructing step edges by Fourier basis.

Figure 6. Experimental results on reconstructing the global face I_H^g using the hard constraint. (a) Input 24×32 low-res images. From (b) to (e) are global faces inferred using the hard constraint with different eigenspace dimensions. (b) $r = 1000$, (c) $r = 1500$, (d) $r = 2000$, (e) $r = 2500$. (f) Original 96×128 high-res images. With fewer eigenvectors the reconstructions were noisy, because there is not much freedom to maximize the probability in the eigenspace given the low-res constraint. With more eigenvectors, however, most of the errors diminish. In (b) and (c) we may observe similar ghosting effect to the reconstruction using the soft-constraint.

by the enlarged I_L. Theoretically this model is similar to that of Baker and Kanade [1]. As we have pointed out in the introduction and related work, we may see that even though the Markov Network is doing even a better job in hallucinating the local facial feature details, the global facial features, such as symmetry, are missing. We also evaluate peak signal to noise ratio (PSNR) between the hallucinated and the original images by the three approaches, namely soft constraint, hard constraint and Markov Network, in Table 1. Clearly the proposed approaches (soft and hard constraints) outperform the Markov Network in terms of PSNR, and the hard constraint produces better results than soft constraint, though this is perceptually debatable.

PSNR	Soft constraint	Hard constraint	Freeman et. al.
mean	26.81	27.40	26.42
max	32.25	32.67	30.85
min	22.80	23.28	22.85

Table 1. The statistics of PSNR for three face hallucination approaches.

In this section, we have used the toy domain of manually aligned face images to understand effects of parameter variations independently of alignment issues. In the next section we return to the more general problem of unaligned faces and fully automatic processing.

5. Accurate Alignment of Low-Res Face Image

Face alignment is key to the success of an automatic face hallucination algorithm. In practice, we cannot assume that any low-res face has been accurately aligned although the approximate localization of the low-res face is given by face detection. We have used the face detector presented in [43, 41] to detect all the possible faces from a single image. The face detector outputs the top-left and bottom-right coordinates of each face. This is the initialization for the face alignment algorithm, which has two components, affine warping and multiple randomized initializations.

5.1 Alignment by Affine warping

Let the input image be I. Let $\mathbf{z} = \{z_i\} = \{(x_i, y_i)\}$ be the coordinate of the face template. We want to know a warping function $\mathbf{W}(z, \mathbf{p})$ so that the warped $I(\mathbf{W}(\mathbf{z}, \mathbf{p}))$ is close to a face image. Here affine transformation is chosen as the warping function

$$\mathbf{W}(z, \mathbf{p}) = \begin{bmatrix} p_1 & p_3 & p_5 \\ p_2 & p_4 & p_6 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{34}$$

Figure 7. Comparison of different algorithms. (a) Input low-res face images. (b) Reconstruction by global face model (soft constraint). (c) Reconstructed by combining the global with the local face model, obtained from the nonparametric Markov network. Many local facial details are added from (b) to (c), e.g. highlights in eyeballs, teeth. (d) Results after post-processing by smart bilateral filtering. Some noise and ghosting artifacts caused by PCA reconstruction are removed. (e) Hallucinated results by hard constraint. We also compare our algorithm with others. (f) Bicubic interpolation. (g) Freeman et. al. [14]'s approach, adapted to be inhomogeneous to meet [1]. Although facial detailed features can be reconstructed locally, the global facial features are in general lost via this approach. (h) Original high-res images.

where $\mathbf{p} = (p_1, \cdots, p_6)^T$. Let the mean and covariance of the faces on the template be μ and $\mathbf{\Sigma}$. We want to find the optimal affine warp parameter $\mathbf{p}^*$ so that

$$\mathbf{p}^* = \arg\min_{\mathbf{p}} (I(\mathbf{W}(\mathbf{z}, \mathbf{p})) - \mu)^T \mathbf{\Sigma}^{-1} (I(\mathbf{W}(\mathbf{z}, \mathbf{p})) - \mu) \tag{35}$$

This objective function is nonlinear because function $I(\mathbf{W}(\cdot))$ is nonlinear. The problem can be addressed by a gradient descent algorithm. Base on current $\mathbf{p}$, we want to compute an update

$$\mathbf{p} \leftarrow \mathbf{p} + \Delta\mathbf{p} \tag{36}$$

so that the objective function can be optimized. Similar to Lucas-Kanade approach [26, 3] the objective function in Eq. (35) is linearized by first order Taylor expansion

$$I(\mathbf{W}(\mathbf{z}, \mathbf{p} + \Delta\mathbf{p})) = I(\mathbf{W}(\mathbf{z}, \mathbf{p})) + \nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}} \Delta\mathbf{p} \tag{37}$$

The optimization problem becomes

$$\Delta\mathbf{p}^* = \arg\min_{\Delta\mathbf{p}} (\nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}} \Delta\mathbf{p} + I(\mathbf{W}(\mathbf{z}, \mathbf{p})) - \mu)^T \mathbf{\Sigma}^{-1} (\nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}} \Delta\mathbf{p} + I(\mathbf{W}(\mathbf{z}, \mathbf{p})) - \mu) \tag{38}$$

For affine motion, the Jacobian $\frac{\partial \mathbf{W}}{\partial p}$ is

$$\frac{\partial \mathbf{W}}{\partial \mathbf{p}} = \begin{bmatrix} x & 0 & y & 0 & 1 & 0 \\ 0 & x & 0 & y & 0 & 1 \end{bmatrix} \tag{39}$$

Let matrix $\mathbf{D} = \nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}} \in \mathbb{R}^{M \times 6}$. The solution to Eq. (38) is

$$\Delta\mathbf{p}^* = (\mathbf{D}^T \mathbf{\Sigma}^{-1} \mathbf{D})^{-1} \mathbf{D}^T \mathbf{\Sigma}^{-1} (\mu - I(\mathbf{W}(\mathbf{z}, \mathbf{p}))) \tag{40}$$

Note that it is not necessary to compute $\mathbf{\Sigma}^{-1}$ which is normally ill-conditioned. From the low-res images of the training data we get the principal components $\mathbf{B}_L$ and the corresponding eigenvalues $\mathbf{\Lambda}_L$. The inverse covariance matrix can be approximated by

$$\mathbf{\Sigma}^{-1} \approx \mathbf{B}_L \mathbf{\Lambda}_L^{-1} \mathbf{B}_L^T \tag{41}$$

Low-Res Face Alignment Algorithm

- Given initial guess of centroid z_0, scale s_0, orientation θ_0 (*from face detector*), and number of iterations n and number of samples m (*from computational considerations*).
- Set $z^* = z_0$, $s^* = s_0$, $\theta^* = \theta_0$. Error $J^* = \infty$.
- For i=1:m
- $z_0 = z^*$, $s_0 = s^*$, $\theta_0 = \theta^*$
- For i=1:n
- Sample $z \sim \mathcal{N}(z_0, \sigma_z^2 \mathbf{I})$, $s \sim \mathcal{N}(s_0, \sigma_s^2)$, $\theta \sim \mathcal{N}(\theta_0, \sigma_\theta^2)$.
- Initialize affine parameter $\mathbf{p}$ using z, s and θ.
- Optimize parameter $\mathbf{p}$ and get the minimal error J.
- If $J < J^*$ then $J^* = J,\ \mathbf{p}^* = \mathbf{p},\ z^* = z,\ s^* = s,\ \theta^* = \theta$.
- Output $\mathbf{p}^*$.

Figure 8. The algorithm of robustly aligning faces at low-res. It outputs reliable alignment when $n = 4$ and $m = 20$. The parameter setting is $\sigma_z = 1$, $\sigma_\theta = 0.05$, and $\sigma_s = 0.06$.

5.2 Robust alignment by randomization

The algorithm above is very sensitive to the initialization. It works well if the scale and orientation are nearly correct. Unfortunately, the initialization given by face detection algorithm contains errors in position, scale and orientation of the face. Our alignment algorithm needs to take the inaccuracy of initialization into account. Therefore, we have designed a randomized algorithm for the alignment with the pseudo code shown in Figure 8. The basic idea is to randomize the position, scale and orientation from the initialization, find the best transform $\mathbf{p}^*$, and restart randomization again from $\mathbf{p}^*$. Even though this algorithm is not efficient, we find it robust enough to align faces in low-res images. This is essential for automatic operation.

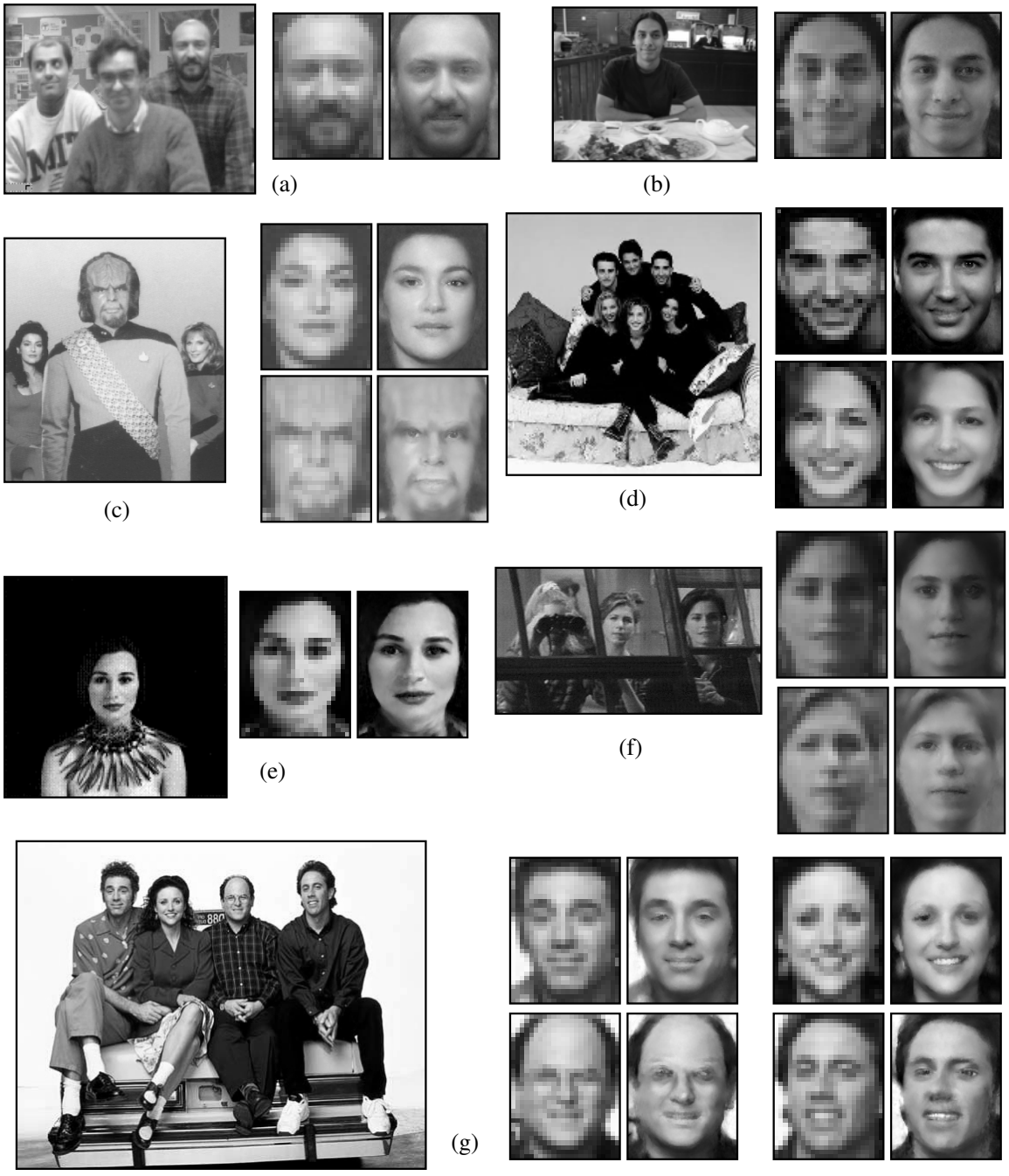


Figure 9. High-res hallucination from low-res faces using automatic detection and alignment of low-res face images. For each example, the input image is at left, the extracted, aligned low-res in the middle, and the high-res hallucinated at the right. All processing was using the soft constraint except for (e) and the bottom row of (g), which used the hard constraint (see text).

Figure 10. Our system is applied to hallucinating low-res faces in a group picture. Both the detection and synthesis processing was entirely automatic.

Figure 11. A failure case. The input picture is of very low quality and the face hallucination system cannot overcome such degradation, which are significantly different than those modeled in the training set. The hallucinated faces contain artifacts.

6 Experimental results

6.1 Face hallucination

We use CMU face database [33] and some other images to test the face hallucination system. We first run the system on a number of images and the result for a collection of test images is shown in Figure 9. The pairs of low-res and hallucinated high-res are displayed to the right of the original image from which the low-res faces are detected, registered and extracted. The results are shown at 128×96 resolution. The results in (e) and the bottom two in (g) are generated using the hard constraint, and the rest of them are generated using the soft constraint. The input images might be noise contaminated, and thus it is not necessary to use the hard constraint to enforce the hallucinated image to be exactly the same as the input when smoothed and downsampled, e.g. (a), (b) and (f). Meanwhile, the soft constraint tends to produce sharper facial features, producing clear eyelids and teeth, e.g. in (c) and (d). We also find that the soft constraint is more robust to the misalignment of the face. In (d) we may see that the registered

faces are not exactly upright, but the soft constraint is able to hallucinate reasonable results. However, for the top right example of (g), the soft constraint fails in hallucinating the details of the eyes.

The soft constraint tends to hallucinate results more like the mean face, as shown in the first row of (d). The hard constraint, on the other hand, faithfully represents the information in the low-res image, e.g. in (e) and the bottom row of (g). In (e) we see very strong facial features from the hard constraint. In the bottom row of (g), the face on the left wears eyeglasses and the right has a non-frontal pose. Because these two cases do not appear in the training, the soft constraint again tries to rectify the faces to the mean face, whereas the hard constraint is able to reproduce the information even though there are artifacts.

Our system is able to produce reasonable results even though the test images are drastically different from the training examples. An interesting example is the second row of (c) where an artificial mask is on the upper part of the face. Our face modeling handles well the unusual textures on the mask, as well as successfully hallucinating the facial details.

What if you have forgotten some faces of your classmates yet the old class photo is small and blurred? Our face hallucination system may be able to help, as shown in Figure 10. All the results are generated using the soft constraint. Our system is able to hallucinate the details of facial features, particularly eyes, eyebrows, mouth and nose though they are not visible in the low-res. However, we may observe that the symmetry of eyes is sometimes broken as in (a3), (b2) and (d1), which might be caused by the inaccurate registration.

As the image quality deteriorates further, or the size of the low-res face is significantly lower than 32×24, or the image contains some faces very different from those in the training, as shown in Figure 11, the hallucinated results do not improve the resolution as before. This shows again the characteristics of a learning-based vision system which requires certain amount of similarity between test and training samples.

6.2 Random Face Synthesis

Our probabilistic model for face appearance is not restricted to super resolution application. For instance, if we can model $p(I_L)$ and draw sample $I_L \sim p(I_L)$ in Eq. (12), then our model can be applied to synthesizing random faces[1].

To model $p(I_L)$ we first apply PCA to the low-res images and reduce the dimension to 40 which preserves 92.58% energy. A gaussian mixture model with 6 kernels is estimated in this subspace using the EM algorithm. In the sampling part, the Gaussian kernel is first sampled

[1]Note that this is not strictly the correct way to sample faces, which should also draw samples from $I_H \sim p(I_H|I_L)$ instead of Bayesian MAP inference.

Figure 12. Random synthesized faces. The male and female samples are arranged manually.

according to the weight, and then a sample is drawn by the Gaussian kernel. Projecting this sample to the low-res image space by the eigenvectors, we obtain a sample I_L. Then we use our face hallucination system (hard constraint) to hallucinate the high-res face I_H.

We randomly select 64 examples out of 1,000 random faces samples and display them in Figure 12. We manually arrange the males samples at the top four rows and females at bottom four for the sake of better comparison. The synthesized faces cover different race, lighting and expression, though the boundary part of the face is blurred. This is caused by the simple $p(I_L)$ model. We believe that more sophisticated model will further improve the quality of the synthesized faces.

7. Discussion

7.1 Face resolution

What resolution of face is needed to do face hallucination? Obviously there are two extreme cases. When the input is only one pixel, then face hallucination becomes a problem random face synthesis constrained by that the average intensity is given. When the input has very high resolution, e.g. 128×96, then there is no need to do hallucination, either. Therefore, there exists a range of resolutions in which face hallucination makes sense. In this paper we have chosen low-res at 32×24 for face hallucination. Most face detection systems have been designed using a 20×20 or 24×24 templates [33, 41, 43].

7.2 Why global and local modeling?

Theoretically we can solve face hallucination by Bayesian MAP in one step $I_H^* = \arg\max_{I_H} p(I_L|I_H)p(I_H)$, then why do we bother to decompose $p(I_H)$ into two steps of global and local modeling? The resolution of the high-res face image is $128\times96=12288$, requiring too many training examples to estimate a reasonable probability distribution $p(I_H)$ even using advanced machine learning algorithms, e.g. [25]. We designed a hybrid model for $p(I_H)$, i.e. a global face model by eigenfaces to capture the global facial features, and a local face model by Markov network to capture the local facial details.

7.3 Soft constraint vs. hard constraint

The likelihood model in the Bayesian MAP framework can be either formulated as a soft constraint, which implies Gaussian noise to the observation, or a hard constraint, which emphasizes that the reconstruction should be exactly the same as the input after being smoothed and downsampled. From a different perspective, when the number of eigenvectors is fewer

than the dimension of the low-res face, we can only apply the soft constraint. When there are more eigenvectors, we can enforce the hard constraint. The experimental results show that three times more eigenvectors are needed for the hard constraint than for the soft constraint.

From the experimental results we observe that the soft constraint tends to generate sharp facial details, be less sensitive to inaccurate registration, pose variation and noise, but smooth out the distinguished facial features of the input low-res face. The hard constraint, on the other hand, faithfully reproduces the distinguished facial features, but is very sensitive to any inaccurate registration and noise.

7.4 Face hallucination of a single person?

In our current system we use a database containing all kinds of faces. What if only a database of one person is applied? Since we are able to get a database for one person from daily digital pictures, face hallucination might be integrated with a face identification system to synthesize high-res image for a particular person. We feel that this would be an interesting direction for both face recognition and computational photography.

8. Conclusion

We have designed a two-step approach to hallucinating low-res face images by decomposing face appearance into a global eigenface model and a local Markov network model. To apply the hallucination system to real images we designed a low-res face registration tool to follow face detection so that high-res faces can be automatically hallucinated from low-res images with no manual intervention. We have both developed a theoretical framework for our hybrid approach, as well as addressed implementation details to solve practical issues that affect synthesis quality. The successful experimental results prove that face hallucination can be applied in real applications to enhance the resolution of face for both face recognition and face image editing. We also showed other applications of our face appearance modeling, e.g. random face synthesis.

9. Acknowledgement

The authors appreciate the help from Lin Liang of MSRA for aligning the training faces and running face detector for the test images. Ce Liu would like to thank Edward Adelson, Antonio Torralba and Bryan Russell for the insightful discussions. Heung-Yeung Shum thanks Takeo Kanade for helpful discussion on face hallucination and computer vision.

References

[1] S. Baker and T. Kanade. Hallucinating faces. In *IEEE International Conference on Automatic Face and Gesture Recognition*, March 2000.

[2] S. Baker and T. Kanade. Limits on super-resolution and how to break them. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2000.

[3] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *International Journal on Compter Vision*, 56(3):221–255, March 2004.

[4] A. Blake, B. Bascle, and A. Zisserman. Motion deblurring and super-resolution from an image sequence. In *Proc. European Conference on Computer Vision*, pages 312–320, 1996.

[5] J. De Bonet. Multiresolution sampling procedure for analysis and synthesis of texture images. *Proceedings of SIGGRAPH 97*, pages 361–368, August 1997.

[6] D. Capel and A. Zisserman. Super-resolution from multiple views using learnt image models. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 627–634, 2001.

[7] H. Chen, Y.Q. Xu, H.Y. Shum, S.C. Zhu, and N.N. Zheng. Example-based facial sketch generation with non-parametric sampling. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 433–438, 2001.

[8] T. F. Cootes and C. J. Taylor. Statistical models of appearance for computer vision. Technical report, University of Manchester, 2000.

[9] G. Dedeoglu, S. Baker, and T. Kanade. Resolution-aware fitting of active appearance models to low-resolution images. In *Proc. European Conference on Computer Vision*, pages 83–97. Springer, May 2006.

[10] 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*, volume 2, pages 151 – 158, June 2004.

[11] M.I. Jordan (ed.). *Learning in graphical models*. MIT Press, 1998.

[12] A. A. Efros and W. T. Freeman. Quilting for texture synthesis and transfer. In *Proceedings of SIGGRAPH 2001*, pages 341–346, August 2001.

[13] A. W. Fitzgibbon, Y. Wexler, and A. Zisserman. Image-based rendering using image-based priors. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 1176–1183, 2003.

[14] W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-based super-resolution. *IEEE Computer Graphics and Applications*, 22(2):56–65, 2002.

[15] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *International Journal on Compter Vision*, 40(1):25–47, 2000.

[16] H. Greenspan, C. Anderson, and S. Akber. Image enhancement by nonlinear extrapolation in frequency space. *IEEE Trans. on Image Processing*, 9(6), 2000.

[17] A. Hertzmann, C.E. Jacobs, N. Oliver, B. Curless, and D.H. Salesin. Image analogies. *Proceedings of SIGGRAPH 2001*, August 2001.

[18] H. H. Hou and H. C. Andrews. Cubic splines for image interpolation and digital filtering. *IEEE Trans. Acoust. Speech Signal Proc.*, 26(6):508–517, 1978.

[19] T. S. Huang and R. Y. Tsai. Multi-frame image restoration and registration. *Advances in Computer Vision and Image Processing*, 1:317–339, 1984.

[20] K. Jia and S. Gong. Multi-modal tensor face for simultaneous super-resolution and recognition. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 1683–1690, 2005.

[21] Y. Li and X. Lin. Face hallucination with pose variation. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 723–728, 2004.

[22] L. Liang, C. Liu, Y. Q. Xu, B. Guo, and H. Y. Shum. Real-time texture synthesis by patch-based sampling. *ACM Trans. Graph.*, 20(3):127–150, 2001.

[23] C. Liu, W. T. Freeman, R. Szeliski, and S. B. Kang. Noise estimation from a single image. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 901–908, 2006.

[24] C. Liu, H. Y. Shum, and C. S. Zhang. A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 192–198, 2001.

[25] C. Liu, S.C. Zhu, and H.Y. Shum. Learning inhomogeneous Gibbs model of faces by minimax entropy. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 281–287, 2001.

[26] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *the 7th International Joint Conference on Artificial Intelligence (IJCAI '81)*, pages 674–679, April 1981.

[27] A. Martinez and R. Benavente. The AR face database. Technical report, CVC Techinical Report No.24, June 1998.

[28] D. Martinez. *Model-based motion estimation and its application to restoration and interpolation of motion pictures*. PhD thesis, Massachusetts Institute of Technology, 1986.

[29] B. Morse and D. Schwartzwald. Image magnification using level set reconstruction. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 333–341, 2001.

[30] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab. *Signal and Systems*. Prentice Hall, Inc., 2nd edition, 1997.

[31] P. Philips, H. Moon, P. Pauss, and S. Rivzvi. The feret evaluation methodology for face-recognition algorithms. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 137–143, 1997.

[32] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes in C*. Cambridge University Press, second edition, 1992.

[33] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.

[34] R. R. Schultz and R. L. Stevenson. A bayesian approach to image expansion for improved definition. *IEEE Trans. Image Processing*, 3(3):233–242, 1994.

[35] E. Shechtman, Y. Caspi, and M. Irani. Space-time super-resolution. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(4):531–545, April 2005.

[36] G. Strang. *Linear Algebra and Its Applications*. Thomson Learning, Inc., 3rd edition, 1988.

[37] J. Sun, N. N. Zheng, H. Tao, and H.-Y. Shum. Generic image hallucination with primal sketch prior. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 729–736, 2003.

[38] M. F. Tappen, B. C. Russell, and W. T. Freeman. Efficient graphical models for processing images. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 673–680, 2004.

[39] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 839–846, 1998.

[40] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neurosciences*, 3:71–86, 1991.

[41] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 511–518, 2001.

[42] X. G. Wang and X. Tang. Hallucinating face by eigentransformation. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 35(3):425–434, 2005.

[43] R. Xiao, M.J. Li, and H.J. Zhang. Robust multipose face detection in images. *IEEE Trans. Circuits Syst. Video Techn.*, 14(1):31–41, 2004.

[44] Y. Zhou, L. Gu, and H.J. Zhang. Bayesian tangent shape model: Estimating shape and pose parameters via bayesian inference. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 109–118, 2003.

[45] S. Zhu, Y. Wu, and D. Mumford. Filters random fields and maximum entropy (FRAME): To a unified theory for texture modeling. *International Journal on Compter Vision*, 27:1–20, 1998.