Bayesian Tensor Inference for Sketch-based Facial Photo Hallucination

Wei Liu[†] **Xiaoou Tang**[†‡] **Jianzhuang Liu**[†]
[†]Dept. of Information Engineering, The Chinese University of Hong Kong, Hong Kong
[‡]Visual Computing Group, Microsoft Research Asia, Beijing, China
{wliu5, jzliu}@ie.cuhk.edu.hk and xitang@microsoft.com

Abstract

This paper develops a statistical inference approach, Bayesian Tensor Inference, for style transformation between photo images and sketch images of human faces. Motivated by the rationale that image appearance is determined by two cooperative factors: image content and image style, we first model the interaction between these factors through learning a patch-based tensor model. Second, by introducing a common variation space, we capture the inherent connection between photo patch space and sketch patch space, thus building bidirectional mapping/inferring between the two spaces. Subsequently, we formulate a Bayesian approach accounting for the statistical inference from sketches to their corresponding photos in terms of the learned tensor model. Comparative experiments are conducted to contrast the proposed method with state-of-the-art algorithms for facial sketch synthesis in a novel face hallucination scenario: sketch-based facial photo hallucination. The encouraging results obtained convincingly validate the effectiveness of our method.

1 Introduction

Recently, machine learning becomes more and more popular applied to the computer vision community. Various applications and methods that learn low-level vision have been proposed in the classical literature [Freeman *et al.*, 2000]. In this paper, we focus on a fascinating vision topic: automatic image sketching which automatically generates alike sketch images from photo images. Sketches are the simplest form of drawings because they consist of only drawing lines. The artists can distill the identifying characteristics of a photo and highlight them with a small number of critical strokes.

Implementing the great idea of learning vision, successful image sketching techniques try to observe and learn from the artist's works, and hence generate vivid and expressive sketches. As a result, example-based methods are widely studied in the latest years. Given a set of training images and their associated sketches drawn by artists, it is of interest to generate a sketch image automatically from an input image with the help of machine learning techniques.

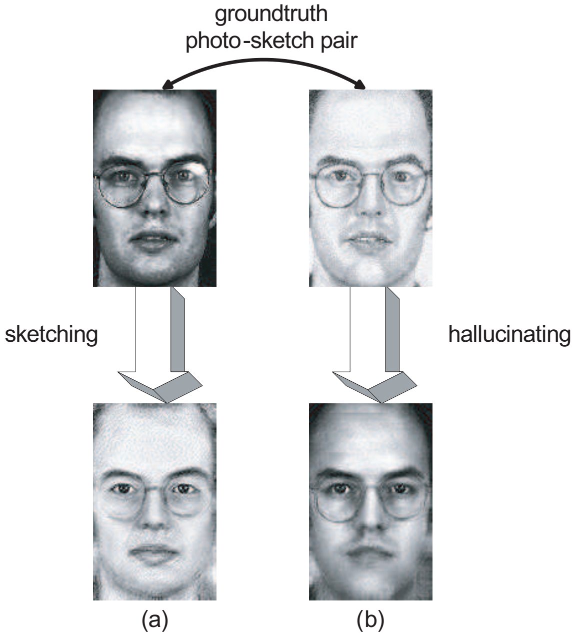

Figure 1: Bidirectional transforms on photo-sketch pairs. (a) Forward transform: synthesizing a sketch image from a photo image; (b) backward transform: hallucinating a photorealistic image from a sketch image.

For the particular image class of human faces, the transform between photo-realistic faces and their associative sketch drawings has shown promising applications in the future. In recent years, some works have been done to address the issues of synthesizing sketches from photos [Chen *et al.*, 2001; Tang and Wang, 2004; Liu *et al.*, 2005a] and sketch-based face recognition [Tang and Wang, 2004; Liu *et al.*, 2005a]. However, to the best of our knowledge, a more difficult issue - the backward transform from sketches to photos - has not been seriously addressed. In this paper, we develop a novel research topic: hallucinating photorealistic faces from sketches, which is termed as *sketch-based facial photo hallucination*. We design a new face hallucination technique to fulfill the intractable backward transform. We also consider the forward and backward transforms together in order to explore the inherent relation between the bidirectional transforms shown in Fig. 1.

Considering the complexity of image spaces and the conspicuous distinction between photos and sketches, global linear models such as [Tang and Wang, 2004; Liu *et al.*, 2005a] tend to oversimplify this problem. Therefore, we try to extract the local relations by explicitly establishing the connection between two feature spaces formed by a patch-based tensor model. To hold the statistical dependencies between pairwise patches with two styles more precisely and flexibly, we present a *Bayesian Tensor Inference* approach that incorporates the advantages of multilinear analysis techniques based on tensor into Bayesian statistics.

The rest of this paper is organized as follows. In Section 2, we elicit a tensor model for a facial image ensemble. The detailed rationale and algorithm for Bayesian Tensor Inference are presented in Section 3. Experimental results are shown in Section 4 and conclusions are drawn in Section 5.

2 Tensor Model

Recently, multilinear algebra and tensor modeling have attracted considerable attention in both computer vision and computer graphics communities. Research efforts applying tensor cover a broad range of topics including face modeling and synthesis [Vasilescu and Terzopoulos, 2002; Wang and Ahuja, 2003; Vlasic *et al.*, 2005], super-resolution [Liu *et al.*, 2005b], etc.

Motivated by previous multilinear approaches, we make use of a novel tensor model to exclusively account for the representation of images with two styles: photo-style and sketch-style. As small image patches can account for high-level statistics involved in images, we take patches as constitutive elements of the tensor model. Based on a corpus containing image patches with photo- and sketch-styles, we arrange all these patches into a high-order tensor which will suffice to encode the latent connection between the two styles.

2.1 TensorPatches

Based on the observation that both styles, i.e. photo- and sketch-styles, share some common characteristics and each style possesses its special traits, we assume the existence of decomposition of the patch feature space into the *common variation space*, which reflects the commonalities shared by both styles, and the *special variation space*. Relying on this rationale, we employ multilinear algebra to perform tensor decomposition on one large patch ensemble carrying two modalities.

Let us divide training pairwise images into overlapping small square patches which are assumed to be of the same size. m pairs of patches within a spatial neighborhood located in images are collected to form a high-order tensor. Following the non-parametric sampling scheme [Chen *et al.*, 2001], we allow m to be smaller than the length of each patch feature vector d. Resulting from the confluence of three modes related to patch examples, patch styles and patch features, a 3-order tensor $\mathcal{D} \in \Re^{m\times 2\times d}$ is built by grouping pairwise patches pertaining to the same neighborhood.

This paper adopts the High-Order SVD [Lathauwer *et al.*, 2000] or N-mode SVD [Vasilescu and Terzopoulos, 2002], both of which are the generalizations of SVD, to decompose the higher-order tensor $\mathcal{D}$ as follows[1]

$$\begin{aligned}\mathcal{D} &= \mathcal{C} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3 \\ &= \mathcal{C} \times_1 \mathbf{U}_{patches} \times_2 \mathbf{U}_{styles} \times_3 \mathbf{U}_{features}, \quad (1)\end{aligned}$$

where $\mathcal{C}$, known as the *core tensor*, governs the interaction between the mode matrices $\mathbf{U}_1, \cdots, \mathbf{U}_N$. Mode-$n$ matrix $\mathbf{U}_n$ contains the orthonormal vectors spanning the column space of matrix $\mathbf{D}_{(n)}$ resulting from mode-n flattening (unfolding) $\mathcal{D}$. Defining a tensor $\mathcal{T} = \mathcal{C} \times_3 \mathbf{U}_{features}$ as *TensorPatches*, then we have

$$\mathcal{D} = \mathcal{T} \times_1 \mathbf{U}_{patches} \times_2 \mathbf{U}_{styles}. \quad (2)$$

So far, we succeed in decomposing the two-modal patch feature space into the common variation space spanned by $\mathbf{U}_{patches} \in \Re^{m\times m}$ and the special variation space spanned by $\mathbf{U}_{styles} \in \Re^{2\times 2}$.

For any patch whether it is photo- or sketch-style, its corresponding tensor representation is

$$\mathcal{P} = \mathcal{T} \times_1 \mathbf{w}^T \times_2 \mathbf{s}_k^T, \ k = 1, 2 \quad (3)$$

where $\mathbf{w}$ contains patch parameters encoding the common characteristics of pairwise patches, and $\mathbf{s}_k$, capturing style parameters reflecting the special properties of style k (1 denotes photo-style, 2 denotes sketch-style), is the k-th row vector of $\mathbf{U}_{styles}$. When $\mathbf{w}^T$ is the i-th row vector of $\mathbf{U}_{patches}$, the tensor representation of the i-th training patch with k-th style is obtained. Its vector representation can be derived via mode-1(or 2, 3) flattening the subtensor $\mathcal{P}$, that is, $(f_1(\mathcal{P}))^T$.

For a new patch pair $(\mathbf{x}, \mathbf{y})$, $\mathbf{x}$ represents the photo-style and $\mathbf{y}$ denotes the sketch-style. Their tensor representations can be learned in line with eq. (3) by

$$\begin{aligned}\mathcal{P}(\mathbf{x}) &= \mathcal{T} \times_1 \mathbf{w}^T \times_2 \mathbf{s}_1^T = \mathcal{A}_1 \times_1 \mathbf{w}^T \\ \mathcal{P}(\mathbf{y}) &= \mathcal{T} \times_1 \mathbf{w}^T \times_2 \mathbf{s}_2^T = \mathcal{A}_2 \times_1 \mathbf{w}^T, \quad (4)\end{aligned}$$

where both $\mathcal{A}_1 = \mathcal{T} \times_2 \mathbf{s}_1^T$ and $\mathcal{A}_2 = \mathcal{T} \times_2 \mathbf{s}_2^T$ are constant tensors. Mode-1 flattening eq. (4) results in

$$\begin{aligned}\mathbf{x} &= (f_1(\mathcal{A}_1))^T \mathbf{w} = \mathrm{A}_x \mathbf{w} \\ \mathbf{y} &= (f_1(\mathcal{A}_2))^T \mathbf{w} = \mathrm{A}_y \mathbf{w}. \quad (5)\end{aligned}$$

$\mathrm{A}_x \in \Re^{d\times m}$ and $\mathrm{A}_y \in \Re^{d\times m}$ are called *inferring matrices*. The shared parameter vector $\mathbf{w}$, expected to be solved, exists in the common variation space. We solve it through mapping pairwise patches into the common variation space as follows

$$\begin{aligned}\mathbf{w} &= \mathrm{B}_x \mathbf{x} \\ \mathbf{w} &= \mathrm{B}_y \mathbf{y}, \quad (6)\end{aligned}$$

where $\mathrm{B}_x = (\mathrm{A}_x^T \mathrm{A}_x)^{-1} \mathrm{A}_x^T$ and $\mathrm{B}_y = (\mathrm{A}_y^T \mathrm{A}_y)^{-1} \mathrm{A}_y^T$ are called *mapping matrices*. Fig. 2 illustrates the relations among $\mathbf{x}, \mathbf{y}, \mathbf{w}$.

When one of $(\mathbf{x}, \mathbf{y})$ is unknown, inferring the counterpart style from the known style of a patch is the objective of image sketching or hallucination. Combining eq. (5) and eq. (6), a coarse estimation for $\mathbf{y}$ can be derived as

$$\mathbf{y} \approx \mathrm{A}_y \mathrm{B}_x \mathbf{x}, \quad (7)$$

[1] Details for tensor decomposition can be found in [Lathauwer *et al.*, 2000; Vasilescu and Terzopoulos, 2002], we will not elaborate.

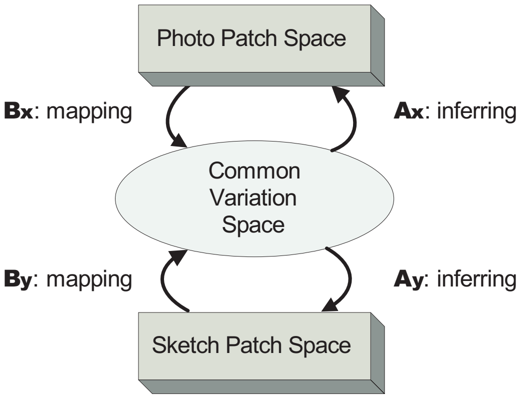

Figure 2: Illustration of relations among common variation space, photo patch space and sketch patch space.

which coincides with the path shown in Fig. 2: mapping the “Photo Patch Space” to the “Common Variation Space” from which inferring the “Sketch Patch Space” .

Importantly, the coarse estimation offered by eq. (7) is consistent with sketch synthesis methods [Tang and Wang, 2004][Liu *et al.*, 2005a] under particular conditions. Considering $(\mathbf{x}, \mathbf{y})$ as the holistic photo-sketch image pair of faces, when A_x and A_y are sample matrices whose columns are vectors recording the training sketch and photo images, eq. (7) is equivalent to the eigentransform method [Tang and Wang, 2004] where $\mathbf{w}$ contains linear combination coefficients w.r.t. training faces. In the case that A_x contains K nearest neighbors of input photo patch $\mathbf{x}$ in the training set and B_x collects associative sketches of these neighbors, eq. (7) resembles the local geometry preserving method [Liu *et al.*, 2005a]. Furthermore, eq. (7) can be thought as the unconstrained solution of LLE [Roweis and Saul, 2000] where $\mathbf{w}$ is the weighting vector for locally linear reconstruction.

2.2 TensorPatches for Facial Image Ensembles

Due to the structural characteristics of face images, we adopt a higher-order tensor rather than the 3-order tensor modeled on generic images. For the facial image ensemble, it necessitates modeling both global structures and local details of face images together.

Given the multi-factor (people, spatial positions of patches) and multi-modal (styles) natures of these patches, we develop a 4-order tensor model to explicitly account for both styles and constituent factors of facial patches. Suppose we have n pairs of face images available for training, of which each is divided into m overlapping square patches. With two styles, the length of each patch feature vector is d. Therefore, patches are influenced by four modes: people, patches, styles and features. A 4-order tensor $\mathcal{D} \in \Re^{n\times m\times 2\times d}$ is naturally built by grouping all photo-sketch patches sampled from the training faces. Perform tensor decomposition on $\mathcal{D}$

$$\begin{aligned}\mathcal{D} &= \mathcal{C} \times_1 \mathbf{U}_{people} \times_2 \mathbf{U}_{positions} \times_3 \mathbf{U}_{styles} \times_4 \mathbf{U}_{features} \\ &= \mathcal{T} \times_1 \mathbf{U}_{people} \times_2 \mathbf{U}_{positions} \times_3 \mathbf{U}_{styles},\end{aligned}$$

where the core tensor $\mathcal{C}$ governs the interaction between 4 modes encoded in 4 mode matrices: $\mathbf{U}_{people} \in \Re^{n\times n}$, $\mathbf{U}_{positions} \in \Re^{m\times m}$, $\mathbf{U}_{styles} \in \Re^{2\times 2}$ and $\mathbf{U}_{features} \in \Re^{d\times d}$, and TensorPatches $\mathcal{T}$ is obtained by forming the product $\mathcal{C} \times_4 \mathbf{U}_{features}$.

Significantly, the two factors (people, position) are crucial to determine a concrete patch and encoded in row vector spaces of mode matrices $\mathbf{U}_{people}$ and $\mathbf{U}_{positions}$. To remove factor (people, position) redundancies and suppress the effects of noise, a truncation of the two mode matrices is needed. In this paper, we adopt the N-mode orthogonal iteration algorithm in [Vasilescu and Terzopoulos, 2003] to perform factor-specific dimensionality reduction, outputting converged truncated mode matrices $\hat{\mathbf{U}}_{people} \in \Re^{n*r_1}(r_1 < n)$, $\hat{\mathbf{U}}_{positions} \in \Re^{m*r_2}(r_2 < m)$ along with the rank-reduced approximation of tensor $\mathcal{D}$

$$\hat{\mathcal{D}} = \hat{\mathcal{T}} \times_1 \hat{\mathbf{U}}_{people} \times_2 \hat{\mathbf{U}}_{positions} \times_3 \hat{\mathbf{U}}_{styles}. \quad (8)$$

In analogy to the last subsection, for a new patch pair $(\mathbf{x}_j, \mathbf{y}_j)$ residing in the j-th spatial position of face images, their tensor representations can be derived with the similar form as eq. (4)

$$\begin{aligned}\mathcal{P}(\mathbf{x}_j) &= \hat{\mathcal{T}} \times_1 \mathbf{w}^T \times_2 \mathbf{v}_j^T \times_3 \mathbf{s}_1^T = \mathcal{A}_{1j} \times_1 \mathbf{w}^T \\ \mathcal{P}(\mathbf{y}_j) &= \hat{\mathcal{T}} \times_1 \mathbf{w}^T \times_2 \mathbf{v}_j^T \times_3 \mathbf{s}_2^T = \mathcal{A}_{2j} \times_1 \mathbf{w}^T, \quad (9)\end{aligned}$$

where $\mathbf{v}_j^T$ and $\mathbf{s}_k^T$ is the j-th and k-th row vector of the mode matrix $\hat{\mathbf{U}}_{positions}$ and $\hat{\mathbf{U}}_{styles}$, respectively. Both $\mathcal{A}_{1j} = \hat{\mathcal{T}} \times_2 \mathbf{v}_j^T \times_3 \mathbf{s}_1^T$ and $\mathcal{A}_{2j} = \hat{\mathcal{T}} \times_2 \mathbf{v}_j^T \times_3 \mathbf{s}_2^T$ are constant tensors. The people parameter vector $\mathbf{w} \in \Re^{r_1\times 1}$ maintains to be solved for new patch pairs.

Mode-1 flattening eq. (9) results in

$$\begin{aligned}\mathbf{x}_j &= (f_1(\mathcal{A}_{1j}))^T \mathbf{w} = \mathrm{A}_x^j \mathbf{w} \\ \mathbf{y}_j &= (f_1(\mathcal{A}_{2j}))^T \mathbf{w} = \mathrm{A}_y^j \mathbf{w}, \quad (10)\end{aligned}$$

where $\mathrm{A}_x^j, \mathrm{A}_y^j \in \Re^{d\times r_1}$ are position-dependent inferring matrices. Note that the people parameter vector $\mathbf{w}$ is shared in all patch pairs $(\mathbf{x}_j, \mathbf{y}_j)_{j=1}^m$ appearing in the same person. Defining a concatenated photo feature vector $\mathbf{I}_x = [\mathbf{x}_1; \cdots; \mathbf{x}_m] \in \Re^{md\times 1}$ whose sketch counterpart is $\mathbf{I}_y = [\mathbf{y}_1; \cdots; \mathbf{y}_m] \in \Re^{md\times 1}$ and two enlarged $md \times r_1$ matrices $\overline{\mathrm{A}}_x = [\mathrm{A}_x^1; \cdots; \mathrm{A}_x^m]$ and $\overline{\mathrm{A}}_y = [\mathrm{A}_y^1; \cdots; \mathrm{A}_y^m]$, we have

$$\begin{aligned}\mathbf{I}_x &= \overline{\mathrm{A}}_x \mathbf{w} \\ \mathbf{I}_y &= \overline{\mathrm{A}}_y \mathbf{w}. \quad (11)\end{aligned}$$

So we can solve the parameter vector $\mathbf{w}$ by the least square method

$$\begin{aligned}\mathbf{w} &= \left(\overline{\mathrm{A}}_x^T \overline{\mathrm{A}}_x\right)^{-1} \overline{\mathrm{A}}_x^T \mathbf{I}_x = \overline{\mathrm{B}}_x \mathbf{I}_x \\ \mathbf{w} &= \left(\overline{\mathrm{A}}_y^T \overline{\mathrm{A}}_y\right)^{-1} \overline{\mathrm{A}}_y^T \mathbf{I}_y = \overline{\mathrm{B}}_y \mathbf{I}_y, \quad (12)\end{aligned}$$

where both $\overline{\mathrm{A}}_x^T \overline{\mathrm{A}}_x$ and $\overline{\mathrm{A}}_y^T \overline{\mathrm{A}}_y$ are invertible because $md >> r_1$ which is always satisfied throughout this paper.

3 Bayesian Tensor Inference

The learned tensor representations eq. (9) model the latent connection between $\mathbf{I}_x$ and $\mathbf{w}$, or $\mathbf{I}_y$ and $\mathbf{w}$. The concluded relations eq. (5), (6), (11), and (12) originate from factor decomposition which is implemented through tensor decomposition. Although eq. (7) gives a direct inference from $\mathbf{x}$ to $\mathbf{y}$, we lack analysis in statistical dependencies between $\mathbf{x}$ and $\mathbf{y}$.

Utilizing the merits of super-resolution techniques [Freeman *et al.*, 2000; Baker and Kanade, 2002], we incorporate the learned relations eq. (11) and (12) in which the tensor model entails into a Bayesian framework. The entire inference process is thus called the *Bayesian Tensor Inference*.

3.1 Formulation

We fulfill the backward transform $\mathbf{I}_y \rightarrow \mathbf{I}_x$ through the people parameter vector $\mathbf{w}$ by taking these quantities as a whole into a global optimization formulation. Our inference approach is still deduced from canonical Bayesian statistics, exploiting PCA to represent the photo feature vector $\mathbf{I}_x$ to be hallucinated. The advantage of our approach is to take into account the statistics between the latent variable $\mathbf{a} \in \Re^l$ and $\mathbf{I}_x$. We write the eigenface-domain representation after performing PCA on the training photo vectors $\{\mathbf{I}_x^{(i)}\}_{i=1}^n$

$$\mathbf{I}_x = \mathrm{U}\mathbf{a} + \mu + \varepsilon \approx \mathrm{U}\mathbf{a} + \mu, \tag{13}$$

where PCA noise ε may be ignored. The prior $p(\mathbf{a})$ is easier to acquire

$$p(\mathbf{a}) \propto \exp\{-\mathbf{a}^T \Lambda^{-1} \mathbf{a}\}, \tag{14}$$

where Λ is a $l \times l$ diagonal matrix with eigenvalues at its diagonal.

Due to the Bayesian MAP (maximum a posterior) criterion, we find the MAP solution $\mathbf{a}^*$ for hallucinating the optimal $\mathbf{I}_x^*$ as follows

$$\begin{aligned} \mathbf{a}^* &= \arg\max_{\mathbf{w},\,\mathbf{a}} p(\mathbf{w}, \mathbf{a} | \mathbf{I}_y) \\ &= \arg\max_{\mathbf{w},\,\mathbf{a}} p(\mathbf{I}_y | \mathbf{w}, \mathbf{a}) p(\mathbf{w}, \mathbf{a}) \\ &= \arg\max_{\mathbf{w},\,\mathbf{a}} p(\mathbf{I}_y | \mathbf{w}) p(\mathbf{w} | \mathbf{a}) p(\mathbf{a}). \end{aligned} \tag{15}$$

We take the statistics between $\mathbf{w}$ and $\mathbf{I}_y$ into consideration. Due to the learned relation (see eq.(11)) $\mathbf{I}_y = \overline{\mathrm{A}}_y \mathbf{w}$, we assume the representation error of $\mathbf{I}_y$ w.r.t $\mathbf{w}$ is i.i.d. and Gaussian. The conditional PDF $p(\mathbf{I}_y|\mathbf{w})$ is thus obtained as follows

$$p(\mathbf{I}_y|\mathbf{w}) \propto \exp\left(-\frac{\|\mathbf{I}_y - \overline{\mathrm{A}}_y \mathbf{w}\|^2}{\lambda_1}\right), \tag{16}$$

where λ_1 scales the variance in the photo space. In analogy to above, we obtain the condition PDF $p(\mathbf{w}|\mathbf{a})$ in doing $\mathbf{w} = \overline{\mathrm{B}}_x \mathbf{I}_x$ and using eq. (13)

$$p(\mathbf{w}|\mathbf{a}) \propto \exp\left(-\frac{\|\mathbf{w} - \overline{\mathrm{B}}_x(\mathrm{U}\mathbf{a} + \mu)\|^2}{\lambda_2}\right), \tag{17}$$

where λ_2 scales the variance in common variation space.

Substituting PDFs given in eq. (14), (16), and (17) into eq. (15), maximizing $p(\mathbf{w}, \mathbf{a}|\mathbf{I}_y)$ is equivalent to minimizing the following object function

$$\begin{aligned} E(\mathbf{w}, \mathbf{a}) = \frac{\|\mathbf{I}_y - \overline{\mathrm{A}}_y \mathbf{w}\|^2}{\lambda_1} &+ \frac{\|\mathbf{w} - \overline{\mathrm{B}}_x(\mathrm{U}\mathbf{a} + \mu)\|^2}{\lambda_2} \\ &+ \mathbf{a}^T \Lambda^{-1} \mathbf{a}. \end{aligned} \tag{18}$$

Significantly, function (18) is quadratic, and consequently minimizing it is a quadratic optimization problem (unconstrained) where a globally optimal solution exists.

Taking the derivative of E with respect to $\mathbf{w}$ and $\mathbf{a}$ respectively, the partial gradients of E can be calculated as

$$\frac{\partial E}{\partial \mathbf{w}} = 2\left(\frac{\overline{\mathrm{A}}_y^T \overline{\mathrm{A}}_y}{\lambda_1} + \frac{\mathrm{I}}{\lambda_2}\right)\mathbf{w} - 2\left(\frac{\overline{\mathrm{A}}_y^T \mathbf{I}_y}{\lambda_1} + \frac{\overline{\mathrm{B}}_x \mathrm{U}\mathbf{a} + \overline{\mathrm{B}}_x \mu}{\lambda_2}\right),$$

$$\frac{\partial E}{\partial \mathbf{a}} = 2\left(\frac{\mathrm{U}^T \overline{\mathrm{B}}_x^T \overline{\mathrm{B}}_x \mathrm{U}}{\lambda_2} + \Lambda^{-1}\right)\mathbf{a} - \frac{2\mathrm{U}^T \overline{\mathrm{B}}_x^T (\mathbf{w} - \overline{\mathrm{B}}_x \mu)}{\lambda_2}.$$

By $\partial E/\partial \mathbf{w} = \mathbf{0}$ and $\partial E/\partial \mathbf{a} = \mathbf{0}$, we obtain a pair of solutions $\{\mathbf{w}^*, \mathbf{a}^*\}$ that is unique, and so must be globally optimal. The optimal MAP solution $\mathbf{a}^*$ is hereby given as

$$\begin{aligned} \mathbf{a}^* = &\left(\mathrm{U}^T \overline{\mathrm{B}}_x^T \mathrm{C}\, \overline{\mathrm{B}}_x \mathrm{U} + \lambda_2 \Lambda^{-1}\right)^{-1} \mathrm{U}^T \overline{\mathrm{B}}_x^T \\ &\left(\frac{\lambda_2 (\mathrm{I} - \mathrm{C}) \overline{\mathrm{A}}_y^T \mathbf{I}_y}{\lambda_1} - \mathrm{C}\, \overline{\mathrm{B}}_x \mu\right), \end{aligned} \tag{19}$$

where the $r_1 \times r_1$ matrix C is predefined as

$$\mathrm{C} = \mathrm{I} - \left(\frac{\lambda_2 \overline{\mathrm{A}}_y^T \overline{\mathrm{A}}_y}{\lambda_1} + \mathrm{I}\right)^{-1}. \tag{20}$$

So far, we accomplish the task of backward transform, i.e. hallucinating the photorealistic photo feature vector $\mathbf{I}_x^* = \mathrm{U}\mathbf{a}^* + \mu$ given a sketch feature vector $\mathbf{I}_y$ based on an available set of myriad training photo-sketch image pairs.

3.2 Algorithm

Our algorithm tackles the intractable problem: sketch-based facial photo hallucination with a learning-based scheme. There have been several successful relevant efforts on facial sketch synthesis. Example-based sketch generation [Chen *et al.*, 2001] takes pixels as rendering elements, so the computational efficiency is somewhat low. Linear methods [Tang and Wang, 2004; Liu *et al.*, 2005a] tend to trivialize the synthesized results without any statistical consideration. The fact that our work improves upon the existing techniques is that patch-based tensor learning and Bayesian inference are coordinated in our algorithm, which is illustrated in Fig. 3.

Collect a large corpus of n training facial image pairs $\{I_x^{(n)}, I_y^{(n)}\}_{n=1}^N$ with two styles, and divide each image into m overlapping patches. Because the sketch image space and photo image space are disparate in terms of gray-level intensities, to bring the two spaces into correspondence, we use the horizontal and vertical image derivatives as features of both photo- and sketch-style patches for learning the TensorPatches model. Concretely, we envision Bayesian Tensor Inference involving two phases: the training and testing phase.

Training Phase

1. Based on the derivative feature representation, build a 4-order tensor on $n * m$ photo-sketch patch pairs sampled from n training image pairs and learn the matrices $\overline{\mathrm{A}}_y$ and $\overline{\mathrm{B}}_x$.

2. Construct concatenated photo feature vectors $\{\mathbf{I}_x^{(i)}\}_{i=1}^n$, and run PCA on them to achieve the eigen-system (U, Λ) and the mean vector μ.

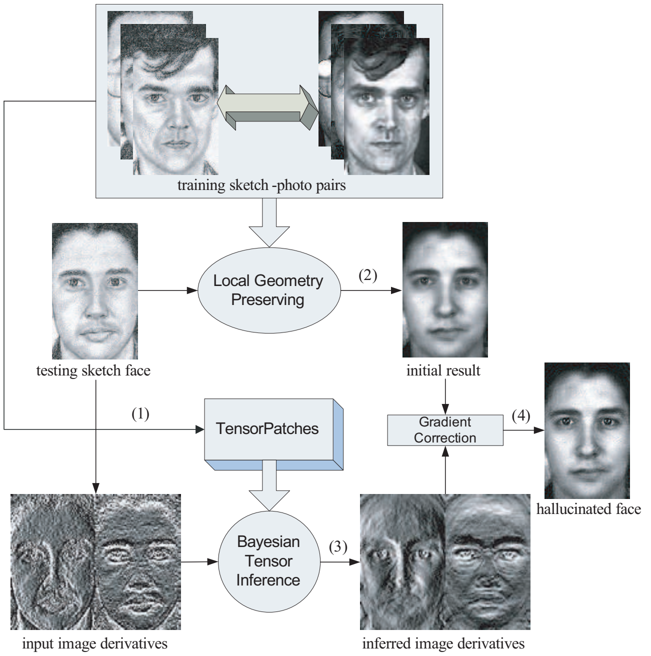

Figure 3: Architecture of our sketch-based facial photo hallucination approach. (1) Learn the TensorPatches model taking image derivatives as features in the training phase; (2) obtain the initial result applying the local geometry preserving method; (3) infer image derivatives of the target photo using the Bayesian Tensor Inference method, given input image derivatives extracted from the test sketch face; (4) conduct gradient correction to hallucinate the final result.

Testing Phase

1. For a new sketch image I_y, divide it into m patches the same way as in the training phase. Using grayscale intensities for image features, apply the local geometry method [Liu *et al.*, 2005a] to estimate an initial hallucinated result I_x^0.

2. Construct the concatenated sketch feature vector $\mathbf{I}_y$ using m overlapping patches, and exploit it and $\overline{\mathrm{A}}_y, \overline{\mathrm{B}}_x, \mathrm{U}, \Lambda, \mu$ to infer the target photo feature vector $\mathbf{I}_x^*$ according to eq.(19).

3. Break $\mathbf{I}_x^*$ into m patches and afterward integrate these patches into two holistic derivative maps I_x^h and I_x^v, with pixels in the overlapping area blended.

4. Conduct gradient correction on I_x^0 using the derivative maps I_x^h, I_x^v (i.e. gradient field) to render the final result I_x^f.

4 Experiments

Sketches often have many styles, for example, a cartoon sketch often represents a person very exaggeratedly. In this paper, we only focus on sketches of plain styles without exaggeration, as for applications in law enforcement or photo search, no much exaggeration is allowed in sketch images so that the sketches can realistically describe the real subjects. Some sketch examples are shown in Fig. 4 and Fig. 5.

We conduct experiments on a database over 600 persons of which 500 persons are for training and the rest is for testing. Each person has a frontal face photo image and a sketch image with a plain style drawn by an artist. All the training and testing images are normalized by affine transform to fix the position of centers of eyes and mouths. Each 160×100 face image is divided into 160 overlapping 13×13 patches in the same way, and the overlap between adjacent patches is three pixels.

To display the effectiveness of our Bayesian Tensor Inference, we compare it with the representative methods, which have been applied in facial sketch synthesis, including the global method– eigentransform [Tang and Wang, 2004] and the local method– local geometry preserving (LGP) [Liu *et al.*, 2005a]. Fig. 4 and Fig. 5 illustrates the results of sketch-based face hallucination for Asians and Europeans, respectively. In both the two groups of results, patch-based methods (c), (d) consistently and notably outperform the global method (b) in high fidelity of local details, which indicates that a local model is more suitable for modeling complex distributions such as images. By comparing the results yielded by LGP (c) and those produced by our method (d), we can clearly see that our inference algorithm performs sharper and more realistic with higher image quality than LGP. One reason is that employing the patch-based tensor model can recover both the common face structure and local details.

We set the scale parameters λ_1 and λ_2 to be 0.02 and 0.2, respectively. We use standard SVD to do PCA on the training photo feature vectors, with 97% of the eigenvalue total retained. When applying the LGP method to obtain the initial hallucinated result in step 1 of the testing phase, we choose $K = 5$ to be the number of nearest neighbors.

From the quantitative perspective, Table 1 lists the average root-mean-square pixel errors (RMSE) for 4 different methods. The baseline is the nearest neighbor method. As might be expected, the performance of our method does improve as the number of training sample pairs increases from 250 (Training Set I) to 500 (Training Set II). Although the RMSE of our method is larger than that of LGP, the hallucinated images look better, as detailed as that in the groundtruth face images. To sum up, our sketch-based face hallucination method acquires the best visually perceptual quality.

Table 1: Comparison of Hallucination Methods.

Hallucination Method	Training Set I		Training Set II	
	RMSE	Red.	RMSE	Red.
Nearest Neighbor	52.49	-	47.20	-
Eigentransfrom	52.18	0.59%	43.14	8.60%
Local Geometry Preserving	50.74	3.33%	44.24	6.27%
Bayesian Tensor Inference	51.47	1.94%	44.90	4.87%

5 Conclusions

In this paper, we present a statistical inference approach called Bayesian Tensor Inference between the photo patch space and the sketch patch space. Based on the principle that only commonalities contribute to the inference process and by virtue of the patch-based tensor model, we capture and learn the inter-space dependencies. The Bayesian MAP framework integrates tensor modeling and statistical optimization to ensure the inference performance of the difficult vision problem: sketch-based facial photo hallucination. The potency of our method is sufficiently shown in the comparative experiments, achieving surprising photo rendering quality.

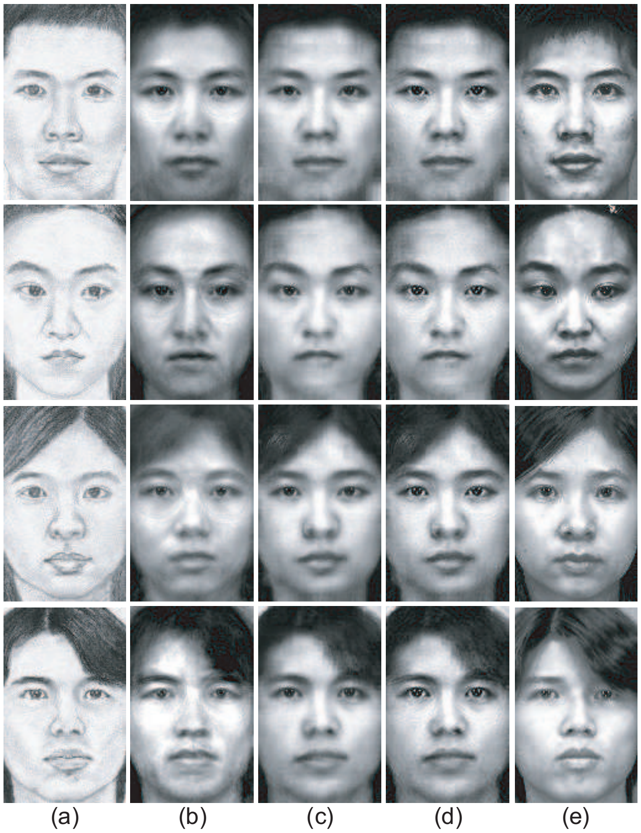

Figure 4: Photo hallucination results for Asian faces. (a) Input sketch images, (b) eigentransform method, (c) local geometry preserving method, (d) our method, (e) groundtruth face photos.

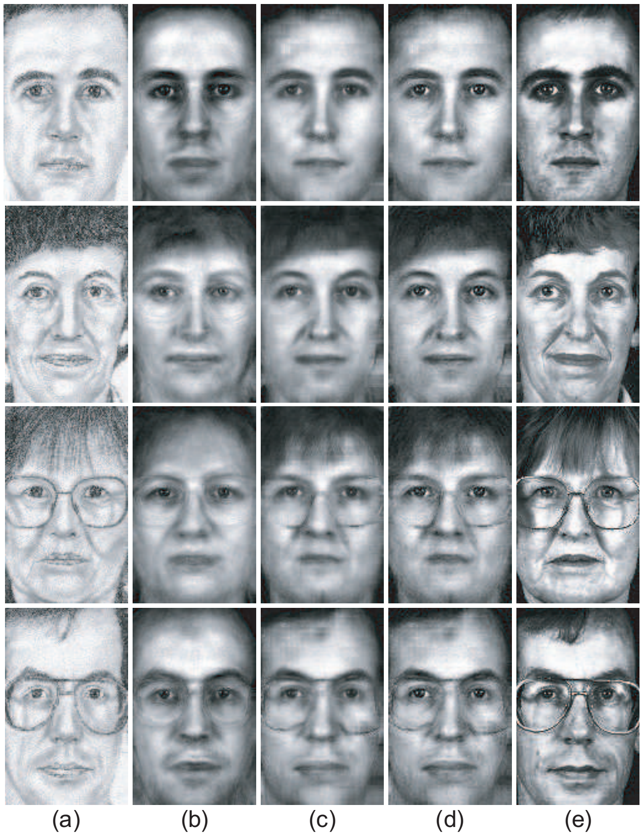

Figure 5: Photo hallucination results for European faces. (a) Input sketch images, (b) eigentransform method, (c) local geometry preserving method, (d) our method, (e) groundtruth face photos.

Acknowledgement

The work described in this paper was fully supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region and a joint grant (N_CUHK409/03) from HKSAR RGC and China NSF. The work was done at the Chinese University of Hong Kong.

References

[Baker and Kanade, 2002] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Trans. on PAMI*, 24(9):1167–1183, 2002.

[Chen *et al.*, 2001] H. Chen, Y. Xu, H. Shum, S-C. Zhu, and N. Zheng. Example-based facial sketch generation with non-parametric sampling. In *Proc. of ICCV*, 2001.

[Freeman *et al.*, 2000] W. Freeman, E. Pasztor, and O. Carmichael. Learning low-level vision. *IJCV*, 40(1):25–47, 2000.

[Lathauwer *et al.*, 2000] L. D. Lathauwer, B. D. Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[Liu *et al.*, 2005a] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma. A nonlinear approach for face sketch synthesis and recognition. In *Proc. of CVPR*, 2005.

[Liu *et al.*, 2005b] W. Liu, D. Lin, and X. Tang. Hallucinating faces: Tensorpatch super-resolution and coupled residue compensation. In *Proc. of CVPR*, 2005.

[Roweis and Saul, 2000] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[Tang and Wang, 2004] X. Tang and X. Wang. Face sketch recognition. *IEEE Trans. on CSVT*, 14(1):50–57, 2004.

[Vasilescu and Terzopoulos, 2002] M. Vasilescu and D. Terzopoulos. Muiltilinear analysis of image ensembles: Tensorfaces. In *Proc. of ECCV*, 2002.

[Vasilescu and Terzopoulos, 2003] M. Vasilescu and D. Terzopoulos. Multilinear subspace analysis for image ensembles. In *Proc. of CVPR*, 2003.

[Vlasic *et al.*, 2005] D. Vlasic, M. Brand, H. Pfister, and J. Popovic. Face transfer with multilinear models. In *Proc. of ACM SIGGRAPH 2005*, pp. 426–433, 2005.

[Wang and Ahuja, 2003] H. Wang and N. Ahuja. Facial expression decomposition. In *Proc. of ICCV*, 2003.