Learning 3D Shape from a Single Facial Image via Non-linear Manifold Embedding and Alignment

Xianwang Wang and Ruigang Yang
Center for Visualization & Virtual Environments
University of Kentucky
xwang,ryang@cs.uky.edu

Abstract

The 3D reconstruction of a face from a single frontal image is an ill-posed problem. This is further accentuated when the face image is captured under different poses and/or complex illumination conditions. In this paper, we aim to solve the shape recovery problem from a single facial image under these challenging conditions. The local image models for each patch of facial images and the local surface models for each patch of 3D shape are learned using a non-linear dimensionality reduction technique, and the correspondences between these local models are then learned by a manifold alignment method. By combining the local shapes, the global shape of a face can be reconstructed directly using a single least-square system of equations. We perform experiments on synthetic and real data, and validate the algorithm against the ground truth. Experimental results show that our method can yield accurate shape recovery from out-of-training samples with a variety of pose and illumination variations.

1. Introduction

Human face modeling, due to its vast application from biometric authentication to human-computer interactions, is a very active topic in computer vision research. Recovering the 3-dimensional shape from a single facial image is an under-constrained problem, since the same 2D intensity images may be generated from different shapes. Many algorithms have been developed to address this problem. Many approaches can be thought of as an extended Shape-from-Shading approach, in which the 3D shape is optimized so that its rendering matches the input image (e.g., [1, 2, 5, 23]. Domain-specific constraints are typically added to reduce the solution space so that meaningful results can be obtained. While some very impressive results have been obtained, one of the biggest challenges of these methods is that the optimization could be trapped in a local minimum.

Another class of methods use machine learning techniques to reconstruct the 3D shape(e.g., [4, 11, 16]). These Learning-based methods take advantage of the availably of prior training data, i.e., face images with the corresponding shapes, from which the relationship of shapes and facial images can be inferred. The reconstruction quality depends heavily on the training data sets. Given the need for high-quality 3D models and accurate data labeling, obtaining or reproducing good results is always difficult. In addition, they suffer from the curse of dimensionality problem, i.e., the requirement of a vast amount of training data to achieve accurate reconstruction. As a result, most of these methods focus solely on frontal images taken under ambient (or fixed) illuminations to reduce the amount of training data needed.

We present a novel approach in this paper to address these issues in current single-view 3D model methods. More specifically, the technical contributions of our method include:

- Instead of relying on explicit 2D-3D correspondences in the training database, we apply *manifold alignment* techniques to find the appropriate mapping between a 2D image and its corresponding shape. This eliminates the need for tedious and manual labeling in the training database.
- We introduce a new parametrization of the face model. Rather than recording the absolute position of vertices, we record the per-triangle affine transformation between an individual model and a reference model. This parametrization is invariant to pose changes of 3D shapes and implicitly encodes the fact that the vertices cannot move independently from one another.
- We divide the image and 3D shapes into overlapping patches and apply non-linear dimensionality reduction (DR) method to each patch. Working on the patch level has two advantages over the whole face: it is easier to compensate illumination locally, and the images and

shapes within a patch have considerably smaller variance [7]. Non-linear DR methods have been shown to be more effective for deformation [19].

Using these novel components, our approach is able to deal with the face images with both varying illumination and very large pose variation–up to 90° profile view, which we believe has not been demonstrated before. Note that the results are based on a training database without 2D-3D labeling or any illumination variations, making our method more accessible. Furthermore the global reconstruction is achieved by solving a linear system in closed form. No iterative step is needed.

The rest of the paper is organized as follows. Section 2 reviews the related work on 3D shape recovery from a single image. Section 3 introduces the preprocessing of the training data, images of face and the corresponding shapes. Section 4 describes the fundamentals of the Gaussian Process Latent Variable Model (GP-LVM) and its application for learning the local image models and the local surface models. The learning of correspondences between these models using the manifold alignment method is detailed in section 5. Section 6 presents the reconstruction procedure of the global shape by combining the learned local surface shapes. The experimental results and analytic analysis are shown in section 7, and the conclusion is made in section 8.

2. Related Work

A classic method to recover 3D shape from a single image is Shape-from-Shading (SFS) [3, 28]. Direct application of SFS to face modeling has limited success since a face has large albedo variation and both concave and convex regions. Some SFS-based methods have been developed to improve the shape recovery using specific domain constraints. The symmetric SFS method [29, 20] reconstructs the faces by exploiting the bilateral symmetry of faces. However, it is difficult to establish the point-wise correspondence between the symmetric parts. Prados *et al.* [15, 14] use a unique critical point over the face image to enforce convexity for the shape recovery. All the parameters of the light source, the surface reflectance and the camera have to be known.

Kemelmacher and Basri [8] presented an example-based SFS method for 3D shape recovery of a face from a single image using a single 3D reference model of a different person's face. To achieve a desired reconstruction, the method seeks the shape, albedo and lighting that best fit the input image while preserving the rough shape and albedo of the reference model. While this method provides accurate reconstruction of novel faces, it makes the assumption of Lambertian reflectance and rough alignment of the input image and the reference model. Similar methods include [22], results from only frontal face images are demonstrated.

Statistical SFS methods [1, 23, 5] represent face shapes in the parametric eigenspace by applying PCA to a training set of 3D faces. [1] seeks the shape-coefficients by fitting the PCA model to satisfy the image irradiance constraints, while [23] recovers the shape by fitting the PCA model to image brightness data using constraints on the surface normal direction provided by Lambert's Law. Dovgard and Basri [5] reconstruct the shape by combining the geometric constraint [29] and the statistical constraints [1]. These methods are computationally expensive in the fitting procedure for minimizing the error between the rendered facial surface and the intensity of the input face. And the optimization may not converge.

3D Morphable Model (3DMM) [2] developed by Blanz and Vetter is a well-known face reconstruction method. It applies to the images and shapes separately to derive the linear models. The 3D shape reconstruction is an optimization process which aims to minimize the difference between the rendered model image and the input image. However, 3DMM suffers from the same problem as Statistical SFS methods, long runtime and multiple local minima. The approach is presented to accelerate fitting procedure of 3DMM in [18].

Some learning-based methods have been developed for the shape reconstruction. Reiter *et al.* [16] recover the 3D shape from a NIR facial image by learning the canonical correlation analysis (CCA) mapping from near infrared (NIR) facial images to 3D shape, which are both transformed to vectors. Lei *et al.* [11] present an approach (Tensor+CCA) similar to [16], while the mapping is learned from the NIR tensor space to the 3D shapes. Castelan and Hancock [4] apply coupled statistical models (CSM) to recover surfaces from brightness images of faces. However, these statistical learning approaches can handle the shape recovery only from a frontal face image. Georghiades et al [6]. developed a generative method to handle pose and illumination varations for face recognition. The change of pose is limited to be less than +/- 30 degrees, while we can deal +/- 90○.

3. Training Data Preprocessing

2D Image Preprocessing All the training facial images are first automatically aligned to the reference images I_i^r using the method in [26]. The index i denotes the pose variation. We use different reference images for different poses. Estimating a 3D shape from a facial image with M pixels can be viewed as a generic non-linear M-dimensional regression problem. Even for small images, this dimensionality is still too large. To overcome this dimensionality issue, we adopt local, low-dimensional estimation based on small image patches. For each facial image of a specific pose, we divide it into N_z overlapping $p \times q$ rectangular patches. This patch representation not only reduces the problem dimen-

sion, but also makes illumination correction easier. Instead of applying global and complex methods (such as [17]), we can use simply local image normalization to correct non-uniform illumination or shading artifacts for each patch by:

$$J(x,y) = \frac{I(x,y) - m_I(x,y)}{\sigma_I(x,y)} \quad (1)$$

where $I(x,y)$ is the original image patch, $m_I(x,y)$ and $\sigma_I(x,y)$ are, respectively, the mean and the variance of $I(x,y)$, and $J(x,y)$ is the output image patch. After that histogram equalization is performed on $J(x,y)$. Figure 1 demonstrates the effectiveness of this approach. The corrected patches show little effect of lighting. Applying the same approach to an entire image is unlikely to be effective.

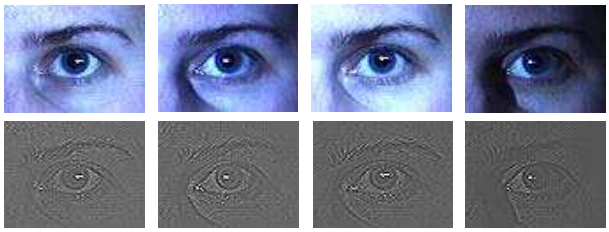

Figure 1. Local image patches before and after illumination correction.

As shown in Figure 2, after image subdivision and the normalization, we construct the data $\{\mathbf{Y}_{i,j} = [\mathbf{y}_{i,j,1}, \cdots, \mathbf{y}_{i,j,k}, \cdots, \mathbf{y}_{i,j,N}], j = 1 \cdots N_z\}$, where $\mathbf{y}_{i,j,k}$ is the transformed column vector from the facial image region with pose i and patch index j of the k^{th} person, and N is the number of subjects.

3D Shape Preprocessing We select one 3D facial shape $\mathbf{M}_r$ as a reference model, and every other facial model $\mathbf{M}_h$, $h = 1 \ldots N$, is registered to $\mathbf{M}_r$ using the coherent point drift (CPD) algorithm. CPD is a probabilistic method for non-rigid registration of point sets; details can be found in [13]. After the registration, each facial shape has the same number of vertices and triangles (in our experiments, 2500 vertices and 4624 triangles for each facial shape), which provides us convenience for later processing. Then, we parameterize the 3D shape model with deformation transfer, which describes the shape transformation from the source ($\mathbf{M}_r$) to the target ($\mathbf{M}_h$) [24]. The source deformation is represented as a collection of affine transformations tabulated for each triangle of $\mathbf{M}_r$, e.g., $\mathbf{T}_h = [\mathbf{q}_1, \cdots, \mathbf{q}_m]^T$, where m is the number of triangles, and $\mathbf{q}_v$ denotes the affine transformation of the v^{th} triangle. We also decompose $\mathbf{T}_h$ into N_z overlapped parts. With this, we construct the representations of 3D shapes $\widetilde{\mathbf{Y}}_j = [\widetilde{\mathbf{y}}_{j,1}, \cdots, \widetilde{\mathbf{y}}_{j,k}, \cdots, \widetilde{\mathbf{y}}_{j,N}]$, as shown in Figure 2, where $\widetilde{\mathbf{y}}_{j,k}$ is from the j^{th} patch of $\mathbf{T}_k$, corresponding to the facial image patches with patch index j of the k^{th} person.

4. The Local Image and Surface Models

In the previous sections, we explain how we gathered data as patches of facial images and 3D shapes. We will show how to learn the local image and surface models from such data. Generally, it is difficult to work with the data in the original high-dimensional space, since the number of training examples needed to to fully cover the space of possible deformations grows exponentially with the number of dimentions. A large amount of research work on non-linear manifold embedding has been done to handle the *curse of dimensionality*. We adopt the Gaussian Process Latent Variable Model (GP-LVM) [9], which provides a good generalization from very small data sets using non-linear models. An important characteristic of the GP-LVM is the reconstruction of a new point in the latent space with ease and accuracy. GP-LVM represents a Gaussian process (GP) mapping from the latent space $\mathbf{X}$ (low-dimensional embedding) to the data space $\mathbf{Y}$ (high-dimensional data set), where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]^T \in \Re^{N \times d}$ is the non-linear embedding matrix whose rows represent the corresponding positions in the latent space, $\mathbf{x}_i \in \Re^d$, and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N]^T \in \Re^{N \times D}$ is the data matrix in which each row is a single training sample, $\mathbf{y}_i \in \Re^D$. For a detailed discuss on GP and GP-LVM, see [9, 12]. Given a kernel function for the GP, $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)$, the likelihood of the data given the latent positions is

$$p(\mathbf{Y}|\mathbf{X}, \Theta) = \frac{1}{\sqrt{(2\pi)^{ND}|\mathbf{K}|^D}} \exp(-\frac{1}{2} tr(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^T)) \quad (2)$$

where $\mathbf{K}$ denotes the kernel matrix whose elements are defined by the kernel function $(\mathbf{K})_{i,j} = \mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)$, and Θ is the kernel hyper-parameters. In our experiments we use the form of the radial basis function (RBF) kernel, which controls the output variance, the RBF support width, the bias and the variance of the additive noise. GP-LVM learning consists of maximizing the posterior $p(\mathbf{X}, \Theta|\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{X}, \Theta)p(\mathbf{X})p(\Theta)$ with respect to the latent space $\mathbf{X}$, and the hyper-parameters Θ.

To reduce the computational complexity from an often prohibitive $O(N^3)$ to $O(Nk^2)$, sparse approximation techniques were proposed [10] and were proven more accurate than simply using a subset of the data. k is the number of points specified by the user in the sparse representation. All approximations involve augmenting the function values at the training points, $\mathbf{F} \in \Re^{N \times d}$, with $\mathbf{F} = [\mathbf{f}_1, \cdots, \mathbf{f}_N]^T$ and the function values at the test points, $\mathbf{F}_* \in \Re^{\infty \times d}$, by an additional set of variables, $\mathbf{X_u} \in \Re^{k \times d}$, called inducing variables. Learning the sparse GP-LVM involves maximizing with respect to $\mathbf{X}$, $\mathbf{X_u}$ and Θ the posterior

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{X_u}, \Theta)) = \mathrm{N}(\mathbf{K_{f,u}}\mathbf{K_{u,u}^{-1}}\mathbf{X_u}, \Lambda + \sigma^2\mathbf{I}) \quad (3)$$

where $\Lambda = diag[\mathbf{K_{f,f}} - \mathbf{K_{f,u}}\mathbf{K_{u,u}^{-1}}\mathbf{K_{f,f}}\mathbf{K_{u,f}}]$ and $diag(A)$ is a diagonal matrix whose elements match the diagonal of A,

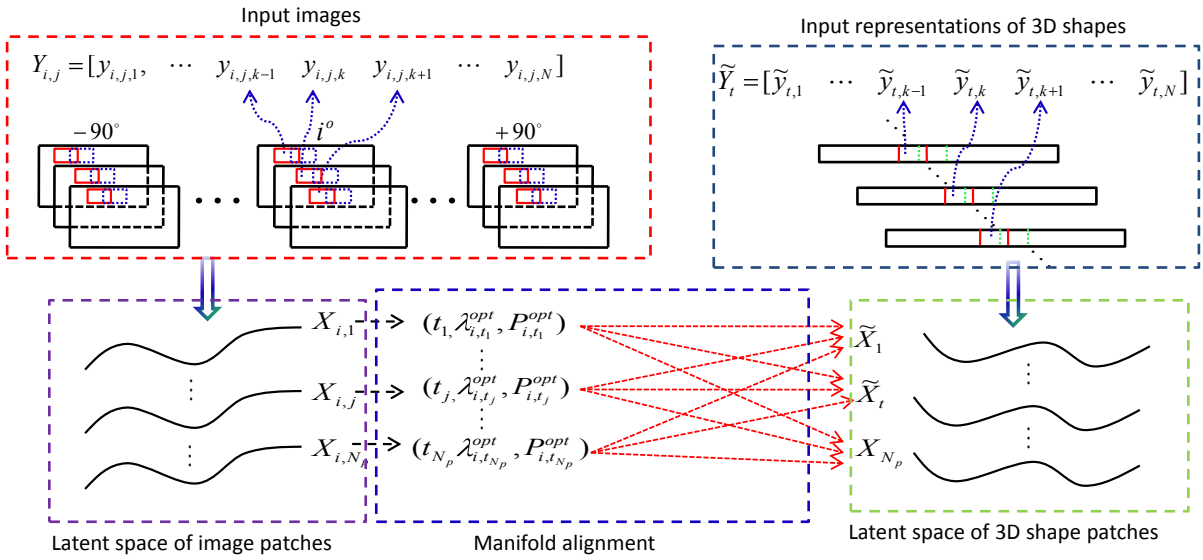

Figure 2. Data flow chart of our algorithm. $\mathbf{Y}_{i,j}$ is constructed from the image regions of all training faces with pose i and patch index j, while $\widetilde{\mathbf{Y}}_j$ is from the representations of all 3D shapes with patch index j. $\mathbf{Y}_{i,j}$ and $\widetilde{\mathbf{Y}}_j$ are projected into the low dimensional space using GP-LVM and generate $\mathbf{X}_{i,j}$ and $\widetilde{\mathbf{X}}_j$. For each $\mathbf{X}_{i,j}$, its correspondence, $\widetilde{\mathbf{X}}_{t_j}$, is found as the one with the minimal alignment error by the manifold alignment algorithm.

$\mathbf{K_{f,u}}$ denotes the covariance function computed between $\mathbf{X}$ and $\mathbf{X_u}$, $\mathbf{K_{u,u}}$ is the kernel matrix for the elements of $\mathbf{X_u}$, $\mathbf{K_{f,f}}$ is the symmetric covariance between $\mathbf{X}$, and σ^2 is the noise variance.

Given a new test point $\mathbf{x}_*$, the predictive distribution of its high-dimensional position $\mathbf{y}_*$ can be obtained [9] by

$$p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{Y}, \mathbf{X_u}, \Theta) = N(\mu_*, \sigma_*^2) \tag{4}$$

where the mean and variance are

$$\mu_* = \mathbf{Y}^T \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{K}_* \tag{5}$$

$$\sigma_*^2 = \mathbf{K}_{**} - \mathbf{K}_*^T \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{K}_* \tag{6}$$

where $\mathbf{K}_*$ is a vector with elements $\mathbf{K}(\mathbf{x}_*, \mathbf{x}_i)$ for latent positions $\mathbf{x}_i \in \mathbf{X_u}$, and $\mathbf{K}_{**} = \mathbf{K}(\mathbf{x}_*, \mathbf{x}_*)$.

Given a new test point $\mathbf{y}_*$, its latent position can be inferred in the sparse GP-LVM by minimizing $-\ln p(\mathbf{y}_*, \mathbf{x}_*|\mathbf{Y}, \mathbf{X_u}, \Theta)$, up to an additive constant [19],

$$\ell(\mathbf{x}_*, \mathbf{y}_*) = \frac{\|\mathbf{y}_* - \mu(\mathbf{x}_*)\|^2}{2\sigma^2(\mathbf{x}_*)} + \frac{D}{2} \ln \sigma^2(\mathbf{x}_*) + \frac{1}{2}\|\mathbf{x}_*\|^2 \tag{7}$$

with the mean and variance given by

$$\mu(\mathbf{x}_*) = \mathbf{Y}^T \mathbf{K}_{\mathbf{f},\mathbf{u}}^T \mathbf{A}^{-1} \mathbf{K}_* \tag{8}$$

$$\sigma^2(\mathbf{x}_*) = \mathbf{K}_{**} - \mathbf{K}_*^T (\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} - \sigma^2 \mathbf{A}^{-1}) \mathbf{K}_* \tag{9}$$

where $\mathbf{A} = \sigma^2 \mathbf{K_{u,u}} + \mathbf{K_{u,f}} \mathbf{K_{f,u}}$.

The local image model $\Theta_{i,j}^I$ and the low dimensional embedding $\mathbf{X}_{i,j} = [\mathbf{x}_{i,j,1}, \cdots, \mathbf{x}_{i,j,N}]$ are learned by the input image patches $\mathbf{Y}_{i,j}$, where $j = 1 \cdots N_z$. Similarly, we can get the local surface model Θ_t^S and the low-dimensional embedding $\widetilde{\mathbf{X}}_t = [\widetilde{\mathbf{x}}_{t,1}, \cdots, \widetilde{\mathbf{x}}_{t,N}]$ from the input 3D patches $\widetilde{\mathbf{Y}}_t$.

5. Learning the Correspondences

Previous methods in single-image 3D face modeling usually require explicit registration between the 2D images and the 3D models. Registration between different modality is a difficult problem. Typically this is done with user interaction. However, given our intention to deal with both pose and illumination variations, manually labeling all the images in the combinatory space is too time-consuming. Rather we develop an automatic procedure to estimate the correspondences via manifold alignment with procrustes analysis [25].

More specifically, we have two collections of low-dimensional embeddings, 2D image patches $\{\mathbf{X}_{i,j}\}$ and 3D shape patches $\{\widetilde{\mathbf{X}}_t\}$. We estimate a transformation (i.e. procrustes analsyis) to best align one data configuration ($\mathbf{X}_{i,j}$) to another ($\widetilde{\mathbf{X}}_t$). Each element of $\mathbf{X}_{i,j}$ and $\widetilde{\mathbf{X}}_t$ is first translated so that its centroid is at the origin, by

$$\begin{aligned} \mathbf{x}_{i,j,k} &= \mathbf{x}_{i,j,k} - \textstyle\sum_{k=1}^N \mathbf{x}_{i,j,k}/N, \quad j = 1 \cdots N_z \\ \widetilde{\mathbf{x}}_t &= \widetilde{\mathbf{x}}_{t,k} - \textstyle\sum_{k=1}^N \widetilde{\mathbf{x}}_{t,k}/N, \quad t = 1 \cdots N_z \end{aligned} \tag{10}$$

Then, we try to align $\mathbf{X}_{i,j}$ to all $\widetilde{\mathbf{X}}_t$. The alignment error of matching $\mathbf{X}_{i,j}$ and $\widetilde{\mathbf{X}}_t$ is defined by $\|\mathbf{X}_{i,j} - \lambda_{i,t} \widetilde{\mathbf{X}}_t \mathbf{P}_{i,t}\|_F$, where $\|\cdot\|_F$ denotes Frobenius norm, $\lambda_{i,t}$ is a re-scaling

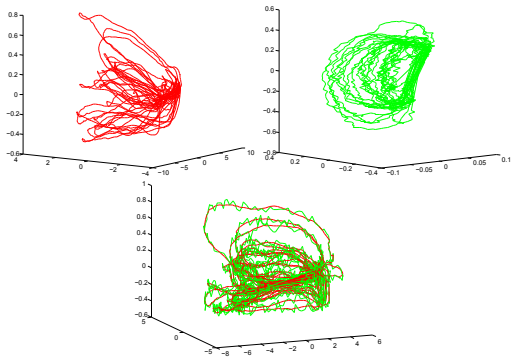

Figure 3. (Top) Example low-dimensional embeddings of 2D image patches and 3D shape patches from different subjects with pose changes; They are in different coordinate systems. (Bottom) the two embeddings after alignment.

factor to either stretch or shrink $\widetilde{\mathbf{X}}_{i,t}$, and $\mathbf{P}_{i,t}$ is an orthonormal matrix, defining a rotation and possibly a reflection. We denote the correspondence of $\mathbf{X}_{i,j}$ as $\widetilde{\mathbf{X}}_{t_j}$ with patch index t_j, which has the minimal alignment error with $\mathbf{X}_{i,j}$. That is, the problem is simplified to find the patch index t_j of 3D shape representations, λ_{i,t_j}^{opt} and the transform $\mathbf{P}_{i,t_j}^{opt}$ such that

$$\{t_j, \lambda_{i,t_j}^{opt}, \mathbf{P}_{i,t_j}^{opt}\} = \arg \min_{t \in \{1 \cdots N_z\}, \lambda_{i,t}, \mathbf{P}_{i,t}} \|\mathbf{X}_{i,j} - \lambda_{i,t} \widetilde{\mathbf{X}}_t \mathbf{P}_{i,t}\|_F \tag{11}$$

It is shown that $\lambda_{i,t_j}^{opt} = trace(\Sigma)/trace(\widetilde{\mathbf{X}}_{t_j}^T \widetilde{\mathbf{X}}_{t_j})$ and $\mathbf{P}_{i,t_j}^{opt} = \mathbf{U}\mathbf{V}^T$ in [25], where $\mathbf{U}$, $\mathbf{V}$ and Σ are given by the SVD of $\widetilde{\mathbf{X}}_{t_j}^T \widetilde{\mathbf{X}}_{t_j}$, that is, $\mathbf{U}\Sigma\mathbf{V}^T = SVD(\widetilde{\mathbf{X}}_{t_j}^T \widetilde{\mathbf{X}}_{t_j})$.

Our method is based on the assumption that corresponding 2D and 3D embedding will have similar shape, yielding the minimum amount of registration error. Figure 3 shows that this assumption is likely to be valid.

Give a new point $\mathbf{x}_{i,j,*}$ in the embedding space of $\mathbf{X}_{i,j}$, the point $\widetilde{\mathbf{x}}_{t_j,*}$ corresponding to $\mathbf{x}_{i,j,*}$ can be computed by

$$\mathbf{x}_{i,j,*} = \lambda_{i,t_j}^{opt} \widetilde{\mathbf{x}}_{t_j,*} \mathbf{P}_{i,t_j}^{opt} \tag{12}$$

6. Shape Recovery from a Single Image

In summary, from the training data set, the local image models, the surface models, and their correspondences are learned using manifold embedding and alignment techniques, as outlined in the learning phase of Algorithm 1. Then we can recover the corresponding local shapes, $\widetilde{\mathbf{y}}_1^*, \cdots, \widetilde{\mathbf{y}}_{N_z}^*$ of a single image in the reconstruction phase of algorithm 1. We first estimate the pose i of this image. We use the algorithm in [27], which is robust to large illumination and pose variations. The facial image is then aligned to the reference facial image I_i^r with the estimated pose using the method in [26] and divided into N_z overlapped patches, $\mathbf{y}_1^*, \cdots, \mathbf{y}_{N_z}^*$. We have to correct the illumination of these patches before the shape recovery via Eq. 1. After preprocessing, the local shape for each patch can be estimated as outlined in the reconstruction phase of Algorithm 1.

Algorithm 1. Locally Estimating the Shape

I. Learning phase

Input: A set of N_z training examples, $(\mathbf{Y}_{i,1}, \widetilde{\mathbf{Y}}_1)$, $\cdots$, $(\mathbf{Y}_{i,N_z}, \widetilde{\mathbf{Y}}_{N_z})$

Output: the N_z local models of image patches and shape patches, $\Theta_{i,1}^I, \cdots, \Theta_{i,N_z}^I$ and $\Theta_1^S, \cdots, \Theta_{N_z}^S$; the correspondences (patch index) and the optimal mapping parameters between the models: $(t_1, \lambda_{i,t_1}^{opt}, \mathbf{P}_{i,t_1}^{opt}), \cdots, (t_{N_z}, \lambda_{i,t_{N_z}}^{opt}, \mathbf{P}_{i,t_{N_z}}^{opt})$

1: **for** j = $1 \cdots N_z$

2: Learn the local image and shape models, $\Theta_{i,j}^I$ and Θ_j^S, and get the low-dimensional embeddings of $\mathbf{Y}_{i,j}$ and $\widetilde{\mathbf{Y}}_j$, $\mathbf{X}_{i,j}$ and $\widetilde{\mathbf{X}}_j$ by maximizing the posterior of Eq. 3.

3: **end for**

4: **for** j = $1 \cdots N_z$

5: Learn the correspondence of t_j and the optimal mapping parameters: λ_{i,t_j}^{opt}, $\mathbf{P}_{i,t_j}^{opt}$, between $\mathbf{X}_{i,j}$ and $\widetilde{\mathbf{X}}_{i,t_j}$ via Eq. 10 and 11.

6: **end for**

II. Reconstruction phase

Input: the test facial image patches, $\mathbf{y}_1^*, \cdots, \mathbf{y}_{N_z}^*$

Output: the recovered N_z local shapes, $\widetilde{\mathbf{y}}_1^*, \cdots, \widetilde{\mathbf{y}}_{N_z}^*$

1: **for** j = $1 \cdots N_z$

2: Compute the low-dimensional embedding, $\mathbf{x}_j^*$, of $\mathbf{y}_j^*$ by minimizing the negative log likelihood of Eq. 7 with the learned local image model, $\Theta_{i,j}^I$.

3: Map $\mathbf{x}_j^*$ into $\widetilde{\mathbf{x}}_j^*$ of low-dimensional space using the learned λ_{i,t_j}^{opt} and $\mathbf{P}_{i,t_j}^{opt}$ via Eq. 12.

4: Recover the local shape of $\widetilde{\mathbf{x}}_j^*$, $\widetilde{\mathbf{y}}_j^*$, by computing the mean of posterior in Eq. 4 with the learned local shape model, $\Theta_{t_j}^S$.

5: **end for**

Global Reconstruction The recovered representative of the local shapes, $\widetilde{\mathbf{y}}_1^*, \cdots, \widetilde{\mathbf{y}}_{N_z}^*$, need to be combined into the representative of a global shape, $\mathbf{s}^* = [\widetilde{\mathbf{y}}_1^* \cdots \widetilde{\mathbf{y}}_{N_z}^*]$. $\mathbf{s}^*$ can be considered as a collection of vectorized affine transformations of the triangles of the reference models $\mathbf{M}_r$. The problem that we need to solve here is to find the target shape $\mathbf{M}_u = \{\widetilde{\mathbf{v}}_1, \cdots, \widetilde{\mathbf{v}}_n\}$ to satisfy the constraints $\mathbf{s}^*$. For each target triangle of $\mathbf{M}_u$, and the affine transformation can be written as $\mathbf{T} = \widetilde{\mathbf{V}}\mathbf{V}^{-1}$ in terms of the original and deformed vertices. The elements of $\mathbf{V}^{-1}$ are coordinates of the known, original vertices of $\mathbf{M}_r$, while the elements of $\widetilde{\mathbf{V}}$ are coordinates of the unknown deformed vertices of $\mathbf{M}_u$. From this definition, we see that the elements of $\mathbf{T}$ are linear combinations of the coordinates of the unknown de-

formed vertices. Thus we can formulate the problem as a minimization problem [24]:

$$\min_{\widetilde{\mathbf{v}}_1 \ldots \widetilde{\mathbf{v}}_n} \sum_{j=1}^{|M|} \|\mathbf{S}_j - \mathbf{T}_j\|_F^2 \tag{13}$$

where $\mathbf{S}_j$ is the known source transformation, $|M|$ is the number of transformations in the constraint, and $\mathbf{T}_j$ is the unknown target transformation. Since the target transformations are defined in terms of the unknown deformed target vertices, the problem can be rewritten in the matrix form,

$$\min_{\widetilde{\mathbf{v}}_1 \ldots \widetilde{\mathbf{v}}_n} \|\mathbf{s}^* - \mathbf{A}\widetilde{\mathbf{x}}\|_2^2 \tag{14}$$

where $\widetilde{\mathbf{x}}$ is a vector of unknown deformed vertex coordinates, and A is a large, sparse matrix that relates $\widetilde{\mathbf{x}}$ to $\mathbf{s}^*$. Thus, all the vertices of the target shape $\mathbf{M}_u$ can be solved in the least-square sense.

7. Experimental Results

Data Sets To evaluate the performance of our approach, we employed two data sets in our experiments. The first one is a 3D face scans database [2], which contains shapes and textures of 120 real faces obtained with a laser scanner. We generate the synthetic facial images from them with pose and illumination changes. The pose changes horizontally from $-90°$ to $+90°$ at 5 degree increments. The illumination varies horizontally from $-45°$ to $+45°$ with a granularity of 5. The resolution of facial images is 256×256. Notice that the images provided in this database are not identical to the real albedos of the faces, due to noticeable effects of the lighting conditions. The second one is the CMU-PIE database [21], which contains 68 individuals with 9 horizontal and 3 vertical pose variations and 21 illumination variations.

Among these images, we use 2052 synthetic facial images as the training data set. They correspond to 108 subjects under 19 pose variations. The illumination condition is fixed at a natural (ambient) setting. To learn the local image and surface models, we use 60 inducing variables, and the latent dimension $d = 8$.

Experiments Our first experiment shows the effectiveness of our patch-based method for illumination variations. Note that our training database contains no sample under changing illumination. Figure 4 shows a comparison without and with illumination normalization. We use $m_I(x, y) = \sigma_I(x, y) = 2$ for synthetic images and $m_I(x, y) = \sigma_I(x, y) = 0.5$ for real images in Eq. 1 to correct the illumination variation, and divided all the images into N_z overlapped patches with the size of 7×7. The value of N_z depends on the pose of images, e.g., 252 patches for the frontal faces in our experiments.

Synthetic Inputs We use the images and shapes from the remaining 12 subjects in the first database as the testing data to run a controlled experiment. Our method is used to recover their shapes from the synthetic images. This experiment allows us to show comparisons of our reconstructions to the ground truth shapes. The quantitative accuracy of reconstruction can be defined as [11]:

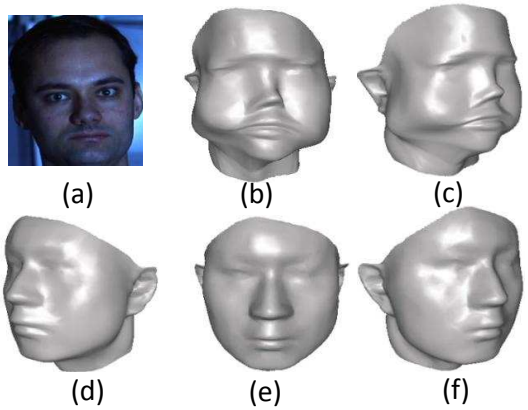

Figure 4. Shape recovery from a single frontal image w/o local illumination normalization. (a) the input frontal images with illumination; (b,c) different views of the reconstruction result without illumination normalization; (d,e,f) different views of the reconstruction result with illumination normalization.

$$\varepsilon = \frac{1}{n} \sum_{i=1}^{n} |D_r(i) - D_t(i)| \tag{15}$$

where D_r is the recovered shape and D_t is the ground truth shape, and n is the number of vertices in the shape. Figure 5 shows a few results. For comparison we show the reconstructed shapes and the ground truth, and plot the alignment of the reconstructed shapes (in gray) with the ground truth shapes (in blue). It can be seen that our algorithm can obtain accurate reconstructions in spite of illumination and pose variations. The reconstructed error in each pose is shown in Figure 7, which shows that our algorithm is fairly insensitive to pose variations and achieves the same level of accuracy as the methods [11, 4, 16] in all poses. The recovery accuracy for the frontal facial images in our method is slightly better than that of those methods, but our method can handle illumination and pose changes.

Real Inputs We apply our method to several real images from CMU database using the same training data set. The reconstructed results are shown in Figure 6.

8. Conclusion

In this paper we proposed a novel approach to the shape recovery from a single side-view image. We studied the limitation of related approaches in shape recovery for facial images with illumination and pose variations and addressed the problem using non-linear embedding and alignment. We conducted experiments to evaluate our approach by comparing the reconstructed results to ground truth shapes and

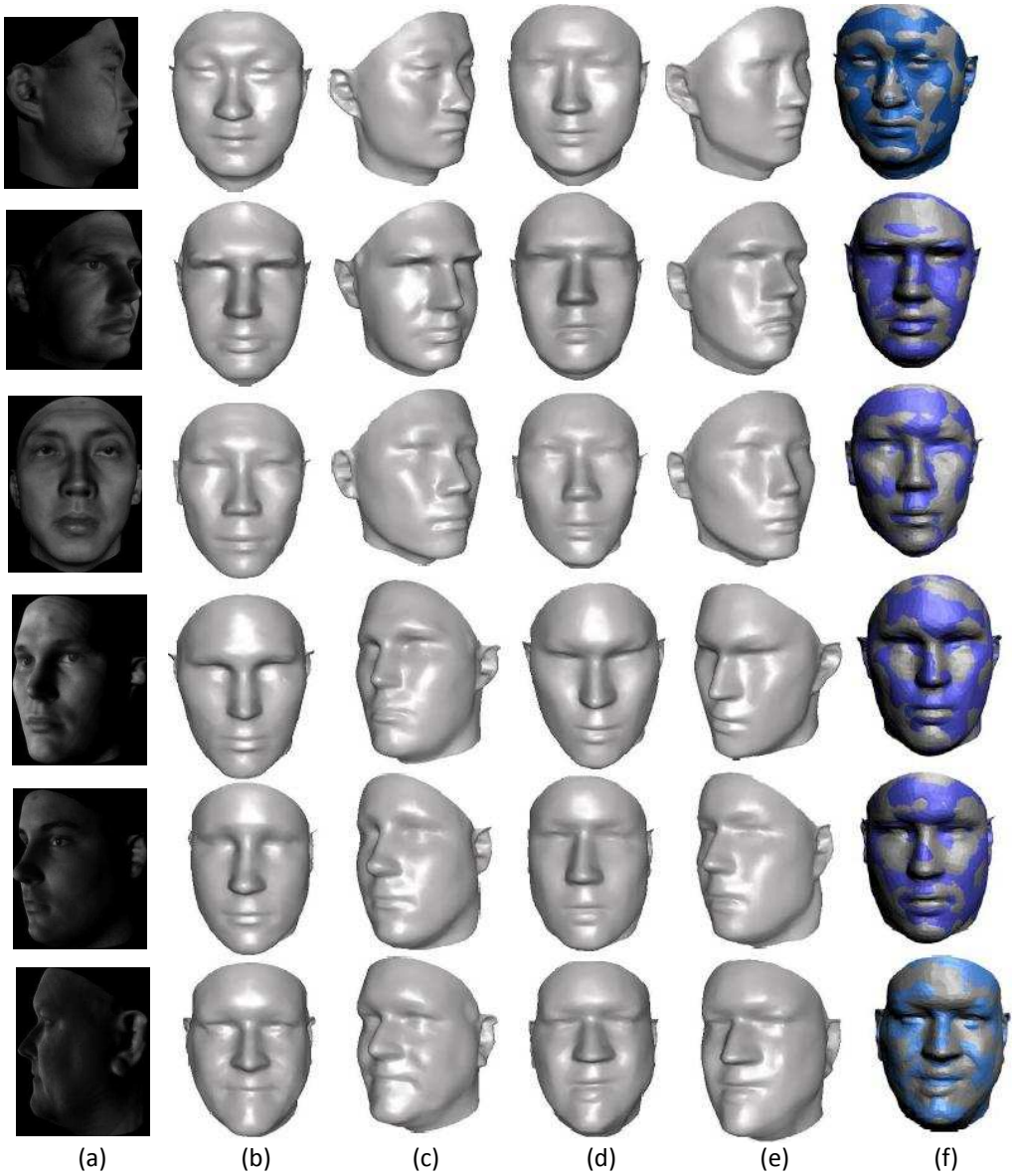

Figure 5. Results of shape recovery for the synthetic facial images. (a) the input image rendered from 3D face scan database; (b,c) different views of the ground truth shape; (d,e) the frontal view and side view of the recovered 3D shape; (f) the aligned image of the ground truth shape (in blue) and the recovered shape (in gray), which is used for measuring the reconstruction accuracy.

by applying the method to various real images. The experimental results demonstrated that our method is robust to variation in pose, illumination and identity of individuals. Looking into the future, we would like to further evaluate the performance of our approach with more appropriate real training data. In addition, we plan to extend our approach to reconstruct the shapes of other objects, such as the human body.

Acknowledgements This research was performed as part of a contract with the Maryland Advanced Simulation, Training, and Innovation Center at the University of Maryland Medical Center, and supported in part by NSF grants HCC-0448185 and CPA-0811647, and Open Project of State Key Lab of CAD& CG, Zhejiang University (No.A0812).

References

[1] J. J. Atick, P. A. Griffin, and A. N. Redlich. Statistical approach to shape from shading: Reconstruction of 3d face surfaces from single 2d images. *Neural Computation*, 8:1321–1340, 1996.

[2] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[3] M. J. Brooks and B. K. P. Horn. *Shape from shading*. MIT Press, Cambridge, MA, USA, 1989.

Figure 6. Results of shape recovery for the real images. (a) the input image from CMU database; (b) the reconstruction from (a) using our approach; (c) The image of the same person as (a) in a different pose that was not used for the reconstruction; (d) the profile view of the reconstruction corresponding to the pose in (c).

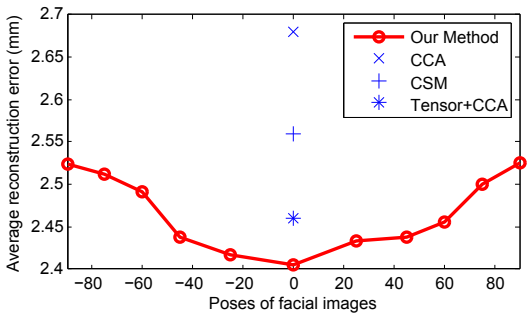

Figure 7. 3D reconstruction error vs. pose variation in our method. For comparison we also show the best reconstruction errors of CCA [16], CSM [4] and Tensor+CCA [11]. These methods can only deal with the frontal images of face without illumination variation. Note that these methods measure the reconstruction error with the specific training data set and testing data set.

[4] M. Castelan and E. R. Hancock. A simple coupled statistical model for 3d face shape recovery. In *ICPR*, pages 231–234, 2006.

[5] R. Dovgard and R. Basri. Statistical symmetric shape from shading for 3d structure recovery of faces. In *ECCV*, pages 108–116, 2004.

[6] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Generative models for recognition under variable pose and illumination. pages 277–284, 2000.

[7] L. Gu and T. Kanade. 3d alignment of face in a single image. *CVPR*, 1:1305–1312, June 2006.

[8] I. Kemelmacher and R. Basri. Molding face shapes by example. *ECCV*, pages 277–288, 2006.

[9] N. D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *NIPS*, 2004.

[10] N. D. Lawrence. Learning for larger datasets with the gaussian process latent variable model. In *AISTATS*, 2007.

[11] Z. Lei, Q. Bai, R. He, and S. Li. Face shape recovery from a single image using cca mapping between tensor spaces. *CVPR*, pages 1–7, June 2008.

[12] D. J. C. MacKay. Introduction to gaussian processes. In C. M. Bishop, editor, *Neural Networks and Machine Learning*, NATO ASI Series, pages 133–166. Kluwer Academic Press, 1998.

[13] A. Myronenko, X. Song, and M. A. . Carreira-Perpin.an. Non-rigid point set registration: Coherent point drift. In *Advances in Neural Information Processing Systems*, 2006.

[14] E. Prados, F. Camilli, and O. Faugeras. A unifying and rigorous shape from shading method adapted to realistic data and applications. *J. Math. Imaging Vis.*, 25(3):307–328, 2006.

[15] E. Prados and O. Faugeras. Shape from shading: a well-posed problem? *CVPR*, 2:870–877, June 2005.

[16] M. Reiter, R. Donner, G. Langs, and H. Bischof. 3d and infrared face reconstruction from rgb data using canonical correlation analysis. In *ICPR*, pages 425–428, 2006.

[17] T. Riklin-Raviv and A. Shashua. The quotient image: Class based recognition and synthesis under varying illumination conditions. In *CVPR*, pages 2566–, 1999.

[18] S. Romdhani and T. Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *CVPR*, pages 986–993, 2005.

[19] M. Salzmannn, R. Urtasun, and P. Fua. Local deformation models for monocular 3d shape recovery. *CVPR*, June 2008.

[20] I. Shimshoni, Y. Moses, and M. Lindenbaum. Shape reconstruction of 3d bilaterally symmetric surfaces. *International Journal of Computer Vision*, 39:97–110, 2000.

[21] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression (pie) database of human faces. Technical Report CMU-RI-TR-01-02, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, January 2001.

[22] T. Sim and T. Kanade. Combining models and exemplars for face recognition: An illuminating example. In *Proceedings of the CVPR 2001 Workshop on Models versus Exemplars in Computer Vision*, December 2001.

[23] W. A. P. Smith and E. R. Hancock. Recovering facial shape and albedo using a statistical model of surface normal direction. In *ICCV*, pages 588–595, 2005.

[24] R. W. Sumner and J. Popović. Deformation transfer for triangle meshes. In *SIGGRAPH '04: ACM SIGGRAPH 2004 Papers*, pages 399–405, 2004.

[25] C. Wang and S. Mahadevan. Manifold alignment using procrustes analysis. In *ICML*, pages 1120–1127, 2008.

[26] H. Wang, L. Dong, J. O'Daniel, R. Mohan, A. S. Garden, K. K. Ang, D. A. Kuban, M. Bonnen, J. Y. Chang, and R. Cheung. Validation of an accelerated 'demons' algorithm for deformable image registration in radiation therapy. *Physics in Medicine and Biology*, 50:2887–2905, 2005.

[27] X. Wang, X. Huang, J. Gao, and R. Yang. Illumination and person-insensitive head pose estimation using distance metric learning. In *ECCV*, pages 624–637, 2008.

[28] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah. Shape from shading: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):690–706, 1999.

[29] W. Y. Zhao and R. Chellappa. Symmetric shape-from-shading using self-ratio image. *International Journal of Computer Vision*, 45(1):55–75, 2001.