Face Shape Recovery from a Single Image Using CCA Mapping between Tensor Spaces

Zhen Lei Qinqun Bai Ran He Stan Z. Li
Center for Biometrics and Security Research & National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences,
95 Zhongguancun Donglu, Beijing 100080, China.
{zlei,qxbai,rhe,szli}@cbsr.ia.ac.cn

Abstract

In this paper, we propose a new approach for face shape recovery from a single image. A single near infrared (NIR) image is used as the input, and a mapping from the NIR tensor space to 3D tensor space, learned by using statistical learning, is used for the shape recovery. In the learning phase, the two tensor models are constructed for NIR and 3D images respectively, and a canonical correlation analysis (CCA) based multi-variate mapping from NIR to 3D faces is learned from a given training set of NIR-3D face pairs. In the reconstruction phase, given an NIR face image, the depth map is computed directly using the learned mapping with the help of tensor models. Experimental results are provided to evaluate the accuracy and speed of the method. The work provides a practical solution for reliable and fast shape recovery and modeling of 3D objects.

1. Introduction

Shape modeling of human faces has many practical applications, including computer graphics, computer games, human-machine interaction, and movie making. Two possible ways of face shape recovery are depth scanner-based and image-based. While the former is costly, we investigate into an approach for fast, reliable yet cost-effective solution to face shape recovery from a single image.

Recovering object shapes from a single image is a classic problem in computer vision. One popular approach is shape from shading (SFS) [7] (see also a survey [21]), which aims to invert the mapping from surface shape to image intensity by exploiting relationship between surface geometry and image formation. The SFS approach usually recovers object surface in two steps: (1) computing the surface orientation map, such as the normal direction or gradient field, from an intensity image, and (2) reconstructing the surface depth map from the orientation map.

However, SFS is an ill-posed problem because there are more unknown variables than equations. Several algorithms have been proposed to impose additional prior constraints, such as smoothness and integrability, to make the problem well-posed, and minimization, propagation or local techniques are used to find solutions [21]. However, the reliability of SFS solutions remains a problem, and SFS is still an active area of research.

For the 3D object of human face, algorithms are proposed to enhance the recovery accuracy by restricting an SFS method to a particular class of objects using subspace techniques and/or other constraints. Atick, Griffin and Redlich [1] model the problem of extracting SFS as parameter estimation in a low dimensional space, in which an ensemble of laser-scanned 3D heads are used to derive the PCA parameters of head shapes. Zhao and Chellappa [22] present a symmetric SFS (SSFS) approach to recover both shape and albedo for symmetric objects, in which a self-ratio image irradiance equation is introduced to facilitate the direct use of symmetry cue. In [5], Dovgard and Basri use a combination of the methods of [1] and [22] to recover the 3D shape of a human face using a single image. Smith and Hancock [16] fit a PCA model, trained using surface normal data acquired from range images, to intensity images of faces using constraints on the surface normal direction provided by Lambert's law. Kemelmacher and Basri [11] mold the face shape from a single non-frontal lighted image with the global face normal constraints. All these methods need to estimate the light source and reflectance properties of the surface simultaneously with shape which makes the problem more difficult.

Blanz and Vetter [2] propose morphable face model (MFM) method. It produces 3D face model by fitting one intensity face image to a pre-built statistical models of face shape and texture. Romdhani and Vetter [15] accelerate the fitting process and improve the result by combining multiple features. Wang *et al.* [20] modify the MFM process to

recover the face normal instead of depth. Hu, Jiang *et al.* [9, 10] simplify the method [2] and introduce a approach for 3D face reconstruction from a single frontal face image with homogeneous illumination and neutral expression. The 3D shape is recovered by the correspondence between the 2D-3D fiducial feature points using the morphable face model. However, in general, the MFM based methods are time-consuming and prone to local solutions.

Recently, machine learning methods have been successfully used in computer vision areas. It can also be utilized to solve the shape recovery problem. Reiter *et al.* [14] and Mario *et al.* [4] apply canonical correlation analysis (CCA) and coupled statistical model (CSM) respectively to recover 3D face shape from a 2D color image. However, in their methods, both the 2D and 3D images are transformed into vectors, which ignores the image intrinsic structure and the derived vector is usually of high dimension, which are likely to bring about the curse of dimensionality problem due to the limited training data.

In this paper, we develop a fast, reliable and cost-effective approach for face shape reconstruction from a single near infrared (NIR) image. The contributions of this paper include:

- We use the NIR images rather than visual light images. The NIR images are captured using a camera with active NIR LED illumination mounted on it [12]. Near infrared is invisible to human eyes and hence the active illumination causes no disturbance to the human. Such an image provides front-illuminated face image, and the pixel intensities are proportional to the normal component in the viewing direction and subject to albedo variation [12]. The NIR image is much less sensitive to environment lighting variation than the visual light one and therefore can be used for shape modeling more robustly and is more practical in real world application.
- Different from [14, 4], in this paper, we propose tensor models [17, 18] to formulate the NIR and 3D face image ensembles. In tensor modeling, the images are divided into small overlapped regions and are not transformed into vectors which therefore retain the intrinsic image structure and maintain high-order statistical information for modeling. Additionally, the tensor modeling can also efficiently avoid the curse of dimensionality problem due to its much lower dimension for every mode.
- After the construction of tensor models, we propose canonical correlation analysis (CCA) based method to learn the mapping from NIR to 3D tensor space. Experimental results show that our proposed method is fast and significantly reduces the reconstruction errors compared to the existing methods [14, 4].

The rest of the paper is organized as follows. Section 2 introduces the tensor fundamentals and details the tensor modeling of NIR and 3D images. The principle and process of CCA based mapping is detailed in Section 3. Section 4 describes the procedure of shape recovery from NIR images. Experimental results in terms of visual effect as well as quantitative accuracy are demonstrated in Section 5 and in Section 6, we conclude the paper.

2. Tensor models for NIR and 3D spaces

2.1. Tensor algebra fundamentals

We first review the tensor definition and some terminology on tensor operations [17, 18]. For clarity, in this paper, we denote scalars by lower case letters, vectors by bold lower case letters, matrices by bold upper-case letters, and higher-order tensors by calligraphic upper-case letters. Tensor, also known as a multidimensional array or n-way array, is a high-order multidimensional extension of vector (1-order tensor) and matrix (2-order tensor). Let $\mathcal{A} \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_N}$ be a tensor over $\mathbb{R}$. The order of $\mathcal{A}$ is N. The jth dimension of $\mathcal{A}$ is m_j and an element of $\mathcal{A}$ is specified as $\mathcal{A}_{i_1 i_2 \ldots i_n}$. The inner product of two tensors $\mathcal{A} \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_N}$ and $\mathcal{B} \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_N}$ with the same dimensions is defined as

$$< \mathcal{A}, \mathcal{B} >= \sum_{i_1=1,\ldots,i_N=1}^{i_1=m_1,\ldots,i_N=m_N} \mathcal{A}_{i_1,\ldots,i_N} \mathcal{B}_{i_1,\ldots,i_N} \quad (1)$$

The norm of a tensor $\mathcal{A}$ is $||\mathcal{A}|| = \sqrt{< \mathcal{A}, \mathcal{A} >}$, and the distance between two tensors $\mathcal{A}$ and $\mathcal{B}$ is $||\mathcal{A} - \mathcal{B}||$.

Specifically, the product of a tensor and a matrix is extended from the product of two matrices. The mode-k product of a tensor $\mathcal{A} \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_k \times \cdots \times m_N}$ by a matrix $\mathbf{M} \in \mathbb{R}^{m'_k \times m_k}$, denoted as $\mathcal{A} \times_k \mathbf{M}$, is a tensor $\mathcal{B} \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m'_k \times \cdots \times m_N}$ which is computed by

$$\mathcal{B}_{i_1,..,i_{k-1},j,i_{k+1},..,i_N} = \sum_{i=1}^{m_k} \mathcal{A}_{1,..,i_{k-1},i,i_{k+1},..,i_N} \times \mathbf{M}_{ji} \quad (2)$$

The mode-k product can also be expressed in terms of flattened matrices,

$$\mathbf{B}_{(k)} = \mathbf{M}\mathbf{A}_{(k)} \quad (3)$$

where $\mathbf{A}_{(k)}$ and $\mathbf{B}_{(k)}$ are mode-k flattening of tensor $\mathcal{A}$ and $\mathcal{B}$.

Moreover, as a natural generalization of matrix SVD, an alternative formulation of tensor decomposition named "N-mode SVD" [17] is defined as the mode-k product of N orthogonal spaces

$$\mathcal{D} = \mathcal{C} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \cdots \times_N \mathbf{U}_N \quad (4)$$

where $\mathcal{C}$ is known the core tensor, which governs the interaction between the mode matrices $\mathbf{U}_k$, for $k = 1, \ldots, N$.

And mode matrix $\mathbf{U}_k$ contains the orthogonal vectors spanning the column space of the mode-k flattening of D. This N-mode SVD expression can be derived by an iterative procedure of several conventional matrix SVDs.

2.2. Tensor models formulation

In traditional statistical learning involved with image, the image is usually represented by a vector. However, this transformation may disturb the high-order statistical information and spatial structure of the image. Moreover, the derived vector is usually of high-dimension, and hence brings about the curse of dimensionality problem. In this work, we try to solve these issues. We study methods to maintain the intrinsic image structure and avoid the curse of dimensionality. Motivated by this rational, we propose tensor models to formulate the NIR and 3D image ensembles. To keep the intrinsic image structure, we maintain the 2-order structure of an image and divided it into different small patches to avoid the curse of dimensionality. Therefore, two 4-order tensor models, describing the multi-factor (people, spatial positions, height and width of patches) are developed to formulate the NIR and 3D image ensembles respectively. Suppose we have n pairs of face images for training, each of which is divided into m overlapping patches and the numbers of rows and columns of each patch are h and w. The two 4-order tensors $\mathcal{D}_{NIR} \in \mathbb{R}^{n\times m\times h\times w}$ and $\mathcal{D}_{3D} \in \mathbb{R}^{n\times m\times h\times w}$ are naturally built by grouping all NIR and 3D patches sampled from the training faces respectively. By performing tensor decomposition (high-order extension of SVD) [18] on $\mathcal{D}_{NIR}$ and $\mathcal{D}_{3D}$, we have

$$\begin{aligned}\mathcal{D}_{NIR} &= \mathcal{C}_{NIR}\times_1\mathbf{U}_{people}\times_2\mathbf{U}_{positions}\times_3\mathbf{U}_{rows}\times_4\mathbf{U}_{columns}\\ &= \mathcal{T}_{NIR}\times_1\mathbf{U}_{people}\times_2\mathbf{U}_{positions}\\ \mathcal{D}_{3D} &= \mathcal{C}_{3D}\times_1\mathbf{V}_{people}\times_2\mathbf{V}_{positions}\times_3\mathbf{V}_{rows}\times_4\mathbf{V}_{columns}\\ &= \mathcal{T}_{3D}\times_1\mathbf{V}_{people}\times_2\mathbf{V}_{positions}\end{aligned} \tag{5}$$

where $\mathcal{C}_{NIR}$ is the core tensor that governs the interaction between 4 modes encoded in 4 orthogonal mode matrices in NIR sapce: $\mathbf{U}_{people} \in \mathbb{R}^{n\times n'_{NIR}}$, $\mathbf{U}_{positions} \in \mathbb{R}^{m\times m'_{NIR}}$, $\mathbf{U}_{rows} \in \mathbb{R}^{h\times h'_{NIR}}$, $\mathbf{U}_{columns} \in \mathbb{R}^{w\times w'_{NIR}}$, while $\mathcal{C}_{3D}$ is the 3D counterpart that governs the 4 orthogonal mode matrices in 3D space: $\mathbf{V}_{people} \in \mathbb{R}^{n\times n'_{3D}}$, $\mathbf{V}_{positions} \in \mathbb{R}^{m\times m'_{3D}}$, $\mathbf{V}_{rows} \in \mathbb{R}^{h\times h'_{3D}}$, $\mathbf{V}_{columns} \in \mathbb{R}^{w\times w'_{3D}}$, and T_{NIR}, T_{3D} are the tensor patches obtained by performing the mode product $\mathcal{C}_{NIR}\times_3\mathbf{U}_{rows}\times_4\mathbf{U}_{columns}$ and $\mathcal{C}_{3D}\times_3\mathbf{V}_{rows}\times_4\mathbf{V}_{columns}$. The notations $n'_{NIR}, m'_{NIR}, h'_{NIR}, w'_{NIR}, n'_{3D}, m'_{3D}, h'_{3D}, w'_{3D}$ denote the reduced dimensionality of the corresponding space where the eigenvectors associated with the smallest eigenvalues are truncated as the noise components. Significantly, the two factors (people, positions), encoded in row vector spaces of mode matrices $\mathbf{U}_{people}$, $\mathbf{V}_{people}$ and $\mathbf{U}_{positions}$, $\mathbf{V}_{positions}$, are crucial to determine a certain patch. For a patch pair (x_j, y_j) residing in the j-th spatial position of face images, their tensor representations can be derived as

$$\begin{aligned}\mathcal{P}(x_j) &= \mathcal{T}_{NIR}\times_1\mathbf{u}^T_{people}\times_2\mathbf{u}^{jT}_{position} = \mathcal{A}^j_x\times_1\mathbf{u}^T_{people}\\ \mathcal{P}(y_j) &= \mathcal{T}_{3D}\times_1\mathbf{v}^T_{people}\times_2\mathbf{v}^{jT}_{position} = \mathcal{A}^j_y\times_1\mathbf{v}^T_{people}\end{aligned} \tag{6}$$

where $\mathbf{u}^{jT}_{position}$ and $\mathbf{v}^{jT}_{position}$ are the j-th row vectors of the mode matrix $\mathbf{U}_{positions}$ and $\mathbf{V}_{positions}$, respectively. Both $\mathcal{A}^j_x = \mathcal{T}_{NIR}\times_2\mathbf{u}^{jT}_{position}$ and $\mathcal{A}^j_y = \mathcal{T}_{3D}\times_2\mathbf{v}^{jT}_{position}$ are constant tensors for the j-th patch pair. The people parameter vectors $\mathbf{u}_{people} \in \mathbb{R}^{n'_{NIR}\times 1}, \mathbf{v}_{people} \in \mathbb{R}^{n'_{3D}\times 1}$ which depict the individual characteristics need to be solved.

By mode-1 flattening Equ. 6, we have

$$\begin{aligned}\mathbf{x}_j &= (f_1(\mathcal{A}^j_x))^T\mathbf{u}_{people} = \mathbf{A}^{jT}_x\mathbf{u}_{people}\\ \mathbf{y}_j &= (f_1(\mathcal{A}^j_y))^T\mathbf{v}_{people} = \mathbf{A}^{jT}_y\mathbf{v}_{people}\end{aligned} \tag{7}$$

where $\mathbf{A}^j_x \in \mathbb{R}^{n'_{NIR}\times hw}, \mathbf{A}^j_y \in \mathbb{R}^{n'_{3D}\times hw}$ are position-dependent flattening matrices. It is obvious all patch pairs $\{x_j, y_j\}^m_{j=1}$ from the same person should share the same people parameter vectors $\mathbf{u}_{people}$ and $\mathbf{v}_{people}$. By defining a concatenated NIR feature vector $\mathbf{x} = [\mathbf{x}^T_1, \ldots, \mathbf{x}^T_m]^T \in \mathbb{R}^{mhw\times 1}$ and its depth counterpart $\mathbf{y} = [\mathbf{y}^T_1, \ldots, \mathbf{y}^T_m]^T \in \mathbb{R}^{mhw\times 1}$ and the enlarged matrices $\mathbf{A}_x = [\mathbf{A}^1_x, \ldots, \mathbf{A}^m_x] \in \mathbb{R}^{n'_{NIR}\times mhw}$, $\mathbf{A}_y = [\mathbf{A}^1_y, \ldots, \mathbf{A}^m_y] \in \mathbb{R}^{n'_{3D}\times mhw}$, we have

$$\mathbf{x} = \mathbf{A}^T_x\mathbf{u}_{people} \tag{8}$$

$$\mathbf{y} = \mathbf{A}^T_y\mathbf{v}_{people} \tag{9}$$

The parameter vectors $\mathbf{u}_{people}, \mathbf{v}_{people}$ can then be solved in the least squares sense:

$$\mathbf{u}_{people} = (\mathbf{A}_x\mathbf{A}^T_x)^{-1}\mathbf{A}_x\mathbf{x} \tag{10}$$

$$\mathbf{v}_{people} = (\mathbf{A}_y\mathbf{A}^T_y)^{-1}\mathbf{A}_y\mathbf{y} \tag{11}$$

Note that generally $mhw >> n'_{NIR}, n'_{3D}$, so the solutions of the above equations are commonly available. Because each NIR and 3D depth pair is acquired from the same person, the people parameter vectors $\mathbf{u}_{people}, \mathbf{v}_{people}$ should have strong correlative relationship. On the other hand, there may also exist some noise and redundant information among these vectors. So it is not the best way to learn the mapping between them directly. In this paper, we propose CCA-based mapping formulated in the next part to build the relationship between NIR and 3D image spaces.

3. CCA based mapping

In this part, we concentrate on exploring the relationship between the people spaces of NIR and 3D tensor models. Specifically, we want to predict the parameter vector in 3D people space from the input one in NIR space. It is known

that not all the component variables in the parameter vector have the same contribution to the mapping task and there exists redundance and noise among them which may even have negative effects for the mapping. Therefore, we first apply CCA approach on two spaces to find the most correlative and complementary factors and then build the mapping based on them.

Canonical correlation analysis (CCA) is a powerful tool for correlating two sets of multi-variate measurements in their leading factor subspaces [3]. Suppose the training data pairs $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n], \mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$[1]. The leading factor subspaces are the linear subspaces of the training data sets $\mathbf{X}$ and $\mathbf{Y}$, both of a reduced dimensionality d. CCA takes into account the two data sets simultaneously and finds the optimal linear projective matrices, also called canonical projection pairs, $\mathbf{W}^x = [\mathbf{w}_1^x, \mathbf{w}_2^x, \cdots, \mathbf{w}_d^x]$ and $\mathbf{W}^y = [\mathbf{w}_1^y, \mathbf{w}_2^y, \cdots, \mathbf{w}_d^y]$, from the corresponding data $\{\mathbf{X}, \mathbf{Y}\}$, such that $\mathbf{x}'_i = \mathbf{X}^T\mathbf{w}_i^x$ and $\mathbf{y}'_i = \mathbf{Y}^T\mathbf{w}_i^y$ are most correlated. This is done by maximizing the following correlation

$$\begin{aligned} \rho(\mathbf{w}_i^x, \mathbf{w}_i^y) &= \frac{E[\mathbf{x}_i'^T\mathbf{y}'_i]}{\sqrt{E[||\mathbf{x}'_i||^2]E[||\mathbf{y}'_i||^2]}} \\ &= \frac{\mathbf{w}_i^{xT}\mathbf{C}_{xy}\mathbf{w}_i^y}{\sqrt{\mathbf{w}_i^{xT}\mathbf{C}_{xx}\mathbf{w}_i^x\mathbf{w}_i^{yT}\mathbf{C}_{yy}\mathbf{w}_i^y}} \\ \text{subject to} \quad & \rho(\mathbf{w}_j^x, \mathbf{w}_i^y) = \rho(\mathbf{w}_i^x, \mathbf{w}_j^y) = 0 \\ & \text{for} \quad j = 1, \cdots, i-1 \end{aligned} \tag{12}$$

where $\mathbf{C}_{xy}$, $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are the correlation matrices computed from the training data sets $\mathbf{X}$ and $\mathbf{Y}$. Let

$$\mathbf{A} = \begin{pmatrix} 0 & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & 0 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{C}_{xx} & 0 \\ 0 & \mathbf{C}_{yy} \end{pmatrix} \tag{13}$$

It can be shown [13] that the solution $\mathbf{W} = (\mathbf{W}^{xT}, \mathbf{W}^{yT})^T$ amounts to the extremum points of the *Rayleigh quotient*:

$$r = \frac{\mathbf{W}^T\mathbf{A}\mathbf{W}}{\mathbf{W}^T\mathbf{B}\mathbf{W}} \tag{14}$$

The solution $\mathbf{W}^x$ and $\mathbf{W}^y$ can then be obtained by solving the generalized eigenproblem:

$$\mathbf{A}\mathbf{W} = \mathbf{B}\mathbf{W}\Lambda \tag{15}$$

After performing CCA on two data sets, we can extract the most correlative component pairs from the original data. Denote sample pairs from the data sets by random vectors $\mathbf{x}$ and $\mathbf{y}$. Let $\tilde{\mathbf{x}} = \mathbf{W}^{xT}\mathbf{x}$, where $\mathbf{W}^x$ is the CCA transformation matrix, so $\tilde{\mathbf{x}}$ is the most correlative components of $\mathbf{x}$ to $\mathbf{y}$. At the next step, our purpose is to learn the relationship between $\tilde{\mathbf{x}}$ and $\mathbf{y}$. Specifically, we assume the variables $\mathbf{y}$ and $\tilde{\mathbf{x}}$ have a linear relationship as

$$\mathbf{y} = \mathbf{R}\tilde{\mathbf{x}} + \epsilon \tag{16}$$

[1]The sample pairs in CCA are the parameter vectors in NIR and 3D people spaces. The notations can be deduced from the context and will not be confused.

where ϵ is the noise item which obeys the Gaussian distribution, $\epsilon \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I})$, where $\mathbf{I}$ is the identity matrix. Thus, we have

$$P(\mathbf{y}|\tilde{\mathbf{x}}, \mathbf{R}) = \frac{1}{Z}\exp\{-\frac{(\mathbf{y} - \mathbf{R}\tilde{\mathbf{x}})^T(\mathbf{y} - \mathbf{R}\tilde{\mathbf{x}})}{2\sigma^2}\} \tag{17}$$

where Z is a normalization coefficient. By maximizing the log-likelihood in the training set with respect to $\mathbf{R}$, we have

$$\begin{aligned} \mathbf{R}^* &= \arg\max_{\mathbf{R}}\{-\frac{1}{2\sigma^2}\sum_i(\mathbf{y}_i - \mathbf{R}\tilde{\mathbf{x}}_i)^T(\mathbf{y}_i - \mathbf{R}\tilde{\mathbf{x}}_i)\} \\ &= \arg\min_{\mathbf{R}} tr((\mathbf{Y} - \mathbf{R}\tilde{\mathbf{X}})(\mathbf{Y} - \mathbf{R}\tilde{\mathbf{X}})^T) \end{aligned} \tag{18}$$

and we can get the solution by putting the derivative of objective function w.r.t. $\mathbf{R}$ to zero as

$$\mathbf{R}^* = \mathbf{Y}\tilde{\mathbf{X}}^T(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T)^{-1} \tag{19}$$

Moreover, in order to improve the generalization of result, we can impose regularized penalty, also known the prior knowledge onto the log-likelihood in Equ. 18 as

$$\mathbf{R}^* = \arg\min_{\mathbf{R}} tr((\mathbf{Y} - \mathbf{R}\tilde{\mathbf{X}})(\mathbf{Y} - \mathbf{R}\tilde{\mathbf{X}})^T + \lambda\mathbf{R}\mathbf{R}^T) \tag{20}$$

where λ controls the trade-off between the accuracy in the training set and the generalization. We can then obtain the optimal result as

$$\mathbf{R}^* = \mathbf{Y}\tilde{\mathbf{X}}^T(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T + \lambda\mathbf{I})^{-1} \tag{21}$$

which is essentially equivalent to the ridge regression [6].

Given a new input vector $\mathbf{x}_{new}$, $\tilde{\mathbf{x}}_{new}$ is computed using CCA transformation matrix,

$$\tilde{\mathbf{x}}_{new} = \mathbf{W}^{xT}\mathbf{x}_{new} \tag{22}$$

and the prediction of the output vector is then obtained by

$$\mathbf{y}_{new} = \mathbf{R}^*\tilde{\mathbf{x}}_{new} \tag{23}$$

4. Shape Recovery from NIR

4.1. Data Processing

In the training phase, the NIR and 3D training sets should be properly prepared. This consists of the following steps:

- **Preprocessing**. The NIR images are taken using a commercially available NIR web-camera with NIR LED lights, where The LED lights are approximately co-axial to the lens direction. The 3D faces are acquired with a Minolta vivid 910 laser scanner. The laser scanner provides the depth of the visible parts of the faces which are actually 2.5D data. Regions not belonging to the face are discarded and then the 3D data is preprocessed by removing high noise and filling holes using an interpolation algorithm.

- **Face Alignment**. For each face image both in NIR and 3D, 68 landmark points are labeled manually. Note that this process is implemented automatically in the test phase using an DAM [8] model learned from the training data set. The NIR and 3D faces enclosed by the convex hull of the landmark points are denoted as $\mathbb{U}^0$ and $\mathbb{Z}^0$.
- **Warping**. The NIR and 3D faces $\mathbb{U}^0$ and $\mathbb{Z}^0$ are warped to an uniform shape in the image plane based on the landmark points, where the mean shape of NIR faces, bounded by the box of size 112×118, is used as the uniform shape. The warping operations are expressed as follows:

$$\mathbb{U}^0 \overset{warp}{\Longrightarrow} \mathbb{U}^w, \quad \mathbb{Z}^0 \overset{warp}{\Longrightarrow} \mathbb{Z}^w$$

and illustrated in Fig. 1. The deforming operations of the warping are nonlinear.

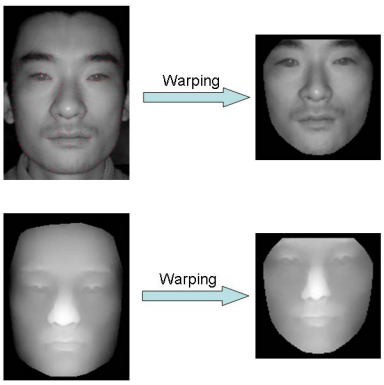

Figure 1. NIR and 3D face warping according to 68 landmark points

4.2. Shape Recovery from a Single NIR Image

In the reconstruction phase, shape recovery can be done following the procedure shown in Figure 2. The input is an NIR face image. The face is detected using an AdaBoost face detector [19], and the landmark points are located by a DAM model [8]. Face warping is then performed from the aligned shape to the uniform shape according to the located landmark points. The warping parameter W is memorized for later use of a reverse-warping. After that, the NIR face under uniform shape $\mathbb{U}^w$ is divided into m overlapped patches which are rearranged and projected into the NIR tensor space via Equ. 10 to get the personalized parameter vector in NIR people space. Then the CCA based mapping learned from the training set is utilized to predict the corresponding personalized vector in 3D tensor space. After that, the whole 3D face in the uniform shape $\mathbb{Z}^w$ is reconstructed using specific personalized vector with the help of 3D tensor model via Equ. 9 and rearrangement operation. In the final, the 3D face image in the true shape can be recovered by the reverse-warping of W

$$\mathbb{Z}^w \overset{W^{-1}}{\longrightarrow} \mathbb{Z}^0$$

A reliable and fast facial shape recovery system can be built using the present hardware and the proposed algorithm. The system consists of five modules: face detection, alignment, warping, mapping and reverse warping. The essential engine of the system is a mapping from NIR to 3D, just using some multilinear algebra operations, therefore the proposed algorithm is very fast and it can achieve reliable results as the experiment shows.

5. Experimental Results

In the experiments, 400 pairs of NIR images and 3D laser scans (Minolta vivid 910) of 200 persons, including male and female are collected, with 2 pairs per person. All the faces are without accessories, prominent makeup and facial hair. The database is divided into training set and testing set randomly. Training set contains 200 NIR-3D pairs of 100 persons while testing set includes the remaining 200 NIR-3D pairs of 100 persons. So the training set and the testing set have no intersection of persons and images either.

The quantitative accuracy of reconstruction result is evaluated in terms of the mean absolute error (MAE) defined as follows

$$e = \frac{1}{n}\sum_{i=1}^{n} |D_r(i) - D_t(i)| \tag{24}$$

where D_r is the reconstructed depth and D_t the ground-truth depth, and n is the total number of the effective facial points in the uniform shape.

In this experiment, for the proposed method, each image is divided into 14×14 overlapped patches with the size of 16×16. The dimensions of people spaces for NIR and 3D tensor models are retained 150 and 100 respectively for maintaining the 98% energy. The regularized coefficient λ in the CCA based mapping is set to 0.01 empirically. For ease of representation, our method is denoted as Tensor+CCA in the following experiments. We also implement two recently developed methods named CCA [14] in which the images of NIR and 3D are vectorized first and CCA is taken to establish the relationship between the two data sets, and CSM [4] where a simple coupled statistical model is constructed for 3D image inference to make a comparison with our proposed method.

Table 1 lists reconstruction results on 10 differently split test sets and Figure 3 illustrates the average reconstruction error on the 10 test sets with respect to the reduced dimension of CCA based mapping. It shows our proposed method (Tensor+CCA) achieves significantly better results compared to the CCA and CSM methods, which indicates the accuracy and effectiveness of the proposed method. From Figure 3, we can see the best result can be achieved by only remaining about 52 dimensions which reduces the computation cost and improves the reconstruction results simultaneously, and proves the effectiveness of the proposed CCA

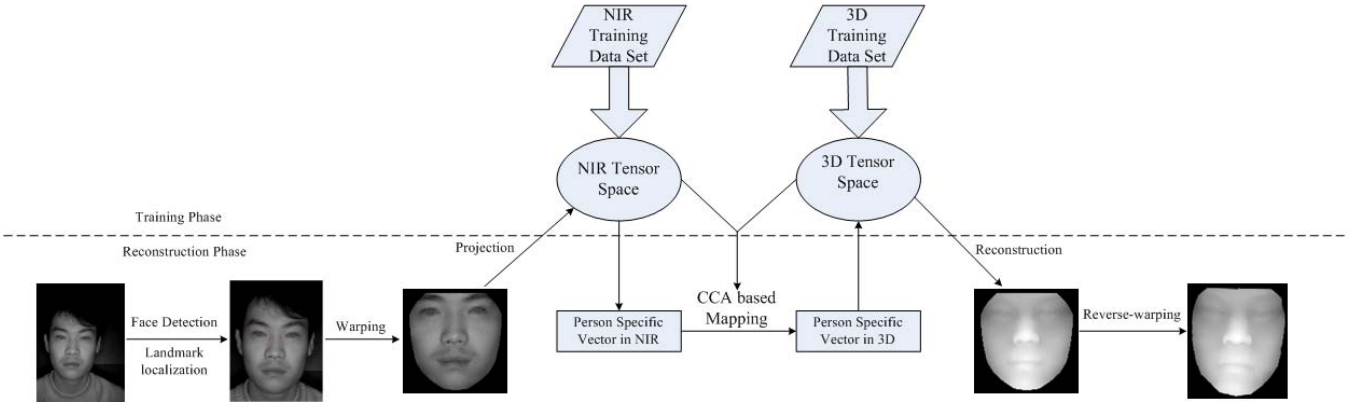

Figure 2. Reconstruction of a 3D face from a single NIR face image.

based mapping by exploiting the most correlative components. Furthermore, in our experiments, the results of CCA are obviously better than the results in [14], less than $1/2$ of reconstruction errors reported in their paper. It may be ascribed to the use NIR image rather than visual light image as the 2D input, which is much less sensitive to environment lighting variations. This reflects the superiority of NIR image over visible light image in practice.

Table 1. Reconstruction errors (mm) of different methods on different split test sets. (The number in bracket is the reduced dimension corresponding to the minimum error)

	1	2	3	4	5
CCA	2.73 (46)	2.68 (24)	2.90 (8)	2.79 (27)	2.78 (36)
CSM	2.78	2.56	2.85	2.75	2.82
Tensor+CCA	2.59 (24)	2.46 (50)	2.69 (40)	2.59 (30)	2.57 (67)
	6	7	8	9	10
CCA	2.85 (26)	2.68 (47)	2.79 (36)	2.75 (24)	2.80 (30)
CSM	2.88	2.76	2.75	2.88	2.83
Tensor+CCA	2.68 (31)	2.58 (95)	2.58 (40)	2.58 (38)	2.66 (19)

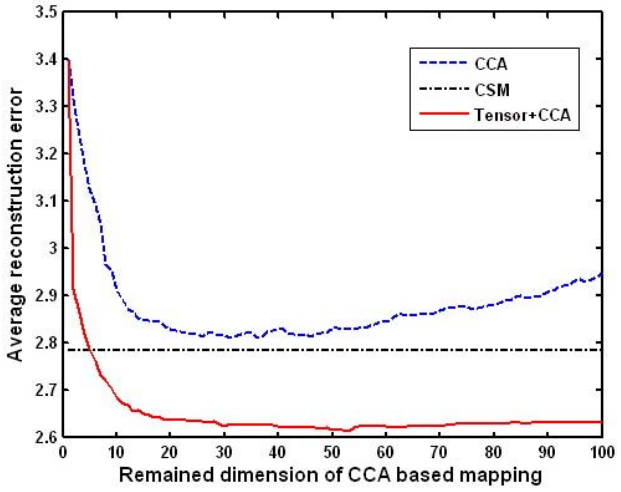

Figure 3. 3D reconstruction error curves of three methods.

Figure 4 shows some qualitative reconstruction results of the testing data out of the training set. There, the depth reconstruction result obtained by Tensor+CCA is compared with the ground truth data for each input NIR image. Column 1 is the input NIR image, and column 2-4 are the ground-truth depth illustrated from different views. The last three columns are the reconstructed results by the proposed method. The surface are colored from blue to red according to the depth. It shows the reconstructed results of Tensor+CCA approximate relatively well to the ground-truth.

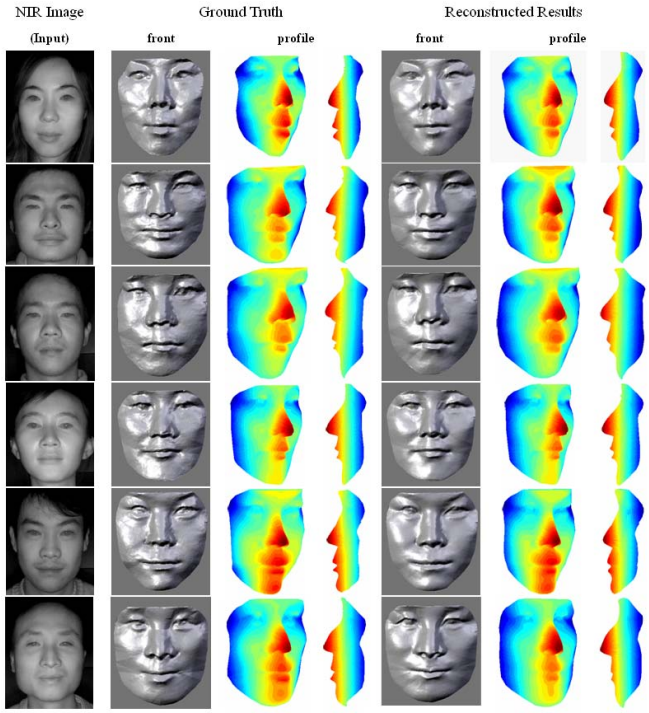

Figure 4. Shape recovery from a single NIR image by the proposed method.

Finally, regarding the reconstruction computation, the process of three main steps of the proposed method: warp-

ing, mapping, reverse warping, takes only about one second on average on a P4 3.0 GHz computer.

6. Conclusions

In this paper, we have proposed a NIR imaging and a statistical learning approach of tensor modeling with CCA based mapping for facial shape recovery from a single image. The key component is the tensor modeling of NIR and 3D spaces and a CCA based NIR to 3D mapping learned from a training set of NIR-3D pairs. Once the mapping is learned, the depth map can be reconstructed from a single NIR image with the help of tensor models analytically. The solution is reliable and accurate and is proved to be effective and efficient to exploit the relationship between NIR and 3D images. The future work will be to develop better mapping learning method and to train the model using a larger training set, so that the learned mapping can generalize to unseen faces better.

Acknowledgements

This work was supported by the following funds: Chinese National Natural Science Foundation Project #60518002, Chinese National 863 Program Projects #2006AA01Z192, #2006AA01Z193, and #2006AA780201-4, Chinese National Science and Technology Support Platform Project #2006BAK08B06, and Chinese Academy of Sciences 100 people project, and AuthenMetric R&D Funds.

References

[1] J. J. Atick, P. A. Griffin, and A. N. Redlich. "Statistical approach to shape from shading: Reconstruction of three-dimensional face surfaces from single two-dimensional images". *Neural Computation*, 8(6):1321–1340, 1996.

[2] V. Blanz and T. Vetter. "A morphable model for the synthesis of 3d faces". In *SIGGRAPH'99 Conference Proceedings*, pages 187–194, 1999.

[3] L. Breiman and J. Friedman. "Predicting multivariate responses in multiple linear regression". *Journal of the Royal Statistical Society*, 59(1):3–54, 1997.

[4] M. Castelan and E. R. Hancock. "A simple coupled statistical model for 3d face shape recovery". In *Proceedings of the 18th International Conference on Pattern Recognition 2006*, pages 231–234, 2006.

[5] R. Dovgard and R. Basr. "Statistical symmetric shape from shading for 3d structure recovery of faces". In *Proceedings of the European Conference on Computer Vision*, pages 108–116, 2004.

[6] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification, Second Edition*. John Wiley and Sons, 2001.

[7] B. K. P. Horn and M. J. Brooks, editors. *Shape from Shading*. MIT Press, Cambridge, MA, June 1989.

[8] X. W. Hou, S. Z. Li, and H. J. Zhang. "Direct appearance models". In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 828–833, Hawaii, December 11-13 2001.

[9] Y. Hu, D. Jiang, S. Yan, L. Zhang, and H. zhang. "Automatic 3d reconstruction for face recognition". In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 843–848, 2004.

[10] D. Jiang, Y. Hu, S. Yan, L. Zhang, H. Zhang, and W. Gao. "Efficient 3d reconstruction for face recognitionn". *Pattern Recognition*, 38(6):787–798, 2005.

[11] I. Kemelmacher and R. Basri". "Molding face shapes by example". *"European Conference on Computer Vision"*, "2006".

[12] S. Z. Li, R. Chu, S. Liao, and L. Zhang. "Illumination invariant face recognition using near-infrared images". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):627–639, April 2007.

[13] T. Melzer, M. Reiter, and H. Bischof. "Appearance models based on kernel canonical correlation analysis". *Pattern Recognition*, 36(9):1961–1971, 2003.

[14] M. Reiter, R. Donner, L. Georg, and B. Horst. "3D and infrared face reconstruction from RGB data using canonical correlation analysiss". In *Proceedings of International Conference on Pattern Recognition*, 2006.

[15] S. Romdhani and T. Vetter. "Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior". In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition 2005*, pages 986–993, 2005.

[16] W. A. P. Smith and E. R. Hancock. "Recovering facial shape using a statistical model of surface normal direction". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1914–1930, 2006.

[17] M. A. O. Vasilescu and D. Terzopoulos. "Multilinear analysis of image ensembles: Tensorfaces". In *ECCV (1)*, pages 447–460, 2002.

[18] M. A. O. Vasilescu and D. Terzopoulos. "Multilinear subspace analysis of image ensembles". In *Computer Vision and Pattern Recognition (2)*, pages 93–99, 2003.

[19] P. Viola and M. Jones. "Robust real time object detection". In *IEEE ICCV Workshop on Statistical and Computational Theories of Vision*, Vancouver, Canada, July 13 2001.

[20] Y. Wang, Z. Liu, G. Hua, Z. Wen, Z. Zhang, and D. Samaras. "Face re-lighting from a single image under harsh lighting conditions". In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition 2007*, June 2007.

[21] 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.

[22] W. Zhao and R. Chellappa. "Symmetric shape-from-shading using self-ratio image". *International Journal of Computer Vision*, pages 55–75, 2001.