Single-view-based 3D facial reconstruction method robust against pose variations

Jaeik Jo a, Heeseung Choi b, Ig-Jae Kim b, Jaihie Kim a,⁎,1

a School of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749, Republic of Korea
b The Image Media Research Center, Korea Institute of Science and Technology, Seoul 136-130, Republic of Korea

Abstract

The 3D Morphable Model (3DMM) and the Structure from Motion (SfM) methods are widely used for 3D facial reconstruction from 2D single-view or multiple-view images. However, model-based methods suffer from disadvantages such as high computational costs and vulnerability to local minima and head pose variations. The SfM-based methods require multiple facial images in various poses. To overcome these disadvantages, we propose a single-view-based 3D facial reconstruction method that is person-specific and robust to pose variations. Our proposed method combines the simplified 3DMM and the SfM methods. First, 2D initial frontal Facial Feature Points (FFPs) are estimated from a preliminary 3D facial image that is reconstructed by the simplified 3DMM. Second, a bilateral symmetric facial image and its corresponding FFPs are obtained from the original side-view image and corresponding FFPs by using the mirroring technique. Finally, a more accurate 3D facial shape is reconstructed by the SfM using the frontal, original, and bilateral symmetric FFPs. We evaluated the proposed method using facial images in 35 different poses. The reconstructed facial images and the ground-truth 3D facial shapes obtained from the scanner were compared. The proposed method proved more robust to pose variations than 3DMM. The average 3D Root Mean Square Error (RMSE) between the reconstructed and ground-truth 3D faces was less than 2.6 mm when 2D FFPs were manually annotated, and less than 3.5 mm when automatically annotated.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Researchers have focused considerable attention on 3D facial reconstruction technologies because they are useful in various applications such as pose-invariant facial recognition [1,2,38], age-invariant facial recognition [3], 3D person-specific game and movie-character generation [4,5], frontal facial synthesis of subjects in surveillance camera systems, teleconferencing, surgical simulation, etc.

These 3D facial reconstruction technologies can be divided into two approaches. The first is a hardware-based approach that reconstructs a 3D face by using additional hardware such as a stereographic camera [39], structured light [40], depth sensors [41], or a 3D laser scanner [6]. This method can be used to obtain accurate 3D facial data; however, it incurs additional costs and requires image preprocessing and camera calibration. To address these issues, a monocular camera-based approach was proposed to reconstruct the 3D face. This approach can be categorized into single-view and multiple-view-based approaches. The representative methods of these approaches are the Structure from Motion (SfM) and the 3D Morphable Model (3DMM), respectively. Shape-From-Shading (SFS) is a method to reconstruct a 3D face from the brightness variations in a single image. Although considerable research has been done on SFS-based methods, they have impractical constraints because the Lambertian reflectance model and a known light source direction need to be assumed to produce accurate results [32–35,42–44].

The multiple-view-based SfM method calculates the 3D facial shape and projection matrix using the corresponding 2D Facial Feature Points (FFPs), which are extracted from 2D facial images captured from various angles [8,9,13–16,22–24,26,45]. The 2D FFPs can be factorized into the 3D facial shape and projection matrix using Singular Value Decomposition (SVD) and the rank theorem [17]. SfM can accurately reconstruct a user-specific 3D face because more 2D facial information is available than the single-view-based approach. Moreover, a training process is not required, and the reconstructed face is not biased toward the 3D mean face model. However, SfM requires multiple-view images from various angles and the corresponding points between these facial images have to be located.
The single-view-based approach using the 3D model builds the 3D face model in a training process to represent facial shape, texture, illumination, and camera geometry with a number of model parameters [1,7,29–30,36–37,46]. Given a 2D facial image, the model-based method continuously optimizes 3D facial model parameters to minimize the shape and texture residuals between the 2D facial image input and a 2D facial image synthesized from model parameters. The final 3D face can be reconstructed from the obtained optimal parameters. This method enables reconstruction of a 3D face from a single facial image; moreover, it can work relatively well under various constraints. However, this method has some limitations. For one, it incurs high computational costs because the process identifies the numerous parameters through an iterative parameter optimization process [8,9]. In addition, the reconstructed 3D face can be biased toward a 3D mean face from the training data. Furthermore, the reconstruction accuracy decreases as the rotation angle of the input face increases. This decrease in accuracy is because of fitting errors caused by occluded FFPs due to head rotation.

To reduce computational complexity, which is the primary limitation of the 3DMM method, the Simplified 3DMM (S3DMM) method was proposed [3,10–12,19–21,25]. Unlike 3DMM, which reconstructs 3D facial shape, texture, and illumination, S3DMM reconstructs only a 3D facial shape. Additionally, S3DMM uses a sparse shape composed of dozens of vertices, whereas 3DMM uses a dense shape comprised of thousands of vertices. Therefore, the S3DMM method is not required to search dense FFPs; consequently, it has low computational complexity due to the decreased number of model parameters. One disadvantage of S3DMM is that it is limited by the constraint that its input facial image should be a frontal view image [11,19]. Although previous research methods [3,20,21] have reconstructed a 3D face from arbitrary views, the work remains insufficiently robust to pose variations due to vulnerability to self-occlusion errors.

For example, when a face is rotated, one side of the face is self-occluded by the other side; the FFPs in the occluded region are invisible. Therefore, the located FFPs in the self-occluded region always include location errors. This decreases the accuracy of 3D facial reconstruction by S3DMM. Lee et al. [25] decreased errors due to self-occlusion by excluding FFPs detected in the occluded region, and by using only residual FFPs that were visible after detection. However, their approach did not resolve the problem of decreased 3D reconstruction performance arising from pose variations. Many fitting errors are generated in the process of fitting the residual FFPs to a 3D model because the number of occluded FFPs to be excluded increase as the face rotation angle increases.

To address the issues mentioned above, we propose a novel single-view-based 3D facial reconstruction method. In our proposed method, the 3D facial shape is reconstructed by SfM using mirrored side-view image and frontal face image generated via bilateral symmetry and S3DMM. The primary contributions of the proposed method are as follows. First, the proposed method is robust to pose variations, unlike previous model-based methods that are vulnerable to pose variations because of fitting errors caused by occluded FFPs due to head rotation. In the proposed method, the occluded FFPs in the original view-image are revealed in the different view-images generated by bilateral symmetry and S3DMM. Subsequently, these corresponding 2D FFPs are used to calculate the 3D facial shape using SfM. Thus, the proposed method is more robust to pose variations than the previous model-based methods. Second, the proposed method is person-specific and not biased toward a mean face. The reconstruction results of the previous model-based methods, which are obtained by optimizing the parameters from the 3D mean face, can be biased toward the 3D mean face when parameter optimization falls into local minima. In contrast, the reconstruction results of the proposed method is person-specific because reconstruction is done by calculating the 3D facial shape using 2D FFPs in the different view-images of the corresponding subject. Table 1 provides a comparison of the advantages and disadvantages of our proposed method and previous 3D facial reconstruction methods.

The rest of this paper is organized as follows. In Section 2, the proposed 3D facial reconstruction method is described. In Section 3, our experimental environment and results are outlined. Conclusions and future work are presented in Section 4.

### 2. Proposed method

The proposed method is composed of several steps, which include FFP extraction, head pose estimation, occlusion point compensation, 3D facial shape reconstruction by combining S3DMM and SfM methods, dense 3D facial shape reconstruction, and texture mapping. A schematic diagram and flowchart of the proposed method are presented in Figs. 1 and 2, respectively.

<table>
<thead>
<tr>
<th>Table 1 Comparison of proposed method and previous methods.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overview</strong></td>
</tr>
<tr>
<td><strong>Strength</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Weakness</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
First, FFPs are extracted by automatic or manual annotation. The head pose is then estimated using a cylinder head model \[18\]. If the estimated yaw angle is greater than $\pm 30^\circ$ or less than $-30^\circ$, the locations of extracted FFPs are compensated because it is difficult to accurately localize the FFPs in the occluded region due to rotation. The compensation is performed by applying a Shape Conversion Matrix (SCM) \[26\] to decrease location errors. Next, a 2D frontal face is estimated from an initial 3D face that has been reconstructed using S3DMM. Based on the assumption that a human face has bilateral symmetry, a mirrored facial image and its FFPs are generated from the original facial image. A more accurate 3D facial shape is then reconstructed by SfM using the frontal, original, and symmetric FFPs obtained in the previous step. If the estimated yaw angle is within the predetermined angle ($\pm 30^\circ$), then the 3D facial shape is reconstructed by S3DMM using only the original face image as in the usual manner. In this case, the 3D face reconstructed by SfM is less accurate because differences between the corresponding points of the frontal, original, and symmetric FFPs are small when the yaw angle is small. The reconstructed sparse 3D shapes are then converted into dense 3D shapes by a dense 3D mean model adaptation. Finally, the 3D face is reconstructed after applying texture mapping. These processes are described in details as follows.
2.1. Facial feature point extraction

For facial feature point extraction, 80 FFPs are extracted from facial images in various poses, as shown in Fig. 3. These FFPs are extracted automatically using Active Appearance Model (AAM) [31], and manually. The AAM is one of the most widely used automatic FFPs annotation methods [47–52]. The AAM based on the simultaneous inverse compositional algorithm is employed in our work, because this algorithm showed a relatively good performance for occlusion [61]. While it is true that there are many FFPs extraction methods [53–60], none of them are error-free. When the 3D face is reconstructed using FFPs obtained by the automatic FFPs extraction method, it is unclear whether the finally measured 3D reconstruction error is caused by the FFPs extraction error or by the 3D reconstruction error. Therefore, we also manually extracted FFPs, which can be regarded as ground truth FFPs, to minimize the influence of error from FFPs extraction algorithms.

Automatic FFPs extraction method such as AAM is generally used in practical applications. However, some applications prefer high reconstruction accuracy over processing time and manual FFPs extraction can be favored for such applications (3D game, movie, plastic surgery, surveillance). For example, synthesis of frontal face of a suspect in surveillance requires a high quality 3D face reconstruction result and manually annotated FFPs are used with the help of the user.

2.2. Pose estimation

As shown in Fig. 2, head pose estimation is required to decide which of the two 3D face reconstruction methods should be used. This is because the choice of the reconstruction method depends on the variation in the rotation of the head. In this paper, the cylindrical head model method [18] proposed by Ohue et al. is used. Ohue et al. used the cylindrical head model to estimate the direction of a driver’s head in a real-time pre-crash safety system. This method is based on the assumption that a human head is generally cylindrical shape, as shown in Fig. 4. The head pose is estimated by using three facial points: the right facial edge \((x_R)\), left facial edge \((x_L)\), and center \((x_C)\). The yaw angle \((\theta)\) of the head pose is calculated by using the following equation:

\[
\theta = \arcsin \left( \frac{(x_L + x_R/2) - x_C}{(x_L - x_R/2) + x_C} \right)
\]

2.3. Occlusion point compensation

When a face is rotated, one side of the face is self-occluded by the inverse region. As shown in Fig. 5, some FFPs (such as the eyes, eyebrows, nose, and facial contours) in the self-occluded region are invisible. Therefore, the detected FFPs in the self-occluded region always have some location errors. This decreases the accuracy of 3D facial reconstruction by S3DMM.

By employing SCM [26] to compensate for the location errors of self-occluded FFPs, the accuracy of FFP extraction and 3D facial reconstruction is improved. First, 3D faces are obtained by a 3D scanner [27]. Ground-truth 3D FFPs are set through manual annotation, and then all 3D faces are aligned. The ground-truth 2D FFPs are obtained by rotating and projecting the ground-truth FFPs.
3D FFPs onto the aligned 3D faces. The 2D facial shapes (2D FFPs) are acquired through manual or automatic annotation (AAM) on the projected 2D images from the 3D faces. The ground-truth 2D FFPs and the observed 2D FFPs obtained by annotation can be represented by the following equations, respectively:

\[ G_{v}^{2D} = A_0 + \sum_{i = 1}^{n} \alpha_v A_i \]

\[ O_{v}^{2D} = B_0 + \sum_{i = 1}^{m} \beta_v B_i \]

where \( G_{v}^{2D} \) is the ground-truth 2D facial shape of the \( v \)th projected image, and \( O_{v}^{2D} \) is the observed facial shape of \( G_{v}^{2D} \). \( C_{v}^{2D} \) and \( O_{v}^{2D} \) are \( 2N \times 1 \) matrices composed of the \( x \) and \( y \) positions of \( N \) FFPs. \( A_0 \) and \( B_0 \) are the respective mean and shape variations obtained by using the ground-truth 2D facial shapes in training data and the Principal Component Analysis (PCA). \( \alpha_v = [\alpha_1, \ldots, \alpha_m]^T \) is the ground-truth shape parameter of the \( v \)th projected image, and \( n \) is the dimension of the parameter. Similarly, \( B_0 \) and \( B_i \) are the respective mean and shape variations obtained by using the observed 2D facial shapes in training data and the PCA. \( \beta_v = [\beta_1, \ldots, \beta_m]^T \) is the ground-truth shape parameter of the \( v \)th projected image, and \( m \) is the dimension of the parameter.

Lee et al. [26] experimentally demonstrated that the observed shape parameters are linear to the ground-truth shape parameters in Eqs. (2) and (3). More detailed derivation is given in [26]. Therefore, these two shape parameters can be represented as given in Eq. (4).

\[ \alpha_v = C \beta_v \]

Finally, \( C \), which denotes the shape conversion matrix, can be calculated by the least square method as given in Eq. (5).

\[ C = [\alpha_1 \ldots \alpha_v][\beta_1 \ldots \beta_v]^T \left( \beta_1 \ldots \beta_v \right)^{-1} \left( \beta_1 \ldots \beta_v \right)^T \]

\( V \) is the total number of projected facial images in the training data. For example, given the 2D facial scans for 150 subjects and 7 facial pose images for each 3D face, \( V \) is calculated as 1050. FFPs are compensated by using the SCM calculated in Eq. (5). The SCM converts the observed 2D facial shapes into 2D facial shapes that are similar to the ground-truth 2D facial shapes. Self-occluded FFPs can be fairly accurately converted into the ground-truth FFPs by SCM, without the need for iterative calculations.

2.4. 3D facial shape reconstruction

The 3D facial shape is reconstructed from the extracted 2D FFPs obtained in the previous steps as follows. If the pose variation is not large, S3DMM is used to reconstruct the 3D facial shape. S3DMM is used in this case because the FFPs occluded because of head rotation are typically not generated; moreover, S3DMM fitting errors by the occluded FFPs rarely occur. On the other hand, if the pose variation is large, the S3DMM and SfM fusion method is used to reconstruct the 3D facial shape because of the large fitting errors due to occluded FFPs.

The proposed fusion method is described as follows. First, 2D frontal FFPs are estimated from an initial 3D face, which was temporarily reconstructed by S3DMM. Bilateral symmetric FFPs from the original FFPs are generated under the assumption that the human face exhibits bilateral symmetry. Finally, a more accurate 3D facial shape is reconstructed by SfM using the frontal, original, and symmetric FFPs obtained in the previous steps. Each process is described in detail below.

2.4.1. Bilateral symmetric face generation

To accurately reconstruct a 3D face, SfM requires various 2D pose images such as the frontal, left, and right facial images. To reconstruct a 3D face from only one input-rotated image, another rotated image should be generated. A bilateral symmetric facial image from an input-rotated face should be generated by using the bilateral symmetry of a human face, as shown in Fig. 6.

Assuming that \( N \) FFPs and \( F \) facial images are given, the method to produce bilateral symmetric FFPs can be described by the following equations.

\[ \tilde{W}_{original} = \begin{bmatrix} x_{11} & \ldots & x_{1N} \\ y_{11} & \ldots & y_{1N} \\ \vdots & \ddots & \vdots \\ x_{F1} & \ldots & x_{FN} \\ y_{F1} & \ldots & y_{YN} \end{bmatrix} = \begin{bmatrix} x_1 \\ y_1 \\ \vdots \\ x_F \\ y_F \end{bmatrix} \]

\[ \tilde{W}_{bilateral} = \begin{bmatrix} 2 \cdot \text{Max}(x_1) - x_{11}^{\text{flipped}} & \ldots & 2 \cdot \text{Max}(x_1) - x_{1N}^{\text{flipped}} \\ y_{11} & \ldots & y_{1N} \\ \vdots & \ddots & \vdots \\ 2 \cdot \text{Max}(x_F) - x_{F1}^{\text{flipped}} & \ldots & 2 \cdot \text{Max}(x_F) - y_{FN}^{\text{flipped}} \end{bmatrix} \]

where \( \text{Max}(x) \) is the maximum value of vector \( x \). The order of FFPs should be inverted.

\[ I_{bilateral}(2 \cdot \text{Max}(x) - x, y) = I_{original}(x, y) \]

The bilateral symmetric facial texture can be obtained by Eq. (8), \( I(x, y) \) denotes an RGB color in the \( x \) and \( y \) coordinates.

2.4.2. Frontal FFP estimation by S3DMM

To obtain frontal FFPs from the input-rotated FFPs, S3DMM is used [3,25]. In S3DMM, the facial shape can be represented as a shape vector that consists of the \( x \), \( y \), and \( z \) coordinates of \( n \) vertices.

\[ S = (x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_n, y_n, z_n) \]
The S3DM is constructed by applying PCA to the shape vectors (S) of training 3D faces. Thus, the shape of a certain face can be represented by a linear combination of the eigenvectors of the shape as follows:

$$S = \pi + \sum_{i=1}^{m} \alpha_i s_i$$  \hspace{1cm} (10)

where $\pi$ is the mean shape, $s_i$ is the shape variation, $\alpha = [\alpha_1, \alpha_2, ..., \alpha_m]$ are the shape parameters, and $m$ is the dimension of the shape parameter. A new 3D facial shape can be generated by changing the shape parameter ($\alpha$).

A simple flowchart for S3DM is shown in Fig. 7. The 3D facial shape of the target face is reconstructed by optimizing the shape parameter to make the modeled facial shape as similar as possible to the input facial shape. To achieve optimization, the shape parameter is determined to minimize the shape residual between the input 2D facial shape and the modeled facial shape parameter. The optimal shape parameter can be obtained from the cost function given in Eq. (11); the iterative process is performed until it satisfies a predetermined value.

$$\min_{\pi \in \mathbb{R}^T} ||P(\pi S + \tilde{T}) - \tilde{S}_{2d}||^2$$  \hspace{1cm} (11)

where $\tilde{S}_{2d} = (x_1, y_1, x_2, y_2, ..., x_n, y_n)^T$ is a $2 \times n$ matrix that represents the 2D facial shape obtained from the input facial image. $S$ is a $3 \times n$ matrix that is acquired by reshaping the $3n \times 1$ model shape vector $S$ obtained using Eq. (10). $P$ is a $2 \times 3$ orthographic projection matrix, $T$ is a $3 \times n$ translation matrix, and $R$ is a $3 \times 3$ rotation matrix.

### 2.4.3. Structure from motion

The SFM technique is not biased towards a mean face; moreover, for 3D facial reconstruction, it is more robust to pose variations than 3DMM. However, SFM cannot accurately reconstruct a 3D face using only a single-view facial image; it requires corresponding points from facial images in various poses [26]. Therefore, we obtained the bilateral symmetric FFPs and frontal FFPs in the previous steps. At this point, the 3D face can be reconstructed by SFM using these FFPs, which can be represented by the measurement matrix shown in Eq. (12).

$$W = \begin{bmatrix} x_{n1} & \cdots & x_{nN} \\ y_{n1} & \cdots & y_{nN} \\ \vdots & \vdots & \vdots \\ x_{f1} & \cdots & x_{fN} \\ y_{f1} & \cdots & y_{fN} \end{bmatrix}$$  \hspace{1cm} (12)

where $x_{nN}$ and $y_{nN}$ are the x and y coordinates of the Nth FFP in the $i$th image. Matrix $W$ can be expressed by a scale-orthographic camera model, as shown in Eq. (13).

$$W = PS + t^T 1$$  \hspace{1cm} (13)

$P$ is a $2 \times 3$ projection matrix, and $S$ is a $3 \times N$ matrix consisting of 3D FFPs. $t$ is a $1 \times 2$ translation matrix, and $1$ is a $1 \times N$ matrix in which all elements are equal to 1. Let us suppose that the center of the 3D face is the same as the center of the “world” coordinates. Based on this assumption, the translation vector $t$ can be described by the following equation:

$$t = \begin{bmatrix} 1 \ N_i=1 \frac{1}{N} \sum_{i=1}^{N} W_{11} \cdots 1 \ N_i=1 \frac{1}{N} \sum_{i=1}^{N} W_{2i} \end{bmatrix}$$  \hspace{1cm} (14)

where $w_{ab}$ is an element in the $a$th row and $b$th column of the matrix $W$. The SFM algorithm then calculates the registered measurement matrix $\tilde{W}$ by subtracting $t$ from $W$ as follows:

$$\tilde{W} = W - t^T 1 = PS$$  \hspace{1cm} (15)

From the given matrix $\tilde{W}$, the SFM algorithm (also known as the factorization method) finds $S$ and $P$ by using the rank constraint and SVD. To be specific, the registered measurement matrix $\tilde{W}$ is decomposed into a $2F \times 2F$ unitary matrix $O_1$, a $2F \times N$ diagonal matrix $\Sigma$, and an $N \times N$ unitary matrix $O_2$ by using the SVD as follows:

$$\tilde{W} = O_1 \Sigma O_2$$  \hspace{1cm} (16)

According to the rank theorem [17], the rank of $\tilde{W}$ is at most three if we assume the target object is rigid and there is no measurement noise in $\tilde{W}$. Even if there is some noise in $\tilde{W}$, the best possible rank-three approximation to the ideal registered measurement matrix ($\tilde{W}$ without noise) can be expressed as follows:

$$\tilde{W} = O_1 \Sigma O_2 \approx O_1 \Sigma^{1/2} O_2^{1/2}$$  \hspace{1cm} (17)

where the $2F \times 3$ matrix $O_1^{1/2}$ consists of the first three columns of the $2F \times 2F$ unitary matrix $O_1$, the $3 \times 3$ matrix $\Sigma$ is a sub-matrix of the $2F \times N$ diagonal matrix $\Sigma$, and the $3 \times N$ matrix $O_2^{1/2}$ consists of the first three rows of the $N \times N$ unitary matrix $O_2$. Then, $S$ and $P$ can be obtained as follows:

$$S = [\Sigma]^{1/2} O_2^{1/2}$$  \hspace{1cm} (18)

$$P = O_1^{1/2} [\Sigma]^{1/2}$$  \hspace{1cm} (19)

The factorization method is a closed-form solution; therefore, it incurs minimal computational cost. In addition, this method is proven to provide an optimal solution if we assume $W$ has the equivalent Gaussian noise for each component of $\tilde{W}$.

### 2.5. Dense 3D facial shape reconstruction

To reconstruct a complete 3D face, a dense 3D shape should be reconstructed from a sparse 3D shape, as obtained in Section 2.4. The dense 3D shape is reconstructed by adapting the FFPs reconstructed by SFM to a dense 3D mean face obtained from the training data. The dense 3D mean face is composed of more than 0.1 million vertices and mesh information, and 80 3D FFPs are manually extracted from the dense 3D mean face. A nonlinear relationship exists between the 3D FFPs in the previous steps. At this point, the 3D face can be reconstructed by SFM using these FFPs, which can be represented by the measurement matrix shown in Eq. (12).
nonlinear transformation parameter, \( t \) is a translation matrix, and \( s \) is a spline function. These parameters are calculated by the least square method \([12]\).

Thus, we show that the FFPs of the 3D mean face \((u)\) are converted into 3D FFPs reconstructed from 2D FFPs by the nonlinear function \( F \). Residual vertices are converted by the nonlinear function \( F \); a dense 3D facial shape can then be reconstructed. Fig. 8 depicts the process by which the dense 3D facial shape is generated from a sparse 3D facial shape.

2.6. Texture mapping

The last step of the 3D facial reconstruction process involves mapping the texture information of the input facial image onto the dense 3D facial shape. However, in the case of a rotated facial image, some regions of facial texture can be self-occluded. To recover the texture of the self-occluded region, the texture of the bilateral symmetric facial image generated using a mirroring process is used as shown in Fig. 9.

The texture mapping method is described as follows. First, a corresponding texture is obtained by projecting the dense 3D facial shape onto the 2D facial image using the projection matrix \( P \) obtained in Section 2.4.3. This involves projecting one vertex of the dense 3D facial shape onto three 2D facial images. The value of the corresponding vertex color is obtained by calculating the weighted mean of the color information of the projected vertices. The weight is determined by using the mean value of the normal vectors for the meshes that contain the corresponding vertices. For example, if the mean value of a normal vector indicates the left side, more weight is given to the color value obtained in the left facial image; if the mean value of a normal vector indicates the right side, more weight is given to the color value obtained in the right facial image.

3. Experimental results

3.1. Database

To test our proposed method, 3D facial scans from 150 subjects comprising 86 males and 64 females were obtained by a 3D laser scanner \([25,26]\). The ages of the subjects ranged from 10 to 60 years. The 3D facial scans were composed of more than 0.1 million vertices, meshes, and texture coordinates. Eighty ground-truth 3D FFPs were extracted by manual annotation of the obtained 3D facial scans, and 35 poses (yaw angles: \(0°, \pm 15°, \pm 30°, \pm 45°\); pitch angles: \(0°, \pm 15°, \pm 30°\)) of 2D facial images were obtained by rotating and projecting the 3D facial scans, as shown in Fig. 10. The total number of ground-truth 2D facial shapes was therefore 5250 (150 subjects \(\times\) 35 views). The resolution of the projected 2D facial images was \(1200\times900\), and the mean facial width was approximately \(300\) pixels. 2D facial shape data was automatically extracted (AAM) and manually annotated on the projected 2D facial images. This process involved the extraction of the 2D facial shape data consisting of 2D FFPs by considering the ideal case and practical cases.

For the ideal case, Test Data 1 was provided based on the ground-truth 2D FFPs. The ground-truth 2D FFPs were acquired by rotating the ground-truth 3D FFPs (manually annotated points on the 3D facial scan) and projecting them onto the \(x-y\) plane. For the practical case, Test Data 2 was derived by automatically detecting the FFPs through the AAM fitting algorithm. The AAM used is based on the simultaneous inverse compositional algorithm. This data can have AAM fitting errors as well as errors caused by self-occlusion.

To measure the similarity between the reconstructed facial shape and the corresponding ground-truth facial shape, we used the Root Mean Square Error (RMSE) method. Let the reconstructed...
Fig. 9. Texture mapping procedure.

Fig. 10. 2D facial images in different poses.
facial shape vector be \( S_x = [X_1, Y_1, Z_1, X_2, \ldots, Y_n, Z_n]^T \) and the corresponding ground-truth facial shape vector be \( S_T = [X_1, Y_1, Z_1, X_2, \ldots, Y_n, Z_n]^T \). Given these two shape vectors, the RMSE is calculated by the following equation, where \( n \) is the number of FFPs:

\[
e = \frac{1}{n} \sum_{i=1}^{n} \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2 + (Z_i - Z_j)^2}
\]  

(21)

3.2. Experimental results

3.2.1. Quantitative 3D facial reconstruction results

The proposed method was compared and evaluated against the previous method that used 3DMM in 35 2D facial pose images. The performance for each method was compared and evaluated by the RMSE between the reconstructed 3D FFPs and the ground-truth 3D FFPs manually annotated in the 3D scans. The RMSE was measured in millimeters after adapting the reconstructed 3D facial shape onto the 3D FFPs on a 3D facial scan. The 3D facial scan was obtained from a 3D laser scanner using a similarity transform. The FFPs obtained through manual annotation were used; the mean values of the 3D RMSE for the previous and proposed methods are shown in Table 2 and Fig. 11, respectively. We find that the proposed method greatly reduced the 3D RMSE compared to the previous method. In case the yaw angle is within ±30°, the previous method using 3DMM demonstrates better performance.

Table 2

<table>
<thead>
<tr>
<th>Pitch</th>
<th>Yaw</th>
</tr>
</thead>
<tbody>
<tr>
<td>-45°</td>
<td>5.4153</td>
</tr>
<tr>
<td>-30°</td>
<td>5.4205</td>
</tr>
<tr>
<td>0°</td>
<td>5.4305</td>
</tr>
<tr>
<td>15°</td>
<td>5.4355</td>
</tr>
<tr>
<td>30°</td>
<td>5.4405</td>
</tr>
<tr>
<td>45°</td>
<td>5.4455</td>
</tr>
</tbody>
</table>

Fig. 11. 3D RMSE graph of previous and proposed methods when manually annotated feature points were used (①: S3DMM+SfM, ②: S3DMM [25], ③: S3DMM [3]).

Table 3

<table>
<thead>
<tr>
<th>Pitch</th>
<th>Yaw</th>
</tr>
</thead>
<tbody>
<tr>
<td>③: 5.4386</td>
<td>4.6867</td>
</tr>
<tr>
<td>0°</td>
<td>①: 3.4253</td>
</tr>
<tr>
<td>③: 5.5012</td>
<td>4.5034</td>
</tr>
</tbody>
</table>

Fig. 12. 3D RMSE graph of previous and proposed methods when automatically annotated feature points were used.

Fig. 13. 3D RMSEs in local facial regions.

Table 4

<table>
<thead>
<tr>
<th>SFM method</th>
<th>Manually annotated FFPs (mm)</th>
<th>Automatically annotated FFPs (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marques et al. [22]</td>
<td>2.6168</td>
<td>3.5328</td>
</tr>
<tr>
<td>Wiberg et al. [23]</td>
<td>2.8954</td>
<td>3.8558</td>
</tr>
<tr>
<td>Buchanan et al. [24]</td>
<td>2.9122</td>
<td>3.8892</td>
</tr>
</tbody>
</table>
Fig. 14. Examples of qualitative 3D facial reconstruction results: (a) SfM method using an input-rotated facial image and a bilateral-symmetric image without the frontal facial image; (b) S3DMM method [25]; and (c) the proposed method.
when the yaw angle is greater than 10°. When the yaw angle is greater than 30° or less than −30°, the proposed method (which employs a combination of SfM and S3DMM) demonstrates better performance than the previous method. In particular, we find that the 3D RMSE gap increases as the face rotation angle increases.

The 3D RMSEs resulting when FFPs automatically annotated by AAM were used in the previous and proposed methods are shown in Table 3 and Fig. 12. These results demonstrate that the proposed method is more robust to pose variations than the previous method when automatically annotated FFPs are used. Therefore, the proposed method can effectively perform 3D facial reconstruction with pose variations in both online and offline applications.

Fig. 13 presents the 3D RMSEs of each local region, i.e., the eyebrows, eyes, nose, mouth, and the facial contour. We find that the proposed method provides better performance in 3D facial reconstruction than the previous method for each local region. Additionally, the previous method demonstrated significant 3D RMSEs in the facial contour region. This region has greater variations due to head rotation compared to the center region (eyebrows, eyes, nose, and mouth), which has relatively small variations caused by head rotation. Therefore, the previous method is vulnerable to head rotation. The proposed method greatly reduced the 3D RMSEs in the facial contour region, thereby decreasing the difference of 3D RMSEs in the facial contour and center regions.

Table 4 shows the average 3D RMSEs of the proposed method at ±30° and ±45° yaw angles when different types of SfM methods were used. Our experiments confirmed that the SfM proposed by [22] demonstrated the best performance in terms of 3D RMSEs. Therefore, we used the method [22] in our proposed method.

3.2.2. Qualitative 3D facial reconstruction results
To visually demonstrate the 3D facial reconstruction results, we obtained qualitative results from the previous SfM method, S3DMM method, and the proposed method. Fig. 14 shows the reconstructed results for six subjects from our database when using manually annotated FFPs. The first column lists the input-rotated images. Column A shows the reconstructed faces using the previous SfM method with an input-rotated facial image and a bilateral symmetric image without the frontal face image. Because the corresponding FFPs were not extracted in various poses, the 3D facial shapes reconstructed by the SfM method using these FFPs were not accurate. Column B shows the reconstructed faces using the previous S3DMM method. This method has the disadvantage of being biased toward a mean face and generating relatively large 3D reconstruction errors in the facial contour region where occluded FFPs caused by head rotation exist. Therefore, as shown in Fig. 14, we find that an oval (thin) face can be reconstructed into a rounder face that is similar to a mean face, while a round face can be reconstructed into a thin face.

3.2.3. Evaluation though facial recognition
The performance of 3D face reconstruction can also be evaluated by matching the frontal face image of the reconstructed 3D face with the frontal face image of the ground truth 3D face. To demonstrate the 3D facial reconstruction results, we obtained qualitative results from the previous SfM method, S3DMM method, and proposed method.

As described in Section 3.1, 35 poses of 2D facial images were obtained by rotating and projecting the ground-truth 3D facial scans from 150 subjects. The frontal images of 35 facial pose images for each subject were enrolled in a gallery dataset. The probe dataset consists of 150 synthesized frontal face images from the reconstructed 3D faces. The 3D faces of each subject are reconstructed from the pose image with the greatest angular variation (yaw: 45°, pitch: 30°) among all 35 ground truth images of the subject.

As shown in Fig. 15, when an image of the probe dataset was inputted, the 3D face was reconstructed by the proposed and previous methods. After a frontal facial image was synthesized from the reconstructed 3D face, facial recognition was performed by matching the synthesized frontal facial image with the images in the gallery dataset. The software development kit called VeriLook 4.0 standard [28] was used to recognize faces. The results are presented as a Receiver Operating Characteristic (ROC) curve in Fig. 16. It is evident that the proposed method outperformed the
previous S3DMM method. The Equal Error Rate (EER), i.e., the rate at which both a False Accept Rate (FAR) and False Reject Rate (FRR) are equal, for the proposed method was 2.80%. The EER for the previous method was 4.62%. The True Accept Rate (TAR) is equal to 1 minus the FRR (TAR = 1 – FRR). If the full face and both eyes cannot be detected, a Failure To Acquire (FTA) error has occurred. In this case, the frontal face recognition was not achieved, and the test image was not counted as a recognition error. The FTA rate was approximately 1.33%.

4. Conclusions

In this work, we proposed a single-view-based pose-invariant 3D facial reconstruction method that combines 3DMM and SFM techniques. The proposed method is comprised of the following steps. FFPs are first extracted from the input-rotated facial image, and bilateral-symmetric FFPs with these FFPs are generated. The 2D frontal FFPs are then estimated from an initial 3D face temporarily reconstructed by S3DMM. The 3D facial shapes are finally reconstructed by SFM using the frontal, left, and right FFPs obtained in the previous steps. We tested our proposed method with an image database that contained images in 35 different poses. Based on the experimental results, it was evident that the proposed method is more robust to pose variations than the previous method. As shown in Table 4, the average 3D RMSE between the reconstructed 3D face and ground-truth 3D face obtained from the 3D scanner was less than 2.6 mm when the 2D FFPs were manually annotated, and less than 3.5 mm when the 2D FFPs were automatically annotated.

However, in the proposed method, two virtual facial shapes, frontal and mirrored facial shapes, are estimated from a single side facial shape. The original facial shape and two estimated shapes are used as an input to the SFM to reconstruct a 3D facial shape. Although such 3D face reconstruction approach was used in this paper considering the difficulty of 3D facial shape reconstruction from a single facial shape, a method to simplify the proposed 3D reconstruction approach needs to be studied in the future.

In future work, we will also investigate pose-invariant facial recognition and frontal-face synthesis from rotated faces in CCTV environments by using the proposed single-view-based 3D facial reconstruction method. In addition, we will conduct a study on 2D-to-3D facial conversion in 3D content services, such as 3D TV and movies.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MEST) (No. 2011-0015321).

References


