International Journal of Computer Vision 46(3), 179–199, 2002

Determination of Head Pose and Facial Expression from a Single Perspective View by Successive Scaled Orthographic Approximations*

CHIN-CHUN CHANG AND WEN-HSIANG TSAI
Department of Computer and Information Science, National Chiao Tung University, Hsinchu, Taiwan 300, Republic of China
whtsai@cis.nctu.edu.tw

Received June 2, 2000; Revised June 2, 2000; Accepted July 6, 2001

Abstract. Human faces are the main organs for expressing human emotion. In this study, a new iterative approach to analyzing the head pose and the facial expression of a human face from a single image is proposed. The proposed approach extends the concept of successive scaled orthographic approximations, which was used to estimate the pose of a rigid object, to develop a method to estimate the parameters for a non-rigid object, namely, a human face. The implementation of the proposed method is simple; furthermore, no initial guess is required. The convergency property of the proposed method is also analyzed theoretically and experimentally. Experimental results show that the proposed method is robust and has a high percentage of convergency, and thus prove the feasibility of the proposed approach.

Keywords: head pose, facial expressions, scaled orthographic projection, human-computer interaction

1. Introduction

Human faces are the main organs to express and convey human emotion. Because of growing demand for applications of the internet and human-computer interaction such as computer facial animation, distant learning, talking agents (Thalmann et al., 1998), model-based video coding (Li et al., 1993), etc., knowing about the head pose and facial expression of a human becomes important. In this study, a new method for estimating the head pose and facial expression of a human face from a single view is proposed.

In general, there exist two vision-based approaches to estimating the parameters for the head pose and the facial expression of a human face. The first approach (Li et al., 1993; Choi et al., 1994; Tao and Huang, 1998) tries to estimate the parameters of the head pose and the facial expression simultaneously but this approach faces a complicated non-linear problem. Thus, most existing systems of this approach are based on instantaneous motion. That is, they process the motion of the head pose and the facial expression with short sampling periods, and thus the change of the head pose and the facial expression in a short period of time can be approximately described by a simplified form, for example, by a linear or quadratic form. Due to this simplification, the parameters for describing the head pose and the facial expression can be solved easier. However, these systems need high speed hardwares to achieve estimating and tracking the head pose and facial expression in short sampling periods. Instead of using full perspective projection, Bascle and Blake (1998) proposed a method with a simpler mathematical formulation to estimate the head pose and the facial expression based on an affine with a parallax model. However, the estimated head pose and facial expression may be inaccurate.

The second approach (Li et al., 1996; Zhang, 1998; http://www.ina.fr/Recherche/TV) estimates the head pose and the facial expression separately. First, the head

*This work was supported by National Science Council under Grant NSC-89-2213-E-009-097.

pose is estimated by some explicitly determined rigid features on the face. After obtaining the head pose, the head movement can be compensated and thus the parameters for the facial expression can be solved. However, this approach may be unstable and inaccurate (Bascle and Blake, 1998). The first reason is that it is sensitive to noise because the head pose is estimated from only a few rigid features on the face such as eyes' corners and nostrils; in addition, stretched expressions might cause these rigid feature points nonrigid. The second is that the errors in estimating the head pose might propagate to the estimation of facial parameters. In stead of using explicitly determined rigid features, Li and Forchheimer (1994) estimated the head motion between two consecutive frames by an M-estimator which excludes the feature points violating rigidity from computing the head motion. However, this method cannot work well if there do not exist enough rigid features between two consecutive frames.

In this study, to analyze the head pose and the facial expression robustly, a new method is proposed to estimate the parameters for the head pose and the facial expression simultaneously from a single perspective view. The proposed method estimates the head pose and the facial expression from the relation between the features of a 3-D face and their respective perspective projections. The advantages of the proposed method are as follows. First, since no instantaneous motion techniques are adopted, the proposed method can be implemented on commercial hardwares. Second, because the proposed method utilizes all detected features at a time and because the proposed method does not reply on explicitly determined rigid features, the proposed method is more robust than the second approach mentioned above.

Shown in Fig. 1 is a diagram of the system proposed in this study for analyzing the head pose and the facial expression of a human face from a single view. This system consists of four main modules, namely, the 3-D face model creation module, the analysis module, the computer graphics module, and the animation module. The analysis module is designed to estimate the parameters for the head pose and the facial expression of a person using his 3-D face model established by the 3-D face model creation module. According to the estimated parameters, the animation module animates the facial motion using a computer graphics model established by the computer graphics module. For applications aiming at talking agents and video games, etc., the estimated parameters need not be very accurate, so some of the parameters for the 3-D face model for an individual human can be obtained by a transformation from a generic model. In this paper, we focus on the analysis module.

The proposed method for analyzing the head pose and the facial expression of a human face is based on DeMenthon and Davis's method (DeMenthon and Davis, 1995), which computes the pose of a rigid object with a perspective model by successive scaled orthographic approximations (Horaud et al., 1997). This

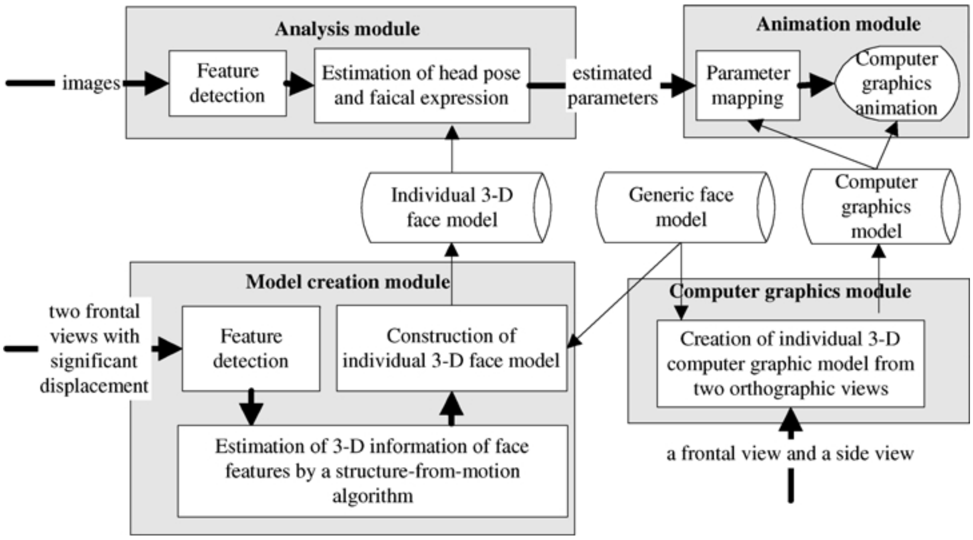

Analysis module
images
Feature detection
Estimation of head pose and faical expression
estimated parameters
Animation module
Parameter mapping
Computer graphics animation
Individual 3-D face model
Generic face model
Computer graphics model
Model creation module
two frontal views with significant displacement
Feature detection
Construction of individual 3-D face model
Estimation of 3-D information of face features by a structure-from-motion algorithm
Computer graphics module
Creation of individual 3-D computer graphic model from two orthographic views
a frontal view and a side view

Figure 1. System diagram.

method has good convergency property, needs no initial guesses, and uses fewer floating-point operations than classical approaches such as Newton's method for every iteration (DeMenthon and Davis, 1995). The proposed method deals with the motion of a human face, which is non-rigid, and the proposed method also inherits the merits of DeMenthon and Davis's method. Furthermore, it must be pointed out that the advantages of DeMenthon and Davis's method are very valuable for analysis of human faces. Especially, since there need tens to twenties parameters to be estimated for a human face in general, it is difficult to generate good initial guesses for the classical non-linear method. Therefore, a method without initial guesses is desired. In this study, the convergency property for the proposed method is analyzed based on some proposed sufficient criteria for the convergency property of DeMenthon and Davis's method. It is shown that the proposed method has high percentage of convergency without initial guesses in general configurations.

The remainder of this paper is organized as follows. In Section 2, the adopted 3-D face model is described first. From the 3-D face model, the mathematical formulation for analysis of the head pose and the facial expression is derived. The proposed method is based on the technique of successive scaled orthographic approximations, which is so reviewed in Section 3. In Section 4, the proposed method is presented. Analysis of convergency and complexity of the proposed method is discussed in Section 5. Experimental results including those of computer simulation and processing real images are presented in Section 6. Discussions and concluding remarks are given in the last section.

2. Face Model and Problem Formulation

2.1. Face Model

In this study, the pose of a human head is described by six parameters: three parameters for the orientation and three for the position. In addition, some parameters for describing the facial expressions such as the amount of jaw rotation, the displacement of eyebrows, etc., are also included. The facial expression can be described by a linear combination of some key expressions (Li et al., 1993; Li and Forchheimer, 1994; Choi et al., 1994; Ullman and Basri, 1991; Tao and Huang, 1998; Bascle and Blake, 1998) or muscle vectors (Terzopoulos and Waters, 1993; Lei et al., 1996). In this study, the latter approach is adopted. Specifically, the muscle-based face model conducts facial expressions based on the properties of the facial muscle and skin. In this study, a human face is represented by the muscle-based face model described in Appendix A of Parke and Waters's book (Parke and Waters, 1996) with a little modification. The main modification is modeling the amount of mouth opening like the muscle contraction value for a muscle vector.

Let $\mathbf{x}_i$ and $\mathbf{x}_i^{'}$ represent the position of the ith feature point before and after muscle activities, respectively. Suppose there exist m muscle fibers in the face model. When the muscle fibers activate, the displacement of the ith feature point is a weighted sum of m muscle activities acting on the ith feature point as follows:

$$\mathbf{x}_i^{'} = \mathbf{x}_i + \sum_{j=1}^{m} c_j b_{ij} \mathbf{m}_{ij}, \qquad (1)$$

where b_{ij} is a muscle blend function that specifies the influence of the jth muscle fiber on the ith feature point, c_j is the contraction factor for the jth muscle fiber, and $\mathbf{m}_{ij} = \mathbf{x}_i - \mathbf{m}_j^e$ where $\mathbf{m}_j^e$ is the point that the jth muscle fiber emerges from the bone. The muscle vector $\mathbf{m}_j$ for the jth muscle fiber is defined by $\mathbf{m}_j = \mathbf{m}_j^a - \mathbf{m}_j^e$, where $\mathbf{m}_j^a$ is the point that the jth muscle fiber attaches to the skin tissue. The definition of b_{ij} is as follows:

$$b_{ij} = \begin{cases} w \cos\left(\dfrac{\|\mathbf{m}_{ij}\|_2 - r_j^{inner}}{r_j^{outer} - r_j^{inner}} \dfrac{\pi}{2} \right) & \text{for } r_j^{inner} \le \|\mathbf{m}_{ij}\|_2 \le r_j^{outer} \text{ and } \theta_{ij} \ge \theta_j^m; \\ w & \text{for } \|\mathbf{m}_{ij}\|_2 < r_j^{inner} \text{and } \theta_{ij} \ge \theta_j^m; \qquad j = 1, 2, \ldots, m, \\ 0 & \text{otherwise;} \end{cases}$$

where $w = 1 - \frac{\theta_{ij}}{\theta_j^m}$, θ_j^m is the cosine of the angle of the influence zone for the jth muscle, r_j^{inner} and r_j^{outer}, respectively, are the inner radius and the outer radius of influence of the consine blend profile for the j th muscle, and θ_{ij} is the cosine of the angle between $\mathbf{m}_{ij}$ and $\mathbf{m}_j$. In this study, the eighteen muscle vectors defined in Appendix A of the book (Parke and Waters, 1996) are adopted (i.e. $m = 18$). For simplicity, let $\mathbf{v}_{ij} = b_{ij}\mathbf{m}_{ij}$.

Opening the mouth makes the feature points on the lower part of the face rotate about a jaw pivot axis

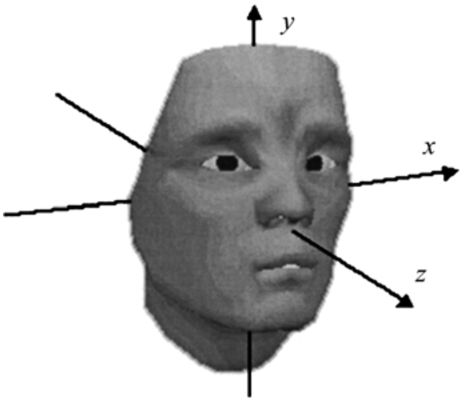


Figure 2. Face coordinate system.

(Parke and Waters, 1996). Since the feature points of the lower part of the face are also affected by some muscle activities, estimation of the amount of jaw rotation becomes complicated. For simplicity, as shown in Fig. 2, the x-axis of the face coordinate system is a jaw pivot axis. Accordingly, opening the mouth moves the feature points on the lower part of the face by rotating them about the x-axis. Since the amount of jaw rotation for a normal person is not large, the rotation matrix for jaw rotation can be approximately represented by an identity matrix $\mathbf{I}$ plus an anti-symmetric matrix (Kanatani, 1993). Define $b_{i(m+1)}$ as an indicator for the ith feature point being influenced by jaw rotation; that is,

$$b_{i(m+1)} = \begin{cases} 1 & \text{if feature point } i \text{ is in the lower part of the face;} \\ 0 & \text{otherwise.} \end{cases}$$

Hence, a rotation matrix $\mathbf{R}_{ix}$ denoting the amount of the ith feature point rotating around the jaw pivot axis can be defined as follows:

$$\mathbf{R}_{ix} \cong \mathbf{I} + c_{m+1} b_{i(m+1)} \mathbf{K}, \tag{2}$$

where c_{m+1} represents the amount of jaw rotation and $\mathbf{K} = \left[\begin{smallmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{smallmatrix}\right]$. By considering jaw rotation as well as the muscle activities, Eq. (1) becomes

$$\mathbf{x}_i' = \mathbf{R}_{ix} \left(\mathbf{x}_i + \sum_{j=1}^{m} c_j \mathbf{v}_{ij} \right).$$

Substituting $\mathbf{R}_{ix}$ by Eq. (2), we can obtain

$$\mathbf{x}_i' = \mathbf{x}_i + \sum_{j=1}^{m} c_j \mathbf{v}_{ij} + c_{m+1} b_{i(m+1)} \mathbf{K}\mathbf{x}_i + c_{m+1} b_{i(m+1)} \mathbf{K} \sum_{j=1}^{m} c_j \mathbf{v}_{ij}. \tag{3}$$

By ignoring the right-most term, the least significant term, and letting $\mathbf{v}_{i(m+1)} = b_{i(m+1)}\mathbf{K}\mathbf{x}_i$, Eq. (3) becomes

$$\mathbf{x}_i' = \mathbf{x}_i + \sum_{j=1}^{m+1} c_j \mathbf{v}_{ij}. \tag{4}$$

Accordingly, estimation of the amount of jaw rotation can be done by the same way as estimation of the contraction factors of muscle fibers, and Eq. (4) is the fundamental equation for analyzing the facial expression in this study.

2.2. Problem Formulation

Let the orientation and the position of the face with respect to a camera be described by a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$, respectively. Let $\mathbf{x}_i''$ obtained by

$$\mathbf{x}_i'' = \mathbf{R}\mathbf{x}_i' + \mathbf{t}, \tag{5}$$

denote the position of the ith feature point with respect to the camera coordinate system. Let $\mathbf{u}_i = [u_{ix} \ u_{iy}]^t$, $i = 1, 2, \ldots, n$, be the perspective projections of n feature points on the face; that is, let

$$\mathbf{u}_i = \left[\frac{f x_{ix}''}{x_{iz}''} \quad \frac{f x_{iy}''}{x_{iz}''} \right]^t, \tag{6}$$

where $\mathbf{x}_i'' = [x_{ix}'' \ x_{iy}'' \ x_{iz}'']^t$ and f is the focal length of the camera. Without loss of generality, f can be assumed to be one. In addition, we have some prior knowledge about the ranges of the muscle contraction values of muscle fibers and the amount of jaw rotation as follows:

$$G = \{\alpha_j \le c_j \le \beta_j, j = 1, 2, \ldots, m+1\}. \tag{7}$$

According to the 3-D face model, $\mathbf{x}_i$ and $\mathbf{v}_{ij}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m+1$, are known. Hence, the problem that we deal with is to find $\mathbf{R}$, $\mathbf{t}$, and

c_j, $j = 1, 2, \ldots, m+1$, from the perspective projections of the n feature points subject to the inequalities (7). In this study, the 3-D face model of a human face contains eighteen muscle fibers (i.e., $m = 18$); therefore, we have to estimate twenty-five parameters.

3. Review of Estimating Pose of Rigid Object by Successive Scaled Orthographic Approximations

It is well known that to obtain a better estimated pose by the scaled orthographic projection model, it is appropriate to align the optical axis of the camera to pass through the gravity center of the projections of the feature points (Aloimonos, 1990). The gravity center of the projections of n feature points can be computed by $[\bar{u}_x \ \ \bar{u}_y]^t = \frac{1}{n}\sum_{i=1}^{n} \mathbf{u}_i$. The ith feature point after this alignment process becomes $\frac{1}{\hat{\mathbf{r}}_3[\mathbf{u}_i^t \ \ 1]^t}[\hat{\mathbf{r}}_1[\mathbf{u}_i^t \ \ 1]^t \ \hat{\mathbf{r}}_2[\mathbf{u}_i^t \ \ 1]^t]^t$, where

$$\hat{\mathbf{R}} = \begin{bmatrix} \hat{\mathbf{r}}_1 \\ \hat{\mathbf{r}}_2 \\ \hat{\mathbf{r}}_3 \end{bmatrix} = \begin{bmatrix} \frac{\bar{u}_y}{\sqrt{\bar{u}_x^2+\bar{u}_y^2}} & \frac{-\bar{u}_x}{\sqrt{\bar{u}_x^2+\bar{u}_y^2}} & 0 \\ \frac{\bar{u}_x}{\sqrt{\bar{u}_x^2+\bar{u}_y^2+1}\sqrt{\bar{u}_x^2+\bar{u}_y^2}} & \frac{\bar{u}_y}{\sqrt{\bar{u}_x^2+\bar{u}_y^2+1}\sqrt{\bar{u}_x^2+\bar{u}_y^2}} & \frac{-\sqrt{\bar{u}_x^2+\bar{u}_y^2}}{\sqrt{\bar{u}_x^2+\bar{u}_y^2+1}} \\ \frac{\bar{u}_x}{\sqrt{\bar{u}_x^2+\bar{u}_y^2+1}} & \frac{\bar{u}_y}{\sqrt{\bar{u}_x^2+\bar{u}_y^2+1}} & \frac{1}{\sqrt{\bar{u}_x^2+\bar{u}_y^2+1}} \end{bmatrix}.$$

After estimating the object pose, the desired rotation matrix and translation vector can be obtained by multiplying $\hat{\mathbf{R}}^t$ and the estimated rotation matrix and the estimated translation vector together, respectively. In the following sections, the optical axis of the camera is assumed to have been aligned to pass through the gravity center of the projections of the feature points. In addition, we regard an object to be in an *ambiguous state* if the perspective projections of the feature points of the object are all identified and if there exist at least two sets of parameters for the object which yield the same perspective projection.

3.1. Review of DeMenthon and Davis's Method

DeMenthon and Davis's method (DeMenthon and Davis, 1995) is a fast iterative method for estimating the pose of a 3-D rigid object. The principle of DeMenthon and Davis's method is stated as follows. If the image of a rigid object is a scaled orthographic projection, the pose of the object can be solved by a simple linear method. However, the image of the rigid object is a perspective projection. Hence, DeMenthon and Davis proposed an iterative scheme to compute the object pose by adjusting the perspective projection of the object to a scaled orthographic projection with the same pose parameters. The following is a review of DeMenthon and Davis's method.

From the perspective projection equation for the feature $\mathbf{x}_i$, we have

$$\mathbf{u}_i = \frac{1}{\mathbf{r}_3\mathbf{x}_i + t_z}\left(\begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix}\mathbf{x}_i + \begin{bmatrix} t_x \\ t_y \end{bmatrix}\right), \tag{8}$$

where $\mathbf{r}_1, \mathbf{r}_2$, and $\mathbf{r}_3$ are the three row vectors of $\mathbf{R}$ and $\mathbf{t} = [t_x \ t_y \ t_z]^t$. Dividing the numerators and denominators of the right-hand side of Eq. (8) by t_z, and rearranging the resulting equations, we can obtain

$$\mathbf{u}_i\,(1+\varepsilon_i) = \begin{bmatrix} \mathbf{r}_1' \\ \mathbf{r}_2' \end{bmatrix}\mathbf{x}_i + \begin{bmatrix} t_x' \\ t_y' \end{bmatrix}, \tag{9}$$

where

$$\begin{aligned} \mathbf{r}_i' &= \frac{\mathbf{r}_i}{t_z}, \quad i = 1, 2, 3, \\ t_x' &= \frac{t_x}{t_z}, \\ t_y' &= \frac{t_y}{t_z}, \\ \varepsilon_i &= \mathbf{r}_3'\mathbf{x}_i. \end{aligned} \tag{10}$$

It should be noticed that $\mathbf{u}_i(1+\varepsilon_i)$ is a scaled orthographic projection of the ith feature point, and Eq. (9) is called a scaled orthographic projection equation for the feature $\mathbf{x}_i$. In addition, it is possible to estimate the object pose by minimizing the residue in Eq. (9) for every feature point, and this minimization process can be achieved in the following manner.

Arranging Eq. (9) for every feature point on the rigid object to be a matrix form, we can obtain

$$\bar{\mathbf{U}} = \begin{bmatrix} \mathbf{r}_1' \\ \mathbf{r}_2' \end{bmatrix}\bar{\mathbf{X}}, \tag{11}$$

where

$$\begin{aligned} \bar{\mathbf{U}} &= [\mathbf{u}_1(1+\varepsilon_1) - \bar{\mathbf{u}}_\varepsilon \ \cdots \ \mathbf{u}_n(1+\varepsilon_n) - \bar{\mathbf{u}}_\varepsilon], \\ \bar{\mathbf{X}} &= [\mathbf{x}_1 - \bar{\mathbf{x}} \ \cdots \ \mathbf{x}_n - \bar{\mathbf{x}}], \end{aligned}$$

with

$$\bar{\mathbf{u}}_\varepsilon = \frac{1}{n}\sum_{i=1}^{n}\mathbf{u}_i(1+\varepsilon_i),$$

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i.$$

To obtain unique $\mathbf{r}'_1$ and $\mathbf{r}'_2$ by solving Eq. (11), the rank of $\bar{\mathbf{X}}$ must be three; that is, the configuration of the feature points cannot be coplanar. In the following are the steps of DeMenthon and Davis's method.

Algorithm 1. DeMenthon and Davis's method.

Input. $\mathbf{u}_i, \mathbf{x}_i, i = 1, 2, \ldots, n$.

Output. $\mathbf{R}$ and $\mathbf{t}$.

Step 1. For $i = 1, 2, \ldots, n, \varepsilon_i = 0$. Compute $\bar{\mathbf{X}}^+ = \bar{\mathbf{X}}^t(\bar{\mathbf{X}}\bar{\mathbf{X}}^t)^{-1}$.

Step 2. Estimate $\mathbf{r}''_1$ and $\mathbf{r}''_2$ by solving the overconstrained linear system Eq. (11). That is, compute

$$\begin{bmatrix}\mathbf{r}''_1\\ \mathbf{r}''_2\end{bmatrix} = \bar{\mathbf{U}}\bar{\mathbf{X}}^+. \tag{12}$$

Step 3. Compute $\mathbf{R} = [\,\mathbf{r}_1^t\ \mathbf{r}_2^t\ \mathbf{r}_3^t\,]^t$ and t_z by

$$\mathbf{r}_i = \frac{\mathbf{r}''_i}{\|\mathbf{r}''_i\|_2}, \quad i = 1, 2,$$

$$\mathbf{r}_3 = \frac{\mathbf{r}_1 \times \mathbf{r}_2}{\|\mathbf{r}_1 \times \mathbf{r}_2\|_2},$$

$$t_z = \sqrt{\frac{2}{\|\mathbf{r}''_1\|_2^2 + \|\mathbf{r}''_2\|_2^2}},$$

where $\|\cdot\|_2$ is the vector 2-norm.

Step 4. Update $\mathbf{r}'_i$ by $\mathbf{r}'_i = \frac{\mathbf{r}_i}{t_z}$ for each $i = 1, 2, 3$.

Step 5. Compute $[t_x\ t_y]^t = t_z(\bar{\mathbf{u}}_\varepsilon - [\mathbf{r}_1'^t\ \mathbf{r}_2'^t]^t\bar{\mathbf{x}})$.

Step 6. For all $i = 1, 2, \ldots, n$, compute ε_i by Eq. (10). If $\varepsilon_i, i = 1, 2, \ldots, n$, moves a small distance in this iteration, then stop; otherwise go to Step 2.

The rotation matrix estimated by the original DeMenthon and Davis's method does not guarantee orthonormality, however. To ensure the orthonormality of the estimated rotation matrix, a fitting process (Kanatani, 1993; Horaud et al., 1997) can be included into Step 3, and the modified method is called the modified DeMenthon and Davis's method in this study. The modified Step 3 is as follows.

Step 3′. Compute the rotation matrix $\mathbf{R} = [\mathbf{r}_1^t\ \mathbf{r}_2^t\ \mathbf{r}_3^t\,]^t$ and t_z that minimize

$$\left\|\frac{1}{t_z}\left[\mathbf{r}_1^t\mathbf{r}_2^t\right]^t - \left[\mathbf{r}_1''^t\ \ \mathbf{r}_2''^t\right]^t\right\|_F$$

by using the singular value decomposition as follows:

$$\begin{bmatrix}\mathbf{r}_1\\ \mathbf{r}_2\end{bmatrix} = \mathbf{P}\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\end{bmatrix}\mathbf{Q}^t,$$

$$\mathbf{r}_3 = \mathbf{r}_1 \times \mathbf{r}_2,$$

$$t_z = \frac{2}{S_{11} + S_{22}},$$

where $\|\cdot\|_F$ represents the Frobenius matrix norm (Horn and Johnson, 1985), and $\mathbf{PSQ}^t$ is the singular value decomposition of $[\,\mathbf{r}_1''^t\ \mathbf{r}_2''^t\,]^t$ with singular values S_{11} and S_{22}.

4. Estimation of Head Pose and Facial Expression

In this section, the proposed method for estimating the head pose, the amount of jaw rotation, and the muscle contraction values is presented first. Upper eyelids are also important organs for facial expressions. In this study, for simplicity, upper eyelids are defined to have three states, namely, the opened state, the partially opened state, and the closed state. The method for determining the states of upper eyelids is introduced at last.

4.1. Estimation of Head Pose, Amount of Jaw Rotation, and Muscle Contraction Values

From Eqs. (4)–(6), we have the perspective equations for the feature points on the face as follows:

$$\mathbf{u}_i = \frac{1}{\mathbf{r}_3\left(\mathbf{x}_i + \sum_{j=1}^{m+1}c_j\mathbf{v}_{ij}\right) + t_z} \times \begin{bmatrix}\mathbf{r}_1\left(\mathbf{x}_i + \sum_{j=1}^{m+1}c_j\mathbf{v}_{ij}\right) + t_x\\ \mathbf{r}_2\left(\mathbf{x}_i + \sum_{j=1}^{m+1}c_j\mathbf{v}_{ij}\right) + t_y\end{bmatrix}, \quad i = 1, \ldots, n. \tag{13}$$

Rearranging Eq. (13), we can obtain

$$\mathbf{u}_i\,(1+\varepsilon_{di}) - \begin{bmatrix}\xi_{ix}\\ \xi_{iy}\end{bmatrix} = \begin{bmatrix}\mathbf{r}'_1\\ \mathbf{r}'_2\end{bmatrix}\mathbf{x}_i + \begin{bmatrix}t'_x\\ t'_y\end{bmatrix}, \quad i = 1,\ldots,n, \qquad (14)$$

where

$$\varepsilon_{di} = \mathbf{r}'_3\left(\mathbf{x}_i + \sum_{j=1}^{m+1} c_j \mathbf{v}_{ij}\right), \qquad (15)$$

$$\xi_{ix} = \mathbf{r}'_1 \sum_{j=1}^{m+1} c_j \mathbf{v}_{ij}, \qquad (16)$$

$$\xi_{iy} = \mathbf{r}'_2 \sum_{j=1}^{m+1} c_j \mathbf{v}_{ij}. \qquad (17)$$

It should be noticed that $[\xi_{ix}\ \xi_{iy}]^t$ is a scaled orthographic projection of the local motion of the ith feature point, and $\mathbf{u}_i(1+\varepsilon_{di}) - [\xi_{ix}\ \xi_{iy}]^t$ is a scaled orthographic projection of the ith feature point without local motion. Obviously, for a rigid object, $\xi_{ix} = \xi_{iy} = 0$. From Eq. (14), we have the matrix form similar to Eq. (11) for the DeMenthon and Davis's method as follows:

$$\bar{\mathbf{U}}_d = \begin{bmatrix}\mathbf{r}'_1\\ \mathbf{r}'_2\end{bmatrix}\bar{\mathbf{X}}; \qquad (18)$$

where

$$\bar{\mathbf{U}}_d = [\mathbf{u}_1(1+\varepsilon_{d1}) - \boldsymbol{\xi}_1 - \bar{\mathbf{u}}_{\varepsilon d} \cdots \mathbf{u}_n(1+\varepsilon_{dn}) - \boldsymbol{\xi}_n - \bar{\mathbf{u}}_{\varepsilon d}],$$

with

$$\boldsymbol{\xi}_i = [\xi_{ix} \quad \xi_{iy}]^t,$$

$$\bar{\mathbf{u}}_{\varepsilon d} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}_i(1+\varepsilon_{di}) - \boldsymbol{\xi}_i).$$

In this study, the parameters for the head pose and the facial expression are obtained by minimizing the residue in Eq. (14) iteratively, and the steps are described as follows.

Algorithm 2. Estimation of the head pose and facial expression by successive scaled orthographic approximations.

Input. $\mathbf{x}_{i,}$ $\mathbf{u}_i$, and $\mathbf{v}_{ij}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m+1$.

Output. $\mathbf{R}$, $\mathbf{t}$, and c_i, $i = 1, 2, \ldots, m+1$.

Steps.

Step 1. Initially, for $i = 1, 2, \ldots, n$, $\varepsilon_{di} = 0$, $\xi_{ix} = \xi_{iy} = 0$.

Step 2. Estimate $\mathbf{r}''_1$ and $\mathbf{r}''_2$ by the same operations as Step 2 in Algorithm 1 with $\bar{\mathbf{U}}$ being replaced by $\bar{\mathbf{U}}_d$.

Step 3. Compute $\mathbf{R} = [\,\mathbf{r}^t_1\ \mathbf{r}^t_2\ \mathbf{r}^t_3\,]^t$ and t_z by the same operations as Step 3′ in Algorithm 1.

Step 4. Update $\mathbf{r}'_i$ by the same operations as Step 4 in Algorithm 1.

Step 5. Compute $[\,t_x\ t_y]^t$ by the same operations as Step 5 in Algorithm 1 with $\bar{\mathbf{u}}_\varepsilon$ being replaced by $\bar{\mathbf{u}}_{\varepsilon d}$.

Step 6. According to the newly estimated $\mathbf{r}'_1$, $\mathbf{r}'_2$, and $\mathbf{t}$, estimate the parameters for the facial expression c_i, $i = 1, 2, \ldots, m+1$ (see Section 4.1.1).

Step 7. For all $i = 1, 2, \ldots, n$, compute ε_{di}, ξ_{ix}, and ξ_{iy} by Eqs. (15), (16), and (17), respectively. If ε_{di}, $i = 1, 2, \ldots, n$, move a small distance in this iteration, then stop; otherwise go to Step 2.

In this study, the termination criterion for the proposed method is the average of the distances which all ε_{di}, $i = 1, 2, \ldots, n$, advance in an iteration is smaller than 10^{-6}.

4.1.1. Solving the Parameters for Facial Expression. After obtaining the head pose from the scaled orthographic projection model for an iteration, the muscle contraction values and the amount of jaw rotation can be obtained by solving a system of linear equations formed by the scaled orthographic projections of all feature points without the terms for the global motion. Specifically, from Eq. (14), we have

$$\begin{bmatrix}\mathbf{r}'_1\mathbf{v}_{i1} & \mathbf{r}'_1\mathbf{v}_{i2} & \cdots & \mathbf{r}'_1\mathbf{v}_{i(m+1)}\\ \mathbf{r}'_2\mathbf{v}_{i1} & \mathbf{r}'_2\mathbf{v}_{i2} & \cdots & \mathbf{r}'_1\mathbf{v}_{i(m+1)}\end{bmatrix}\mathbf{c} = \mathbf{u}_i\,(1+\varepsilon_{di}) - \begin{bmatrix}\mathbf{r}'_1\\ \mathbf{r}'_2\end{bmatrix}\mathbf{x}_i - \begin{bmatrix}t'_x\\ t'_y\end{bmatrix}, \quad i = 1, 2, \ldots m+1,$$

or equivalently,

$$\mathbf{G}\mathbf{c} = \mathbf{g}, \qquad (19)$$

where

$$\mathbf{G} = \begin{bmatrix} \mathbf{r}_1'\mathbf{v}_{11} & \mathbf{r}_1'\mathbf{v}_{12} & \cdots & \mathbf{r}_1'\mathbf{v}_{1(m+1)} \\ \mathbf{r}_2'\mathbf{v}_{11} & \mathbf{r}_2'\mathbf{v}_{12} & \cdots & \mathbf{r}_2'\mathbf{v}_{1(m+1)} \\ \vdots & \vdots & & \vdots \\ \mathbf{r}_1'\mathbf{v}_{n1} & \mathbf{r}_1'\mathbf{v}_{n2} & \cdots & \mathbf{r}_1'\mathbf{v}_{n(m+1)} \\ \mathbf{r}_2'\mathbf{v}_{n1} & \mathbf{r}_2'\mathbf{v}_{n2} & \cdots & \mathbf{r}_2'\mathbf{v}_{n(m+1)} \end{bmatrix},$$

$$\mathbf{c} = [c_1 \quad c_2 \quad \cdots \quad c_{m+1}]^t,$$

$$\mathbf{g} = \begin{bmatrix} u_{1x}\,(1+\varepsilon_{d1}) - \mathbf{r}_1'\mathbf{x}_1 - t_x' \\ u_{1y}\,(1+\varepsilon_{d1}) - \mathbf{r}_2'\mathbf{x}_1 - t_y' \\ \vdots \\ u_{nx}\,(1+\varepsilon_{dn}) - \mathbf{r}_1'\mathbf{x}_n - t_x' \\ u_{ny}\,(1+\varepsilon_{dn}) - \mathbf{r}_2'\mathbf{x}_n - t_y' \end{bmatrix}.$$

Hence, by solving Eq. (19), the best parameters for the facial expression associated with the newly estimated head pose with respect to the scaled orthographic projections of the feature points estimated in the previous iteration can be obtained. However, to ensure that the estimated parameters for the facial expression are all in the allowed ranges, the priori knowledge about the muscle contraction values and the amount of jaw rotation must be considered, and estimation of the parameters of the facial expression becomes a constrained minimization problem **QP**:

$$\mathbf{QP}: \ \arg\min_{\mathbf{c}} \|\mathbf{Gc} - \mathbf{g}\|_2^2 \\ \text{subject to the inequalities (7).}$$

Since Inequalities (7) are all linear, $\|\mathbf{Gc} - \mathbf{g}\|_2^2$ is bounded, and $\mathbf{G}^t\mathbf{G}$ is a positive semidefinite symmetric matrix, the non-linear programming problem **QP** is a linearly constrained quadratic programming problem, and the minimum of **QP** can be always obtained in polynomial time (Bazaraa et al., 1993; Nesterov and Nemirovskii, 1994).

4.1.2. A Geometric Interpretation of Algorithm 2. The parameters for the head pose and the facial expression estimated at the current iteration are computed from the scaled orthographic projections of the feature points obtained in the previous iteration. Step 2 through Step 5 compute the head pose from the scaled orthographic projections without the terms caused by the local motions estimated in the previous iteration; therefore, these steps can be regarded as computing the pose of a rigid object. After estimating the head pose, the parameters for the facial expression are obtained from the scaled orthographic projections without the terms caused by the newly estimated head pose. These steps repeat until no significant changes between consequent iteration.

4.2. Determination of States of Upper Eyelids

In this study, the state of an upper eyelid is determined by the ratio of the distance w between the two corners of the eye, and the length h of the black region around the eye in the direction perpendicular to the line formed by the two eyes' corners. The decision rule for determining the state of an upper eyelid is defined as follows:

$$\text{state of upper eyelid} \\ = \begin{cases} \text{opened} & \text{if } s_1 < \dfrac{w}{h} \\ \text{partially opened} & \text{if } s_2 < \dfrac{w}{h} \leq s_1, \\ \text{closed} & \text{if } \dfrac{w}{h} \leq s_2 \end{cases}$$

where s_1 and s_2 are determined by experiments.

5. Analysis of Convergency Property and Complexity

In this section, the convergency property and complexity of the proposed method are analyzed. First, the convergency property of the proposed method is discussed.

5.1. Analysis of Convergency

In DeMenthon and Davis's work (DeMenthon and Davis, 1995), a geometric interpretation for the convergency of DeMenthon and Davis's method was given. In this study, the convergency property of the proposed method is interpreted via proposed properties for the convergency of DeMenthon and Davis's method. For simplicity, the noise effect is not considered in this analysis. First, some new results of the convergency properties of the modified DeMenthon and Davis's method are proposed, including a sufficient condition about reducing the residue in each iteration, and a sufficient criterion for the global convergency. Then, the convergency property of the proposed method is illustrated based on the result of the modified DeMenthon and Davis's method.

5.1.1. Sufficient Criteria for Convergency of Modified DeMenthon and Davis's Method.

Scaled orthographic projection equations for the n feature points with respect to the pose parameters obtained at the kth iteration can be written as

$$\mathbf{u}_i^{(k)}\left(1+\varepsilon_i^{(k)}\right)=\begin{bmatrix}\mathbf{r}_1'^{(k)}\\ \mathbf{r}_2'^{(k)}\end{bmatrix}\mathbf{x}_i+\begin{bmatrix}t_x'^{(k)}\\ t_y'^{(k)}\end{bmatrix},\quad i=1,2,\ldots,n, \tag{20}$$

where $\mathbf{u}_i^{(k)}$ is the perspective projection for the ith feature point with respect to the pose parameters estimated at the kth iteration, and $\varepsilon_i^{(k)}=\mathbf{r}_3'^{(k)}\mathbf{x}_i$. It can be easily obtained that the residue in the kth iteration is

$$\sqrt{\sum_{i=1}^{n}\left\|\mathbf{u}_i\left(1+\varepsilon_i^{(k)}\right)-\mathbf{u}_i^{(k)}\left(1+\varepsilon_i^{(k)}\right)\right\|_2^2},$$

and that the residue in Eq. (9) for the $k+1$th iteration is

$$\sqrt{\sum_{i=1}^{n}\left\|\mathbf{u}_i\left(1+\varepsilon_i^{(k)}\right)-\mathbf{u}_i^{(k+1)}\left(1+\varepsilon_i^{(k+1)}\right)\right\|_2^2}.$$

Obviously, if the residue in each iteration is monotonically decreasing and if the rigid object is in an unambiguous state, the modified DeMenthon and Davis's method can converge to a solution with the desired precision after a finite number of iteration. In this study, a sufficient criteria about reducing the residue in an iteration is derived as follows.

Lemma 1. *In Algorithm* 1, *the residue in the* $k+1$*th iteration is not larger than that in the kth iteration if the following inequality is satisfied*:

$$\left\|\mathbf{r}_3'^{(k+1)}-\mathbf{r}_3'^{(k)}\right\|_2\sqrt{\sum_{i=1}^{n}\|\mathbf{u}_i\|_2^2\|\mathbf{x}_i\|_2^2}\leq h^{(k+1)}, \tag{21}$$

where

$$h^{(k+1)}=\sqrt{\sum_{i=1}^{n}\left\|\mathbf{u}_i\left(1+\varepsilon_i^{(k)}\right)-\mathbf{u}_i^{(k)}\left(1+\varepsilon_i^{(k)}\right)\right\|_2^2}-\sqrt{\sum_{i=1}^{n}\left\|\mathbf{u}_i\left(1+\varepsilon_i^{(k)}\right)-\mathbf{u}_i^{(k+1)}\left(1+\varepsilon_i^{(k+1)}\right)\right\|_2^2}.$$

Proof: See Appendix A.1. □

In Inequality (21), $\sqrt{\sum_{i=1}^{n}\|\mathbf{u}_i\|_2^2\|\mathbf{x}_i\|_2^2}$ is constant. Moreover, the smaller $\sqrt{\sum_{i=1}^{n}\|\mathbf{u}_i\|_2^2\|\mathbf{x}_i\|_2^2}$ is, the more relaxed Inequality (21) becomes. In the following, we show that $\sqrt{\sum_{i=1}^{n}\|\mathbf{u}_i\|_2^2\|\mathbf{x}_i\|_2^2}$ is related to the global convergency property of the modified DeMenthon and Davis's method. First, some definitions are given as follows.

Let $\mathbf{M}_{2\times 3}(\Re)$ denote all 2×3 real matrices. Let Γ denote the set of matrices in $\mathbf{M}_{2\times 3}(\Re)$ such that for every $\mathbf{W}\in\Gamma$, the two row vectors of $\mathbf{W}$ are orthogonal and the 2-norms of the two row vectors are equal. Since ε_i is equal to $\|\mathbf{r}_1'\|_2^{-1}(\mathbf{r}_1'\times\mathbf{r}_2')^t\mathbf{x}_i$, it can be noticed that Eq. (12) can be regarded as a function Φ mapping Γ to $\mathbf{M}_{2\times 3}(\Re)$. Similarly, Step 3′ and Step 4 can also be regarded as a function Ψ mapping $\mathbf{M}_{2\times 3}(\Re)$ to Γ. Therefore, when $k\geq 1$, the iterative formula of the modified DeMenthon and Davis's method can be written as

$$\mathbf{W}^{(k+1)}=\Psi\left(\Phi\left(\mathbf{W}^{(k)}\right)\right), \tag{22}$$

where $\mathbf{W}^{(k)}$ denotes the value of the matrix $\mathbf{W}$ at the kth iteration. From Eq. (22), we know that the modified DeMenthon and Davis's method can be regarded as a fixed-point iteration method (Stoer and Buliesch, 1993). For further analysis, let $\mathbf{R}^*=[\mathbf{r}_1^{*t}\ \mathbf{r}_2^{*t}\ \mathbf{r}_3^{*t}]^t$ and $\mathbf{t}^*=[t_x^*\ t_y^*\ t_z^*]$ denote the actual rotation matrix and translation vector, respectively, and let $\mathbf{W}^*=\frac{1}{t_z^*}[\mathbf{r}_1^{*t}\ \mathbf{r}_2^{*t}]^t$. It can be easily shown that $\mathbf{W}^*=\Phi(\mathbf{W}^*)$ and $\mathbf{W}^*=\Psi(\mathbf{W}^*)$.

To analyze the global convergency property of the modified DeMenthon and Davis's method, Lemma 2 is derived for illustrating the relation between the two matrices mapped by Φ from two matrices in Γ as follows.

Lemma 2. *For all* $\mathbf{W}$ *and* $\hat{\mathbf{W}}\in\Gamma$, *if the rank of* $\bar{\mathbf{X}}$ *is three*, *we have*

$$\|\Phi(\mathbf{W})-\Phi(\hat{\mathbf{W}})\|_F\leq C(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{u}_1,\ldots,\mathbf{u}_n)\|\mathbf{W}-\hat{\mathbf{W}}\|_F, \tag{23}$$

where

$$C(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{u}_1,\ldots,\mathbf{u}_n)=\|\bar{\mathbf{X}}^+\|_s\sqrt{\sum_{i=1}^{n}\|\mathbf{u}_i\|_2^2\|\mathbf{x}_i\|_2^2},$$

with $\|\cdot\|_s$ *representing the spectral matrix norm.*

Proof: See Appendix A.2. □

Based on Lemma 2, three corollaries about the convergency properties of the modified DeMenthon and Davis's method are derived in this study as follows.

Corollary 1. *If* $C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n) < 1$, *the state of the rigid object is unambiguous.*

Proof: See Appendix A.5. □

Corollary 2. *If the state of the rigid object is ambiguous, the ambiguity can be resolved by place the object away from the camera until* $C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n) < 1$.

Proof: Note that the optical axis has been aligned to pass through the gravity center of the projections of the feature points. Hence, moving the object away from the camera decreases $\|\mathbf{u}_i\|_2, i = 1, 2, \ldots, n$, and $C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n)$ becomes smaller. According to Corollary 1, once $C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n)$ is smaller than one, the ambiguity is resolved. □

Corollary 3. *If* $C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n) < 0.5$, *the modified DeMenthon and Davis's method is globally convergent.*

Proof: See Appendix A.6. □

Corollary 1 states a sufficient criterion for unambiguity of the state of a rigid object. Corollary 2 describes a method for resolving the ambiguity of the state of a rigid object which is useful for configuring a system for pose estimation. The intuitive meaning of this method is to make the projections of the feature points of the rigid object look like scaled orthographic projections. Corollary 3 describes a sufficient criterion for global convergency of the modified DeMenthon and Davis's method. Since the value of $C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n)$ is dependent on the structure and the pose of a rigid object, the convergency property of the modified DeMenthon and Davis's method is related to the pose and the structure of the rigid object.

To verify the relation between $C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n)$ and the convergency property of the modified DeMenthon and Davis's method, computer simulations were conducted. The eight corners of a cube in size of 20^3 were utilized as feature points in the computer simulations where the origin of the object coordinate system is the gravity center of the cube. Six thousand samples of the cube in arbitrary orientations were generated. The x- and y-coordinates of the samples with respect to the camera coordinate system are distributed uniformly in $[-5 \cdots 5]$ and $[-5 \cdots 5]$, respectively, and the z-coordinates of them are ranged from eleven to sixty-six at an interval of five. The maximum number of iterations was one hundred. Shown in Fig. 3(a) are some experimental results of the value of $C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n)$ versus the average number of iterations for the modified DeMenthon and Davis's method to be convergent. It shows that the rate of convergency of the modified DeMenthon and Davis's method is one hundred percent when $C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n)$ is smaller than 0.5. In addition, we can find that the number of iterations increases when $C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n)$ gets larger. Figure 3(b) are the results of $C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n)$ versus the z-coordinate of the center of the cube in the camera coordinate system. It shows that $C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n)$ decreases when the object moves away from the camera. Figure 3(c) shows that $C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n)$ can act as an indicator of the convergency of the modified DeMenthon and Davis's method.

5.1.2. Convergency Analysis of Proposed Method. The aim of the iterative formula of the proposed method is to find the parameters which minimize the residue in Eq. (14). If the proposed method is convergent for an unambiguous state of a face, the estimated parameters are the desired ones because they satisfy Eq. (14). Like Lemma 1 for Algorithm 1, a sufficient criterion for the convergency property of Algorithm 2 about reducing the residue in an iteration can be obtained as follows.

Lemma 3. *In Algorithm* 2, *the residue in the* $k+1$*th iteration is not larger than that in the kth iteration if the following inequality is satisfied*:

$$\begin{aligned}
&\left\|\mathbf{r}_3^{\prime(k+1)} - \mathbf{r}_3^{\prime(k)}\right\|_2 \sqrt{\sum_{i=1}^{n} \|\mathbf{u}_i\|_2^2 \|\mathbf{x}_i\|_2^2} \\
&\quad + \sqrt{\sum_{i=1}^{n} \left\|2\mathbf{u}_i\left(\mathbf{r}_3^{\prime(k+1)} \Delta\mathbf{x}_i^{(k+1)} - \mathbf{r}_3^{\prime(k)} \Delta\mathbf{x}_i^{(k)}\right)\right\|_2^2} \\
&\le \sqrt{\sum_{i=1}^{n} \left\|\mathbf{u}_i\left(1 + \varepsilon_{di}^{(k)}\right) - \mathbf{u}_i^{(k)}\left(1 + \varepsilon_{di}^{(k)}\right)\right\|_2^2} \\
&\quad - \sqrt{\sum_{i=1}^{n} \left\|\mathbf{u}_i\left(1 + \varepsilon_{di}^{(k)}\right) - \mathbf{u}_i^{(k+1)}\left(1 + \varepsilon_{di}^{(k+1)}\right)\right\|_2^2},
\end{aligned} \tag{24}$$

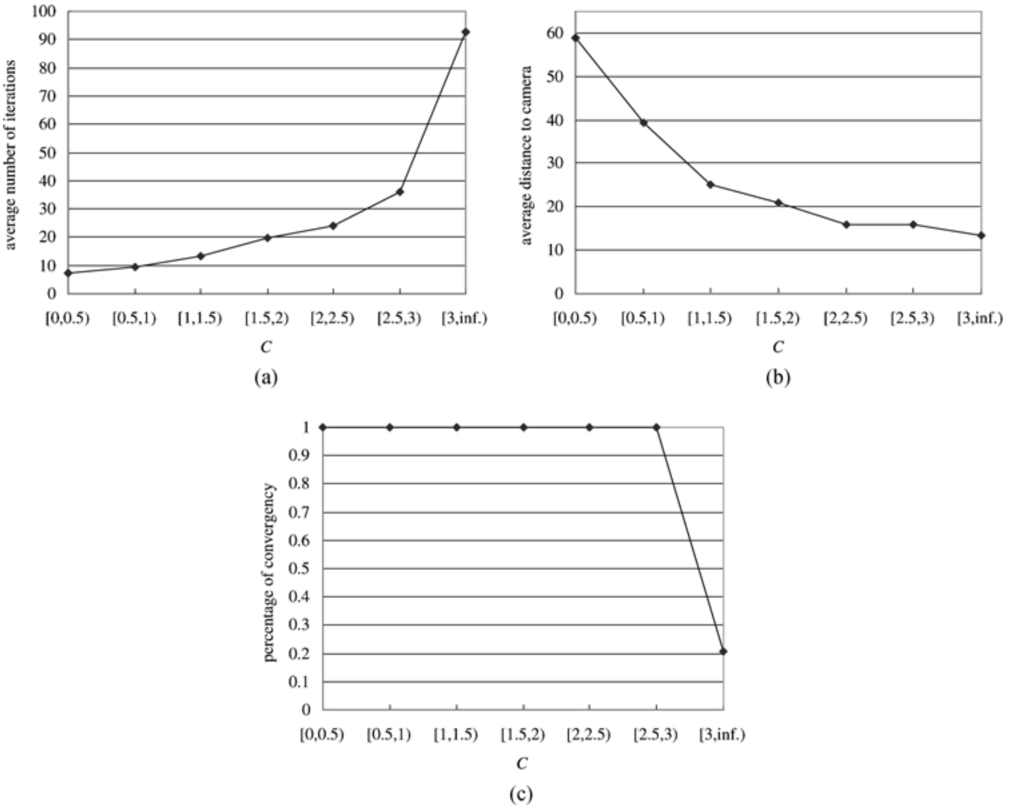


Figure 3. The simulation results of C versus the number of iterations, the average distance to the camera, and the percentage of convergency: (a) illustrates the result of C versus the number of iterations; (b) illustrates the result of C versus the average distance to camera; (c) illustrates the result of C versus the percentage of convergency.

where $\Delta\mathbf{x}_i^{(k)}$ denotes the local movement of the ith feature point estimated at the kth iteration.

Proof: The derivation is omitted because it is similar to that of Lemma 1. □

Currently, the global convergency property of the proposed method similar to that of the modified DeMenthon and Davis's method has not been derived. However, in fact, the maximum amounts of local movements of the feature points except those on the lower jaw are not large, and thus the left-hand side of Inequality (24) is dominated by its first term. Therefore, $\sqrt{\sum_{i=1}^{n}\|\mathbf{u}_i\|_2^2\|\mathbf{x}_i\|_2^2}$ is highly related to the convergency of the proposed method for each iteration, and $C(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{u}_1,\ldots,\mathbf{u}_n)$ can also be an index for global convergency of the proposed method.

5.2. *Analysis of Complexity*

In the proposed method, Step 6 in Algorithm 2 is the critical step which solves the linearly constrained quadratic programming problem **QP**. As shown in Nesterov and Nemirovskii (1994), by the logarithmic barrier algorithm with Karmarkar acceleration, it takes $\ln(\frac{m}{\alpha(G:w)\varepsilon})$ iterations and $O(mn^2+m^{1.5}n)$ arithmetic cost for every iteration to find an ε-solution of a convex quadratic programming problem with n variables and m linear constrains, where $\alpha(G:w)=\max\{t \mid w+t(w-G)\subset G\}$ is the asymmetry coefficient of G with respect to the starting point $w\in\{\alpha_j<c_j<\beta_j,\ j=1,2,\ldots,m+1\}$. Since **QP** has nineteen variables and thirty-eight linear constrains, it takes $O(38\times 19^2+19\times 38^{1.5})\ln(\frac{38}{\alpha(G:w)\varepsilon})$ arithmetic cost to find an ε-solution of **QP**.

By using the successive quadratic programming (SQP) method (Bazaraa et al., 1993), which is a popular method for the non-linear programming problem, to estimate the parameters for the head pose and the facial expression, one has to deal with twenty-five variables and thirty-eight linear constrains and thus needs $O(38 \times 25^2 + 25 \times 38^{1.5}) \ln(\frac{38}{\alpha(G:w)\varepsilon})$ arithmetic cost to get a solution. Moreover, the SQP method needs extra overhead to compute a 25×25 Hessian matrix for every iteration. Hence, the proposed method is at least 1.63 times faster than the SQP method for every iteration.

6. Experimental Results

In this section, we describe the results of testing the proposed method by computer simulations and using real images. Since the sensitivity of the parameters for the head pose and the facial expression is distinct for each other, it is inappropriate to evaluate the estimated results directly via the estimated parameters. Alternatively, the errors of the estimated results are defined as:

(1) *the average image error*: the average of the errors between the estimated positions and the theoretical positions of the feature points in the image;
(2) *the average 3-D relative global error*: the average of the relative errors between the estimated 3-D positions and the theoretical 3-D positions of the feature points; and
(3) *the average 3-D relative local error*: the average of the relative errors between the estimated 3-D positions and the theoretical 3-D positions of the feature points with respect to the face coordinate system.

In the following are the experimental results of the proposed method analyzed by computer simulations, including:

(1) the relation between C and the rate of convergency;
(2) error analysis;
(3) the rate of convergency;

and tested by using real images, including:

(1) the quality of the parameters estimated by the proposed method;
(2) comparisons between the results estimated by the proposed method and the results estimated by the SQP method.

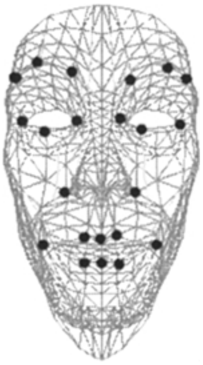


Figure 4. A computer graphic face model used for simulations where the black points indicate the positions of the twenty-two feature points.

6.1. Computer Simulations

In this study, the 3-D face model in Appendix A of the book (Parke and Waters, 1996) was utilized to test the proposed method. As shown in Fig. 4, twenty-two vertices on the face distributed in a $6.57 \times 7.6 \times 2.29$ cm^3 box were selected as the feature points for this simulation. The ranges for yaw, pitch, and roll, were randomly generated within ±22.5°, ±22.5°, and ±30°, respectively. The eighteen muscle contraction values were generated from -1 to 1, and the amount of jaw rotation was produced from 0° to 10° randomly. The x, y-coordinates of the test samples were generated within ±30 cm uniformly. The focal length of the camera in this experiment is set to be 350 pixels.

6.1.1. Relation between C and the Rate of Convergency. To show that $C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n)$ can be an index of convergency of the proposed method, one thousand test samples without noise were generated. The distances between the test samples and the origin of the camera coordinate system were from 10 cm to 80 cm. The experimental results are shown in Fig. 5. They reveal that $C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n)$ is related to the rate of convergency. In addition, from the value of $C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n)$, the proper working distance from the origin of the camera coordinate system to the adopted 3-D face model is suggested not to be smaller than 40 cm. This suggested value was utilized in the subsequent analysis.

6.1.2. Error Analysis. In this experiment, five hundred test samples were generated and the perspective projections of the feature points of every test sample

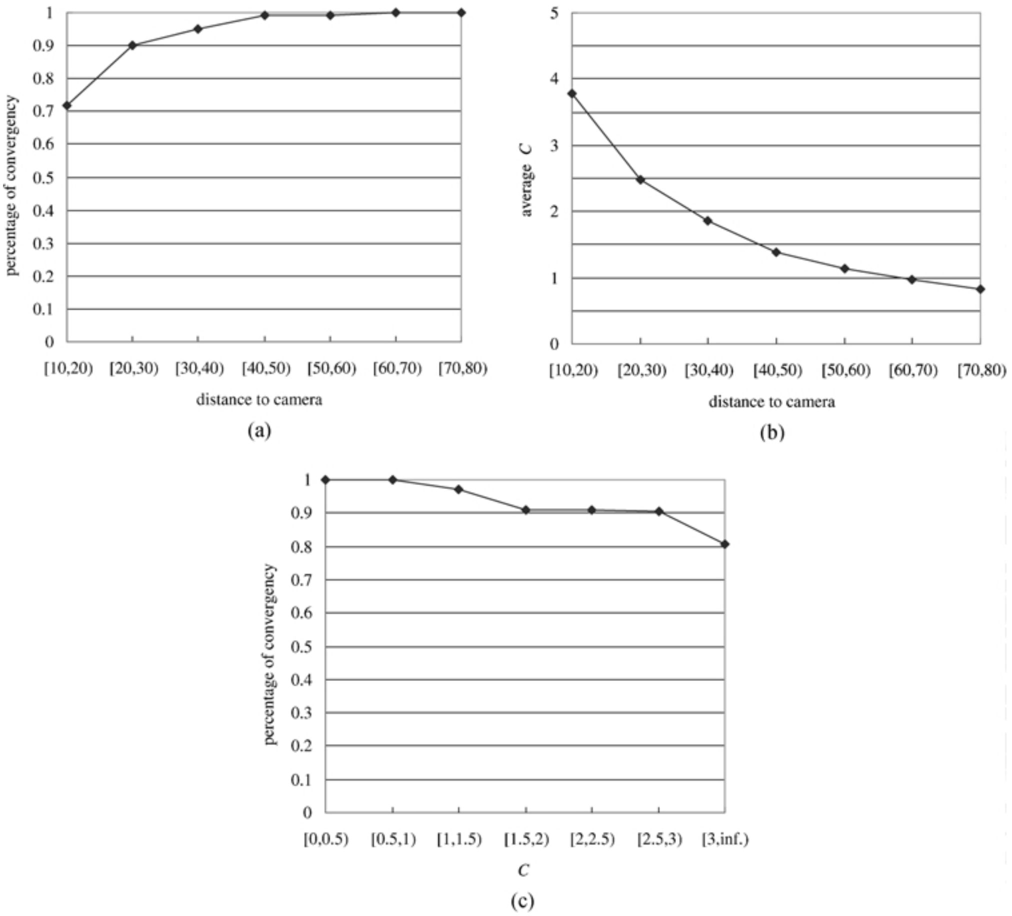


Figure 5. The simulation results of the relation among the distance to the camera, C, and the percentage of convergency: (a) illustrates the result of the distance to the camera versus the average percentage of convergency; (b) illustrates the result of the distance to the camera versus the average C; (c) illustrates the result of C versus the average percentage of convergency.

were perturbed by Gaussian noise with standard deviations 0, 1, 2, 3, 4, and 5. The distances between the test samples and the origin of the camera coordinate system were ranged from forty cm to seventy cm. In this simulation, the average maximum local motion and the average local motion of the feature points of a test sample were 1.75 cm and 0.53 cm, respectively. The experimental results are as follows. The average image errors versus various noise levels are computed and plotted in Fig. 6(a), and the average 3-D relative global errors and the average 3-D relative local errors are analyzed and shown in Fig. 6(b) and (c) where the averages of the errors with their standard deviation bars versus various noise levels are plotted. Figure 6(b) and (c) show that the proposed method gives the parameters with less than five percent of 3-D relative global errors and less than five percent of 3-D relative local errors, in average. The stability of the proposed method in estimation of global motion and local motion is so ensured.

6.1.3. Rate of Convergency. Figure 7(a) shows the average number of iterations versus various noise levels. In general, the test samples perturbed by noise with higher noise levels have larger residues in Eq. (14), so the proposed method takes smaller numbers of iterations to reach the termination condition. Figure 7(b) shows the percentage of convergency for the proposed method to estimate parameters from the test samples versus various noise levels. The overall average percentage of convergency is 99.23%.

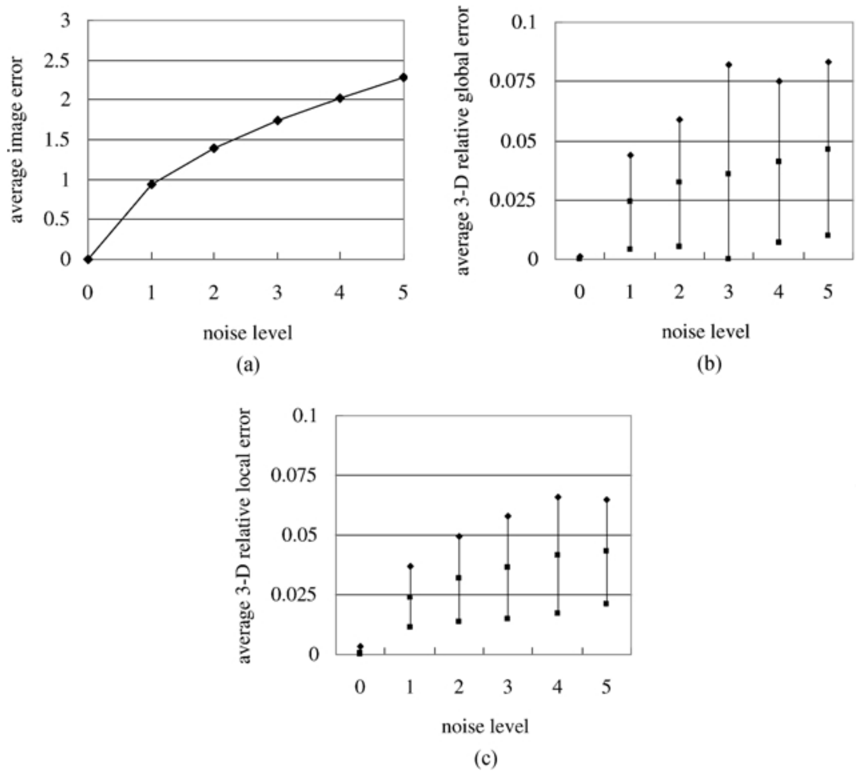


Figure 6. The simulation results of the quality of the estimated results versus various noise levels: (a), (b) and (c) illustrate the result of the average image errors, the average 3-D relative global errors and the average 3-D relative local errors, respectively.

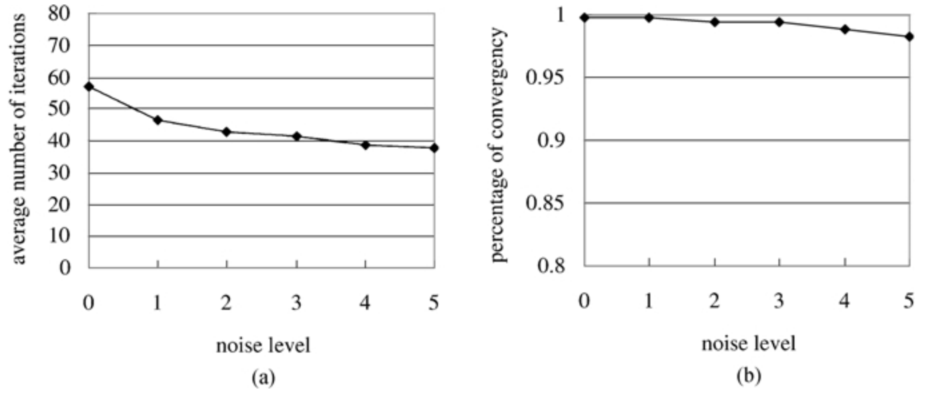


Figure 7. The simulation results of analysis of percentage of convergency versus various noise levels: (a) illustrates the result of the average number of iterations versus various noise levels; (b) illustrates the result of the percentage of convergency versus various noise levels.

6.2. Tests with Real Images

To estimate the parameters for the head pose and the facial expression, the 3-D positions of the feature points and the muscle vectors for an individual human face with respect to the face coordinate system must be known beforehand. In this study, the proposed method was implemented on a PentiumII 400 MHz PC and twenty-two feature points on a human face shown in Fig. 8 were detected automatically. The 3-D positions of the feature points were estimated by the structure-from-motion algorithm proposed by

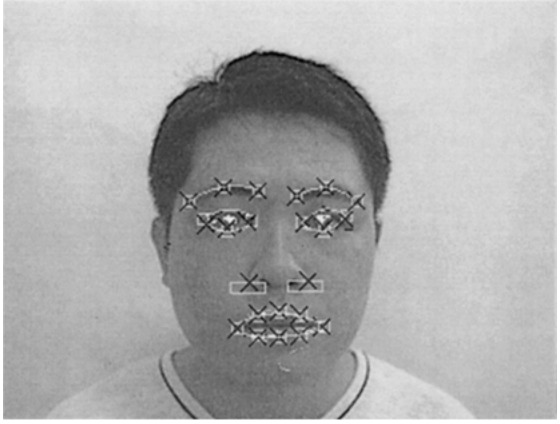


Figure 8. Twenty-two feature points indicated by crosses.

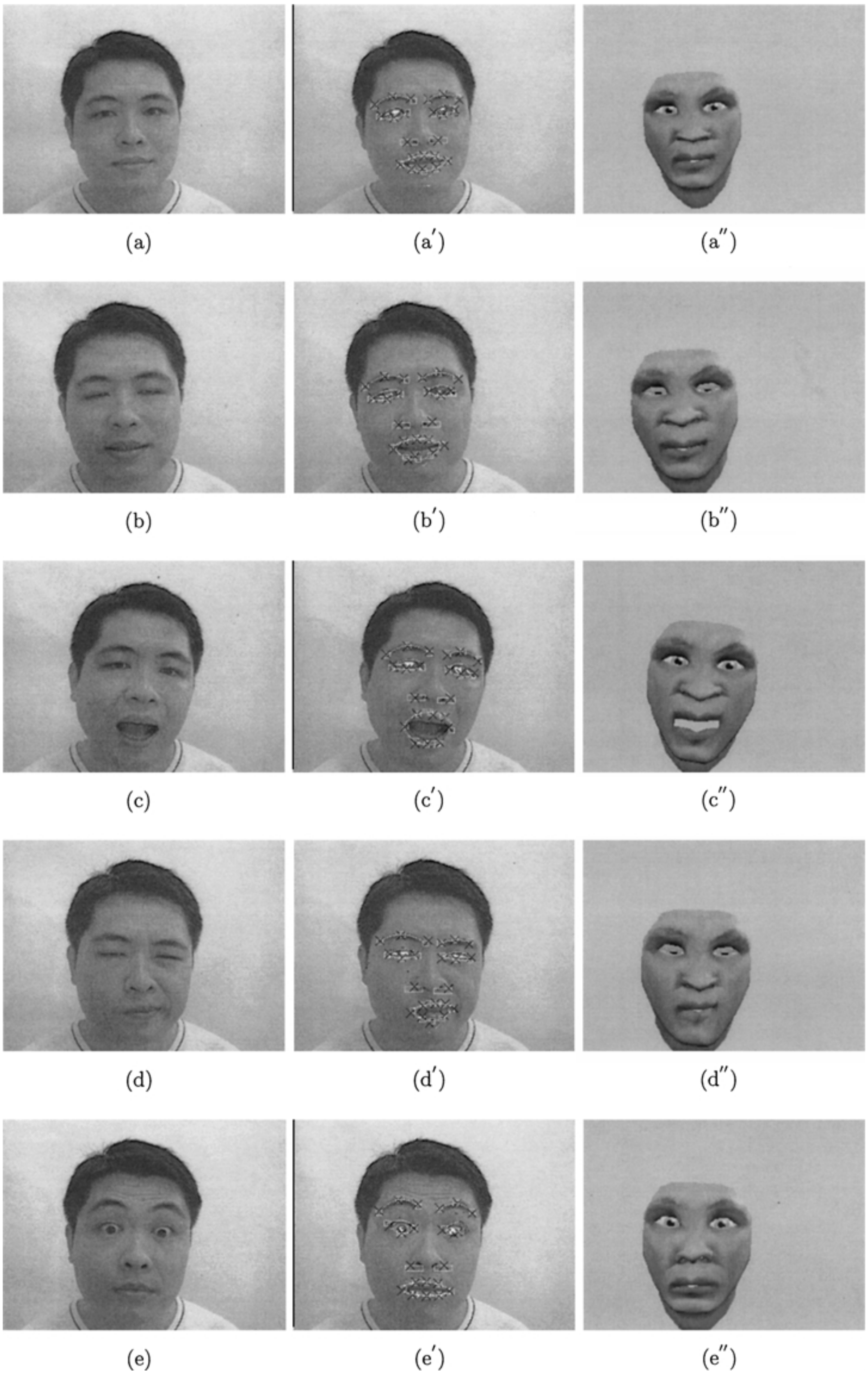
Horn (Horn, 1990). The errors of the 3-D positions of the feature points built in this manner are about 6%. Besides, based on a generic 3-D face model, the points of the muscle fibers attached on the skin tissue were obtained by an interpolation method (Akimoto et al., 1993), and the lengths and directions of the muscle fibers were obtained by an affine transform. In fact, the 3-D face model obtained in this manner is not accurate. Thus, the error sources in this experiment include at least the image error and the model error. For the sake of shortening this paper, the details of constructing an individual 3-D face model and detecting feature points are omitted.

Since it is hard to measure the actual 3-D positions of the feature points with respect to the camera, the quality of the estimated parameters was evaluated by the average image error. Figure 9(a)–(j) show ten test images in size of 320 × 240. The focal length of the camera used in this experiment is 366 pixels. For comparisons, the ten test images were utilized to test not only the proposed method but also the SQP method. The initial guess for the SQP method was a face facing to the camera with no expression. The termination criterion for the method is any of the three cases: the average image error is smaller than 0.001, the distance that the estimated parameters advance in an iteration is smaller than 0.01, or the number of iterations exceeds one hundred.

Figure 9(a′)–(j′) show the detected feature points for the test images, and Fig. 9(a″) through (j″) are the synthesized images for the ten test images with the parameters estimated by the proposed method. Table 1 shows the details of the experimental results. The values of C for the ten test images are all smaller than one. It indicates that the configuration of this system is proper for the proposed method. The average image error of the solutions found by the proposed method is about 3.17 pixels, and the average computation time is 0.478 seconds. The average image error of the solutions found by the SQP method is about 12 pixels, and the average computation time is 5.55 seconds. For a single iteration, it takes about 21.3 ms and 55.53 ms for the proposed method and the SQP method, respectively. For the ten test images, the SQP method does not converge to good solutions because it needs better initial guesses. In our experience, although it is hard to obtain good initial guesses, once the SQP method converges to a good solution, the average image error of the solution is smaller than that of the solution found by the proposed method. The reason is that the SQP method estimates the parameters by minimizing the average image error. As a summary, the experimental results show that the proposed method is effective, robust, and better.

Table 1. Comparisons between the results estimated by the proposed method (denoted by SSOA) and the results estimated by the successive quadratic programming method (denoted by SQP).

	Image no.																			
	a		b		c		d		e		f		g		h		i		j	
C	0.83		0.9		0.95		0.89		0.99		0.93		0.84		0.81		0.93		0.9	
Method	SSOA	SQP	SSOA	SQP	SSOA	SQP	SSOA	SQP	SSOA	SQP	SSOA	SQP	SSOA	SQP	SSOA	SQP	SSOA	SQP	SSOA	SQP
Average image error	2.81	8.16	3.19	10	2.7	15.02	3.43	13.62	3.12	11.82	5.5	22.87	2.93	10.15	4.09	8.81	2.18	10.43	1.7	9.1
Number of iterations	20	100	25	100	20	100	22	100	14	100	19	100	24	100	23	100	32	100	25	100
Computation time (s)	0.33	6.15	0.83	4.56	0.5	6.81	0.38	6.1	0.22	3.84	0.44	4.72	0.55	6.27	0.38	5.05	0.49	6.1	0.66	5.93



(a) (a′) (a″)

(b) (b′) (b″)

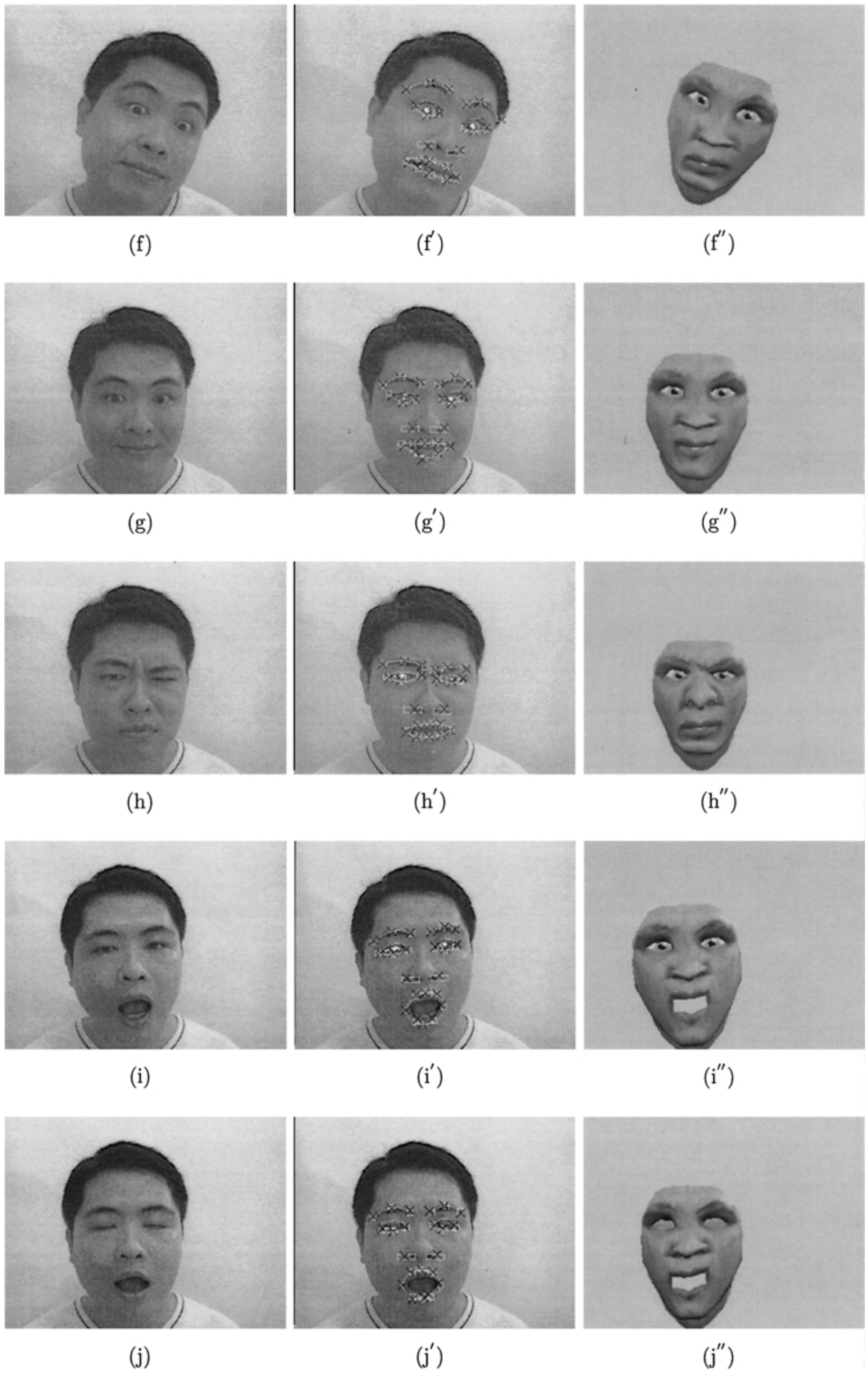
(c) (c′) (c″)

(d) (d′) (d″)

(e) (e′) (e″)

Figure 9. Ten test images: (a) through (j) are the original images; (a′) through (j′) show the detected feature points; and (a′) through (j′) are the synthesized images.

(*Continued on next page.*)



(f) (f′) (f″)

(g) (g′) (g″)

(h) (h′) (h″)

(i) (i′) (i″)

(j) (j′) (j″)

Figure 9. (*Continued*).

7. Conclusions

In this study, a new iterative method for estimating the parameters of the head pose and the facial expression from a single view has been proposed. The proposed method is based on the technique of successive scaled orthographic approximations, which has been successfully applied to estimating the pose and the structure of a rigid object. In this study, we have extended this technique to estimate the parameters for a non-rigid object, namely, a human face. In addition, we have derived a sufficient criterion for the convergency of the proposed method. Our theoretical analysis and the simulations both show that the convergency property of the proposed method is related to the proposed criteria. If the distance between the camera and the face is in the range suggested by the derived criterion, the proposed method without initial guesses can, as shown by our experimental results, have about one hundred percentage of convergency. The experimental results also show that the successive quadratic programming method, a general method for non-linear constrained optimization, often traps into local minima or cannot converge with respect to the test images used in our experiments. Hence, the proposed method is better than the successive quadratic programming method in estimating the head pose and the facial expression. The experimental results show finally that the proposed method is robust and can act successfully as a basis of a system for applications in human-computer interaction.

Appendix

A.1. Proof of Lemma 1

From the triangular inequality of the vector 2-norm, we can derive the following inequality describing the relation between the residue in the $k+1$th iteration and the residue in Eq. (9) for the $k+1$th iteration as follows:

$$\left\| \begin{matrix} \mathbf{u}_1\left(1+\varepsilon_1^{(k+1)}\right) - \mathbf{u}_1^{(k+1)}\left(1+\varepsilon_1^{(k+1)}\right) \\ \vdots \\ \mathbf{u}_n\left(1+\varepsilon_n^{(k+1)}\right) - \mathbf{u}_n^{(k+1)}\left(1+\varepsilon_n^{(k+1)}\right) \end{matrix} \right\|_2$$

$$\leq \left\| \begin{matrix} \mathbf{u}_1\left(1+\varepsilon_1^{(k)}\right) - \mathbf{u}_1^{(k+1)}\left(1+\varepsilon_1^{(k+1)}\right) \\ \vdots \\ \mathbf{u}_n\left(1+\varepsilon_n^{(k)}\right) - \mathbf{u}_n^{(k+1)}\left(1+\varepsilon_n^{(k+1)}\right) \end{matrix} \right\|_2$$

$$+ \left\| \begin{matrix} \mathbf{u}_1\left(\varepsilon_1^{(k+1)} - \varepsilon_1^{(k)}\right) \\ \vdots \\ \mathbf{u}_n\left(\varepsilon_n^{(k+1)} - \varepsilon_n^{(k)}\right) \end{matrix} \right\|_2 . \tag{25}$$

From Inequality (25) and the definition of $h^{(k+1)}$, if the following inequality is satisfied:

$$\left\| \begin{matrix} \mathbf{u}_1\left(\varepsilon_1^{(k+1)} - \varepsilon_1^{(k)}\right) \\ \vdots \\ \mathbf{u}_n\left(\varepsilon_n^{(k+1)} - \varepsilon_n^{(k)}\right) \end{matrix} \right\|_2 \leq h^{(k+1)}, \tag{26}$$

we can conclude

$$\left\| \begin{matrix} \mathbf{u}_1\left(1+\varepsilon_1^{(k+1)}\right) - \mathbf{u}_1^{(k+1)}\left(1+\varepsilon_1^{(k+1)}\right) \\ \vdots \\ \mathbf{u}_n\left(1+\varepsilon_n^{(k+1)}\right) - \mathbf{u}_n^{(k+1)}\left(1+\varepsilon_n^{(k+1)}\right) \end{matrix} \right\|_2$$

$$\leq \left\| \begin{matrix} \mathbf{u}_1\left(1+\varepsilon_1^{(k)}\right) - \mathbf{u}_1^{(k)}\left(1+\varepsilon_1^{(k)}\right) \\ \vdots \\ \mathbf{u}_n\left(1+\varepsilon_n^{(k)}\right) - \mathbf{u}_n^{(k)}\left(1+\varepsilon_n^{(k)}\right) \end{matrix} \right\|_2 ;$$

that is, the residue in the $k+1$th iteration is not greater than that in the kth iteration. Furthermore, it can be easily shown that

$$\left\| \begin{matrix} \mathbf{u}_1\left(\varepsilon_1^{(k+1)} - \varepsilon_1^{(k)}\right) \\ \vdots \\ \mathbf{u}_n\left(\varepsilon_n^{(k+1)} - \varepsilon_n^{(k)}\right) \end{matrix} \right\|_2$$

$$\leq \left\|\mathbf{r}_3'^{(k+1)} - \mathbf{r}_3'^{(k)}\right\|_2 \sqrt{\sum_{i=1}^{n} \|\mathbf{u}_i\|_2^2 \, \|\mathbf{x}_i\|_2^2},$$

so $\|\mathbf{r}_3'^{(k+1)} - \mathbf{r}_3'^{(k)}\|_2 \sqrt{\sum_{i=1}^{n} \|\mathbf{u}_i\|_2^2 \|\mathbf{x}_i\|_2^2} \leq h^{(k+1)}$ is a sufficient condition for Inequality (26). Accordingly, we have completed the proof of Lemma 1. □

A.2. Proof of Lemma 2

For convenience, let $\bar{\mathbf{U}}_\mathbf{W}$, $\bar{\mathbf{u}}_{\varepsilon\mathbf{W}}$, and $\varepsilon_{i\mathbf{W}}$ denote the matrix $\bar{\mathbf{U}}$, the vector $\bar{\mathbf{u}}_\varepsilon$, and the scalar ε_i in Eq. (11) with respect to $\mathbf{W}$, respectively, and let $\mathbf{U}_\mathbf{W}$ denote

$$\mathbf{U}_\mathbf{W} = [\mathbf{u}_1(1+\varepsilon_{1\mathbf{W}}) \cdots \mathbf{u}_n(1+\varepsilon_{n\mathbf{W}})].$$

Now, for every $\mathbf{W} = [\mathbf{r}_1'^t \ \mathbf{r}_2'^t]^t \in \Gamma$ and $\hat{\mathbf{W}} = [\hat{\mathbf{r}}_1'^t \ \hat{\mathbf{r}}_2'^t]^t \in \Gamma$, from Eq. (12) and the following inequality (the proof can be found in Appendix A.3):

$$\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F \|\mathbf{B}\|_s \tag{27}$$

where **A** and **B** are two arbitrary compatible matrices, we have

$$\begin{aligned}\|\Phi(\mathbf{W}) - \Phi(\hat{\mathbf{W}})\|_F &= \|\bar{\mathbf{U}}_{\mathbf{W}}\bar{\mathbf{X}}^+ - \bar{\mathbf{U}}_{\hat{\mathbf{W}}}\bar{\mathbf{X}}^+\|_F \\ &\leq \|\bar{\mathbf{U}}_{\mathbf{W}} - \bar{\mathbf{U}}_{\hat{\mathbf{W}}}\|_F \|\bar{\mathbf{X}}^+\|_s.\end{aligned} \tag{28}$$

Since

$$\begin{aligned}\|\bar{\mathbf{U}}_{\mathbf{W}} - \bar{\mathbf{U}}_{\hat{\mathbf{W}}}\|_F^2 = \sum_{i=1}^{n} \|&(\mathbf{u}_i(1 + \varepsilon_{i\mathbf{W}}) - \bar{\mathbf{u}}_{\varepsilon\mathbf{W}}) \\ &- (\mathbf{u}_i(1 + \varepsilon_{i\hat{\mathbf{W}}}) - \bar{\mathbf{u}}_{\varepsilon\hat{\mathbf{W}}})\|_2^2\end{aligned}$$

and $\bar{\mathbf{u}}_{\varepsilon\mathbf{W}} - \bar{\mathbf{u}}_{\varepsilon\hat{\mathbf{W}}}$ is the mean vector of $\mathbf{u}_i(1 + \varepsilon_{i\mathbf{W}}) - \mathbf{u}_i(1 + \varepsilon_{i\hat{\mathbf{W}}})$, $i = 1, 2, \ldots, n$, we have

$$\|\bar{\mathbf{U}}_{\mathbf{W}} - \bar{\mathbf{U}}_{\hat{\mathbf{W}}}\|_F \leq \|\mathbf{U}_{\mathbf{W}} - \mathbf{U}_{\hat{\mathbf{W}}}\|_F. \tag{29}$$

In addition, it can be easily shown that

$$\|\mathbf{U}_{\mathbf{W}} - \mathbf{U}_{\hat{\mathbf{W}}}\|_F \leq \sqrt{\sum_{i=1}^{n} \|\mathbf{u}_i\|_2^2 \|\mathbf{x}_i\|_2^2 \|\mathbf{r}_3' - \hat{\mathbf{r}}_3'\|_2}. \tag{30}$$

Hence, from Inequalities (28)–(30), we can obtain

$$\begin{aligned}&\|\Phi(\mathbf{W}) - \Phi(\hat{\mathbf{W}})\|_F \\ &\quad \leq C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n)\|\mathbf{r}_3' - \hat{\mathbf{r}}_3'\|_2,\end{aligned} \tag{31}$$

where

$$\begin{aligned}C\,&(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n) \\ &= \|\bar{\mathbf{X}}^+\|_s \sqrt{\sum_{i=1}^{n} \|\mathbf{u}_i\|_2^2\, \|\mathbf{x}_i\|_2^2}.\end{aligned}$$

As proved in Appendix A.4, we have

$$\|\mathbf{r}_3' - \hat{\mathbf{r}}_3'\|_2 \leq \left\| \begin{bmatrix} \mathbf{r}_1' \\ \mathbf{r}_2' \end{bmatrix} - \begin{bmatrix} \hat{\mathbf{r}}_1' \\ \hat{\mathbf{r}}_2' \end{bmatrix} \right\|_F. \tag{32}$$

From Inequalities (31) and (32), we can conclude

$$\begin{aligned}&\|\Phi(\mathbf{W}) - \Phi(\hat{\mathbf{W}})\|_F \\ &\quad \leq C\big(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n\big)\|\mathbf{W} - \hat{\mathbf{W}}\|_F. \quad \square\end{aligned}$$

A.3. Proof of Inequality (27)

Let $\mathbf{A} \in \mathbf{M}_{l\times m}(\Re)$ and $\mathbf{B} \in \mathbf{M}_{m\times n}(\Re)$, and let the row vectors of **A** be denoted by $\mathbf{a}_i, i = 1, 2, \ldots, l$. From the definition of the Frobenius matrix norm, we have $\|\mathbf{AB}\|_F^2 = \sum_{i=1}^n \|\mathbf{a}_i\mathbf{B}\|_2^2$. Since the spectral matrix norm is compatible with the vector 2-norm (Horn and Johnson, 1985), we can obtain

$$\sum_{i=1}^{n} \|\mathbf{a}_i\mathbf{B}\|_2^2 \leq \left(\sum_{i=1}^{n} \|\mathbf{a}_i\|_2^2\right) \|\mathbf{B}\|_s^2 = \|\mathbf{A}\|_F^2\, \|\mathbf{B}\|_s^2.$$

Hence, we have $\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F\|\mathbf{B}\|_s$, and this completes the proof. □

A.4. Proof of Inequality (32)

The proof of the inequality can be simplified by multiplying the matrices and the vectors of the both sides of Inequality (32) by a rotation matrix $\mathbf{R} = [\mathbf{r}_1^t\ \mathbf{r}_2^t\ \mathbf{r}_3^t]$ because this operation does not change the norms of the both sides of Inequality (32). Thus, we can have

$$\|\mathbf{r}_3' - \hat{\mathbf{r}}_3'\|_2^2 = \left\|\begin{bmatrix} 0 & 0 & \frac{1}{t_z} \end{bmatrix} - \check{\mathbf{r}}_3'\right\|_2^2, \tag{33}$$

$$\left\| \begin{bmatrix} \mathbf{r}_1' \\ \mathbf{r}_2' \end{bmatrix} - \begin{bmatrix} \hat{\mathbf{r}}_1' \\ \hat{\mathbf{r}}_2' \end{bmatrix} \right\|_F^2 = \left\| \begin{bmatrix} \frac{1}{t_z} & 0 & 0 \\ 0 & \frac{1}{t_z} & 0 \end{bmatrix} - \begin{bmatrix} \check{\mathbf{r}}_1' \\ \check{\mathbf{r}}_2' \end{bmatrix} \right\|_F^2, \tag{34}$$

where $[\check{\mathbf{r}}_1'^t\ \check{\mathbf{r}}_2'^t\ \check{\mathbf{r}}_3'^t]^t = \frac{1}{\hat{t}_z}[\check{\mathbf{r}}_1^t\ \check{\mathbf{r}}_2^t\ \check{\mathbf{r}}_3^t]^t$ with $[\check{\mathbf{r}}_1^t\ \check{\mathbf{r}}_2^t\ \check{\mathbf{r}}_3^t]^t = [\hat{\mathbf{r}}_1^t\ \hat{\mathbf{r}}_2^t\ \hat{\mathbf{r}}_3^t]^t\mathbf{R}$. From Eqs. (33) and (34), it can be easily shown

$$\begin{aligned}&\left\| \begin{bmatrix} \mathbf{r}_1' \\ \mathbf{r}_2' \end{bmatrix} - \begin{bmatrix} \hat{\mathbf{r}}_1' \\ \hat{\mathbf{r}}_2' \end{bmatrix} \right\|_F^2 - \|\mathbf{r}_3' - \hat{\mathbf{r}}_3'\|_2^2 \\ &\quad = \frac{1}{\hat{t}_z^2} - \frac{2(\check{r}_{11} + \check{r}_{22} - \check{r}_{33})}{\hat{t}_z t_z} + \frac{1}{t_z^2}.\end{aligned}$$

Since $[\check{\mathbf{r}}_1^t\ \check{\mathbf{r}}_2^t\ \check{\mathbf{r}}_3^t]^t$ is a rotation matrix, we have $\check{r}_{33} = \check{r}_{11}\check{r}_{22} + \check{r}_{12}\check{r}_{21}$ and

$$\begin{aligned}&(\check{r}_{11} + \check{r}_{22})^2 + (\check{r}_{12} + \check{r}_{21})^2 \\ &\quad = 1 + \check{r}_{33}^2 + 2(\check{r}_{11}\check{r}_{22} + \check{r}_{12}\check{r}_{21}) \\ &\quad = 1 + \check{r}_{33}^2 + 2\check{r}_{33} \\ &\quad = (1 + \check{r}_{33})^2.\end{aligned}$$

Hence, we have $(\check{r}_{11}+\check{r}_{22})^2 \leq (1+\check{r}_{33})^2$. Furthermore, since $-1 \leq \check{r}_{33} \leq 1$, we also have $\check{r}_{11}+\check{r}_{22} \leq 1+\check{r}_{33}$. Accordingly, we obtain

$$\frac{1}{\hat{t}_z^2} - \frac{2(\check{r}_{11}+\check{r}_{22}-\check{r}_{33})}{\hat{t}_z t_z} + \frac{1}{t_z^2} \geq \frac{1}{\hat{t}_z^2} - \frac{2}{\hat{t}_z t_z} + \frac{1}{t_z^2} \geq 0.$$

That is, we have

$$\left\| \begin{bmatrix} \mathbf{r}'_1 \\ \mathbf{r}'_2 \end{bmatrix} - \begin{bmatrix} \hat{\mathbf{r}}'_1 \\ \hat{\mathbf{r}}'_2 \end{bmatrix} \right\|_F^2 - \|\mathbf{r}'_3 - \hat{\mathbf{r}}'_3\|_2^2 \geq 0.$$

Hence, we have completed the proof of Inequality (32). □

A.5. Proof of Corollary 1

If the pose of the object is ambiguous, there exist at least two matrices $\mathbf{W}$ and $\hat{\mathbf{W}} \in \Gamma$, and $\mathbf{W} \neq \hat{\mathbf{W}}$ such that $\Phi(\mathbf{W}) = \mathbf{W}$ and $\Phi(\hat{\mathbf{W}}) = \hat{\mathbf{W}}$. From Eq. (23), we can have

$$\begin{aligned}\|\mathbf{W} - \hat{\mathbf{W}}\|_F &= \|\Phi(\mathbf{W}) - \Phi(\hat{\mathbf{W}})\|_F \\ &\leq C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n)\|\mathbf{W} - \hat{\mathbf{W}}\|_F.\end{aligned}$$

The above inequality holds only if $C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n) \geq 1$. Accordingly, we have proved the corollary. □

A.6. Proof of Corollary 3

From Eq. (23), we can have

$$\begin{aligned}&\|\mathbf{W} - \mathbf{W}^*\|_F \\ &\geq \frac{1}{C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n)}\|\Phi(\mathbf{W}) - \mathbf{W}^*\|_F,\end{aligned} \tag{35}$$

for an arbitrary $\mathbf{W} \in \Gamma$. From the definition of Ψ, we have that $\Psi(\Phi(\mathbf{W}))$ is the matrix in Γ closest to $\Phi(\mathbf{W})$ in the Frobenius matrix norm and obtain

$$\|\Phi(\mathbf{W}) - \mathbf{W}^*\|_F \geq \|\Phi(\mathbf{W}) - \Psi(\Phi(\mathbf{W}))\|_F.$$

Thus, we can have

$$\begin{aligned}\|\Phi(\mathbf{W}) - \mathbf{W}^*\|_F \geq \frac{1}{2}(&\|\Phi(\mathbf{W}) - \mathbf{W}^*\|_F \\ &+ \|\Phi(\mathbf{W}) - \Psi(\Phi(\mathbf{W}))\|_F).\end{aligned}$$

In addition, from the triangle inequality for the Frobenius matrix norm, we can have

$$\|\Phi(\mathbf{W}) - \mathbf{W}^*\|_F \geq \frac{1}{2}\|\Psi(\Phi(\mathbf{W})) - \mathbf{W}^*\|_F;$$

in addition, from Inequality (35), we obtain

$$\|\mathbf{W} - \mathbf{W}^*\|_F \geq \alpha\|\Psi(\Phi(\mathbf{W})) - \mathbf{W}^*\|_F,$$

where $\alpha = \frac{1}{2C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n)}$. Hence, for an arbitrary starting point $\mathbf{W}^{(0)} \in \Gamma$, we have

$$\left(\frac{1}{\alpha}\right)^k \left\|\mathbf{W}^{(0)} - \mathbf{W}^*\right\|_F \geq \left\|\mathbf{W}^{(k)} - \mathbf{W}^*\right\|_F.$$

Thus, if $\alpha > 1$; i. e., $C(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{u}_1, \ldots, \mathbf{u}_n) < 0.5$, we have

$$\lim_{k\to\infty}\left\|\mathbf{W}^{(k)} - \mathbf{W}^*\right\|_F \leq \lim_{k\to\infty}\left(\frac{1}{\alpha}\right)^k \left\|\mathbf{W}^{(0)} - \mathbf{W}^*\right\|_F = 0,$$

and conclude that the modified DeMenthon and Davis's method guarantees to converge to the actual object pose from every starting point. □

References

Akimoto, T., Suenaga, Y., and Wallace, R.S. 1993. Automatic creation of 3D facial models. *IEEE Computer Graphics & Applications*, 13(5):16–22.

Aloimonos, J.Y. 1990. Perspective approximations. *Image and Vision Computing*, 8(3):179–192.

Bascle, B. and Blake, A. 1998. Separability of pose and expression in facial tracking and animation. In *Proceedings of the Sixth International Conference on Computer Vision*, pp. 323–328.

Bazaraa, M.S., Sherali, H.D., and Shetty, C.M. 1993. *Nonlinear Programming Theory and Algorithms* (2nd edn.). John Wiley & Sons: New York.

Choi, C.S., Aizawa, K., Harashima, H., and Takebe, T. 1994. Analysis and synthesis of facial image sequences in model-based image coding. *IEEE Trans. on Circuits and Systems for Video Technology*, 4(3):257–275.

DeMenthon, D.F. and Davis, L.S. 1995. Model-based object pose in 25 lines of code. *International Journal of Computer Vision*, 15:123–141.

Horaud, R., Dornaika, F., Lamiroy, B., and Christy, S. 1997. Object pose: The link between weak perspective, paraperspective and full perspective. *International Journal of Computer Vision*, 22(2):173–189.

Horn, B.K.P. 1990. Relative orientation. *International Journal of Computer Vision*, 4:59–78.

Horn, R.A. and Johnson, C.R. 1985. *Matrix Analysis*. Cambridge Univ. Press: New York.

Kanatani, K. 1993. *Geometric Computation for Machine Vision*. Oxford Univ. Press: New York.

Lei, Y.W., Wu, J.L., and Ouhyoung, M. 1996. A three-dimensional muscle-based facial expression synthesizer for model-based image coding. *Singal Processing: Image Communication*, 8:353–363.

Li, H. and Forchheimer, R. 1994. Two-view facial movement estimation. *IEEE Trans. on Circuits and Systems for Video Technology*, 4(3):276–287.

Li, H., Roivainen, P., and Forchheimer, R. 1993. 3-D motion estimation in model-based facial image coding. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(6):545–555.

Nesterov Yurii and Nemirovskii Arkadii. 1994. *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM Studies in Applied Mathematics. SIAM: Philadelphia, PA.

Parke, F.I. and Waters, K. 1996. *Computer Facial Animation*. A.K. Peters: Wellesley, MA.

Stoer, J. and Bulirsch, R. 1993. *Introduction to Numerical Analysis* (2nd edn.). Springer-Verlag: New York.

Tao, H. and Huang, T.S. 1998. Bézier volume deformation model for facial animation and video tracking. In *Proceeding of International Workshop, CAPTECH'98: Modelling and Motion Capture Techniques for Virtual Environments*, Berlin, Germany, Springer-Verlag: Berlin, pp. 242–253.

Terzopoulos, D. and Waters, K. 1993. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(6):569–579.

Thalmann, N.M., Kalra, P., and Escher, M.1998. Face to virtual face. *Proceedings of the IEEE*, 86(5):870–883.

Ullman, S. and Basri, R. 1991. Recognition by linear combinations of models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(10):992–1006.

WWW page:http://www.ina.fr/Recherche/TV.

Zhang, L. 1998. Automatic adaptation of a face model using action units for semantic coding. *IEEE Trans. on Circuits and Systems for Video Technology*, 8(6):781–795.