The 13th IPPR Conference on
Computer Vision, Graphics and Image Processing

Knowledge-Based Tracking and Modeling of Facial Expressions by Stereo Vision Techniques

Chia-Yang Huang(黃家揚), Wen-Hsiang Tsai(蔡文祥)

Department of Computer & Information Science
National Chiao-Tung University
1001 Ta Hsueh Rd., Hsinchu, Taiwan 300, R.O.C.
Tel: 886-3-5712121 Ext. 56650
Email:whtsai@cis.nctu.edu.tw

Abstract

In this study, a stereo vision system for knowledge-based tracking and modeling of facial expressions is proposed. Image processing techniques are used to extract feature points on a human face and calculate related facial parameters, including the head pose caused by rigid motion and the displacements of feature points caused by local motion. And computer graphics techniques are used to re-animate the facial expressions. Three cameras are used in the system, and feature-point correspondences are employed to calculate the 3-D coordinates of the feature points. The velocity and acceleration of the head pose caused by rigid motion and the displacements of feature points caused by local motion are calculated from the 3-D coordinates of the feature points in the current and the previous image frames. The locations of feature points in the next frame are predicted by the use of the motion parameters, with help from the knowledge of the geometric properties of the feature points on the face. The knowledge-based prediction result is used to speed up search of the corresponding feature points in the next image frame. Finally, a muscle-based face model is used to re-animate the facial expressions. Experimental results show the feasibility and practicability of the proposed approaches.

(KEYWORD: Computer Animation, Motion Capture, Feature-Based, Local Tracking)

1. Introduction

In recent years, applications of multimedia and virtual reality are popular. On one hand, we can play virtual roles in a virtual environment generated by computer programs. On the other hand, we can observe the virtual environment of buildings through the network, like we observe real buildings in the real world. Besides, there are many other computer vision based applications such as videophony and teleconferencing. Recently, IBM has developed a robot, with the technique of computer vision, which is capable of observing the deformation of a human face and then simulating the facial expressions.

Feature points for use to describe the facial expression must be chosen properly to facilitate the extraction of the feature points. For this purpose, some artificial methods have been employed, such as make-up [8] or attaching marks [17]. In some other systems, the natural feature points on the face are adopted. For example, in the INA system [14], the corner of eyes, eyebrows, and the shape of the mouth are chosen to be the feature points.

After feature points are chosen, many methods have been proposed to analyze the parameters of the head motion and the

deformation of the facial expression. In Terzopoulos and Waters's approach [12], feature points are extracted and their correspondences are computed for estimating the motions parameters. In their approach, if the correspondence of the feature points cannot be found correctly, the parameters of the head motion and the facial expression will be inaccurate. In the method by Choi et al. [6], the parameters are directly estimated without finding the correspondence of the feature points. This method incorporates analysis and synthesis, and integrates the head motion and facial actions in the 2-D image sequences. In the method by Fischl et al. [13], the parameters of the head motion and the facial expressions cannot be directly estimated. The parameters are divided into four groups and tracked with the generate-and-test steepest-descent approach.

A larger motion of the human face may cause difficulty in estimating the parameters of the head motion and the facial expression [5]. One way out is to employ the approach of prediction of the parameters. Some methods have been proposed to predict the 3-D data of the feature points in the next frame in terms of the parameters of the head motion and the facial expression. For example, the Kalman filter was adopted to perform such prediction in [8]. In some other methods, the parameters of the head motion and the facial expression are separated to predict the head pose and the deformation of the facial expression. The predicted parameters must be refined to fit the real positions of the feature points. Finally, the Candide model [5] and the muscle-based face model [13] have also been employed to re-animate facial expressions.

2. Overview of Proposed System

2.1. System Configuration

The setup of multiple cameras is a key problem of stereo vision applications. In the proposed system, the setup is shown in Fig. 1. The three CCD cameras are used as input devices. They are mounted in the following way: Two of the three cameras are mounted nearly parallel in the same horizontal position and the distance of two cameras is approximately 40cm. The third is mounted near the central position of the first two cameras in the vertical direction. The distance between the central position of the first two cameras and the third is approximately 20cm. The way for mounting the cameras in this study is chosen by trial and error. If the distance between any two cameras is too small, the effective view scopes will be too narrow. If the distance between any two cameras is too large, the view scopes of the two cameras may not capture the entire region of the human face. In most cases, interesting feature points appear in at least two of the three views of the three cameras in the adopted mounting method. Furthermore, a Matrox Meteor/RGB frame grabber, which can grab the three images of the three cameras synchronously, is used to capture the head motion of the human face. The synchronous data are used to avoid the problem of the inaccurate correspondence of the feature points.

2.2. Procedure of System Operations

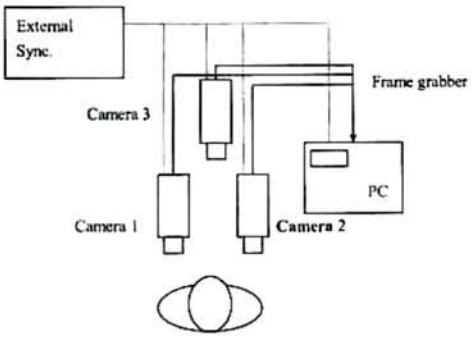


Fig. 1. The architecture of the imaging system.

The procedure of the proposed system operations will be described. Before tracking facial expressions, camera calibration should be performed in the system setup procedure to

obtain the interior and exterior parameters of the cameras. In this study, the camera parameters are estimated by using the calibration algorithm of [2]. A flowchart of the proposed procedure of extracting feature points is shown in Fig. 2. And a flowchart of the proposed system is shown in Fig. 3. More detailed system operations are described in the following.

The proposed system operations consist two major stages: the initial stage and the tracking stage. There are three major steps in the initial stage: feature extraction, feature correspondence, and calculation of the 3-D coordinates of feature points. In the initial stage, the locations of the important feature points on the human face are concerned. In the proposed system, a technique of edge detection, namely, the sobel filter is applied to locate the boundaries of the marks. Because the gaps of the detected boundaries affect the result of component labeling, it is thus desired to fill the gaps of the broken boundaries. Then, the round region inside the boundary can be obtained by a gap filling method. Each round region can be assigned an order number to represent the feature point by component labeling, and the centroids of the round regions can be calculated to represent the 2-D coordinates of the feature points. After the 2-D coordinates of the feature points are obtained, the next step is to find out the relation of the view-to-view point correspondence. In computer vision, it is a common sense that a point in a image, when projected onto another image to be, becomes a straight line. Each point on this straight line is possibly corresponding to this point. This ambiguity problem can be solved by the use of an epipolar line and a method of computer vision. The goal of feature correspondence is to find out the corresponding feature points among the three views. The next step in the initial stage is to calculate the 3-D coordinates of the feature points using the 2-D coordinates and the correspondence results of the feature points in any two views.

After all steps of the initial stage are completed, the proposed system enters the tracking stage. There are four steps in the tracking stage: motion parameter estimation, prediction, detection, and calculation of 3-D

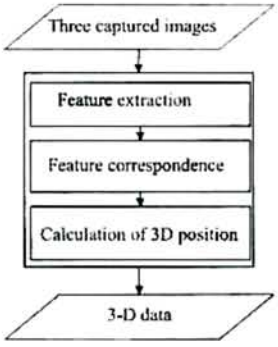


Fig. 2. The flowchart of the proposed procedure of extracting feature points.

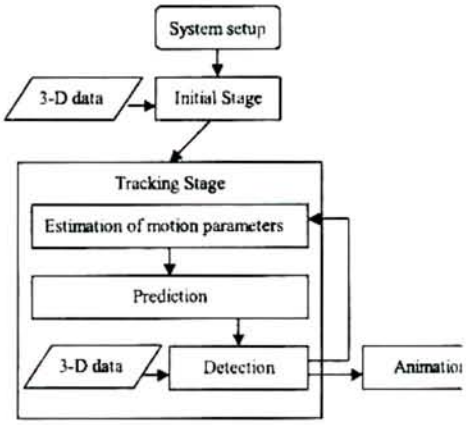


Fig. 3. The flowchart of system operations.

coordinates. Before the next frame of facial expressions is tracked, two motion parameters, namely, the velocity and the acceleration, of the head motion and the facial expressions are estimated in the parameter estimation step. The estimation is based in the use of the recorded information of the three previous frames. In the prediction step, the estimated parameters are used to predict the 3-D coordinates of the feature points in the next frame. The predicted positions may be not correct. In the detection step, the specified area of the predicted position is checked to see whether a real feature point exists

or not. Erroneous prediction may result in loss of feature points. To solve this situation, some knowledge-based methods are proposed in this study to recover the lost feature points. However, if there are too many lost feature points, the recovery may not be complete, and in this case the information of this frame will be discarded. Once all the 3-D positions of the feature points are obtained, a muscle-based face model is used to display the facial expressions in the animation step. The steps of the tracking stage are performed iteratively to track the facial expressions and display it continuously.

3. Acquisition of Feature Point Data

3.1. Edge Detection and Gap Filling

In the proposed approach to extracting feature points, standard 3x3 sobel filtering is firstly performed to detect the boundaries of the marks. Then, the edge image is bi-level thresholded with a pre-selected threshold value. To obtain a better result of edge detection, the threshold value should be well selected by experiments. After the procedure of edge detection, some detected boundaries may be broken, and thus a gap-filling procedure is performed to obtain close-up boundaries.

3.2. Component Labeling and Calculation of Central Position

After the regions of the marks are extracted from the sensed images, a component labeling procedure is employed to assign an order number to each of the regions. Meanwhile, the count of the pixels inside each region is calculated. If the pixel count of a region is too large or too small, this region is ignored.

After the order numbers of the round regions are assigned, the centroids of the regions can be calculated. If the gray-level value of the centroid of a region is smaller than a pre-selected threshold value, this round region is ignored. Because some feature points may be ignored in the procedures mentioned above, the order numbers must be sorted to facilitate the subsequent processes of the proposed system.

3.3. Correspondence of Feature Points

Because a feature point in one image might be projected onto another image ambiguously, to find the correspondence between two images, the ambiguity problem must be solved. In our approach, the technique using the epipolar line in computer vision is employed to solve the correspondence of feature points.

As mentioned previously, the centroids of the feature points can be obtained, but no information about the depth of each centroid is available. In the proposed system, the range of the depth is assumed within the working area. If this assumption stands, the range of the epipolar line can be obtained. The symbols used in the description of the procedure for estimating the correspondence are enumerated as follows:

S_A, S_B and S_C: The set of the feature points in images A, B and C, respectively.

$E_A^B(S)$: The set of the feature points in image B passed by the epipolar line associated with the feature point set S in image A.

Now we describe the feature point correspondence procedure in the following. First, with the feature points in the images A, B and C obtained by the procedures mentioned above, we project a feature point x of S_A onto images B and C to obtain the sets $E_A^B(\{x\})$ and $E_A^C(\{x\})$. Next, we project the set $E_A^B(\{x\})$ onto image C to obtain the set $E_B^C(\{E_A^B(\{x\})\})$. We then compute the union of the sets $E_A^C(\{x\})$ and $E_B^C(\{E_A^B(\{x\})\})$. If the result is non-empty and unique (i. e., the union set includes only one point), then the correspondence of the feature point x in image A to those feature points in images B and C are said to be established. Accordingly, the 3D coordinates of the point x can be computed using the correspondence.

In some cases, it might happen that certain feature points only can be observed in two of the three camera views, and so the above procedure cannot be applied. However, after we complete the above procedure, it is still possible that unique correspondence of a feature point may be established between two camera views. For such cases, the unique correspondence still can be utilized to compute the 3D coordinates of the feature point, according to the stereo computer vision principle.

Finally, if a feature point is observed only in a camera view, then correspondence of the feature point is hopeless and its 3D data are discarded.

3.4. Calculation of 3D Data of Feature Points

Before the 3-D coordinates of the feature points are calculated, the positions of the extracted feature points must be estimated by the interior parameters of the cameras computed by the calibration algorithm in [2]. Then, the estimated positions of the feature points are used to calculate the 3-D coordinates according to the estimated feature-point correspondence. Let rotation matrix $\mathbf{R}_A$ and translation matrix $\mathbf{t}_A$ be the exterior parameters of camera A, and rotation matrix $\mathbf{R}_B$ and translation matrix $\mathbf{t}_B$ be the exterior parameters of camera B. f_A and f_B are defined to be the focal lengths of the cameras A and B, respectively. And $[u_A\ v_A]$ and $[u_B\ v_B]$ are the 2-D coordinates of the feature points in the views A and B, respectively. Define [illegible] and [illegible]. Then, the 3-D coordinates of a feature point in camera A can be obtained by Eq. (1) below:

$$\begin{bmatrix} z_A \end{bmatrix}_{CO} \tag{1}$$

where

$$\begin{bmatrix} 0 & d_2 & -d_1 \end{bmatrix},$$
$$\begin{bmatrix} d_2 t_1 - d_1 t_2 \end{bmatrix}.$$

All 3-D coordinates of the feature points in camera A can be calculated from the coordinates of the centroids of the feature points and the correspondence. However, some corresponding feature points might not appear in the view of camera A. In such cases, the 3-D coordinates of these feature points can only be calculated with the information from the other two cameras. The 3-D coordinates of the feature points in the other cameras can be transformed into the coordinate system in camera A by coordinate system transformation. With the procedures mentioned above, the 3-D coordinates of the feature points can be calculated with respect to all camera coordinate systems.

4. Knowledge-Based Tracking of Facial Expressions

4.1. Estimation of Motion Parameters

After extracting the feature points in the initial stage, the deformation of the feature points is tracked by the technique of prediction. The facial expression and the head motion result in the changes of the positions of the feature points. In the proposed system, the motion parameters of the head and the facial expression are estimated with the recorded information of the three previous frames. The motion parameters can be employed to predict the positions of the feature points in the next frame and to estimate the correspondence between the new positions and the original positions of the feature points. The motion parameters are firstly estimated by dividing the parameters into two parts: the rigid motion and the local motion, of the facial expressions and analyzing these parameters. The parameters, including the velocity v and the acceleration a, are estimated by the rotation matrix and the translation matrix of the head pose. The motion parameters of the rigid motion including the rotation angels and the translation displacements in the face coordinate system, and the parameters of the local motion of the facial expressions, including the velocity v and the acceleration a, are

estimated by the positions of each feature point in the three previous frames.

4.2. Prediction of Feature Points

The position of the head can be predicted with the motion parameters, including the velocity v and the acceleration a, and the predicted position can be calculated as follows:

$$S = S_0 + vt + (1/2)at^2 \quad (2)$$

where S is the predicted position, S_0 is the original position of the head, v is the velocity, a is the acceleration, and t is the cycle time of tracking.

The velocity v and the acceleration a of the displacement, caused by the local motion, of each feature point can be calculated. Therefore, the predicted displacement of each feature point can be calculated by Eq. (2). Finally, the predicted position of each feature point can be calculated by adding the predicted position of the head pose to the predicted displacement of the local motion of each feature point.

Besides the prediction method mentioned above, there are two other methods to predict the rigid motion of the head pose. First, it is desired to find at least three corresponding feature points between two frames. For this reason, to estimate the position of the head, it is assumed that there always exist three feature points which keep the rigid motions between two frames. But there are many combinations of three correspondent feature points in 3-D space between two frames, and smaller triangles of three corresponding feature points are not appropriate for use to estimate the rigid motion because they will cause inaccuracy in the result. Instead, the largest congruence triangle of three corresponding feature points in 3-D space between two frames is searched in this study. Then, the rotation matrix and the translation matrix in the face coordinate system is calculated using the data of the corresponding feature points of the largest congruence triangle. The estimated position of the head is computed as follows:

Let P and Q be two sets of the feature points with Q being the result of P after a rigid motion. The relation of P and Q is shown as follows:

$$\mathbf{q}_n = \mathbf{R}\mathbf{p}_n + \mathbf{t}, n = 1,\ldots,N$$

where R is the rotation matrix and t is the translation vector, $\mathbf{m}_p$ and $\mathbf{m}_q$ are the central positions of P and Q, respectively. R and t can be computed as follows [1]:

$$\mathbf{t} = \mathbf{m}_q - \mathbf{R}\mathbf{m}_p$$

where $\mathbf{q} = (q_0, q_1, q_2, q_3)$ is the unit eigenvector with the largest eigenvalue in $\mathbf{K}$. And $\mathbf{K}$ is described as follows:

$$\begin{bmatrix} k_{11} & k_{12} & k_{13} \\ k_{21} & k_{22} & k_{23} \\ k_{31} & k_{32} & k_{33} \end{bmatrix} = \sum_{n=1,\ldots,N} (\mathbf{p}_n - \mathbf{m}_p)(\mathbf{q}_n - \mathbf{m}_q)^t$$

where

$$\mathbf{K} = \begin{bmatrix} k_{11}+k_{22}+k_{33} & k_{32}-k_{23} & k_{13}-k_{31} & k_{21}-k_{12} \\ k_{32}-k_{23} & k_{11}-k_{22}-k_{33} & K_{12}+k_{21} & k_{31}+k_{13} \\ k_{13}-k_{31} & k_{12}+k_{21} & -k_{11}+k_{22}-k_{33} & k_{23}+k_{32} \\ k_{21}-k_{12} & k_{31}+k_{13} & k_{23}+k_{32} & -k_{11}-k_{22}+k_{33} \end{bmatrix}$$

Second, the prediction of the rigid motion of the head pose with statistic rigid motion estimation is described as follows. The covariance matrices are calculated with the 3-D coordinates of the initial frame P and the current frame Q firstly:

$$\mathbf{M}_p = \sum_{n=1,\ldots,N} (\mathbf{p}_n - \mathbf{m}_p)(\mathbf{p}_n - \mathbf{m}_p)^t$$

$$\mathbf{M}_q = \sum_{n=1,\ldots,M} (\mathbf{q}_n - \mathbf{m}_q)(\mathbf{q}_n - \mathbf{m}_q)^t$$

Then, the normalized eigenvectors $\mathbf{U}_X$ and $\mathbf{U}_Y$ are calculated with the and by the Jacobi method. And the approximation rotation matrix R and the translation matrix t of the head in the face coordinate system are calculated by the normalized eigenvectors $\mathbf{U}_X$ and $\mathbf{U}_Y$ as follows:

$$\mathbf{R} = \mathbf{U}_Y \mathbf{U}_X^t,$$
$$\mathbf{t} = \mathbf{M}_q - \mathbf{R}\mathbf{M}_p.$$

4.3. Knowledge-Based Detection of Predicted Feature Points

The deformation of the facial expressions is not regular. For example, the feature points corresponding to the eyewink might not be detected when the eyes open or close quickly. If the situation of losing some feature points occurs, knowledge-based detection will be employed to recover the unpredicted feature points.

For example, as shown in Fig. 4, all extracted feature points are firstly projected onto a 2-D image plane in the face coordinate system. Because the feature points corresponding to the canthus change in comparison with the original positions, their positions can be estimated easily. The unpredicted feature points are in the area corresponding to the right-eye. Among these unpredicted ones, two feature points corresponding to the canthus are firstly estimated. Once the positions of the feature points corresponding to the canthus are estimated, the positions of the other two feature points can be estimated with the geometric property of the right-eye. The feature points corresponding to the eye appear in the form of a convex quadrangle. When unmatched feature points are found near the right eye, two feature points corresponding to the canthus and any other two candidate feature points can form a smallest-perimeter convex quadrangle. However, there are many permutation combinations of these feature points, and a search procedure is performed to find the smallest-perimeter convex quadrangle among these combinations of feature points. If the smallest-perimeter convex quadrangle can be found, the positions of feature points corresponding to the right eye can be estimated. In Fig. 4, the canthus corresponding to the right eye is shown as solid cycles, and the candidate feature points of the right eye are shown as hollow circles. The convex quadrangle

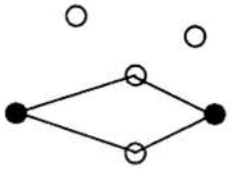


Fig. 4. The feature points corresponding to the right-eye are estimated by the geometric property.

in Fig. 4 is the feature points corresponding to the right eye by the geometric relation. The feature points corresponding to the left eye can be estimated by the same procedure using the geometric relation, too.

If the smallest-perimeter convex quadrangle is found, establish the correspondence of the feature points.

The geometric property of the feature points can also be employed to search the unpredicted feature points corresponding to the mouth. An example for recovering the unpredicted feature points corresponding to the mouth is shown in Fig. 5. Given n feature points found by the predicted positions and other 6-n candidate feature points failed to be found in the prediction stage, there are many permutations of these six feature points. To estimate the correspondence by comparing with the original positions, the six feature points corresponding to the mouth are projected onto the image plane and form a convex hexagon. Then the convex hexagon with the largest perimeter is found by searching all permutations. Furthermore, the angle formed by a feature point and its two adjacent feature points, like $\angle ABC$ as shown in Fig. 5, is used to check the validity of the feature points. It is observed that, besides the angles corresponding to the corners of the mouth, the angles formed by the feature points and two adjacent feature points, including $\angle ABC$, $\angle CDE$, $\angle DEF$ and $\angle FAB$, are always larger then 90° .

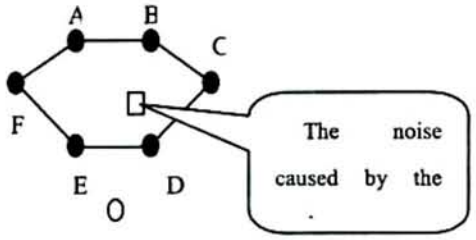


Fig. 5. The geometric relation for estimating the feature points of the mouth. The feature points of ABCDEF must be a convex hexagon. The angles of $\angle ABC$, $\angle CDE$, $\angle DEF$ and $\angle FAB$ must be larger than 90° .

4.4. Proposed Procedure for Facial Expression Tracking

As mentioned previously, the 3-D coordinates of the feature points can be predicted with the motion parameters. However, the real positions of the feature points are required to test the quality of the prediction. Firstly, the predicted positions of the feature points in the next frame are calculated, and then the real positions of the detected feature points are compared with the predicted ones to establish the correspondence. The "match count" is a variable that is used to measure the quality of the prediction. It is initially set to be zero. And if a detected feature point is found within the neighborhood of the predicted position, the value of the "match count" will be increased by one. Three situations, For the proposed procedure for facial expression tracking can be identified. A flowchart of the proposed procedure for facial expression tracking is shown in Fig. 6. The tracking procedure is as follows.

Case 1:

Step 1. Predict the positions of feature points in the next frame by adding the head pose to the displacement of each feature point.

Step 2. Compare the predicted positions with the real positions of the feature points.If the match count is larger than a pre-selected threshold value, then go to Step 3. If not, then regard the predicted positions to fail to detect the positions of the feature points, and go to Case 2.

Step 3. If all feature points are detected, then go to Step 5. If not, go to Step 4.

Step 4. Recover the unpredicted positions of feature points corresponding to the eyes and the mouth. If all unpredicted positions of feature points are recovered, then go to Step 5. If not, ignore the information.

Step 5. Calculate the head pose with the correspondence of the detected feature points and the reference ones.

Case 2:

Step 1. Predict the positions of the feature points in the next frame by calculating the head pose with the largest-congruence-triangle rigid motion estimation method.

Step 2. Compare the predicted positions with the real positions of feature points. If the match count is larger than a pre-selected thresholding value, then go to Step 3. If not, regard the predicted positions to fail to detect the positions of the feature points, and go to Case 3.

Step 3. If all feature points are detected, apply the predicted positions to detect the real positions of the feature points. If not, go to

Step 4.

Step 4. Recover the unpredicted positions of the feature points corresponding to the eyes and the mouth. If all unpredicted positions of the feature points are recovered, then finish the procedure. If not, ignore the information.

Case 3:

Step 1. Predict the positions of the feature points in the next frame by calculating the head pose with the statistic rigid motion estimation method.

Step 2. Compare the predicted positions with the real positions of the feature points. If the match count is larger than a pre-selected threshold value, then go to Step 3. If not, regard the predicted positions to fail to detect the positions of the feature points, and ignore the information.

Step 3. If all feature points are detected, apply the predicted positions to detect the real positions of the feature points. If not, go to Step 4

Step 4. Recover the unpredicted positions of the feature points corresponding to the eyes and the mouth. If all unpredicted positions of the feature points are recovered, then finish the procedure. If not, ignore the information.

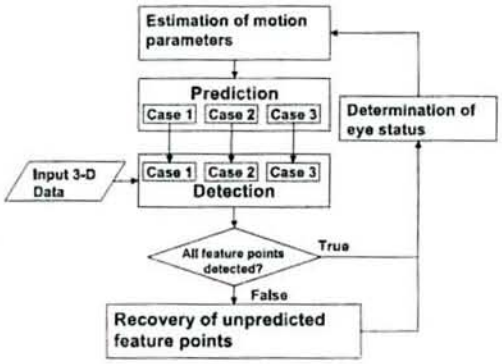


Fig. 6. The flowchart of the proposed procedure for facial expression tracking.

5. Animation of Facial Expressions

5.1. Review of An Adopted Model for Facial Muscles

In this study, the muscle-based face model described in Parke and Waters [15] is adopted. In this model, facial expressions are described based on the properties of the facial muscle and skin, but the amount of mouth opening which is necessary to analyze facial expressions is not modeled. For analyzing facial expressions completely, the amount of mouth opening is modeled as another parameter of the muscle contraction [16].

In this model, assume that there are m muscle fibers, and x_i represents the position of the ith feature point before muscle activities in the face coordinate system. Let x_i' be the position of the ith feature point plus the influence of the m muscle activities acting on the ith feature point. The relation between x_i and x_i' can be described as follows:

$$x_i' = x_i + \sum_{j=1} c_j v_{ij} m_{ij} \tag{3}$$

where b_{ij} is a muscle blend function that specifies the influence of the jth muscle fiber on the ith feature point, and c_j and m_j are the contraction factor and muscle vector for the jth muscle fiber, respectively. The function b_{ij} is defined as follows:

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

For simplicity, let $v_{ij} = b_{ij} m_j$. Eq. (3) becomes

$$x_i' = x_i + \sum_{j=1} v_{ij} c_j \tag{4}$$

Arranging Eq. (4) for every feature point into a matrix form, we can obtain:

$$\mathbf{g} = \mathbf{Mc} \tag{5}$$

where

$$\begin{bmatrix} x_1' - x_1 \\ \vdots \\ x_{m+1}' - x_{m+1} \end{bmatrix} \quad \begin{bmatrix} V_{(m+1)1} & V_{(m+1)2} & \cdots & V_{(m+1)(m+1)} \end{bmatrix} \quad \begin{bmatrix} c_1 \\ \vdots \\ c_{m+1} \end{bmatrix}$$

By considering some priori knowledge about the ranges of the muscle contraction values, Eq. (5) becomes a constrained optimization problem QP:

Since QP is a linearly constrained quadratic

$$\text{QP}: \min \|\mathbf{Mc} - \mathbf{g}\|_2^2$$

subject to

$$-0.25 \le c_i \le 1, i = 1, \cdots, m,$$

$$-0.01 \le c_{m+1} \le 1.$$

programming problem that can be solved by the quadratic programming, the parameter c_i of each muscle fiber can always be obtained in polynomial time [15].

5.2. Determination of Eye Status

The procedure for estimating the parameters, including the muscle fibers and the amount of mouth opening, is described. However, the parameters of the eye status are not considered. In the initial stage, order numbers are assigned to represent feature points automatically. By the user interface, the feature points of eyewinks are chosen manually. With them, the distance d_e of eye-opening between the eyewinks can be calculated. Once the positions of feature points are estimated in tracking, the distance d_e' of eyewinks can be also calculated. Then the eye status are considered as follows:

"Open": $d_e' \geq 0.8 \times d_e$,

"Semi-Open": $0.6 \times d_e \leq d_e' < 0.8 \times d_e$;

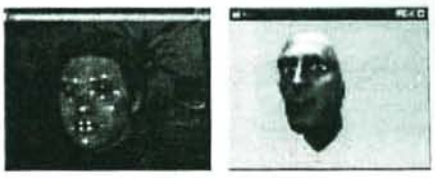
"Close": $d_e' < 0.6 \times d_e$;

6. Experimental Results

Several examples of tracking and modeling of facial expressions are shown in Fig. 7 through Fig. 15. In Fig. 7(a), an initial image is shown which is captured from one of the cameras and animation of facial expressions is shown in Fig. 7(b). A results of tracking is shown In Fig. 8. In Fig. 9, a result of estimating the rigid motion of feature points by largest-congruence-triangle rigid motion estimation is shown. In Figs. 10, 11 and 12, some results of recovery of unpredicted feature points corresponding to the mouth by knowledge-based detection are shown. Some further results of tracking are shown In Fig. 13. In Figs. 14 and 15 some results of recovery of unpredicted feature points corresponding to eyes by knowledge-based detection are shown.

(a) (b)

Fig. 7. An initial image captured from one of the cameras and animation of facial expressions. (a) Initial image. (b) Face model.

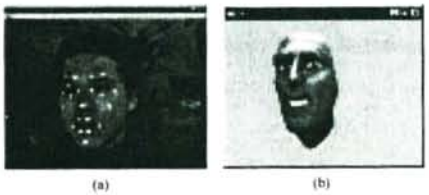


(a) (b)

Fig. 8. A result of tracking.

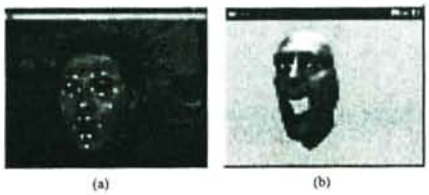
(a) (b)

Fig. 9. A result of estimating the rigid motion of feature points by largest-congruence-triangle rigid motion estimation.



(a) (b)

Fig. 10. A result of recovery of unpredicted feature points corresponding to the mouth by knowledge-based detection.



(a) (b)

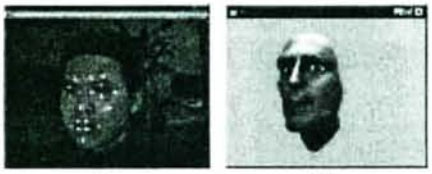
Fig. 11. A result of recovery of unpredicted feature points corresponding to the mouth by knowledge-based detection.

(a) (b)

Fig. 12. A result of recovery of unpredicted feature points corresponding to the mouth by knowledge-based detection.

7. Discussions

After observing the experimental results shown previously, some issues are discussed as follows.



(a) (b)

Fig. 13. A result of tracking.

(a) (b)

Fig. 14. A result of recovery of unpredicted feature points corresponding to eyes by knowledge-based detection.

(a) (b)

Fig. 15. A result of recovery of unpredicted feature points corresponding to eyes by knowledge-based detection.

In the initial stage of the extraction processes, the gap filling approach is useful to recover broken boundaries caused by edge detection. Most broken gaps can be filled well, and feature points can be extracted completely. The noise, which was caused by the teeth when opening the mouth, cannot be avoided in the detection processes, but the procedure of estimating the correspondence of feature points can solve this problem.

In the tracking stage, if the trajectory of the deformation of facial expressions is continuous, the feature points can be tracked well by the prediction procedure. If a large motion of the head occurs, the prediction procedure might fail to track feature points. However, some methods, including largest-congruence-triangle rigid motion estimation and statistic rigid motion estimation, are employed to estimate the position of the head approximately. Meanwhile, the facial expressions might be not predicted well, and this is a problem. This situation is often caused by eye closing and mouth opening. Thus knowledge-based detection is proposed to solve this problem, which can recover the unpredicted feature points corresponding to the eye and the mouth. The goal of solving the problem caused by unpredicted feature points is achievable.

In our proposed system, the frame rate in tracking is 1.5 frame/sec, and this result is not satisfactory for real-time applications. Ten times faster is desirable.

8. Conclusions

A proposed system for tracking and modeling of facial expressions has been successfully implemented. It is based on the use of many techniques: digital image processing, computer vision, and computer graphics, etc. They are summarized as follows.

In the aspect of digital image processing, the feature points in the captured images from the cameras were extracted completely. The boundaries of the adopted marks are detected by the standard sobel filter and a bi-level image are generated by thresholding from the edge-value image. In the process of edge detection, the boundaries may be broken and we use a technique of gap filling to fill the gaps. Therefore, round regions of feature points can be obtained and labeled by the technique of component labeling.

In the aspect of computer vision, the correspondence of feature points can be established by utilizing the relation of the epipolar line. The ambiguity problem of correspondence can be solved by this technique. In the proposed system, 3-D coordinates are calculated with the correspondence of feature points in any two of three images.

In the aspect of computer graphics, a muscle-based face model is employed to re-animate facial expressions. By the contraction of muscle fibers, the face model can be adapted to animate alive facial expressions.

In the aspect of tracking, the simple concept of physics to estimate motion parameters is used. In tracking, feature points might be unpredicted by the large motion, thus knowledge-based detection by the use of the geometric property of feature points is employed to recover this problem.

The experimental results shown in the previous chapters have revealed the feasibility of the proposed system.

9. References

[1] K. Kanatani , *Geometric Computation for Machine Vision*, Oxford Univ. Press, New York, NY, 1993.

[2] S.W Shih, Y.P Hung, and W.S Lin, "Accurate linear technique for camera calibration considering lens distortion by solving an eigenvalue problem," Optical Engineering, Vol. 32, no. 1, 1993, pp. 128-149.

[3] T. Minagawa., H. Saito, and S. Ozawa, "Face-direction estimating system using stereo vision," IECON'97: 23rd International Conference on Industrial Electronics, Control and Instrumentation, Vol. 3, 1997, pp. 1454 -1459.

[4] M. Okubo,and T. Watanab., "Lip motion capture and its application to 3-D molding," Proceedings of Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 187 -192.

[5] H. Li, P. Roivainen, and R. Forchheimer, "3-D motion estimation in model-based facial image coding," IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 15, No.6, 1993, pp. 545-555.

[6] C. S. Choi, K. Aizawa, H. Harashima, and T. Takebe, "Analysis and synthesis of facial image sequences in model-based image coding," IEEE Trans. On Circuits and Systems for Video Technology, Vol. 4, No 3, 1994, pp. 257-275.

[7] H. Tao and T. S. Huang, "Bezier volume deformation model for facial animation and video tracking," in Processing of International Workshop, CAPTECH'98: Modeling and Motion Capture Techniques for Virtual Environments, Berlin, Germany, 1998, pp. 242-253, Springer-Verlag.

[8] B. Bascle and A. Blake, "Separability of pose and expression in facial tracking and animation," in Proceedings of the Sixth International Conference on Computer Vision, 1998, pp. 323-328.

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

[10] L. Zhang, "Automatic adaptation of a face model using action units for semantic coding," IEEE Trans. On Circuits and Systems for Video Technology, Vol. 8, No. 6, 1998, pp. 781-795.

[11] H. Li and R. Forchheimer, "Two-view facial movement estimation," IEEE Trans. On Circuits and Systems for Video Technology, Vol. 4, No. 3, 1994, pp. 276-287.

[12] D. Terzopoulos and K. Waters, "Analysis and synthesis of facial image sequences using physical and anatomical models," IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 15, No. 6, 1993, pp. 569-579.

[13] J. Fischl, B.Miller and J. Robinson, "Parameter tracking in a muscle-based analysis/synthesis coding system," In Proc. Picture Coding Symp. (PCS93) 2.3, Mar 1993.

[14] WWW page: http://www.ina.fr/INA/Research/TV/

[15] F. I. Parke and K. Waters, Computer Facial Animation, A. K. Peters, Ltd., Wellesley, MA, 1996.

[16] C. C. Chang, and W. H. Tsai, "Determination of head pose and facial expression from a single perspective view by successive scaled orthographic approximations," submitted.

[17] B. Guenter, C. Grimm, D. Wood, H. Malvar and F. Pighin, "Making Faces," Technique Report.