Realistic 3D Facial Animation Parameters from Mirror-reflected Multi-view Video

I-Chen Lin, Jeng-Sheng Yeh, Ming Ouhyoung
Dept. of CSIE, National Taiwan University
Email: {ichen, jsyeh, ming}@cmlab.csie.ntu.edu.tw

Abstract

In this paper, a robust, accurate and inexpensive approach to estimate 3D facial motion from multi-view video is proposed, where two mirrors located near one's cheeks can reflect the side views of markers on one's face. Nice properties of mirrored images are utilized to simplify the proposed tracking algorithm significantly, while a Kalman filter is employed to reduce the noise and to predict the occluded markers positions. More than 50 markers on one's face are continuously tracked at 30 frames per second. The estimated 3D facial motion data has been practically applied to our facial animation system. In addition, the dataset of facial motion can also be applied to the analysis of co-articulation effects, facial expressions, and audio-visual hybrid recognition system.

1. Introduction

Pouting lips, raising eyebrows, and grinning on the face, these delicate facial expressions and lip motions are critical factors for a human being to understand or express one's meanings or feelings. Therefore, for decades, a lot of researches have been undertaken or even underway to synthesize facial animation for new communication methods such as talking heads or virtual conferencing. However, the spatio-temporal relation of facial motions are nonlinear and do not have rigid body properties; furthermore, there are a multitude of subtle expressional variations on the face and mouth. Up to the present, synthesizing realistic facial animation is still a tedious and difficult work. In addition, during speaking and pronunciation, the facial and lip motion variations can be much more complex. The motions at the transition between articulations, so called co-articulation effects [1], are also nonlinear. To animate realistic facial expression, this should be taken into account.

The goal of our project is to collect an accurate dataset of facial motion according to audio articulations, and to develop a system for realistic facial animation. We proposed a complete procedure from semi-automatic marker tracking in a video sequence, 3D position and motion estimation, to facial animation driven by estimated 3D motion trajectories. In the first step, an adaptive Kalman filter [31, 36] is utilized to improve the stability of marker tracking. Most of the jitters and "derailment", caused by intensity noise, estimation errors, interlaced effects, and even some short-term occlusions of markers, can be diminished or removed after filtering. For 3D position and motion estimation, we propose an approach that analyzes video clips with frontal and mirror-reflected images. In the results of simulation, the proposed approach can be more reliable than that of general-purpose stereovision approaches in this specific situation. In the phase of facial animation, a generic head model is deformed according to range images acquired by 3D laser scanner. Scatter data interpolation function is then applied to smoothly scatter the effects of estimated feature points to non-estimated points.

This paper is organized as following. Some representative related researches are discussed in section 2. In section 3, we introduce the application of adaptive Kalman filter to marker tracking in a video sequence. In section 4, the proposed approach of 3D facial motion estimation is described, and some comparisons with general-purpose 3D position estimation approach via R, t estimation are also discussed. A face synthesis system will be mentioned in section 5. Finally, we will conclude our paper and mention our future work.

2. Related work

Researches for synthesis of human face and animation can be approximately classified into three categories: feature point-driven, physical-based, and image-sample-based approach.

The most representative researches of physical-based approach are Waters et al's work [3, 5, 6, 34]. They use a physical or procedural model to synthesize facial motion. In an ideal case, this approach should realistically manifest the facial motion from the dynamics or kinetics evaluation. However, human faces are so subtle that many fine variations on a face cannot be simulated by an approximate model.

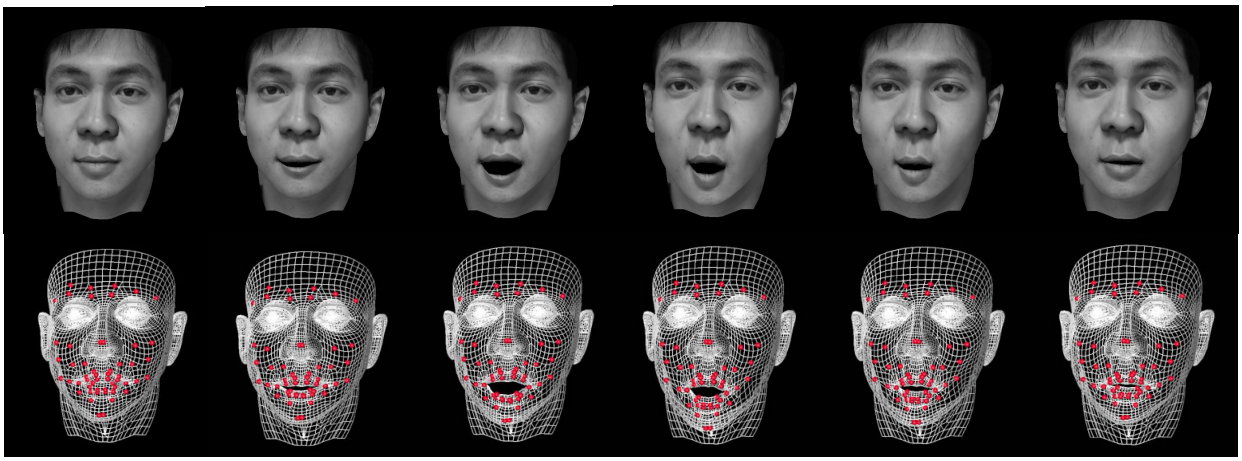


Figure 1. Motion trajectories of control points estimated by the proposed method and the synthesized head that is pronouncing the sound "au".

Recently, many researchers adopt feature-point driven approaches. Some of them produce facial animation by morphing 2D key frame images according to the feature point displacement, such as [7, 8, 9]. The 2D morphing approaches' disadvantages are that the view directions are limited and difficult to be combined with a 3D graphics environment. Other research uses 3D head models instead [10, 11]. Nevertheless, most of these kinds of research still use only 2D key frames and some hypotheses to drive a 3D model. Pighin et al. [12], Guenter et al. [13] developed remarkably lifelike realistic facial animations from 3D data. In Guenter's approach, a large numbers of markers are placed on an actor's face, and facial motions are faithfully estimated from multiple view sequence. Our work is similar to Guenter's work; moreover, we do not only focused on reproducing the facial motion of a certain performer but also collecting a dataset according to voice articulation for further analysis.

"Video Rewrite" proposed by Bregler [14] synthesizes video realistic facial animation by combining image samples of faces and mouths according to input phonemes. Cosatto et al. [15, 16] further decompose the samples into smaller facial parts and let the process of synthesis with more flexibility and efficiency. Nevertheless, the image sample-based approach suffers the same disadvantage of the 2D morphing approach, where the view direction is limited. Besides, it requires a large database of image samples for each performer.

There are some other related researches on synthetic human faces. Z. Liu et al. [38] proposed to synthesize delicate details on a face with expression ration images (ERI). Blanz et al [17] established an excellent system to build head model from only single face image by statistic human head information. Voice Puppetry [18] applied the Hidden Markov Model (HMM) to simulate facial motions driven by various audio features. Our previous work [19] is also a speech driven talking head system.

3D motion can be estimated from optical or magnetic motion tracking devices, or video sequences. Optical or magnetic marker tracking devices can provide extremely precise 3D position data, but they are also highly expensive. Moreover, because the special markers may obstruct some subtle motions, most of these tracking devices are unsuitable for motion tracking on a lip surface.

Most of the stereo video motion-tracking approaches are based on the epipolar constraint and the 8 points algorithm [20]. Images with multiple view directions are taken to estimate the 3D positions of feature points. [21, 22, 23] provide a good reference and discussion for 3D motion and structure estimation.

In addition to capturing stereo videos by multiple cameras, Patterson et al. [39] proposed to use a mirror to acquire multiple views for facial motion recording. Basu et al. [24, 25] employed mirror views to capture the lip motion. In our works, we also used mirrors to get new images with different view directions. However, unlike the related works, we proposed a more robust and simpler algorithm to estimate accurate 3D position and motion from mirrored and front view video sequences, since there are nice properties of mirrored images that can be used.

3. Tracking markers in video with adaptive Kalman filter

3.1. Marker tracking

In our face synthesis system, we separate a face into 11 regions. While regarding each region as a smoothly deformable surface, we find that there are 50 points (10 for lip contours, 12 for the lip surfaces, 10 for the mouth, 8 for cheeks, and 10 for the forehead) on a face, where the variations are the most representative to control the surface deformation. Therefore, we take these 50 positions as feature points to drive facial animation.

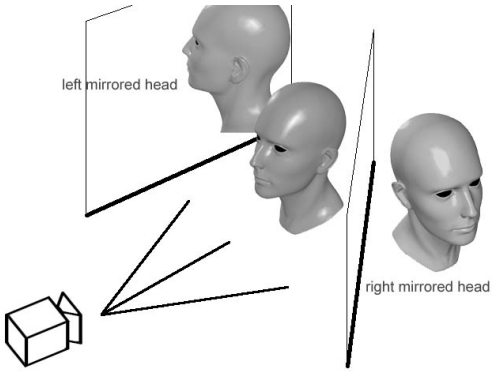


Figure 2. A diagram of our capture equipment. Two mirrors are placed next to a subject's face, and the front view and mirror-reflected images are captured simultaneously.

In order to get precise 3D positions and motions of feature points on the face of a subject, colorful dot markers are stuck onto feature points. With these markers, tracking of feature point movement is much easier and more accurate.

It is well known that multiple view images (at least two images of different view directions for a target) are required for 3D position reconstruction. In our work, we didn't use multiple cameras to capture images from different view directions. Instead, we placed two mirrors next to one subject's face (as shown in Fig.2), and used only one camera to capture the front view image and two mirrored images (as shown in Fig. 3).

Before calculating the 3D positions of markers, the locational variations of markers in each frame of a video clip and the correspondence of markers in front and mirrored images should be determined in advance. We adopted a semi-automatic approach to do this. Once a video has been prepared for tracking, users have to initially select the position of each marker and their correspondence in front and mirrored images. Our system then searches for the most probable motion trajectories of markers in the following frames.

3.2. Adaptive Kalman filter for marker tracking

The Kalman filter is a linear, unbiased, and minimum error variance recursive algorithm to optimally estimate the unknown state of a linear dynamic system from noisy data at discrete time intervals, and it is widespreadly applied to control system, radar tracking and etc. [31, 32, 37]. Here we briefly mention the concept of the Kalman filter.

Let $s(t)$ denote an M-dimensional state vector of a dynamic system at time t, and the propagation of the state in time can be expressed as a linear equation

$$s(t) = As(t\text{-}1) + w(t), \quad t = 1,2,\ldots,T_{limit},$$

Figure 3. The image data captured by DV camera (resolution: 720x480 pixels). 55 markers are placed on the subject's face and lips.

where A is a state-transition matrix and $w(t)$ is a zero-mean, random sequence with covariance matrix $Q(t)$, representing the state model error.

Suppose that a time-series of measurements $h(t)$, are available, which are linearly related to the state variable as

$$h(t) = Cs(t) + v(t), \quad t = 1,\ldots,T_{limit}.$$

where C is the observation matrix and $v(k)$ denotes a zero-means, noise sequence, with covariance matrix $R(k)$.

Given the measurement $h(t)$, the state vector can be estimated as

$$s(t) = As(t\text{-}1) + K(t)[h(t)\text{-}CAs(t\text{-}1)],$$

where the $K(t)$ is so called the Kalman gain matrix. And the $s(t+1)$ can be predicted as

$$s(t+1|\,t) = As(t).$$

In our work, we adopt an adaptive Kalman filter [36] to improve the stability of marker tracking in video. We assume the state transition equation to be

$$\begin{bmatrix} s_{px}(t) \\ s_{vx}(t) \\ s_{py}(t) \\ s_{vy}(t) \end{bmatrix} = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s_{px}(t-1) \\ s_{vx}(t-1) \\ s_{py}(t-1) \\ s_{vy}(t-1) \end{bmatrix} + \begin{bmatrix} 0 \\ w_{vx}(t-1) \\ 0 \\ w_{vy}(t-1) \end{bmatrix}, \quad (1)$$

where $s_{px}(t)$, $s_{vx}(t)$, $s_{py}(t)$, and $s_{vy}(t)$ represent the state values of position and velocity in x and y axial directions at time t respectively. And $w_{vx}(t)$, $w_{vy}(t)$ represent the change of velocity in x and y axial directions respectively over interval T with variance $\sigma_{vx}^2(t)$ and $\sigma_{vy}^2(t)$.

The relation between measurement and state vector can be written as

$$\begin{bmatrix} h_{px}(t) \\ h_{py}(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} s_{px}(t) \\ s_{vx}(t) \\ s_{py}(t) \\ s_{vy}(t) \end{bmatrix} + \begin{bmatrix} v_{px}(t) \\ v_{py}(t) \end{bmatrix}, \quad (2)$$

where $v_{px}(t)$ and $v_{py}(t)$ represent the position measurement

error in x and y axis with variance $\sigma_{px}^2(t)$ and $\sigma_{py}^2(t)$. $\sigma_{px}^2(t)$ and $\sigma_{py}^2(t)$ are variables and can be adjusted according to the confidence of measurement. The details of Kalman filter are well described in the reference book [31, 33].

The whole procedure of marker tracking is as following:

1. Users have to designate the location $h_i(0)$ of feature point i in the first frame (at $t=0$), where $h_i(0) = [h_{pxi}(0), h_{pyi}(0)]^t$, for $i = 1,2...,N$.
Set $s_{pxi}(0) = h_{pxi}(0)$, $s_{pyi}(0) = h_{pyi}(0)$, $s_{vxi}(0) = s_{vyi}(0)$, $t=0$.

2. Predict the position at time $t+1$ as
$s_i(t+1|t) = As_i(t)$, for $i = 1,...,N$.
and update the time stamp, set $t = t+1$.

3. Within the searching range centered by $(s_{pxi}(t), s_{pyi}(t))$, find the measurement position $h_i(t)$ by searching the position with minimum $Cost_i(t)$, for $i = 1,2,...,N$.
$Cost_i(t) = CostR_i(t) + CostG_i(t) + CostR_i(t)$, (3)

where $CostR_i(t)$, $CostG_i(t)$, $CostB_i(t)$, represent the correlation of color component R, G, B between $s_i(t-1)$ and a candidate position in frame t.

4. Set $\sigma_{pxi}^2(t) = \sigma_{base}^2 + \alpha Cost_i(t)$ and $\sigma_{pyi}^2(t) = \sigma_{base}^2 + \alpha Cost_i(t)$, where α is a weighted value, and σ_{base}^2 is the constant base variance.
Calculate the state vector $s_i(t)$ by Kalman filtering.

5. Record $(s_{pxi}(t), s_{pyi}(t))$ as the 2D position of marker i in time t.

6. If $t < T_{limit}$, go to step 2.

As the adjustment of $\sigma_{pxi}^2(t)$ and $\sigma_{pyi}^2(t)$ in step 4, when an image of marker is occluded or interfered by interlace effect or intense specular-lighting noise, the value of cost function should be dramatically high, and the variances of measurement error $\sigma_{pxi}^2(t)$ and $\sigma_{pyi}^2(t)$ will be large; then, the Kalman gain values will be decreased. With this design, the effects of noise or occlusion are diminished.

4. 3D facial motion estimation

As the conceptual diagram in Fig. 4, the mirrored image can be regarded as a "flipped" image taken by a "virtual camera", which is in a distinct view direction comparing to physical one. With two mirrors next to a subject's face, we can acquire three different views of the face image data simultaneously and can also avoid the problem of synchronization between data among different cameras.

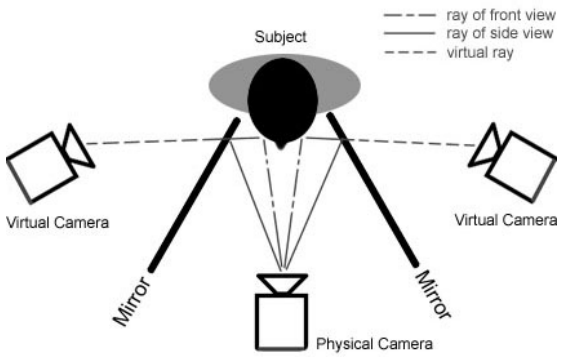


Figure 4. The conceptual diagram of "virtual camera".

In some related researches [24], the 3D positions of the aforementioned situation were estimated by modified general-purposed 3D structure reconstruction approaches, which estimate affine transformation (rotation matrix R, translation vector t) between two cameras from fundamental matrix [23]. After getting the location and orientation of two cameras, the target point 3D positions can then be approximated by the closest points to all projection rays from lens of different cameras.

However, there are some special properties of mirrored images that can be applied to get a more accurate result. We present our approach in subsection 4.1. A flexible camera calibration method proposed by Zhang et al [26] is utilized to calculate the camera intrinsic parameters. With these parameters, we can calibrate the video captured by camera.

4.1. 3D position estimation from front and mirrored images

After the motion trajectories of markers in videos of front and mirror-reflected views are acquired with the method described in section 3, 3D motion trajectories can be calculated by first calculating the orientation and location of the mirror in video, and then estimating the 3D positions of markers as a minimization problem.

In the first step, we can assume that a mirror is flat without distortion, and we only use the image data within the range of mirrors. The location and orientation of the mirror can be represented by a plane equation:

$$ax + by + cz = d \quad (4)$$

$u = (a, b, c)^t$, $\| u \| = 1$, where u is the unit normal of the plane, and there are two possible directions of vector u. Without loss of generality, we take the direction of $c < 0$. In the following discussion, we assume that I is the image

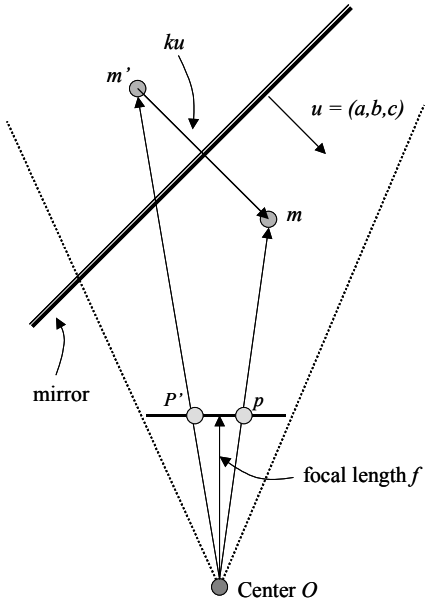


Figure 5. The geometric representation of the physical point *m*, the reflected point *m'*, and the projection points *p*, *p'*.

plane of camera film, *f* is the focal length, the camera lens center *O* is assumed as the origin in the coordinate, and the view direction of the camera is the *Z* axis.

As shown in Fig. 5, m_i is the physical 3D position of marker *i*, $m_i = (x_{mi}, y_{mi}, z_{mi})^t$, m_i' is the virtual 3D position of marker *i* in the mirrored image, $m_i' = (x_{mi}', y_{mi}', z_{mi}')^t$, p_i is the projection of m_i on *I*, $p_i = (f\frac{x_{mi}}{z_{mi}}, f\frac{y_{mi}}{z_{mi}}, f)^t = (x_{pi}, y_{pi}, z_{pi})^t$,

p_i' is the projection of m_i' on *I*, $p_i' = (f\frac{x'_{mi}}{z'_{mi}}, f\frac{y'_{mi}}{z'_{mi}}, f)^t = (x_{pi}', y_{pi}', z_{pi}')^t$. (x_{pi}, y_{pi}) and (x_{pi}', y_{pi}') are the estimated 2D marker positions as mentioned in section 3.

Owing to the property of mirrors,

$$m_i' = m_i + ku, \tag{5}$$

where *k* is a scale value. Vector m_i, m_i', *u* are co-plane, and thus

$$m_i' \cdot (u \times m_i) = 0\,, \tag{6}$$

• is dot product, and × is cross product.

From Eq. (6), we reformulate in terms of p_i, p_i',

$$\frac{z'_{mi}}{f} p_i' \cdot \left[u \times \left(\frac{z_{mi}}{f} p_i \right) \right] = 0\,, \tag{7}$$

and it can be simplified as

$$p_i' U p_i = 0, \text{ where } U = \begin{bmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{bmatrix}. \tag{8}$$

Eq. (8) can then be represented in terms of *u* as,

$$\begin{bmatrix} (y_{pi} - y'_{pi})f & (-x_{pi} + x'_{pi})f & (x_{pi}y'_{pi} - y_{pi}x'_{pi}) \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = 0\,, \tag{9}$$

For each marker and the rest stationary points for rigid body calibration, we can form a matrix *M*,

$$Mu = 0,$$

where

$$M = \begin{bmatrix} (y_{p1} - y'_{p1})f & (-x_{p1} + x'_{p1})f & (x_{p1}y'_{p1} - y_{p1}x'_{p1}) \\ (y_{p2} - y'_{p2})f & (-x_{p2} + x'_{p2})f & (x_{p2}y'_{p2} - y_{p2}x'_{p2}) \\ \vdots & \vdots & \vdots \\ (y_{pn} - y'_{pn})f & (-x_{pn} + x'_{pn})f & (x_{pn}y'_{pn} - y_{pn}x'_{pn}) \end{bmatrix} \tag{10}$$

Since there is noise to perturb the shape and position of markers on image plane *I*, the least square method is applied to estimate the vector *u* with least error. It is well-known that solution of

$$\min_u \| Mu \|, \qquad for \| u \| = 1, \tag{11}$$

is the eigenvector corresponding to the smallest eigenvalue of the matrix M^tM [27].

There is another property of mirror is that

$$(m_i' - \Theta) = H_u (m_i - \Theta)\,, \tag{12}$$

where Θ is an arbitrary point on the mirror plane *Mirror*. $H_u = (I_{3x3} - 2uu^t)$ is the *Householder matrix*, where I_{3x3} is the identity matrix. We choose that $\Theta = (0,0,\frac{d}{c})^t$, and deduce the equation

$$\begin{bmatrix} \left(\frac{2a^2-1}{2f}\right)x_{pi} + \left(\frac{ab}{f}\right)y_{pi} + ac & \frac{x'_{pi}}{2f} \\ \left(\frac{ab}{f}\right)x_{pi} + \left(\frac{2b^2-1}{2f}\right)y_{pi} + bc & \frac{y'_{pi}}{2f} \\ \left(\frac{ac}{f}\right)x_{pi} + \left(\frac{bc}{f}\right)y_{pi} + \frac{2c^2-1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} z_{mi} \\ z'_{mi} \end{bmatrix} = d \begin{bmatrix} a \\ b \\ c \end{bmatrix}, \tag{13}$$

From Eq. (13), we can find that once vector *u* has been determined, z_{mi} and z'_{mi} is proportional to variable *d*. The value *d* can be determined by comparing the scaled data with a reference ruler in real world. Thus, along the above-mentioned steps, vector *u* should be first estimated by Eq. (11); then the position of $[x_{mi}, y_{mi}, z_{mi}]^t$ for each marker and stationary points can be calculated by the least

square method of the form

$$\min_z \|Gz - du\|, \tag{14}$$

based on Singular Value Decomposition (SVD) or QR factorization [27].

Furthermore, to reduce the influence of errors of the marker position estimation in the front view image, we mirror the virtual marker m_i' back to physical world, set as m_i'',

$$m_i'' = H_u^{-1}(m_i' - \Theta) + \Theta, \tag{15}$$

and take $m_i''' = \frac{(m_i + m_i'')}{2}$ as the 3D position of marker i.

4.2. Head motion removal

In the previous step, 3D marker positions have been estimated. However, a subject under test may swing or nod his head when speaking and making facial expressions, and thus the motions of 3D markers are composed of both facial motions and global head motions. To get precise facial motion, the head motion must be estimated and removed from 3D facial expression data.

As mentioned in [22], with 3 non-colinear 3D points, the movement of rigid object can be uniquely determined by a rotation matrix R, and translation vector t.

$$r_{ij+1} = Rr_{ij} + t, \tag{16}$$

where r_{ij} is the 3D position of point i on a rigid object at time j, and where r_{ij+1} is the 3D position of point i on a rigid object at time $j+1$.

Therefore, the 3D data of 4 additional markers placed on the performer's ears are regarded as points on rigid head, and we applied the SVD (singular value decomposition) based algorithm proposed by K. Arun et al. [28] to determine the head rotation R and head translation t. After the rotation and translation of successive time stamps are determined, we can obtain the displacement of marker i caused by facial motion as $disp_i = R^{-1}(v_{i(j+1)} - t) - v_{ij}$, where v_{ij} is the estimated 3D position of marker i at time j.

4.3. Discussion of proposed 3D estimation approach

Intuitively, in the case of 3D position estimation from the mirror-reflected mutli-view images, the proposed 3D estimation approach should be much more robust than approaches that apply some other general-purposed 3D estimation approaches which calculate rotation matrix R and translation vector t of the virtual camera from the fundamental matrix [24]. One of the reasons is that the degrees of freedom of the rotation matrix R and the translation vector t are both three. In our case, we evaluate the mirror plane equation, which has only 4 degrees of freedom. The fewer degrees of freedom roughly mean that we can use much fewer information to reach the accuracy of the same magnitude.

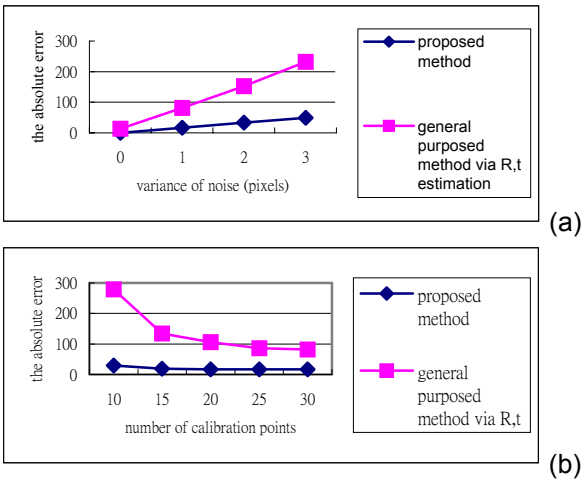


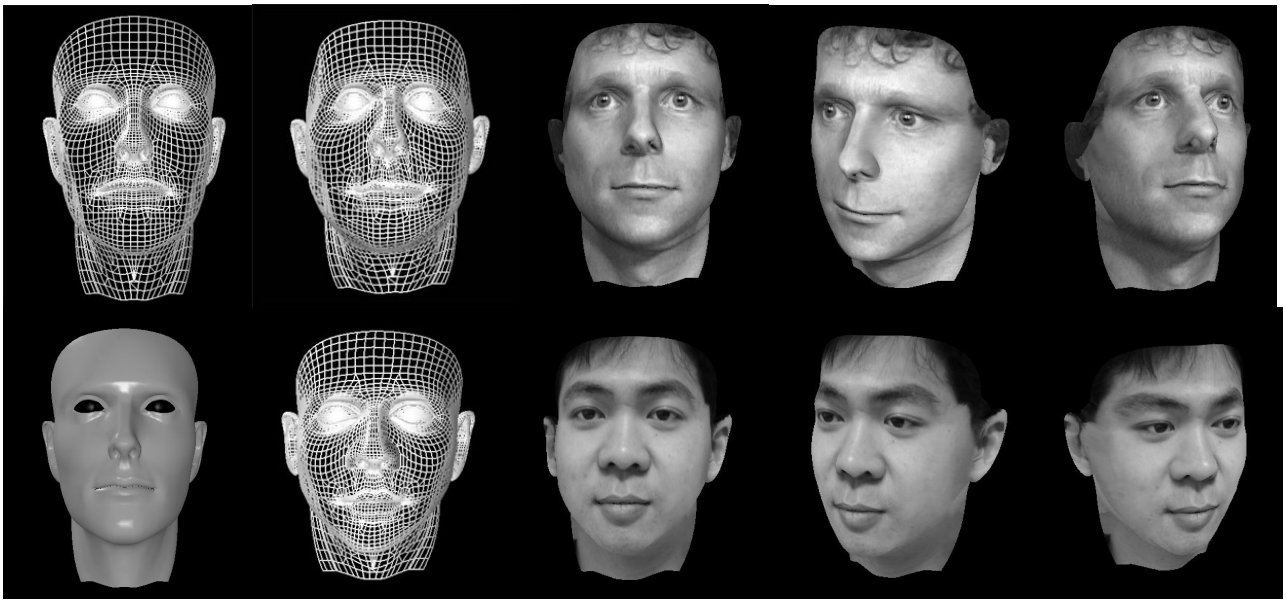
Figure 6. Error estimation of two different solutions. We simulated the situation where normal-distributed noise perturbed the estimation of marker motion in video. The target subject is a virtual object (about 1000x2000x1000 pixel^3) 4000 pixels apart from the lens center.

(a) absolute mean-square error versus variance of normal-distributed noises.(mean = 0)

(b) absolute mean-square error versus number of calibration points when noise variance = 1, mean = 0.

Secondly, when estimating R and t from the fundamental matrix [21], it first has to evaluate the fundamental matrix, which is of 8 degrees of freedom, and then analogous rotation matrix W is estimated. However, the matrix W usually may not be of the properties of rotation matrix, such as orthogonality, etc. In that situation, the matrix W is adjusted to fit the properties, and then the vector t can be evaluated. Each of the steps involves a lot of numerical matrix computations, such as the smallest eigenvalue and eigenvector estimation, singular value decomposition, and quaternion reformulation, etc. The errors are progressively accumulated by each step. [21] provides a detailed discussion of error analysis and estimation of 3D position and structure reconstruction from R, t.

We also simulated the situation where normal-distributed errors perturbed the measurement of 2D marker positions by computer. Fig. 6 is the figure about the error distribution for our proposed approach and the approach via the virtual camera R, t estimation. The figure manifests that the virtual camera approach requires more feature points or calibration points to reach the same accuracy of the proposed approach. Our proposed approach is also more robust in the noisy situation.



(a) (b) (c) (d) (e)

Figure 7. The reconstructed 3D face model and texture mapping. There are 6144 polygons and 5902 vertices on the face model. (a) the generic model. (b) the deformed model. (c)~(e) synthetic faces in different view directions.

5. Synthetic face

5.1. Face modeling

The approach mentioned in subsection 4.1 for 3D position estimation can also be applied to construct a realistic head model. However, a 3D scanner can provide 3D models of error less than 1 millimeter. Thus, we exploit a 3D scanner to get 3D head information. Nevertheless, the 3D scanned data cannot be applied for facial animation directly for three main reasons. The first one is that the topology of face model generated by 3D scanner is arbitrary and does not fit the characteristics of human face; for example, a topology on the lip should be distinct from the mouth. The second one is that there are always a lot of "holes" in 3D scanned data. The third reason is that the number of polygons generated by a 3D scanner is extremely large, and that is too many for near real-time animation. For these reasons, a generic face model with a suitable polygon topology is employed and deformed to fit the 3D scanned range data.

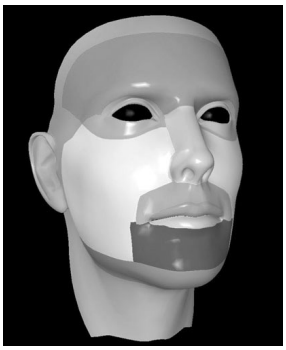


Figure 8. The 11 regions of head model: jaw, lower mouth, lower lip, upper lip, upper mouth, left cheek, right cheek, nose, left eye, right eye, and forehead

Fig 7(a) is the figure of the generic model, and fig 7(b) is the deformed model. In our current work, to fit one new generated 3D scanned range data, users have to manually specify the corresponding features such as the mouth corners, nose tip, eye corners etc. in the scanned face data. The deformation method we applied is the so-called "scatter data interpolation", which is a smooth interpolation function that can scatter the effects of feature points to non-recorded points. Supposed that p_i is the 3D position of feature point i, p_{oi} is the corresponding point on the generic model, and $u_i = p_i - p_{oi}$ is the displacement. We should construct a function that finds the unknown displacement u_j of unconstrained vertex j from u_i.

In our case, a method based on radial basis functions is adopted to represent the influence of constrained points. We chose $\phi(r) = e^{-r/64}$. The scatter data function is then of the form

$$f(p) = \sum_i c_i \phi(\|p - p_i\|) + Mp + t \qquad (17)$$

where p_i is the constrained vertex; low-order polynomial terms M,t are added as affine basis. Many kinds of function for $\phi(r)$ have been proposed [29].

To determine the unknown coefficients c_i and the affine components M and t, we must solve a set of linear equations that includes $u_i = f(p_i)$, the constraints $\sum_i c_i = 0$

Figure 9. Subtle facial expression of the synthetic face twisting his mouth.

and $\sum_i c_i p_i^t = 0$. In general, if there are n feature point correspondences, we will have n+4 unknowns and n+4 equations with the following form:

$$\begin{bmatrix} \cdot & \cdot & \cdots & p_{1x} & p_{1y} & p_{1z} & 1 \\ \cdot & e^{-\|p_i - p_j\|/64} & \cdots & p_{2x} & p_{2y} & p_{2z} & 1 \\ \vdots & \vdots & \cdot & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \cdots \ \vdots & p_{nx} & p_{ny} & p_{nz} & 1 \\ 1 & 1 & \cdots \ 1 & 0 & 0 & 0 & 0 \\ p_{1x} & p_{2x} & \cdots \ p_{nx} & 0 & 0 & 0 & 0 \\ p_{1y} & p_{2y} & \cdots \ p_{ny} & 0 & 0 & 0 & 0 \\ p_{1z} & p_{2z} & \cdots \ p_{nz} & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \\ a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad (18)$$

where $1 \le i, j \le 3 \quad P_i = (p_{ix}, p_{iy}, p_{iz})$.

5.2. Facial Animation

A general face is separated into 11 regions: jaw, lower mouth, lower lip, upper lip, upper mouth, left cheek, right cheek, nose, left eye, right eye, and forehead (as shown in Fig. 8). Control points within a region can only affect vertices in that region, and interpolation is applied to smooth the jitter effect at the boundary of two regions.

These control points consist of feature points, "fixed points" and "hypothetical points". As mentioned in subsection 3.1, feature points are the positions where markers are placed. "Fixed points" are the points where the position is always stationary no matter what the facial motion, such as the points near ears and points near the bottom of the neck etc. "Hypothetical points" are the points which are hard to capture well by view point of the video; for example the points of jaw near the ear, etc. We use a hypothesis to derive the hypothetical points according to related feature points. Eyelids and some of the points on the jaw are hypothetical points. The blink of eyelid is approximately once per 2.5 seconds as a random process. During blinking, the vertices on the eyelid move downward along the model of the eyeballs. The action of the jaw is given as the following pseudo code:

If (current jaw tip higher than the position in neutral face)
{
  Teeth should be clamp together.
  Vertices of jaw, except the neighbor area near the jaw tip, are at neutral position.
} else if (current jaw tip is lower than the one in neutral face)
{
  Jaw, which is now a rigid object, rotates and stretches around the hypothesis axis near the ears.
}

After determining the displacement of all control points, a face can be deformed by the radial basis scatter data interpolation function mentioned in subsection 5.1. Once we repeat the above similar process frame by frame, we can generate realistic facial animation according to estimated 3D facial motion data.

6. Experiment

The collection of dataset for facial and lip motions according to articulation is still under way. Three languages, English, French, and Mandarin Chinese, are adopted to be included in our dataset. At this moment, data of 6 French subjects (3 males, 3 females), and 2 Taiwanese subjects (2 males) have been recorded. For records of French, the videotaping is focused on the mouth. Each French subject performed 20 French visemes, 14 consonant-vowel articulations, 10 vowel-vowel articulations, and read a paragraph about 2 minutes long. The speech group of Loria, France suggests the decision of visemes and articulations. For Taiwanese subjects, all markers described in subsection 3.1 are applied. They did 14 MPEG4 visemes [30], 40 consonant-vowel articulations, and 10 vowel-vowel articulations.

In addition, we also developed an experiment to acquire the accuracy of the proposed 3D estimation approach. A plastic dummy head was attached with markers mentioned in section 3.1, and the diameter of a marker is about 3 mm. A 3D laser scanner is applied to measure the position of each marker; then, the 3D positions were also estimated by the proposed method. Since the measurement error bound of a 3D scanner is less than 0.1 mm, we assumed that the

data acquired by the 3D scanner are exact. Comparing with the 3D scanned data, the root mean square error of positions estimated by the proposed method is 1.95 mm, and the maximal error of 2.94 mm occurs at a marker position beneath the lower lip.

7. Result and conclusion

In this paper, we have presented a realistic facial animation system and proposed a procedure to estimate 3D facial motion trajectory from front view and mirror-reflected video clips. We have discussed the benefits of the proposed procedure to estimate 3D position and motion, and compared the approach with general-purpose 3D position estimation method via R, t evaluation. Our facial animation system can synthesize realistic facial expression with a frame rate of more than 30 frames per second on a Pentium-4 1.5GHz PC with a Nvidia Geforce 2 GTS Ultra OpenGL acceleration card.

The collection of facial motion dataset is still in progress. We hope that this data will be published soon through the web and applied for further research for the analysis and synthesis of the human face.

Acknowledgement

We would like to appreciate Nathalie. P. Valles, Yves Lapire, Dominique Fohr, Michel Pitermann and other researchers of speech group of Loria, France. They help our experiment in French and provide knowledge of French visemes. Especially, we would like to thank Michel Pitermann, since we use his face as one of the face model exhibited in this paper and he also gave us a lot of valuable suggestions for the paper. Besides, we would also like to thank Professor Mary Flanagan of University of Oregon. She helped us revise the language style.

References

[1]. M.M. Cohen and D.W. Massaro. “Modeling co-articulation in synthetic visual speech.” N.M. Thalmann and D. Thalmann, editors, Models and Techniques in Computer Animation. Springer-Verlag, 1993.

[2]. S. Dupont and J. Luettin. “Audio-Visual Speech Modeling for Continuous Speech Recognition”, IEEE Trans. Multimedia, vol. 2, No.3, pp.141-149, 2000.

[3]. K. Waters “A Muscle Model for Animating Three-Dimensional Facial Expresson”, ACM SIGGRAPH’87, vol.21, pp.17-24, July, 1987.

[4]. K. Waters and T. Levergood. “An Automatic Lip-Synchronization Algorithm for Synthetic Faces.” Proceeding of ACM Multimedia, pp. 149-156, San Francisco, CA, USA, 1994, ACM Press.

[5]. F. I. Parke, K. Waters, *Computer Facial Animation*, A K Peters, Wellesley, Massachusetts, 1996.

[6]. Y. Lee, D. Terzopoulos, and K. Waters, “Realistic Modeling for Facial Animation”, SIGGRAPH conference proceedings, pp. 55-62, ACM SIGGRAPH, August 1995.

[7]. T. Beier, and S. Neely, “Feature-based Image Metamorphosis”, SIGGRAPH 92 Conference Proceedings, pp. 35-42. ACM SIGGRAPH, July 1992.

[8]. T. Ezzat and T. Poggio, “Miketalk: A talking facial display based on morphing visemes”, Proc. Computer Animation, pp. 96-102, June 1998.

[9]. S.M. Seitz, C.R. Dyer. “View Morphing”, Proc. SIGGRAPH 96, pp. 21-30.

[10]. J. Ostermann, “Animation of Synthetic Faces in MPEG-4”, Proc. of Computer Animation, pp.49-51, Philadelphia, Pennsylvania, USA, June 8-10, 1998.

[11]. W. Lee, N.M. Thalmann. “Head Modeling from Picutes and Morphing in 3D with Image Metamorphosis Based on Triangulation”, Proc. CAPTECH’98, Geneva, pp.354-267, 1998.

[12]. F. Pighin, J. Hecker, D. Lischinski, P. Szeliski, D.H. Salesin. “Synthesizeng Realistic Facial Expressions from Photographs”, Proceedings of ACM Computer Graphics (SIGGRAPH 98), pp. 75-84 Aug-1998.

[13]. B. Guenter, c. Grimm, D. Wood, H. Malvar, F. Pighin. “Making Face”, Proc. of Computer Graphics (SIGGRAPH ’98), pp. 55-66, Aug. 1998.

[14]. C. Bregler, M.Covell, M.Slaney. “Video Rewrite: Driving Visual Speech with Audio”, Proc. SIGGRAPH’97, pp.353-360, 1997.

[15]. E. Cosatto, H.P. Graf. “Sample-Based Synthesis of Photo-Realistic Talking Heads”, Proc. of Computer Animation 98, pp. 103-110, Philadelphia, Pennsylvania, June 8-10, 1998.

[16]. E. Cosatto and H. P. Graf, “Photo-Realistic Talking-Heads from Image Samples”, IEEE Trans. Multimedia, vol. 2, no. 3, pp. 152-162, 2000.

[17]. V. Blanz and T. Vetter, “A morphable model for the synthesis of 3D faces, “, Proc. SIGGRAPH’99, pp. 353-360, July 1999.

[18]. Matthew Brand. “Voice Puppetry”, Proc. SIGGRAPH’99, pp.21-28, 1999.

[19]. T.J. Yang, I.C. Lin, C.S. Hung, C.F. Huang and M. Ouhyoung. “Speech Driven Facial Animation”, pp. 99-108, Proc. of Eurographics workshop on Computer Animation and Simulation’99 (CAS’99), Milan, Italy, Sept. 1999.

[20]. H.C. Longuet-Higgins. “A computer algorithm for reconstructing a scene from two projections”, Nature, 293:133-135, Sept. 1981.

[21]. J. Weng, T. S. Huang, N. Ahuja, “Motion and Structure from Image Sequences”, Springer-Verlag, 1993.

[22]. T.S. Huang, and A.N. Netravali. “Motion and Structure from Feature Correspondences: A Review”, Proceedings of the IEEE, 82(2), pp. 252-268, Feb. 1994.

[23]. R. I. Hartley, “In Defence of the 8-point Algorithm”, Proc. of IEEE pp. 1064-1069, 1995.

[24]. S. Basu and A. Pentland, “A Three-Dimensional Model of Human Lip Motions Trained from Video”, Proc. of IEEE Non-Rigid and Articulated Motion Workshop at CVPR’97, San Juan, June 16, 1997.

[25]. S. Basu, N. Oliver, and A. Pentland, “3D Modeling and Tracking of Human Lip Motions”, Proc. of International Conference on Computer Vision (ICCV’98), Bombay, India,

Jan. 1998.
[26]. Z. Zhang, “A Flexible New Technique for Camera Calibration”, Technical Report Microsoft MSR-TR-98-71, 1998.
[27]. G. Golub, and C. F. Van Loan, *Matrix Computation third edition*, The John Hopkins Univ. Press, Baltimore and London, 1996.
[28]. K. S. Arun, T. S. Huang, and S. D. Blostein, “Least Square Fitting of Two 3D Point Sets”, IEEE Trans. Pattern analysis and machine intelligence, vol. 9, no. 5, pp. 698-700, sept. 1987.
[29]. G.M. Nielson. “Scattered data modeling”, in IEEE Computer Graphics and Applications, 13(1), pp. 60-70, Jan. 1993.
[30]. MPEG4 Video “Text for ISO/IEC FCD 14496-2 video”, ISO/IEC JTC1/SC29/WG11 N3056, Dec. 1999.
[31]. S.M.Bozic. *Digital and Kalman Filter*, Edward Arnold Ltd, London, 1979.
[32]. Z.Zhang, and O.Faugeras. *3D Dynamic Scene Analysis*, Springer-Verlag, Berlin and Heidelberg, 1992.
[33]. A.M. Tekalp, *Digital Video Processing*, Prentice Hall PTR, 1995.
[34]. J.C. Lucero and K.G.Munhall, “A Model of Facial Biomechanics for Speech Production”, the Journal of Acoustical Society of America, vol. 106, No. 5, pp. 2834-2842, 1999.
[35]. T. Kuratate, H. Yehia, and E. Vatikiotis-Bateson, “Kinematics-based synthesis of realistic talking faces”. In Proc. of International Conference on Auditory-Visual Speech Processing (AVSP’98), Terrigal-Sydney, Australia, pp. 185-190.
[36]. Y. Altunbasak, A.M. Tekalp, and G. Bozdagi, “Simultaneous Stereo-motion Fusion and 3D Motion Tracking”, Proc. of International Conference on Acoustics, Speech, and Signal Processing 1995(ICASSP’95), vol.4, pp.2277-2280.
[37]. K. Nickels, S. Hutchinson, “Model-based Tracking of Complex Articulated Objects”, IEEE Trans. Robotics and Automation, vol.17, No.1, Feb. 2001.
[38]. Z. Liu, Y. Shan, Z. Zhang, “Expressive Expression Mapping with Ratio Images”, to appear in Proc. of Computer Graphics (SIGGRAPH 2001), LA, US. 2001.
[39]. E.C. Patterson, P.C. Litwinowicz, N. Greene, “Facial Animation by Spatial Mapping”, Proc. Computer Animation ‘91, N.M. Thalmann, D. Thalmann (Eds.), Springer-Verlag, pp 31 – 44, 1991.