Extraction of 3D Facial Motion Parameters from Mirror-Reflected Multi-View Video for Audio-Visual Synthesis

I-Chen Lin, Jeng-Sheng Yeh, Ming Ouhyoung

Dept. of CSIE, National Taiwan University

ABSTRACT

The goal of our project is to collect the dataset of 3D facial motion parameters for the synthesis of talking head. However, the capture of human facial motion is usually an expensive task in some related researches, since special devices must be applied, such as optical or electronic trackers.

In this paper, we propose a robust, accurate and inexpensive approach to estimate human facial motion from mirror-reflected videos. The approach takes advantages of the characteristics between original and mirrored image, and can be more robust than most of other general-purposed stereovision approach in the motion analysis for mirror-reflected videos. A preliminary dataset of facial motion parameters of MPEG-4 and French visemes and with voice data has been acquired, the estimated data are also applied to our facial animation system.

1. INTRODUCTION

Research about relation between acoustic speech and human facial motion is an important topic for several areas in computer science, such as audio-visual speech recognition, computer graphics, and etc. However, so far, the analysis or synthesis of facial motion is still a difficult work since it is not easy to acquire accurate 3D facial motion data.

For 3D facial or lip motion estimation, stereovision motion tracking, optoelectronical or magnetic space trackers are the mainstream approach. Magnetic and optoelectronical space trackers can provide extremely precise 3D position data, but the accurate devices are highly expensive, and they are unsuitable for lip surface motion tracking. Most of the stereovision motion tracking is based on the epipolar constraint and 8 points algorithm [18]. Images with multiple view directions are taken to estimate the 3D positions of feature points. [19][20][21] provide a good reference and discussion base for 3D motion and structure estimation.

A number of techniques have been developed for audio-visual synthesis of facial animation. This research about synthetic face can be approximately classified into three categories: feature-point-driven, muscle-based, and image-sample-based approach. Waters et al's work [3][4] use a physical or procedural model to synthesize facial motion. 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 [5] [6]. Pighin et al. [9], Guenter et al. [10] 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, our work is not only focused on reproducing the facial motion of a certain performer but also collecting a dataset according to articulation for further analysis. "Video Rewrite" proposed by Bregler et al. [11] synthesizes video realistic facial animation by combining image samples of faces and mouths according to input phonemes. Cosatto et al. [12] further decompose the samples into smaller facial parts and let the process of synthesis with more flexibility and efficiency.

Some researches focus on the analysis of relation between acoustic features and facial motions. Voice Puppetry [14] applied the Hidden Markov Model (HMM) to simulate facial motions driven by various audio features. Neural Networks are also adopted in some researches [15][16].

Our previous work [17] is also a speech driven talking head system, which utilized interpolation between key frames to animate smooth animation. However, there are some subtle motions cannot be simulated realistically. Thus, in our works, we used mirrors to get new images with different view directions. Besides, we proposed a more robust and simpler algorithm to estimate 3D position and motion from mirrored and front view video sequence.

2. FACIAL MOTION TRAJECTORY

From the observation and simulation by our synthesis module, we find that there are 50 points (12 for the mouth, 6 for cheeks, and 10 for the forehead, 10 for lip contours, and 12 for the lip surfaces) on a face, where the variations are the most typical of the whole motions of human faces and lips. Therefore, we take these positions as feature points to drive facial animation.

Figure 1: The image data captured by DV camera (resolution 720x480). 54 markers are placed on the subject's face and lips.

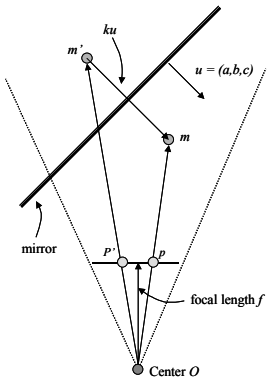


Figure 2: The geometric representation of the physical point *m*, the reflected point *m'*, and the projection points *p*, *p'*.

In order to get precise 3D position and the motion of feature points on the faces of one subject, colorful dot markers are stuck on the positions of feature points. With these markers, tracking feature point movements is much easier and more accurate.

It is well known that multiple view images 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, and used only one camera to capture the front view image and two mirrored images (as shown in Fig. 1). 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 this approach, we can acquire three different views of the image data simultaneously, and it can also avoid the problem of synchronization between data among different cameras.

In the aforementioned situation, the 3D position determination can be solved by general-purposed 3D structure reconstruction approaches, which estimate rotation matrix *R*, translation vector *T* between two cameras from fundamental matrix [21]. After getting the location and orientation of two cameras, the target point 3D positions can be approximated by the closest points to all projection rays from 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 2.2. Moreover, a flexible camera calibration method proposed by Zhang et al [23] is utilized to calculate the intrinsic parameters of the camera. With the parameters, we can calibrate the video captured by camera.

2.1. Marker Tracking in Video

A simple semi-automatic approach is applied to estimate the location variations of markers in each frame of the video clip. Once a video has been prepared for tracking, users have to initially select the position of each marker and their correspondence for markers in front and mirrored images. Our system then searches for the most probable new position of markers in the following frames by finding the positions in the searching area with the maximum correlation to the previous one. Moreover, the geometrical center of a tracked marker will be regarded as the position in that frame.

2.2. 3D Position estimation from front and mirrored images

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

$$ax + by + cz = d \qquad (1)$$

$u = (a, b, c)^t$, $\| u \| = 1$, 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 plane of camera film, *f* is the focal length, and we take the value of *f* as the unit in measurement which can be omitted in equations. 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 2, 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*, $p_i' = (f\frac{x_{mi}'}{z_{mi}'}, f\frac{y_{mi}'}{z_{mi}'}, f)^t = (x_{pi}', y_{pi}', z_{pi}')^t$.

Owing to the property of mirrors,

$$m_i' = m_i + ku, \qquad (2)$$

where k is a scale value. Thus, vector m_i, m_i', u are coplanar, and thus

$$m_i' \cdot (u \times m_i) = 0, \quad (3)$$

• is the dot product, and × is the cross product.

It can be simplified as

$$\begin{bmatrix}(y_{pi}-y_{pi}') & (-x_{pi}+x_{pi}') & (x_{pi}y_{pi}'-y_{pi}x_{pi}')\end{bmatrix}\begin{bmatrix}a\\b\\c\end{bmatrix}=0, \quad (4)$$

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

$$Mu = 0, \text{where}$$

$$M=\begin{bmatrix}(y_{p1}-y_{p1}') & (-x_{p1}+x_{p1}') & (x_{p1}y_{p1}'-y_{p1}x_{p1}')\\(y_{p2}-y_{p2}') & (-x_{p2}+x_{p2}') & (x_{p2}y_{p2}'-y_{p2}x_{p2}')\\\vdots & \vdots & \vdots\\(y_{pn}-y_{pn}') & (-x_{pn}+x_{pn}') & (x_{pn}y_{pn}'-y_{pn}x_{pn}')\end{bmatrix} \quad (5)$$

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.

There is another property of mirror is that

$$(m_i'-\Theta) = H_u(m_i-\Theta), \quad (6)$$

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

$$\begin{bmatrix}a^2x_{pi}+aby_{pi}+ac-\frac{1}{2}x_{pi} & \frac{1}{2}x_i'\\abx_{pi}+b^2y_{pi}+bc-\frac{1}{2}y_{pi} & \frac{1}{2}y_i'\\acx_{pi}+bcy_{pi}+c^2-\frac{1}{2} & \frac{1}{2}\end{bmatrix}\begin{bmatrix}z_i\\z_i'\end{bmatrix}=d\begin{bmatrix}a\\b\\c\end{bmatrix}, \quad (7)$$

From equation 7, we can find that once vector u has been determined, z_i and z_i' is proportional to variable d. Thus, after the above steps, the vector u can be estimated first, and then the scaled position of $s(x_i, y_i, z_i)$, where s is the scale value, for each marker and stationary points can be calculated by the least square method

$$min_z \, ||Gz - du||, \quad (8)$$

such as approach based on Singular Value Decomposition (SVD) or QR factorization [24]. The scale value s can be determined by comparing the scaled data with a reference ruler in real world.

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'',

$$m_i'' = H_u^{-1}(m_i'-\Theta)+\Theta, \quad (9)$$

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

2.3. Head Motion Removal

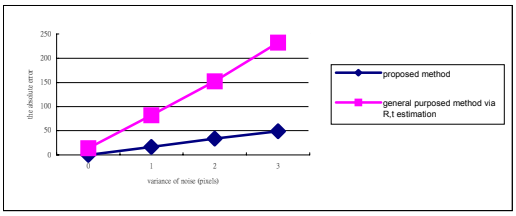
In the previous step, 3D marker positions have been estimated. However, a subject under test may swing or nod his head when speaking and making facial expressions, and thus the motions of 3D markers are composed of both facial motions and global head motions. To get precise facial motion, the head motion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

As mentioned in [20], with 3 non-colinear 3D points, the movement of rigid object can be uniquely determined by a rotation matrix R, and translation vector t.

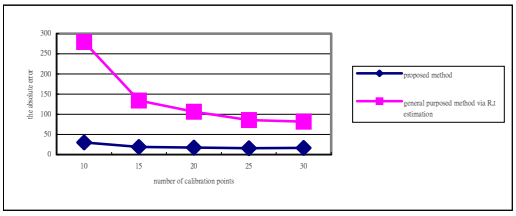
$$r_{i\,j+1} = Rr_{i\,j} + t, \quad (10)$$

where $r_{i\,j}$ is the 3D position of point i on a rigid object at time j, and where $r_{i\,j+1}$ 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 adapted as points on rigid head, and we applied the SVD (singular value decomposition) based algorithm proposed by *t*. Arun et al. [25] to determine the R and t. After the



(a)



(b)

Figure 3: Error estimation of two different solutions. We simulate the situation where normal-distributed noise perturbed the captured image. The target subject is a virtual object (about 1000x2000x1000 pixel^3) 4000 pixels far 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.

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.

2.4. Discussion of the proposed approach

Intuitively, in the case of 3D position estimation from the mirror-reflected multi-view images, the proposed 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 [22]. One of the reasons is that the degree of freedom of the rotation matrix R and the translation vector t both have 3 degrees of freedom. In our case, we evaluate the mirror plane equation, which has only 4 degree of freedom. The fewer degrees of freedom roughly mean that we can use much fewer information to reach the accuracy of the same magnitude.

Secondly, when estimating R and t from the fundamental matrix, it first has to evaluate the fundamental matrix, which is of 8 degree 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 computation, such as eigenvector estimation, SVD, and quaternion reformulation, etc. The errors are progressively accumulated by each step. [19] provides a detailed discussion of error analysis and estimation of 3D position and structure reconstruction from R, t.

We also simulate the situation where normal-distributed errors are perturbed in the captured image by computers. Fig 3 is the figure about the error distribution for our proposed approach and the approach via the virtual camera R, t estimation. The figure manifest that the virtual camera approach requires many more feature points or calibration points to reach the accuracy of the proposed approach. Our proposed approach is also more robust in the noisy situation.

3. FACIAL ANIMATION

3.1. Face Modeling

The approach mentioned in subsection 2.2 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

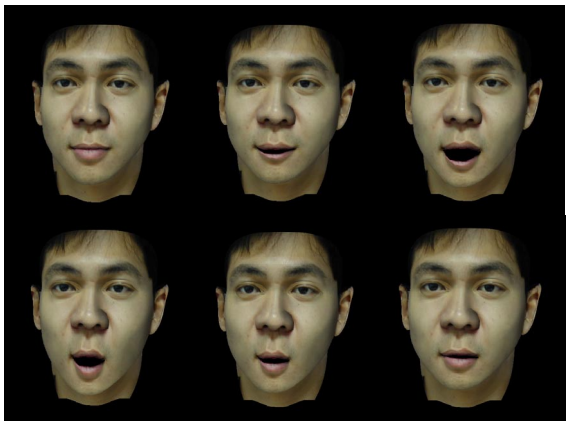


Figure 4: The synthetic facial expression when pronouncing "au".

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. 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 adopted and deformed to fit the 3D scanned range data.

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 "scatter data interpolation", which is a smooth interpolation function that gives the 3D displacements between the original points positions and the new position for every vertex. The details are described in [12][26]

3.2. Animation Issues

A synthetic 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. Control points within a region can only affect vertices in that region. These control points consist of feature points, "fixed points" and "hypothetical points". As mentioned in subsection 2.1, 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. "Hypothetical points" are the points that are difficult to capture well by viewpoint of the video; for

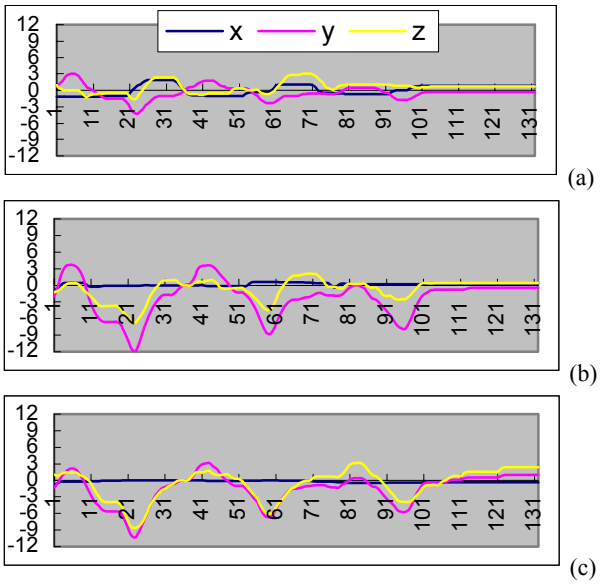


Figure 5: the part of estimated 3D motion variation (viseme O 3 times). (a) the motion of the right lip corner; (b) the motion of the lower lip tip; (c) the motion of the jaw tip. Frame (33ms) is the unit of x-axis, and mm is the unit of y-axis.

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 3 seconds as a random process. After determining the displacement of all control points, a face can be deformed by the radial basis scatter data interpolation function mentioned in subsection 3.1. Once we repeat the above similar process frame by frame, we can generate realistic facial animation according to estimated 3D facial motion data.

4. 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 [27], 40 consonant-vowel articulations, and 10 vowel-vowel articulations.

5. CONCLUSION

In this paper, we proposed an inexpensive and robust procedure to estimate 3D facial motion trajectories from front view and mirror-reflected video clips. A preliminary dataset of MPEG-4 and French visemes has been acquired. The collection of facial and lip motion dataset is still in progress. We hope that this data will be published soon through the web and applied to further research for the analysis and synthesis of the human face.

6. ACKNOWLEDGEMENT

We would like to appreciate N. P. Valles, Y. Lapire, D. Fohr, M. Pitermann and other members of speech group of Loria, France. They help our experiment of French and provide knowledge of French visemes, and also give us a lot of valuable suggestions.

7. REFERENCE

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. Y. Lee, D. Terzopoulos, and K. Waters, “Realistic Modeling for Facial Animation”, SIGGRAPH conference proceedings, pp. 55-62, ACM SIGGRAPH, August 1995.
5. T. Beier, and S. Neely, “Feature-based Image Metamorphosis”, SIGGRAPH 92 Conference Proceedings, pp. 35-42. ACM SIGGRAPH, July 1992.
6. S.M. Seitz, C.R. Dyer. “View Morphing”, Proc. SIGGRAPH 96, pp. 21-30.
7. J. Ostermann, “Animation of Synthetic Faces in MPEG-4”, Proc. of Computer Animation, pp.49-51, Philadelphia, Pennsylvania, USA, June 8-10, 1998.
8. 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.
9. 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.

10. B. Guenter, c. Grimm, D. Wood, H. Malvar, F. Pighin. "Making Face", Proc. of Computer Graphics (SIGGRAPH '98), pp. 55-66, Aug. 1998.

11. C. Bregler, M.Covell, M.Slaney. "Video Rewrite: Driving Visual Speech with Audio", Proc. SIGGRAPH'97, pp.353-360, 1997.

12. E. Cosatto and H. P. Graf, "Photo-Realistic Talking-Heads from Image Samples", IEEE Trans. Multimedia, vol. 2, no. 3, pp. 152-162, 2000.

13. V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces, ", Proc. SIGGRAPH'99, pp. 353-360, July 1999.

14. M. Brand. "Voice Puppetry", Proc. SIGGRAPH'99, pp.21-28, 1999.

15. D.W. Massaro, J. Beskow, M.M. Cohen, C.L. Fry, and T. Rodriguez, "Picture My Voice: Audio to Visual Speech Synthesis using Artificial Neural Networks", Proc. of Audio-Visual Speech Processing (AVSP'99), Santa Cruz, US, 1999.

16. E.Agelfors, J. Beskow, B. Granstrom, M. Lundeberg, G. Salvi, K. Spens, T. Ohman, "Synthetic Visual Speech driven from Auditory Speech", Proc. of Audio-Visual Speech Processing (AVSP'99), Santa Cruz, US, 1999.

17. 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.

18. H.C. Longuet-Higgins. "A computer algorithm for reconstructing a scene from two projections", Nature, 293:133-135, Sept. 1981.

19. J. Weng, T. S. Huang, N. Ahuja, *Motion and Structure from Image Sequences*, Springer-Verlag, 1993.

20. 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.

21. R. I. Hartley, "In Defence of the 8-point Algorithm", Proc. of IEEE pp. 1064-1069, 1995.

22. 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.

23. Z. Zhang, "A Flexible New Technique for Camera Calibration", Technical Report Microsoft MSR-TR-98-71, 1998.

24. G. Golub, and C. F. Van Loan, *Matrix Computation third edition*, The John Hopkins Univ. Press, Baltimore and London, 1996.

25. 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.

26. G.M. Nielson. "Scattered data modeling", in *IEEE Computer Graphics and Applications*, 13(1), pp. 60-70, Jan. 1993.

27. MPEG4 Video "Text for ISO/IEC FCD 14496-2 video", ISO/IEC JTC1/SC29/WG11 N3056, Dec. 1999.