FACIAL MOTION ANALYSIS DURING MASTICATION BASED-ON FACTORIZATION

Toshio KAWASHIMA*, Masashi TODA*, Yoshinao AOKI*
Kiwamu SAKAGUCHI**, and Takao KAWASAKI**

{School of Engineering*, School of Dentistry**}
Hokkaido University
Kita-13, Nishi-8, Sapporo, 060-8628, JAPAN
kawasima@media.eng.hokudai.ac.jp

Commission V, Working Group 4

Key words: Mastication, Motion Analysis, Factorization

Abstract:

We propose a direct facial motion estimation method based-on factorization. In the method, we can measure the facial motion of a subject masticating without marker. The measurement process is divided into two stages; learning stage and measurement stage. In the learning stage, we attach markers to a set of measurement points on a subject's face. We capture several examples of facial motion image sequences with the marker location. Once the matrix equation is derived, we can directly estimate the location of measurement points from facial image without markers. In the report, we state the detail of the method, and discuss the limitation of this approach.

1. INTRODUCTION

Face and gesture image analysis is an attractive area because the information contained in the motion data is essential for communication between human and machine. Facial motion analysis is also important for medical diagnosis. In dentistry, facial motion around lips, *perioral motion*, is an index of stomatognathic function.

Most studies[1] in dental application attach markers to subject's face. This is because precise measurement requires exact localization of characteristic points. In addition, the head of a subject must be fixed to the special chair to prevent perturbation. These restriction limits the clinical application of facial motion analysis.

In this report, we tried a direct estimation of feature points without attaching any markers to face. In the recent work of Covell [2], he proposes "*eigenpoints*" approach to locate control points from an unmarked image. His method were applied to morphing and used to match corresponding points of two images.

We follow this approach. Instead of sample face images of subjects, we preliminary measure sample image sequence of facial motion with markers. The sequence and the location of marker points are used as a training sample. From the relation ship between gray levels of an image and its marker location, we construct an estimation equation using SVD (singular value decomposition). The SVD decompose an observation into an orthonormal basis of observation and a potential motion parameter. From the result of the SVD, we form an estimation equation.

In section 2, we outline the principle of the method. Experimental results of the method are shown in section 3. A simple experiment of mastication analysis is shown in the section.

2. DIRECT ESTIMATION OF CHARACTERISTIC POINTS FROM IMAGES

Problem Definition: Estimate the location of virtual feature points of a subject from an image sequence around lips without markers.

The term "virtual feature point" is the place where a mark to be expected. In [2], they divide the measurement into two stages. The first stage calculates the estimation equation using SVD. In the stage, they mark a set of control points where geometrical correspondence between images is explicitly defined by

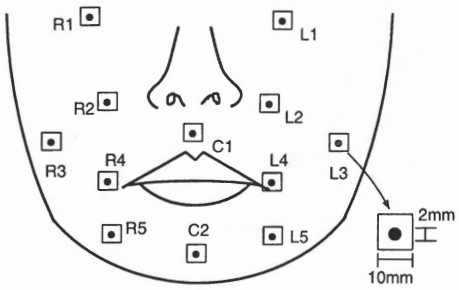


Figure 1: Fisrt stage observes a facial image sequence with markers. The trace P of the markers are measured by a tracking program. In our experiment thirteen points are located around lips. A marker is a white square (10mm×10mm) sticker in the center of which a black circle (2mm) is printed.

the user, to a sample face image set of subjects.

Instead of an image set , we prepare an image sequence which observes a facial motion. We put n markers on the face of a subject (Fig.1). A tracking program traces the two-dimensional position of the markers on the camera plane. A $2n$-length vector p_i denotes the location of these n points at time i. An $N_x \times N_y$ vector f_i is the gray scale values of the pixels around lips(figure 2) where $N_x \times N_y$ is the size of area. Matrixes F and P are the times-series measurements of vectors f_i and p_i subtracted by the expectation value.

$$F = \{f_1 - \bar{f}, \ldots, f_n - \bar{f}\}$$
$$P = \{p_1 - \bar{p}, \ldots, p_n - \bar{p}\}$$

If we suppose f and p are controlled by a potential parameter set x which governs the facial affine model, we can model the mechanism by the following equations.

$$f = M_F x + \bar{f}$$
$$p = M_P x + \bar{p}$$

Consequently, the matrix F of the image sequence can be factorized into the following equation.

$$\begin{bmatrix} F \\ P \end{bmatrix} = \begin{bmatrix} M_F \\ M_P \end{bmatrix} \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}$$
$$= \left[\begin{array}{c|c} U_F & U_\perp \\ U_P & \end{array}\right] \begin{bmatrix} \Sigma_K & 0 \\ 0 & \Sigma_\perp \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V^T \\ V_\perp^T \end{bmatrix}$$

The lower expression of the right-hand side is singular value decomposition of of $[F^T P^T]^T$.

Once $U_F, U_P, \bar{f}$, and $\bar{p}$ are determined, we can compute an estimation $\hat{p}'$ of p' for input image f' using the result of above singular value decomposition.

$$\hat{p}' = U_P U_F^{-1} \begin{bmatrix} f' - \bar{f} \end{bmatrix} + \bar{p},$$

where U_F^{-1} is general inverse of U_F. Thus, it is possible to estimate the position of the markers from

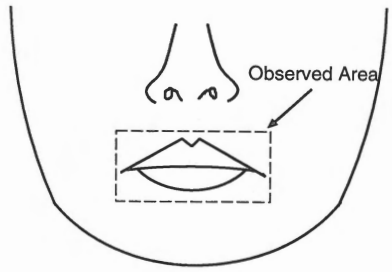


Figure 2: A vector f of gray level values are measured within the area indicated as a square. For reduction of computational cost, the pixel size of the image is re-quantized by 1:4.

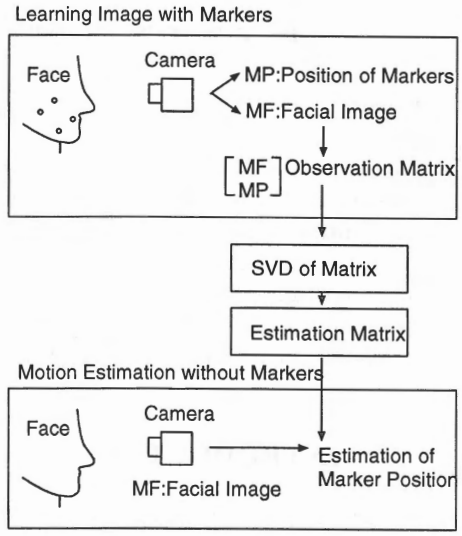


Figure 3: Block-diagram of the process. In the first learning stage, an image vector f_i of the face is measured with the positions p_i of markers. For the sequence $\{i = 1, \ldots, n\}$, singular value decomposition is applied. Using the result the algorithm estimates *virtual marker position* only from f.

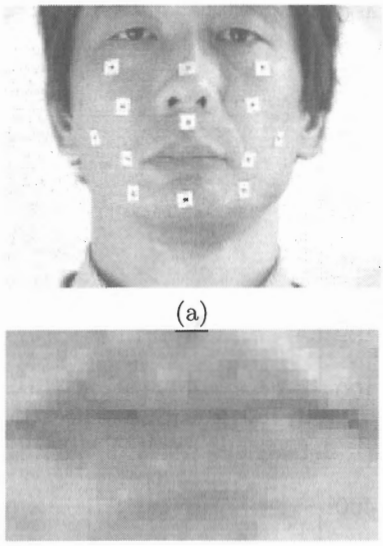
camera input without markers. In the method we do not use explicit representation of x.

In Covell's paper, index i indicates the subject number. They correlate faces $\{f\}$ and control points $\{p\}$ with U_F and U_P. Instead, we correlate perioral image and the location of feature points.

Figure 3 summarizes our process. Image sequence used in the first learning stage must span enough orthonormal basis in order that the estimation functions in the second stage. If the sequence is insufficient, the output of the second stage would be incomplete.

3. EXPERIMENTS

We observed face by a camera located in front of a subject. Thirteen markers are placed on the sub-



(a)

(b)

Figure 4: (a) An image example extracted from an image sequence used in the first stage. The pixel size is 640×480. Thirteen markers are placed around the lips. (b) Cropped area for f. The original pixel size 148×84 is re-quantized to 37×21.

ject's face as shown in figure 1. A marker tracking program extracts the markers, and determines their trajectories.

We tested the following two types of facial motion.

- Horizontal motion of face
- Chewing motion

Each motion was viewed with a CCD camera, and digitized into 640×480 pixels by a video capture. The sampling rate was 30 frames/second. The length of the sequences are between 240 to 250 frames (almost 8 seconds). Image vector f is the gray values within a square area around lips. Figure 4(a) shows the image used in the first stage. Figure 4(b) is the square area for f. The image, originally 148×84, is reduced to 37×21.

Figure 5 compares estimated position and actual position of a marker for horizontal motion. In this estimation, f_i's and p_i's for frame 0 to frame 149 are used as the learning sample. The SVD result of these frames is applied to f_i's for frame 150 to frame 200. The result shows the amplitude of the horizontal motion is reproduced only from gray level images. The marker is located just under lips, and is labeled as C2 in figure 1.

Figure 6 shows the results of proposed method applied to chewing motion. Figure (a) compares the output and the true position of point C2 during mastication. In the experiment, frames from 50 to 150

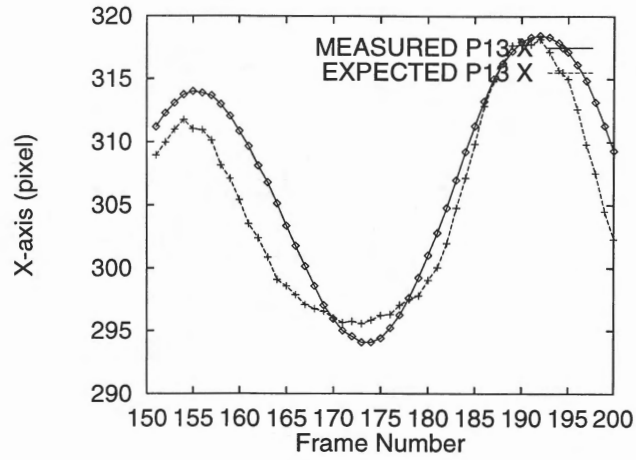


Figure 5: An example of estimation. Frames from 0 to 149 are the learning sample. The position of the markers are estimated from f for 150-200 frames. Crosses: estimation. Diamonds: true position.

are used as the training sample. The estimation error is much greater than figure 5 because the subject's face motion made the basis insufficient. Nevertheless, rough motion is reproduced by the estimation. Figures (b)-(d) are the true trajectories and estimated trajectories for thirteen points. Figure (c) is the result for the training sample. Figure (d) is the case that the SVD result for (c) is applied to another input sequence. Results (a) and (c) show that rough motion can be recovered with the method. However, the motion which cannot be spanned by the training sample will be distorted by the method.

In the example, the first 50 frames of the training sample is not spanned by frames 50-150. This caused the error in the first frames large (figure (a)).

4. CONCLUSIONS

An advantage of our proposed method is that the algorithm measures virtual marker locations without makers once the estimation equation is created in the first stage. In clinical application, an examination is often done periodically. In such a case, it is troublesome to attach the markers to the same location of the face. Our method does not require any preparation after the first measurement.

A problem in the current method is that the orthonormal basis must span sufficient image space. The algorithm outputs incorrect estimates if the input image is not supported by the basis. The sample image sequences must be carefully chosen to satisfy the requirement. Practically, if the facial position is fixed, the variation of sample images will be reduced because no degree of freedom is required for trans-

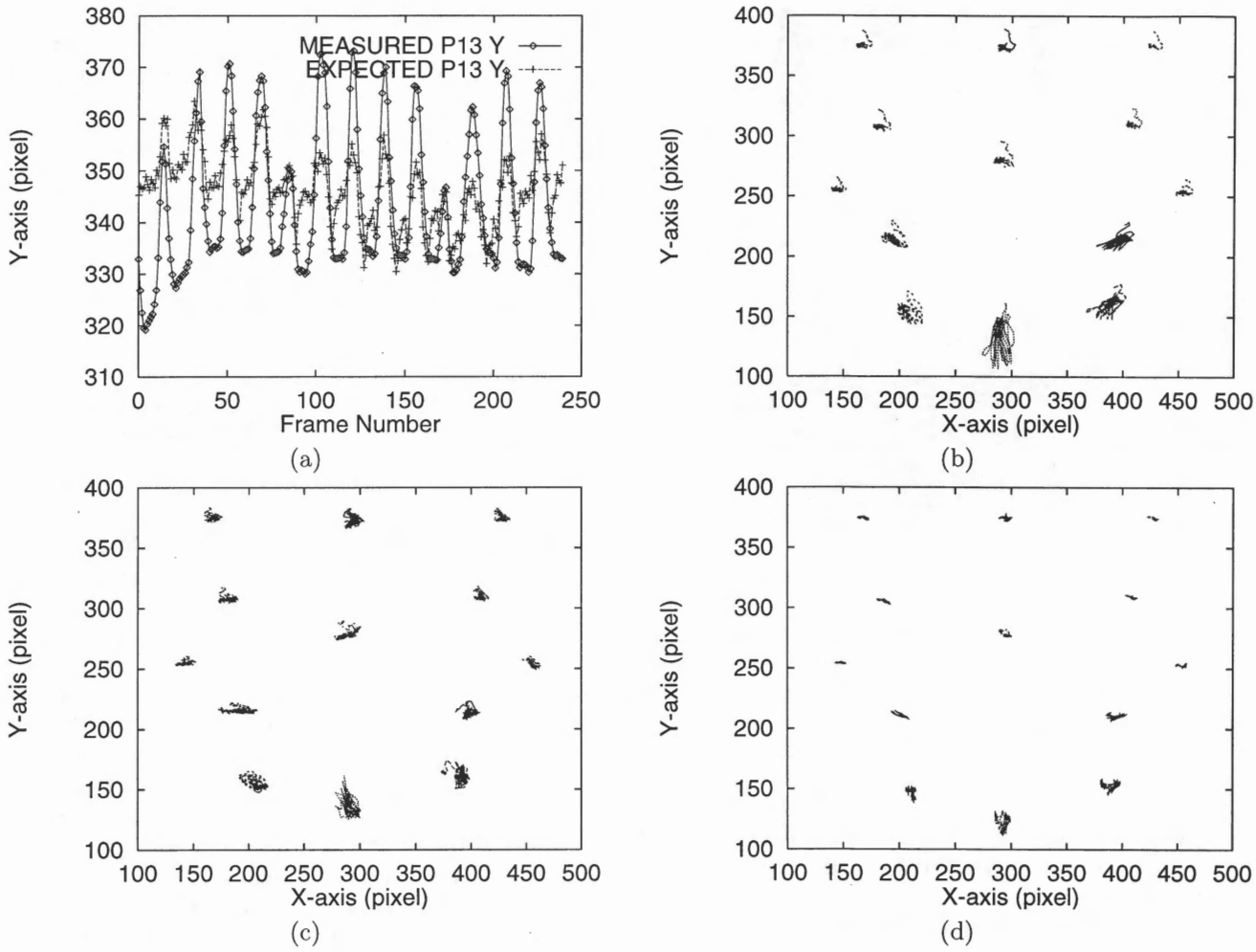


Figure 6: (a)Vertical motion of the point C2. Cross: estimation. Diamond: true position. The amplitude of the expected vertical motion is smaller than actual motion. This is probably caused by in sufficient span of the basis. Figures (b)-(d) are the trajectories of markers. Figure (b) is the output of tracker (actual motion). Figure (c) is the estimated trajectory for the training sample, and (d) is the result for an test image.

lational motion. Distortion shown in the experiment is caused by the insufficiency. However, collecting enough sample images requires subject's effort.

Our future work will concentrate on the reduction of the distortion caused by insufficient orthonormal basis.

5. REFERENCES

[1] Masashi Setaka, Time-series analysis of multi-point movements of lips and adjacent regions during mastication, Hokkaido Journal of Dental Science, 15, pp.89-107, 1994 (in Japanese)

[2] M. Covell, Eigen-points: Control-point Location using Principal Component Analysis, 2nd Int. Conf. on Automatic Face and Gesture Recognition, 1996