Human Hand Tracking from Binocular Image Sequences

Kenichi NIREI†, Hideo SAITO†,
Masaaki MOCHIMARU‡ and Shinji OZAWA†

† Dept. of Electrical Engr., Keio University,
3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223, JAPAN
‡ Mechanical Factors Lab., Nat. Inst. of Biosci. and Human-tech.,
1-1 Higashi, Tsukuba 305, JAPAN

Abstract **– Sensing of human hand motion is very important for a variety of applications, such as CG animation and athletic performance measurement. Tracking a hand is difficult because the hand has high degree of freedom articulated mechanisms. This paper presents a 3-D model-based hand tracking method which is robust to occlusions and local minima. Tracking is performed minimizing estimation error of an optical flow and maximizing the overlap between a projected model and a silhouette image. We employ stochastic optimization to solve them, which are generally difficult. We present experimental results on tracking from synthetic and real image sequences.**

I. INTRODUCTION

Sensing of human hand motion is very important for a variety of applications, such as CG animation and athletic performance measurement. However, it is very hard to track a hand because it has high degree of freedom articulated mechanisms.

General solutions to this problem are divided into two categories. One is use of gloves with sensors [1] and the other is use of computer vision techniques. Although the former can give real-time processing and reliable information, it imposes a burden on the user and makes sensing natural human motion difficult. On the other hand, the latter is suitable for hand tracking since it is passive sensing and captures natural motion.

Previous work on vision-based hand tracking includes [2], [3], [4]. Mochimaru et al.[2] proposes a system for tracking a hand maximizing the overlap between a projected 3-D hand model and a silhouette image on the basis of information at the previous frame. The silhouette image itself has less information and it makes this method weak in occlusions, which often happen in real world. Moreover, this maximization is performed by simple hill-climbing so that it is easy to be trapped in local minima. Kameda et al.[3] estimates poses of a hand transforming a 3-D hand model which consists of patches. Since this method uses a precise model, it takes vast computational time. In [4], Rehg et al. describe a method to use the constraints of global shape features in order to solve kinematic equations of the hand model. They assume that there is no occlusions.

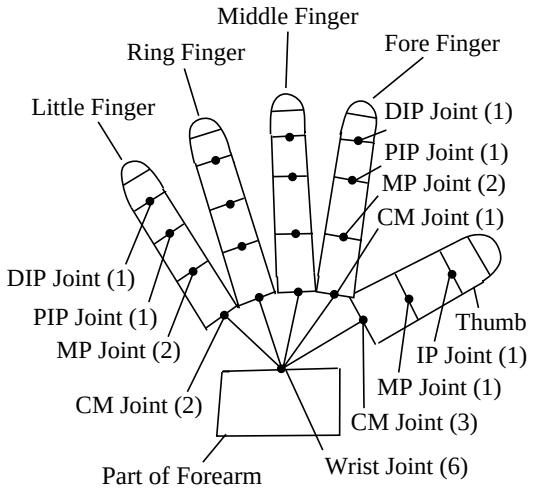


Figure 1: 21 segments and 20 joints. Figures at each joint represent their degree of freedom.

This paper presents a 3-D model-based hand tracking method which is robust to occlusions and local minima. Tracking is performed minimizing estimation error of an optical flow and maximizing the overlap between the projected model and the silhouette image. We employ stochastic optimization to solve them, which are generally difficult.

II. HAND MODEL

Hand tracking is performed fitting the 3-D hand model to the hand in the image. The model represents all possible hand poses.

The hand is modeled as a collection of 21 segments and 20 joints on the basis of anatomy (Fig. 1). The thumb, each of four fingers, and a part of the forearm

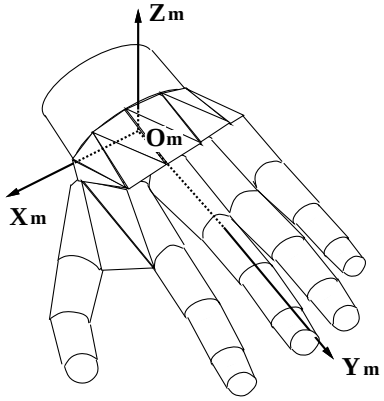


Figure 2: 3-D hand model and its coordinate system.

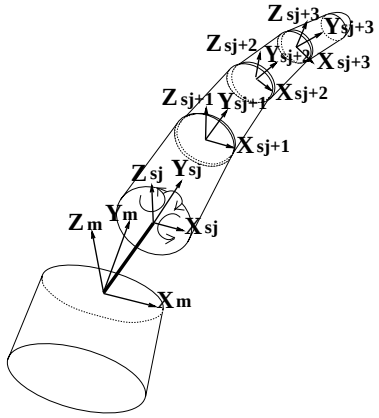


Figure 3: Segment coordinate system.

have 3, 4, and 1 segments, respectively. Each segment is described as a rigid truncated elliptic corn. Furthermore, each fingertip is described as an rigid elliptic hemisphere. The CM joint of each digit and wrist joint is handled only as skeletal model which has no shape. The gaps are covered with triangular patches (Fig. 2).

The degree of freedom for each joint is defined from its function. The wrist joint which describes the original position and pose of the model has 3 translation and 3 rotation degree of freedom (Fig. 2). As shown in Fig. 3, each segment of digits has its own segment coordinate system which has 1 - 3 degree of freedom. After all, the hand pose is represented by the model with 33 degree of freedom.

Since the shapes of are determined by actual measurement, estimation of hand pose results in estimation of 33 parameters.

III. STOCHASTIC OPTIMIZATION

Tracking articulated mechanisms such as a hand is more difficult than that of a single rigid object because their search space is larger and their state equations are nonlinear. In this case, general methods such as a least-squares method [5] and simple hill-climbing tend to be trapped in local minima and cost computational time. Contrary to this, stochastic optimization does not depend on the characteristic of search space very much and gives better solution. We employ two stochastic optimization methods, a genetic algorithm (GA) and a simulated annealing (SA).

The GA is a search algorithm based on the mechanics of natural selection and natural genetics [6]. It quickly approaches neighborhood of the best answer in large search space, while it is weak in local search. The SA is an optimization method based on temperature schedule in annealing. It is good at local search, while it is expensive computationally with large search space.

In our approach, each optimization method is put to proper use.

IV. HAND TRACKING

We estimate the pose of the hand in the same way as [2], that is, maximizing the overlap between the projected model and the silhouette image on the basis of information at the previous frame. This method works well in the case which these is no occluded area and the motion of hand is small but not in the other cases. We solve this problem utilizing the optical flow (OF) to make good use of characteristics of image sequences.

The algorithm is shown in Fig. 4

A. ESTIMATION OF INITIAL PARAMETERS

At first frame, initial parameter is estimated maximizing the overlap between the silhouette image and the projected model. The silhouette image is made binalizing the input frame. To utilize shape information of projection and silhouette, we employ a distance transform.

The overlap is defined as follows:

$$E_o = \frac{\sum_{x,y} f(x,y)g(x,y)}{\sqrt{\sum_{x,y}|f(x,y)|^2}\sqrt{\sum_{x,y}|g(x,y)|^2}} + w\sum_i \frac{O_i}{M_i} \quad (1)$$

where $f(x,y)$ and $g(x,y)$ are distance transform images of the silhouette image and the model projection on the image plane, respectively. As shown in Fig. 5, M_i and O_i are the area of the projected finger model, and the overlap region of the silhouette with M_i,

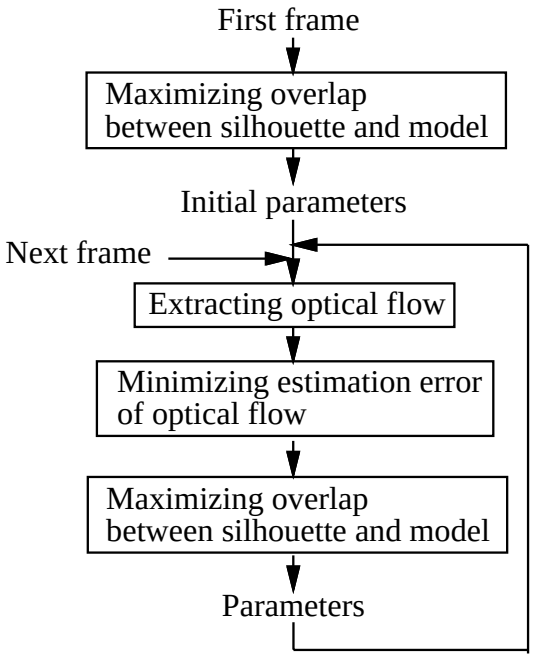


Figure 4: Algorithm of proposed method.

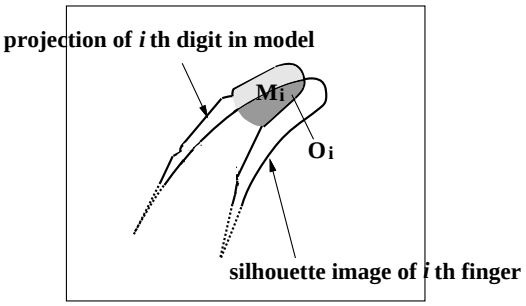


Figure 5: Overlap between the projected model and the silhouette.

respectively. Here, i represents the number of the finger, and w represents the weighting factor. The second term is defined for evaluating the overlap in the finger tip region which is important for tracking the finger motion.

Only at first frame, rough position of the hand is given by mouse input. However, search space is still very large. Therefore, E_o in Eq.(1) is maximized by using the GA which quickly approaches neighborhood of the best answer in large search space, and then sub-optimum solution can be obtained. Then the solution is maximized by the SA which is good at local search to improve the answer.

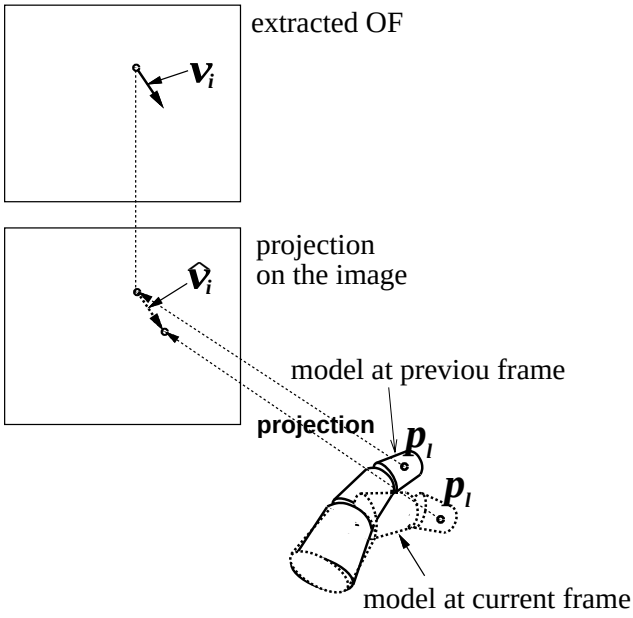


Figure 6: Extracted flow vector and its corresponding estimated flow vector

B. ESTIMATION OF PARAMETERS

After estimation of initial parameters, parameters at each frame are estimated minimizing estimation error of OF. The OF is extracted between the previous frame and the current frame.

The estimation error of the OF is defined as follows:

$$E_f = \frac{\sqrt{\sum_i \|\boldsymbol{v}_i - \widehat{\boldsymbol{v}}_i\|^2}}{n} \qquad (2)$$

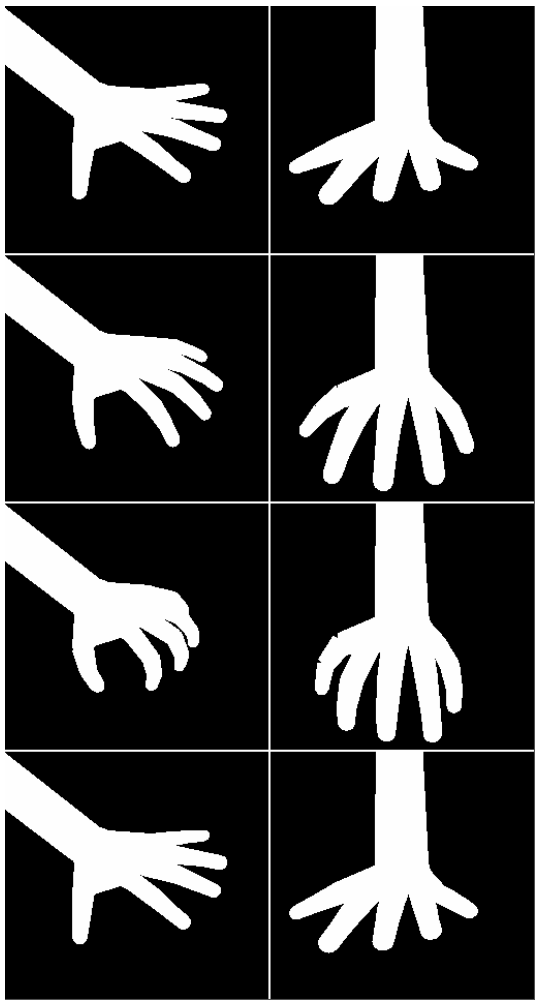
where $\boldsymbol{v}_i$ is the ith extracted flow vector, and $\widehat{\boldsymbol{v}}_i$ is its corresponding estimated flow vector calculated from the model at previous frame and the one at current frame (Fig. 6), n is the number of corresponded flow vector.

The parameter difference between successive two frames is a little and its search space is small. Therefore, (2) is minimized by the SA which is good at local search. Furthermore, the answer is improved maximizing (1) using the SA.

As a result of these, parameters at each frame are estimated. Hand tracking is performed successively implementing this process for image sequences.

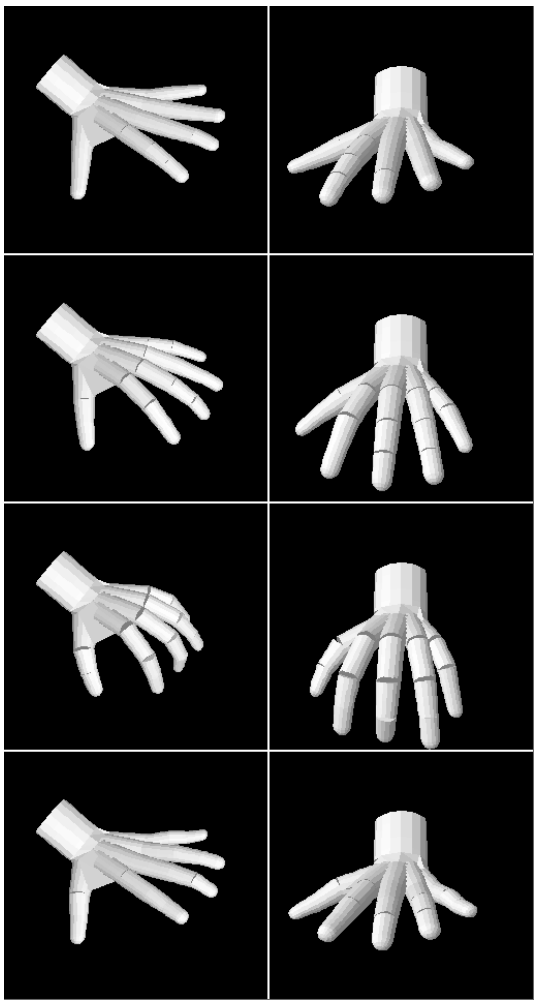
V. EXPERIMENTAL RESULTS

To test the method described above, we made experiments on tracking for both synthetic and real image sequences. Two cameras were used for avoiding the occlusion. In these experiments, the wrist joint was fixed to simplify the process. Then the parameters of joint angle of every finger are estimated. Furthermore,



Camera 1 View Camera 2 View

Figure 7: Synthetic images. Samples were taken at frame 1, 15, 30, 60.



Camera 1 View Camera 2 View

Figure 8: Estimated hand poses for the samples in Fig.7.

the CM joints were fixed after first frame because their motions seem to be small enough to be neglected.

A. EXPERIMENT USING SYNTHETIC IMAGE SEQUENCE

Using the 3-D hand model, we made a synthetic image sequence consists of 60 frames and OFs. This sequence simulates "clenching" and "unclenching." Fig. 7 shows input images and Fig. 8 shows the estimated hand poses.

Generally, tracking occluded area such as the little finger at frame 30 is difficult. We, however, successfully tracked it as shown in Fig. 8. Furthermore, the estimated parameters at the MP joint of the fore finger and the PIP joint of the little finger are shown in Fig. 9. The change of parameters are extracted though they contain errors.

These results demonstrate effectiveness of our tracking method.

B. EXPERIMENT USING REAL IMAGE SEQUENCE

The real image sequence which consists of 50 frames deals with "clenching." Fig. 10 shows input images and the estimated hand poses.

Although the estimation error was accumulated, tracking succeeded on the whole. Furthermore, the estimated parameters at the MP joint of the five digits are shown in Fig. 11. The estimated parameters represent the motion "clenching."

Effectiveness of our hand tracking method against real image sequence is shown by these results.

VI. CONCLUSION

We have proposed a model-based hand tracking method performed by minimizing estimation error of an OF and maximizing the overlap between the projected model and the silhouette image using the GA and the SA. We have demonstrated effectiveness of this method by experiments.

We will improve OF extraction process which has great effect on tracking. Moreover, we will repeatedly make experiments on various image sequences.

References

[1] T. Takahashi and F. Kishino, "A hand gesture recognition method and its application", *Trans. IEICE*, J73-D-II(12), pp.1895-1992, 1990.

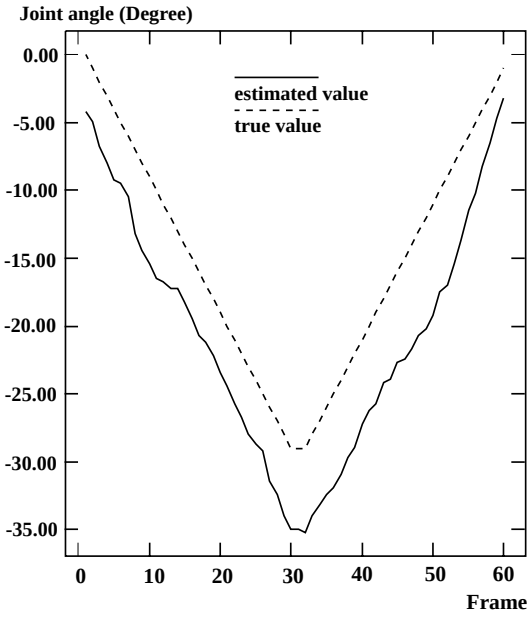
[2] M. Mochimaru and N. Yamazaki, "The three-dimensional measurement of unconstrained motion using a model-matching method", *ERGONOMICS*, vol.37, No.3, pp.493-510, 1994.

[3] Y. Kameda, M. Minoh, and K. Ikeda, "Three Dimensional Pose Estimation of an Articulated Object from its silhouette Image", *ACCV '93*, pp.612-615, 1993.

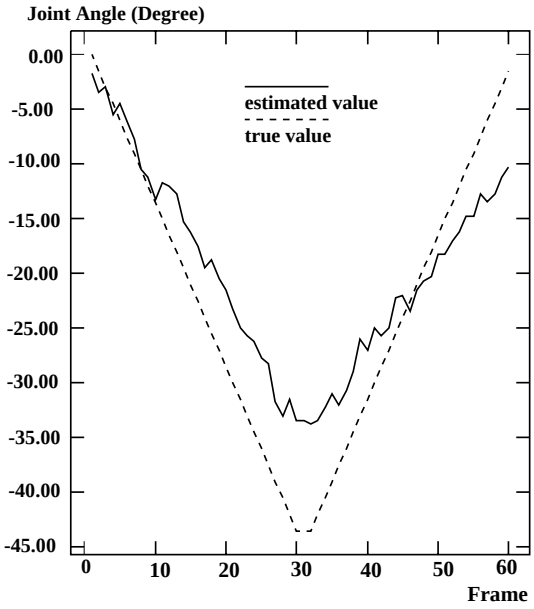
[4] J. M. Rehg and T. Kanade, "Visual Tracking of High DOF Articulated Structures: an Application to Human Hand Tracking", *ECCV '94*, pp.35-46, 1994.

[5] D. G. Lowe, "Fitting Parameterized Three-Dimensional Models to Images: *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(5), pp.441-449, 1991.

[6] D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, 1989.



(a) Rotation(bending) at MP joint of fore finer.



(b) Rotation at PIP joint of little finger.

Figure 9: Estimated parameters for the synthetic image sequence.

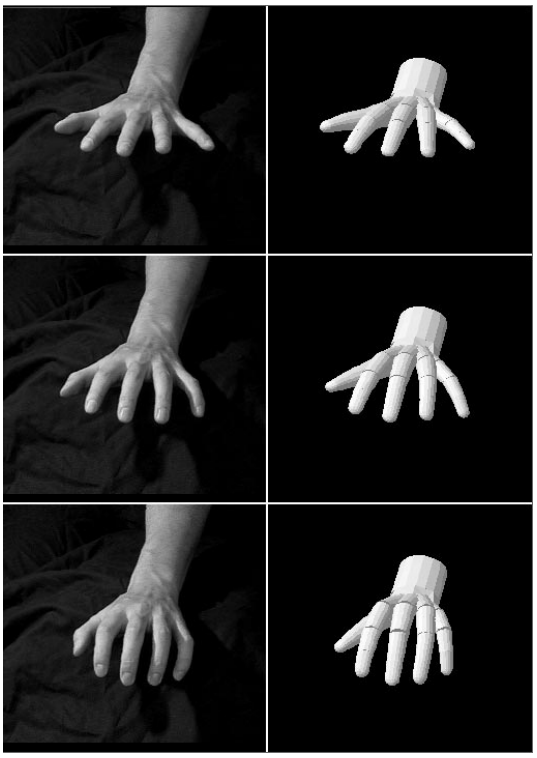


Figure 10: Real images at frame 1, 25, 49 and estimated hand poses.

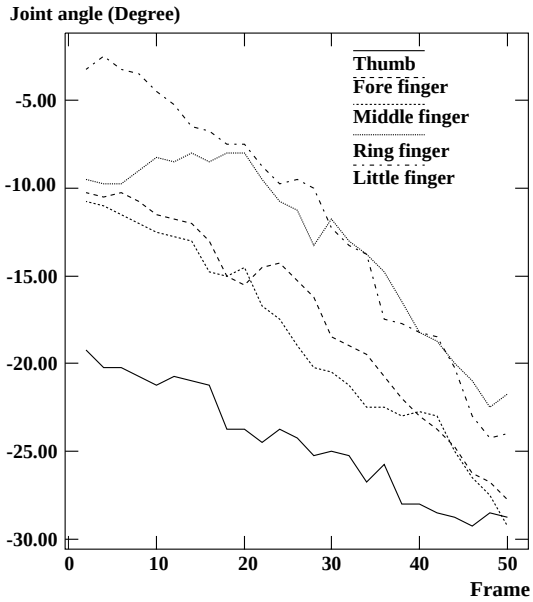


Figure 11: Estimated rotation(bending) at MP joint for the real image sequence.