

# Grasp Recognition using a 3D Articulated Model and Infrared Images

Koichi Ogawara

Institute of Industrial Science,  
Univ. of Tokyo, Tokyo, Japan

Jun Takamatsu

Institute of Industrial Science,  
Univ. of Tokyo, Tokyo, Japan

Kentaro Hashimoto

Fuji Xerox Information Systems Co.,Ltd.  
Tokyo, 150-0031, Japan

Katsushi Ikeuchi

Institute of Industrial Science,  
Univ. of Tokyo, Tokyo, Japan

**Abstract**—A technique to recognize the shape of a grasping hand during manipulation tasks is proposed; which utilizes a 3D articulated hand model and a reconstructed 3D volume from infrared cameras. Vision-based recognition of a grasping hand is a tough problem, because a hand may be partially occluded by a grasped object and the ratio of occlusion changes along the progress of the task. To recognize the shape in a single time frame, a robust recognition method of an articulated object is proposed. In this method, 3D volumetric representation of a hand is reconstructed from multiple silhouette images and a 3D articulated object model is fitted to the reconstructed data to estimate the pose and the joint angles. To deal with large occlusion, a technique to simultaneously estimate time series reconstructed volumes with the above method is proposed, which can automatically suppress the effect from badly reconstructed volumes. The proposed techniques are verified in simulation as well as in a real world.

## I. INTRODUCTION

Robust hand shape recognition technique is a key requirement for flexible human computer interface(HCI), tele-operation and teaching robot behavior. As for teaching robot behavior, much attention is paid to “Learning from Observation” paradigm in recent years, in which a robot system observes a demonstration of manipulation tasks and learns a procedure of a task automatically for purposes of reproduction. To reproduce delicate manipulation, accurate recognition of the hand shape is necessary as well as object recognition, intention understanding, etc.

The approaches to recognize the hand shape can be divided into two classes; one is to utilize contact-based sensing devices, such as data gloves, on top of a hand and measure the hand shape directly[1]. While the other approach is to recognize the hand shape by using computer vision techniques, which has the advantage of being a natural user interface, thus many studies have been attempted toward this direction so far.

The purpose of our research is to recognize the hand shape during manipulation tasks by vision, and the issues to be solved are summarized as follows. (1) What geometric representation should be reconstructed by vision which is appropriate for being recognized. (2) How to estimate

the shape of an articulated object, hand in this case, from the representation. (3) How to deal with occlusion problem in that an acquired image may be partially occluded by a manipulated object.

The basic ideas are (1) to represent a 3D geometry of a hand from multiple infrared cameras by volume intersection approach, (2) to extend our 3D robust pose estimation technique to deal with an articulated object model and (3) to utilize time series reconstructed volumes to avoid occlusion problem on the assumption that the orientation of the hand changes while the joint angles do not.

In section II, the related research is referred and the advantage of our approach is briefly mentioned. In section III, a method to reconstruct 3D volumetric representation from multiple cameras is described. The representation of a 3D articulated hand model is also described here. A robust estimation technique for an articulated object model from a single reconstructed volume is proposed in section IV. To apply this technique to a grasping hand, a simultaneous estimation technique is proposed in section V which utilizes time series reconstructed volumes for automatically selecting well reconstructed, i.e. not much occluded, volumes to be weighted higher in the recognition process. In section VI, some experimental results are shown. We conclude in section VII.

## II. RELATED RESEARCH

### A. Hand Shape Recognition

Vision based recognition techniques of the hand shape are ranging from a 2D model fitting from a single camera system to a 3D model fitting from a multiple cameras system[2]. Among them, Rehg et al. takes a 2D approach and deals with self occlusion by utilizing 2D templates of finger tips which are ordered along z-axis [3]. Delamarre et al. reconstructs a 3D surface of a hand from a stereo vision system and estimates the hand shape by shifting a 3D articulated hand model based on virtual forces between



Fig. 1. Intensity Image(left) and Infrared Image(right)

the model surface and the reconstructed hand surface[4]. Ueda et al. reconstructs the hand shape by volume intersection approach using multiple cameras and applies virtual forces between the 3D reconstructed volume and an articulated hand surface model [5]. However, most of the approaches so far assume an empty hand, i.e. a hand without grasping, so they are not suitable for recognition of the hand shape during manipulation tasks in which a hand may be partially occluded by a manipulated object.

Our approach employs a volume intersection technique similar to [5] to reconstruct 3D representation of a hand. While they aim for real-time gesture recognition and efficiency in computation is their main concern, we developed much robust fitting techniques and time-series hand motion data are utilized to disregard badly reconstructed volumes to estimate the shape of a grasping hand.

### B. 3D Model Fitting

With regard to localization techniques between a rigid 3D geometric model and 3D range data, Besl et al. proposed iterative closest point (ICP) algorithm which iteratively searches for the corresponding model vertex to each 3D range point. Then the optimal correspondence, i.e. the pose of the model, is calculated by solving the least squares method using all the correspondences[6]. However ICP is sensitive to outliers, so Wheeler et al. proposed 3D template matching (3DTM) technique which utilizes M-estimator from robust statistics to exclude the effect of outliers[7]. This technique has been used, for example, to track a manipulated object in 3D space[8]. We extend 3DTM to handle an articulated object model so that the joint angles as well as the pose of a hand can be robustly estimated by applying an articulated hand model to a reconstructed hand volume.

## III. DATA REPRESENTATION

### A. Volumetric Reconstruction from Infrared Images

In this study, infrared cameras are selected because the areas corresponding to a hand silhouette is easily extracted from infrared images as shown in Fig. 1. Three images are taken at a time from three infrared cameras whose optical

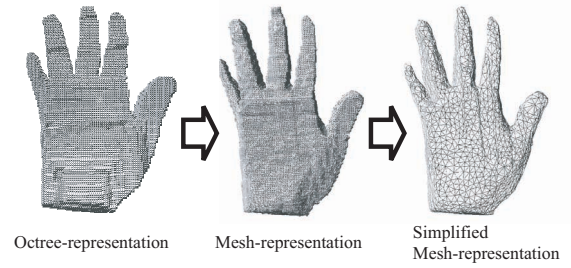


Fig. 2. Reconstructed Hand Volume

axes intersect at right angles to each other. A volumetric representation is generated from three silhouette images by using volume intersection technique enhanced by Octree data structure[9] as shown in Fig. 2(left).

The Octree representation is converted to mesh representation by applying Marching-cubes algorithm[10] as shown in Fig. 2(center). To reduce calculation cost later, this representation is further simplified by applying a mesh simplification method[11] as shown in Fig. 2(right).

### B. Articulated Hand Surface Model

A hand model used in this study is composed of a kinematic bone model and a surface mesh model.

1) *Kinematic Bone Model*: The kinematic bone model is represented by bones in anatomically correct tree-structure. A connection between bones indicates a joint and 15 joints (20 D.O.F.) are modeled as shown in Fig. 3(left). In addition to 20 D.O.F. for joints, the bone model has extra 6 D.O.F., 3 D.O.F. for translation and 3 D.O.F. for rotation, which determines the pose of the entire hand.

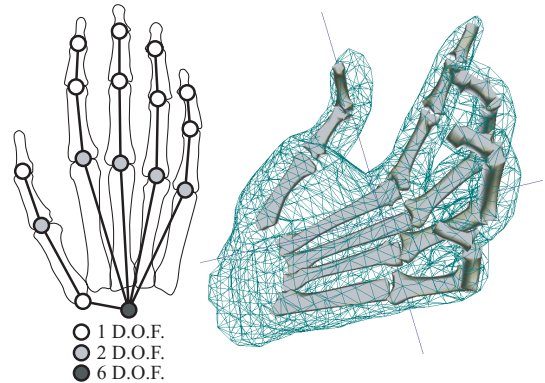


Fig. 3. Articulated Hand Model

2) *Surface Mesh Model*: A surface model of a hand is constructed by measuring a real hand using a range finder and a mesh model is generated by the same way as in the case of hand volume reconstruction.

Then the kinematic bone model and the surface mesh model are overlapped and each vertex in the surface model is linked to the corresponding two bones (the nearest and the second nearest). As a bone rotates about a joint, all the linked vertices move to a new location which is determined by weighted interior division of two relative relations(links); so the mesh model deforms smoothly around each joint(Fig. 3(right)).

#### IV. ESTIMATION OF THE HAND SHAPE

The pose and the shape of a reconstructed hand volume is estimated by fitting the articulated hand model in 3D space. As a fitting algorithm, 3DTM localization algorithm[7] is extended to deal with an articulated object model.

In the subsequent sections, 3DTM algorithm is explained and the extended 3DTM algorithm is presented.

##### A. 3DTM

The point  $\mathbf{r}_i$  in the reconstructed volume corresponding to a point  $\mathbf{m}_i$  in the hand model can be represented with two parameters,  $\mathbf{t}$ , a translation vector, and  $\mathbf{R}$ , a rotation matrix as Eq.(1).

$$\mathbf{r}_i = \mathbf{R}\mathbf{m}_i + \mathbf{t} + \beta \quad (1)$$

where  $\beta$  is random 3D noise. If  $\beta$  follows a gaussian distribution,  $\langle \mathbf{R}, \mathbf{t} \rangle$  can be estimated by minimizing Eq.(2) by solving the least squares method.

$$f(\mathbf{R}, \mathbf{t}) = \sum_i \|\mathbf{R}\mathbf{m}_i + \mathbf{t} - \mathbf{r}_i\|^2 \quad (2)$$

But the real error distribution usually doesn't follow a gaussian distribution and the effect of outliers makes the localization process unstable. Therefore, Wheeler proposed a technique to apply M-estimator to estimate the real error distribution[7]. M-estimator is a generalized form of the least squares method and is formulated as Eq.(3).

$$E(\mathbf{p}) = \sum_i \rho(z_i) \quad (3)$$

where  $\mathbf{p} = (\mathbf{t}^T, \mathbf{q}^T)^T$  is the position and the rotation of the rigid object model and  $\mathbf{q}$  is a 4 D.O.F. quaternion vector which is equivalent to  $\mathbf{R}$ .  $\rho(z_i)$  is an arbitrary function of the error  $z_i (= \beta^2)$ .

$\mathbf{p}$  which minimizes  $E(\mathbf{p})$  can be solved from Eq.(4).

$$\frac{\delta E}{\delta \mathbf{p}} = \sum_i \frac{\delta \rho(z_i)}{\delta z_i} \frac{\delta z_i}{\delta \mathbf{p}} \quad (4)$$

Here, we introduce  $w(z)$  as a weight function which represents an error term as Eq.(5).

$$w(z) = \frac{1}{z} \frac{\delta \rho}{\delta z} \quad (5)$$

With Eq.(5), Eq.(4) can be rewritten as Eq.(6). If we ignore the fact that  $w(z)$  is a function of  $z$ , this is a form of weighted least squares.

$$\frac{\delta E}{\delta \mathbf{p}} = \sum_i w(z_i) z_i \frac{\delta z_i}{\delta \mathbf{p}} \quad (6)$$

In this study, Lorentzian distribution is chosen as a probability distribution of error to exclude the effect of outliers and the weight function is defined as follows.

$$w(z) = \left(1 + \frac{1}{2} \left(\frac{z}{\sigma}\right)^2\right)^{-1} \quad (7)$$

Eq.(6) can be solved by conjugate gradient search algorithm and  $\mathbf{p}$  which minimizes the error is obtained.

##### B. Extension to an Articulated Model

In section IV-A,  $E(\mathbf{p})$  is described as Eq.(8).

$$E(\mathbf{p}) = \sum_i \rho \left( \|\mathbf{R}\mathbf{m}_i + \mathbf{t} - \mathbf{r}_j\|^2 \right) \quad (8)$$

To handle an articulated model, the above equation is rewritten as Eq.(9).

$$E(\mathbf{p}, \boldsymbol{\theta}) = \sum_i \rho \left( \|\mathbf{R}\mathbf{m}_i(\boldsymbol{\theta}) + \mathbf{t} - \mathbf{r}_j\|^2 \right) \quad (9)$$

$$\begin{pmatrix} \mathbf{m}_i(\boldsymbol{\theta}) \\ 1 \end{pmatrix} = \prod_l T_l(\theta_l) \cdot \begin{pmatrix} \mathbf{m}_i \\ 1 \end{pmatrix}$$

where  $T_l(\theta_l)$  is a  $4 \times 4$  homogeneous matrix which takes  $l$ -th joint angle and converts a point in the coordinate frame of a child link to that of the parent link.

Now the estimation process can be written as follows.

- 1) **repeat**
- 2)  $\mathbf{p} = \mathbf{p}', \boldsymbol{\theta} = \boldsymbol{\theta}'$
- 3) Calculate  $E(\mathbf{p}, \boldsymbol{\theta})$
- 4) Calculate gradient  $\frac{\partial E(\mathbf{p}, \boldsymbol{\theta})}{\partial \mathbf{p}}$
- 5) Estimate new  $\mathbf{p}'$  which minimizes  $E(\mathbf{p}', \boldsymbol{\theta})$
- 6) Calculate  $E(\mathbf{p}', \boldsymbol{\theta})$
- 7) Calculate gradient  $\frac{\partial E(\mathbf{p}', \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$
- 8) Estimate new  $\boldsymbol{\theta}'$  which minimizes  $E(\mathbf{p}', \boldsymbol{\theta}')$
- 9) **until**  $|E(\mathbf{p}', \boldsymbol{\theta}') - E(\mathbf{p}, \boldsymbol{\theta})| < \varepsilon$

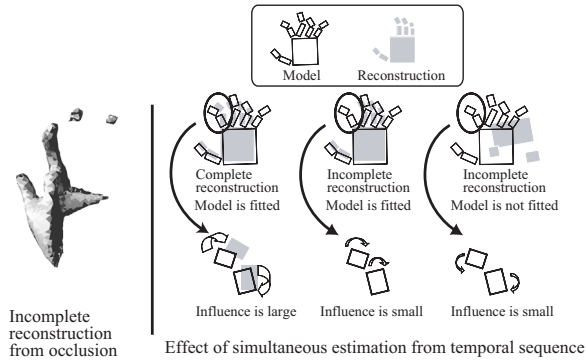


Fig. 4. Effect of Simultaneous Estimation

## V. SIMULTANEOUS ESTIMATION WITH TEMPORAL SEQUENCE

A single reconstructed volume may lead to bad estimation because of occlusion caused by a manipulated object as shown in Fig. 4(left). The ratio of occlusion depends on the orientation of the hand at that time.

However, in typical manipulation tasks, we can assume that a demonstrator moves and rotates the manipulated object without changing grasp itself. In this case, the recognition ratio can be improved by using temporal sequence of reconstructed volumes.

In this study, we assume that the shape of the grasp is not changed during observation and that the rough initial pose, not shape, of the first time frame is given. The rough initial pose can be estimated by a random sample approach or an image template based approach. When the pose of the first frame is known, the pose in the subsequent frames can be estimated by applying the proposed fitting algorithm by setting the previous frame's pose as the initial pose of the current frame.

Here a method to suppress the effect of occlusion by using reconstructed volumes for  $N$  continuous frames simultaneously is proposed.

### A. Simultaneous Estimation

From Eq.(9),  $E(\mathbf{p}_k, \boldsymbol{\theta})$  for each frame  $k$  is described as Eq.(10).

$$E(\mathbf{p}_k, \boldsymbol{\theta}) = \sum_i P \left( \|\mathbf{R}_k \mathbf{m}_{k,i}(\boldsymbol{\theta}) + \mathbf{t}_k - \mathbf{r}_{k,j}\|^2 \right) \quad (10)$$

Because  $\boldsymbol{\theta}$  is maintained as the same for all the data from the assumption, the simultaneous estimation is defined as to estimate  $\mathbf{p}_k, \boldsymbol{\theta}$  which minimizes  $E_{\text{total}}$ .

$$E_{\text{total}} = \sum_{k=1}^n E(\mathbf{p}_k, \boldsymbol{\theta}) \quad (11)$$

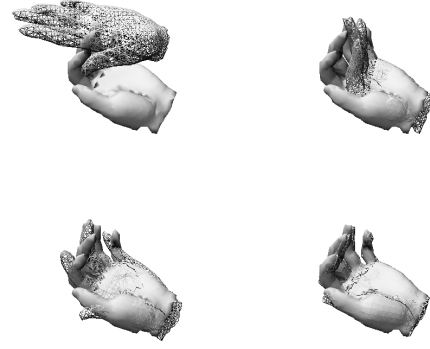


Fig. 5. Procedure of Fitting

The procedure of estimation is as follows.

- 1) For each reconstructed result, calculate  $\mathbf{p}'_k, \boldsymbol{\theta}'_k$ .
- 2) Calculate  $\boldsymbol{\theta}'$  which is the average of  $\boldsymbol{\theta}'_k$ .
- 3) **repeat**
- 4)  $\mathbf{p}_1 = \mathbf{p}'_1, \dots, \mathbf{p}_k = \mathbf{p}'_k, \boldsymbol{\theta} = \boldsymbol{\theta}'$
- 5) Fix  $\boldsymbol{\theta}$  and calculate  $\mathbf{p}_1, \dots, \mathbf{p}_N$  iteratively.
- 6) Fix  $\mathbf{p}_1, \dots, \mathbf{p}_N$  and calculate  $\boldsymbol{\theta}$  iteratively.
- 7) **until**  $|E_{\text{total}}(\mathbf{p}'_1, \dots, \mathbf{p}'_k, \boldsymbol{\theta}') - E_{\text{total}}(\mathbf{p}_1, \dots, \mathbf{p}_k, \boldsymbol{\theta})| < \varepsilon$

### B. Effect of Simultaneous Estimation

State of each part of each reconstructed volume can be divided into one of three cases as shown in Fig. 4(right).

- 1) Complete reconstruction and the accuracy of fitting is good
- 2) Incomplete reconstruction and the accuracy of fitting is good
- 3) Incomplete reconstruction result and the accuracy of fitting is bad

If we see the index finger in Fig. 4(right), the index finger model in the case of 1 is greatly affected by the reconstructed result which is close to it. On the contrary, in the case of 2 and 3, no reconstructed volume exist near the index finger model so the influence of these two cases are reduced compared to the first case by the effect of M-estimator. So, the influence to the index finger from incomplete volumes is suppressed and the accuracy of fitting is improved.

The above three cases are the extreme cases and the actual influence is automatically estimated based on the result of nearest neighbor search at each model point.

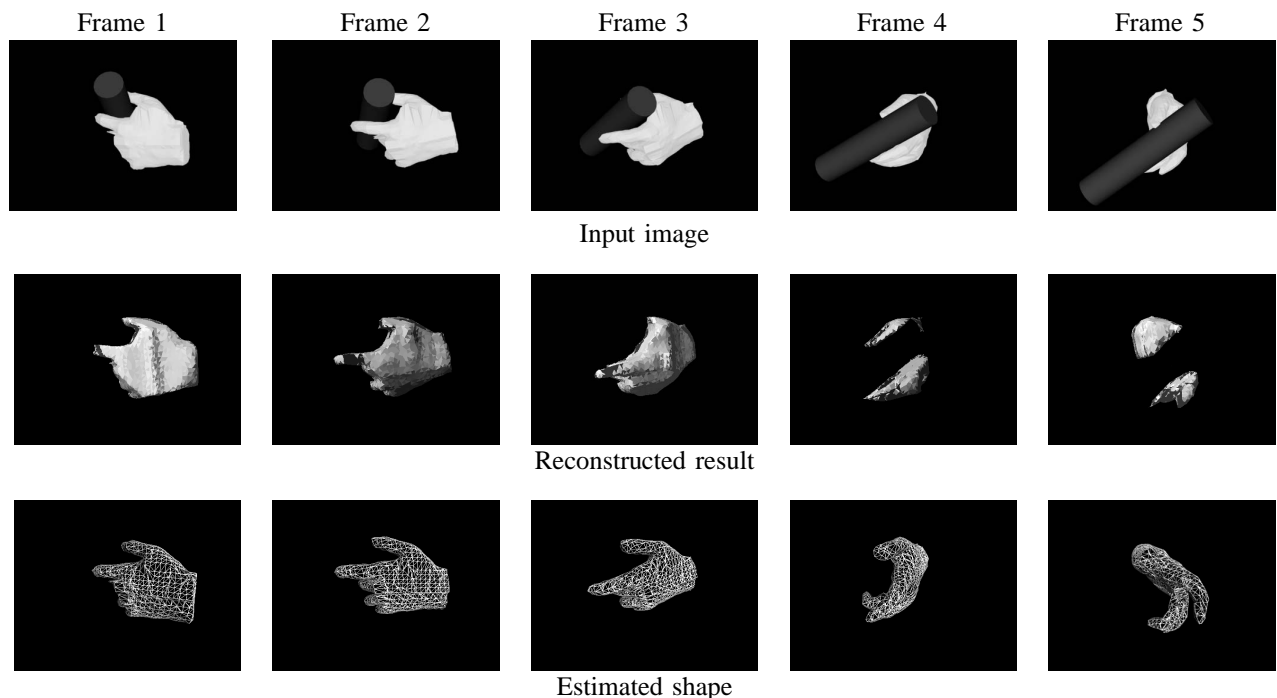


Fig. 6. Result in Simulation

TABLE I  
RESULTS FROM SIMULATION

|         | Average error between surface patches [mm] | Average error between joint angles [deg] |
|---------|--|--|
| Grasp 1 | 8.3  | 13.5                                     |
| Grasp 2 | 10.5                                       | 15.8                                     |
| Grasp 3 | 10.9                                       | 18.2                                     |

## VI. EXPERIMENT

### A. Simulation

As quantitative analysis, the proposed method is applied to data created in a virtual environment. In this case, the true value is known and the error can be evaluated.

Fig. 5 shows the case where the hand model is fitted to the hand volume made from the same hand model. It shows the effectiveness of the extended 3DTM algorithm with an articulated object.

Table I shows the recognition error from three different types of grasp typically appeared in manipulation tasks.

Fig. 6 shows the result of an experiment. With regard to the index finger, frame 1, 2, 3 correspond to the first case in section V-B, frame 4 corresponds to the second case and Frame 5 corresponds to the third case respectively. In the simultaneous estimation process, the influence of the result from Frame 4 and Frame 5 is reduced and the total accuracy keeps higher.

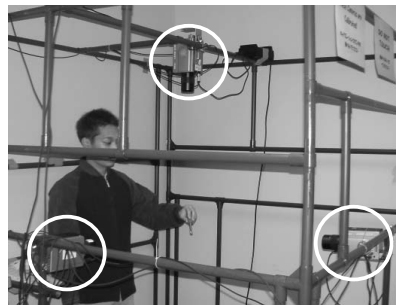


Fig. 7. Experimental Setup

### B. Real environment

As in the virtual environment, an experimental setup whose length is 1500[mm], width is 1500[mm] and height is 2000[mm] is constructed and 3 infrared cameras, Nikon Thermal Vision LAIRD-S270, are mounted (Fig. 7).

Experimental result is shown in Fig. 8. The estimation result seems relatively good, though the accuracy of thumb is lower compared to the other fingers. This is caused by the complex kinematic structure of thumb and some heuristics should be introduced.

## VII. CONCLUSION

In this paper, a robust 3D shape recognition technique of an articulated object is proposed. The main contribution of the proposed technique is that it can recover 3D shape

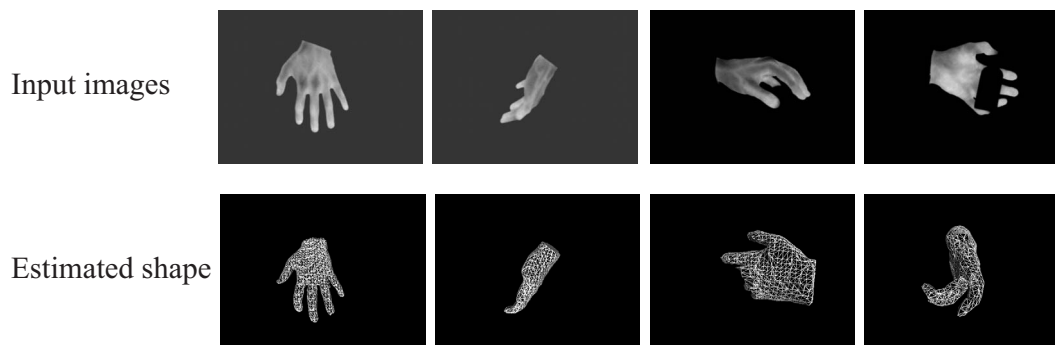


Fig. 8. Results in Real Environment

of a hand, joint angles and pose, even if it is partially occluded by an object. The features of our technique can be summarized as follows. (1) It is a vision-based approach, so the motion of the user is not constrained by any of attached sensors. (2) It estimates large number of D.O.F. of an articulated object model simultaneously by using a method from robust statistics which suppresses the effect of noise and small occlusion. (3) To deal with large occlusion, time series hand motion are utilized and badly reconstructed hand volume are automatically rejected.

The proposed technique is verified in a virtual world as well as in a real world, and the shape of a hand which grasps various objects are robustly recognized.

Although infrared cameras are used to detect silhouette of a hand easily, our technique can be applied to normal color cameras together with a skin detection algorithm.

Future work includes adopting trajectory information of the hand motion to constrain the possible pose of the hand and improve the accuracy of estimation.

### VIII. ACKNOWLEDGMENTS

This work is supported in part by the Japan Science and Technology Corporation (JST) under the Ikeuchi CREST project, and in part by the Grant-in-Aid for Scientific Research on Priority Areas (C) 14019027 of the Ministry of Education, Culture, Sports, Science and Technology.

### IX. REFERENCES

- [1] K. Ogawara, S. Iba, T. Tanuki, H. Kimura, and K. Ikeuchi. Acquiring hand-action models by attention point analysis. In *Int. Conference on Robotics and Automation*, volume 4, pages 465–470, 2001.
- [2] Y. Wu and T. Huang. Human hand modeling, analysis and animation in the context of hci, 1999.
- [3] James M. Rehg and Takeo Kanade. Model-based tracking of self-occluding articulated objects. In *ICCV*, pages 612–617, 1995.
- [4] Q. Delamarre and O. Faugeras. Finding pose of hand in video images: a stereo-based approach, 1998.
- [5] E. Ueda, Y. Matsumoto, M. Imai, and T. Ogasawara. Hand pose estimation for vision-based human interface. In *10th IEEE Int. Workshop on Robot and Human Communication (ROMAN) 2001*, pages 473–478, 2001.
- [6] P. J. Besl and N. D. Mckay. A method for registration of 3-D shapes. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 14(2):239–256, Feb. 1992.
- [7] M. D. Wheeler and K. Ikeuchi. Sensor modeling, probabilistic hypothesis generation, and robust localization for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(3):252–265, 1995.
- [8] K. Ogawara, J. Takamatsu, H. Kimura, and K. Ikeuchi. Extraction of fine motion through multiple observations of human demonstration by dp matching and combined template matching. In *10th IEEE Int. Workshop on Robot and Human Communication (ROMAN) 2001*, pages 8–13, 2001.
- [9] Michael Potmesil. Generating octree models of 3D objects from their silhouettes in a sequence of images. *j-CVGIP*, 40(1):1–29, October 1987.
- [10] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *ACM SIGGRAPH*, pages 163–169, 1987.
- [11] Hugues Hoppe. Progressive meshes. *Computer Graphics*, 30(Annual Conference Series):99–108, 1996.