Real-time 3-D Hand Posture Estimation based on 2-D Appearance Retrieval Using Monocular Camera

Nobutaka Shimada, Kousuke Kimura and Yoshiaki Shirai
Dept. of Computer-Controlled Mechanical Systems, Osaka University,
2-1 Yamadaoka, Suita, Osaka 565-0871, Japan.
E-mail: shimada@mech.eng.osaka-u.ac.jp

Abstract

This paper proposes a system estimating arbitrary 3-D human hand postures in real-time. Differently from most of the previous real-time system, it can accept not only predetermined hand signs but also arbitrary postures and it works in a monocular camera environment. The estimation is based on a 2-D image retrieval. More than 16000 possible hand appearances (silhouette contours) are first generated from a given 3-D shape model by rotating model joints and stored in an appearance database. Every appearance is tagged with its own joint angles which are used when the appearance was generated. By retrieving the appearance in the database well-matching to the input image contour, the joint angles of the input shape can be rapidly obtained. For robust matching, little difference between the appearances and the input is adjusted before matching considering shape deformation and quantization error in appearence samping.

In order to achieve the real-time processing, the search area is reduced by using an adjacency map in the database. In the map, adjacent appearances having similar joint angles are connected with each other. For each appearance, a neighborhood which consists of the adjacent appearances is defined. In each frame, the search area is limited to the neighborhood of the estimated appearance in the previous frame.

The best candidate at one frame is sometimes wrong due to approximation errors, too rapid motions or ambiguities caused by self-occlusion. To prevent tracking failures, a fixed number of the well-matching appearances are saved at every frame. After the multiple neighborhoods of the saved appearances are merged, the unified neighborhood is searched for the estimate efficiently by Beam search.

These algorithms are implemented on a PC cluster system consisting of 6 PCs (Pentium III 600MHz) and real-time processing is achieved. The resulted posture estimates are shown by experimental examples.

1 Introduction

Recently vision-based human interfaces have attracted increasing attention as an alternative way to traditional input devices like mouses and keyboards. Such attempts previously proposed can be divided into two categories: 3-D model-based and 2-D appearance-based approaches. Methods in the first category extract local image features and fit a given 3-D shape model to the features [1][2]. While the methods are able to estimate the object postures accurately based on the least squares criterion, segmentation and correspondence of the feature is not robust for arbitrary hand postures. Even from a simple background, the feature segmentation (finger tips, joint positions, finger axes and a wrist position etc.) often fails due to a great variety of hand appearances and self-occlusion. Methods in the second category register the possible 2-D appearances of the target object and then find the best-matching one to the input image [3][4]. They are robust to self-occlusion since they extract no features and directly compare the intensity property between the input images and the registered ones. In addition, required processing time is short since the images are dimensionally compressed by Principal Component Analysis (PCA). However, they only categorize the inputs into several patterns like the hand signs with no extraction of the 3-D information. Black et al. [5] extended this approach to estimate 2-D position and orientation but not 3-D.

The idea of "Estimation by Synthesis (ES)" [6][7][8][9] is the first bridge connecting the 3-D model-based and the 2-D appearance-based methods. The ES methods generate the possible appearances from a given 3-D shape model. They can estimate the 3-D postures because the 3-D parameters of the generated appearances are known. However, it takes too much computation to process in real-time because they calculate the overlapping ratio of the silhouettes as the matching degree and the search space for arbitrary hand postures becomes huge due to the high degrees of freedom (DOF) of the hand.

2 Overview of our system

We propose an extended ES method by combining 2-D appearance matching and 3-D model-based fitting. Fig.1 illustrates the entire framework of our system.

First the method obtains a rough posture estimate based on the 2-D appearance retrieval. The method generates possible hand appearances (silhouette contours) from a given

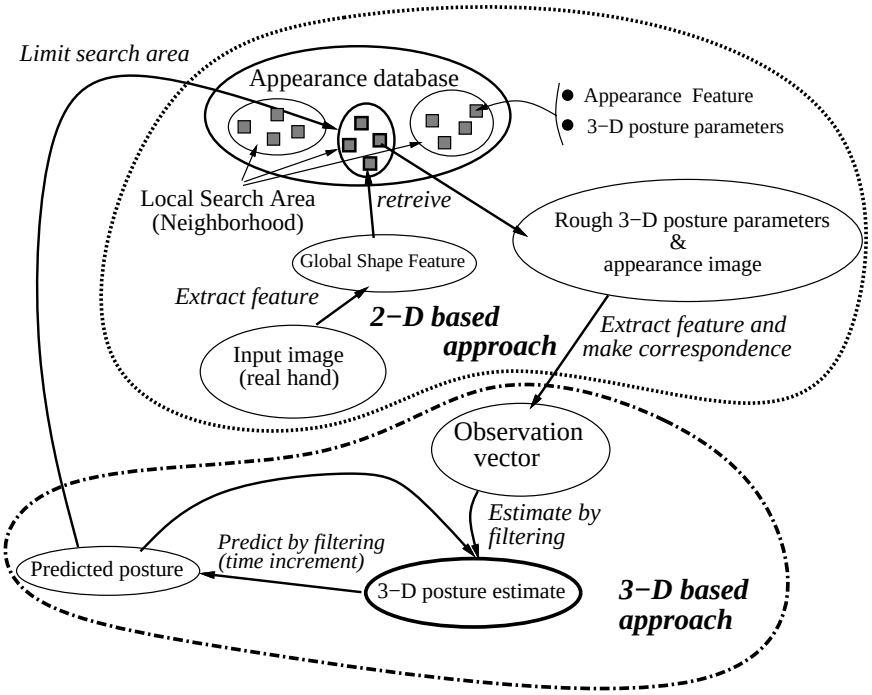


Figure 1. Entire Framework

3-D shape model by rotating model joints and stores them in an appearance database.* Every appearance is tagged with its own 3-D parameters (joint angles) which are used when the appearance was generated.

In the estimation phase, the joint angles of the input shape are rapidly obtained by retrieving the appearance in the database well-matching to the input image contour. In order to achieve the real-time processing, the search area is reduced by using an adjacency map in the database. In each frame, the search area is limited to the neighborhood of the estimated appearance in the previous frame. Additionally, to prevent tracking failure, a fixed number of the well-matching appearances are saved at every frame. The neighborhoods of the multiple appearances are merged and searched for the estimate efficiently by Beam search[9].

Based on the obtained rough estimate, the image features are correctly segmented and made correspondences with every part in the 3-D model even in self-occlusion. The method estimates the precise hand posture and moreover can refine the given initial 3-D model by a model fitting method [10][11] during observing the image sequence.

By repeating both the 2-D appearance matching and the 3-D model-based fitting approaches in every frame, robust and rapid posture estimation is established. In this paper, the implementation of the first part of the framework is discussed. The system implemented on a PC cluster system achieves the real-time processing. The following sections describe the details of the real-time system with experimental results.

*It depends on the resolution in quantization of the joint angles whether appropriate shape variations are contained.

3 Feature Extraction and Matching

3.1 Global Shape Feature

In the retrieval of the appearance, the method compares global shape features in the hand contour image. For simplicity, the hand region is assumed to be brighter than the background and the clothes so that the hand contour is easily obtained.

For reduction of the amount of the database, the global shape features invariant to the position and scale are computed from a hand contour. As shown in Fig.2, let $\{P_i\}(i = 0, \cdots, N-1)$ be N points (N is fixed to 256 in the following experiments) which are placed at a regular interval on the contour in clockwise order and r_i be the distance between P_i and the center of gravity G. The scale-normalized distance $\rho(i)$ is obtained by $\rho(i) = \frac{r_i}{\sqrt{A}}$ where A is the area of the hand region. The shape feature is defined as the list of normalized distance $\{\rho(i)\}(i = 0, \cdots, N-1)$. This feature is calculated in advance for every appearance in the appearance database. For the input image, the feature is also calculated in estimation time.

3.2 Rotation/Deformation-Invariant Matching

While the above feature is normalized in terms of posi-

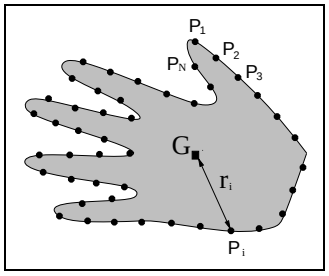


Figure 2. Shape feature of hand contour

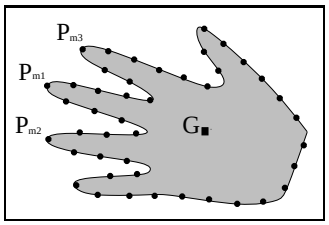


Figure 3. Normalization using remarkable shape feature

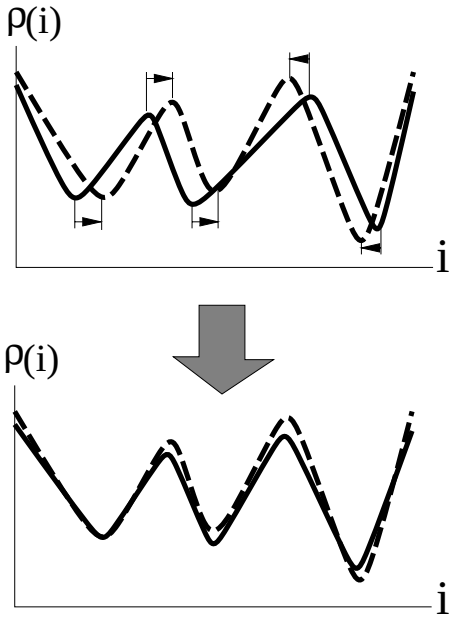


Figure 4. Adjustment of Deformation

tion and scale, it is sensitive to rotation in a image plane. Rotation changes the order of the elements in the feature list. Moreover, the deformation of the contour should be considered in matching.

We use the positions of remarkable points in the global shape feature for alignment of the starting point. First, the positions $\{P_m\}$ of the local maxima/minima of $\{\rho(i)\}$ are extracted from the input feature. For the models in the database, $\{P_m\}$ are extracted in advance. In the matching, if the number of $\{P_m\}$ of a model is more than 2 different from the input one, such a model is immediately rejected for rapid matching. For the remaining models, the most remarkable P_m is selected as the starting position P_0. For the input images, no more than J largest $P_m^j (j = 1, \cdots, J)$ from $\{P_m\}$ are considered as the candidates of the starting point when there are several positions have similar $\rho(i)$ to the maximum/minimum value (Fig.3). By matching the model's P_0 with each P_m^j of the input, $\{\rho(i)\}$ of the model is aligned with the input one. Then the model's $\{\rho(i)\}$ is adjusted to the input by deforming it so that all the local maxima/minima positions coincide between the model and the input. The deformation is achieved by translating the local maxima/minima points and linear interpolation. As the result of the selection of the starting points and the deformation adjustment, J normalized features $\hat{\rho}_j(i)$ are obtained. Finally, the minimum difference

$$e_k = \min_j \sqrt{\sum_{i=0}^{N-1} (\rho_k(i) - \hat{\rho}_j(i))^2} \qquad (1)$$

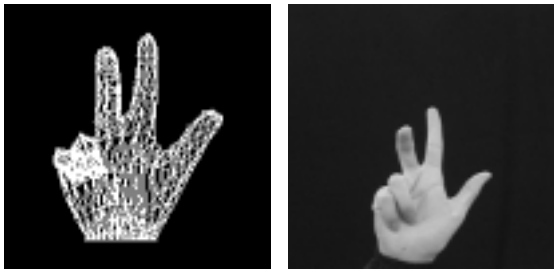
is evaluated as the distance between the input and the kth model features. Posture estimation is achieved by searching for the model minimizing Eq.1. Fig.5 shows a matching result example.

4 Rapid and Robust Search Method

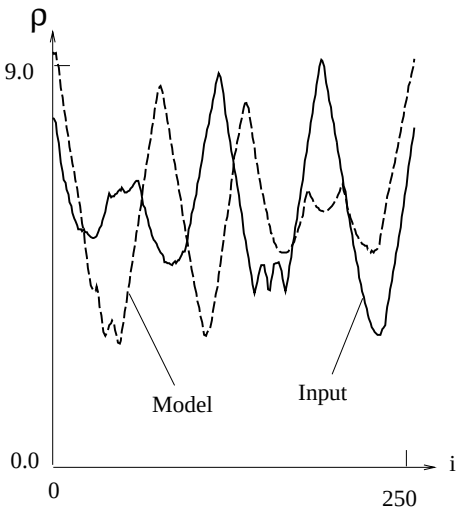
4.1 Reduction of Search Space

In the current system, the appearance database contains more than 16000 possible hand appearances by rotating the 3-D model joints and changing the viewpoint along a spherical surface. Each joint angle is sampled from its full range of rotation but the joints of same finger are assumed to have identical angles. The 128 viewpoints are sampled uniformly from the sphere.

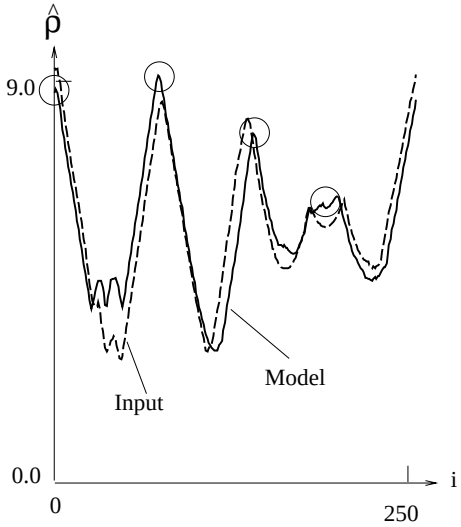
In the estimation time, the model features in the appearance database are matched to the input features one after another. Because there are a number of appearances in the database, it is impossible to search the whole candidates in real-time. In order to limit the search region, we use an adjacency map in the appearance database. In the map, adjacent appearances having similar 3-D posture parameters (i.e. joint angles and viewpoint) are connected with each other. Based on the map, we define a neighborhood of each model as a set of candidates directly connected with the model. Supposing the continuous change of the hand postures, we can limit the next search area to the neighborhood of the posture estimate in the previous frame (see Fig.6).



(a) Input image (b) Matched model



(c) Original Feature Vectors



(d) Adjusted Feature Vectors

Figure 5. Normalized feature vector

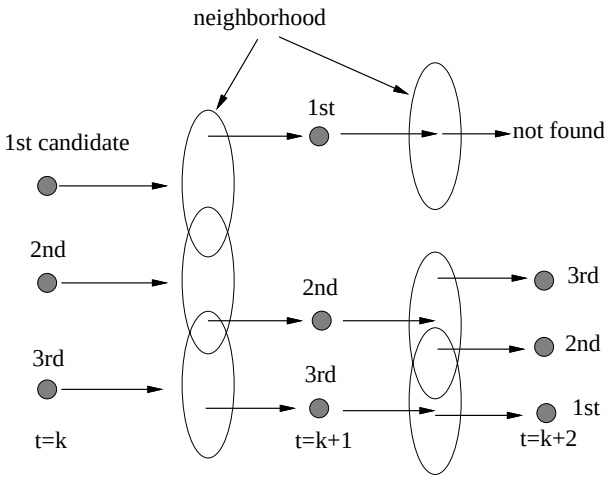


Figure 6. Limitation of Search Area and Beam-Search

If compressing the dimension of each neighborhood using PCA (Principal Component Analysis), the storage amount of the model shape features and the amount of matching computation can be reduced [12][13]. This has not been implemented in the current real-time system because the dimension of the feature is small enough (N=256). If higher order features like image intensities are used, the feature compression is required for real-time processing.

4.2 Beam Search Using Multiple Processors

By the above method, we can find the appearance well-matching to an input image if it is successfully tracked over the time sequence. However, the best candidate at one frame is sometimes wrong due to approximation errors of the 3-D shape model, too rapid motion or ambiguities caused by self-occlusion. In best-first search, tracking failure often happens when the best candidate is actually wrong. Once tracking failure occurs, back tracking is necessary to recover but is causes serious deterioration in response.

To resolve this problem, we utilize beam search (Fig. 6). A fixed number K of the best-matching appearances are saved at every frame. The neighborhoods of the saved multiple candidates are merged and searched for K best-matching ones in the next. Tracking can be successfully done without any backtracking as long as there is a correct appearance in the K estimates. Since this search algorithm can be parallel processed, real-time system is built with PC clusters.

5 Implemented Architecture

This section describes the implementation of the above

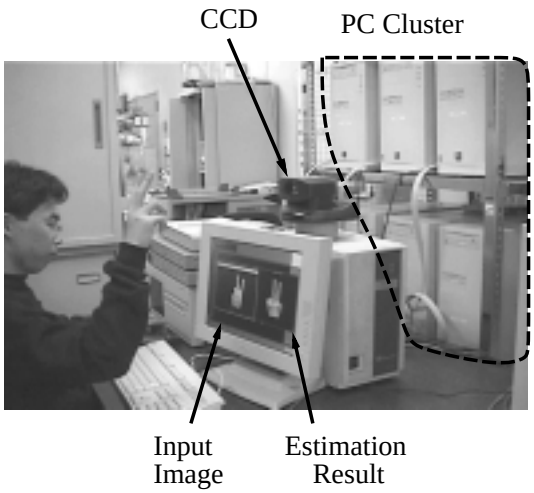


Figure 7. Implemented System

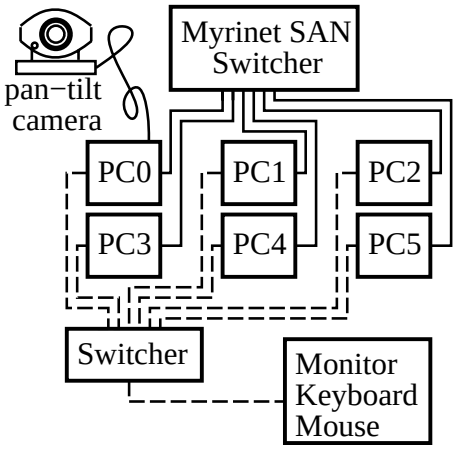


Figure 8. Configuration of PC Cluster

methods. We employ a pan-tilt CCD camera and a cluster of PC with a single 600MHz PentiumIII CPU, 256 M-Bytes memories and a Myrinet network interface (about 1 G-bit/sec transfer) (Fig. 7). Currently our cluster contains 6 PCs (Fig. 8). PC0 is a master controller of the cluster with a image capture device. Its roles are activation / shutdown of the system, image capture and extraction of the contour feature, broadcasting it to the other PCs and display of the finally estimated posture by 3-D computer graphics (see the time-chart in Fig. 9). PC1 determines the search region in every frame from the estimates in the previous frame, divides the region into 5 pieces and dispatches the search in each region to itself and PC2-5. Each of PC1-5 has the whole appearance database on memory and searches the assigned search region for the appearances well-matching to the input feature in parallel. The found appearances are collected to PC1. PC1 selects the K best appearances (K=5 in the following experiments) and send it to PC0. The series of this cycle is optimally processed like a pipe-line structure. Since PC0's idling time is large, display of the $k-1$th estimate and the $k+1$th image capture is done during the kth idling period (Fig.10).

Currently, processing time for one frame is slightly more than 33 m-sec. Latency from image capture to display the result is 1 frame. The major time-consuming process is the parallel matching (20 m-sec on average). Because the actual cycle time is quantized by the NTSC sync timing, the system skips a frame during processing approximitely 15 frames. Additional one or two more PCs can help realize the video-rate processing.

6 Experimental Results

Fig.11 shows examples of the estimated 3-D postures. (a)-(c) are input images and (d)-(f) are the estimates respectively.

Fig.12 illustrates the effect of beam search for robust tracking. Each column shows the input image (top of the column) and the estimated postures displayed by wire-frame (upper posture is the better estimate). Time (frame number) goes from left to right. At the first and second frame, the best estimation at that moment has by no succeeding estimates. The best at the second frame has only one stretched finger but actually one more stretched finger is occluded. Because such a correct estimate has been discovered at the third frame, best-first search fails the tracking in such a case but beam search can tracks without backtracking. Of course, if you want, the system can correct the postures at the past frames by backtracking from the current frame in compensation for some latency of response. This function is useful for motion capturing applications. In Fig.12, "O"marks in the figure show the best estimates and "X"marks show the eliminated estimates obtained by backtracking from the 5th frame.

7 Conclusion

We have presented a robust and rapid method of estimating hand postures based on 2-D image retrieval. It is possible to obtain the 3-D posture from model hand images with known parameters in the appearance database. The largest contribution of this paper is the implementation of the real-time estimation system in a monocular environment. Temporal ambiguity due to self-occlusion is resolved and the robust tracking is done by beam search implemented by parallel processing with a PC cluster.

One of the remaining problems is the difficulty in distinguishing shapes like a fist. In order to resolve it, we consider to use image intensities. It is also necessary to extract hand

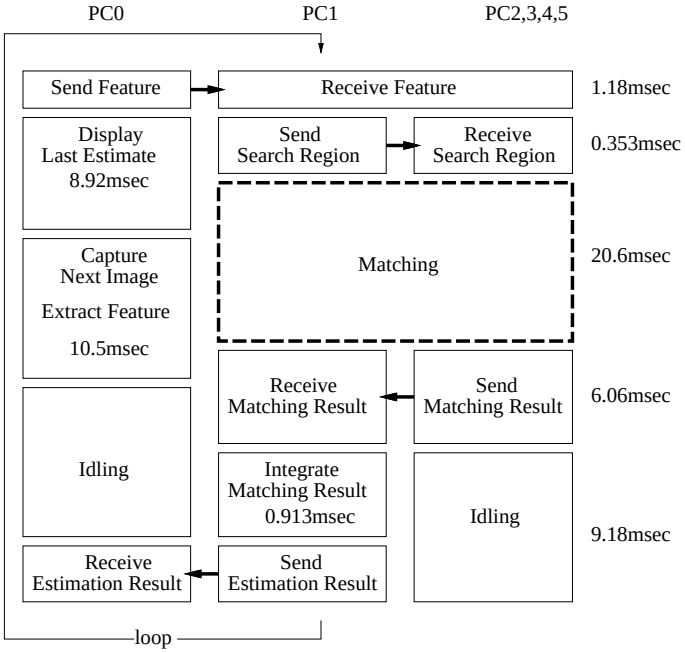


Figure 9. Processing Time-chart

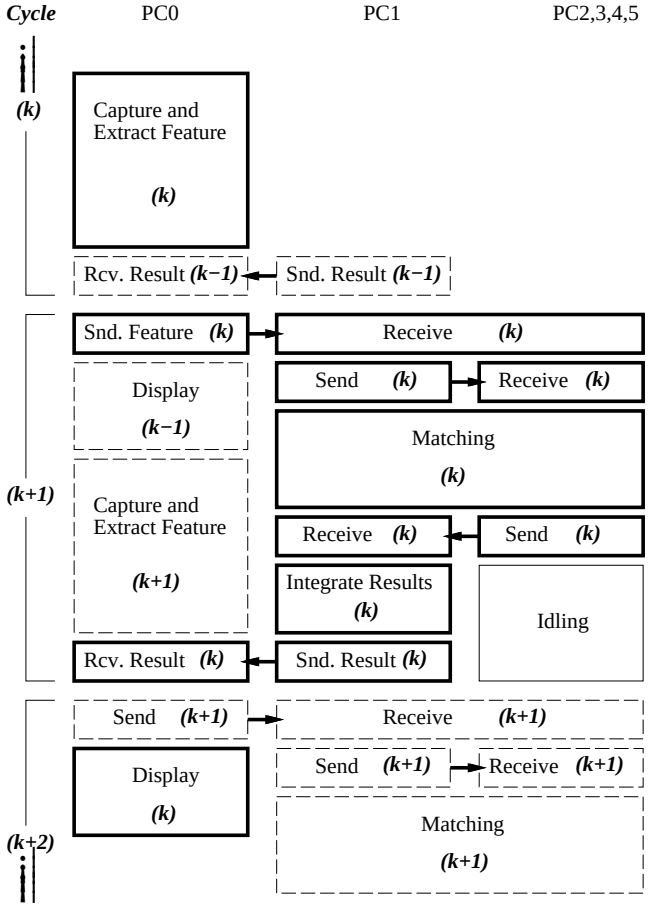


Figure 10. Pipeline Processing

regions from any backgrounds. Currently we are trying to extract hand region using a infrared illumination and a CCD with an infrared filter.

References

[1] 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.

[2] D. Lowe. "Fitting Parameterized Three Dimentional Models to Images". *IEEE Trans., Pattern Anal. Machine Intell.,vol.13,No.5*, pp. 441–450, 1991.

[3] B. Moghaddam and A. Pentland. "Maximum Likelihood Detection of Faces and Hands". *Proc.of Int.Workshop on Automatic Face and Gesture Recognition*, pp. 122–128, 1995.

[4] U. Brockl-Fox. "Realtime 3-D Interaction with up to 16 Degrees of Freedom from Monocular Video Image Flows". *Proc.of Int.Workshop on Automatic Face and Gesture Recognition*, pp. 172–178, 1995.

[5] M. J Black and A. D. Jepson. "EigenTracking: Robust Matching and Tracking of Articlated Objects Using a View-Based Representation". *Int.J.of Computer Vision 26(1)*, pp. 63–84, 1998.

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

[7] J. J. Kuch and T. S. Huang. "Virtual Gun: A Vision Based Human Computer Interface Using the Human Hand". In *MVA'94*, pp. 196–199, 1994.

[8] 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.

[9] N. Shimada, Y. Shirai, and Y. Kuno. "Hand Gestrue Recognition Using Computer Vision Based on Model-matching Method". In *Proc.of 6th Int. Conf. on HCI*, pp. 11–16. Elsevier, 1995.

[10] N. Shimada, Y. Shirai, Y. Kuno, and J. Miura. "Hand Gesture Estimation and Model Refinement using Monocular Camera". In *Proc. of 3rd Int. Conf. on Automatic Face and Gesture Recognition*, pp. 268–273, 1998.

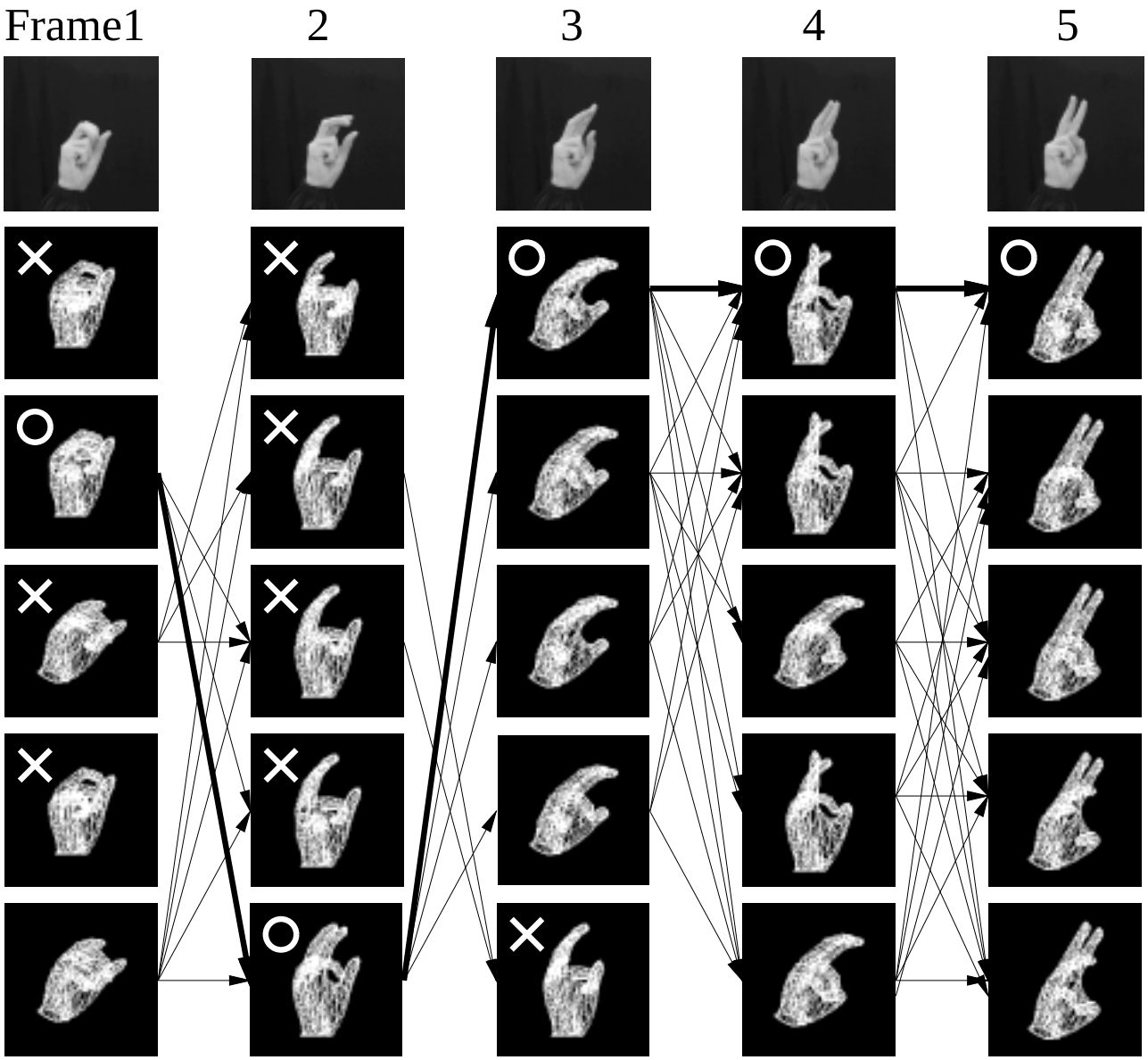
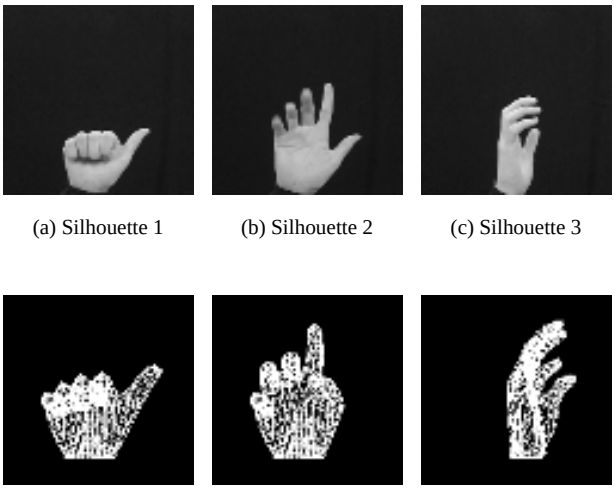


Figure 12. Robust Tracking Result by Beam-Search



(a) Silhouette 1 (b) Silhouette 2 (c) Silhouette 3

(d) Estimate of (a) (e) Estimate of (b) (f) Estimate of (c)

Figure 11. Retrieved 3-D Hand Postures

[11] N. Shimada, Y. Shirai, and Y. Kuno. ”Model Adaptation and Posture Estimation of Moving Articulated Object Using Monocular Camera”. In *Proc. 1st Int’l Workshop on Articulated Motion and Deformable Object (LNCS 1899)*, pp. 159–172. Springer, 2000.

[12] T. Heap and D. Hogg. “Wormholes in Shape Space:Tracking through Discontinuous Changes in Shape”. *Proc. of 6th Int. Conf on Computer Vision*, pp. 344–349, 1998.

[13] N. Shimada, K. Kimura, Y. Shirai, and Y. Kuno. ”Hand Posture Estimation by Combining 2-D Appearance-based and 3-D Model-based Approaches”. In *Proc. of 15th Int’l Conf. on Patter Recognition, vol.3*, pp. 709–712, 2000.

Acknowledgment

This work is supported in part by Grant-in-Aid for Scientific Research from Ministry of Education, Science, Sports, and Culture, Japanese Government, No.11555072 and 09555080.

The 3-D model of the real human hand was provided by courtesy of Prof. Kishino and Prof. Kitamura, Dept. of Electronic, Information Systems and Energy Engineering, Osaka University.