Hand Posture Estimation by Combining 2-D Appearance-based and 3-D Model-based Approaches *

Nobutaka Shimada, Kousuke Kimura, Yoshiaki Shirai and Yoshinori Kuno
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 method for the rapid and precise estimation of human hand postures by combining 2-D appearance matching and 3-D model-based fitting. First a rough posture estimate is obtained by image indexing. Each possible hand appearance generated from a given 3-D shape model is labeled by an index obtained by PCA compression and registered with its 3-D model parameters in advance. By retrieving the index of the input image, the method can obtain the matched appearance image and its 3-D parameters rapidly. Then, starting from the obtained rough estimate, it estimates the posture and moreover refines the given initial 3-D model by model-fitting. The sequential activation of the two processes in every frame gives the precise posture estimate rapidly. The effectiveness of the method is shown by experimental results.

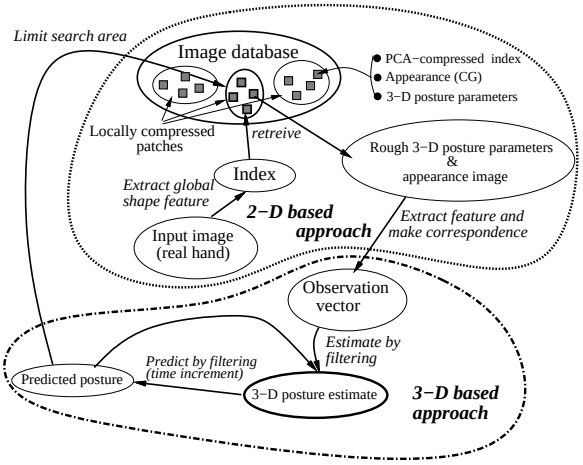


Figure 1. Flowchart

1 Introduction

Recently vision-based human interfaces have attracted increasing attentions 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, failures of segmentation and feature correspondence often occur 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-matched one to the input image [3][4]. They are robust to self-occlusion since they extract no features and directly match the intensity property between the input and the registered images. Required processing time is short since the matched 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 is huge due to the high degrees of freedom (DOF) of the hand. In addition, a precise shape model is required for precise posture estimation because individual users' hands actually have different shapes. We need model adaptation to individual hands.

2 Overview of our method

We propose an extended ES method by combining 2-D appearance matching and 3-D model-based fitting (see Fig.1). First the method obtains a rough posture estimate based on the image indexing. The method generates all possible hand appearances from a given 3-D shape model in advance. Then it generates an index from each appearance by compressing the appearance with PCA and registers the set of the index, the appearance, and its 3-D model parameters. In the estimation phase, the method can retrieve the matched appearance image and its 3-D parameters rapidly by using the index generated from the input image as a search key. Based on the obtained rough estimate, the method can extract the fingertips, finger axes, palm contours and a wrist position. Thus it can correctly find their corresponding parts in the 3-D model even in the case of self-occlusion. Then the method estimates the precise hand posture and moreover refines the given initial 3-D model by a model-fitting method like modified Kalman filter with "distribution truncation" [10] during observing the image sequence. The refined posture estimate conversely helps the 3-D parameter retrieval in the next time-step. Since it is expected that the posture in the next time-step is found near the posture predicted by filtering, the search area can be limited. Since our method divides the registered hand appearances with similar 3-D parameters into subgroups and each subgroup is PCA-compressed respectively, the rough estimate for the

*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.

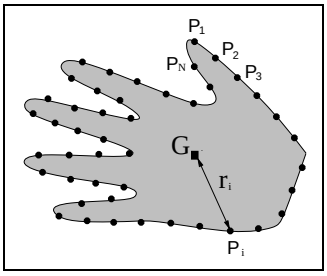
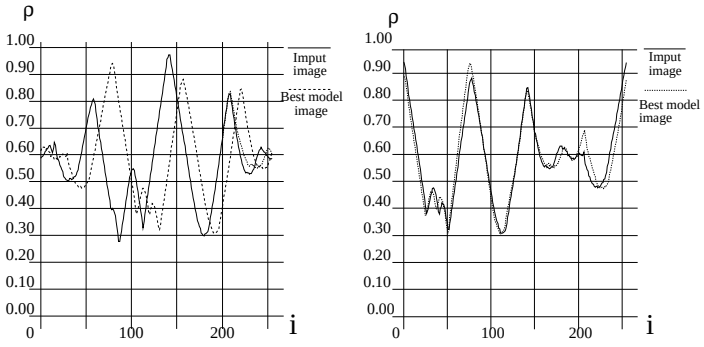


Figure 2. Shape feature of hand contour



(a) By principal axis (b) By remarkable features

Figure 3. Rotation-normalized feature vector

input image at the next time-step can be obtained by searching only the neighbouring subgroup of the posture predicted by filtering. In this framework combining 2-D appearance matching and 3-D model-based fitting approaches, robust and rapid posture estimation is established.

The following sections describe the details of the posture estimation with experimental results.

3 Rapid Estimation of Rough Posture

3.1 Global Shape Feature

In the retrieval of the rough estimate, the method compares global shape features extracted from the hand contour in an input image in stead of direct matching of the image itself. For simplicity, the hand region is assumed to be brighter than the background and the clothes so that the hand region is easily obtained.

For reduction of the number of the models, the shape features invariant to the position, scale and rotation are computed from the contour of the hand region. As shown in figure 2, let $P_i(i = 0, \cdots, N-1)$ be N points which are placed at a regular interval on the contour 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 $\{\rho(i)\}(i = 0, \cdots, N-1)$.

3.2 Rotation-Invariant Matching

The use of the above feature normalizes the position and scale. However, the feature is sensitive to rotation. Rotation changes the order of the elements in the feature list. Thus we need to devise a method of determining a particular point on the contour as the first element in the list. We first tried to use the principal axis of the silhouette. However, the result is not satisfactory as shown in Fig.3(a) because the principal axis is sensitive to the slight shape changes of parts far from the center of gravity.

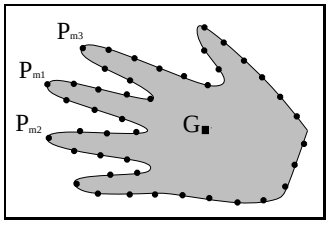


Figure 4. Normalization using remarkable shape feature

Therefore we consider to use the position of remarkable shape in the hand contour. Here the position P_m giving the maximum $\rho(i)$ is used as the starting position P_0 for the model images. If there are several positions giving $\rho(i)$ similar to the maximum value, no more than three largest local maximum positions $P_m^j(j = 1, 2, 3)$, are considered for the input images as shown in Fig.4. By fixing P_0 at each P_m, three normalized features $\hat{\rho}_j(i)(j = 1, 2, 3)$ are obtained. The minimum difference

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

is adopted as the difference between the input and the kth model features. Based on this criterion, we tried a matching experiment using 192 model images and 70 input images not included in the model images. Here, N is fixed to 256. In this experiment, the matching was regarded as successful if there are right shapes decided by human in the best three models. The method successfully matched 95% of the input images. Fig.3(b) shows a matching result example.

3.3 Rapid Matching by Image Indexing Using PCA Compression

In the estimation time, the model shape features stored in advance are matched to the input shape features one after another. For rapid process nearly in real-time, the individual matching computation is required to be reduced as much as possible. For this reason, the extracted shape features are compressed by PCA.

The shape feature $\{\rho(i)\}$ can be also regarded as points on the high-dimensional space by definition of feature vector $\boldsymbol{x}$:

$$\boldsymbol{x} = [\rho(0), \cdots, \rho(N-1)]^T. \quad (2)$$

Performing PCA to a feature vector set of the model images, the eigen vectors $\{\boldsymbol{e}_i\}(i = 1 \cdots M)$ corresponding to the largest M eigen values are obtained ($M < N$). The compressed feature vector $\boldsymbol{y}$ is obtained as follows:

$$\boldsymbol{y} = [\boldsymbol{e}_1 \cdots \boldsymbol{e}_M]^T (\boldsymbol{x} - \bar{\boldsymbol{x}}) \quad (3)$$

where $\bar{\boldsymbol{x}}$ denotes the mean of $\boldsymbol{x}$. Since the input images are supposed to be similar enough to one of the model images, the difference e_k between the feature vectors is approximated by the difference e_{ek} between the compressed feature vectors. In the case that $N = 256$,$M = 30$ and the number of the model images is 200, the amount of compu-

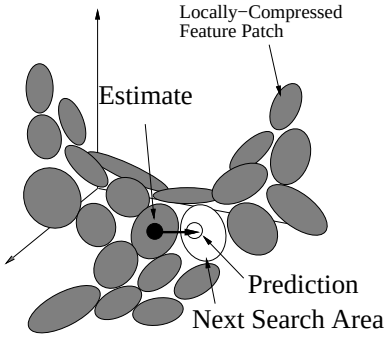


Figure 5. Locally-Compressed Feature Manifold

tation of the matching with the PCA compression is reduced to less than 30% of that without compression.

3.4 Fast Search with Locally-Compressed Feature Manifold

In spite of the above data compression, the number of comparison can become huge due to the great variety of the hand shape. Heap et al. [11] divided the image space into a set of local patches generated by PCA-compression in each local neibourhood. This patch set constructs a manifold in the feature space to be searched for the model matching to the input image. Since the model images in a patch are similar each other, this idea improves the data compression rate (i.e. decrease the number of dimensionality). Additionally it also helps with limiting the search area. After a refined 3-D posture estimate is obtained (described in Sec.4.2), we can limit the next search area to the patches adjacent to the estimate. (see Fig.5). Note that this adjacency degree between patches is defined by the difference of the 3-D shape parameters, not that of the shape feature. Under the assumption of the motion smoothness of the human hand, the well-matched model image is found in a short time.

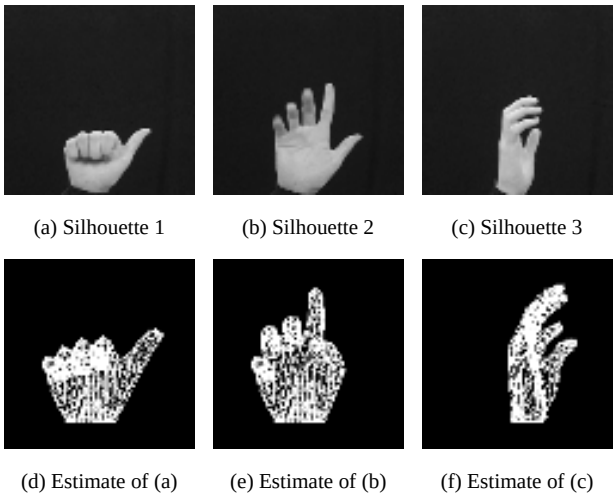
3.5 Experimental Result of Rough Estimation

Since we have a 3-D shape model of human hand, we can easily make a number of CGs as the model images whose 3-D posture parameters (palm orientation and angles of finger joints) are known. Using CGs as the model images, it is therefore possible to retrieve the 3-D posture by search for the best-matched model image by using the PCA-compressed shape feature $\boldsymbol{y}$ as the index. Fig.6 shows examples of the retrieved 3-D postures from the indexed CGs. In the current implementation on Sun SPARC Station 10, the posture can be estimated 10 frames per second including image capturing and the pre-processes.

4 Estimate Refinement by 3-D Model Fitting

4.1 Image Segmentation and Making Feature Correspondence

Based on the rough posture obtained by the above method, we next refine the posture estimate. For more precise estimation, we try to simultaneously adapt the initial 3-D model to the individual hand shape. Given the rough estimate, the feature extraction is easily resolved because it is already clear where each finger is located and which finger is occluding or occluded. Therefore, the refined 3-D parameters are obtained by fitting to the image features with least squares method. However it often fails when some



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

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

Figure 6. Retrieved 3-D Hand Postures using CG Model Images

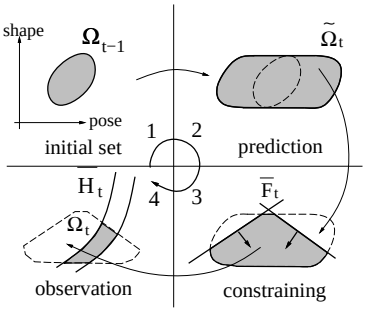
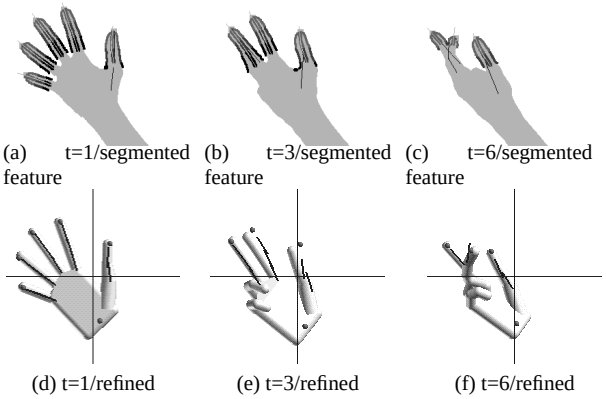


Figure 7. Incremental update of the solution set: $\bar{H}$: observations and $\bar{F}$: constraints

features are missing due to self-occlusion, especially only monocular image is available. The reason of the failures is that the depth cannot be uniquely determined in such situations. The method more robust to this depth ambiguity problem was proposed by Shimada et al.[10] called "Distribution truncation". We briefly explains the method in the next section.

4.2 Ambiguity Limiting using Constraint Knowledge

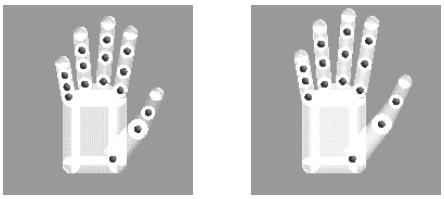
Because we can use the time sequence of images, we consider to apply a filtering method such as Kalman Filter for estimation of the parameter vector $\boldsymbol{x}$ including scale, wrist position, palm orientation, joint angles, lengths and widths of fingers and a palm. Additionally we consider more constraint knowledge of the hand. Our ambiguity-limiting process with the above constraints is shown in Fig.7. It is the same as a normal filtering method except in that the *constraining* phases is inserted between the prediction and observation phases. In the filtering framework, the ambiguity of the estimated parameter is represented as its probability distribution. When a *inequality constraint* such as $-20 \leq \theta \leq 40$ is available, the probability distribution can be *truncated* by removing the parts where the inequalities are not satisfied. Then the covariance ellipsoid



(a) t=1/segmented feature (b) t=3/segmented feature (c) t=6/segmented feature

(d) t=1/refined (e) t=3/refined (f) t=6/refined

Figure 8. Refined posture for real images



(a) Initial shape (b) Refined shape (t=9)

Figure 9. The result of adaptation to individuals for real image)

is incrementally limited, namely the depth ambiguity gets small, during the sequential observations. We use the following knowledge for the truncation.

(a) shape parameters (lengths and widths) are constant over the sequence.

(b) pose parameters (joint angles) change continuously.

(c) each parameter is within a certain range and has relations with the other parameters.

These constraints are represented as $|x_i - x_j| \leq \Delta x_{ij}$.

5 Experimental Results of Posture Refinement and Adaptation

We show an estimation result for the real hand images. In this experiment, only finger lengths are estimated as the shape. Fig.8 shows the refined pose estimates based on the rough estimates obtained by the method of Sec. 3. Before the refinement process, the image features are segmented into each part of fingers using the rough estimation result, and the observation (finger tips and axes) are calculated by line fitting to the segmented features. (Figs.8(a)-(c)). Then the pose estimates are refined ((d)-(f)). The refinement result of the shape model is shown in Fig.9.

6 Conclusion and Discussion

We have presented a method of estimating hand postures robustly and rapidly by combining 3-D model-based and 2-D appearance-based approaches. Rapid matching is realized based on image indexing using the shape feature invariant to the position, scale and rotation and PCA compression. It is possible to obtain the 3-D posture parameters using hand CG images with known parameters as the model images. Based on the obtained rough posture, the posture and the initial 3-D model is refined based on 3-D model fitting. We show the effectiveness of our method by simulation and an application to real hand images.

One of the remaining problems is the difficulty in distinguishing shapes like a fist. In order to resolve it, we consider to use intensities and multiple camera sources. It is also necessary to extract hand regions from any backgrounds.

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 – Ambiguity Limitation by Inequality Constraints". In *Proc. of 3rd Int. Conf. on Automatic Face and Gesture Recognition*, pp. 268–273, 1998.

[11] 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.

Acknowledgment

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