Hand Gesture Estimation and Model Refinement using Monocular Camera – Ambiguity Limitation by Inequality Constraints*

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

Abstract

This paper proposes a method to precisely estimate the pose (joint angles) of a moving human hand and also refine the 3-D shape (widths and lengths) of the given hand model from a monocular image sequence which contains no depth data.

First, given an initial rough shaped 3-D model, possible pose candidates are generated in a search space efficiently reduced using silhouette features and motion prediction. Then, selecting the candidates with high posterior probabilities, the rough poses are obtained and the feature correspondence is resolved even under quick motion and self-occlusion.

Next, In order to refine both the 3-D shape model and the rough pose under the depth ambiguity in monocular images, the paper proposes an ambiguity limitation method by loose constraint knowledge of the object represented as inequalities. The method calculates the probability distribution satisfying both the observation and the constraints. When multiple solutions are possible, they are preserved until a unique solution is determined. Experimental results show that the depth ambiguity is incrementally reduced if the informative observations are obtained.

1 Introduction

Hand is the most functional part of human body: pointing, handling, or expressing some symbols etc. In order to recognize automatically these variety of hand function, it is important to capture the precise hand gesture. Many methods precisely estimating human gestures have been developed in recent years. Some of them estimate by fitting a 3-D model to images based on feature correspondence [2][11][14]. Since the human gestures change quickly and cause self-occlusion in the images, it is difficult to resolve the correspondence in a real hand motion. Under those situations, *the estimation by synthesis* can determine the gestures by generating the possible pose candidates using the 3-D model [6][7][9]. The search space is however so huge due to a high DOF of a human body that it should be roughly quantized to reduce the computation cost. In addition, there are approximation errors in the initial shape of the model since it is difficult to prepare an exact shape model in advance. For these reasons, the estimated poses are quite rough.

The more precise estimate can be obtained by refining both the roughly estimated pose and the initial shape model. Given the rough estimate, the feature correspondence is so easily resolved that the model can be fitted to the image by least squares method. These methods however don't work well for the monocular images due to *depth ambiguity*. For example, in a case shown by Fig.1 where a stick object is projected as a line, there is the ambiguity between the length and the joint angle. Even by multiple cameras [11], the ambiguity still remains if a certain part is visible from only one camera.

Some methods try to limit the ambiguity using *constraint knowledge* of the objects in monocular setting. Effectiveness of the constraints are shown in Fig.1: supposing the joint angle is constrained within a certain range, then the length is limited between the maximum achieved at the candidate 'C' and the minimum at 'D'. In [1][10], the ambiguity is handled by intervals represented as the maxima and the minima of the parameters. In this manner, the ambiguity isn't sufficiently limited because the correlations between the parameters are not considered. (see the square region in Fig.2)

In our method, given an initial approximately shaped 3-D model, rough pose estimate (a position, a orientation and joint angles) is first obtained by estimation by synthesis. Next, both the roughly estimated pose and the shape (lengths and widths of the parts)

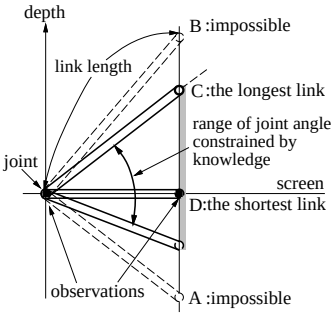


Figure.1 Depth ambiguity of a stick object

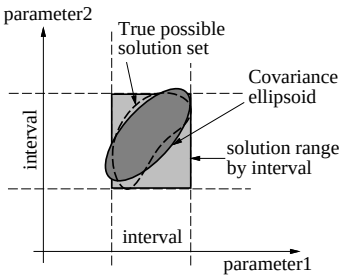


Figure.2 Interval description and possible solution set

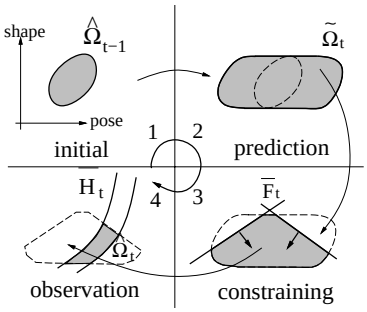


Figure.3 Incremental update of the solution set: $\bar{H}$: observations and $\bar{F}$: constraints

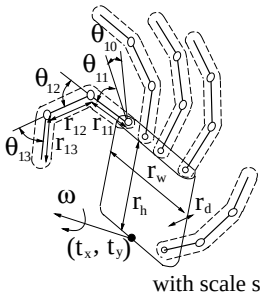


Figure.4 3-D hand model

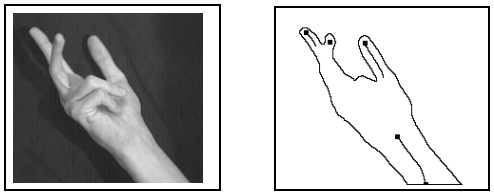
[1] This work is supported in part by Grant-in-Aid for Scientific Research from Ministry of Education, Science, Sports, and Culture, Japanese Government.

are simultaneously refined by representing the parameter ambiguity as a *covariance ellipsoid* of a probability distribution in the multi-dimensional parameter space (see Fig.2) and then limiting the ambiguity ellipsoid using the constraint knowledge.

The our ambiguity-limiting process is shown in Fig.3. It is same as normal filtering method except the *constraining* phase inserted between the prediction and observation phase. In order to deal with *inequality constraints* such as $-20 \leq \theta \leq 40$, a novel method is proposed which modifies the probability distribution by *truncating* the probability of a part where the inequalities are not satisfied. Then the ambiguity ellipsoid is incrementally limited, namely the shape and pose get precise, by various observations over the sequence. Since the correlations between the parameters are considered by the co-variance, the ambiguity is sufficiently limited (see the broken contour in Fig.2).

In the following sections, the rough estimation process is first explained and then the details of the refinement process and experimental results follow.

2 Rough Estimation by Silhouette Matching



(a) Input image (b) Extracted feature

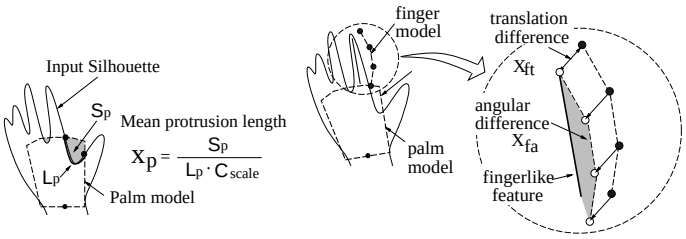
Figure 5. Feature extraction

Our method uses an object model in Fig.4 to estimate the shape and pose. Its shape and pose are represented as a wrist position, palm direction, 3-D shapes of the parts, joint angles and scale of the projection. In the rough estimation, the wrist position and the joint angles are estimated by fixed shape. We briefly explain that process here (refer [13] for details).

At each frame, the following preprocesses precede. As a wrist position, we extract a point in a silhouette image where the width of the arm abruptly changes. In the same way, silhouette features like fingers are extracted (Fig.5).

Next we search for the candidates well-matching to the silhouette. The matching of the hand pose consists of two phases: *generating* appearance candidates from the model and *evaluating* their degrees of matching to the silhouette. For reduction of the total number of generated candidates we utilize the following tactics:

1. *hierarchical estimation* from the palm to fingers
2. *adaptive quantization* of the parameter space considering the degree of image deformation for the variation of each parameter
3. limitation of the search space using *silhouette features*
4. search strategy considering the *prior probability* $P_b(\boldsymbol{\theta})$ based on motion smoothness ($\boldsymbol{\theta}$ denotes an appearance candidate).



(a) Palm candidate (b) Finger candidate

Figure 6. Evaluation of candidates

Because candidates with higher probabilities are earlier tested, a well-matching candidate can be found in a short time.

In evaluation phase, *a mean protrusion length* is evaluated for the generated palm pose as matching degree (Fig.6(a)). For the finger pose, the projected difference is evaluated between the axes of the model fingers and the pre-extracted silhouette features (Fig.6(b)). If any of the above matching degrees are larger than thresholds, such a candidate is rejected. For the rest, these different types of evaluation are integrated. We suppose the likelihood distribution of the matching degrees considering errors of the shape modeling and the quantization of the parameters. Then, the matching degrees of respective parts are integrated by Bayes rule:

$$P(\boldsymbol{\theta}_j|\boldsymbol{x}) = \frac{P_b(\boldsymbol{\theta}_j)\prod_n P(x_n|\boldsymbol{\theta}_j)}{\sum_j [P_b(\boldsymbol{\theta}_j)\prod_n P(x_n|\boldsymbol{\theta}_j)]} \qquad (1)$$

where $\boldsymbol{\theta}_j$ and $\boldsymbol{x} = \{x_n\}$ respectively denote the jth well-matching candidate and the set vector of the matching degrees.

Still, the best candidate at one frame is sometimes wrong when the later observations are considered. There are various causes: model approximation errors, too rapidly motion changes or ambiguities caused by occlusion. To resolve this problem, we utilize one more tactics:

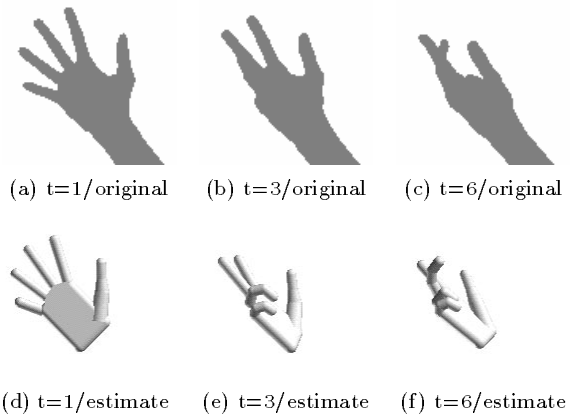
4. *preserving multiple estimates* at one frame by beam-search [8].

A fixed number of candidates are preserved at one frame. Even if the best estimate is actually wrong, the following estimation can be continued based on the rest estimates without back-tracking. If you wish, the globally optimal solutions are obtained over a long sequence with back-tracking. The rough estimation results for a certain sequence is shown in Fig.7.

3 Refinement of Shape and Pose Parameters

Next we refine both the shape and pose using results of the rough pose estimation. In order to resolve the depth ambiguity, we consider the following constraint knowledge of the shape and pose of an articulated object.

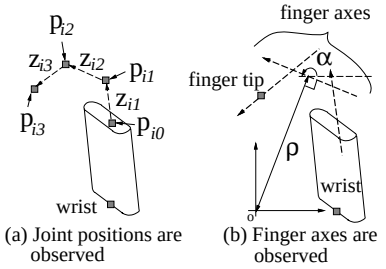
(a) shape parameters (lengths and widths) are constant over the sequence.



(a) t=1/original (b) t=3/original (c) t=6/original

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

Figure 7. Rough estimation results



(a) Joint positions are observed (b) Finger axes are observed

Figure.8 Observed features

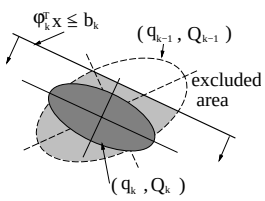


Figure.9 Truncation of the EKF solution

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

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

we describe details in use of the above constraints in following sections.

3.1 Modeling of Articulated Object

Here, we consider the object (Fig.4) is observed by scaled orthogonal projection. We define an m-dimensional state vector of the shape and pose as

$$\boldsymbol{x} = (\boldsymbol{t}^T, s, \boldsymbol{\omega}^T, \theta_{10}, \cdots, \theta_{53}, \dot{\theta}_{10}, \cdots, \dot{\theta}_{53}, r_{11}, \cdots, r_{53})^T \tag{2}$$

where $\boldsymbol{t}$, s, $\boldsymbol{\omega}$, θ, $\dot{\theta}$, and r respectively denote a wrist position (t_x, t_y), scale of the projection, 3-D direction of palm, joint angles, its velocities and lengths of links. Supposing the constancy of the shape ((a) in Sec.3), $\dot{r}$ is not included. The transition and observation formulas are represented as

$$\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{u}_t \tag{3}$$

$$\boldsymbol{y}_t = \boldsymbol{h}(\boldsymbol{x}_t) + \boldsymbol{w}_t \tag{4}$$

where $\boldsymbol{y}_t$ is an n-dimensional observation vector. $\boldsymbol{u}_t$, $\boldsymbol{w}_t$ are white noises with zero mean and variances $\boldsymbol{U}$, $\boldsymbol{W}$. Supposing linear prediction, $\boldsymbol{A}$ is represented as the following $(2m+n) \times (2m+n)$ matrix:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I}_m & \boldsymbol{I}_m & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I}_m & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I}_n \end{bmatrix} \tag{5}$$

where $\boldsymbol{I}_m$ denotes $m \times m$ identity matrix. $\boldsymbol{U}$ is determined by considering the continuity of the pose changes ((b) in Sec.3).

The observation formula of the wrist and joint position and the finger axes is next modeled in detail. In Fig.8, the 2-D projection of the jth joint position of ith finger is described as

$$\boldsymbol{p}_{ij} = (p_{ij}^{(x)}, p_{ij}^{(y)}) = \boldsymbol{L} \cdot \boldsymbol{R}(\boldsymbol{\omega})\boldsymbol{R}(\theta_{i0}) \cdot \sum_{k=1}^{j} r_{ik} \left(\cos(\sum_{l=1}^{k} \theta_{il}), \sin(\sum_{l=1}^{k} \theta_{il}), 0 \right)^T + \boldsymbol{t} \tag{6}$$

where $\boldsymbol{L}$ and $\boldsymbol{R}$ represent projection and rotation matrices. We suppose a straight line (α_{ij}, ρ_{ij}) is extracted as the jth axis of ith finger. α_{ij} and ρ_{ij} respectively denote the direction of the line and the distance from the origin. They are formulated as

$$\alpha_{ij} = \arctan(z_{ij}^{(y)} / z_{ij}^{(x)}) \tag{7}$$

$$\rho_{ij} = (z_{ij}^{(y)} p_{ij}^{(x)} - z_{ij}^{(x)} p_{ij}^{(y)}) / |\boldsymbol{z}_{ij}| \tag{8}$$

$$\boldsymbol{z}_{ij} = (z_{ij}^{(x)}, z_{ij}^{(y)}) = \begin{cases} \boldsymbol{p}_{ij} - \boldsymbol{p}_{ij-1} & \cdots j \neq 0 \\ \boldsymbol{p}_{ij} & \cdots j = 0 \end{cases} \tag{9}$$

The observation function $\boldsymbol{h}$ consists of the $\boldsymbol{p}_{ij}$, α_{ij}, ρ_{ij} observed in one frame. In case of occlusion, $\boldsymbol{p}_{ij}$, α_{ij}, ρ_{ij} of the occluded part are not contained in $\boldsymbol{h}$. For example, if we obtain the wrist, finger tips and the all of the finger axes except the most proximal one, $\boldsymbol{h}$ is described as

$$\boldsymbol{h}(\boldsymbol{x}) = (\boldsymbol{t}^T, \alpha_{12}, \rho_{12}, \alpha_{13}, \rho_{13}, \boldsymbol{p}_{13}^T, \cdots, \alpha_{52}, \rho_{52}, \alpha_{53}, \rho_{53}, \boldsymbol{p}_{53}^T)^T. \tag{10}$$

In addition to above, the following constraints of the object is considered ((c) in Sec.3). They are formulated as

$$x_{min,i} \leq x_i \leq x_{max,i} \tag{11}$$

$$|x_i - x_j| \leq \Delta x_{ij} \tag{12}$$

The above inequalities must be simultaneously satisfied.

3.2 Distribution Truncation with Inequality Constraints

Because the observation function $\boldsymbol{h}$ is non-linear, the current state $\widehat{\boldsymbol{x}}_t$ and variance $\boldsymbol{P}_t$ are approximately estimated by extended Kalman filter (EKF) as

$$\widehat{\boldsymbol{x}}_t = \widetilde{\boldsymbol{x}}_t + \boldsymbol{K}_t\{\boldsymbol{y}_t - \boldsymbol{h}(\widetilde{\boldsymbol{x}}_t)\} \tag{13}$$

$$\boldsymbol{P}_t = (\boldsymbol{I} - \boldsymbol{K}_t \left.\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}_t}\right|_{\widetilde{\boldsymbol{x}}_t})(\boldsymbol{A}\boldsymbol{P}_{t-1}\boldsymbol{A}^T + \boldsymbol{U}) \tag{14}$$

where $\boldsymbol{K}_t$ is the Kalman gain matrix, $\widetilde{\boldsymbol{x}}_t = \boldsymbol{A}\widehat{\boldsymbol{x}}_{t-1}$. In the calculation of $\partial \boldsymbol{h}/\partial \boldsymbol{x}_t$, $\partial \boldsymbol{p}/\partial \boldsymbol{x}_t$ is directly obtained from the derivatives of Eq.(6). $\partial \alpha/\partial \boldsymbol{x}_t$ and $\partial \rho/\partial \boldsymbol{x}_t$ are also obtained from the derivatives of Eq.(7) and (8) which are reduced to $\partial \boldsymbol{p}/\partial \boldsymbol{x}_t$. However, the solution

of Eq.(13) includes errors due to the depth ambiguity.

In order to resolve the ambiguity, we introduce the model constraints Eq.(11)-(12) into EKF. If they are equations, they can be treated as an observation with a Zero variance. Here, they are however *inequalities*. Although a method is proposed in which inequality constraints are modified to quadratic equations with slack variables and linearized [4], the linearized constraints are quite different from original ones. Another way is to introduce the constraints as an initial distribution. But it is also inappropriate because the effect of the initial distribution decreases by filtering at every frame.

In our method, the inequality constrants are integrated as follows. First the mean and variance of prediction $(\hat{\boldsymbol{x}}_t^*, \boldsymbol{P}_t^*)$ which are made by normal EKF way is truncated outside the constraints as shown in Fig.9. Then the observation is integrated to the truncated prediction by Eqs.(13)(14).

Eqs.(11)-(12) are generally represented as

$$\varphi_k^T \boldsymbol{x} \leq b_k \qquad (k = 1 \cdots K). \tag{15}$$

Because of difficulty to exactly compute the distribution truncated with all constraints, it is approximated by sequential truncation with an each single constraint.

Suppose the distribution with a mean $\boldsymbol{q}_{k-1}$ and a variance $\boldsymbol{Q}_{k-1}$ is truncated by the constraint $\varphi_k^T \boldsymbol{x} \leq b_k$, where $\boldsymbol{q}_0 = \hat{\boldsymbol{x}}_t^*$ and $\boldsymbol{Q}_0 = \boldsymbol{P}_t^*$. This computation is reduced to the case where the mean is $\mathbf{o}$, the variance is identity matrix $\boldsymbol{I}$ and the constraint is $(1,0,\cdots,0)^T \boldsymbol{x}' \leq c_k$, by applying the following transformation:

$$\boldsymbol{x}' = \boldsymbol{R}\boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{T}^T(\boldsymbol{x} - \boldsymbol{q}_{k-1}) \tag{16}$$

where $\boldsymbol{R}$, $\boldsymbol{T}$ is orthogonal, $\boldsymbol{W}$ is diagonal and

$$\boldsymbol{T}\boldsymbol{W}\boldsymbol{T}^T = \boldsymbol{Q}_{k-1} \tag{17}$$

$$\boldsymbol{R}\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{T}^T\varphi_k = (1,0,\cdots,0)^T \tag{18}$$

$$c_k = (b_k - \varphi_k^T \boldsymbol{q}_{k-1}) / \left(\varphi_k^T \boldsymbol{Q}_{k-1}\varphi_k\right)^{\frac{1}{2}} \tag{19}$$

In this case, the truncated mean $\boldsymbol{\mu}_k$ and variance $\boldsymbol{S}_k$ is computed as

$$\boldsymbol{\mu}_k = (\nu_k, 0, \cdots, 0)^T \tag{20}$$

$$\boldsymbol{S}_k = \mathrm{diag}\{1 + c_k\nu_k - \nu_k^2, 1, \cdots, 1\} \tag{21}$$

$$\nu_k = -\sqrt{\frac{2}{\pi}}\exp(-\frac{c_k^2}{2}) / (1 + \mathrm{erf}(\frac{c_k}{\sqrt{2}})) \tag{22}$$

where $\mathrm{erf}(\cdot)$ represents the error function and $\mathrm{diag}\{a, b, \cdots\}$ represents a diagonal matrix whose diagonal elements are $a, b, \cdots$. Then the truncated mean and variance are expressed as

$$\boldsymbol{q}_k = \boldsymbol{T}\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{R}^T\boldsymbol{\mu}_k + \boldsymbol{q}_{k-1} \tag{23}$$

$$\boldsymbol{Q}_k = \boldsymbol{T}\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{R}^T\boldsymbol{S}_k\boldsymbol{R}\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{T}^T. \tag{24}$$

Finally, the fully truncated mean and variance are obtained by recursive computation: $\hat{\boldsymbol{x}}_t = \boldsymbol{q}_K$ and $\boldsymbol{P}_t = \boldsymbol{Q}_K$. Because the computation of Eq.(17) takes much time, the kth truncation is skipped for efficiency if c_k, mahalanobis distance from $\boldsymbol{q}_0$ to the plane $\varphi_k^T \boldsymbol{x} = b_k$, is greater than a threshold.

3.3 Multiple Estimation

Estimation by EKF may fail because the distribution becomes multimodal due to the depth ambiguity. Fig.10 shows an example of a 2-D link system in which 1-D joint positions are observed. At the 18th frame(Fig.10(b)), In spite that there are two possible solution (see (e)), normal EKF only produces either. The sampling method [5] can treat the multimodal case. However, it should generate a number of samples for one dimension. That makes the method hard to apply to high dimensional cases like articulated objects.

In our method, we generate and preserve multiple estimates. This means that the multimodal distribution is approximated by sum of multiple gaussian distributions. For the ith link, the multimodal problem arises when the link is nearly parallel to the screen, namely when $\partial \boldsymbol{h}_i / \partial \theta_i |_{\tilde{\boldsymbol{x}}_t} \simeq 0$, where $\boldsymbol{h}_i$ and θ_i respectively denote observations of the ith link and the proximal joint angle. If the prediction $\tilde{\boldsymbol{x}}_t$ satisfies above equation, the following processes are activated.

1. Generate the $\tilde{\boldsymbol{x}}_t^{sym}$ which is identical to $\tilde{\boldsymbol{x}}_t$, except that the ith link is symmetrical to $\tilde{\boldsymbol{x}}_t$ with respect to the screen.
2. At both of $\tilde{\boldsymbol{x}}_t$ and $\tilde{\boldsymbol{x}}_t^{sym}$, calculate $\partial \boldsymbol{h} / \partial \boldsymbol{x}_t$ in Eq.(14)
3. With each $\partial \boldsymbol{h} / \partial \boldsymbol{x}_t$, calculate the estimate by modified EKF respectively with the original prediction $\tilde{\boldsymbol{x}}_t$ and its variance.

At most, 2^n (n:the number of links) estimates are possible. In case that the area truncated by the constraints is more than a threshold, such a estimates is eliminated as illegal. The rest are also preserved for robustness in the same way of the rough estimation. Only a certain number of candidates with high probabilities (Eq.(1)) are preserved until unique solution is determined by Beam search.

In Fig.10, two estimates are simultaneously generated at the 18th frame (e) and the wrong estimate is eliminated at the 20th frame (f) and the following estimation is successfully continued.

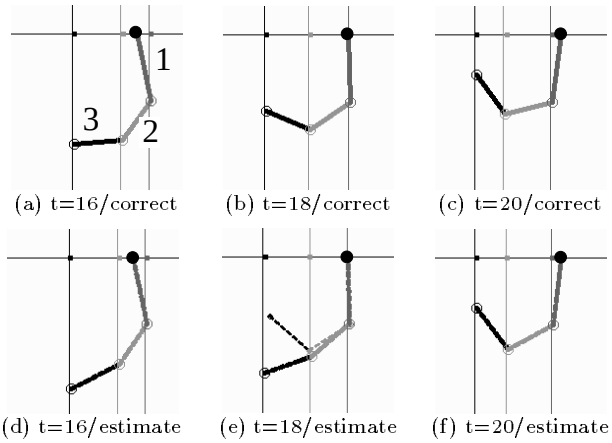
4 Experimental Results

We first show an estimation result by a simulation with a synthesized image sequence (Fig.11). In the sequence, a hand-like object moves to right and left, rotates, and folds its fingers. We utilize the constraints shown in Tab.1. The initial estimate and variance is set so that the correct value is included in 99% confidence region.

In Fig.11, (a)-(c) show correct shapes and poses and (d)-(f) show the estimates. Observed wrist positions, finger tips and finger axes perturbed by Gaussian noise are shown by the black spheres and straight lines in

Table 1. The constraints used in the simulation

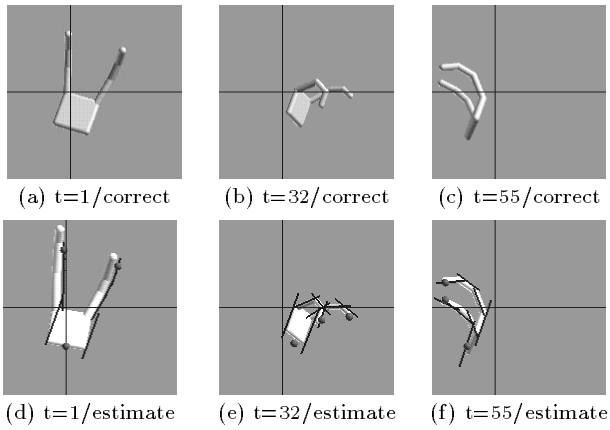
pose constraints		shape constraints	
	$\lvert\theta_{i2} - \theta_{i3}\rvert \leq 10deg$,		$0 \leq r_{i1} - r_{i2} \leq 30$,
	$-20deg \leq \theta_{10} \leq 0deg$,		$\lvert r_{i2} - r_{i3}\rvert \leq 10$,
	$0deg \leq \theta_{20} \leq 20deg$,		$55 \leq r_{i1} \leq 70$,
	$0 \leq \theta_{ij} \leq 90deg$		$30 \leq r_{i2} \leq 50$,
	$\lvert\theta_{11} - \theta_{21}\rvert \leq 35deg$		$25 \leq r_{i3} \leq 50$
	for $i = \{1,2\}$, $j = \{1,2,3\}$		for $i = \{1,2\}$



(a) t=16/correct (b) t=18/correct (c) t=20/correct

(d) t=16/estimate (e) t=18/estimate (f) t=20/estimate

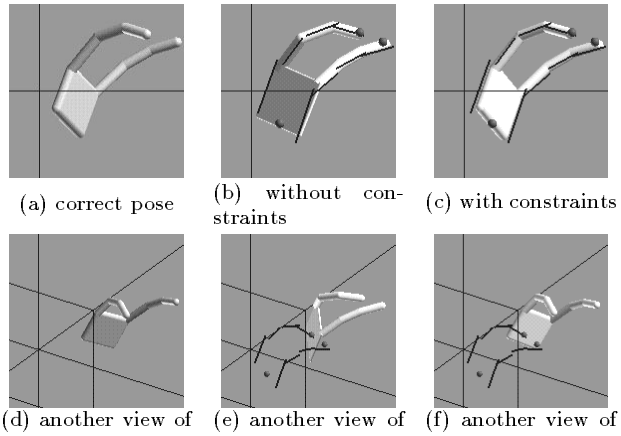
Figure 10. Multiple estimation: In (e), solid line shows the wrong estimate obtained by normal EKF and broken line shows the alternative estimate.



(a) t=1/correct (b) t=32/correct (c) t=55/correct

(d) t=1/estimate (e) t=32/estimate (f) t=55/estimate

Figure.11 Estimation results by simulation



(a) correct pose (b) without constraints (c) with constraints

(d) another view of (a) (e) another view of (b) (f) another view of (c)

Figure.12 Comparison of estimates with / without constraints (t=23)

(d)-(f). Although the hand poses are wrongly estimated without constraints due to the depth ambiguity (Fig.12(b),(e)), our method correctly estimates the pose with the constraints in the Tab.1 (Fig.12(c),(f)).

The correct and estimated parameters (scale,a joint angle and a finger length) are plotted in Fig.14-16, in which the solid lines, small circles and the vertical lines respectively show the correct values, estimated mean and a twice of standard deviations. The joint angles are well-estimated and the possible range of the finger lengths (r_{ij}) are correctly limited. Fig.13 shows that two different shapes (b) and (c) are correctly identified using the same initial shape shown in (a) and the same constraints.

We next show an estimation result for the real hand images. In this experiment, only finger lengths are estimated as the shape. Fig.17 shows the refined pose estimates for the rough estimates shown in Fig.7. Before the refinement process, the image features are segmented into each part of fingers using the rough estimation result, and the observation α and ρ are calculated by line fitting to the segmented features. (Fig.17(a)-(c)). Then the pose estimates are refined ((d)-(f)). The refinement result of the shape model is shown in Fig.18.

5 Conclusion and Discussion

In this paper, we propose a method to simultaneously estimate the shape and pose of a human hand from a monocular image sequence. First, the pose is roughly estimated by silhouette matching. Various candidates are generated and matched to an input silhouette. The search space is efficiently reduced using the silhouette features and motion prediction. The posterior probabilities of the candidates are evaluated in order to integrate the prior probabilities and the likelihoods of different sorts of the matching degrees. Next, we refine the rough pose and the initial shape model using the modified EKF with distribution truncation by inequality constraints. Then the depth ambiguity is incrementally limited with informative observations over the sequence. In addition, we resolve the ambiguity of symmetrical poses by generating and preserving multiple solutions. We show the effectiveness of our method by simulation and an application to a real hand images. This method is applicable to gesture estimation and model acquisition of other articulated objects.

However, we still have a problem. In some cases, the

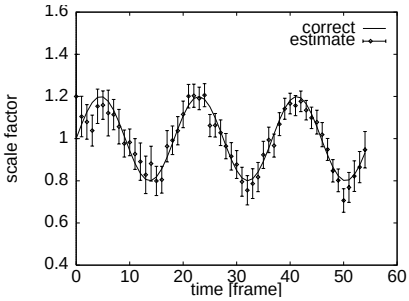


Figure.14 Mean and variance of s

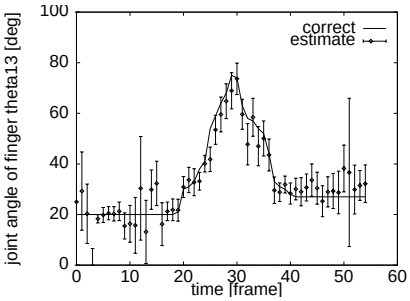


Figure.15 Mean and variance of θ_{13}

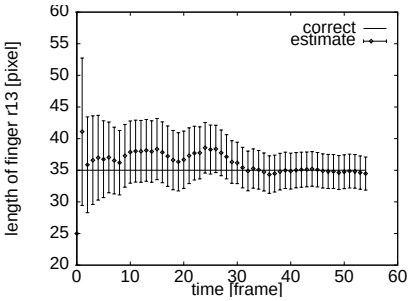
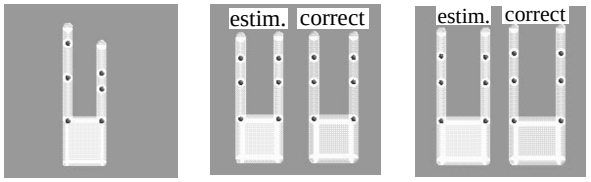
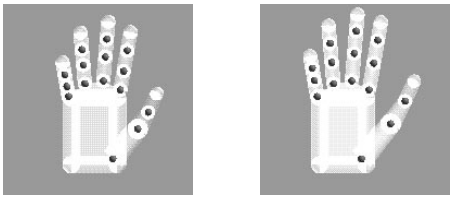


Figure.16 Mean and variance of r_{13}



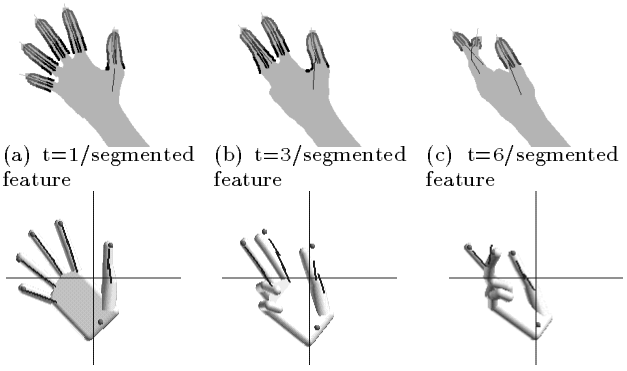
(a) Initial shape (b) Refined shape1 (t=55) (c) Refined shape2 (t=55)

Figure 13. The result of shape refinement (simulation)



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

Figure 18. The result of shape refinement (real image)



(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 17. Shape and pose estimation result from real images

variance of the estimate improperly decreases. This is caused by the error of linearization of the observation. One way to solve this problem is to use a boundary description instead of the probability distribution. In general, however, the computation of the boundary in multi-dimensional space is almost impossible. An approximation method was proposed [3][12] which approximate the boundary by an ellipse in a multi-dimensional space and updates the ellipse with each observation iteratively. There remains for the future works to apply that method to our problem and to estimate not only lengths and widths but also shape of surfaces.

References

[1] R. A. Brooks. "Symbolic Reasoning Among 3-D Models 2-D Images". *Artificial Intelligence, Vol.17, No.1-3*, pages 285–348., 1981.

[2] J. Davis and M. Shah. "Recognizing Hand Gestures". *ECCV'94.*, pages 331–340, 1994.

[3] E. Fogel and Y. F. Huang. "On the Value of Information in System Identification – Bounded Noise Case". *Automatica, vol.18, No.2*, pages 229–238, 1982.

[4] Y. Hel-Or and M. Werman. "Recognition and Localization of Articulated Objects". In *Proc. of Workshop on Motion of Non-Rigid and Ariticulated Objects '94*, pages 116–123. IEEE, 1994.

[5] M. Isard and A. Blake. "Contour Tracking by Stochastic Propagation of Conditional Density". *ECCV'96.*, pages 343–356, 1996.

[6] Y. Kameda, M. Minoh, and K. Ikeda. "Three Dimensional Pose Estimation of an Articulated Object from its Silhouette Image". In *ACCV'93*, pages 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*, pages 196–199, 1994.

[8] B. T. Lowerre and R. D. Reddy. "The Harpy Speech Understanding System". In W. A. Lea, editor, *Trends in Speech Recognition*, pages 340–360. PresticeHall, Englewood Cliffs, NJ, 1980.

[9] M. Mochimaru and N. Yamazaki. "The Three-dimensional Measurement of Unconstrained Motion Using a Model-matching Method". *ERGONOMICS, vol.37, No.3*, pages 493–510, 1994.

[10] J. O'Rourke and N. I. Badler. "Model-Based Image Analysis of Human Motion Using Constraint Propagation". *IEEE Trans. of Pattern Anal. and Machine Intell.,PAMI-2, No.6*, pages 522–536, 1980.

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

[12] F. C. Schweppe. "Recursive State Estimation: Unknown but Bounded Errors and System Inputs". *IEEE Trans. on Automatic Control, vol.AC-13, No.1*, pages 22–28, 1968.

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

[14] M. Yamamoto and K. Koshikawa. "Human Motion Analysis Based on A Robot Arm Model". In *CVPR'91*, pages 664–665. IEEE, 1991.