21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Real-Time and Precise 3-D Hand Posture Estimation Based on Classification Tree Trained with Variations of Appearances

Sho MIYAMOTO, Tadashi MATSUO, Nobutaka SHIMADA, Yoshiaki SHIRAI
Ritsumeikan University
miyamoto@i.ci.ritsumei.ac.jp

Abstract

We propose a method for estimating 3-D hand postures from 2-D monocular images in real-time. The estimation is based on finding the best matched posture from typical postures whose appearances are learned in advance. For high accuracy, conventional methods require high computational cost for comparing an input with many typical postures. In our method, a tree is automatically generated and trained with typical postures and their variations. Efficient search with the tree brings about real-time estimation. We show the effectiveness of our method by some experimental results.

1. Introduction

Gesture recognition has been attracting attention as one of the human interfaces. It is very difficult to estimate complex hand shape and movement such as sign language by simple appearance matching. Conventional methods for hand posture estimation are roughly divided into two types: three-dimensional model and two-dimensional appearance.

The former is a method to find the hand shape by changing parameters of the three-dimensional hand model to match the input image features [1, 2, 3]. It is possible to estimate with high accuracy by using three-dimensional model. If the self occlusion occurred, which is in the state, where fingers and palm are hidden in other fingers, it is generally not robust. In order to solve this problem, the techniques using multi-viewed information with two or more cameras are also proposed [4, 5].

The latter is an approach to estimate the three-dimensional shapes by exploring the hand posture with high degree of similarity to the input. Each hand posture and its two-dimensional appearance is stored into a shape database. These approaches can overcome the self occlusion [6, 7, 8]. However it is difficult to estimate in real time by the simple search because these approaches require to register huge variations of postures for improving accuracy. Although the techniques limiting the candidate postures by smooth finger motion constraint are proposed, they are established in real time using restriction of degree of freedom and large size computers [9, 6, 7]. Hoshino et al. [10] uses learning of relations between the actual 2D appearance and the 3D hand parameters using Higher Local Auto Correlation of hand contours and the self-organizing map [11] for efficiency and optimizing estimation.

We propose a method of pose estimation based on 2-D appearance matching based on 3-D model via CG generation: *Estimation by Synthesis*. A number of 3-D hand shape CG samples are generated by changing the joint angle and then we create a classification tree based on visual features like the concavity and convexity of 2-D hand CG shapes for efficient matching. An input image feature is first extracted, is second classified into a leaf node of the tree. Then the feature is precisely matched to the hand shapes in the node by time-consuming matching process.

It is too expensive process to directly build such a classification tree with huge number of shape samples for precise accuracy. Otherwise, due to too coarsely sampled joint parameters, the input shape of the user's hand substantially differs from original CG model and thus is difficult to match.

Thus, we first build a primal classification tree with very coarsely sampled CG hand shapes (*typical postures*), and then we more precisely make samples which slightly deforms from each typical postures, and try to classify each of them based on the primal tree. If the precise samples (*variations*) are classified to the same leaf node as its original typical postures, those variations are stored in the typical postures. Otherwise, we make a new typical postures by clustring the variations classified to an identical leaf node and register the new postures in the leaf node. This procedure works as the re-segmentation of too coarsely divided typical postures in the primal classification tree. Although the number of postures per leaf node increases by learning more variations, highly precise estimation can be performed. Our system can process 30 or more frames per second with Intel Core i7 2.66GHz PC.

978-4-9906441-1-6 ©2012 IAPR

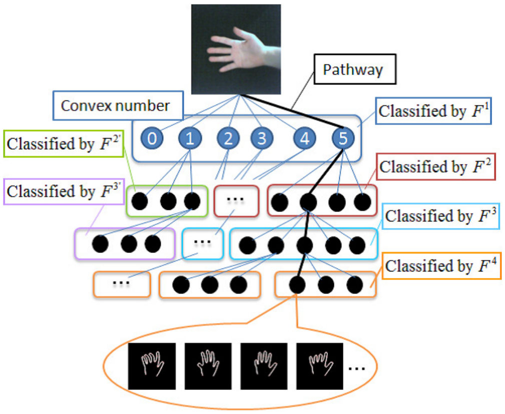

Figure 1. Classification tree

2. Hand Posture Estimation

2.1. Generate Typical Postures

It is impossible to learn the appearances of all hand postures since hand posture changes continuously. Therefore learning is conducted about the postures sampled by a combination of quantized joint angles. We determined to call these quantized postures as ”typical postures”. We use a simplified hand model because the usual gesture does not need high DOF.

We assume the following relationship between joint angles: DIP, PIP and flexion/extension rotation of MP joint angles are all the same for each finger, abduction/adduction rotation of MP joint angle is fixed to a certain natural position. Thus the all parameters of hand posture description consist of the following 7 DOFs: $\theta_{wrist1}, \theta_{wrist2}$ for the palm orientation except the rotation in a image (the 3rd rotation DOF is normalized in the preprocess), $\theta_{thumb}, \theta_{index}, \theta_{middle}, \theta_{ring}, \theta_{pinky}$ for the finger flextion. These parameters are quantized in 4 steps for θ_{thumb}, θ_{index}, θ_{middle}, θ_{ring}, θ_{pinky}, 12 steps for θ_{wrist1}, 7 steps for θ_{wrist2}. The total number of sampled typical postures are $86016 \left(= 4^5 \times 12 \times 7\right)$.

2.2. Coarse Classification Tree

A classification tree (Fig. 1) is created based on the features of the typical postures sampled by Sec. 2.1's quantization.

Child nodes of the first level of the tree are generated by classifying typical postures by the number of finger-like convex protrusions of the shape contour (0 to 5). The nodes in the first layer having 2 - 5 convex protrusions are further classified in the second layer based on relative offset vector between the protrusions which is assumed to follow a Gaussian Mixture distribution by applying unsupervised clustering [12]. The number of clusters in each layer is determined by the minimum description length (MDL) method. For the further classification of subclusters in the 3rd and 4th layers other heuristic features describing on the position and the shape of the protrusions and the concave parts of the contour are repeatedly employed to build the classification tree.

2.3. Learning Micro Variations

The classification tree is created by considering only the typical postures sampled very coarsely. Therefore, some posture having a similar joint parameters to one of the typical postures may be classified into the different nodes due to the drastic appearance change not cared by the coarse shape sampling. In those cases such a posture should be represented as typical postures more than one by the more precise clustering in advance.

To solve this problem, we employ the coarse classification tree for the judgement whether such precise clustering is required around each coarse typical postures and the clustering itself. Specifically, the following process is performed about each typical posture p_t, for the micro variation learning. First, micro variations $U_t = \{u_t^j\}$ are generated around p_t by changing the posture parameters. Then each u_t^j is classified by the equivalence relation $p \equiv q \Leftrightarrow H(p) = H(q)$ where $H(p)$ shows the results of classification by the coarse classification tree. The resulted classification destinations (the histogram of the leaf nodes into which each variations are classified) are numbered in descending order of the number of the elements in the node, as $A_{t,0}, A_{t,1}, \cdots, A_{t,N-1}$. These leaf nodes are all possible candidates when a real hand shape similar to p_t are input. If all of them are registered into the refined classification tree the amount of computation of detail matching is much increased while the estimation (classification) accuracy improves. Therefore we omit the rare destinations from the refined classification tree. τ is the percentage of the total classification, N_t which is the employed number of classes are as follows.

$$N_t = \min\left\{ n \in [0, N-1) \,\middle|\, \tau < \frac{\sum_{k=0}^{n} |A_{t,k}|}{\sum_{k=0}^{N-1} |A_{t,k}|} \right\}$$

The new typical posture having an average joint angles for each resulted cluster of $A_{t,0}, A_{t,1}, \cdots, A_{t,N_t}$ is respectively registered as the new typical posture of $p_{t,n}$. By repeating this process for every typical postures stored in the coarse classification tree, a refined classification tree is bootstrapped.

2.4. Hand Posture Estimation with Classification Tree

Given the refined classification tree, the hand posture estimation for an input image can be done with the following steps.

1. Extract the hand contour from the input image. Then calculate the feature vector s from the extracted contour points.

2. Set the variable n to the root node, and the depth of tree layer $L \leftarrow 0$.

3. Determine the node m that minimizes the Mahalanobis distance (see below) between the classification feature $F^L(s)$ and the set of child nodes C_n of the node n The classification feature F^L is defined for each tree layer L, and its average μ_k and the covariance Σ_k of each node k are calculated based on the all micro variations classified into the node n in the micro variation learning process (Sec.2.3).

$$m = \arg\min_{k \in C_n} \left(F^L(s) - \mu_k\right)^{\mathrm{T}} \Sigma_k^{-1} \left(F^L(s) - \mu_k\right)$$

$$\mu_k = \mathop{\mathrm{E}}_{v \in k} \left[F^L(v)\right]$$

$$\Sigma_k = \mathop{\mathrm{E}}_{v \in k} \left[\left(F^L(v) - \mu_k\right)\left(F^L(v) - \mu_k\right)^{\mathrm{T}}\right]$$

4. If m is not a leaf node, set $n \leftarrow m$, $L \leftarrow L + 1$ and back to 3.

5. If m is a leaf node, the best-matched typical posture $p_{\hat{t}}$ is determined by minimizing the feature distance $D(v, s)$ by Imai et al. [8] about typical postures $p_{\hat{t}}$ belonging to m.

$$T(m) = \left\{(t, n) \,\middle|\, {}^{\exists}n \leq N_t, {}^{\exists}p \in A_{t,n} \text{ s.t. } m = H(p)\right\}$$

$$(\hat{t}, \hat{n}) = \mathop{\mathrm{argmin}}_{(t,n) \in T(m)} D(s, p_{t,n})$$

6. Estimated result is the representative posture $p_{\hat{t},\hat{n}}$ which is the average of $A_{\hat{t},\hat{n}}$.

3. Experiment

3.1. Generate Classification Tree

The classification tree for limiting the candidate postures is created based on concavity and convexity features of 86016 CG images which are generated by the three-dimensional hand model. Table 1 shows the measure of classification ability. Each node of the first layer (Classified with convex number F^1 (0 to 5)), Total number of belonging postures, Number of leaf nodes (Nodes of F^4 layer classified by the feature F^4), Average number of postures per leaf nodes (Equal to the number of detail matching [8]) are shown as a measure of classification ability. Classified Input contour feature with the classification tree, if convex number is 3, the candidate postures are limited to average 165 postures for detail matching [8]. From these, most of the leaf nodes (0.1-0.5% of the total number of postures) can be reduced in 100-500 postures.

Table 1. The number of leaf nodes

Convex number	Number of postures	Number of leaf nodes	Postures per leaf node
0	2561	1	2561
1	18476	40	462
2	37954	75	500
3	36976	224	165
4	16707	209	81
5	3410	38	89

Table 2. Number of typical postures in leaf nodes.

τ	0.5	0.7	0.9	1.0
Postures/leaf node	710.7	1319.5	2923.2	9336.2

Table 2 is the average numbers of the postures per leaf node by changing τ introduced in section 2.3 to consider the micro variations. The average numbers of the postures per leaf node is increased corresponding to the value of τ.

3.2. Recognition Result

Experiments were performed to estimate hand posture about real hand input image of 1170 frame. Some of the estimated results shown in Fig. 2. Judging subjectively that each frame estimation result is correctly recognized or not, estimation accuracy is 76.50%. Recognition accuracy is 100% in experiments with the CG model used to learning.

Fig. 3 shows the recognition result for CG created with different joint angle from used in learning. In the figure, an estimated posture is “correct” if it has the nearest joint angles to those of the input CG. The experiment was performed with $1/32, 1/16, 1/8$ and $1/4$ of the maximum quantization interval width of the joint angle. If $0.9 < \tau$, hand posture estimation is performed with high recognition accuracy.

3.3. Processing Time

Experiment was performed to register processing time by estimating hand posture in parallel about 1000 frame with a Core i7 2.66GHz 8core machine. In our previous method using simple search, average processing time per frame is 509 msec. In proposed method with $\tau = 0.0$, the candidate postures are limited to average 464 postures for detail matching, average processing time per frame is 9.01 msec. 110 fps system was realized. The result learned about micro variations is Fig. 4. Processing time for detail matching is increased corresponding to the value of τ. If $\tau = 0.0$, hand posture estimation can be performed with small computa-

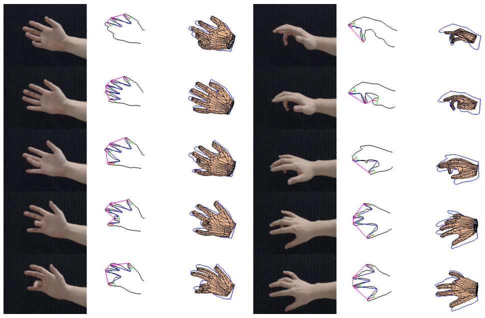

Figure 2. Recognition result.

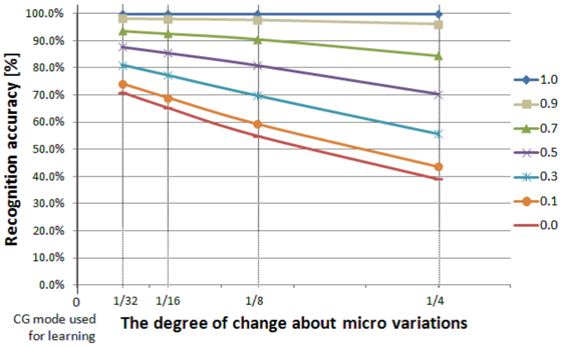

Figure 3. Recognition result about CG for each τ value

tional cost in 110 fps. In contrast, if $\tau = 0.9$, hand posture estimation can be performed with high recognition accuracy in 30 fps.

4. Conclusion

This study introduced a hierarchical classification tree based on the concavity and convexity feature of hand contour. It is possible to estimate real time without large computers because the classification tree is designed to reduce computational cost. Classifying the appearance of micro variations of about 86000 typical posture into 800 classes, hand posture estimation is performed with high recognition accuracy in 30 fps.

Although the proposed method for the CG is very good recognition accuracy, a slightly different presumed result is also seen about the actual image taken of human hand. This is probably due to the roughness of the joint angle between the quantization. Future problem is selection of a proper joint angle.

References

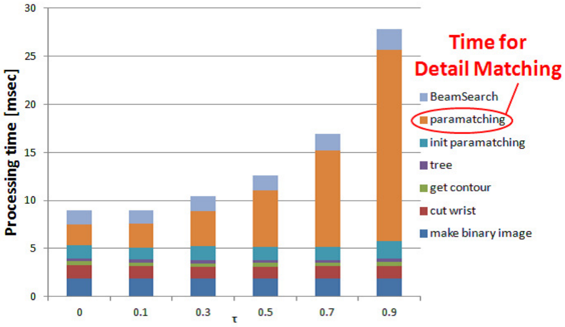

Figure 4. Time for recognition for each τ value

[1] S. U. Lee and I. Cohen. 3d hand reconstruction from a monocular view. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 310 – 313 Vol.3, aug. 2004.

[2] J. Rehg and T. Kanade. Visual tracking of high dof articulated structures: An application to human hand tracking. In J.-O. Eklundh, editor, *Computer Vision – ECCV '94*, volume 801 of *Lecture Notes in Computer Science*, pages 35–46. Springer Berlin / Heidelberg, 1994. 10.1007/BFb0028333.

[3] D. Lowe. Fitting parameterized three-dimensional models to images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 13(5):441 –450, may 1991.

[4] W. Chen, R. Fujiki, D. Arita, and R. ichiro Taniguchi. Real-time 3d hand shape estimation using multiple cameras. In *Proceedings of 13th Japan-Korea Joint Workshop on Frontiers*, volume 2007, pages 15–20, jul. 2007.

[5] E. Ueda, Y. Matsumoto, M. Imai, and T. Ogasawara. A hand-pose estimation for vision-based human interfaces. *Industrial Electronics, IEEE Transactions on*, 50(4):676 – 684, aug. 2003.

[6] B. Stenger, A. Thayananthan, P. Torr, and R. Cipolla. Model-based hand tracking using a hierarchical bayesian filter. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(9):1372 –1384, sept. 2006.

[7] K. Hoshino, E. Tamaki, and T. Tanimoto. Copycat hand robot hand imitating human motions at high speed and with high accuracy. *Advanced Robotics*, 21(15):1743–1761, 2007.

[8] A. Imai, N. Shimada, and Y. Shirai. 3-d hand posture recognition by training contour variation. *Proc. of 6th Int. Conf. on Automatic Face and Gesture Recognition*, pages 895–900, 2004.

[9] N. Shimada, Y. Shirai, Y. Kuno, and J. Miura. Hand gesture estimation and model refinement using monocular camera – ambiguity limitation by inequality constraints. *Proc. of The 3rd Int. Conf. on Automatic Face and Gesture Recognition*, pages 268–273, 1998.

[10] K. Hoshino and M. Tomida. 3d hand pose estimation using a single camera for unspecified users. *Journal of Robotics and Mechatronics*, 21(6):749–757, 2009.

[11] T. Kohonen. The 'neural' phonetic typewriter. *Computer*, 21(3):11 –22, march 1988.

[12] C. A. Bouman. Cluster: An unsupervised algorithm for modeling gaussian mixtures. `https://engineering.purdue.edu/~bouman/software/cluster/`.