Real-time 3D Hand Shape Estimation using Multiple Cameras

Chen, Weiying
Department of Intelligent Systems, Kyushu University

Fujiki, Ryuji
Department of Intelligent Systems, Kyushu University

Arita, Daisaku
Department of Intelligent Systems, Kyushu University

Taniguchi, Rin-ichiro
Department of Intelligent Systems, Kyushu University

https://hdl.handle.net/2324/6160

出版情報：Proceedings of 13th Japan-Korea Joint Workshop on Frontiers, pp.15-20, 2007-01
バージョン：
権利関係：

Real-time 3D Hand Shape Estimation using Multiple Cameras

Weiying CHEN, Ryuji FUJIKI, Daisaku ARITA and Rin-ichiro TANIGUCHI †

† : Department of Intelligent Systems, Kyushu University
{chen,fujiki,arita,rin}@limu.is.kyushu-u.ac.jp

Abstract Vision-based hand shape estimation is a challenging task, since the hand presents a motion of high degrees of freedom and since self-occlusions of different fingers bring a lot of uncertainty for the occluded parts. Considering that the influence of self-occlusions may be reduced by observing multiple images, we propose a multiple view system to obtain hand features, with using previous information that facilitates very robust feature extraction. The extracted features are then used to compute approximate global state of hand and perform preliminary estimation of the local state of each finger by Inverse Kinematics (IK). By minimizing the estimation error between groups of model features and groups of image features, some model parameters are refined and IK is recomputed, which contributes to enhance estimation accuracy. To reduce the estimation complexity due to the high degrees of freedom of the hand, we combine IK with the motion constraints of articulated hand. The effectiveness of our approach is demonstrated with experiments on a number of different hand motions with finger articulation and global hand rotation under complex background.

1 Introduction

In recent years hand shape recognition has attracted increasing attention, because of its crucial role in the design of new human computer interaction method as an alternative way to traditional input devices like keyboards and mice. Special markers/attachments can induce direct hand shape estimation, but those devices are generally expensive and cumbersome. On the other hand, vision-based techniques offer an inexpensive and non-invasive alternative. In this paper we present an approach for 3D hand shape estimation based on computer vision. However, a hand has high degrees of freedom and causes a lot of self-occlusions of different fingers, which introduces significant challenges for the 3D hand shape estimation.

One strategy for estimating the 3D hand shape is using the appearance-based approach, which attempts to estimate the hand shape directly from the image features: a nonlinear mapping is learned from a large amount of training images. For example, Stenger et al. [1] present a tree-based estimator based on Bayesian filter. The approach achieves coarse to fine search by approximating the posterior distribution at multiple resolutions, and hopeless sub-trees are not further evaluated. Shimada et al. [2] represent variations of possible shape appearances as the Locally-Compressed Feature Manifold in an appearance feature space. It is effective to prevent the system from tracking failures and reduce the search area. This approach can quickly estimate the hand shape once the mapping is learned. However, it is difficult to determine the set of optimal training data and deal with a large amount of templates.

Model-based approach is another alternative for estimating hand motions, which extracts local hand features from images and estimates hand shapes by fitting a 3D hand model to the features. Ueda et al. [3] demonstrate a voxel representation reconstructed from silhouette images by a multi -viewpoint camera system, and then a 3D hand shape is estimated using model fitting between a 3D hand model and the voxel model. Lu et al. [4] compute 2D data-based forces from edge, optical flow and shading cue constraints and then use a dynamic hand model to convert the 2D forces into 3D forces that drive the hand model. This model-based

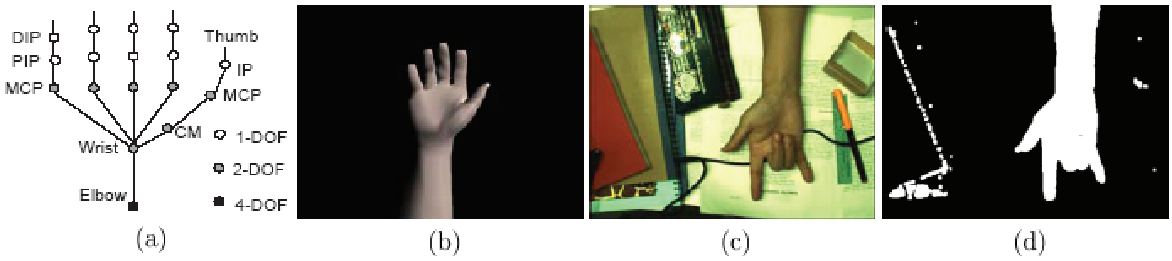


Fig. 1 (a) skeleton model, (b) skin model, (c) input image, (d) binary image

approach often requires reliable feature detection which is plagued by self-occlusion. Another problem lies in the extremely huge search space due to the high articulation of the hand, resulting in the difficult implementation of real-time system.

In this paper, we propose a novel method of estimating 3D hand shapes based on shape features acquired from multiple camera images. In principle, it is enough to get depth information by using a stereo image pair, but a naive stereo camera system is not tolerant of the self-occlusion. Alternatively, we employ multiple cameras and automatically select two proper cameras with fewer occlusions. Then, hand shape features such as wrist and finger feature points are extracted from silhouettes and contours. Considering correspondences of detected finger feature points between successive image frames, finger feature points utilized for kinematics estimation can be robustly detected. The extracted features are used to compute approximate global state of hand and perform preliminary estimation of the local state of each finger by IK. By minimizing the estimation error between groups of model features and groups of image features, some model parameters are refined and IK is recomputed, which contributes to enhance estimation accuracy. Although the hand is highly articulated, hand motion is highly constrained. Therefore, we combine hand constraints with IK to reduce the complexity and ambiguity of estimation.

2 Hand Model

In this paper, a hand is approximated by a 3D rigid articulated model object. The 3D hand model consists of a skin model and a skeleton model (see Fig.1 (a, b)).The skin model is not currently used for hand shape estimation, except for visualization of results of the estimation. The hand skeleton can be abstracted as a stick figure with each finger as a kinematical chain with base frame at the palm and each fingertip as the end-effecter. It has 27 degrees of freedom (DOFs) in total, including the translation and rotation of the elbow. In each finger except for the thumb, DIP joint and PIP joint has 1 DOF and MCP joint has 2 DOFs due to flexion and abduction. The thumb has a different structure from the other fingers and it has 5 DOFs, one for IP joint and two for the thumb MCP joint and CM joint due to flexion and abduction. Altogether, the fingers have 21 DOFs. The remaining 6 DOFs are from the motion of the wrist and elbow. The wrist joint has 2 DOFs from yaw and pitch rotation of the palm while the elbow joint has 4 DOFs from translation and roll rotation.

3 Feature Extraction

3.1 Color Segmentation

We extract a hand region under complex background using segmentation in HSV color space. In our system, samples of a known hand region are prepared beforehand to calculate the 2D histogram over HS components of the hand region $h_{skin}(h,s)$ .The probability $p_{skin}(h,s)$ of a pixel with color (h,s) being skin color is given by the Bayesian rule, and here, it is reduced into a simple ratio of the two histograms, where $h_{input}(h,s)$ denotes a HS histogram of the input image [5]:

$$p(skin \mid h,s) = \frac{h_{skin}(h,s)}{h_{input}(h,s)} \qquad (1)$$

A pixel is labeled as a skin pixel when the probability $p_{skin}(h,s)$ is higher than a given threshold. Fig.1 (d) shows an example of binary image acquired by the skin pixel labeling. However, the noise in the region boundaries is quite noticeable and small skin-like colored regions are unnecessarily labeled. To remove the noise, morphological operators are used to smooth

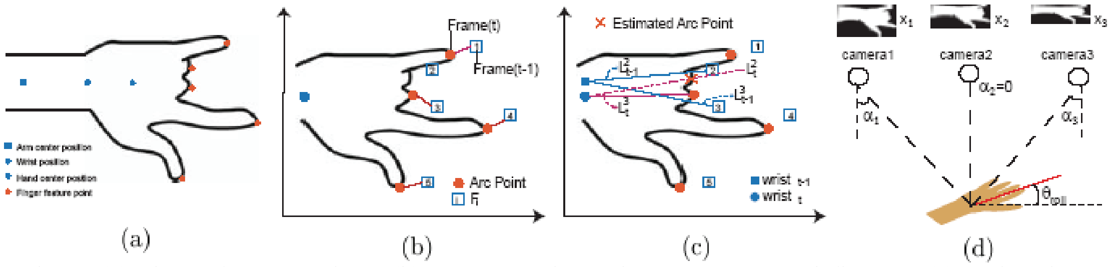


Fig. 2 (a) shape feature points, (b) correspondence between arcs and fingers, (c) estimation of undetected arc points, (d) camera layout

the binary image. Furthermore, to make the following processing more efficiently, a minimal-sized rectangle of the hand contour is retrieved and the rectangle is then normalized with rotation where the X axis is set to coincide with the arm direction.

3.2 Shape Features

In our multi-view camera system, a criterion must be defined to select two proper cameras to be used for obtaining 3D information. Since the self-occlusion affects the feature extraction, the camera full-face to the palm which arises few self-occlusions is desirable. According to the fact that a hand region extracted from the camera image gets bigger when the palm turns to face the camera, two cameras with bigger hand regions in the images are automatically selected. However, the unselected cameras are also used for obtaining 3D hand information if the two selected cameras can not offer enough information. The selected stereo camera images are used to calculate 3D position of shape features which consists of non-finger feature points and finger feature points (see Fig.2 (a)). The non-finger feature points refer to wrist, arm center and hand center. They are considered to approximately calculate parameters that present global hand pose in section 4. The finger feature points as significant features for finger pose estimation are described in the following section.

3.3 Finger Feature Detection

As major features of 3D hand shape estimation, we detect arcs on the contour of the extracted hand regions. These arcs projected outer-most in 2D image space correspond to five fingers' joints and they are used as the positions of the end-effecters in IK. The problem of estimating 3D position of each finger feature point can be decomposed into the following three sub-problems.

1. Detection of Arcs on the Contour: We detect arcs on the contour by curvature information, i.e., we detect contour points with high curvatures. However, it is difficult to get five arcs rightly corresponding to the five fingers. It results in two challenges for the correspondence between the arcs and fingers and correspondence between arcs extracted in two cameras. Since there are many combinations about the correspondence, hand constraints, the finger order and intervals between fingers are considered to reduce the combination. Especially, we make use of correspondence of detected finger feature points between successive image frames. This contributes to significantly better performance for establishing new correspondences and estimating undetected arc points. It can help to reach the solution even if five fingers are disordered or some finger points are occluded.

2. Correspondence between Arcs and Fingers: When just five arcs are detected, it is easy to relate their order with the finger order. In other cases, i.e., when only $k < 5$ arcs are detected, we heuristically decide which finger corresponds to each arc detected. Here, we suppose that in the first frame the correspondence between arcs and fingers can be established.

As shown in Fig.2 (b), the five corresponded finger points F_i at $t-1$ are compared with arc points detected at t. C_j is defined as a finger set which corresponds to selected k fingers from $F_i^{t-1} : C_j = \{c_j^m, m = 1,2,\cdots,k\}, c_j^m \in \{F_i^{t-1}(x,y), i = 1,2,\cdots,5\}$. The best combination $C_{\min}$ is determined as follows.

$$C_{\min} : \arg\min_j \sum_{m=1}^{k} \| Arcpoint_m^t(x,y) - c_j^m \|$$

According to the best combination, which finger the arc corresponds to can be recognized. Our experiments show that using neighboring frames makes the corresponding quite robust.

3. Estimation of Undetected Arc Points: Since some significant arcs are sometimes missed, it is necessary to estimate them by other information. After establishing the correspondence between detected arcs and fingers, the relative position of undetected arc point to detected finger feature points can be recognized. Then the missing arc points can be estimated by utilizing the relation of two neighboring fingers in two neighboring frames as Fig.2(c). We define L_i^t as the bone line of finger i at t. Considering the fact that the angle between neighboring bone lines changes proportionately with the width of hand region, the unknown bone line can be computed:

$$\frac{\varphi(L_t^i, L_t^{i+1})}{\varphi(L_{t-1}^i, L_{t-1}^{i+1})} \propto \frac{width_t}{width_{t-1}} \tag{2}$$

The missing arc point refers to the intersecting point of the bone line and contour.

4 Global Pose Estimation

In our hand model, the wrist joint has 2 DOFs from yaw and pitch rotation of the palm while the elbow joint has 4 DOFs from translation and roll rotation. After fitting the real elbow position to the model, three rotation parameters of the global hand state are approximately calculated as follows.

4.1 Roll Angle

In our system, three cameras are used and Fig.2 (d) shows the camera positions and an example of the hand regions obtained by the three cameras. Since two cameras are selected based on the above method, in this example, camera1 and camera 2 are selected and the roll angle θ_{roll} is calculated from their images. Here, we assume that each camera projection can be approximated as weak perspective. Therefore, assuming that the width of the hand is x_0, the widths of the hand region x_i are represented in the following equations.

$$x_1 = k_1 x_0 \cos(\theta_{roll} - \alpha_1)\ ,\ x_2 = k_2 x_0 \cos(\theta_{roll} - \alpha_2)$$

Here k_i is a scaling parameter which relates to the position of the respective camera and α_i is the angle at which each camera is tilted to the upright. As a result, θ_{roll} can be derived from the equations:

$$\theta_{roll} = \tan^{-1}\frac{x_1 k_2 \cos\alpha_2 - x_2 k_1 \cos\alpha_1}{x_2 k_1 \cos\alpha_1 - x_1 k_2 \cos\alpha_2} \tag{3}$$

Where $k_1 = 0.95k_2$ is defined according to the distances from the cameras to the hand. When camera2 and camera3 are selected, θ_{roll} is similarly calculated.

Feature Point Correction based on the Roll Angle: In our presentation of the feature extraction, we have so far assumed that a palm and an arm each are a planar object. However, due to the thickness of the palm and the arm, the obtained non-finger feature points tend to have some errors according to the roll rotation of the hand. Therefore, currently we correct the estimated 2D positions of related feature points in the image which are seriously influenced by the roll rotation. In fact, the wrist position is approximately corrected by equation-4 and the correction of other feature points can be similarly defined.

$$wrist_{2D}' = wrist_{2D}(1 - \gamma \sin\theta_{roll}) \tag{4}$$

Here γ is a correction factor which depends on the thickness of the object.

4.2 Yaw and Pitch Angles

The yaw and pitch angles due to the rotation of the wrist are easily calculated from the 3D position of three non-finger feature points.

5 Model-based Hand Shape Estimation

Since a hand motion can be divided into global pose and individual finger motion, the 3D hand shape can be rendered if finger joint angles are all determined. Considering that the hand motion is highly constrained and the estimation complexity can be reduced by hand constraints [6, 7], we combine the constraints with IK for the estimation.

5.1 Hand Constraints

Limits of the Range of Finger Motions: The movement of each finger is limited by movable ranges of joints.

$$\Phi_{joint_k}^{finger^i} \in [\Phi_{low_k}^i, \Phi_{high_k}^i], \Psi_{MCP}^{finger^i} \in [\Psi_{low}^i, \Psi_{high}^i]$$

Where Φ refers to the limits of the finger motion as a result of flexion by each finger joint and Ψ refers to the limits of the adduction/ abduction (AA) motion by MCP joint. However, we neglect the AA motion of the thumb MCP joint. As a result, the thumb motion is characterized by 4 DOFs, being the same with other fingers.

Intra-Finger Constraints: The constraints are the ones between joints of the same finger. A commonly used constraint states that the DIP joints have a linearity relationship with the PIP joints on grasping motions $\theta_{DIP} = 2\theta_{PIP}/3$. Another one adopted in our estimation refers to MCP joints and PIP joints. As MCP joints get to bend, PIP joints will bend slowly at first and then do heavily with MCP joints in common flexion. Finally PIP points again bend slowly with MCP joints in strong flexion. In brief, the relations can be approximated by an S-curve. However, a human hand can extend MCP joints under PIP joints bended. Therefore, we assume that a PIP joint has a constant angle when an MCP joint angle is fewer than 10^0 .

5.2 Preliminary Finger Pose Estimation by IK

After estimating the global translation and rotation, we apply IK with hand constraints to estimate each joint angle of a finger. In general, IK is to determine joint angles θ_i of a manipulator, given the goal point of the manipulator. Here, the goal refers to a finger feature point and the target is a fingertip or a finger joint of the hand model.

To solve the IK problem, we use the Cyclic Coordinate Descent (CCD) method which works fine for hand structures [8]. By combining hand constraints to limit the search space, only few passes are enough to achieve a sufficient precision in many cases and therefore this method could be used in this real time application.

The correspondence between joints/fingertip and finger feature points is first determined, which offers the position of the end-effecter of the finger manipulator. Each finger feature point is identified using TIP to MCP joint as Fig.3 (a). The joint which has the least error will be selected as the right joint corresponding to the given finger feature point. As the process above, angles of parent joints of the target are determined. However, angles of its child joints can not be estimated by IK. To solve this problem, the intra-finger constraints are used to approximate the angles.

5.3 Model Parameter Refinement

For each finger, only 4 joint angles are unknown and the kinematics chain constrains the movement of each joint. Therefore, the joint angles could be estimated well by combining IK and hand constraints if the target and goal of IK are accurately given. However, the global parameters are approximately calculated. Accordingly, as the child of wrist, MCP joint has somewhat position error. Meanwhile, as the root of IK, MCP position error directly causes position error of the target of IK. Since it is difficult to correct the three global rotational angles efficiently, we try to adjust the wrist position and MCP joint position according to the error vector between goal and target. Then they can be refined by minimizing the error.

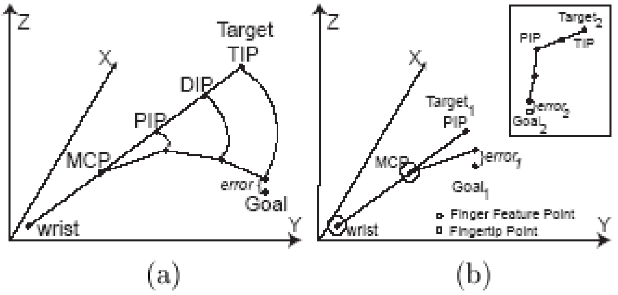
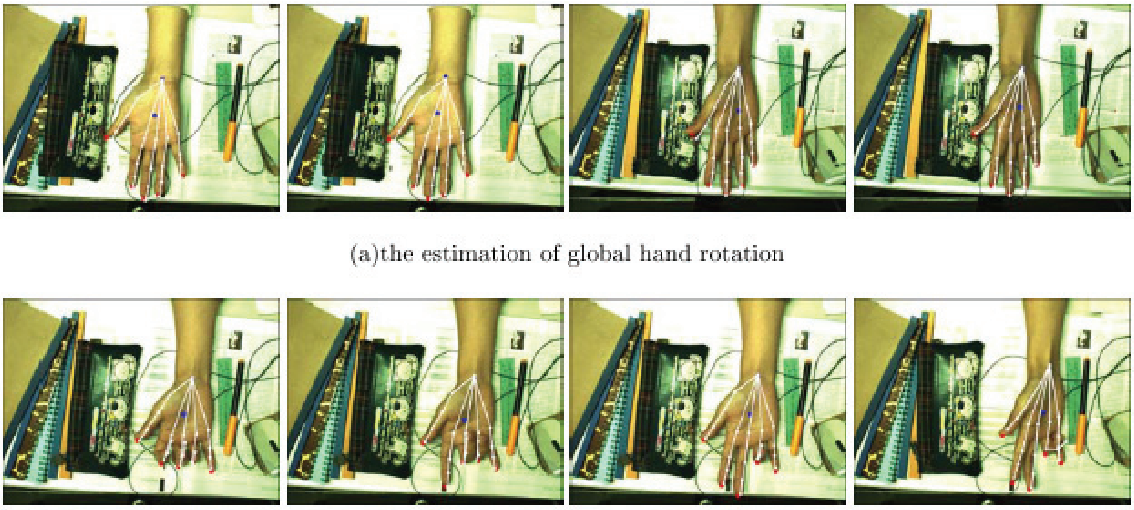


Fig. 3 (a) kinematics of the finger model, (b) updating the estimation

Furthermore, the results can be improved by two-stage estimation when finger feature point extracted is not the fingertip, e.g. PIP joint, as shown in Fig.3 (b). In this case, the fingertips extracted from edges of input image are given as the second goal of IK, and IK is recomputed while modifying the model parameters. Considering both the two error vectors helps to reach a global minimum.

6 Experiment

We demonstrate the effectiveness of our method by estimating a number of hand motions under cluttered background using 3 IEEE-1394-based color cameras (Point Grey Research Inc; Flea) with f:8 mm lenses, which are geometrically calibrated in advance. The images are captured with the size of 640×480 pixels. Fig.4 shows the results of the estimation using real images, with the skeleton of the estimated model projected to the original image. The experimental results indicate that the proposed method works successfully with real images, although the processing speed is still slow because of the color segmentation done for all cameras. The processing time is shown in Table.1 using PC with Pentium Ⅳ (3.2GHz). Currently, skin-color region extraction using OpenCV takes much time,



(a)the estimation of global hand rotation

(b)the estimation of hand motion with both finger articulation and global rotation

Fig. 4 the result of estimation

Table 1: processing time

Algorithm	Time
Convert image to HS color space	11*3msec
Detect skin color region	43*3msec
Extract hand features	6msec
IK calculation	26msec
Total	194msec

but it can be reduced easily by refining the program.

7 Conclusion

This paper proposes a novel method to estimate 3D hand shape by addressing three key issues: 1) to handle the global rotation of the hand and reduce the influence of the self-occlusion by multiple cameras and 2) to extract hand features robustly by utilizing relations between successive image frames and 3) to complement hand shape information by using IK combined with hand constraints. We have experimented with our proposed method under complex background. Although only three cameras are used in our experiment, it is easy to increase the number of cameras to handle larger variation of hand poses.

At present, there are two major problems in our system. One is the accuracy problem, which is caused by the approximated hand constraints. A more general representation of the hand constraints is required. The other problem refers that our 3D hand model can not deal with the hand of every person. Acquisition of the geometrical parameters of 3D hand model, such as lengths of finger bones, from the first pose is also an important issue. As future research, we are aiming to solve the problems above and build an estimation system for natural and practical human interfaces.

References

[1] B.Stenger, A. Thayananthan, P. H. S. Torr, and R. Cipolla, Filtering using a tree-based estimater, in ICCV, pp.1063-1070, 2003.

[2] N. Shimada and Y. Shirai, Hand Posture Estimation based on 2D Appearance Retrieval Using Monocular Camera, In Proc. Int, Workshop on RATFG-RTS, pp. 23-30, 2001.

[3] Etsuko Ueda, Yoshio Matsumoto, Masakazu Imai and Tsukasa Ogasawara, Hand Pose Estimation for Vision-based Human Interface, IEEE Transactions on Industrial Electronics, Vol.50, No.4, pp.676-684, August 2003.

[4] Shan Lu, Dimitris Metacas, Dimitirs Samaras, John Oliensis, Using multiple cues for hand tracking and model refinement, IEEE Conf. pp443-450, 2003.

[5] K. Schwerdt and JL Crowley, Robust face tracking using color, in Proc. of 4th International Conference on Automatic Face and Gesture Recognition, Grenoble, France, pp90-95, 2000.

[6] John Lin, Ying Wu and Thomas S. Huang, Modeling the Constraints of Human Hand Motion, in Proc. 5th ARL Symposium, Maryland, 2001.

[7] George ElKoura, Karan Singh Handrix: Animating the Human Hand, SIGGRAPH Symposium on Computer Animation, 2003.

[8] Wang, L.T. and Chen, C. C, A combined optimization method for solving the inverse kinematics problem of mechanical manipulators, IEEE J. Robotics and Automations, pp489-499, 1991.