Multi-view Appearance-based 3D Hand Pose Estimation

Haiying Guan, Jae Sik Chang, Longbin Chen, Rogerio S. Feris, and Matthew Turk

Computer Science Department
University of California at Santa Barbara, CA, USA
{haiying, jschang, lbchen, rferis, mturk}@cs.ucsb.edu

Abstract

We describe a novel approach to appearance-based hand pose estimation which relies on multiple cameras to improve accuracy and resolve ambiguities caused by self-occlusions. Rather than estimating 3D geometry as most previous multi-view imaging systems, our approach uses multiple views to extend current exemplar-based methods that estimate hand pose by matching a probe image with a large discrete set of labeled hand pose images. We formulate the problem in a MAP (maximum a posteriori) framework, where the information from multiple cameras is fused to provide reliable hand pose estimation. Our quantitative experimental results show that correct estimation rate is much higher using our multi-view approach than using a single-view approach.

1. Introduction

Hand gestures provide abundant degrees of freedom (DOF) and are nature and intuitive for communication. Such advantages make hand gestures a potential input modality for 3D (or even higher) free-form input interfaces. Moreover, gesture-based interfaces are especially suitable for large scale displays, 3D volumetric displays or wearable devices such as PDAs or cell phones. Visual sensing provides a passive and non-intrusive way for computers to acquire gesture input information. Vision-based hand gesture interpretation is therefore a promising technique for HCI and will definitely be part of next generation user interfaces.

In this paper, we address the particular problem of 3D hand posture estimation. Previous approaches to 3D hand pose estimation are generally classified into two categories: model-based and appearance-based approaches. Model-based methods [1] rely on estimating the parameters of a 3D hand model to fit a given hand image. Although they provide more precise estimation than appearance-based approaches, the high DOF of hand configurations impose a search in a high dimensional space, which requires good initialization and often leads to inaccurate local minima solutions. Appearance-based, or exemplar-based, or "Estimation-by-Synthesis" approaches estimate pose by matching a hand image with a discrete set of labeled hand pose images [2]. These methods in general rely on image retrieval approaches to obtain a set of K possible hand pose candidates (which are defined by joint angles and the viewpoints) corresponding to the probe image.

Although appearance-based approaches have received a great deal of attention recently [2] [3], they suffer from major problems such as self-occlusion and non-reliable feature extraction. In most of the hand images, fingers occlude each other and may be occluded by the palm. The hand is very flexible and its appearance changes dramatically from different viewpoints. If we consider the pose estimation in any given viewpoint, it is critical to resolve the occlusion problem in order to obtain the correct estimation. Also, reliable hand feature extraction is another challenge problem, especially when the hand is performed in complex backgrounds. Skin color is an important cue for hand detection; however, for posture recognition more detailed features are needed.

Our approach addresses these challenges and extends the framework of previous appearance-based approaches [2]. For previous monocular exemplar-based systems, if the side view of the hand is captured, it is far less sufficient for us to figure out the actual hand configuration. Figure 1 shows an extreme case. Given the side view of a static gesture, it is unlikely to figure out which count gesture (1, 2, 3 or 4) it actually is. Due to this reason, our algorithm utilizes multi-view images to resolve the ambiguities caused by occlusions. A MAP framework is proposed to estimate the hand posture given a set of images captured from multiple views. An active contour extraction algorithm is also presented to extract the hand shape contour and the hand regions from noisy backgrounds. Differently from other multi-view pose estimation algorithms [4] [5], our approach bypasses the 3D model reconstruction and utilizes the camera viewpoint constraints to estimate the hand pose with the MAP framework. The experimental results give the quantitative comparison of the retrieval performance using a sin-

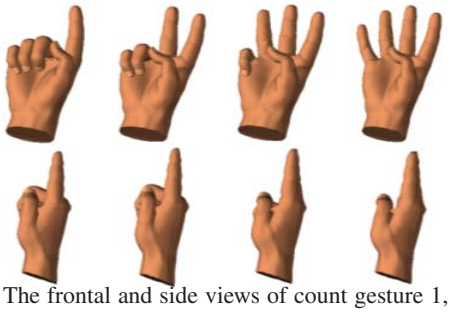

Figure 1. The frontal and side views of count gesture 1, 2, 3 or 4 with the similar side views

gle image and two-view images. It shows that the retrieval rate is greatly increased with two-view images.

This paper is organized as follows: The related works are discussed in Section 2. Section 3 proposes a multi-view pose estimation using a MAP framework. Section 4 describes the transformation matrices among frames. Section 5 focuses on hand contour feature extraction based on active contour model and a shape context descriptor. Experimental results are shown in Section 6. Finally, conclusions and future work are discussed in Section 7.

2. Related work

Athitsos and Sclaroff [2] formulated the pose estimation problem to a database indexing or image retrieval problem. The hand is considered as an articulated object with high DOF. In order to estimate joint angles and viewpoint, a large number of exemplars need to be considered. Given a single hand image from a cluttered background, their algorithm successfully provides a set of possible 3D hand postures and the corresponding camera viewpoints. Zhou and Huang [3] proposed an Okapi-Chamfer matching algorithm for the fast indexing given a single hand image. Generally, the appearance-based frameworks are restricted by the viewpoint. If the side view is given, the retrieval rate decreases. Rehg et al. [6] try to resolve the ambiguity caused by the occlusion, which occurs when the motion is directed along the viewing axis of a single camera during the hand tracking or estimation.

Because single-view algorithm are limited by the occlusion problem, several multi-view algorithms are also proposed using 3D hand reconstruction. Ueda et al. [4] propose a method to reconstruct the hand pose as a "voxel model" from silhouette images of the hand obtained form the multi-view camera system. The major problem of the system is how to precisely represent the finger details and reduce the modeling errors which causes the wrong estimation. The system is tested on a simple background. Delamarre and Faugeras [5] propose a method to estimate the pose of a hand in a sequence of stereo images. They try to fit a 3D finger model to a real finger model reconstructed by the disparity map. In general, reconstruction the 3D hand model by stereo matching is a tough problem since hand region has less texture, in addition, the depth among fingers or between finger and palm are relatively small. Hamada et al. [7] presents a hand posture estimation method from silhouette images taken by multiple cameras. A transition network is build for hand pose matching. The system is tested on hand images with a uniform background.

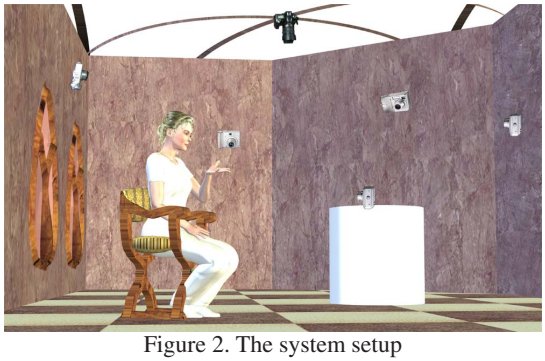

Figure 2. The system setup

In our paper, we try to reduce the influence of self-occlusion by capturing the hand image from different viewpoints. Because it is a changeling task to obtain the 3D hand model (or voxel model) with the precise finger details, we bypass the 3D reconstruction and propose a MAP framework to fuse multi-view 2D hand images information. Furthermore, an efficient hand contour extraction algorithm is presented, which is applicable to noisy background with different lighting conditions.

3. Multiple viewpoints pose estimation using MAP framework

In this section, a MAP framework is given to estimate the top K nearest matches by combining the multi-view information. Figure 2 illustrates the system setup, where multiple cameras could be installed at any location, and the system captures multiple hand gesture images from different viewpoints. According to Bayesian rule [8], given the input images from multiple view points observations; that is, a set of the captured images, $I = \{I_1, I_2, ..., I_n\}$, we would like to maximize the posterior probability and obtain the most likely K hand state (which contains hand configuration and the global rotation in the world coordinate) retrieved from the synthesis database.

Our synthesis dataset contains a set of N hand states, $S = \{s_1, s_2, ..., s_N\}$. If cameras are calibrated in the world coordinate, these N states also determine the hand configuration and the camera viewpoints. A set of images are rendered according to these states, which are clustered into a set of L model images, $I^M = \{I_1^m, I_2^m, ..., I_L^m\}$. In another words, for a given hand state, we define I^m as the image of the hand (at a specific hand configuration) generated at a particular camera viewpoint.

Here, we purposely use two variables to represent the model state, s, and the corresponding image, I^m , respectively. The reason is that different model poses could generate the similar model images, so the probability distribution

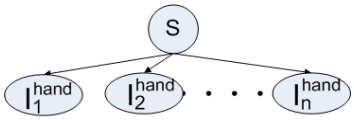

Figure 3. The causal relationship

of variables, s and I^m are not the same. The sizes of the two sets, S and I^M are not necessary the same also.

The Bayesian equation is given in Eq. 1.

$$P(s|I_1, I_2, ..., I_n) = \frac{P(I_1, I_2, ..., I_n|s)P(s)}{P(I_1, I_2, ..., I_n)}. \quad (1)$$

If we only consider the hand region $I_i^h \in \{I_1^h, I_2^h, ..., I_n^h\}$ in the images, which could be extracted by object extraction algorithms, the relationship among the images $I_1^h, I_2^h, ..., I_n^h$ and the state; that is the hand configuration with the global rotation in the world coordinates, could be described by diverging connections in the belief graph (Figure 3). Given the state, we thus assume $I_1^h, I_2^h, ..., I_n^h$ are conditionally independent with each other. The posteriori probability of s given I^h is given as follows:

$$P(s|I^h) \propto \{P(I^h|s)P(s) \approx (\prod_i^n P(I_i^h|s))P(s)\} \quad (2)$$

, where $I_i^h \in \{I_1^h, I_2^h, ..., I_n^h\}$ represents the hand region image. The objective of the research is to find the optimal subset of the S composed of K high rank states with the posteriori probability given by Eq. 2.

For each individual $I_i^h, i = 1, 2, ..., n$, we have

$$\begin{aligned} P(I_i^h|s) &= \frac{P(I_i^h, s)}{P(s)} = \frac{1}{P(s)} \sum_{I_j^m} P(I_i^h, I_j^m, s) \\ &= \frac{1}{P(s)} \sum_{I_j^m} P(I_i^h, I_j^m|s)P(s) \\ &= \sum_{I_j^m} P(I_i^h|I_j^m, s)P(I_j^m|s). \end{aligned} \quad (3)$$

We assume given I_j^m, variable I_i^h and s are conditionally independent and obtain the following equation:

$$P(I_i^h|I_j^m, s) = P(I_i^h|I_j^m) \quad (4)$$

Thus, Eq. 3 is simplified to

$$P(I_i^h|s) = \sum_{I_j^m} P(I_i^h|I_j^m)P(I_j^m|s). \quad (5)$$

Therefore, the posteriori probability in Eq. 2 is shown as

$$P(s|I^h) \propto \prod_i^n \{\sum_{I_j^m} P(I_i^h|I_j^m)P(I_j^m|s)\}P(s). \quad (6)$$

The likelihood $P(I_i^h|I_j^m)$ in Eq. 6 can be defined as

$$P(I_i^h|I_j^m) = \frac{1}{T} \exp(-Dist(I_i^h, I_j^m)) \quad (7)$$

, where T is a normalization factor to make Eq. 7 be a probability distribution. $Dist(I_i^h, I_j^m)$ is the distance of the observation image I_i^h and the synthesis image I_j^m in feature space. The feature used in this paper a variance of shape context feature descriptor obtained from hand contour. It is a scale invariant vector with 256 components. Please refer [9] for detail information.

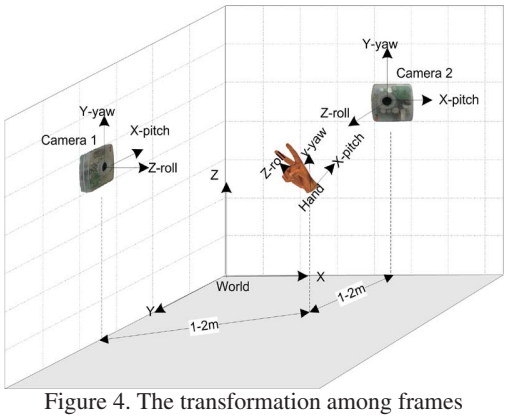

Figure 4. The transformation among frames

The $P(I^m|s)$ can be shown as

$$P(I^m|s) = w_{I^m} = \frac{P(I^m, s)}{P(s)} \quad (8)$$

, where s follows a the uniform distribution and $P(s)$ is a constant; and each discrete probability value of $P(I^m|s)$ is estimated from the synthesis dataset,

$$P(I_j^m \in I^M | s_k \in S) = \frac{\#\ of\ the\ cases\ with\ both\ I_j^m\ and\ s_k}{\#\ of\ the\ s_k}. \quad (9)$$

Let Eq. 6 take a minus logarithmic function, then the maximization of a posteriori probability problem can be formulated as the minimization of the following posteriori energy function:

$$E(s) = \sum_i^n \sum_{I_j^m} (w_{I_j^m} Dist(I_i^h, I_j^m)) - \log P(s). \quad (10)$$

Since $P(s)$ is a constant, the final energy equation is simplified to:

$$E(s) = \sum_i^n \sum_{I_j^m} (w_{I_j^m} Dist(I_i^h, I_j^m)). \quad (11)$$

Finally, our objective is to find K states that minimize the given energy in Eq. 11.

4. Homogeneous transformations among camera, hand and world frames

The correlations of the global hand transformation with camera 1 and camera 2 are obtained by the homogenous transformations among the cameras, the hand and the world frames. Without lost of generality, Figure 4 shows the case of two cameras and the hand in the world space [1].

The transformation between the coordinate frames can be derived from the combined homogenous transformation matrices [10]. Let ${}^{frame_1}T_{frame_2}$ represents the homogenous transformation from $frame_2$ to $frame_1$, we have the following equations:

[1] The frame coordinates here are defined for easy understanding, but in general, they are not limited to horizontal or vertical direction. The transformation among the cameras and the hand could be any arbitrary rotations and translations. The equation given here is suitable for any arbitrary transformations.

$$^{World}T_{Hand} = {}^{World}T_{Cam_1} \cdot {}^{Cam_1}T_{Hand} \quad (12)$$

$$^{World}T_{Hand} = {}^{World}T_{Cam_2} \cdot {}^{Cam_2}T_{Hand} \quad (13)$$

From Eq. 12 and Eq. 13, the relationship of the transformation between hand and two cameras are given below.

$$^{Cam_1}T_{Hand} = {}^{Cam_1}T_{Cam_2} \cdot {}^{Cam_2}T_{Hand} \quad (14)$$

, where

$$^{Cam_1}T_{Hand} = \begin{bmatrix} I_{3\times3} & Trans(d_x,d_y,d_z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} RPY(\phi_z,\phi_x,\phi_y) & 0 \\ 0 & 1 \end{bmatrix} \quad (15)$$

, where

and

$$Trans(d_x, d_y, d_z) = [d_x, d_y, d_z]^T \quad (16)$$

$$RPY(\phi_z^{roll},\phi_x^{pitch},\phi_y^{yaw}) = Rot(z,\phi_z)Rot(x,\phi_x)Rot(y,\phi_y). \quad (17)$$

$RPY(\phi_z^{roll}, \phi_x^{pitch}, \phi_y^{yaw})$ represents the relative global rotation between the hand and the camera. They correspond to the three global rotation parameters of the hand pose. The transformation matrix from Cam_2 to Cam_1 could be obtained by camera calibration [11]. If the scale invariant feature is adopted, we can neglect the scale factor in all homogeneous matrices. However, if we use translation invariant feature, we cannot neglect the translation factor since the translation will cause viewpoint changes. In such case, the hand depth information obtained by stereo algorithms could be used to estimate the translation matrix.

5. Hand contour extraction in noisy background

In this paper, we assume that a given image is composed of two regions: hand region and background region, and each region is characterized by a probability distribution $P(I|\alpha)$, where I is a color value and α is a parameter of the distribution. For hand probability distribution $P(I|\alpha_o)$, we use skin-color information which is represented by a 2D-Gaussian model in normalized $r-g$ space. In the RGB space, color representation includes both color and brightness. Therefore, RGB is not necessarily the best color representation for detecting pixels with skin color. Brightness can be removed by dividing the three components of a color pixel (R, G, B) according to intensity and this space is known as normalized $r-g$ space. The model parameters are chosen manually. Unlike in the hand region, we use a constant value as the background probability distribution $P(I|\alpha_b)$ to discriminate foreground from background.

To extract hand contour, we use an active contour model based on level sets. Given an initial curve manually selected, the curve is evolved until it finds the accurate hand contour in a region competition framework [12]. In this case, the curve evolution can be a special case of region competition in which there are two regions (hand region R_o and background region R_b) and a common boundary $\vec{v}$ as shown in following motion equation:

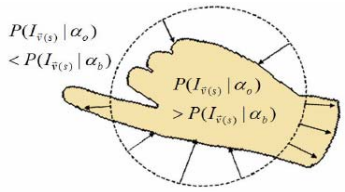

Figure 5. The motion of the curve

$$\frac{d\vec{v}(s)}{dt} = -\mu k_{\vec{v}(s)}\vec{n}_{\vec{v}(s)} + (\log P(I_{\vec{v}(s)}|\alpha_o) - \log P(I_{\vec{v}(s)}|\alpha_b))\vec{n}_{\vec{v}(s)} \quad (18)$$

, where $k_{\vec{v}(s)}$ is the curvature of $\vec{v}$ at point $\vec{v}(s)$ and $\vec{n}_{\vec{v}(s)}$ is the outer unit normal to $\vec{v}$ at point $\vec{v}(s)$. The motion of the curve is shown in Figure 5. Besides the smoothing term $(-\mu k_{\vec{v}(s)}\vec{n}_{\vec{v}(s)})$, the motion of v is determined by the likelihood ratio test. If $P(I_{\vec{v}(s)}|\alpha_o) > P(I_{\vec{v}(s)}|\alpha_b)$, the curve $\vec{v}$ moves outer normal($\vec{n}_{\vec{v}(s)}$) direction. Otherwise the curve moves inner normal($-\vec{n}_{\vec{v}(s)}$) direction.

The curve evolution was implemented using the level set method. We represent curve $\vec{v}$ implicitly by the zero level set of function $u : \Re^2 \rightarrow \Re$, with the region inside $\vec{v}$ corresponding to $u > 0$. Accordingly, Eq. 18 can be rewritten by a level set evolution equation [13]:

$$\frac{du(S)}{dt} = -\mu k_s \|\nabla u\| + (\log P(I_s|\alpha_o) - \log P(I(s)|\alpha_b))\|\nabla u\|. \quad (19)$$

Figure 6 shows the hand contour extraction results in different background with different lighting conditions.

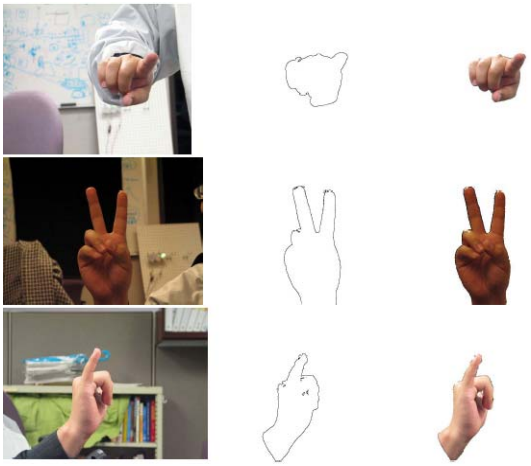

Figure 6. The hand contour extraction

6. Experimental results

6.1. System setup

We use two cameras to capture the hand pose image simultaneously. The two cameras and the hand are nearly in the same horizontal plane. In order to reduce the occlusion area of the hand, two cameras angles relative to hand is as far as possible (close, but not limited to $90°$). Figure 4 illustrates the system setup, the distance between the cameras

and the hand is around $1 \sim 2$ meters. The precise transformation matrix of the two cameras is obtained by camera calibration [11].

6.2. Synthesis dataset

Our synthesis database contains 15 commonly used hand gestures. These 15 gestures are defined at the semantic level. For each hand gesture, small noises are added to their basic hand configurations to create another two hand configurations for the same gesture. In the synthesis dataset, 448 viewpoints (roll pitch, and yaw) are sampled from the surface of the 3D view sphere. Totally, the synthesis dataset contains $448(viewpoints) * 3 * 15(hand\ configurations) = 20160$ images and $224 * 3 * 15 = 10080$ hand states.

For each hand configuration, 57 joint angle parameters are saved as the joint angle vectors. These 57 parameters include 6 global rotation translation parameters for the hand, 3 rotation and 3 translation parameters for the part of forearm, 9 parameters (3 rotation parameters for 3 joints respectively) for each finger and thumb. We fix the 3 rotation and 3 translation parameters for the part of forearm and the 6 global rotation translation parameters for the hand and add the camera rotations to encode the viewpoint changes.

For each synthesis image in the dataset, the hand contour and hand region are extracted. The probability distribution of $P(I^{model}|state)$ are calculated with Eq. 8 and Eq. 9.

6.3. Test dataset

We collect a real hand image dataset with the similar 15 hand gestures. 7 states relative to the world coordinate frame are captured for each gesture. For each state, two sets of images are captured by two cameras from different viewpoint respectively. Each set contains $1 \sim 3$ images of the similar state. The real hand image database totally contains 254 cases. Each case contains two images from Camera 1 and Camera 2 respectively. We label the pseudo-ground truths (the state which contains the hand configuration and the global rotation with respect to the world coordinate frame) by manually identifying the similar images in the synthesis database and assigning its hand configuration and global rotation parameters to the real image. In order to compare with the single image retrieval algorithm, we also label the pseudo-ground truths for each image by manually identifying the similar image in the synthesis database and assigning its hand configuration and global rotation parameters to the real image.

For each real image, the hand contour extraction algorithm presented in Section 5 is executed to extract the hand contour. A variation of shape context descriptor presented in [9] is used to represent the contour. The cost function given in Eq. 7 is evaluated based on the shape context descriptor and the posterior probability in Eq. 11 is evaluated for each state in the synthesis dataset. The set of possible candidates corresponding to the top K maximum probabilities is retrieved.

6.4. Comparisons the retrieval results based on single-view and multi-view algorithm

In our experiments, for the single retrieval algorithm, we define the correct match if the hand gesture and pose of the real image is the same as one of the hand gestures in the top K retrieved images with the same three global hand rotation parameters (roll, pitch, yaw), where K is decided by the application requirements. For the two-view retrieval algorithm, we use the same criterion; that is, we define the correct match if the hand state of the real images is the same as one of the hand states in the top K retrieved states, where, K is decided by the application requirements.

Using such criterion, we compare the performances of the retrieval algorithm using the image captured by Camera 1 only, by the Camera 2 only, and by both of them with the MAP framework (Section 3). The retrieval algorithms of Camera 1 only and Camera 2 only are implemented by directly comparing the query image with each image in the synthesis dataset and retrieving the top K best match. In the dataset, there are 3 cases satisfy such condition, which have the same global rotation and similar hand configuration with small noises. As long as one of them is appeared in the top K, the retrieval is considered to be successful. For the two-view retrieval algorithm, we use the same criterion. For a given hand state, there are three states similar to it in the dataset also. All the 254 cases are tested and the retrieval rate is obtained. The performance of retrieval algorithms (with a single camera and two cameras) in Table 1 shows that the retrieval rate with two cameras almost double compared with a single camera.

Table 1. The comparison of retrieval rates (%) using the hand images captured by Camera 1 only, by Camera 2 only and by both of them

K	50	100	150	200	250	300
Cam. 1 only	19.69	30.71	36.61	39.76	44.88	49.61
Cam. 2 only	14.57	23.23	29.13	32.68	37.40	40.55
Both	48.03	61.42	69.69	74.80	78.74	81.50

The results of the MAP framework with two-view images are shown in Figure 7. The first two images in the first row are the query images captured by two cameras. The next two images in the first row are the synthesis images generated by their corresponding pseudo-ground truth. These two viewpoints satisfy the relationship defined by Eq. 14. The remaining 12 pair images are the top 12 retrieval results from the synthesis dataset. It shows that for this particular case, the pseudo-ground truth is successfully retrieved by the 3^{rd}, 4^{th} and the 5^{th} pair [2], which are all correct match cases contained in the dataset.

[2] Although it is hard to distinguish the difference between them from the images, they are not duplicate cases because some parameters of their hand configurations are different by 10°.

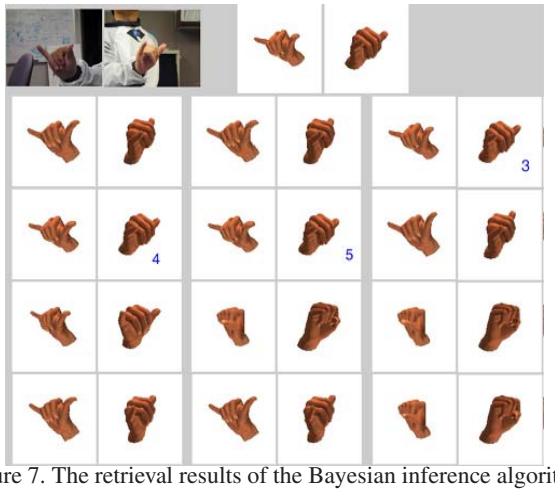

Figure 7. The retrieval results of the Bayesian inference algorithm with two-view images

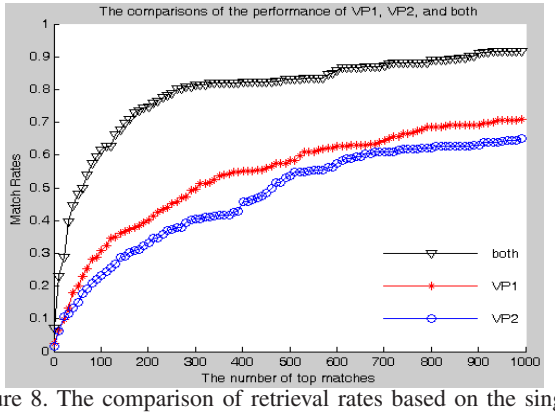

Figure 8. The comparison of retrieval rates based on the single-view algorithm (VP 1 or VP 2) and two-view algorithm (both)

Figure 8 draws the relationship between the retrieve rates and the retrieve number, K. It shows that the correct match rate of the MAP framework using two viewpoint images increases dramatically compared to the retrieval rate only using the single viewpoint image. It also shows that the retrieval results of the images captured by Camera 1 only (VP 1) are higher than Camera 2 only (VP 2). The main reason is that the dataset of VP 1 contains more "frontal" view of the hand than dataset of VP 2, which contains more "side" view. This indication is also consistent with the results given in [2] (Table 2).

7. Conclusions and future work

Static hand gesture recognition is a highly challenging task due to the large number of DOF of the hand kinematic structure. One technology bottleneck for vision-based gesture estimation lies in the occlusion problem. In this paper, we propose a multi-view gesture estimation algorithm. Differently from previous gesture recognition methods, our method doesn't need the 3D model reconstruction. The quantitative experimental results with the clutter backgrounds show that the MAP framework based on two-view images greatly improves the retrieval performance from a single camera.

Currently, we only use the contour feature, some important hand pose information, such as internal hand finger edges, hand texture, hand region, are not considered. In future, more efficient features are expected for better retrieval performance.

References

[1] J. Lee and T. L. Kunii, "Model-Based analysis of hand posture," *IEEE Computer Graphics and Applications*, vol. 15, no. 5, pp. 77–86, 1995. 1

[2] V. Athitsos and S. Sclaroff, "Estimating 3D hand pose from a cluttered image," in *Computer Vision and Pattern Recognition (CVPR)*, pp. 432–439, June 2003. 1, 2, 6

[3] H. Zhou and T. Huang, "Okapi-chamfer matching for articulated object recognition," in *Tenth IEEE International Conference on Computer Vision (ICCV'05)*, vol. 2, pp. 17–20, 2005. 1, 2

[4] E. Ueda, Y. Matsumoto, M. Imai, and T. Ogasawara, "Hand pose estimation using multi-viewpoint silhouette images," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 4, pp. 1989 – 1996, 2001. 1, 2

[5] Q. Delamarre and O. Faugeras, "Finding pose of hand in video images: a stereo-based approach," in *Automatic Face and Gesture Recognition*, pp. 585 –590, Apr 1998. 1, 2

[6] J. Rehg, D. D. Morris, and T. Kanade, "Ambiguities in visual tracking of articulated objects using two- and three-dimensional models," *International Journal of Robotics Research*, vol. 22, pp. 393 – 418, June 2003. 2

[7] Y. Hamada, N. Shimada, and Y. Shirai, "Hand shape estimation using sequence of multi-ocular images based on transition network," in *VI 2002*, 2002. 2

[8] F. V. Jensen, *Bayesian Networks and Decision Graphs*. Springer-Verlag, 2001. 2

[9] R. Feris, M. Turk, R. Raskar, K. Tan, and G. Ohashi, "Exploiting depth discontinuities for vision-based fingerspelling recognition," in *IEEE Workshop on Real-time Vision for Human-Computer Interaction (in conjunction with CVPR'04)*, 2004. 3, 5

[10] S. B. Niku, *Introduction to Robotics Analysis, Systems, Applications*. 3

[11] C. Mei, "Camera calibration toolbox for matlab," Aug. 2005. 4, 5

[12] S. C. Zhu and A. Yuille, "Region competition: unifying snakes, region growing, and bayes/mdl for multiband image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 884 –900, Sept. 1996. 4

[13] A. R. Mansouri, "Region tracking via level set pdes without motion computation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 947 –961, Sept. 2002. 4