3D Hand Pose Reconstruction with ISOSOM

Haiying Guan and Matthew Turk

Department of Computer Science, University of California, Santa Barbara, CA 93106
{haiying, mturk}@cs.ucsb.edu

Abstract. We present an appearance-based 3D hand posture estimation method that determines a ranked set of possible hand posture candidates from an unmarked hand image, based on an analysis by synthesis method and an image retrieval algorithm. We formulate the posture estimation problem as a nonlinear, many-to-many mapping problem in a high dimension space. A general algorithm called ISOSOM is proposed for nonlinear dimension reduction, applied to 3D hand pose reconstruction to establish the mapping relationships between the hand poses and the image features. In order to interpolate the intermediate posture values given the sparse sampling of ground-truth training data, the geometric map structure of the samples' manifold is generated. The experimental results show that the ISOSOM algorithm performs better than traditional image retrieval algorithms for hand pose estimation.

1 Introduction

Despite the rapid advances in computing communication and display technologies, the development of Human Computer Interaction (HCI) still lags behind. Gesture is a good candidate for the next generation input devices. It has the potential ability to relieve the interaction bottleneck between users and the computer. Vision-based gesture interpretation is a promising research area for this problem due to its passive and non-intrusive sensing properties.

Many approaches of 3D hand pose estimation to support gesture recognition may be classified into two categories: model-based approaches with 3D data [1] and appearance-based approaches with 2D data [2] [3]. Athitsos *et al.* [4] formulated the problem of hand pose estimation to a problem of image database indexing. Shimada *et al.* [5] generated 125 possible candidate poses with 128 view points with a 3D model of 23 degrees of freedom (DOF). The real input hand image was matched to pre-computed models with a transition network, and possible pose candidates were found for their hand tracking system.

In this research, we take an image retrieval approach based on analysis by synthesis method. It utilizes a 3D realistic hand model and renders it from different viewpoints to generate synthetic hand images. A set of possible candidates is found by comparing the real hand image with the synthesis images. The ground truth labels of the retrieved matches are used as hand pose candidates. Because hand pose estimation is such a complex and high-dimensional problem, the pose estimate representing the best match may not be correct. Thus, the retrieval

G. Bebis et al. (Eds.): ISVC 2005, LNCS 3804, pp. 630–635, 2005.

is considered successful if at least one of the candidates in top N matches is sufficiently close to the ground-truth (similar to [4]). If N is small enough, with additional distinguishable contextual information, it may be adequate for automatic initialization and re-initialization problems in hand tracking systems or sign language recognition systems, where the correct estimation could be found and the incorrect ones could be eliminated in the later tracking.

The hand is modeled as a 3D articulated object with 21 DOF of the joint angles (hand configuration) [1] and 6 DOF of global rotation and translations [1]. A hand pose is defined by a hand configuration augmented by the 3 DOF global rotation parameters. The main problem of analysis by synthesis is the complexity in such a high dimension space. The size of the synthesis database grows exponentially with respect to the parameter's accuracy. Even though the articulation of the hand is highly constrained, the complexity is still intractable for both database processing and image retrieval. Wu *et al.* [6] and Zhou *et al.* [7], for example, reduced the dimensionality to 6 or 7 based on data collected with the data glove.

In this paper, we formulate hand pose reconstruction as a nonlinear mapping problem between the angle vectors (hand configurations) and the images. Generally, such mapping is a many-to-many mapping in high dimension space. Due to occlusions, different hand poses could be rendered to the same images. On the other hand, the same pose is rendered from the different view points and generates many images. To simplify the problem, we eliminate the second case by augmenting the hand configuration vector with the 3 global rotation parameters. The mapping from the images to the augmented hand configurations becomes a one-to-many mapping problem between the image space and the augmented hand configuration space (the hand pose space). The dimensionality of image space can be reduced by feature extraction. Finally, we establish the one-to-many mapping between the feature space and the hand pose space with the proposed ISOSOM algorithm. The experimental results shows that our algorithm is better than traditional image retrieval algorithms.

The paper is organized as follows. The ISOSOM algorithm is proposed in Section 2. The experimental results are shown in Section 3. Finally, the conclusions are given in Section 4.

2 ISOSOM

Instead of representing each synthesis image by an isolated item in the database, the idea of this research is to cluster the similar vectors generated by similar poses together and use the ground-truth samples to generate an organized structure in low dimension space. With such structure, we can interpolate the intermediate vector. This will greatly reduce the complexity. Based on Kohonen's [8] Self-Organizing Map (SOM) and Tenenbaum's ISOMAP algorithm [9], we propose an ISOmetric Self-Organizing Mapping algorithm (ISOSOM). Instead of

[1] The translation parameters could be estimated by hand segmentation algorithms or neglected if translation and scale invariant features are adopted.

organizing the samples in the 2D grids by Euclidian distance, it utilizes the topological graph and geometric distance of the samples' manifold to define the metric relationships between samples and enable the SOM to follow better the topology of the underlying data set. The ISOSOM algorithm compresses information and automatically clusters the training samples in a low dimension space efficiently.

2.1 ISOSOM Initialization, Training, and Retrieval

Although the ISOSOM algorithm is robust with respect to the initialization, the appropriate initialization allows the algorithm to converge faster to the solution. Before the training, initial vectors associated with neurons on the ISOSOM map are linearly interpolated by the sample nodes of the manifold's topological graph.

We generate the topological graph of the manifold using the approach described in the first two steps of the ISOMAP algorithm. We define the graph G over all data points by connecting node i and j if i is one of the k nearest neighbors of j. We set the edge lengths equal to the Euclidean distance of i and j. On the graph, the distance of any two nodes is defined by the cost of shortest path between them. This distance is approximately the geometric distance on the manifold. Such distance preserves the high-dimension relationship of the samples in low dimension space.

The neurons of the ISOSOM map are connected with their neighbors on the low dimension manifold. In each training step, we randomly choose a sample vector from the training samples and find its Best-Matching Unit (BMU) in the ISOSOM map. Instead of using Euclidean distance, we measure the similarities of the input sample vector with each neuron vector in the ISOSOM map by the geometric distance between the two nodes on the manifold of the training samples. In order to measure this similarity, the nearest nodes of input vector and the neuron vector on the topological graph are retrieved by the Euclidean measurements. The shortest path between the two retrieved nodes on the manifold graph is approximated as the distance of the input vector and the neuron vector of the ISOSOM map. The BMU is the neuron on the ISOSOM map with the smallest geometric distance to the input training vector.

After obtaining the BMU, its associated vectors and its topological neighbors on the ISOSOM map are updated and moved closer to the input vector in the input space as in the classical SOM algorithm.

Given a vector with full components or partial components, similar neurons are found and sorted by the similarity measurement described above. The mask is used for the vector with partial components and the similarity measurements are also modified to handle such cases.

2.2 ISOSOM for Hand Pose Reconstruction

Due to the projection and the feature extraction, feature samples of the different poses are highly overlapping in the feature space. In order to separate the mixed-up features, we form large vectors consisting of both feature vectors and their

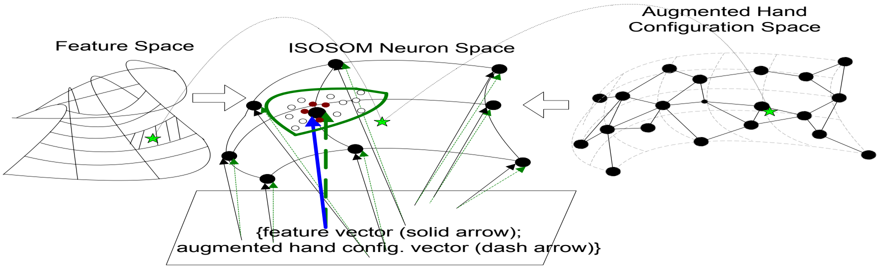
Feature Space
ISOSOM Neuron Space
Augmented Hand Configuration Space
{feature vector (solid arrow);
augmented hand config. vector (dash arrow)}

Fig. 1. The ISOSOM for Hand Pose Reconstruction

corresponding hand pose vectors. These vectors are used as training samples in our algorithm. In other words, after training, each neuron of the ISOSOM is associated with two vectors: the feature vector and the hand pose vector. Figure 1 gives an intuitive depiction of the ISOSOM map.

In the retrieval step, for a given input hand image, we calculate its feature vector. Using a mask to handle the missing hand pose components, we compare the similarity of this feature vector with all feature vectors associated with the ISOSOM neurons. The possible candidates of the hand pose are retrieved by the top n best matches. Because the mapping from the feature space to the hand pose space is a one-to-many mapping, one feature vector could have several possible hand pose candidates. This is desirable because it reflects the intrinsic nature of the mapping. The confidence of each candidate is also measured by the error measurement of ISOSOM.

3 Experimental Results

We generate a training database containing 25 commonly used poses. Three camera parameters (roll: $0° - 360°$, pitch: $-90° - 90°$, and yaw: $0° - 360°$, interval: $36°$) control the global rotation of the camera viewpoints. For each pose, 726 images are rendered in different view points and there are totally 18150 synthesis images in the database. For each hand configuration, 48 joint angle parameters are saved as the joint angle vector, which are 3 rotation parameters for hand, 9 parameters (3 rotation parameters for 3 joints respectively) for each finger and thumb. In addition to the 3 global rotation parameters of the camera, the hand pose vector is composed of these 51 parameters.

In the experiments, we adopt traditional Hu moments features [10] to represent given images. It should be pointed out that ISOSOM is a general algorithm for nonlinear mapping and dimension reduction, it does not specify the features. The following experimental shows that even with less descriptive Hu features, our algorithm still can achieve good estimation results.

We generate another synthesis database containing the same 25 poses but more dense samples for the testings. Because Hu features are invariant to in-plane rotation, translation, and scale, we focus on pitch and yaw rotation of

Table 1. The comparisons of the reconstruction performance by the traditional retrieval algorithm, the SOM and the ISOSOM algorithms. (The number of images measured: 21525.)

The percentages of hits of the similar hand pose (except roll rotation)									
	Threshold 10			Threshold 20			Threshold 40		
Number	IR	SOM	ISOSOM	IR	SOM	ISOSOM	IR	SOM	ISOSOM
Top 40	23.21%	16.13%	36.73%	36.25%	34.27%	59.97%	47.93%	55.64%	78.01%
Top 80	27.13%	23.20%	48.71%	44.48%	44.52%	72.11%	58.36%	67.03%	86.85%
Top 120	29.28%	28.63%	55.77%	49.14%	50.97%	78.82%	63.69%	73.04%	91.64%
Top 160	31.41%	32.39%	61.14%	53.17%	55.37%	83.32%	68.39%	76.28%	94.33%
Top 200	33.35%	35.28%	65.26%	56.98%	58.62%	86.13%	72.78%	78.64%	95.56%

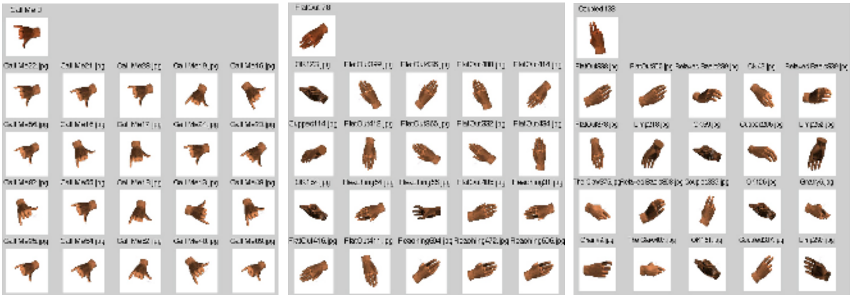

Fig. 2. The ISOSOM retrieval results (The name of each image indicates the hand configuration. The number for query image is the index number in the query dataset. The number for retrieved image is the index number in the ISOSOM neuron graph.)

the camera. Instead of using 36° as a interval, we use 9° as a interval for the pitch and yaw rotations. In this dense sampling database, 21525 hand images are generated as testing data set. It turns out that 92.33% of the testing images in the dense sampling database cannot find the exact match in the sparse sampling database (the training data set). Because the training samples are from the sparse sampling training set and the testing samples are from the dense sampling set, the difficulty for retrieving the correct hits increases.

We compare the performance of the traditional image retrieval algorithm (IR), SOM, and ISOSOM in Table 1. For a given testing image in the dense set, we count a hit if there are one or more hand poses in the retrieved poses which satisfy two criteria: first, they have a similar hand configuration as the query image; that is, the hand configuration's parameter components changes within a small range without changing the physical meaning of the posture. Second, the global rotation parameters are close to the query image within evaluation threshold 10°, 20° and 40° respectively. Table 1 shows that the performance of ISOSOM is the best among the three algorithms. Compared to the traditional image retrieval algorithm for the top 40 matches, the hit rates of ISOSOM increase around $13\% - 31\%$. This also indicates that the ISOSOM algorithm not only has the clustering ability, but also has interpolation ability. The ISOSOM retrieval results are shown in figure 2. The first image is the query image. The rest 20 images are the retrieval results from the ISOSOM neurons.

4 Discussion and Conclusion

We have investigated a nonlinear mapping approach for 3D hand pose estimation from a single image. Traditional image retrieval algorithms just compare the image features in feature space and retrieve the top matches. Our approach utilizes both the feature vectors and their corresponding augmented hand configuration vectors to avoids the feature overlapping problem in nature. To deal with the complexity, we reduced the redundancy by clustering the similar feature vectors generated by similar poses together and represented them by neurons in our low dimension ISOSOM map. The ISOSOM algorithm could be considered as a variant of the SOM algorithm which aims at enabling SOM to follow the better topology of the underlying data set. The experimental results confirm that the ISOSOM algorithm greatly increase the hit rates in the pose retrievals.

References

1. Lee, J., Kunii, T.L.: Model-Based analysis of hand posture. IEEE Computer Graphics and Applications **15** (1995) 77–86.
2. Rosales, R., Athitsos, V., Sclaroff, S.: 3d hand pose reconstruction using specialized mappings. In: International Conference on Computer Vision (ICCV). (2001) 378–385.
3. Nolker, C., Ritter, H.: Visual recognition of continuous hand postures. IEEE Transactions on Neural Networks **13** (2002) 983–994.
4. Athitsos, V., Sclaroff, S.: An appearance-based framework for 3d hand shape classification and camera viewpoint estimation. In: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, Washington, DC, USA, IEEE Computer Society (2002) 45–50.
5. Shimada, N., Kimura, K., Shirai, Y.: Real-time 3d hand posture estimation based on 2d appearance retrieval using monocular camera. In: IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems. (2001) 23–30.
6. Wu, Y., Lin, J.Y., Huang, T.S.: Capturing natural hand articulation. In: ICCV. (2001)
7. Zhou, H., Huang, T.S.: Tracking articulated hand motion with eigen dynamics analysis. In: Ninth IEEE International Conference on Computer Vision. (2003) 1102–1109.
8. Kohonen, T.: Self-Organizing Maps. Springer Series in Information Sciences (2001)
9. Tenenbaum, J.B., Silva, V.D., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science (2000) 2319 –2323.
10. Hu, M.K.: Visual pattern recognition by moment invariants. IRE Trans. on Information Theory (1962) 179 –187.