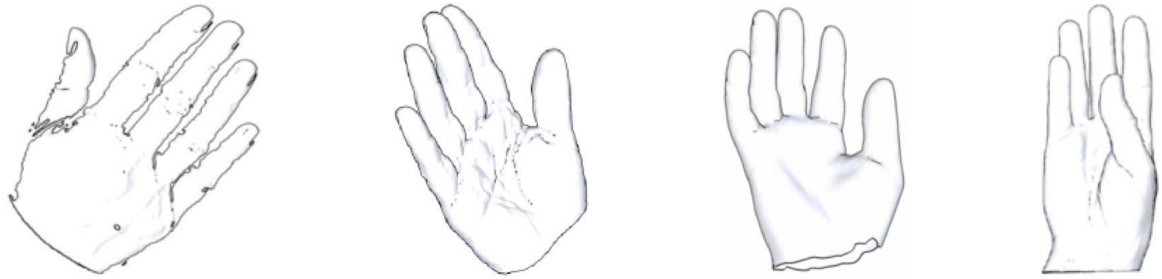
Prototype-based Intraclass Pose Recognition of Partial 3D Scans

Jacob MONTIEL* Hamid LAGA† Masayuki NAKAJIMA‡

Graduate School of Information Science and Engineering, Tokyo Institute of Technology*‡
Global Edge Institute, Tokyo Institute of Technology†



(a) (b) (c) (d)

Figure 1: *In the pose recognition problem we want to measure the degree of similarity between a partial scan (a) and full prototypes (b), (c) and (d). Our method is able to regard the pose of (a) as similar to those of (b) and (c), while differentiating it from (d).*

Abstract

We propose a new algorithm for recognizing the pose of partial scans of objects of the same class such as hands. We formulate the recognition problem as a problem of matching the partial scan to a set of prototypes, each one representing a key pose. The key poses are first indexed offline using a set of local descriptors. Next, given a partial range scan of an object, which can be acquired by any 3D sensor such as stereo or Time Of Flight (TOF) cameras, we start by computing a set of local descriptors, in the same manner as the offline process. The recognition is based on estimating the similarity between the descriptors of the scan and the descriptors of the prototypes. We introduce a comparison scheme that allows the estimation of the similarity between a partial scan and a full 3D model. This allows also an accurate localization of the partial scan onto the full 3D model. Our experiments show that the algorithm is able to: (1) find similar key poses for a given partial scan even in the presence of subtle changes only, and (2) it casts aside the models that have different pose than the partial scan.

Keywords: 3D shape matching, pose recognition, similarity measurement, partial 3D scan, shape context.

1 Introduction

Recognizing objects from 2D images has been extensively investigated in the field of computer vision and pattern recognition, yet it remains a challenging task as the loss of the third dimension at the acquisition stage makes the recognition prune to ambiguities and errors. Recent advances in 3D technologies have given computers the ability to sense the world in 3D. This opens new perspectives to take common applications in Computer Vision that have been constrained to the 2D realm to upper dimensions.

3D Object Recognition from 2.5D data such as partial scans is a challenging task in Computer Vision. It is useful in applications such as scan registration, robot navigation, and surveillance systems. It involves recognizing the shape and the pose given a partial scan of the object. In this paper we focus on intra-class recognition of shapes that undergo rigid as well as non-rigid transformations. We propose a new method for recognizing the pose of partial scans of the same class of objects. The method uses a database of 3D models as prototypes. In our experiments we focus on partial scans of hands but the approach extends easily to other shape classes.

The rest of the paper is organized as follows: Sections 1.1 and 1.2 review the related work and outline the main contributions of the paper. Section 2 describes our algorithm for intra-class pose recognition. Experimental results are presented in Section 3. We conclude in Section 4.

1.1 Related work

Object recognition is a well studied subject in Computer Vision. Existing methods can be classified according to the representation used in: (1) methods based on 2D images and (2) methods working on 3D models.

In the 2D space, appearance-based methods use the brightness value of each pixel in the image. Recognition is typically accomplished by template matching, standard pattern recog-

*e-mail: jacob@img.cs.titech.ac.jp
†e-mail: hamid@img.cs.titech.ac.jp
‡e-mail: nakajima@img.cs.titech.ac.jp

nition, and neural networks. Main limitations of such approaches is sensibility to illumination/pose variations. On the other hand, feature-based approaches work directly with information concerning the global structure of the object. Here, shape recognition is based on the spatial configuration of features such as silhouette, edges and keypoints/landmarks. Regional point descriptors have shown good results in 2D. The Shape Context Descriptor [Mori et al. 2005] is a popular approach due its robustness to noise and invariance to translation, scale and rotation. Nevertheless, the loss of depth information and the lack of a view-invariant representation limits the application of such 2D based approaches.

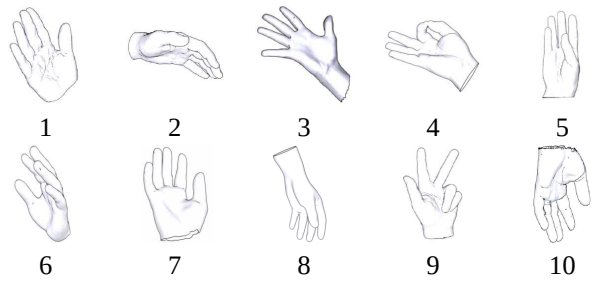
In contrast, 3D methods are view-invariant. Moreover, current developments in 3D sensors have made easier to recover the spatial structure of real objects. Some approaches work exclusively in the 3D space [Pekelny and Gotsman 2008], while others combine 2D and 3D information. Some approaches to 3D object recognition use markers, model templates [Vlasic et al. 2008] , medial axis representation, skeletons [Au et al. 2008] and spherical harmonics. Main limitations for our purpose are the difficulty to work with subtle shape deformations and inadequacy to match partial representations of the objects. On the other hand, the Spin Image [Johnson and Hebert 1999] and 3D Shape Context [Frome et al. 2004] descriptors cope with such limitations. Both approaches can match partial scans with full 3D models in cluttered scenes. Spin-images enable the use of 2D Image Processing and Matching techniques while Shape Context use surface information to describe the shape. Main differences with our method, are that these approaches do not allow partial deformations such as those due to articulated parts of the object. Moreover, while our method uses an intraclass approach to the recognition problem, they focus on finding only one specific object in the database, which makes the size and search space of the database restrictive.

1.2 Overview and contributions

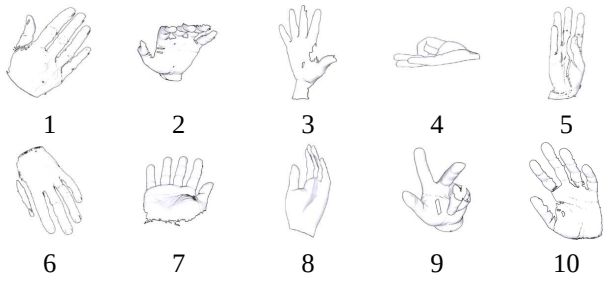
In this paper we focus on the problem of intra-class recognition. Particularly, given a partial scan S_p and a set of full 3D models $C = \{S_i, i = 1, \ldots, n\}$ sampled from the same class of shapes C, the goal is to find the shape $S_i \in C$ that is most similar to S_p. The underlying assumption is that the objects in the database as well as the partial scan S_p have similar intrinsic geometric properties but may undergo rigid as well as non-rigid transformations.

Our proposal to solve this problem consists of:

- **Building a database of prototypes:** we collect a set of 3D models representing key poses of the class of objects we are considering. In this paper we focus on different poses of human hands, but the approach can be extended in a straight forward manner to other classes of shapes.
- **Offline shape indexation:** we describe each shape in the database with a set of local descriptors in the same manner as in the *Bag of Words (BoW)* approach, but carefully sampled in order to capture the main properties of the shapes.
- **Online scan processing:** given a partial scan as query, we extract a set of local descriptors in the same manner as in the offline procedure. We propose a new similarity estimation scheme that allows robust matching of partial scans to 3D models in the presence of subtle shape variations due to: (1) articulated parts, such as fingers, in the objects, and (2) the presence of noise and holes which are very common when dealing with range data.



(a) Examples of 3D models in our database.



(b) Examples of partial scans used as query.

Figure 2: *Examples of prototype models (database samples) and partial scans. The database contains hand models from adult male and female subjects including left and right hands with different poses.*

The proposed method allows the matching in the presence of large pose variations, discarding dissimilar shapes, and also localizing the partial scan on the most similar 3D model.

2 Intra-class Pose Recognition

Given a range scan S_p of an articulated object, we are interested in recognizing its pose. This can be formulated as an intra-class 3D shape matching problem: consider a class of shapes $C = \{S_i : i = 1, ..., n\}$ where S_i are instances of the same object that differ only in the pose (e.g. hand poses), and a given partial scan S_p, the pose of the scan S_p is the pose of the shape S_i that is the most similar to S_p.

We use local descriptors to index the prototype models and the partial scans (Section 2.1). For efficient matching we propose a new similarity estimation scheme (Section 2.2) which is able to deal with subtle changes and is invariance to translation, rotation and scale.

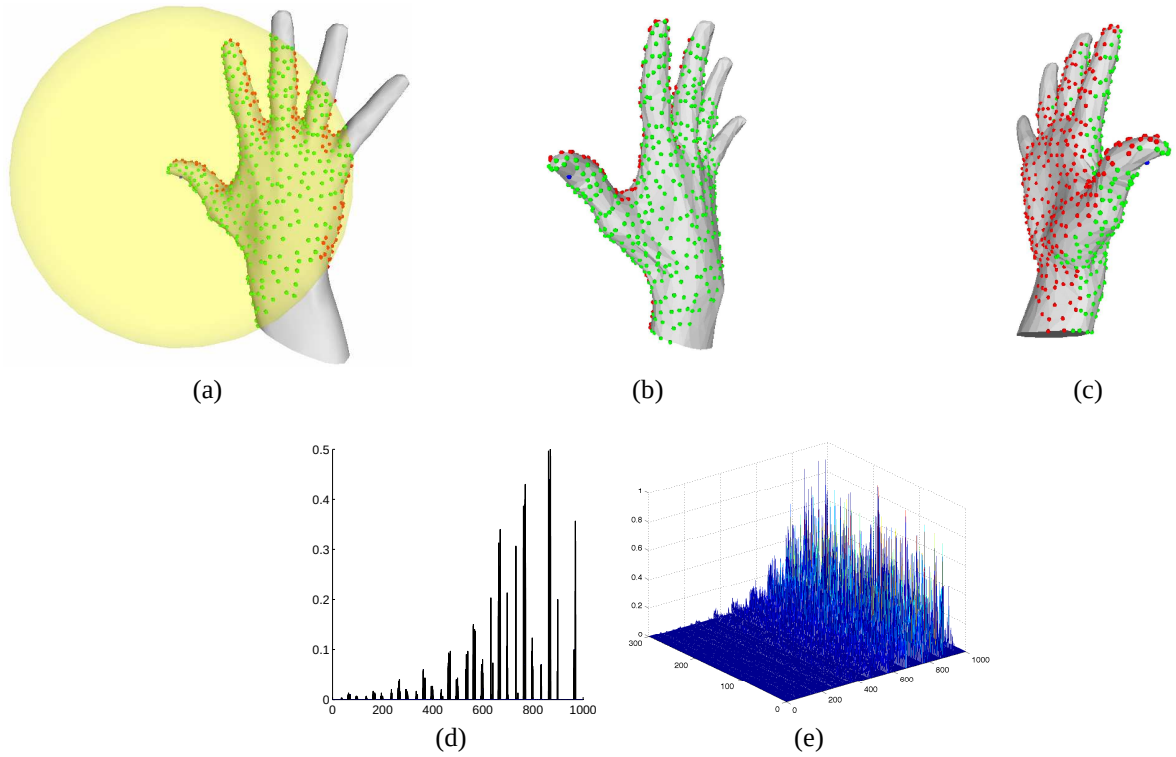


Figure 3: *The shape context descriptor computed at a certain location (the blue point on the thumb). (a) The sphere around the feature point. (b) The green points inside the sphere are accounted in the descriptor, while (c) the red points are discarded since they have been recognized as belonging to the other side of the shape. (d) The descriptor computed at the feature point. (e) The descriptor of the entire model which is a collection of local descriptors computed at the sampled feature points.*

2.1 Shape Description

Similarity is an important abstract concept in human perception. Its definition is a semantic question [Bronstein et al. a]. The similarity function should capture the main shape characteristics which depend on the type of models, the intended application, and the users.

In the hand recognition problem the similarity can be defined based on features such as color, texture, or size. In our case, we are interested in the shape and pose. To capture these features we use local shape descriptors computed at several locations on the shape surface. Using local descriptors allows surface correspondence, intermediate and high-level feature detection, and shape segmentation. In addition it allows robust matching in cluttered scenes.

There are several robust local shape descriptors that have been proposed for matching 3D shapes. Spin images [Johnson 1997] and shape contexts [Frome et al. 2004; Mori et al. 2005] are among the most popular and extensively used in many applications. Particularly, a shape context is a geometric histogram that describes the surface properties around a given location. Shape contexts allow partial matching and are robust to noise. Using the shape context descriptor, the shape matching problem is reduced to sampling, normalization and comparison of probability distributions.

In the first step, we select feature points that are the most representative of a region in the scan and model. This reduces the number of points necessary to compare the shapes. We sample points using quadric error metrics which guarantees that the chosen feature points represent its neighbors accurately. Then we compute a shape context descriptor at these locations. This procedure is executed offline on the models in the database and online when processing a partial 3D scan.

To build the 3D shape context, we consider a sphere of radius r centered at a point on the shape surface and is oriented using the normal of the point as the *z-axis*. The sphere is then divided into bins with logarithmically spaced shells in the radius direction and evenly spaced sectors in the elevation θ and azimuth ϕ directions. Figure 3 illustrates the shape context. Figure 3(c) shows the probability distribution obtained for a given object, computed at a specific location. Table 1 shows the values used in our experiments, where δ corresponds to the mean Euclidean distance between the feature points.

At the end of this step, every prototype model S_i in the database is represented with a set of M shape context descriptors. Similarly given a query scan S_p, we extract $K, K \leq M$ local descriptors and use them to find the most similar prototype in the database.

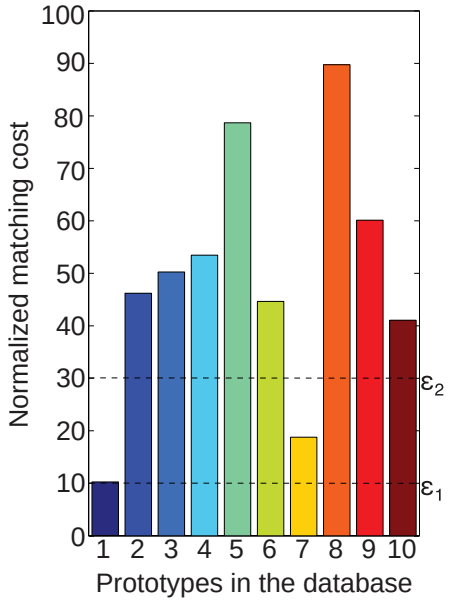


Figure 4: *Similarity measurement. Pose similarity between a prototype and a partial scan increases inversely with the cost. These graph corresponds to the cost of matching the partial scan showed in Fig. 1(a) against the prototypes in the database. The similarity cost for Fig. 1(b) (Prototype 1) indicates that it has the same pose, while Fig. 1(c) (Prototype 7) has similar pose. The rest have different poses, Fig. 1(d) corresponds to the Prototype 5.*

2.2 Similarity estimation

Once we have computed shape descriptors from the partial scans, the next step is to compare them with the ones from the 3D Models in our database. To measure the similarity of two shapes we compare the descriptors of a partial scan S_p and a 3D model S_i by the following equation:

$$E(S_p, S_i) = \sum_{k \in 1,\ldots,K} \min_{m \in 1,\ldots,M} \chi_2(q_k, p_m) \qquad (1)$$

where q_k and p_m are points on the query and on the reference shape respectively. χ_2 is the Chi-square similarity metric. The pose of the partial scan S_p is similar to a 3D model S_i when $E(S_p, S_i) \approx 0$. The best match is the reference model that minimizes the energy function E.

In practice, to measure the similarity degree between partial scans and full models, we define two thresholds ϵ_1 and ϵ_2. The pose of S_p is then considered:

- Same as the pose of S_i if $E(S_p, S_i) < \epsilon_1$,
- Similar to the pose of S_i if $\epsilon_1 \leq E(S_p, S_i) \leq \epsilon_2$,
- Different from the pose of S_i if $E(S_p, S_i) > \epsilon_2$.

2.3 Efficient shape context computation

Often partial range scans represent only the parts of the objects that are visible to the acquisition sensor. The 3D shape context descriptor when computed from a 3D model captures both sides of the shape. However, when computed on a partial scan, the other side is often missing and therefore cannot be captured by the descriptor.

To make the shape descriptor more suitable for scan to shape comparison we introduce the concept of the *descriptor's support angle* α as follows; Suppose we have an oriented point p with normal vector $\vec{n_p}$ for which we are creating a shape context. Consider another point q in the object with normal vector $\vec{n_q}$. The support angle constraint states that q will contribute to the computation of the shape context at p if and only if:

$$\langle \vec{n_p}, \vec{n_q} \rangle < \alpha \qquad (2)$$

where $\langle \cdot, \cdot \rangle$ is the inner dot product and $\alpha \geq 0$. Using this constraint we ensure that the information stored in the descriptor is the most likely to appear in a partial scan. The support angle is only used when computing the descriptors of the prototypes in the database. The descriptors of the partial scans are computed without considering it. In this paper we use an angle of ± 90 degrees in our experiments.

2.4 Implementation issues

Processing time is an important aspect to consider in the implementation. The complexity of computing one shape context is in the order of $O(n)$, where n is the total number of vertices. However, the resolution of the 3D models can be in the order of thousands or million of points, making the descriptors computationally expensive. Our We use spatial data structures such octrees for nearest neighbor search. This reduces the computation complexity of one descriptor to a logarithmic time. Furthermore, the support Angle constraint reduces the time necessary for computing the descriptors of the prototypes.

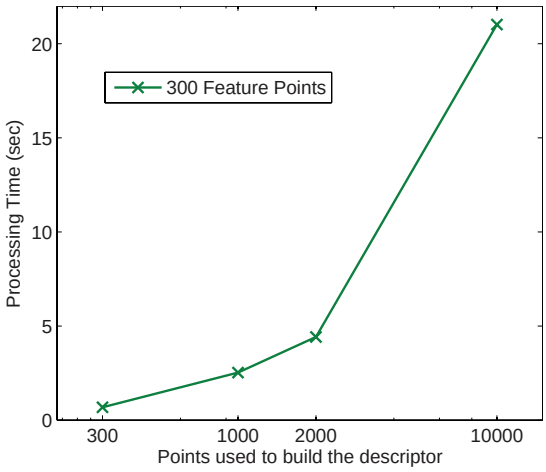
Finally, we experimented with different resolutions to build the descriptors: $300, 1000, 2000,$ and 10000 points for the 3D models, and $50, 100, 300, 600,$ and 3000 for the 3D scans.

3 Results

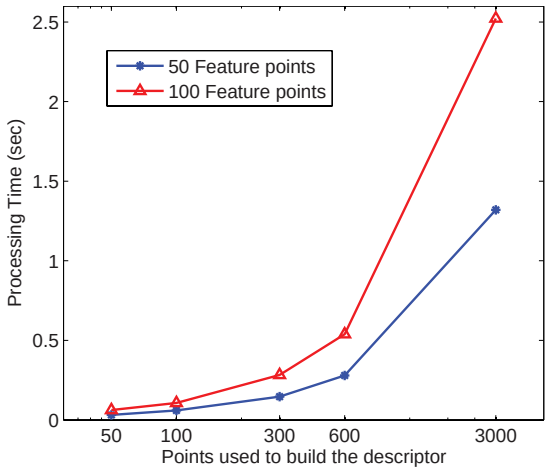
The database used in our experiments contains ten full 3D models of hands taken from the AIM@SHAPE repository, and 20 partial scans. The database include models of male-female, left-right hands, and different poses. Prototypes in

Table 1: *Descriptor Configuration. δ is the mean Euclidean distance between the feature points.*

Parameter	Value
$radius_{max}$	2.5δ
$radius_{min}$	0.1δ
r_{div}	10
θ_{div}	10
ϕ_{div}	10



(a) 3D Model



(b) Partial scan

Figure 5: *Variation of the processing time with respect to the size, in number of vertices (N), of (a) the prototype models in the database, and (b) the partial scans. The horizontal axis indicates the number of vertices while the* y-axis *is the processing time in seconds. Our experiments show that using $M = 300$ feature points with models of size $N = 1000$ vertices, and $K = 100$ feature points with partial scans of size $N = 300$ is a good compromise between matching precision and processing time.*

the database were tested using four different resolutions while for the partial scans we used five different resolutions. The number of feature points used was $M = \{300\}$ and $K = \{50, 100\}$.

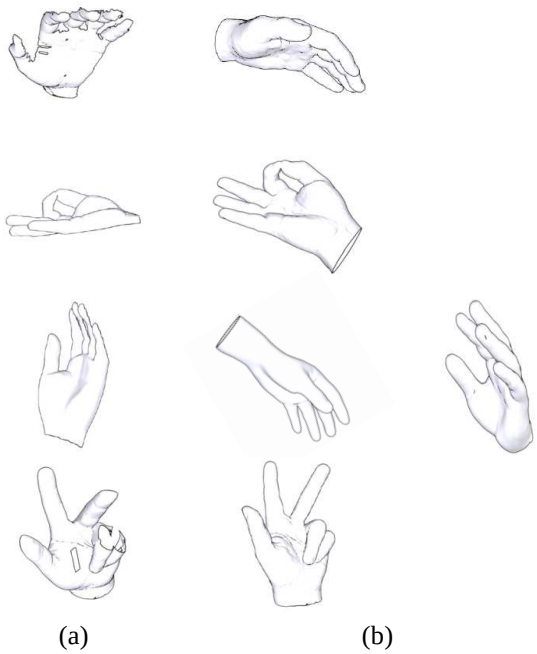
Table 2 and the corresponding graph plot of Figure 5 show the time required for computing the descriptors. In Table 2 δ is the mean Euclidean distance between the feature points. As shown in Figure 5(a), the time necessary to estimate the descriptors of the prototypes is in the order of seconds while the time for partial scans is below 1 second for most of the configurations. Although a high number of points is required for obtaining dense histograms, our experiments show that the probability distribution is not significantly affected.

We found experimentally that using M=*300* with p=*1000* and K=*100* with q=*300* provides a good compromise between processing time and the similarity measurement. The processing time for the partial scan is less than 0.3 seconds.

Figure 4 shows the estimated similarity between the partial scan of Figure 1(a) and all the ten prototype objects in the database. The lowest cost, bin 1 of the graph, corresponds to the prototype of Figure 1(b) and is classified as the same as the range scan. Notice also that the distance between the same partial scan and the prototype of Figure 1(c), corresponding to the bin 7, falls in the range of ϵ_1 and ϵ_2. It is then classified as similar. However, the distance between the partial scan and the prototype of Figure 1(d), depicted by the bin 5, is

Table 2: *Processing Time (sec).*

Pts. in the object	Number of Feature Points		
	50	100	300
50	0.031	0.062	-
100	0.060	0.107	-
300	0.147	0.282	0.683
600	0.279	0.538	-
1000	-	-	2.526
2000	-	-	4.414
3000	1.320	2.251	-
10000	-	-	21.02



(a) (b)

Figure 6: *Samples of the pose recognition results. (a) Partial scans. (b) Best matches retrieved, pose similarity decreases to the right.*

higher than the threshold ϵ_2. This means that the scan and the prototype are different.

Experimentally, we found that $\epsilon_1 = 10\%$ and $\epsilon_2 = 30\%$ are a good compromise for our purpose. The choice of ϵ_2 is highly related to the degree of similarity between key poses in the database. Other matching results are shown in Figure 6, where the objects in column (a) are the partial scans and in (b) are the retrieved prototypes.

These results were obtained on a MacBook Pro with MacOS 10.5.6, 2.2 GHz Intel Core 2 Duo and 2GB SDRAM.

4 Conclusion

We have introduced a new approach to the intraclass object recognition problem. Based on the shape context descriptor, our algorithm is able to match partial 3D scans with full models in the database. The algorithm is able to handle subtle changes due to rigid and non-rigid deformations of articulated parts of the object, such as fingers. An important feature of our approach is the use of keyposes from different subjects, this means that it is not necessary to have in the database the model that corresponds to a given partial scan for the algorithm work. We experimented with hand shapes but the approach can be easily extended to other classes of shapes. Our system is fully automatic and does not require user interaction.

One limitation of the current approach is that it does not handle efficiently symmetric objects. We are currently considering this issue. We plan (1) to experiment with larger databases by adding more hand keyposes and also other shape classes, (2) find the best matching regions for localizing partial scans on the 3D model and (3) measure the partial similarity [Bronstein et al. a] between scans and prototypes .

Acknowledgments

This research is carried out by the support of the JSPS Grant-in-Aid for Scientific Research Wakate-B Number 21700096.

References

AU, O. K.-C., TAI, C.-L., CHU, H.-K., COHEN-OR, D., AND LEE, T.-Y. 2008. Skeleton extraction by mesh contraction. *SIGGRAPH '08: ACM SIGGRAPH 2008 papers*, 1–10.

BRONSTEIN, A., BRONSTEIN, M., BRUCKSTEIN, A., AND KIMMEL, R. Partial similarity of objects, or how to compare a centaur to a horse. *International Journal of Computer Vision*.

BRONSTEIN, A., BRONSTEIN, M., AND KIMMEL, R. Topology-invariant similarity of nonrigid shapes. *International Journal of Computer Vision*.

CHUA, C.-S., AND JARVIS, R. 1997. Point signatures: A new representation for 3d object recognition. *International Journal of Computer Vision 25*, 1, 63–85.

CNR, V. C. L. I. Meshlab. http://meshlab.sourceforge.net/.

FROME, A., HUBER, D., KOLLURI, R., BULOW, T., AND MALIK, J. 2004. Recognizing objects in range data using regional point descriptors. *Proceedings ECCV 3*, 224–237.

JOHNSON, A. E., AND HEBERT, M. 1999. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Trans. Pattern Anal. Mach. Intell. 21*, 5, 433–449.

JOHNSON, A. 1997. *Spin-Images: A Representation for 3-D Surface Matching*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

KING, B. J., MALISIEWICZ, T., STEWART, C. V., AND RADKE, R. J. 2005. Registration of multiple range scans as a location recognition problem: Hypothesis generation, refinement and verification. *In Proceedings of the Fifth Intl. Conf. on 3D Digital Imaging and Modeling*.

LI, H., SUMNER, R. W., AND PAULY, M. 2008. Global correspondence optimization for non-rigid registration of depth scans. *Comput. Graph. Forum 27*, 5, 1421–1430.

MITRA, S., AND ACHARYA, T. 2007. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics 37*, 3 (May), 311–324.

MORI, G., BELONGIE, S., AND MALIK, J. 2005. Efficient shape matching using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell. 27*, 11, 1832–1837.

PEKELNY, Y., AND GOTSMAN, C. 2008. Articulated object reconstruction and markerless motion capture from depth video. *Computer Graphics Forum 27* (April), 399–408.

RUGIS, J., AND KLETTE, R. 2006. Surface registration markers from range scan data. *IWCIA06*, 430–444.

VELTKAMP, R. C., AND HAGEDOORN, M. 2001. *State of the art in shape matching*. Springer-Verlag, London, UK, 87–119.

VLASIC, D., BARAN, I., MATUSIK, W., AND POPOVIĆ, J. 2008. Articulated mesh animation from multi-view silhouettes. *ACM Trans. Graph. 27*, 3, 1–9.

WANG, S., WANG, Y., JIN, M., GU, X., AND SAMARAS, D. 2006. 3d surface matching and recognition using conformal geometry. *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on 2*, 2453–2460.

ZHANG, H., SHEFFER, A., COHEN-OR, D., ZHOU, Q., VAN KAICK, O., AND TAGLIASACCHI, A. 2008. Deformation-driven shape correspondence. *Computer Graphics Forum (Special Issue of Symposium on Geometry Processing 2008) 27*, 5.