3—25

Robust Object Recognition based on Regular Framing and Depth Aspect Image

Tomoyuki Takeguchi*
Graduate School of Engineering
Hokkaido University

Shun'ichi Kaneko†
Graduate School of Engineering
Hokkaido University

Abstract

A method of model-based object recognition for a cluttered depth scene including multiple objects is proposed. A novel model representation, named *depth aspect image*, is also defined as an orientation standardized appearance from the original depth data of objects, which is transformed through tuples of three barycenters defined within regularly defined voxels. A robust matching scheme, named *least quantile of residuals*, can achieve not only object recognition with depth aspect images but also verification with candidate models. The sparsely distributed barycenters and the robust matching make the ICP-based rough registration and the following verification process much faster and more reliable. In this paper, we show recognition experiments on 100 scenes contained multiple objects from a library of 4 models.

1 Introduction

Object recognition and position estimation of given models in objective scenes have been one of the important technologies in the area of robotics and automations. To deal with complex range data including multiple objects, segmentation helps finding the known models [1], but accurate localization of free-form shapes is generally difficult. Recently, some approaches have been proposed for free-form shapes, which have novel data representations based on local structures for object recognition. *Splash* represents relations of surface normals by the three-dimensional segments [2]. *Point signature* is a curve which expresses depths of surfaces from the tangent plane of a reference point [3]. In these methods, the surfaces around a reference point are accumulated along a three-dimensional curve. Another representation called *spin image* [4] using intensity images is proposed, and as an advanced version, *spherical spin image* [5] is also proposed. Matching with these representations finds the relation between points of the model and ones of the scene. Then some pairs of them are required to obtain sound congruent transformations. They are almost fine point-wise representations causing large computation and the resolution of representation is strongly dependent on the one of row range data. So we propose a novel representation called *depth aspect image* [7], which is a controllable two dimensional representation with local depth distribution and cooperated with a distinct coordinate scheme called *regular framing*.

Furthermore, robust algorithms have been proposed and applied to image matching with noises and outliers [9], and such ill-conditions are occurred due to occlusion, clutter and non-overlap parts in depth aspect images. Then we introduce a robust statistical estimator called *least quantile of residual*, which can be utilized for not only the image matching but also the three-dimensional model verification.

References for constructing representations of range data are generally selected from vertices or some geometric features in range data. Since the features have rather sparse data structures than the vertices in general, feature-based algorithms can achieve low computational cost, but these depend on the capability of reappearance [8]. Then we introduce regular framing, which can derive regularly arranged points for reference from range data automatically. Moreover, the sparsely distributed references are useful for the verification.

The rest of our paper is composed as follows: The proposed method is outlined in Section 2, including definitions of voxels and depth aspect images. Matching and verification through depth aspect images and a robust estimator are explained in Section 3, and then how to use ICP algorithm in the proposed method is also denoted. Section 4 shows experiments and analysis of them. Section 5 concludes the paper.

2 Representation

2.1 Outline

Fig.1 shows an outline of the proposed method which is divided into two processes: model learning and object recognition. In the learning process, a set of model range data is segmentalized into voxels by regular framing. It is standardized by a local three-dimensional coordinates, which is called *aspect coordinate frame*, hereafter ACF in short. According to the ACF, the model range data is projected into a *depth aspect image*, hereafter DAI. Multiple model DAIs construct a database, which contains all possible DAIs with corresponding ACFs for each models. In the recognition process, a DAI of the scene is created in the same way of leaning and matched with DAIs in the database. A candidate model is chosen from the database according to the optimality and the transformation between them is calculated from the pair of their ACFs. ICP algorithm can make correspondences between the candidate and the scene range data. And then the correspondence is robustly verified by evalu-

*Address: Kita 13 Nishi 8, Kita-ku, Sapporo 060-8286 Japan. E-mail: take@mee.coin.eng.hokudai.ac.jp

†Address: Kita 13 Nishi 8, Kita-ku, Sapporo 060-8286 Japan. E-mail: kanekos@coin.eng.hokudai.ac.jp

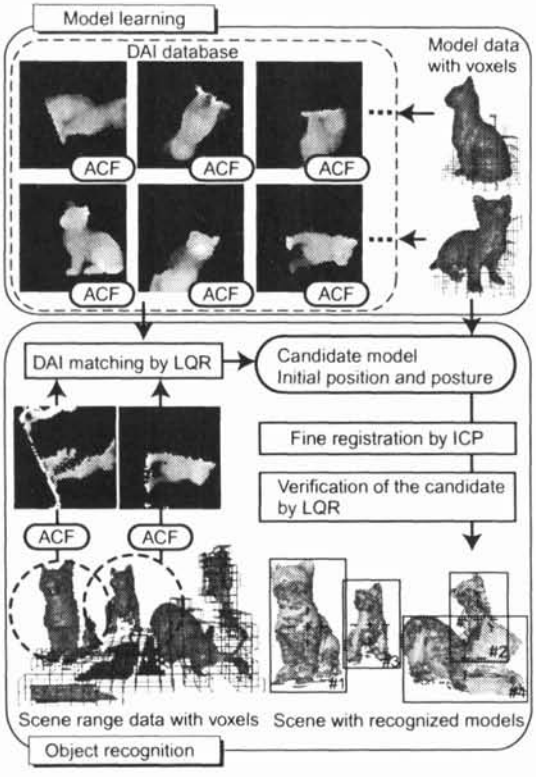


Figure 1: Outline of the proposed method.

ating a fit. After renewal of ACF, in order to recognize multiple objects in the scene, these processes are iterated until ACFs can not be constructed.

2.2 Regular framing

A set of range data consists of points $P = \{\boldsymbol{p}_i = (x_i, y_i, z_i)^T, 1 \leq i \leq |P|\}$ and patches $R = \{R_i, 1 \leq i \leq |R|\}$, where $|A|$ is the number of elements in a set A. Each patch $R_i = \{\boldsymbol{r}_{ij} | \boldsymbol{r}_{ij} \in P, 1 \leq j \leq |R_i|\}$ is formed by three or four vertices. A surface normal calculated by cross product is defined on each patch as $\boldsymbol{n}_{R_i}$, $|\boldsymbol{n}_{R_i}| = 1$. Matching is to obtain a congruent transformation that can make a set of range data correspond another range data. In order to realize such a transformation, we need coordinate frames possibly defined by a set of three axes or four points. Therefore we introduce *regular framing* for this aim. Regular flaming is a segmentation of a set of range data into cubes called voxels, each of which is arranged in regular order and encloses the vertices. Fig.2 shows the outline of the regular framing. The voxel is characterized by two contents: a barycenter and a normal, which are calculated as averages of vertices and vertex normals. The voxel width Δ_v is determined through preliminary experiments and the direction of voxel arrangement is arbitrary. Let $V = \{\boldsymbol{v}_i = (\tilde{x}_i, \tilde{y}_i, \tilde{z}_i)\}$ be a set of all voxels, whose integer indices $(\tilde{x}_i, \tilde{y}_i, \tilde{z}_i)$ represent a cubic piece $v_i = (\tilde{x}_i \times \Delta_v \leq x < (\tilde{x}_i + 1) \times \Delta_v, \tilde{y}_i \times \Delta_v \leq y < (\tilde{y}_i + 1) \times \Delta_v, \tilde{z}_i \times \Delta_v \leq z < (\tilde{z}_i + 1) \times \Delta_v)$. Let $P_i = \{\boldsymbol{p}_{ij} = \boldsymbol{p}_i | \boldsymbol{p}_i \in v_i, 1 \leq j \leq |P_i|\}$ be a set of points included in v_i, and $B = \{\boldsymbol{b}_i = \boldsymbol{v}_i || P_i| \neq 0\}$ be the set of non-null voxels. Next, the voxel barycenter $\boldsymbol{u}_i$ and normal $\boldsymbol{n}_i$ are calculated as follows: $\boldsymbol{u}_i = \sum_{j=1}^{|P_i|} \frac{\boldsymbol{p}_{ij}}{|P_i|}$,

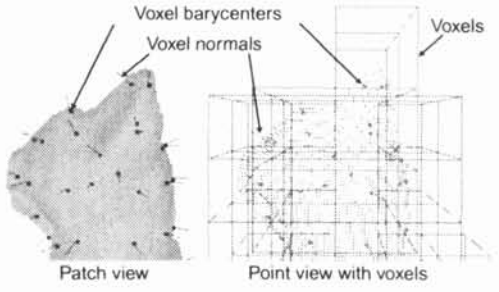


Figure 2: Regular framing of range data.

and $\boldsymbol{n}_i = \frac{\bar{\boldsymbol{n}}_i}{|\bar{\boldsymbol{n}}_i|}$, where $\bar{\boldsymbol{n}}_i = \sum_{\boldsymbol{r}_{ij} \in b_i} \phi(R_i) \boldsymbol{n}_{R_i}$, where $\phi(R_i)$ expresses how many nodes of R_i are included in b_i.

2.3 Aspect coordinate frame (ACF)

To define basic coordinate frames for estimating the congruent transformation, we generally need some reference points in P or B. Searching the reference points in P needs high computational cost, and in another approach, some feature points, edges or corners, are possibly used as such a reference [8], however, it is not sufficient for free-form surfaces. Therefore we propose an approach of using voxels in B as references. We call such a basic coordinate frame as *aspect coordinate frame*, hereafter ACF. A 3-tuple of voxels is utilized here to define an ACF. Then the number of 3-tuples becomes $O(|B|^3)$, which is rather high from a point of view of a processing cost, and we have a risk of combinatorial explosion. So we introduce some restrictions on a 3-tuple, which can eliminate ill-defined combination like collinear voxels or those which do not construct any feasible plane. Let $(\boldsymbol{b}_i, \boldsymbol{b}_j, \boldsymbol{b}_k)$, $(i \neq j \neq k)$ be a 3-tuple of voxels, where a distance between voxels can be defined as $|\boldsymbol{b}_i - \boldsymbol{b}_j| = |\tilde{x}_i - \tilde{x}_j| + |\tilde{y}_i - \tilde{y}_j| + |\tilde{z}_i - \tilde{z}_j|$. The restrictions on the 3-tuple are designed as:

$$|\boldsymbol{b}_j - \boldsymbol{b}_i| \geq A_T. \quad (1)$$

$$\boldsymbol{n}_i \cdot \boldsymbol{n}_j \geq 0, \boldsymbol{n}_j \cdot \boldsymbol{n}_k \geq 0, \boldsymbol{n}_k \cdot \boldsymbol{n}_i \geq 0. \quad (2)$$

$$|\boldsymbol{b}_j - \boldsymbol{b}_i| = |\boldsymbol{b}_k - \boldsymbol{b}_i|. \quad (3)$$

$$(\boldsymbol{b}_j - \boldsymbol{b}_i) \cdot (\boldsymbol{b}_k - \boldsymbol{b}_i) = 0. \quad (4)$$

$$((\boldsymbol{b}_j - \boldsymbol{b}_i) \times (\boldsymbol{b}_k - \boldsymbol{b}_i)) \cdot \boldsymbol{n}_i \geq 0. \quad (5)$$

Condition (1) is called *minimum distance restriction*. Condition (2) is *observability restriction*, which shows simultaneous observability of the three voxels. Condition (3) is called *equidistance restriction*, which requests of the third voxel that it should lie in the equidistance to the other voxels. Condition (4), which is called *orthogonality restriction*, eliminates collinear voxels. We call condition (5) *direction restriction*, which requires the surface orientation around the ACF to be similar to the upper orientation of the ACF (the positive orientation of z axis). The ACF can be defined through a 3-tuple $u = (\boldsymbol{u}_i, \boldsymbol{u}_j, \boldsymbol{u}_k)$ that are modified from the 3-tuple $(\boldsymbol{b}_i, \boldsymbol{b}_j, \boldsymbol{b}_k)$ satisfying all the conditions from (1) to (5). A set of 3-tuples is represented as $U = \{u\}$ and the number of element in U is K. The xy plane of the ACF is called *base plane*, hereafter BP, onto which the depth around the ACF are

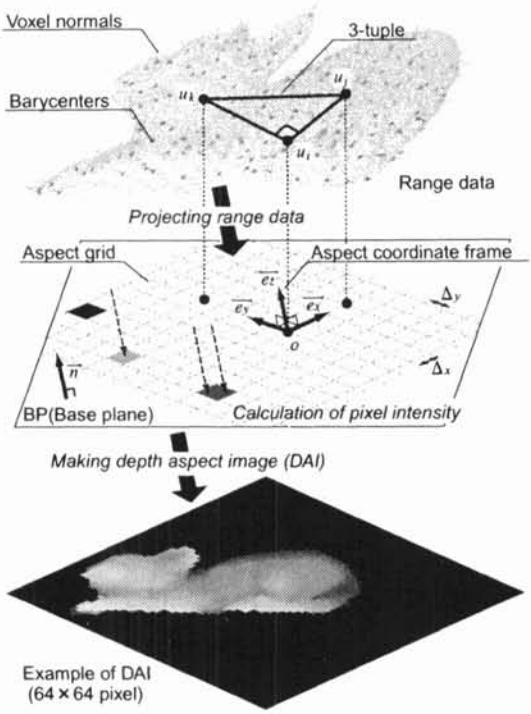


Figure 3: Procedure of making DAI.

mapped. Fig.3 shows the definitions of ACF and BP. The origin is $\boldsymbol{u}_i$ and the unit vector $\boldsymbol{e}_x$ passes through it and $\boldsymbol{u}_j$. The unit vector $\boldsymbol{e}_z$ is defined as the normal vector of the BP, and then the last vector $\boldsymbol{e}_y$ is set so that it is orthogonal to both $\boldsymbol{e}_x$ and $\boldsymbol{e}_z$. Let $C = \{c = (o, \boldsymbol{e}_x, \boldsymbol{e}_y, \boldsymbol{e}_z)\}$ be a set of ACFs. Therefore the number of element C is $|C| = |U| = K$.

2.4 Depth aspect image (DAI)

Each component of the set P is converted with respect to the ACF, resulting $\{P'_k\}_{k=1}^{K}$ as follows:

$$P' = \{\boldsymbol{p}'_i = [\boldsymbol{e}_x, \boldsymbol{e}_y, \boldsymbol{e}_z]^{-1}[\boldsymbol{p}_i - \boldsymbol{o}] = (x'_i, y'_i, z'_i)\}. \quad (6)$$

As shown in Fig.3, an aspect grid $A = \{a_{lm}\}$ with the width Δ_x and Δ_y on the BP are defined. A DAI is defined on the BP by projecting the depths on the aspect grid along with the z axis of the ACF. The DAI is extracted from the BP in the range of $\pm\frac{1}{2}L\Delta_x$ and $\pm\frac{1}{2}M\Delta_y$. So the size are given by $L\Delta_x \times M\Delta_y$ and its resolutions along with each axis are given by Δ_x, Δ_y and Δ_z. P' is projected along the z axis orthogonally onto the BP. Each pixel of the DAI can be assigned its virtual brightness by the following procedure. If some depths are projected into a grid, the value is $\max\{\bar{z}_i = \lceil \frac{z'_i}{\Delta_z} + 128 \rceil\}$. Then $\bar{z}_i$ is clipped into $0 \leq \bar{z}_i \leq 255$. Fig.4 shows some examples made by the these procedures. The figure shows that DAI is a visualization of depth structure from BP.

For each of models, multiple DAIs are constructed according to ACFs, and are registered into a database with corresponding c, u and s, where s means the maximal side length of u.

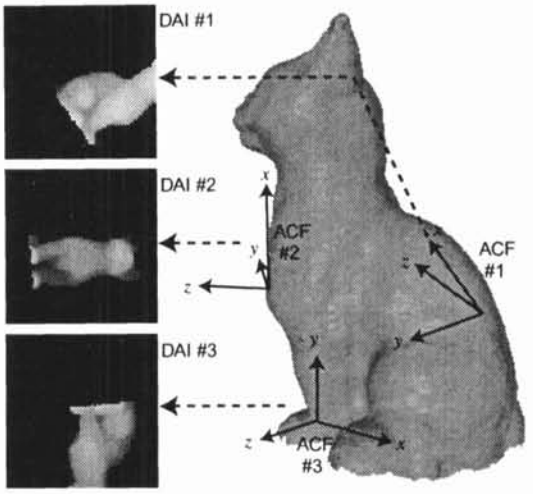


Figure 4: Examples of DAI and its ACF.

3 Object recognition

3.1 Recognition process

Fig.5 shows the recognition process which can be classified into two phases. In the search phase, after selecting a 3-tuple, the ACF is constructed and then the corresponding DAI is generated. It is matched with DAIs in the database in order. The optimal DAI gives a candidate model and the pair of ACFs constructs a congruent transformation from the candidate to the scene. If the DAI matching is failed, another DAI is remaked by another ACF repeatedly until we find a solution. In the verification phase, the candidate fitted by ICP algorithm is verified by LQR. After a part of recognized range data are removed, the above procedures recursively applied to remains.

3.2 Least quantile of residual (LQR)

For comparing two histograms, the least median of squares principle can be well utilized for obtaining robust matching between them in spite of gross or outlying components. In the proposed method, we introduce the robust statistical evaluator called *least quantile of residual* (LQR), which is defined as:

$$f_Q(H) = \arg\min_q \left\{ \frac{\sum_{i=0}^{q} h_i}{\sum_{i=0}^{|H|-1} h_i} \geq Q \right\}, \quad (7)$$

where h_i is the i-th level of a residual histogram and Q is the quantile parameter. If $Q = 0.5$, the above rule generates the least median of residuals estimate. This evaluator can be applied to different histograms because we can design the level of Q relative to problems at hand.

3.3 DAI matching

Let $A = \{a_i\}$ and $M = \{m_i\}$ be the DAI of the scene and the one from the database, respectively. Zero valued pixels in the DAI have to be eliminated before

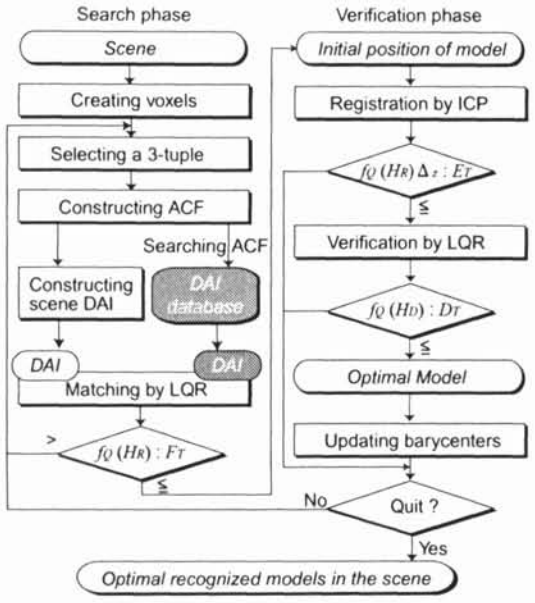


Figure 5: Recognition process.

matching because those have no information on depth. So we define overlapped pixels on the two DAIs as:

$$A' = \{a'_i\} = \{a_i | a_i \neq 0 \cap m_i \neq 0\}, \quad (8)$$
$$M' = \{m'_i\} = \{m_i | a_i \neq 0 \cap m_i \neq 0\}, \quad (9)$$

where the number of overlapped pixels is defined as $|A'| = |M'| = N_{am}$. Some noises and outliers due to occlusion by other objects may be appeared on the DAI as unexpected intensities. Then LQR is employed in order to match the DAIs robustly under these ill-conditions. Let $r_i = |a'_i - m'_i|$ be a residual of each pair of pixels, then the histogram is written as:

$$H_R = \{h_i | h_i = \sum_{j=1}^{N_{am}} \delta(r_j - i)\}, \quad (10)$$

where $\delta(\cdot)$ is Kronecker's delta. The optimal DAI in the database indicates the candidate model so that it has the minimum $f_Q(H_R)$ less than threshold F_T. If not the case, another 3-tuple is tried to be investigated. We introduce some procedures to speed up the search of DAI. Through the search, the model DAIs with largely difference from s are skipped over, and those having small area of overlap are also ignored. Due to the DAI matching by LQR, feasible initial positioning and point-wise relations can be roughly obtained without any extra expensive computation. Moreover the closest point search based on barycenters is faster than the original version. In the experiments, these parameters are set as $Q = 0.6$ and $F_T = 10$.

3.4 Verification by ICP and LQR

Let c_r and c_s be the ACF of the candidate and the one of the scene, and $P_r^s = \{\boldsymbol{p}_r^s\}$ be the set of points of the candidate transformed to the scene. Each transformed point $\boldsymbol{p}_r^s$ is calculated as:

$$\boldsymbol{p}_r^s = A(c_s)A(c_r)^{-1}[\boldsymbol{p}_r - \boldsymbol{o}_r] + \boldsymbol{o}_s. \quad (11)$$

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

Figure 6: Models.

The fitted candidate is registered precisely by ICP algorithm [6]. Original ICP algorithm has some problems such as local minima due to the initial positioning and/or a set of non-overlapping data. The result of DAI matching handles them. The initial position and posture are obtained through the DAI matching successfully. And we can select points of the candidate, which are used for ICP algorithm according to the small residual pixels of matched DAIs. The voxel segmentation has another merit of fast computation of the closest point search although it provides an approximate solution.

After ICP algorithm is applied, the candidate is verified by the amount of total movement along fitting and by LQR with closest point distances. The histogram H_D is made from the closest point distances $D = \{d_i\}$ between the transformed candidate and the scene. And we set $Q = 0.25$ in the verification phase because the whole model surface cannot be measured from one direction and the model surface may be blinded partially by other objects. Let E be the sum of the mean distances of movement in ICP algorithm, and if the candidate satisfies $E \leq E_T$ and $f_Q(H_D) \leq 2$, then it is accepted. In the other case, another DAI is tried to match. In the experiments, we set $E_T = f_Q(H_R) \times \Delta_z$. After getting the verified candidate, the range data congruent with it are modified so that they can be ignored by any following iteration.

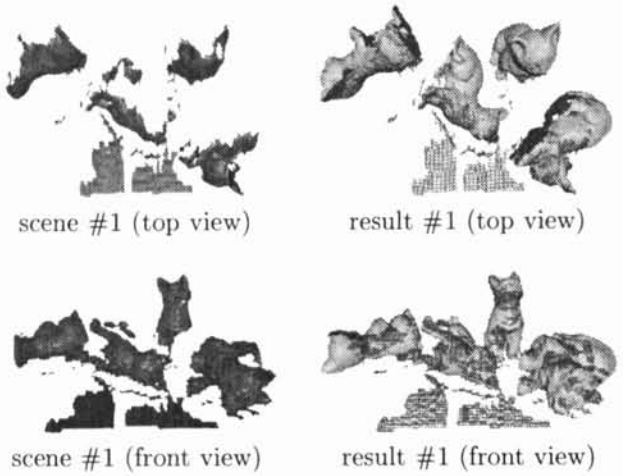
4 Experiments

All the range data we used were measured by the laser range finder (MINOLTA VIVID300). Fig.6 shows the models of toy cats, each of which was merged from the eight set of range data equally measured around every 45 degrees. Small holes and ill-defined patches on the integrated range data are modified by polygon editing tools.

Table 1 shows the parameters for making DAI in the experiments. Table 2 shows the number of DAIs for each model created with under the parameters shown in Table 1, resulting the total number 6490. We have used 100 scenes including two through four models with random placement, which were put at from 600 to 1200 mm far from the range finder. These scenes also contained unknown objects like a floor, human hands, and so on. Fig.7 shows one example of recognition results.

Table 1: Parameters for DAI.

Δ_v	15mm
A_T	3
$\Delta_x \times \Delta_y$	3×3 mm
Δ_z	0.3 mm
$L \times M$	64×64



scene #1 (top view) result #1 (top view)

scene #1 (front view) result #1 (front view)

Figure 7: An example of recognition result.

The left figures in Fig.7 are the scene range data observed from the top and from the front respectively, and the right figures are the visualization of the four recognized models in the scene. Recognition and position estimation of each model were successful, and they were recognized at 599th, 1403rd, 1967th and 3133rd steps, respectively, and then the DAIs of the scene was made 3432 times in total. 30 from them were matched successfully. Furthermore, two of them were rejected due to the total movement in ICP algorithm beyond E_T, and 24 of them were also rejected in the verification by LQR and the other four found their candidates successfully. It took 232 seconds in total using a PC with a Pentium 4 processor at 1.7 GHz. The resolution of each model in Fig.6 were (a)1.04mm, (b)1.06mm, (c)1.09mm and (d)1.10mm, respectively. On the other hand, the resolutions corresponding to each model were (a)2.22mm, (b)2.25mm, (c)2.36mm and (d)2.19mm, respectively. The effectiveness of the proposed method could be shown by the success on the data with these various resolutions.

In order to evaluate the performance of the method in the case of occlusion and clutter [4], occlusion is defined as:

$$occlusion = 1 - \frac{model\ surface\ patch\ area}{total\ model\ surface\ area}. \quad (12)$$

The occlusion rate is beyond 50 % due to the integrated range data. Next, clutter is defined as:

$$clutter = 1 - \frac{points\ of\ models\ in\ the\ scene}{total\ points\ in\ the\ scene}. \quad (13)$$

The clutter is an appearance rate of points in the scene excluding relevant points to models. Then the clutter becomes zero if all the vertices in the scene are relevant to the models. In the case of Fig.7, the occlusions

Table 2: Number of voxels and DAIs for each model.

Model	(a)	(b)	(c)	(d)
Point	27068	31405	25233	25950
Patch	28819	33556	27474	27655
Voxel	181	225	189	204
DAI	1217	2121	1408	1744

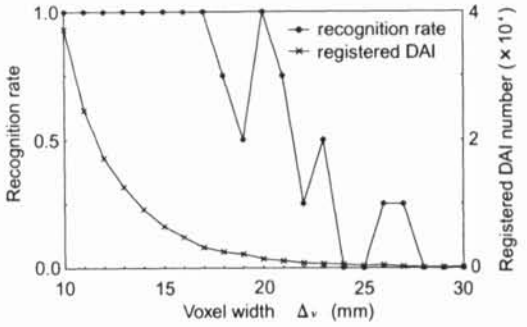


Figure 8: Recognition rate and number of registered DAIs for voxel width.

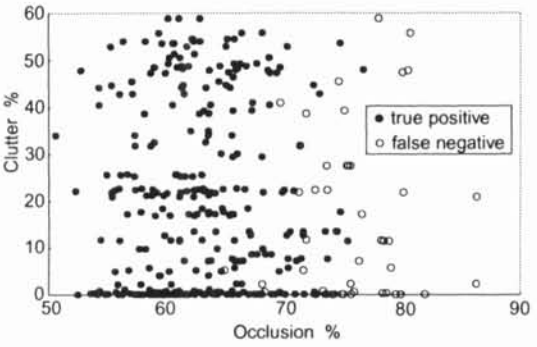


Figure 9: Recognition states for clutter and occlusion.

for each model were (a)74.5%, (b)70.5%, (c)64.2% and (d)71.6%, respectively, and the clutter of the scene was 13.4%.

Fig.8 shows the recognition rate and the number of DAIs in the databases for several voxel widths. In order to examine a basic behavior for each voxel widths, the scene including the four models was chosen so that the occlusions were (a)61.8%, (b)63.1%, (c)56.3% and (d)64.3%, respectively, and the clutter was 0.0%. The small voxel widths increased the number of DAIs in the databases and raise up the recognition rate. If the widths were less than 17mm, all the four models were recognized successfully. Then the voxel width used in the experiments for the scenes was sufficient to recognize four models, and the database could be relatively compact.

Fig.9 shows the result of recognition for 100 scenes. 356 models were appeared in all the scenes, and 312 of them were recognized successfully (true positive), and then the rest could not be found in the scenes (false negative). It was worth noting that there was no false positives in the results. We found that increasing occlusion rate causes an increase of the false negatives, but it was independent of the clutter. Relation between the occlusion and the recognition rate is shown in Fig.10. The average rate calculated from all the models in the scenes was 87.6%, however, it became 99.6% except models whose occlusions were over 70%. Therefore the false negatives were caused by the insufficiency of relevant surfaces to the models in the scenes. Other examples of experiments are shown in Fig.11. The processing time were spread from some seconds to

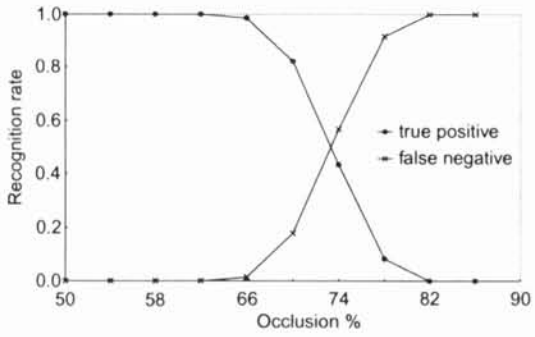


Figure 10: Recognition rate versus occlusion.

around a thousand seconds. From the point of view of processing time, the efficiency of our method was not so high. We will consider about some other possible procedures for skipping inrelevant DAIs of the scene.

5 Conclusions

We proposed the recognition method for rigid objects based on the depth aspect image constructed by regular framing. And we introduced the robust evaluator LQR which is effective both in the image matching and in the verification. Voxels given by the regular framing of range data are also useful for quick search of the closest points in ICP algorithm and verification. Through the experiments with 100 scenes including multiple objects with clutter and occlusion, the effectiveness of our method is confirmed.

References

[1] A.Hoover, et al. , "An Experimental Comparison of Range Image Segmentation Algorithms, "IEEE Trans. on PAMI, vol.18, no.7, pp.673-689, 1996.

[2] F.Stein and G.Medioni, "Structural indexing: Efficient 3-D object recognition, "IEEE Trans. on PAMI, vol.14, no.2, pp.125-145, 1992.

[3] C.S.Chua and R.Jarvis, "Point Signatures: A New Representation for 3D Object Recognition, "International Journal of Computer Vision, Vol.25, No.1, pp.63-85, 1997.

[4] A.E.Johnson and M.Hebert, "Using Spin Images for Efficient Object Recognition in Cluttered 3D Scenes, "IEEE Trans. on PAMI, vol.21, no.5, pp.433-449, 1999.

[5] S.Ruiz-Correa, L.G.Shapio and M.Melia, "A New Signature-Based Method for Efficient 3-D Object Recognition, "Proc. CVPR, I, pp.769-776, 2001.

[6] P.J.Besl and N.D.McKay, "A Method for Registration of 3-D Shapes, "IEEE Trans. on PAMI, vol.14, no.2,1992.

[7] T.Takeguchi, T.Kondo, S.Kaneko and S.Igarashi, "Object Recognition based on Depth Aspect Image Matching, " Proc. International Workshop on Machine Vision Applications, pp.476-480, 2000.

[8] T.Takeguchi, T.Kondo, S.Kaneko and S.Igarashi, "Robust Object Recognition based on Depth Aspect Image Matching, " trans. IEICE, vol.J-84-D-II, no.8, pp.1710-1721, 2001. [in japanese]

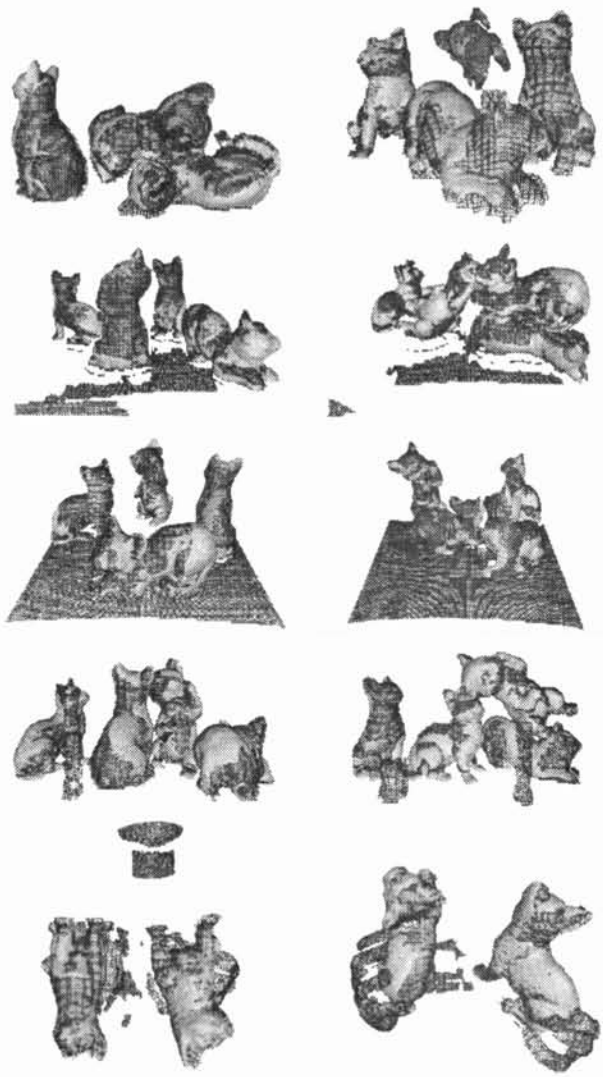


Figure 11: Examples of recognition result.

[9] S.Kaneko, I.Murase and S.Igarashi, "Robust Image Registration by Increment Sign Correlation, ", Pattern Recognition, vol.35, no.10, pp.2223-2234, 2002.