MVA2000 IAPR Workshop on Machine Vision Applications, Nov. 28–30, 2000, The University of Tokyo, Japan

13—5

Object Recognition based on Depth Aspect Image Matching

Tomoyuki TAKEGUCHI, Tsukasa KONDO, Shun'ichi KANEKO and Satoru IGARASHI *
Faculty of Engineering, Hokkaido University

Abstract

A method for three dimensional object recognition based on depth image information is proposed. A depth aspect image is defined as an orientation standardized appearance from the original depth data of the object, which is transformed by the rigid transformation drawn by each possible basis pair of every three feature points of the object depth data. They are made from the original depth images of models and then learned in the system as the database for retrieval of any instances on the models. Matching between an object aspect and the ones from models can be performed by two-dimensional image comparison, which is based on the least quantile of residuals and is robust against occlusion possibly occurred in cluttered scene. The paper includes a formalization of the proposed method and some experimental results with real objects.

1 Introduction

Study on recognition and estimation of three-dimensional posture and motion based on depth images is one of fundamental problems in computer vision for many applications, in the fields of robotic vision and factory automation. There have been proposed some methods for measuring depth, such as laser range finders [1], binocular and multinocular stereo visions [2], factorization [3], depth from defocus or focus [4]. We can roughly classify these approaches into two categories as follows: the depth image based approaches and the point set based one. In the former approach, depth images of rather dense spatial data are utilized for detection some geometric features, such as curvature or edges which are used for data compression and/or geometrical coordination.[5] They are also effective, for example, in model-based matching, for decreasing computational cost, but they are strongly dependent on stability and repeatability of features and their extraction procedures. Otherwise, in the latter approaches, even sparse but less position data can be used for merging algorithms [6] solid object matching[7].

In this paper, a model-based method is proposed for realizing robust object recognition, which is based on 2-D depth aspect images in a registered model database and handles dense depth data with fine resolution of partial shapes of models[8] A novel matching scheme is also presented, which is based on a robust statistic and hence model instances with occluded part can be searched in the real scene. Depth aspect images are fundamental in the method and they are made through relative coordinates by the feature points on the model surfaces, enabling position and posture estimation of the models in the scene. The proposed method is suitable for hardware realization and so fast and real time processing.

The paper consists as follows: In Section 2, an outline of the method is given together with definition of depth aspect images and the algorithm for generating them. In Section 3, a robust recognition algorithm using image-based matching is given for handling complex scenes with multiple objects. In Section 4, experiments with real scenes are presented, and then we conclude the paper with some remarks in Section 5.

* Address: Kita 13, Nishi 8, Kitaku, Sapporo 060-8628 Japan. E-mail: take@mee.coin.eng.hokudai.ac.jp

2 Depth aspect image

2.1 Outline of processing

Fig.1 shows an outline of the proposed method. The method consists of the following two components: model registration and object recognition. In the model registration processing, depth images of models are measured by a range sensor, such as a laser range finder, and geometrical feature points are extracted from them through curvature evaluation. Local coordinate frames called 'Aspect coordinate frame', hereafter ACF, are defined to convert depth data to depth aspect images, hereafter DAI. For every model, multiple DAIs can be derived corresponding each three tuple of measured points of the model surface, and then they are registered into the DAI database with the corresponding information of ACF, which is utilized to reconstruct the object position and posture in the scene. In the object recognition processing, a partial set of depth data can be converted to the DAI through the ACF consisting of the selected three tuple of feature points. The DAI is compared with each of possible DAI from the model database. The model and its position and posture can be obtained as solution.

2.2 Extraction of feature points

In this paper, depth images are measured by a laser range finder, hereafter LRF, with fine resolu-

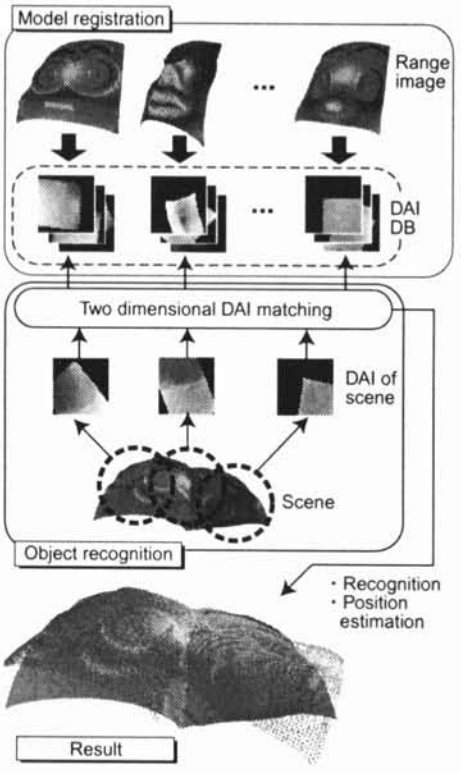


Figure 1: Overview of the proposed method

tion. For each model, let $P = \{\boldsymbol{p}_{ij} = (x_{ij}, y_{ij}, z_{ij})\}$ be a measured point set of total point number N. $\boldsymbol{p}$ shows a vector. Feature points are defined based on their curvature features [9] which are expected to be independent of viewing directions and repeatable enough for defining ACF in the scene. The depth data equally arranged in x, y directions by the LRF used in our measurement. Peripheral points $\{\boldsymbol{b}_k\}_{k=0}^{15}$ in a 5×5 square neighborhood are used for calculating eight curvatures corresponding each pair of points on opposite sides as follows:

$$\rho_k(\boldsymbol{p}_{ij}) = \frac{\cos^{-1}\left\{\frac{(\boldsymbol{p}_{ij}-\boldsymbol{b}_k)\cdot(\boldsymbol{b}_{k+8}-\boldsymbol{p}_{ij})}{|\boldsymbol{p}_{ij}-\boldsymbol{b}_k|\cdot|\boldsymbol{b}_{k+8}-\boldsymbol{p}_{ij}|}\right\}}{\frac{1}{2}\left(|\boldsymbol{p}_{ij}-\boldsymbol{b}_k| + |\boldsymbol{b}_{k+8}-\boldsymbol{p}_{ij}|\right)} \quad (1)$$

where $k = 0, 1, \cdots, 7$. The maximal curvature for each point $\boldsymbol{p}$ is defined as its feature $S(\boldsymbol{p})$, and define the set of points having larger feature values than a threshold S_T as the feature point set $T = \{\boldsymbol{u}_{ij}\} = \{\boldsymbol{p}_{ij} | \, S(\boldsymbol{p}_{ij}) > S_T\}$.

2.3 Three tuple of feature point

In order to do any matching, a certain reference is necessary for registration of positions and postures of two objects on interest. Any three tuple of feature points may be such reference, however, the computation costs in both model registration and object recognition become enormous, so we have to introduce some limitation to the condition as reference. From T, we select a set of three tuple of feature

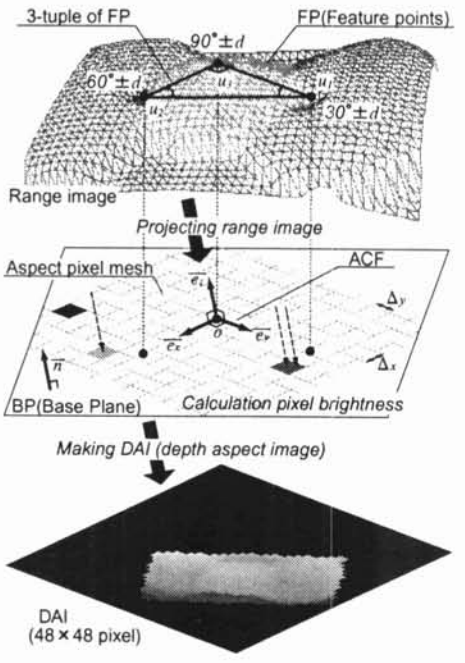


Figure 2: Definitions of 3-tuple, ACF and DAI.

points U that construct the right triangle with a limited range of its three angles as follows:

$$U = \{u = (\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3), \boldsymbol{u}_i \in T, \gamma(u) = 1\} \quad (2)$$

$$\gamma(u) = \begin{cases} 1 & (30i - D_T \leq \alpha_i \leq 30i + D_T \\ & \quad , \forall i = 1, 2, 3) \\ 0 & (\text{otherwise}) \end{cases} \quad (3)$$

$$\alpha_i = cos^{-1}\left(\frac{(\boldsymbol{u}_i - \boldsymbol{u}_{(i+1)}) \cdot (\boldsymbol{u}_i - \boldsymbol{u}_{(i+2)})}{|\boldsymbol{u}_i - \boldsymbol{u}_{(i+1)}||\boldsymbol{u}_i - \boldsymbol{u}_{(i+2)}|}\right) \quad (4)$$

where $(1) = 1, (2) = 2, (3) = 3, (4) = 1, (5) = 2$, α_i means vertex angles at $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, $\boldsymbol{u}_3$, D_T is a range width with respect to angle value. We set $|U| = K$ and omit suffices like u_k for components of U for simplicity. The threshold D_T has to satisfy a condition $D_T < 15(°)$ so that each range of angles is not overlapped each other. In Equation(3), the function $\gamma(u)$ determines if three angles of the tuple u are within the range of 30°, 60°, 90°. The condition for the possible three tuple of feature points can limit the total number of ACF and it enable to discriminate three selected points for localize a coordinate frame on the tuple, in comparison to the case that any three tuple of feature points is possible as the candidate for ACF.

2.4 Aspect coordinate frame

An ACF can be defined and localized on each of the selected tuple under the condition abovementioned, all of which are included in U. The xy plane of an ACF is called a base plane, hereafter BP, which is used for mapping of depth values. Fig.2 shows the definitions of ACF and BP. The origin and the axes of ACF can be defined using three

feature points, respectively, and they serve as a coordinate system for converting the measured depth values with respect to it and for making the DAI of the model. The DAI is defined on the BP of the ACF, which includes all the feature points in the tuple, by projection along the z axis of the ACF. The origin is $\boldsymbol{u}_3$, the unit vector $\boldsymbol{e}_x$ passes through it and $\boldsymbol{u}_2$, 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 $\boldsymbol{e}_x$ and $\boldsymbol{e}_z$. The set of ACF $C = \{c = (o, \boldsymbol{e}_x, \boldsymbol{e}_y, \boldsymbol{e}_z)\}, |C| = K$, thus, is defined as

$$\boldsymbol{o} = \boldsymbol{u}_3 \tag{5}$$

$$\boldsymbol{e}_x = \frac{\boldsymbol{u}_2 - \boldsymbol{u}_3}{|\boldsymbol{u}_2 - \boldsymbol{u}_3|} \tag{6}$$

$$\boldsymbol{e}_z = \frac{(\boldsymbol{u}_2 - \boldsymbol{u}_3) \times (\boldsymbol{u}_1 - \boldsymbol{u}_3)}{|\boldsymbol{u}_2 - \boldsymbol{u}_3||\boldsymbol{u}_1 - \boldsymbol{u}_3|} \tag{7}$$

$$\boldsymbol{e}_y = \boldsymbol{e}_z \times \boldsymbol{e}_x \tag{8}$$

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

$$P' = \{\boldsymbol{p}'_{ij} = (x'_{ij}, y'_{ij}, z'_{ij})\} \tag{9}$$

$$\boldsymbol{p}'_{ij} = A(c)^{-1}(\boldsymbol{p}_{ij} - \boldsymbol{o}) \tag{10}$$

$$A(c) = [\boldsymbol{e}_x, \boldsymbol{e}_y, \boldsymbol{e}_z] \tag{11}$$

2.5 Depth aspect image

All the components of P'_k are converted with respect to the ACF and mapped onto the BP. As shown in Fig.2, an aspect grid $A = \{a_{lm}\}$ with the width Δ_x and Δ_y on the BP.

$$A = \{a_{lm}\}_{l=0,1,\cdots,L-1,\ m=0,1,\cdots,M-1} \tag{12}$$

P'_k are projected along the z axis orthogonally onto the BP. Each pixel of the DAI can be assigned its virtual brightness by the following procedure. Let $P''_{lm} = \{\boldsymbol{p}'_{ij}\} \subseteq P'_k$ be the portion which is projected onto a grid a_{lm}, and then the points satisfy the conditions as

$$l\Delta_x - b_x \leq x'_{ij} < l\Delta_x - b_x + \Delta_x \tag{13}$$

$$m\Delta_y - b_y \leq y'_{ij} < m\Delta_y - b_y + \Delta_y \tag{14}$$

where (b_x, b_y) is the offsets of the DAI origin from the one of the ACF. The pixel value is derived from the z values of P''_{lm}. When it includes multiple points, the maximal projected value is selected as follows:

$$a_{lm} = \begin{cases} max\{\bar{z}_{ij}\} & (|P''_{lm}| \geq 1) \\ 0 & (|P''_{lm}| = 0) \end{cases} \tag{15}$$

where $\bar{z}$ shows the quantized depth after conversion with respect to the ACF as

$$z''_{ij} = \lceil \frac{z'_{ij}}{\Delta_z} + 128 \rceil \tag{16}$$

$$\bar{z} = \begin{cases} 255 & (z'' > 255) \\ z'' & (1 \leq z'' \leq 255) \\ 1 & (z'' < 1) \end{cases} \tag{17}$$

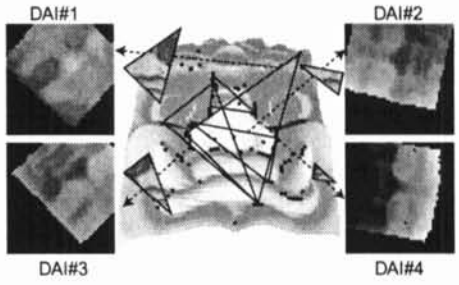


Figure 3: Examples of ACF and DAI

where Δ_z is the quantization width for z'. Brightness is of the eight-bit representation, and the pixels corresponding to the real depth have the brightness in the range of one to 255, which involves the intermediate value 128 that corresponds to the pixel values for the points just on the BP. Fig.3 shows some examples made by the abovementioned procedures. The figure shows that the DAI is a visualization of depth structure with respect to the ACF, and it represents imaginary three dimensional measurement according to the three tuple of feature points on the object surface. The spatial resolution of DAI is set higher than the one of the original depth image to some extent, and it effects a certain smoothing.

2.6 Model registration

For each of models, multiple DAIs are constructed according to three tuples of feature points and then ACFs, and they are registered as items with the following terms into a DAI database.

- model identifier:q
- $S(\boldsymbol{p})$ threshold:S_T
- ACF:$\{\boldsymbol{e}_x, \boldsymbol{e}_y, \boldsymbol{e}_z\}$
- three tuple:$\{\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3\}$
- maximal side length:$s = |\boldsymbol{u}_1 - \boldsymbol{u}_2|$
- DAI:$\{a_{lm}\}$

3 Robust matching of DAI

Fig.4 shows the processing flow in the object recognition. We call the object image of interest by the scene. From the scene, a DAI can be constructed in the same way of the model registration, except for the parameter of S_T which is set to the minimum of the values for each model. The candidate ACF have to be chosen as some likely one for efficient search of the object. Partial search is utilized for this task in terms of the maximal side length s for each ACF. Let s^a and s^m be the one for the scene and a model. the selected ACFs which satisfy the condition: $s^a - d_L \leq s^m \leq s^a + d_L$ can be candidates for matching processing. The parameter d_L can control the range of search in the DAI database. Candidate ACFs for the best match are

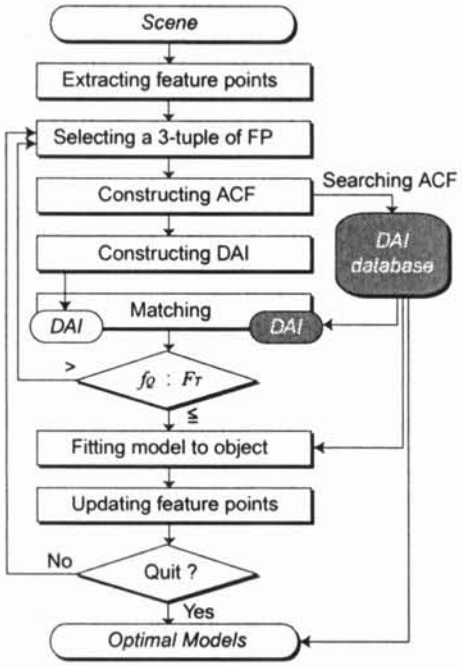


Figure 4: Procedure of recognition

selected from the database in order, and the DAI corresponding to the ACF is matched to the object DAI and evaluated the similarity. Let $A = \{a_{ij}\}$ and $M = \{m_{ij}\}$ be the DAI of the scene and the model DAI from the database, respectively. The following preprocessing selects the overlapped pixels between them.

$$A'_M = \{a'_{ij}\} = \{a_{ij} | a_{ij} \neq 0 \cap m_{ij} \neq 0\} \quad (18)$$

$$M'_A = \{m'_{ij}\} = \{m_{ij} | a_{ij} \neq 0 \cap m_{ij} \neq 0\} \quad (19)$$

The occlusion in depth images is the one of problems which generally occur in the real world situation [10].

In the paper, we introduce Least median of squares (LMedS) principle for solving the problem, which is one of effect approaches in robust statistics such as robust regression. In this approach, we can select the best model M'_A which achieves the minimum of the medians between the candidates M'_A and the object A'_M. In LMedS, it is guaranteed for the best model to have at least a half of residuals each of which is less than the value of the least median. In this paper, the condition of median is relaxed to adopt a quantile of residuals in evaluation of errors. A quantile of Q means the (number of overlapped pixel) $\times Q$th item of the ordered population, for example, the quantile of $Q = 0.5$ is the median. We call this approach by LQR(Least Quantile of Residual) where the best model can be selected so as to have the minimum of the Q quantile of the residuals.

Let E and $h(E)$ be the residuals between two DAIs and its histogram, respectively.

$$E(A'_M, M'_A) = \{r_{ij} = |a'_{ij} - m'_{ij}|\} \quad (20)$$

$$h(E) = (h_0, h_1, h_2, \cdots, h_{H-1}) \quad (21)$$

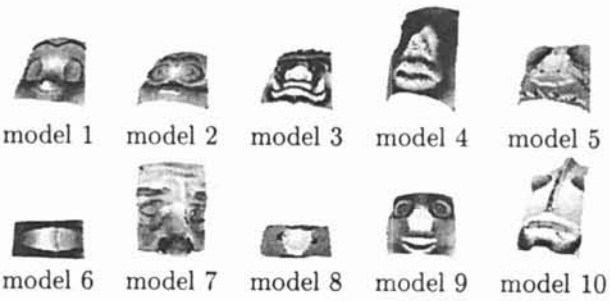


Figure 5: Models

$$h_k = \sum_{(i,j)} \delta(r_{ij}, k) \quad (22)$$

where the ith height in the histogram is given by h_i, the number of classes by H, and $\delta(\cdot)$ means an extended Kronecker's delta function which counts the number of residuals.

The quantile of Q is represented by the next equation, which is evaluated one after another..

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

For all the pairs of the object DAI from the scene and the DAI from the database, each of the quantile values is evaluated for searching the minimum which satisfies the condition of being less than a threshold w. If not the case, another three tuple of feature points and the ACF is tested to generate the DAI and match the models in the database.

Define the ACF of the best model selected and the one of the object by c_r and c_s, respectively, then an expected position and posture of the model in the scene can be estimated by transformation of the point set $P'_r = \{\boldsymbol{p}'_r\}$ through the following:

$$\boldsymbol{p}'_r = A(c_s)A(c_r)^{-1}[\boldsymbol{p}_r - \boldsymbol{o}_r] + \boldsymbol{o}_s \quad (23)$$

In order to search other possible objects after detection of any object, the same procedure is repeatedly applied to the scene after eliminating any feature points with the constant distance d_A from the transformed points P'_r. When a three tuple can not constructed any more or all the tuples have tested, recognition is quitted.

4 Experiments

Depth images are all measured with the pitch of $2mm$ in both x, y directions. Fig.5 shows ten models used for the experiments. We can find that the feature points on their surfaces illustrated by dots are distributed near on the edges. Table 1 shows the specifications for making DAIs. Table 2 shows the numbers of DAI and ACF for each model. These

Table 1: Specifications of DAI

$\|T\|$	100
D_T	1°
$L \times M$	48×48
$\Delta_x \times \Delta_y$	$3 \times 3\ mm^2$
Δ_z	$0.3\ mm$

Table 2: Numbers of DAI for each model

q	1	2	3	4	5
$\|U\|$	163	126	186	165	197
q	6	7	8	9	10
$\|U\|$	134	216	214	174	94

numbers are very small so that they are around one thousandth of the possible supreme number. The effect of filtering by right triangles can be verified by these numbers.

Table 3 involves thresholds used in the experiments.

The scenes with a single object and without occlusion were tested as fundamental experiments, where the objects were observed from the different orientations from the ones of models. Fig.6(a) shows the results. Wire frame versions of recognized models were overlapped on the scene so that partial objects with different orientations and occlusion could be recognized by the proposed method.

Next, the scene including multiple objects with occlusion were searched by the method as shown in Fig.6(b),(c). Some irregular points with deeper depth show lack of measurement. In the figure, black dots show the feature points detected in terms of their curvature features.

5 Conclusions

We proposed the rigid object recognition method based on the depth aspect image. And we also introduced robust image matching using the LQR method. Feature points reappearance, good application for partial data and scenes including multiple objects are confirmed through recognition experiments using 10 models.

References

[1] M.Riovx: Laser range finder based on synchronized scanners, Applied Optics, vol.23, no.21, pp.3837-3844, 1984.

[2] T.Kanade,et al.: A stereo machine for video-rate dense depth mapping and it new application, Proc. CVPR, pp.196-202, 1996.

[3] C.Tomasi,T.Kanade: Shape and motion from image streams under orthography: a factorization method, IJCV, 9:2, pp.137-154, 1992.

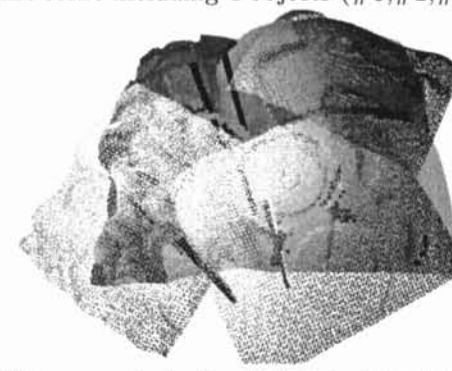
[4] S.K.Nayar and Y.Nakagawa: Shape from focus: An effective approach for rough sarfaces, Proc. ICRA, pp.218-225, 1990.

Table 3: Specifications of matchingn experiments

Q	0.6
d_L	$2mm$
w	4
d_A	$4mm$

(a)the scenes including single object (#1,#2,#5)

(b)the scene including 4 objects (#8,#2,#3,#9)



(c)the scene including 3 objects (#1,#2,#3)

Figure 6: Results of recognition experiments

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

[6] P.J.Besl and N.D.McKay: A Method for Registration of 3-D Shapes, IEEE Trans. on PAMI, vol.14, no.2, pp.239-256, 1992.

[7] 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.

[8] T.Takeguchi, S.Kaneko, T.Kondo, and S.Igarashi: Robust Object Recognition based on Depth Aspect Image Matchig, Proc. MIRU2000 , II pp.235-240, 2000. [in japanese]

[9] Y.Miyake, T.Kondo, S.Kaneko, S.Igarashi, H.Narahara: Reconstraction of Three Dimensional Surface from Slice Positional Data, Rapid Product Development, Chapman & Hall, pp.587-592, 1997.

[10] X.Yu,T.D.Bui and A.Krzyzak: Range image segmentation and fitting by residual consensus, Proc. CVPR, pp.657-660, 1992.