Object Recognition Using Planar Surface Segmentation and Stereo Vision

Dowan Kim, Sungil Kim and Sangchul Won

Department of Electrical Engineering Pohang University of Science and Technology, Pohang, Korea
(Tel: +82-54-279-5576; Fax: +82-54-279-8119; Email:{dwkim78,moceon,won}@postech.ac.kr)

Abstract: This paper describes a new method for 3D object recognition which used surface segment-based stereo vision. The position and orientation of an objects is identified accurately enabling a robot to pick up, even though the objects are multiple and partially occluded. The stereo vision is used to get the 3D information as 3D sensing, and CAD model with its post processing is used for building models. Matching is initially performed using the model and object features, and calculate roughly the object's position and orientation. Though the fine adjustment step, the accuracy of the position and orientation are improved.

Keywords: object recognition, vision, segmentation, model-based, stereo camera

1. Introduction

Vision is the best ability among human's perception. It provides us information of the environment without contacting an object. Computer vision is application of human's vision to a machine. One of the active research fields in computer vision is the object recognition in robotics and the computer vision researcher keep growing their interest. Computer vision gives an efficient and simple 3D object sensing tool to autonomous robot system without extra equipment and make the robot recognize an object.

When designing a recognition system, it is very important to decide what type of sensor will be used. We used stereo vision as the the sensor for 3D object recognition systems. Although stereo vision is a typical technique for sensing 3D information from intensity images, it is not often used in 3D object recognition because it has been considered inadequate for reconstructing the dense and accurate 3D data. However, stereo vision is suitable for object recognition if it is designed well, that is fast enough to create the range image containing 3D information of the scene. Yuns Oh suggest a new fast stereovision algorithm for stereo vision using VLSI method[2]. It makes our vision system to be directly used at 3D object recognition.

Even though it is reasonable to use stereo vision for object recognition, there have been very few researches. TINA[6] and VVV[5] systems are the two of few examples. Both systems use edge-based stereo vision. Two intensity images from the two camera with different view points are obtained and the edge images are extracted. Edge-based depth map is reconstructed using the correspondence between them. It is matched to a 3D wire frame model. But our system is different. We construct the disparity map using our stereo vision system, and reconstruct whole 3D data of a scene. Using the planar surface segmentation method of the 3D scene data, we recognize the pose and orientation of an object. The advantage of our system is that we use whole data from a stereo vision not a local information,i.e. edge.

The object recognition begins with designing appropriate models. It depends on the object to be represented and the algorithms how to choose the method of model representations. Generally, model-based object recognition uses two kind of model representation, feature-based model and appearance-based model.

Feature-based models represent 3D objects through features, their type, and their spatial relations. The identification means finding a set of features which is uniquely distinctive for an object. And, location is to match a number of image and object features and solve for the position and orientation of the 3D object. The advantage of feature-based models is that they generate compact object descriptors, offer some robustness against occlusion, and some invariance against illumination and pose variations. A disadvantage is that they cannot be compared directly with images and require feature extraction and object descriptions is obviously time-consuming and requires detailed knowledge of the internal structure of the object recognition system.

Appearance-based models represent an object through one or more images in eigenspaces method. Models are constructed using prototypical features and extracted from images of the to be model. Recognition means to find the image in a model set which is most similar to the one to recognize. The advantage is that images and models can be compared directly, and objects with no features. Disadvantages is that illumination, pose and location variations change the images.

In this paper, we choose the feature-based model. The objects,coil shaped object and cube, of our system are simple and they can be represented easily using the commercial CAD program. CAD provides a quick and compact object representation.

Our object recognition method is 3D planar surfaces extraction by segment-based stereo vision using range image that contains 3D information. Four research groups from University of South Florida(USF), Washington State University(WSU), University of Bern(UB) and University of Edinburgh(UE) have invented their range segmentation algorithm. The USF and UE algorithms are one the common approach to region segmentation by iteratively growing from seed regions. The WSU algorithm uses a powerful clustering and merging algorithm using surface properties. The UB algorithm uses another approach that exploits the scan line structure of the image. However, in this case no higher level processing for matching was proposed. In this paper,

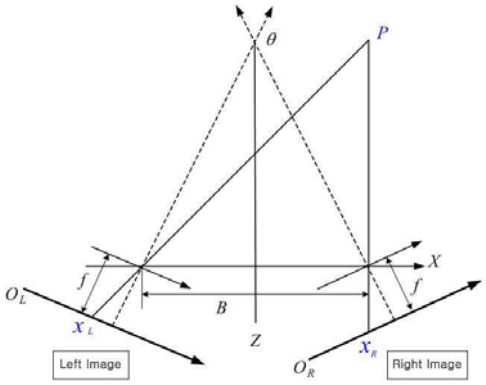

Fig. 1. The Geometry of Stereovision

we proposed a new range segmentation algorithm and 3D object recognition method. The range image is segmented using the two-dimensional histogram method based on a direction of normal vector of each 3D data point.

2. 3D Reconstruction

2.1. Stereo vision

Stereo vision refers to the ability to infer information on the 3D structure and distance of scene from two or more images taken from different viewpoints. It is one of the techniques to calculate the depth or distance information of the scene. Using the geometrical relationship between the two cameras and the location of the feature in each image, the disparity is obtained from solving the correspondence problem between the observed images. Correspondence problem is to determine the corresponding points among the images in the same scene with different view points. Yuns Oh suggests a new fast stereovision algorithm with Stereo Matching Chip using trellis-based stereo matching method [2]. It solves the correspondence problem and calculate the values of the disparities in 3D scene optimally. Two digital images are obtained together and produce the disparity image.

The geometry of stereo vision is shown in Fig.1 x_L and x_R are projected points in left and right camera image coordinate respectively. And let (x_C, y_C) is the middle image coordinate.

$$x_C = x_R + x_L + 1 \tag{1}$$

and y_C is assumed same as left and right image plane.
If the scene point P is mapped into two image planes, the disparity is the difference between the two imagery points[2].

$$d = x_R - x_L \tag{2}$$

The disparity map can be converted to a 3D data points. Using the center imaginary coordinate and the disparity (x_C, y_C, d), the data point,(x, y, z), in the world coordinate is calculated. The scene data shown in the disparity map is calculated by the following geometric equation.

$$\begin{aligned} A &= \tan\theta\left(\frac{\lambda_x^2}{4F^2}(N - x_C)^2 + 1\right) + (1 - \tan^2\theta)\frac{\lambda_x}{2F}d \\ &\quad - \tan\theta\frac{\lambda_x^2}{4F^2}d^2 \\ x &= \frac{B}{2}\frac{\frac{\lambda_x}{2F}(N - x_C)(\tan^2\theta + 1)}{A} \\ y &= \frac{|x\sin\theta + z\cos\theta|}{F}\lambda_y\left(\frac{M}{2} - 0.5 - y_C\right) \\ z &= \frac{B}{2}\frac{(\tan\theta\frac{\lambda_x}{2F}d - 1)^2 - (\frac{\lambda_x}{2F}(N - x_C)\tan\theta)^2}{A} \\ (0 &\le x_C \le 2I_w + 1, 0 \le y_C \le M - 1) \end{aligned} \tag{3}$$

In this equation, N and M is image and hight respectively, λ_x and λ_y means the image cell x-size and y-size, F is the focal length, B is the base line length that is the length between the left and right focal point, the each unit cell size of the camera image is α_x, α_y
. The distance measurement resolution is following.

$$\Delta z = \frac{\alpha_x(z)^2}{FB - \alpha_x z} \tag{4}$$

If the distance z is almost invariant, the distance measurement resolution is changed on by adjusting (F, B), because α_x is the horizontal size of the camera unit cell parameter that it can not be changed. If F is increased, the resolution is improved but the scene dimension of camera is decreased. If B is enlarged, the resolution also is improved. But it can make the value of d bigger than the maximum value allowed in the system.

2.2. 3D data reconstruction from disparity image

Using the two stereo cameras and stereo matching board, we get the disparity image. The image size is 2561×1000 pixels. It contains a lot of salt and pepper noise. To remove the noise we perform median filtering as a low-level processing. We exclude 200 columns from the both the end of the left and right side, because there is no disparity. Fig.2.
The 3D data is reconstructed using the Eqn. 3. The camera parameters are already calibrated. To reduce the consuming time, we sample the image data every five column and row. Fig.3.


Fig. 2. Disparity image.

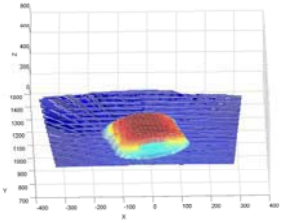
Fig. 3. Reconstructed 3D data from disparity image.

3. Planar Surface Segmentation and Object Features

3.1. Planar surface propriety

In this section, we describe a method of calculating surface normals that will be used in segmentation. If a surface can be represented by a function which has derivatives within a certain boundary, the normal can be calculated within the boundary. A tangent plane is defined at an object surface point $P = (x, y, z)$. Assume that the surface satisfies a specific conditional equation $F(x, y, z) = 0$ in the local neighborhood of point P, and that this equation is partially differentiable with respect to $P(x, y, z)$. In a small neighborhood of a point (x, y, z), the unit normal $N = N(x, y, z)$ can be calculated by the equation,

$$N(x, y) = \frac{\nabla F(x, y, z)}{\|\nabla F(x, y, z)\|} \tag{5}$$

where ∇ represents the gradient operator. Since we assumed $F(x, y, z)$ is C^1, first order Taylor approximations can be made within a small boundary around point (x, y, z). In this case, the first order equations describe a plane. If x is a point on the plane, then $x \cdot N = 1$. Then the plane equation can be represented as $F(x, y, z) = ax + by + cz = 1$. The constants a,b and c can be found by the equation,

$$\begin{aligned} a &= \frac{\partial F(x,y,z)}{\partial x} = \lim_{\Delta x \to 0} \frac{F(x+\Delta x, y, z) - F(x, y, z)}{\Delta x} \\ b &= \frac{\partial F(x,y,z)}{\partial y} = \lim_{\Delta y \to 0} \frac{F(x, y+\Delta y, z) - F(x, y, z)}{\Delta y} \\ c &= \frac{\partial F(x,y,z)}{\partial z} = \lim_{\Delta z \to 0} \frac{F(x, y, z+\Delta z) - F(x, y, z)}{\Delta z} \end{aligned} \tag{6}$$

then an approximation to a is,

$$a = P(x(i+1, j), y(i, j), z(i, j)) - P(x(i, j), y(i, j), z(i, j)) \tag{7}$$

, in this equation, i and j are the index of row and column of disparity image and $P(x(i, j), y(i, j), z(i, j))$ is calculated from $p_{i,j}$ that is a point of disparity image. Since this calculation is noise sensitive, we modify the operator so that it reduces the noise sensitivity by using linear least-squares methods of four locations around $p_{i,j}$,i.e. $p_{i-2,j}$, $p_{i+2,j}$, $p_{i,j-2}$ and $p_{i,j+2}$.

3.2. Histogram method for planar surface segmentation

The purpose of a histogram is to graphically summarize the distribution of a univariate data set. The most common form of the histogram is obtained by splitting the range of the data into equal-sized bins,an interval into which a given data point does or does not fall. Then for each bin, the number of points from the data set that fall into each bin are counted. The bins can either be defined arbitrarily by the user or via some systematic rule. The histogram graphically shows center, spread, skewness and presence of outliers data. These features provide strong indications of the proper distributional model for the data set.

Surface normals can be easily calculated from linear least-squares methods. After the normalization, only two components of the resulting vector are relevant. The surface normals are represented as a pair of angles (φ, ϕ) in sphere coordinates, as shown in Fig.4.

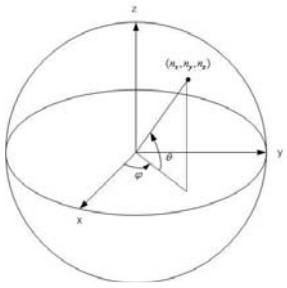

Fig. 4. Representation of normal vector in sphere coordinates.

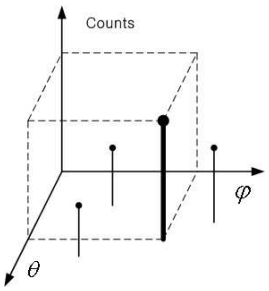

Fig. 5. Two-dimensional histogram.

The angles can be calculated as follows,

$$\begin{aligned} \varphi &= \arctan(\frac{n_y}{n_x}) \\ \theta &= \arcsin(n_z) \end{aligned} \tag{8}$$

By combining them in a two-dimensional histogram, we can obtain highly discriminative classifiers without having to solve a segmentation problem. All pairs of angles (φ, ϕ) of scene data is calculated and voted at the two-dimensional histogram shown in Fig.5. If some data points are from same surface plane, they have similar normal vectors. The points are fell into same bins and coarsely segmented into same region. The main motivation of histogram method is its low computational cost. Since similar surface patches are assigned to the same histogram cells, there is no need to extra segmentation process.

3.3. Building two-dimensional histogram

Before building the histogram, the normal vector should be assigned at each point. We use $P_{i,j} = (x(i, j), y(i, j), z(i, j))$ and around four neighborhood points, i.e. $P_{i-2,j}, P_{i,j-2}, P_{i+2,j}$ and $P_{i,j+2}$, where 'i' and 'j' are row and column index of the disparity image. The surface normal vector is uniquely identified using the linear least-squares method. If a point, $P_{i,j}$, is jump edge or crease edge, the patch surface plane from the $P_{i,j}$ and its neighborhood is not a planar surface. In this case, the point $P_{i,j}$ is disregarded. Fig.6. shows 3D data points with their normal vectors.

We use the two-dimensional histogram to coarsely segment the sets of points which are in the same planar surface. The two bins of histogram are θ and φ from Eqn.8, and they are digitized according to one degree. All the points except for edge points are voted into histogram. All counts data

that is exceeded a certain threshold are selected for the next step. We select counts data that is over 50. Fig.7 shows two-dimensional histogram.

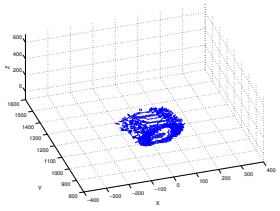

Fig. 6. 3D data with normal vector.

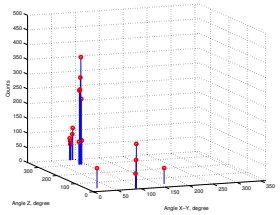

Fig. 7. Two-dimensional histogram with real data.

3.4. Planar surface segmentation

Using the distribution data sets of histogram, the planar surface is segmented. A data set represents a part of data points which are in a same planar surface.

Step1 All data points from a point of two-dimensional histogram are selected. They are formed a point set. Fig.8 shows a selected data points from histogram.

Step2 If all points in the set do not lie in the same planar surface, the set is disregarded. Otherwise a planar surface is built using all points of the point set.

Step3 Among all 3D scene data points, all points are selected if they are inside the planar surface.

Step4 *Step1-3* is applied to all points in a two-dimensional histogram. Fig.9 shows the result.

Step5 Similar planar surfaces are merged. Even though many planar surfaces are segmented, some of the surfaces are from same planar surface. They are merged into one surface. A surface normal vector and mean value of the data set are used in this step. Fig.10 shows the result.

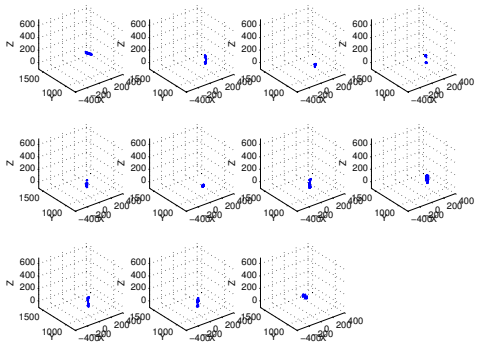

Fig. 8. Candidate surface from histogram.

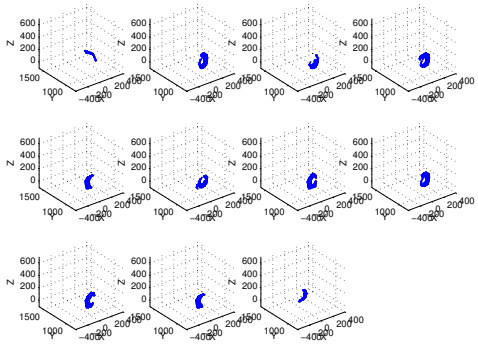

Fig. 9. Candidate surface from all data.

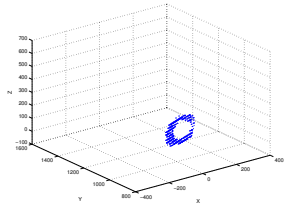

Fig. 10. Selected plane.

3.5. Assigning object features

In this step, object features are assigned to the segmented surface plane. Using the object features, the position and orientation of scene object is roughly determined. The object features consist of one feature point and three feature vectors. The feature point is defined at the most outer point from the mean point. The first feature vector is the surface normal. It is already known. The second feature vector is the unit vector from the feature point to mean point. The last feature vector is the cross product of the previous two features vectors. Fig.1 shows object features and segmented planar surface data points.

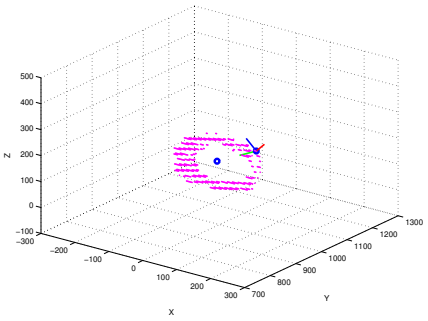

Fig. 11. Assigning object features.

4. Matching

4.1. Object Model

Once the appropriate descriptions are derived from the object scene data and the appropriate model, matching is done to recognize the object and its pose. It is performed in two steps, initial matching and fine adjustment.

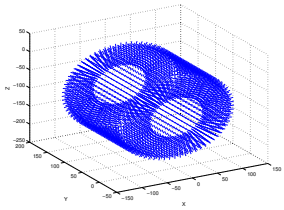

Fig. 12. Coil object model.

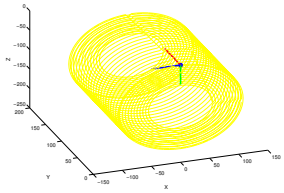

Fig. 13. Coil model features.

4.2. Initial matching

The purpose of the initial matching is to generate hypotheses of correspondences between model features and data features. The candidates for the position and orientation of an object are roughly calculated from the hypotheses. The object's position and orientation are expressed as a 4×4 transformation matrix, $\mathbf{T} = \begin{pmatrix} \mathbf{R} & \mathbf{t} \\ 0\ 0\ 0 & 1 \end{pmatrix}$, where $\mathbf{R}$ is a 3×3 rotation matrix and $\mathbf{t}$ is a 3D translation vector. Fig.14(Left) shows the movement of a model features(with subscript M) to an object features(with subscript D), where each features is expressed by three unit vectors $V1$, $V2$ and $V3$ from a feature point P. The rotation matrix R' and the translation vector t' are uniquely calculated from the following formulas,

$$\begin{aligned} t' &= P_D - P_M \\ V1_D &= R'\ V1_M \\ V2_D &= R'\ V2_M \\ V3_D &= R'\ V3_M \end{aligned} \tag{9}$$

A single pair of model features fixes a transformation. Fig.15

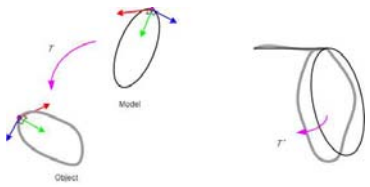

Fig. 14. Initial matching(Left) and Fine adjustment(Right).

shows a result of initial matching.

4.3. Fine adjustment

The fine adjustment process evaluates the validity of the hypotheses generated in the initial matching stage, and improves the accuracy of the transformation by an iteration method.

After the model is moved by an initial matching result T', all model points on the observable side of the object are selected

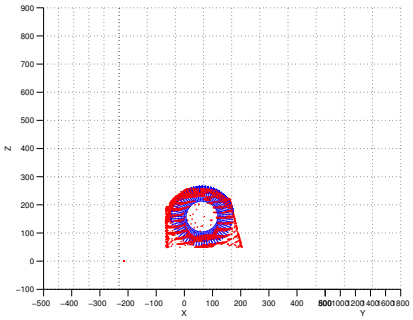

Fig. 15. Initial matching.

with satisfying the following equation,

$$(P - O) \cdot N \leq 0 \tag{10}$$

where P is the position of the model point, N is its normal direction, and O is camera position.

Next, we search for data points corresponding to the selected model points. If the 3D distance between the object point and the model point does not exceed a threshold value, the object point which has minimum distance among the model points is regarded as corresponding point to the model point.

The optimum transformation parameters $\mathbf{T''} = \begin{pmatrix} \mathbf{R''} & \mathbf{t''} \\ 0\ 0\ 0 & 1 \end{pmatrix}$ which move P_{M_i}, i.e. initial matched model points, to P_{D_i} can be estimated by minimizing the following error ε,

$$\varepsilon = \frac{1}{n} \sum_{i=1}^{n} dist(R''\ P_{M_i} + t'' - P_{D_i}) \tag{11}$$

$T = T''\ T'$ denotes the adjusted transformation. When the error ε is not small enough or the number of pairs n is few compared to the total number of selected model points, T is verified. T derived from the minimum ε is the final optimum transformation. From the result transformation T, we know the position and pose of the object. Fig.15 and Fig.16 show the fine adjustment result.

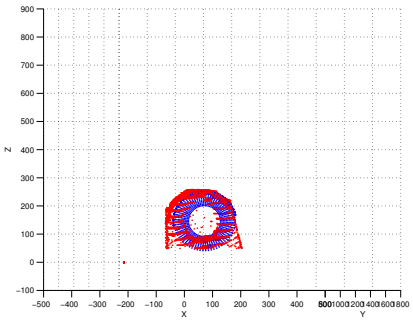

Fig. 16. Fine adjustment.

5. Experiment

5.1. Experimental equipment

Our recognition system consists of three parts,i.e. stereo CCD camera set, stereo matching part and 3D object recognition algorithm part.

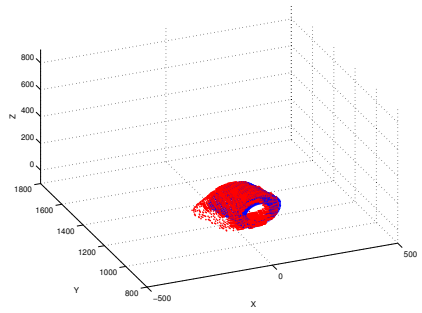

Fig. 17. Result.

Table 1. Positional errors.

	x(mm)	y(mm)	z(mm)	φ (deg.)	θ (deg.)
Error	2.9	4.0	6.2	2.1	1.6

The stereo matching board is made up of integration of FPGA, PCI interface and camera interface. It can be put in PCI slot at personal computer. FPGA module play an important role, computing and transferring a data to generate the disparity image, of the stereo matching board. And, It communicate with PC and camera using PCI and camera interface. Disparity image is produced at 15 frame per second of speed, 1280×100 of size and 200 level of 8bit gray scale disparity. The stereo CCD camera set consist of two zoom lens, the CCD cameras and camera mount equipment. Our CCD camera is PULNIX TM1320-15CL. It support high-resolution, high-speed progressive scan $1300(h) \times 1030(v)$ interline transfer CCD imager and 15 frames per second. Camera mount equipment is manually designed

5.2. Experimental results

Models for this experiment were built using the CAD-based modelling. The two models have around 2000 points and normal vectors at each point. The distance between neighborhood two points is about 5 mm.

We used one coil shaped object to evaluate the positional error of our recognition algorithm. The center of the planar surface is used. We measured the spacial the position of XYZ-coordinate and the rotation of planar surface normal vector. The rotation of the object is represented as two angles that is follows the same way with representation of normals in sphere coordinates, Fig.4. The results are shown in Table 1. The average position error of the XYZ-positions are 2.9 mm, 4.0 mm and 6.2 mm. The rotational error φ and ϕ are 2.1 deg. and 1.6 deg. We let our PARA robot pick and place the object, then we know that the result is accurate enough for a robot manipulator to treat the recognized object.

6. Conclusion

We propose a new method for 3D object recognition. We used the stereo vision system to sense a 3D scene object. The 3D points of the object are reconstructed from the disparity image. Planar surface is segmented and object features are assigned for matching. We used CAD-based modelling.

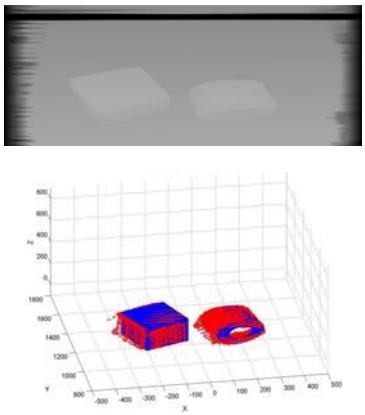

Fig. 18. Experimental result: A cube and a coil.

Matching process of our recognition system divided into two step. In initial matching, the position and orientation of object is roughly estimated. And they are improved through fine adjustment step. Our recognition system can recognize multiple objects at a same time and partially occluded object.

Our recognition system can be improved by extending the planar surface segmentation to generalized surface segmentation, and developing fast matching algorithm.

References

[1] Adam Hoover,Gillian Jean-Baptiste, "An Experimental Comparison of Range Image Segmentation Algorithms", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 18, no 7, pp.673-689, 1996.

[2] Hong Jeong and Yuns Oh, "Fast stereo matching using costraints in discrete space", IEICE Trans. on Information and Systems, vol.7, 2000

[3] M. Kalta and B.J. Davies, "Converting 80-Character ASCII IGES Sequential Files into More Conveniently Accessible Direct-Access Files," The International journal, advanced manufacturing technology ,vol 8,no 3 1993.

[4] Farshid Arman, J.K.Aggarwal, "Model-Based Object Recognition in Dense-Range Images - A Review," ACM Computing Surveys, vol.25, no 1, pp.5-43, 1993.

[5] Y.Sumi, Y.Kawai,T.Yoshimi,F.Tomita, "3D Object Recognition in Cluttered Environments by Segment-Based Stereo Vision," Internation Journal of Computer Vision, vol.46 , no 1, pp.5-23, 2001.

[6] J.Porrill,S.B.Pollard,T.P.Pridmore, "TINA: A 3D Vision System For Pick And Place," Image and Vision Computiog, vol.6 , no 2, pp.91-99, 1998.