A Particle Filter Based Probabilistic Fusion Framework for Simultaneous Recognition and Pose Estimation of 3D Objects in a Sequence of Images

Jeihun Lee, Seung-Min Baek, Changhyun Choi and *Sukhan Lee
Intelligent Systems Research Center
School of Information and Communication Engineering
Sungkyunkwan University, Suwon, KOREA
{ jeihun81, lsh}@ece.skku.ac.kr

Abstract — **We describe a framework for robust object recognition and pose estimation of 3D object which is using a sequence of images and probabilistic method. Using a sequence of images in multiple views has great advantage for robust object recognition and pose estimation in noisy and ill-conditioned environment - texture, occlusion, illuminate and camera pose. For recognizing an object and estimating its pose, we present it as a particle filter based probabilistic method with information of a sequence of images. It means that object pose represents probability distribution by particles in 3D space, and updated particles by consecutive observation in a sequence of images are converged to a single particle. The proposed method allows an easy integration of multiple evidences such photometric and geometric features as SIFT, color, 3D line, 2D square, and etc. The experimental results with stereo camera show the validity of the proposed method in an environment containing both textured and texture-less objects.**

I. Introduction

The object recognition has been one of the major problems in computer vision and intensively investigated for several decades. Although to recognize object have some problems, it has been developed toward real complex objects in cluttered scenes. There are several approaches to solve the problems of object recognition in real environment.

One of the approach for recognizing object is model based recognition method which is most general method. It recognizes the objects by matching features extracted from the scene with stored features of the object [1][2][3]. There are several methods to recognize object using predefined model information.

The method proposed by Fischler and Bolles [4] uses RANSAC to recognize objects. It projects points of all models on the scene and determines if projected points are close to those of detected scene and recognizes the object through this. This method isn't so efficient because of hypothesis and verification tasks several times. Olson [5] proposed pose clustering method for object recognition. This method recognizes object by producing pose space discretely and finding cluster including the object to search. As for disadvantages of this method, data size is quite big because pose space is 6-dimentional and pose cluster can be detected only when sufficient accurate pose is generated. In the next, David et al. [6] proposed recognition method that matching and pose estimation are solved simultaneously by minimizing energy function. But it may not be converged to minimum value in functional minimization method due to high non-linearity of cost function.

In addition, Johnson and Herbert [7] proposed a spin image based recognition algorithm in cluttered 3D scenes and Andrea Frome et al. [8] compared the performance of 3D shape context with spin-image. Jean Ponce et al. [9] introduced the 3D object recognition approach using affine invariant patches. Most recently, several authors have proposed the use of descriptor in image patch [10].

Another approach to recognition of object is local shape features based method which is inspired by the shape contexts of Belongie *et al.* [11]. At each edge pixel in an image, a histogram, or “shape context,” is calculated then each bin in the histogram counts the number of edge pixels in a neighborhood near the pixel. Nearest neighbor search and histogram distance measures then determine correspondences between shape contexts from a test image and shape contexts from model images [12]. But this method may not be effective when the background is concerned. To solve this problem, assessing shape context matching in high cluttered scene have studied [13] recently.

Except for above method, there are many of object recognition research. However, these almost methods are working well only at the condition with accurate 3D data or fully textured environments in single scene information with limited feature, while our approach makes us be able to recognize the object and estimate its pose overcoming a lot of noises and uncertainties from low-quality sensor in probabilistic method based on a sequence of images with multiple features.

The reminder of this paper is organized as follows : Section II outlines proposed framework for object recognition and its pose estimation. Section III describes our experiments for emphasizing advantage of our proposed framework. Section IV summarizes the framework within which we evaluate this and other methods. Section V concludes by discussing scalability and implementation issues along with directions for future works.

II. Proposed Framework Overview

A. Outline of proposed Approach

In the object recognition literature some of probabilistic approaches to object recognition or pose estimation have been reported [15][16][17]. The works of [17] and [18] use maximum a posteriori (MAP) estimation under a Markov random field (MRF) model. Especially the former uses MRF as a probabilistic model to capture dependencies between features of the object model and employs MAP estimation to find the match between the object and scene. Schiele and Crowley [19] have developed a probabilistic object recognition technique using multidimensional receptive field histograms. Although this technique has been shown to be somewhat robust in the face of change in rotation and scale with low cost of computation, it only computes the probability of the presence of an object.

We proposed a probabilistic method based on a sequence of images to recognize an object and to estimate its pose in our previous work [20]. But the previous framework simply uses a ratio of matched features to total features when it assigns a similarity weight to each particle. The main contribution of this paper is to propose a more systematic recognition framework which considers not only matched features but also matched pose errors. The proposed method handles the object pose probabilistically. The probabilistic pose is drawn by particles and is updated by consecutive observations extracted from a sequence of images. The proposed method can recognize not only textured but also texture-less objects because the particle filtering framework of the proposed method can deal with various features such as photometric features (SIFT-Scale Invariant Feature Transform [10], color) and geometric features (line, square) [14].

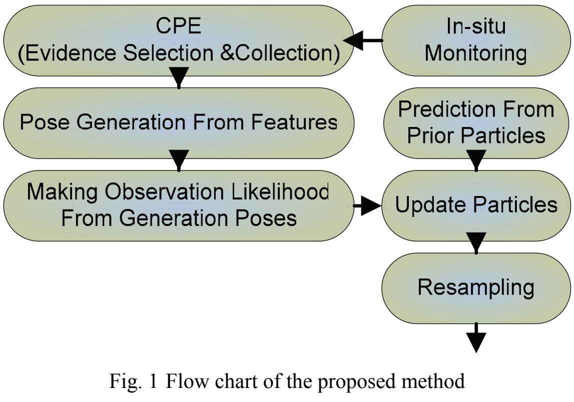

CPE
(Evidence Selection &Collection)
In-situ Monitoring
Pose Generation From Features
Prediction From Prior Particles
Making Observation Likelihood From Generation Poses
Update Particles
Resampling

Fig. 1 Flow chart of the proposed method

Fig. 1 illustrates a flow chart of the proposed method composed of seven procedures. First of all, the information of circumstance - density of texture, illumination and distance of expected object pose - is calculated from input image and 3D point cloud in In-situ Monitoring. Then the valid features in an input image are selected by the Cognitive Perception Engine (CPE) which perceives automatically an environment from information of In-situ Monitoring and keeps the evidences of all objects for their recognition. We assume that the valid features for each object in a current scene are already defined to the CPE. The multiple poses are generated by features extracted from a scene and 3D point cloud. These poses are used for making observation likelihood. The particles representing the object pose are propagated from the previous state using the motion information. The weights are assigned to the predicted particles. Finally, we resample the particles according to their weights for obtaining important particles. These procedures are repeated until the particles are converged to a single pose.

B. In-situ Monitoring

The main role of In-situ Monitoring is simply to check the changes of environment such as illumination, a mount of texture and distance of between robot and assumed object. In this paper, we divide uniformly input image into 25 areas, 5 columns and 5 rows, and calculated values are used for selection of valid feature or feature set. The illumination means intensity information in current image that calculates not absolute value but relative such as changes of environment. Amount of texture in each block is counted pixel which is processed by Canny edge image of current frame. Lastly, we assume that existence possibility of object is high if amount of texture is abundant in particular block. So distance of each block is calculated using processed image pixel with valid 3D point cloud and average those values.

C. Cognitive Perception Engine

We are currently developing the CPE. We assume that the valid features for each object in a current scene are already defined to the CPE. But the main role of CPE is selection of proper feature or feature set using distance information from In-situ monitoring. We have 3 valid features - color, line and SIFT - which are selected one or more automatically for target object. Process of strategy for feature selection is very simple which use only information of distance from texture up to now. If the distance were far, then CPE selects color feature to recognize. On the other hand, SIFT or line feature is used for recognizing in near alternatively. Sometimes all features are used to recognize. What is important thing is asynchronous control of whole processes by the circumstance.

D. Particle Filtering Framework

Basically, Particle Filtering Framework is almost same as mentioned in [20]. The recognized object pose is estimated by particle filtering in a sequence of images over time in order that we represent the object pose with an arbitrary distribution. We keep a formulation of Motion model and Observation model in [20] which is most important parts in proposed particle filter based framework.

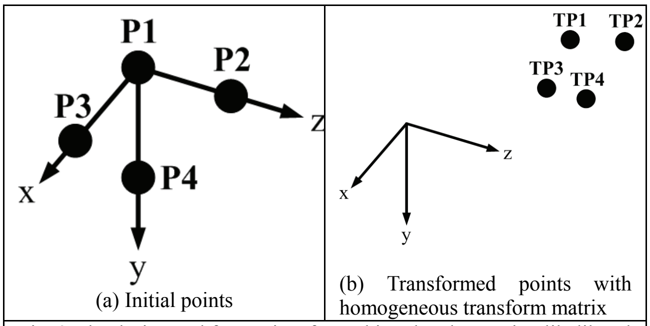
But similarity assignment of Observation model is not only very heuristic but also so experimental in the previous research. In this paper, we improved it using Bayesian theorem and probabilistic approach.

1) Observation Likelihood

We define the observation likelihood $p(\mathbf{Z}_t \mid \mathbf{O}_t^{[i]})$ in previous work [20] :

$$p(\mathbf{Z}_t \mid \mathbf{O}_t^{[i]}) = \sum_{j=1}^{m} w_j \exp\left[-0.5 \times \sum_{l=1}^{4} \left\{ \begin{array}{l} (\mathrm{Ob\_TP}j - \mathrm{St\_TP}j)^T \\ \times \mathbf{S}_j^{-1} (\mathrm{Ob\_TP}j - \mathrm{St\_TP}j) \end{array} \right\} \right] \quad (1)$$

Where w_j is the similarity weight related to transformed points with $\mathbf{O}^{[j]}$. Where *m* is the number of generated its poses at time *t*. Here, we designate four points (P1, P2, P3, P4) at camera frame as Fig. 2. The four points are transformed by the homogeneous transform matrix parameterized by the six spatial degrees of freedom. Fig. 2 (b) shows the transformed points (TP1, TP2, TP3, TP4) with an arbitrary homogeneous transform matrix. We obtain the set of the four points (TP1, TP2, TP3, TP4) transformed from (P1, P2, P3, P4). Let (Ob_TP1[i], Ob_TP2[i], Ob_TP2[i], Ob_TP2[i]) represent the transformed points with $\mathbf{O}^{[i]}$ while (St_TP1[i], St_TP2[i], St_TP2[i], St_TP2[i]) mean those with $\mathbf{O}_t^{[i]}$.



(a) Initial points

(b) Transformed points with homogeneous transform matrix

Fig. 2 The designated four points for making the observation likelihood

2) Similarity Assignment

To assign similarity, we consider how much correspondence between the recognized object and its estimated pose and real ones, respectively. In probabilistic terms the goal of proposed method is to estimate object pose which yield the best interpretation of object pose generated by multiple features in Bayesian sense. Our particle filter based probabilistic method framework approximate variant of the following posterior distribution.

$$w_j = p(O_{t,object} \mid E) = p(O_{t,id}, O_{t,pose} \mid E) \quad (2)$$

Where O_{object} is an object to recognize, it is divided O_{id} and O_{pose} for information of recognition and pose estimation respectively. The O_{id} means whether recognized object is correct or not and O_{pose} means precision level of estimated object pose. Where E denote the evidence, measurement, redefined $E = \{Z_1, Z_2 \cdots Z_n\}$ indicates multiple features.

In other words, the O_{id} means a process of object recognition whether it is the aimed object to recognize or not. The O_{pose} is generated by accuracy rate of estimated object pose. To represent similarity weight, we assume that O_{id} and O_{pose} are independent events because object identification is considered separately as pose estimation. That means that the very well recognized object does not guarantee accurate estimation of object pose, vice versa. According to this assumption, the similarity is represented as follow :

$$p(O_{t,id}, O_{t,pose} \mid E) = p(O_{t,id} \mid E) p(O_{t,pose} \mid E) \quad (3)$$

3) Resampling

Detail method will be mentioned next paragraph, because the similarity calculation of each evidence is quite different. Resample process of particles is also same method in [20].

III. Object Matching Similarity From Features

A. Similarity assignment from SIFT feature

The object pose can be generated by calculating a transformation between the SIFT features [14] measured at current frame and the corresponding ones in the database. The transformation is represented by a homogeneous transform matrix. The object pose can be generated using corresponded 3D point clouds from depth image if the matched features are 3 or more in 2D image [20].

If one scene has several candidates that have matched SIFT features, then all these candidates generate 3D poses for probabilistic fusion at particle filtering stage, as described in previous section. However, to assign similarity weight to each candidate, posterior distribution should be calculated in equation (2).

For example, when an object is shown in the scene, measured average number of matched SIFT is 23 as $p(O_{t,id} \mid E_{SIFT})$, and average distance error is 5mm with certain variation by many trials as $p(O_{t,pose} \mid E_{SIFT})$. Then, the posterior distribution $p(O_{t,object} \mid E_{SIFT})$ can be obtained by equation (3), and the shape of probability distribution of example case is shown in Fig. 3.

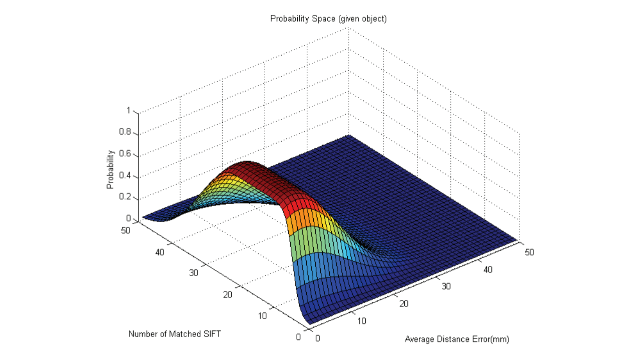


Fig. 3 obtained $p(O_{t,object} \mid E_{SIFT})$ graph by experimental result

B. *Similarity assignment from line feature*

Assigning similarity method of Line feature is conducted the same process with SIFT. But there are two kinds of hypothesis about object identification, $p(O_{t,id} \mid E_{Line})$ and pose accuracy, $p(O_{t,pose} \mid E_{Line})$. We define first one as a Coverage that means how many matched line with information of model line. The Coverage can be calculated by equation (4) as follow :

$$Coverage = \frac{Matched\ line\ distance}{Total\ line\ distance\ of\ model} \tag{4}$$

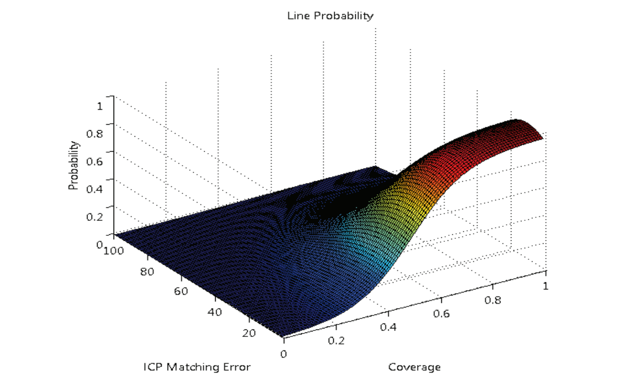


Fig. 4 obtained $p(O_{t,object} \mid E_{Line})$ graph by experimental result

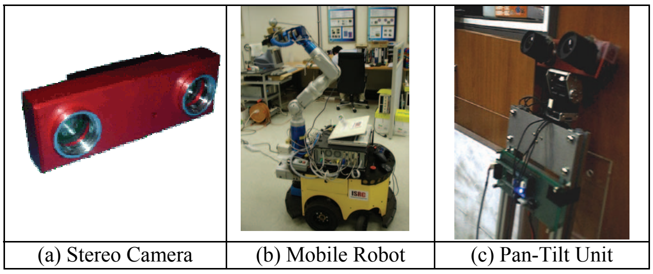
If the Coverage is very high, then the probability of object identification is increased. And the second one is defined as ICP matching error, because we use ICP for line matching. Line matching can find several matched set like SIFT in the single scene. So, $p(O_{t,object} \mid E_{Line})$ can be obtained by equation (3) in each candidate and is represented as a joint probabilistic in Fig. 4.

C. *Similarity assignment from color information*

The object with a particular color can be segmented by the color in the current scene. Although the segmented region can not provide an object's orientation, the object's location can be generated using the segmented region from corresponded depth image. In homogeneous transform matrix, the rotation part is defined by an identity matrix and the translation part represents an object's location as a center of segmented area. Information of translation matrix can be approximated average of valid 3D points in segmented area. If there is no valid point in segmented area, it is not assigned similarity. The similarity weight for *j*th object location, w_j , is denoted as a predefined constant with a small value in comparison with the similarity weight of the object pose generated by the other features. In particular, the color information can be combined with the other features.

IV. Experimental Results

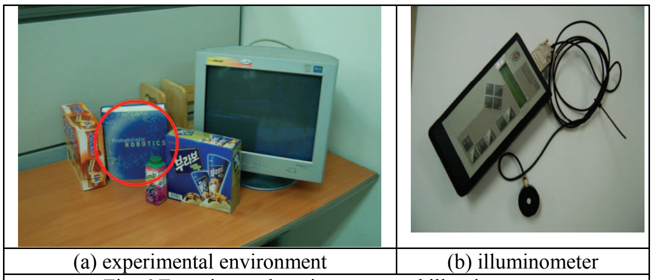
This paper focuses on recognizing an object while estimating its pose concurrently in a sequence of images. The proposed method is tested in textured and texture-less objects. The robot used in the experiment is a PowerBot–AGV with a Videre stereo camera mounted on the pan-tile unit configuration as is seen Fig. 5. The camera motion information is calculated by the internal encoder.



(a) Stereo Camera (b) Mobile Robot (c) Pan-Tilt Unit

Fig. 5 Equipments for Experiment

For our experiments, we decorated clustered environment as Fig. 6(a), and used illuminometer Fig. 6(b) for measuring change of illumination in environment. Our aimed object to recognize is like rectangular parallelepiped, red circled blue book in Fig. 6(a), which has textured front side and texture-less back side. To emphasize advantages of proposed framework, the experiment is conducted various conditions in accordance with changes of illumination, amount of texture and distance. First experiment is conducted as recognizing textured object, front side of the book, with changing illumination, 330 lx and 120 lx, and distances of recognition, about 0.5 meters, 1.0 meters and 1.6 meters and its result illustrates in Fig. 7. We turn front side of the book back in order to carry out second experiment, Fig. 8, recognizing the texture-less side of the target object.



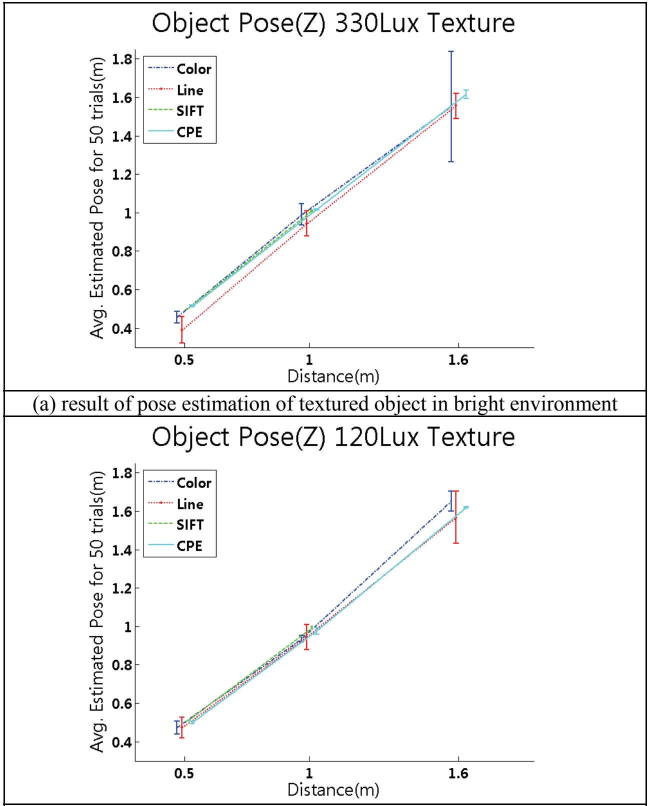
(a) experimental environment (b) illuminometer

Fig. 6 Experimental environment and illuminometer

We make CPE strategy with 3 evidences and its combination. That means that the recognition and pose estimation of 3D object are performed Color, Line, SIFT, Color with Line and Color with SIFT features. These features are selected automatically in accordance with illumination and distance by CPE in the proposed framework. CPE selects SIFT in the close distance to object and bright environment. If the distance is far, CPE use Color with SIFT, Line feature or Color with Line features for object recognition and its pose estimation. On the other hand, if the distance information from in-situ monitor is over the 1.0 meters, Line or Color Line feature are selected by CPE in dark environment.

Fig. 7 and Fig. 8 show that the each evidence - Color, Line and SIFT – has characteristic such as accuracy, effective distance and illumination for recognition. Color feature have the advantage for changes of not only illumination and distance but a number of 3D point clouds. But Color feature cannot estimate object pose because it cannot identify the object. Therefore, variance of Color feature is much wider than others. Line feature has a good performance for recognition and its pose estimation with small variance. But it sometimes mismatch between real object and similar ones because it do not have identification ability of object. Whereas other evidences, SIFT feature has very good performance ability to identify object. So pose estimation combined with 3D point cloud from high precision sensor is very accurate. The results from our experiments in Fig. 7 and Fig. 8 show correctness of recognition and pose estimation, despite low precision and changeable depth information of stereo camera.

But some conditions such as far distance and low illumination are the fact that should be overcome in order to improve the recognition performance. The further robot is from the object, the wider variances are in the result of Color and Line features, Fig. 7 and Fig. 8. In dark environment is also challenging in recognition problem. Note that SIFT method cannot recognize target object when robot locates far from object and in low illuminated place in Fig. 7 (b) and Fig. 8(b).

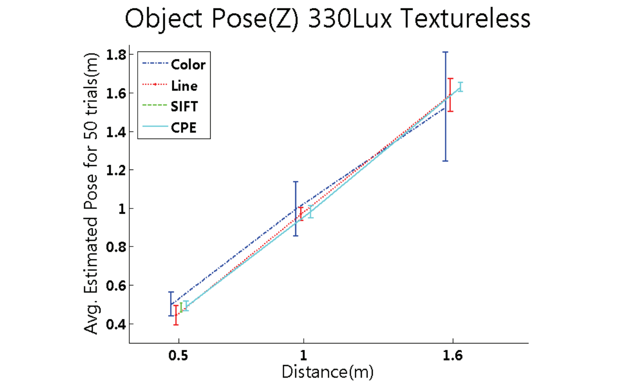


(a) result of pose estimation of textured object in bright environment

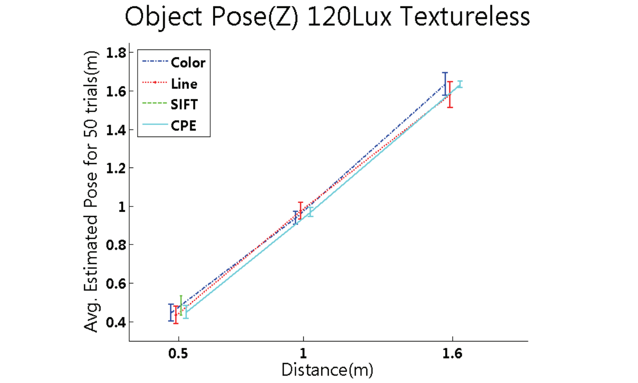
(b) result of pose estimation of textured object in dark environment

Fig. 7 The ARMSE for the textured object pose in accordance with illumination

The results from CPE have good performance in any circumstance. Automatically selected optimal feature or features set are properly achieved according to purpose of proposed framework. The estimated pose is approximated near the real with narrow variances in all cases. In some cases, CPE results better than SIFT.



(a) result of pose estimation of textureless object in bright environment



(b) result of pose estimation of textureless object in dark environment

Fig. 8 The ARMSE for the textureless object pose in accordance with illumination

V. CONCLUSION

We have concentrated on developing a probabilistic method using multiple evidences based on sequence of images to recognize an object and to estimate its pose. Especially in order to design more systematic framework we have improved the previous probabilistic method by considering both the ratio of matched features to total features and matched pose error in assigning similarity weight. The proposed method represents probabilistically the recognized object pose with particles to draw an arbitrary distribution. The particles are updated by consecutive observations in a sequence of images and are converged to a single pose. The proposed method can recognize various objects with individual characteristics because it can incorporates easily multiple features such as photometric features (SIFT, color) and geometric features (line, square) into the proposed filtering framework. We experiment the proposed method with a stereo camera in an experimental environment including textured and texture-less objects with not only changes of illumination but also variation of distance from object. The experiment result demonstrates that the proposed method recognizes robustly various objects with individual characteristics such as textured and texture-less objects in various environments.

ACKNOWLEDGMENT

This work is supported by the Intelligent Robotics Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy of Korea. This work is also supported in part by the Science and Technology Program of Gyeonggi Province as well as in part by the Sungkyunkwan University. And this work was also partly supported by Brain Korea 21 (BK21) project. The authors would like to thank Hyunjun Kim, undergraduate researcher, at ISRC of Sungkyunkwan Univ.

REFERENCES

[1] M. F. S. Farias and J. M. de Carvalho, “Multi-view Technique For 3D Polyhedral Object Rocognition Using Surface Representation,” *Revista Controle & Automacao.*, pp. 107-117, 1999.

[2] Y. Shirai, “Three-Dimensional Computer Vision,” New York: Springer Verlag.

[3] J. Ben-Arie, Z. Wang, and R. Rao, “Iconic recognition with affine-invariant spectral,” *In Proc. IAPR-IEEE International Conference on Pattern an Recognition*, volume 1, pp. 672-676, 1996.

[4] M. A. Fischler and R. C. Bolles. “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography,” *Comm. Assoc. Comp*. Mach, 24(6):381-395, 1981.

[5] C.F. Olson. “Efficient pose clustering using a randomized algorithm,” *IJCV*, 23(2):131-147, June 1997.

[6] P. David, D. F. DelMenthon, R. Duraiswami, and H. Samet. Softposit: Simultaneous pose and correspondence determination. In 7th ECCV, volume III, pages 698-703, Copenhagen, Denmark, May 2002.

[7] A. E. Johnson and M. Hebert, “Using spin images for efficient object recognition in cluttered 3D scenes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, May 1999.

[8] A. Frome, D. Huber, R. Kolluri, T. Bulow, and J. Malik. “Recognizing Objects in Range Data Using Regional Point Descriptors,” *To appear in European Conference on Computer Vision, Prague*, Czech Republic, 2004.

[9] Fred Rothganger, Svetlana Lazebnik, Cordelia Schmid, Jean Ponce, “3D Object Modeling and Recognition Using Affine-Invariant Patches and Multi-View Spatial Constraints,” *CVPR*, volume 2, pp. 272-280, 2003.

[10] D. Lowe. “Object recognition from local scale invariant features,” In Proc. 7th International Conf. Computer Vision (*ICCV’99*), pp.1150–1157, Kerkyra, Greece, September 1999.

[11] S. Belongie, J. Malik, and J. Puzicha, “Shape matching and object recognition using shape contexts”, *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 24(4):509-522, April 2002.

[12] Owen Carmichael and Marial Herbert, “Shape-Based Recognition of Wiry Object,” IEEE PAMI, May 2004.

[13] A. Thayananthan, B. Stenger, P.H.S. Torr, and R. Cipolla, “Shape context and chamfer matching in cluttered scenes,” *In Proc. IEEE Conference On Computer Vision and Pattern Recognition*, 2003.

[14] Sukhan Lee, Eunyoung Kim, and Yeonchool Park, “3D Object Recognition using Multiple Features for Robotic Manipulation,” *IEEE International Conference on Robotics and Automation*, pp. 3768-3774

[15] Clark F. Olson, “A probabilistic formulation for Hausdorff matching,” In IEEE Conference on Computer Vision and Pattern Recognition, pp.150-156,1998.

[16] Jayashree Subrahmonia, David B. Cooper, and Daniel Keren, “Practical reliable bayesian recognition of 2D and 3D objects using implicit polynomials and algebraic invariants,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(5):505-519, May 1996.

[17] Y.Boykov, D.P.Huttenlocher, “A New Bayesian Framework for Object Recognition,” pp.517-523, CVPR99.

[18] S. Z. Li and J. Hornegger. "A two-stage probabilistic approach for object recognition". *In H. Burkhard and B. Neumann, editors, Computer Vision- ECCV98*, vol. II of Lecture Notes in Computer Science, pp 733-747, Heidelberg, 1998.

[19] B. Schiele and J. L. Crowley. “Probabilistic object recognition using multidimensional receptive field histograms”, *ICPR96*, August 1996.

[20] Sukhan Lee, Seongsoo Lee, Jeihun Lee, Dongju Moon, Eunyoung Kim and Jeonghyun Seo, “Robust Recognition and Pose Estimation of 3D Objects Based on Evidence Fusion in a Sequence of Images,” To be appeared in *ICRA07*.