Particle Filter Based Robust Recognition and Pose Estimation of 3D Objects in a Sequence of Images

Jeihun Lee[1], Seung-Min Baek[1], Changhyun Choi[1] and Sukhan Lee[1]

Intelligent Systems Research Center, School of Information and Communication Engineering, Sungkyunkwan University, Suwon, KOREA
{jeihun81,smbaek,lsh}@ece.skku.ac.kr

Summary. A particle filter framework of multiple evidence fusion and model matching in a sequence of images is presented for robust recognition and pose estimation of 3D objects. It attempts to challenge a long-standing problem in robot vision, so called, how to ensure the dependability of its performance under the large variation in visual properties of a scene due to changes in illumination, texture, occlusion, as well as camera pose. To ensure the dependability, we propose a behavioral process in vision, where possible interpretations are carried probabilistically in space and time for further investigations till they are converged to a credible decision by additional evidences. The proposed approach features 1) the automatic selection and collection of an optimal set of evidences based on in-situ monitoring of environmental variations, 2) the derivation of multiple interpretations, as particles representing possible object poses in 3D space, and the assignment of their probabilities based on matching the object model with evidences, and 3) the particle filtering of interpretations in time with the additional evidences obtained from a sequence of images. The proposed approach has been validated by the stereo-camera based experimentation of 3D object recognition and pose estimation, where a combination of photometric and geometric features are used for evidences.

1 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 about object recognition in real environment. One of the most common approach for recognizing object from a measured scene is a model based recognition 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

from all models to the scene and determines if projected points are close to those of detected scene. Then recognizes the object through this. This method is not so efficient because of iterative hypothesis and verification tasks. Olson [5] proposed pose clustering method for object recognition. This method recognizes object by producing pose space discretely and finding the cluster which is including the object. As for disadvantages of this method, data size is quite big since pose space is 6-dimentional and pose cluster can be detected only when sufficient accurate pose becomes 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 by 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 recognize the 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. After searching nearest neighbor and measuring histogram distance, 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 researches. However, most of these methods are working well only at the condition under accurate 3D data or fully textured environments in single scene information with limited feature. It means that 2D or 3D measurement data from real environment contains noisy and uncertain information caused by changes of illumination, amount of texture, distance to the object and etc. Therefore, in this paper, we try to find a solution for simultaneous recognition and pose estimation of 3D object in a real environment conditions. The reminder of this paper is organized as follows : Section II outlines proposed framework for object recognition and its pose estimation. Section III explains how to assign matching similarities from different features. Description of experimental data and a feasibility analysis of our proposed framework are presented in section IV. Section V concludes by discussing scalability and implementation issues along with directions for future works.

2 Proposed Framework Overview

2.1 Outline of proposed Approach

Some of literatures about probabilistic approaches to recognize object or estimate its pose 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 the dependencies between features of the object model, and employs MAP estimation to find the match between the object and a 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 to change of 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].

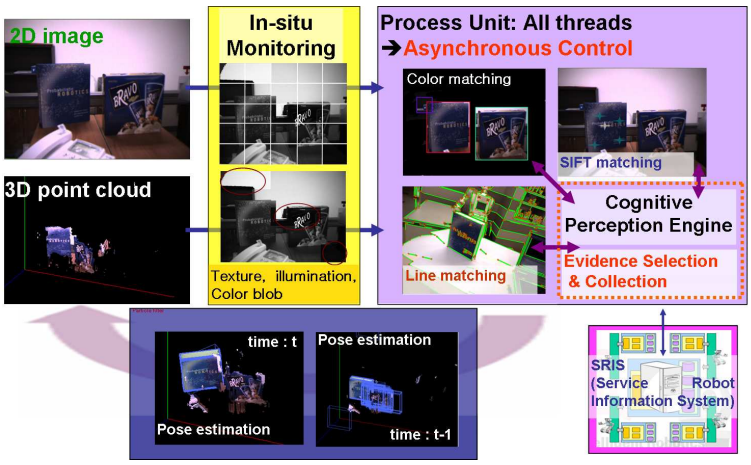


Fig. 1. Block diagram of 3D recognition framework.

Fig. 1 illustrates block diagram of the overall 3D recognition framework. 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 an environment automatically by using information offered by In-situ Monitoring and keeps the evidences of all objects for their recognition. Valid features for recognizing the object are stored in Service Robot Information System (SRIS) and CPE uses this information as a priori-knowledge. The multiple poses are generated by features extracted from a scene and 3D point cloud. And probabilistic estimation is done by particle filter using measured poses and propagated probability distribution of target object in a sequence of images.

2.2 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 between robot and target object. In this paper, we divide input image into 25 areas uniformly, 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 from processed image pixel with valid 3D point cloud and average of those values.

2.3 Cognitive Perception Engine

We assume that the valid features for recognizing each object in a current scene are already defined to the CPE. In Fig. 1 this information could be delivered from SRIS. The main role of CPE is selection of proper feature or evidence set by using information from In-situ monitoring. We have 3 valid asynchronous processing pathes - color, line and SIFT - which are selected one or more automatically based on the target object and information from in-situ monitoring. For example, if the distance is far, then CPE selects color feature. On the other hand, SIFT or line features are more helpful for recognizing near or middle range object. Sometimes all three features could be used for getting maximal information from current scene. However, the processing time is longer than any other set of evidences. So, it is a trade-off problem between performance and time consumption. In this paper, strategy for feature selection is done by using distance and illumination to simplify a feasibility analysis.

2.4 Particle Filtering

Particle filtering procedure is presented in previous papers [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 the most important part of proposed particle filter based framework. In this paper, we improved the way how to assign similarity weight of measured features using Bayesian theorem and probabilistic approach.

Observation Likelihood

We define the observation likelihood $p(Z_t|O_t^{[i]})$ as in previous work [20] :

$$p(Z_t|O_t^{[i]}) = \sum_{j=1}^{m} w_j \cdot exp\left[\frac{-1}{2} \cdot \sum_{l=1}^{4} \left\{ \begin{array}{c} (Ob\_TP_l^j - St\_TP_l^i)^T \\ \times S_l^{-1} \cdot (Ob\_TP_l^j - St\_TP_l^i) \end{array} \right\}\right] \quad (1)$$

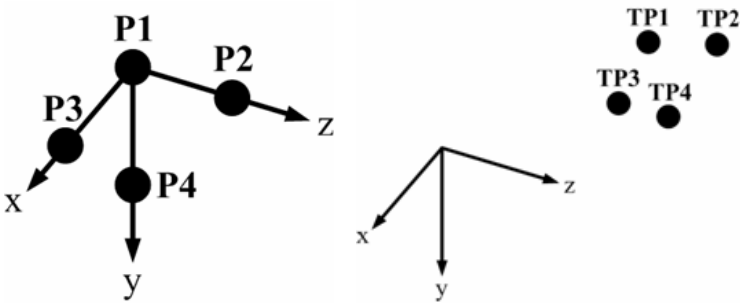
Where w_j is the similarity weight related to transformed points with $O^{[j]}$. Where m is the number of generated poses at time t. Here, we designate four points ($P1$, $P2$, $P3$, $P4$) at camera frame as Fig. 2. The four points are transformed by the homogeneous transform matrix parameterized by 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 ($Ob\_TP1[i]$, $Ob\_TP2[i]$, $Ob\_TP3[i]$, $Ob\_TP4[i]$) represent the transformed points with $O^{[i]}$ while ($St\_TP1[i]$, $St\_TP2[i]$, $St\_TP3[i]$, $St\_TP4[i]$) mean the transformed points with $O_t^{[i]}$.

Similarity Assignment

To assign similarity, we consider how much correspondence exists 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.

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

Where object O_{object} is an object to recognize, it is divided into 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,



(a) Initial points (b) Transformed points with homogeneous transform matrix

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

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 :

$$p(O_{t,id}, O_{t,pose}|E) = p(O_{t,id}|E)p(O_{t,pose}|E) \tag{3}$$

3 Object Matching Similarity from Features

3.1 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 by 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 $p(O_{t,id}|E_{SIFT})$, and average distance error is 5mm with certain variation by many trials as $p(O_{t,pose}|E_{SIFT})$. Then, the posterior distribution $p(O_{t,object}|E_{SIFT})$ can be obtained by equation (3), and the shape of probability distribution of example case is shown in Fig. 3.

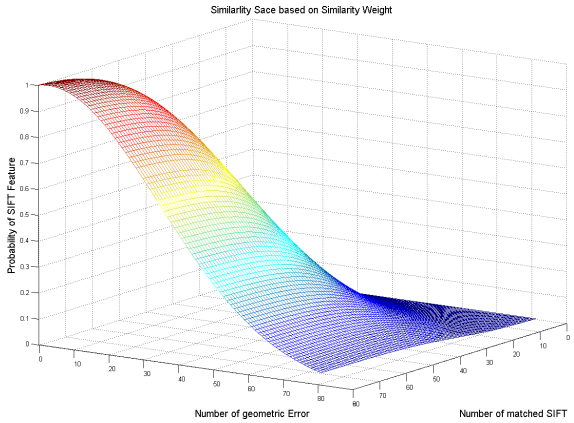


Fig. 3. obtained $p(O_{t,object}|E_{SIFT})$ graph by experimental result.

3.2 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, $p(O_{t,id}|E_{Line})$ and pose accuracy, $p(O_{t,pose}|E_{Line})$. 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 :

$$Coverage = \frac{Matched\_line\_length}{Total\_line\_length\_of\_model} \tag{4}$$

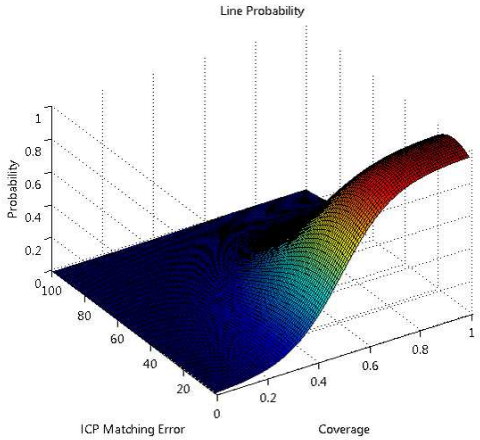


Fig. 4. obtained $p(O_{t,object}|E_{Line})$ graph by experimental result.

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

3.3 Similarity assignment from color information

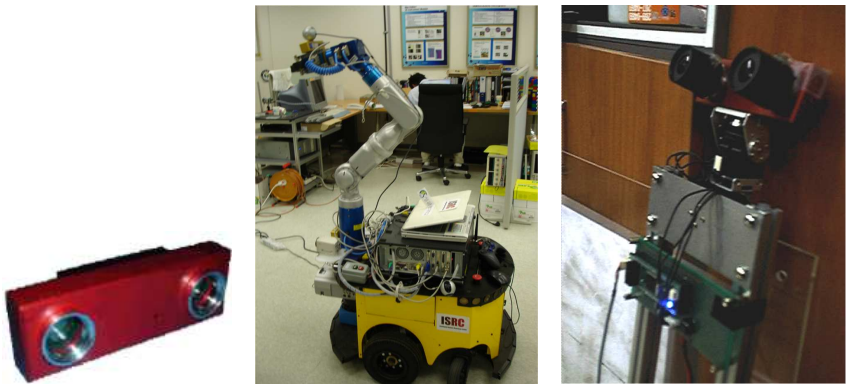
The object with a particular color can be segmented by the color in the current scene. Although the segmented region can not provide an objects orientation, the objects location can be generated by using the segmented region from corresponded depth image. In homogeneous transform matrix, the rotation part is defined by an identity matrix and the translation part represents an objects 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 as similarity. The similarity weight for jth 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.

4 Experimental Results

This paper focuses on simultaneous recognition of target object and estimation of its pose in a sequence of images. The proposed method is tested to recognize textured and textureless objects in various distance and illumination condition. The robot used in the experiment is a PowerBot AGV with a Videre stereo camera mounted on the pan-tile unit configuration as shown in Fig. 5. The camera motion information is calculated by the internal encoder.

For an evaluation of the proposed method, we set up cluttered environment as Fig. 6 (a), and used illuminometer Fig. 6(b) for measuring change of illumination in environment. The target object to recognize is red circled blue book which is rectangular parallelepiped as shown in Fig. 6(a). The book has textured front side and texture-less rear side.

We made CPE strategy for selection of processing passes of three features and its combination. It means that the recognition and pose estimation of 3D object are performed by either basic features such as Color, Line and SIFT, or combined features such as Line + Color, SIFT + Color, and SIFT + Line. 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



(a) Stereo camera (b) Mobile robot (c) Pan/Tilt unit

Fig. 5. Experimental setups.

(a) Experimental environment (b) Illuminometer

Fig. 6. Experimental environment and illuminometer.

from in-situ monitor is over the 1.0 meters, Line or Color + Line features are selected by CPE in dark environment.

Fig. 7 indicates the experimental results of object recognition under bright illumination condition. 'Mean' and 'Var.' in the table represents average distance and variation respectively. 'O' and 'X' indicates amount of detected texture from the target object. 'O' means object has enough textures, whereas 'X' means object does not have enough textures. The cross marked cells mean that the robot is not able to recognize target object by using the selected evidence. Since the number of valid SIFT features are more extracted under the 330 lux illumination than other dark illuminations, the 330 lux can be seen as a proper illumination condition for the recognition using SIFT as a feature. Line features are also more reliably detected at 330 lux case than 120 lux case. Fig. 8 shows the experimental results under the darker illumination

Feature	Distance	0.5		1.0		1.6	
		Texture					
		O	X	O	X	O	X
SIFT	Mean	0.50	0.48	1.01			
	Var.	0.002	0.02	0.002			
Color	Mean	0.46	0.50	0.99	0.94	1.55	1.53
	Var.	0.03	0.06	0.05	0.23	0.29	0.29
Line	Mean	0.39	0.44	1.44	1.04	1.68	1.59
	Var.	0.07	0.05	0.27	0.03	0.07	0.09
Line +Color	Mean	0.47	0.45	0.94	0.98	1.61	1.63
	Var.	0.05	0.04	0.06	0.03	0.02	0.02
SIFT +Color	Mean	0.51	0.49	1.02	0.97		
	Var.	0.005	0.03	0.004	0.04		
SIFT +Line	Mean	0.519	0.49	1.07		1.66	
	Var.	0.017	0.03	0.18		0.20	

*Distance Unit = meter

Fig. 7. Experimental results for object recognition under 330 lux illumination condition.

condition. The gray painted cells are the chosen evidences by CPE for optimal feature selection. This selection aptly shows that CPE's choices are reliable for object recognition.

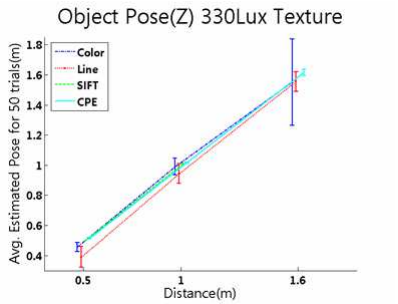
Feature	Distance	0.5		1.0		1.6	
		Texture					
		O	X	O	X	O	X
SIFT	Mean	0.51	0.48	0.99			
	Var.	0.02	0.03	0.003			
Color	Mean	0.47	0.50	0.93	1.00	1.62	1.55
	Var.	0.05	0.07	0.06	0.14	0.13	0.28
Line	Mean	0.47	0.45	0.94	0.97	1.57	1.59
	Var.	0.05	0.05	0.06	0.03	0.13	0.08
Line +Color	Mean	0.50	0.45	0.97	0.97	1.62	1.65
	Var.	0.01	0.03	0.02	0.02	0.02	0.02
SIFT +Color	Mean	0.51	0.49	1.04	0.97		
	Var.	0.03	0.03	0.07	0.04		
SIFT +Line	Mean	0.51	0.48	1.00			
	Var.	0.02	0.03	0.09			

*Distance Unit = meter

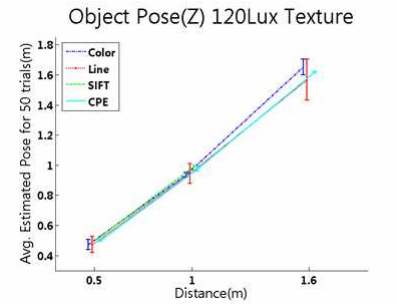
Fig. 8. Experimental results for object recognition under 120 lux illumination condition.

First experiment is conducted as recognizing textured object, front side of the book, with changing illumination, 330 lux and 120 lux, and distances of recognition, about 0.5 meters, 1.0 meters and 1.6 meters and its result illustrates in Fig. 9. The proposed method also tested for recognizing tex-

tureless case, rear side of the book, the results are shown in Fig. 10. Means and variances of estimated poses are described in each figure. So, we can see the variations of performance caused by selected set of evidences in different conditions.

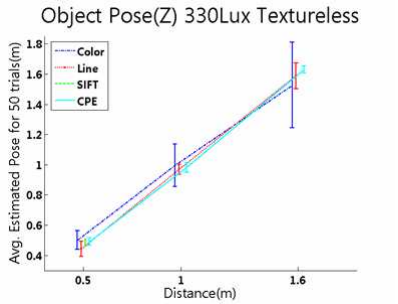


(a) Results of pose estimation of textured object in bright environment

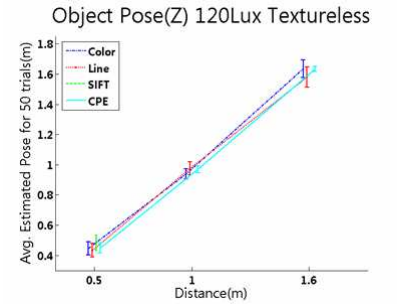


(b) Results of pose estimation of textured object in dark environment

Fig. 9. The ARMSE for the textured object pose in accordance with illumination.



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



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

Fig. 10. The ARMSE for the textureless object pose in accordance with illumination.

Fig. 9 and Fig. 10 show that the each evidence - Color, Line and SIFT - has characteristic such as accuracy, effective distance and illumination for recognition. Color feature has the advantage for changes of not only illumination and distance but also 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 sometimes mismatches between real object and similar ones are detected because it does not have object identification capability. Whereas other evidences, SIFT feature has very good performance 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. 9 and Fig. 10 show feasibility an effectiveness 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, as shown in Fig. 9 and Fig. 10. 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. 9(b) and Fig. 10(b). The results from CPE seem to have better performance in any circumstance. Automatically selected set of features are properly achieved according to the proposed framework.

5 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 and matched pose error in assigning similarity weight. The proposed method probabilistically represents the recognized object's 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 incorporate 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 under 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 robustly recognizes various objects with individual characteristics such as textured and textureless objects in in-door environments.

References

1. M. F. S. Farias and J. M. de Carvalho (1999) Multi-view Technique For 3D Polyhedral Object Rocognition Using Surface Representation. Revista Controle & Automacao., pages 107–117

2. Y. Shirai (1987) Three-Dimensional Computer Vision, Springer, New York
3. J. Ben-Arie, Z. Wang, and R. Rao (1996) Iconic recognition with affine-invariant spectral. In Proc. IAPR-IEEE International Conference on Pattern an Recognition, 1:672–676
4. M. A. Fischler and R. C. Bolles (1981) Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Comm. Assoc. Comp. Mach, 24(6):381–395
5. C.F. Olson (1997) Efficient pose clustering using a randomized algorithm. International Journal of Computer Vision, 23(2):131–147
6. P. David, D. F. DelMenthon, R. Duraiswami, and H. Samet. (2002) Softposit: Simultaneous pose and correspondence determination. 7th European Conference on Computer Vision, 3:698–703, Copenhagen, Denmark
7. A. E. Johnson and M. Hebert (1999) Using spin images for efficient object recognition in cluttered 3D scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(5):433–449
8. A. Frome, D. Huber, R. Kolluri, T. Bulow, and J. Malik. (2004) Recognizing Objects in Range Data Using Regional Point Descriptors. To appear in European Conference on Computer Vision, Prague, Czech Republic
9. Fred Rothganger, Svetlana Lazebnik, Cordelia Schmid, Jean Ponce (2003) 3D Object Modeling and Recognition Using Affine- Invariant Patches and Multi-View Spatial Constraints. IEEE Computer Society Conference on Computer Vision and Pattrn Recognition, 2:272–280
10. D. Lowe. (1999) Object recognition from local scale invariant features. In Proc. 7th International Conf. Computer Vision (ICCV99), pages 1150–1157, Kerkyra, Greece
11. S. Belongie, J. Malik, and J. Puzicha (2002) Shape matching and object recognition using shape contexts. IEEE Trans. On Pattern Analysis and Machine Intelligence, 24(4):509–522
12. Owen Carmichael and Marial Herbert (2004) Shape-Based Recognition of Wiry Object. IEEE Transactions on Pattern Recognition and Machine Intelligence
13. A. Thayananthan, B. Stenger, P.H.S. Torr, and R. Cipolla (2003) Shape context and chamfer matching in cluttered scenes. In Proc. IEEE Conference On Computer Vision and Pattern Recognition
14. Sukhan Lee, Eunyoung Kim, and Yeonchool Park (2006) 3D Object Recognition using Multiple Features for Robotic Manipulation. IEEE International Conference on Robotics and Automation, pages 3768–3774
15. Clark F. Olson (1998) A probabilistic formulation for Hausdorff matching. In IEEE Conference on Computer Vision and Pattern Recognition, pages 150–156
16. Jayashree Subrahmonia, David B. Cooper, and Daniel Keren (1996) 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
17. Y. Boykov, D. P. Huttenlocher (1999) A New Bayesian Framework for Object Recognition. pages517–523
18. S. Z. Li and J. Hornegger (1998) 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, pages 733–747, Heidelberg
19. B. Schiele and J. L. Crowley (1996) Probabilistic object recognition using multidimensional receptive field histograms. ICPR

20. Sukhan Lee, Seongsoo Lee, Jeihun Lee, Dongju Moon, Eunyoung Kim and Jeonghyun Seo (2007) Robust Recognition and Pose Estimation of 3D Objects Based on Evidence Fusion in a Sequence of Images. IEEE International Conference on Robotics and Automation, Pages 3773–3779