Environment Adaptive 3D Object Recognition and Pose Estimation by Cognitive Perception Engine

Hyunjun Kim, Jangwon Lee and Sukhan Lee*

***Abstract*—In this paper, we propose novel evidence selection and collection method based on Bayesian theorem for object recognition and pose estimation in real environment. To recognize and estimate 3D object pose accurately, photometric and geometric evidences such as color blob, SIFT points and lines, can be utilized as single or multiple features in a sequence of images. However, to guarantee dependability in visual perception, the system have to cope with environmental variation that includes change of illumination, amount of texture, and distance to object. So, we made monitoring system to observe the change of environment. The main contribution of this paper is to develop and improve the recognition strategy by proper evidence selection and collection by using Bayesian rule that can be working robustly in various environmental conditions. The experimental results with a single stereo camera show the feasibility and effectiveness of the proposed method in an environment containing both textured and texture-less objects.**

I. INTRODUCTION

The problem of object recognition in complex real environment is dealt long time in the area of computer vision and robotics. Many Researches are in process for long time to develop more robust and faster recognition system because there are many problems - illumination, occlusion, affine, etc. – to object recognition in real environment.

In order to select significant features, Morita et al. [6] have used multi-objective genetic algorithm. They have searched features using two criteria: minimization of a validity index that measures the quality of clusters and minimization of the number of features. Their results seem to be efficient for reducing the number of features and cluster with the recognition rates maintained at the fine level. However, it does not consider automatic selection of valid feature sets. Fan et al. [7] proposed the support vector machines based method for recognition of multi-view faces. This method selects most discriminative features directly without linearly combining the original features. But it contains no mention of the probability of feature integration for recognition. Some research use the approach based on Bayesian theorem to select valid features. Valente and Wellekens [8] proposed the Variational Bayesian learning for model selection criterion. This method has an advantage of fast and robust recognition, but it does not consider mutual information between features.

Several researchers [9][10] use a Bayesian classifier to evaluate the recognition rate with different feature sets. However these approaches have the above-mentioned similar problem. In this paper, we propose novel evidence selection and collection method based on Bayesian theorem for object recognition and pose estimation in real environment. Herein the evidence means features such as SIFT (Scale Invariant Feature Transform) [5], line, color, robot motion and so on. The main advantage of this method is to estimate probability more easily by using Bayesian rule and to select an optimal set of evidences automatically. Therefore this strategy is able to help the robot take advantage of automatic evidence selection in real environment. In addition, we expect that this strategy can be applied to other systems which need robust object recognition.

This paper is organized as follows: Section 2 outlines overall 3D object recognition framework, and section 3 describes proposed evidence selection method based on Bayesian theorem. Evidence collection method is described in section 4 and experimental results for showing feasibility of proposed methodology are presented in section 5. Section 6 concludes by discussing scalability and implementation issues along with directions for future works.

II. Outline of Overall Recognition Framework

We proposed object recognition and pose estimation using image sequence based on probability method in the past paper.[1][2] Overall framework with additional cognitive perception engine is same as Fig. 1.

Stereo vision Camera is used as the input sensor to get 3D information with 2D image. In In-Situ monitoring, the detection of environmental changes based on input information of 2D & 3D, ROI generation for focus of attention and information of environment in ROI is transferred to the evidence selection. The evidence selection selects the best feature to environment based on environmental information using Baysian theorem. Features are extracted and matched using selected feature in evidence selection, and interpretations are generated. Interpretation is the candidate of the object having probability for probabilistic calculation. Each interpretations generated by

This research is performed for the 21st Century Intelligent Robot National Frontier Project sponsored by the Korea Ministry of Knowledge Economy (MKE). This research is also supported in part by the Korea Science and Engineering Foundation (KOSEF) grant sponsored by the Ministry of Education, Science and Technology (MEST), No. R01-2006-000-11297-0, and in part by the Science and Technology Program of Gyeonggi Province,. and also supported by the MKE(Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Advancement) (IITA-2009-(C1090-0902-0046)).

Hyunjun Kim, Jangwon Lee and Sukhan Lee*, the Corresponding Author, are with the Intelligent Systems Research Center, Sungkyunkwan University, Suwon, South Korea (e-mail: kimhj8574@gmail.co.kr, jwlee31@ece.skku.ac.kr, lsh@ece.skku.ac.kr)

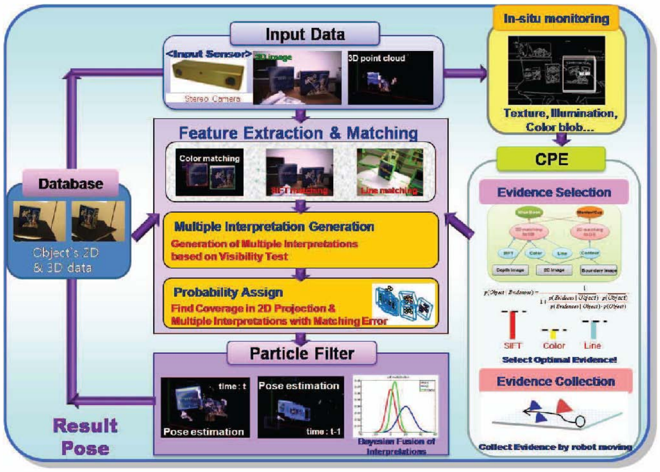
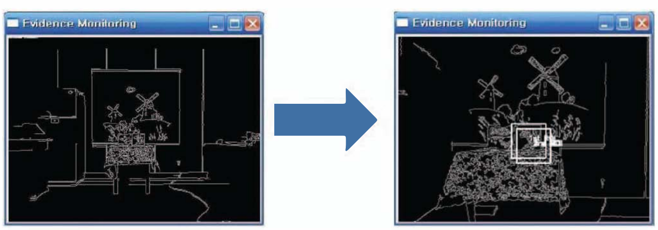


Fig. 1. The Overall 3D Recognition Framework

feature matching are fused in particle filter. Robot behavior is determined for the uncertainty minimization of next cycle object recognition in the evidence collection process using the final results fused by the particle filter.

A. In-Situ Monitoring

In-Situ monitoring generates ROI, measures environment change in ROI, and transfer this data to the evidence selection part of CPE. ROI is generated from projecting the interpretation of high probability to 2D image from previous recognition results. The purpose is to decrease calculation time by extracting only environment factors of the existing object, corresponding to the concept of focus of attention.



(a) (b)

Fig. 2. In-Situ Monitoring System Window. (a) Observe all possible area and (b) Observe region of interest

The detail algorithm of this monitoring system as follows: If there is no previous recognition result, the system observes whole area and calculates average texture, illumination and distance on the region. At this moment if there are candidate objects, probability distributions of object pose are obtained. In our framework, this probability distributions are described particles.

After then this monitoring system makes ROI (Region of Interest) for calculating average texture, illumination and distance to deliver CPE. The ROI is made using projection of vertexes of particles. Amount of texture in ROI is counted pixel which is processed by Canny edge image of current scene. And illumination is calculated using intensity information based on HSI color space in current image. Actually these two values are not the same but relative values about changes of environment. Finally, distance in ROI is calculated using processed image pixel with valid 3D point cloud and average those values. Input, output and summary of this monitoring system algorithm are shown in Algorithm 1.

Algorithm 1 In-Situ Monitoring Algorithm

Input:

2D images, 3D point clouds, previous recognition results

1: Make a canny edge image based on 2D images and get intensity value using HSI color space.
2: Observe whole area and check changes of environment.
3: If previous recognition results are exist, make ROI using the previous particles
Else return to state 2.
4: Calculate average texture, average illumination and average distance.
5: Until recognition mission end, repeat state 1-4.

Output:

average texture, average distance, average illumination

B. Feature Evidence

We use 3 features for the object recognition and pose estimation. Photometric feature (SIFT, Color) and geometric feature (Line) is used for the object recognition robust to the various environment changes. For probabilistic calculation, each feature generate multiple interpretation and assign probability when perform matching.

One of photometric feature is SIFT(Scale Invariant Feature Transform) feature. First, generate matching points between object 2D images previously stored in database and input 2D image using SIFT (Scale Invariant Feature Transform) feature. 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. Details are expressed in [3].

Due to poor accuracy of stereo camera's 3D data, all lines are firstly extracted from 2D image and these 2D lines can be converted to 3D lines through mapping 3D points corresponded to 2D lines. 3D line is generated using 3D point cloud after the extraction of 2D Line and object of box shape can be recognized by the parallel matching of these. Details are expressed in [4].

For color feature, segmentation is performed using representative hue value of color from target object. Color Blob is made by clustering the Segmentation results and color blob's real size in 3D is calculated using 3D data value of each color blob. From these, blob with similar size of target object is selected as the candidate of the object. This method shows low reliability for single use because characteristics of the simple algorithm make many false positive candidates generated by circumstantial environment. Color feature has the advantage of fast calculation, is used for the conjecture because evidence can be extracted in more distant distance than other features (SIFT, Line), and is applied with the fusion of other evidences using particle filter.

C. Cognitive Perception Engine(CPE)

We assume that the valid features for recognizing each object in a current scene are already defined to the CPE. The evidence selection and collection has been accomplished in the CPE part. Among N possible features eligible as evidences for object recognition, where N is possibly too large for all the features to be applied, select M (M < N) features optimal in term of efficacy and efficiency for object recognition. Initially when there is no probability distribution on object pose in space, the scene captured by robot camera is used as a whole to determine a set of optimal features based on its texture content, illumination, and distance. After probability distribution of object pose is obtained, robot is supposed to move to the direction where the probability is high, then the decision on the optimal set of evidences is based on the texture, illumination, and distance information of the scene that is weighed by object probability distribution. Detailed strategy for proper evidence selection by using Bayesian rule is described in next section and evidence collection method is written in section 4.

D. Particle Filter for Evidence Fusion & Estimation

Particle filtering procedure is presented in previous papers [1][2]. 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 [1][2] which is most important parts in proposed particle filter based framework. Probability distribution of the object pose is predicted by using probabilistic motion model and previous particles of the object pose. Multiple poses of the object generated from features at current scene without prior particles and similarity weight of each pose are used for making observation model. Then, the observation likelihood can be calculated by using predicted and measured particles considering similarity weight and Mahalanobis distance metric. According to resulting particle's weights, particles are re-sampled to represent probability distribution of the object pose at current scene. These procedures are repeated until the particles are converged to a single pose.

III. Bayesian Theorem based Evidence Selection

We consider in this section the methodology of evidence selection that is a part of proposed framework for 3D object recognition and its pose estimation by using probabilistic method based on Bayesian theorem [1][2]. But our previous works of proposed method do not conducted the evidence selection for 3D object recognition automatically. It means that we already defined the proper evidence, feature, or feature set, for recognizing target object according to changes of environment. So, in order to select the suitable evidence that is not predefined but adapted to environmental changes, we make use of Bayesian theorem to calculate these probabilities of evidence.

A. Bayesian Formulate for Evidence Selection

To rank feature's suitability, we calculate each feature's $P(O|E)$. This $P(O|E)$ means evidence's confidence. O is the target object and E means evidence or evidence set for recognizing an object. In general, $P(O|E)$ is not easy to get directly. So, in this paper, we try to calculate using the Bayesian theorem. To assign probability, we consider how much correspondence between the recognized object and its available evidence for recognizing. In probabilistic terms the goal of proposed method is to evaluate evidence which yield the best interpretation of evidence generated by proposed hypothesis in Bayesian sense. To calculate each probability that is used for making candidate of recognition result is accordance basic Bayesian theorem as follow:

$$p(O \mid E) = \frac{1}{1 + \dfrac{p(E \mid \bar{O}) \cdot p(\bar{O})}{p(E \mid O) \cdot p(O)}} \tag{1}$$

$P(E|O)$means Positive information, and $P(E|\bar{O})$ means negative information. P(O) and P(Ō) each means target object is exist or not exist in current scene. First of all, both P(O) and P(Ō) are assigned 0.5 because under no information about environment space, we can't predict target object is exist or not. Positive information means that the probability of each evidence when there exist target object in current scene. Positive information is assigned by before experiment result. For more detail explanation of negative information is described in next section. Negative information means that the probability of each evidence when there is no target object in current scene. Initially, this negative information is obtained by each evidence's heuristic value. experimental data of each evidence when there is no target object. And it can be updated by the previous recognition result. For more detail explanation of negative information is described in section 3.D. In addition, we define evidence, E, as information that consists of three kinds of factors: amount of texture, the level of illuminations, and variation of distance to object.

B. Probability Space Models for Each Evidence

In our framework, we have three evidences - SIFT, Line and Color - that are used as features for recognizing 3D object and estimating its pose. These evidences have different characters changes of environment. So we made different probability space models for each evidence.

1) SIFT Model: SIFT feature has such kinds of characters: 1) SIFT is sensitive changes of distance. Actually, SIFT matching is 2D matching and this matching algorithm comparatively robust about changes of distance. But in order to get accurate object pose we use 3D point cloud information and trying 3D matching. It depends on the camera lens size, but 3D Point cloud is incorrect relative to the long distance between robot and the target object. Therefore SIFT feature has weak point in distance changes.

2) The more target object has texture, the easier the object can be matched with model. 3) SIFT points extraction works best in around 330 lux illumination condition and the points has decreased from summit. See these kinds of characters in Fig. 3.

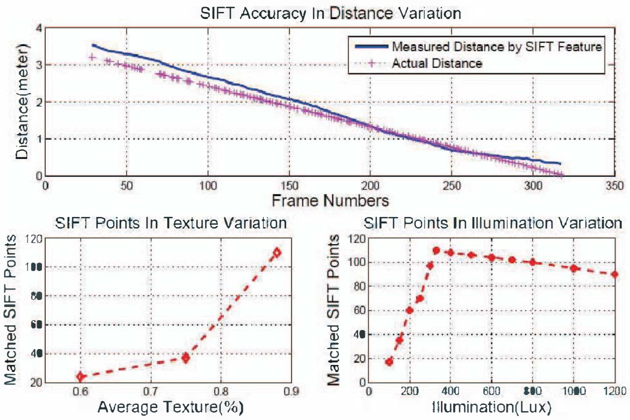


Fig. 3. Characteristic of SIFT feature based on experiment results

Based on these kinds of characters, we made SIFT probability model using two kinds of functions, Gaussian and linear. In case of the probability space in distance, the function is a Gaussian with 0.6 meter mean value, because we usually make model database for SIFT matching with that distance. In addition, the probability space in texture variation is modeled linearly. This is straightforward because the more texture there are, the better SIFT features are extracted. Illumination case is very similar to the probability space in distance and its mean is 330 lux that is the best condition to extract SIFT feature based on experiments. To obtain entire probability of SIFT, each sub probability is multiplied consecutively because each are independent:

$$P(E \mid O) = P(E_{distance} \mid O) \cdot P(E_{texture} \mid O) \cdot P(E_{illumination} \mid O) \qquad (2)$$

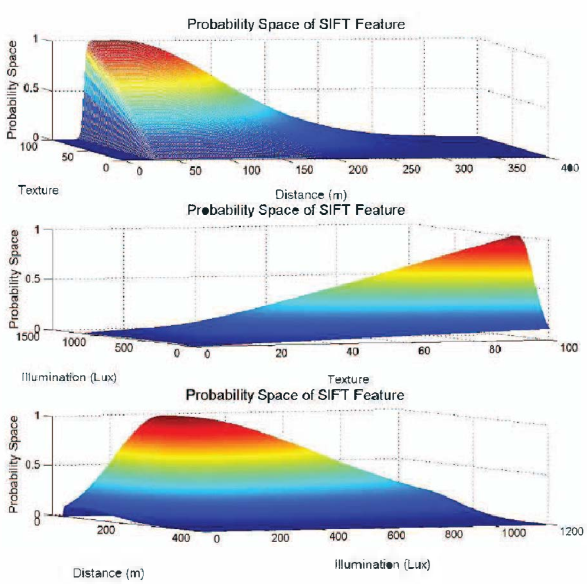


Fig. 4. The Probability Space of SIFT feature

2) Line Model: Although object identification by using line feature is not so good and may lead to mismatch, it can be widely applied to recognize object because of abundance of line feature. Line feature is affected by three environmental factors: 1) If the distance between robot and the target object is so far or so close, line feature is inaccuracy. Valid distance of line feature is around 1.0 meter. 2) Line feature is not sensitive when amount of texture is more than a certain level. 3) Line feature is more sensitive in dark illumination condition than bright illumination condition.

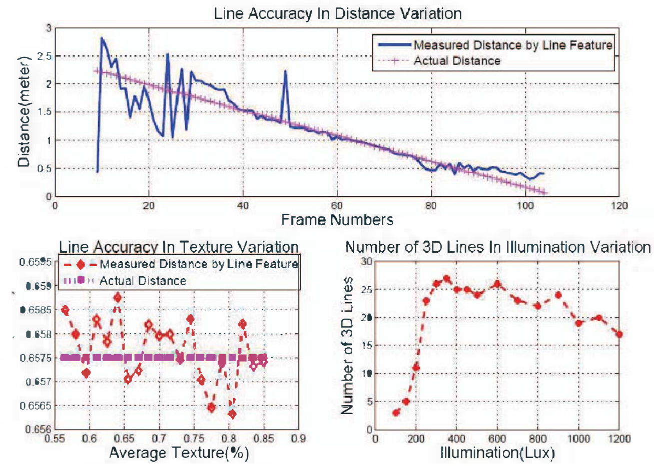


Fig. 5. Characteristic of Line feature based on experiment results

The probability of line feature can be modeled as Gaussian function. Especially, if amount of texture in space is more than a certain level, line feature has nearly little relationship with texture, so its variance should be large enough. The probability space of line feature is depicted in Fig. 6.

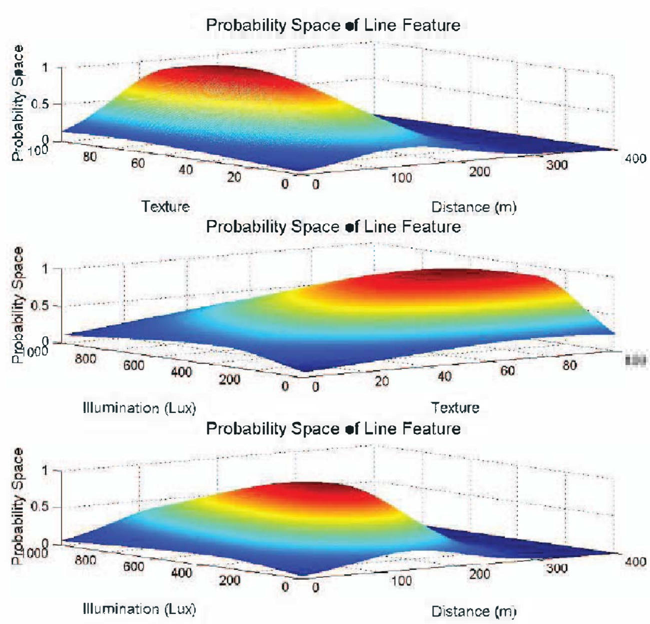


Fig. 6. The Probability Space of Line feature

3) Color Model: Color feature is influenced just by changes of distance and illumination, not by texture. If there are no similar color as the target object, color feature is robust about changes of distance. And this feature accuracy shows Gaussian distribution in illumination changes. See these kinds of characters in Fig. 7.

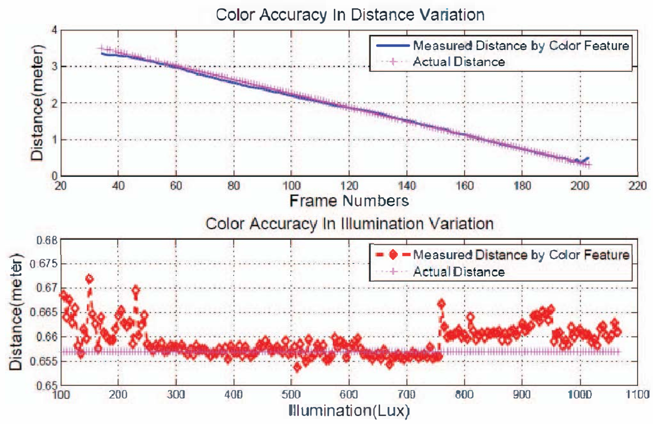


Fig. 7. Characteristic of Color feature based on experiment results

So we made the probability space of color feature using Gaussian function. The probability space in distance variation is made Gaussian with 1.0 meter mean value. If distance is longer than 1.0 meter, we made the probability values are decreased faster than near case. Because if there are similar color object as the target one, mismatches are often caused by long distance. The probability space in illumination is also modeled as a Gaussian function with large variance since illumination is not a dominant factor. In this case, mean value is 330 lux. Fig. 8 shows the calculated result of probability space in Color feature.

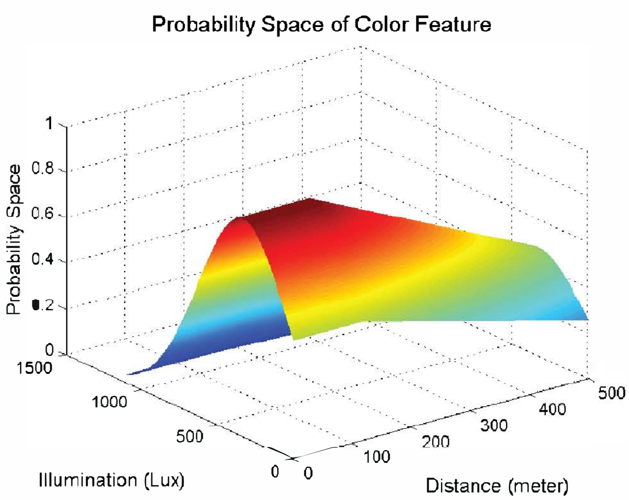


Fig. 8. The Probability Space of Color feature

C. Negative Information in Bayesian Term

As stated above negative information means the probability of each evidence when there is no object. It is represented as $P(E|\bar{O})$. In computing what kind of evidence is best for object recognition ($P(O|E)$), the negative information is a sensitive term. But it is difficult to find the probability of evidence when object do not exist. Thus we use previous recognition result of t-1 in order to get the probability. Primarily, the negative information of each feature is obtained using experimental results. We had experiments by changing environment factors such as distance and illumination, and observe probability of $P(E|O)$ in the situation. And we arranged the initial value of $P(E|\bar{O})$ as lower bound of $P(E|O)$. The reason why we defined the initial value of it that manner is that we think this probability is some kind of error in each evidence. Because we made probability of $P(E|O)$ using environment factors, so the probability of $P(E|O)$ can exist when there is no object. See Fig. 9. The lower bound of each graph is initial value of negative term in our Bayesian equation (3). Then the negative term is updated by previous recognition result which is the probability at t-1. The update algorithm is that when the best evidence or evidence set is selected, but there is lack of the selected evidence in current scene, the negative term of selected evidence is increasing. For example, when the SIFT evidence is selected as the best evidence in the evidence selection engine, however matched SIFT points are short for recognition, negative term of the SIFT evidence is increasing. However, if the recognition success using the selected evidence the negative term is decreasing until the initial value. Using this feedback system, we can make our selection engine more robustly.

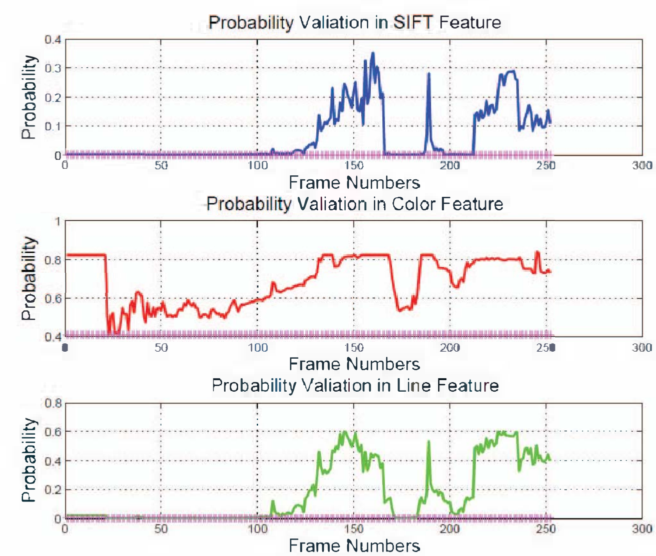


Fig. 9. The Negative Information of Each Feature

IV. EVIDENCE COLLECTION PROCESS

The recognition result only with the analysis of image with static location have not enough reliability because complexity of real environment and imperfection of recognition algorithm. Sometimes there is no recognition result because of multiple environmental limitation – distance, occlusion, object's pose, etc. To solve this, evidence collection process is performed that observes in the different direction from the different location through searching in the case of no recognition result and decrease uncertainty of interpretations by active robot movement to increase reliability of recognition results in the case of recognition results exist.

A. Search Process

Useful evidence does not come out with any feature in the case that the object is too far away or occluded from the robot. Movement to the location with better view is required by changing robot location or view direction in this case. In this research, we simplified problem which possible location of object is robot's forward area. The shape of map is a sector form to the front because the objective is to search the object located somewhere front of the robot by scenario. Only odometry is used because the robot localization does not require much accuracy. Independent platform configuration for the general search and accuracy localization module are planned. Each cell of map represents the robot location and 6 view directions exist as 3 pan values, 2 tilt values for each cell. The view direction of each cell in Fig. 10 has determined the initial motion value and the movement is made searching view direction of low motion value cell, increasing the motion value of view direction of the cell in case of no recognition result. Perform the evidence collection in case that recognition result exists. Algorithm of Search and Evidence Collection process are shown in Fig. 11.

B. Evidence Collection Process

In case that evidence is detected, there is a need to increase the reliability of evidence by minimizing the each uncertainty of the evidence. Evidence collection process minimizes the uncertainty of evidence to determine the robot movement decided by the object. The assumptions to decide movement in evidence collection are followings.

- High probability interpretation has high chance to be an object.

- High uncertainty makes it hard to decide as an object even if probability of the interpretation is high.

- There is a possibility as an object if the uncertainty is high even with low probability.

- It can be decided not to be an object if the probability is low even with minimized uncertainty.

Select some of high probability ones from interpretations result from particle filter, project them into 2-dimensional plane, and decide robot action using each probability of interpretations and uncertainty of distance from robot. Probability of the interpretation is the estimated value by particle filter based on the likelihood by feature matching, is proportional to the probability that the interpretation is real object. Uncertainty of each interpretation is the reversed shape of P(E|O), the probability that feature be extracted when there exist an object to the distance of each feature. Fig. 12 shows this. Use the gradient of uncertainty graph of the current distance as weight because the objective is to decrease the uncertainty of each interpretation. This means

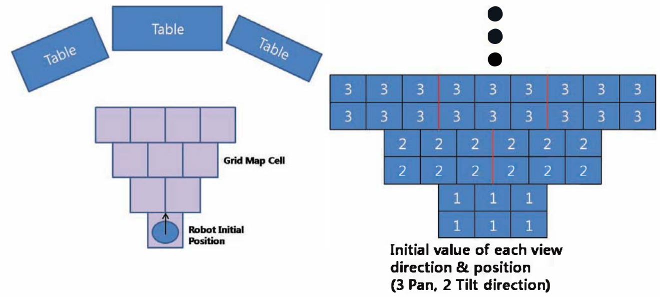


Fig. 10. Grid Map for Search Process

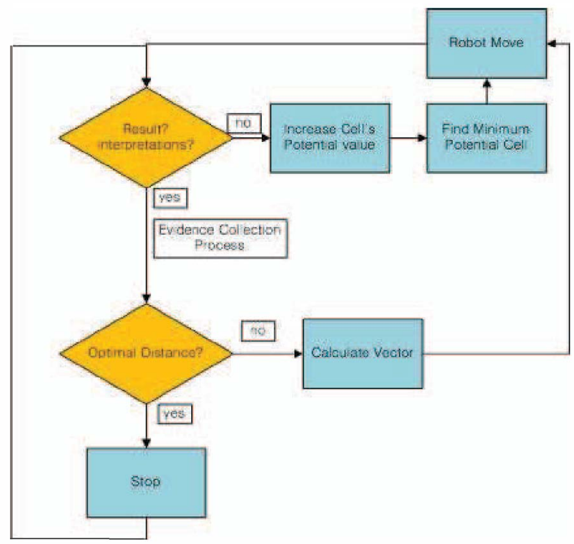


Fig. 11. Flow chart of Evidence Collection

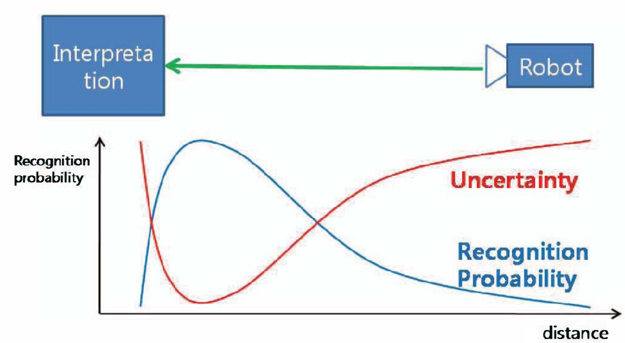


Fig. 12. Interpretation`s Uncertainty about distance

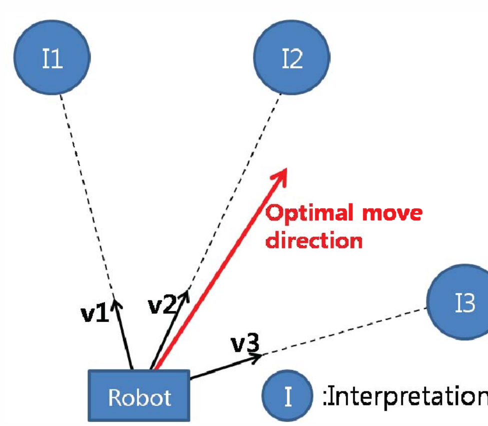


Fig. 13. Optimal Movement Decision

weight is given by how much uncertainty decreases when the movement is made for a unit distance on the distance from current robot to the interpretation. Even if probability of interpretation is relatively low, move to this and check if uncertainty is high and can be easily reduced.

Generate unit vector from robot to each interpretation and multiply each probability and uncertainty weight. Equation (3) decide direction of the sum of these vectors as movement direction of robot. Fig. 13 are shown this process. The process is completed by the decision as 'an object' if probability of interpretation is high enough in situation that uncertainty decreases under the specific value by robot behavior, goes back to the search process by decision of 'not an object' if probability is still high.

$$v_{optimal} = w_{p1}w_{u1}v_1 + w_{p2}w_{u2}v_2 + w_{p3}w_{u3}v_3 \tag{3}$$

w_{pn} : weight of Probability

w_{un} : weight of Uncertainty

v_n : vector

V. Experiment Result

Hardware environment is as following

-Core2Duo Penryn laptop Computer

-Bumblebee2 Stereo Camera (6mm focal length)

-ER Scorpion Mobile Robot

-PanTilt module (Robotiz Dynamixel * 2)

Target object used for recognition and position estimation is selected by one that is recognizable with 3 features (SIFT, Line, Color) that we use. Target object and experiment robot are shown in Fig. 14. Experimental test was made in the environment that there are multiple objects in disorder at the table in front of the initial location of robot. Experiment Environment are shown in Fig. 15. (a) is not occluded case, and (b) is occluded case.

Experiment is performed about 3 condition, not occluded and forward object, occluded and forward object, not occluded diagonal object. Experiment results are shown in Fig. 16, Fig. 17, Fig. 18. Gray squares are grid map, red cross is each sequence's maximum probability interpretation and green line is robot movement when search process and yellow line is robot movement when evidence collection movement. Each case optimal evidence is selected by evidence selection and search and move to interpretations by evidence collection.

VI. Conclusions

The main contribution of this paper is to develop and improve the recognition strategy by proper evidence selection using Bayesian rule can be working robustly in various environmental conditions. Experimental results demonstrated proper evidence selection in real dynamic environment including distance variations and occlusion. And proposed approach can select valid evidence considering processing time of each visual processing. Therefore, it can be one of the feasible solutions to deal with trade-off between performance and time consumption in visual processing. Evidence collection is also improve robustness of visual perception by generation of active behavior. But this evidence collection's search is simplified and specialized. So, our future work is add localization and navigation for general search. So, active evidence selection and collection considering robot motion and real-time feedback from environment will be explored as a future work to complete dependable "behavioral perception" paradigm.

Fig. 14. Experiment Robot and Target Object

Fig. 15. Experiment Environment

References

[1] 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," *IEEE International Conference on Robotics and Automation*, Rome, Italy, 2007

[2] Jeihun Lee, Seung-Min Baek, Changhyun Choi, and Sukhan Lee, "Particle Filter Based Robust Recognition and Pose Estimation of 3D Objects in a Sequence of Images," *Recent Progress in Robotics, Lecture Notes in Control and Information Sciences(LNCIS)*, SpringerVerlag, 2008, pp. 241-253

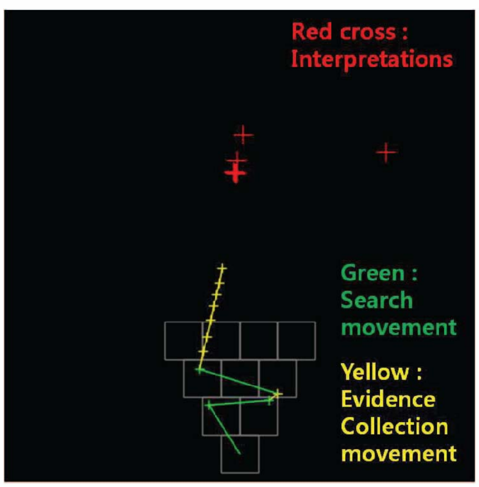
[3] Sukhan Lee, Eunyoung Kim, and Yeonchool Park, "3D Object Recognition using Multiple Features for Robotic Manipulation," *IEEE International Conference on Robotics and Automation*, 2006, pages 3768-3744

[4] Samuel H. Chang, Sukhan Lee, Dongju Moon, WoongMyung Kim, YeungHak Lee, "MODEL BASED 3D OBJECT RECOGNITION USING LINE FEATURES," *The 13th International Conference on Advanced Robotics(ICAR2007)*, Jeju, Korea, 2007

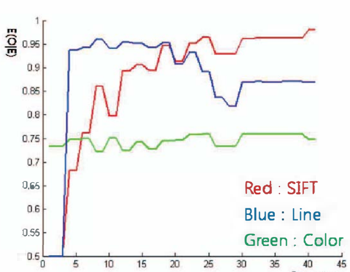
[5] D. Lowe, "Object recognition from local scale invariant features," *In Proc. 7th International Conf. Computer Vision (ICCV'99)*, pp.1150-1157, Kerkyra, Greece, September 1999.

[6] M. Morita, R. Sabourin, F. Bortolozzi, and C. Y. Suen, "Unsupervised feature selection using multi-objective genetic

algorithms for handwritten word recognition," *in 7th IEEE International Conference on Document Analysis and Recognition (ICDAR)*,pp. 666-670, Edinburgh, Scotland, 2003.
[7] Z.-G. Fan and B.-L. Lu, "Fast recognition of multi-view faces with feature selection," *in 10th IEEE International Conference on Computer Vision (ICCV)*, Vol. 1, pp. 76-81, Beijing, China, 2005.
[8] F. Valente and C. Wellekens, "Variational Bayesian feature selection for Gaussian mixture models," *in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.1, pp. I-513-16, Montreal, Quebec, Canada, 2004.
[9] P. C. Ribeiro, J. Santos-victor, "Human activity recognition from video: modeling, feature selection and classification architecture," *in International Workshop on Human Activity Recognition and Modeling (HAREM)*, Oxford, UK, 2005.
[10] B. Krishnapuram, A. J. Hartemink, L. Carin, and M. A. T. Figueiredo, "A Bayesian approach to joint feature selection and classifier design," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26, pp. 1105-1111, September 2004.
[11] S. Russell and P. Norvig,"Artificial Intelligence-A Modern Approach," *Prentice Hall International*, Englewood Cliffs, NJ,USA,1995.

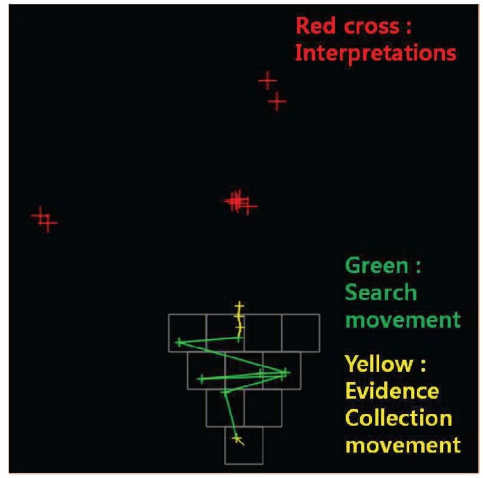


(a) Robot movement & Interpretations position

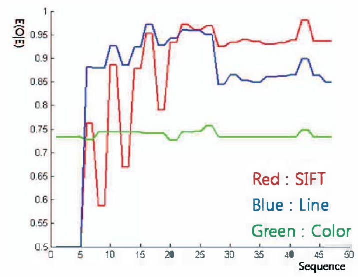


(b)Probability of each feature

Fig. 16. No Occlusion Experiment Result

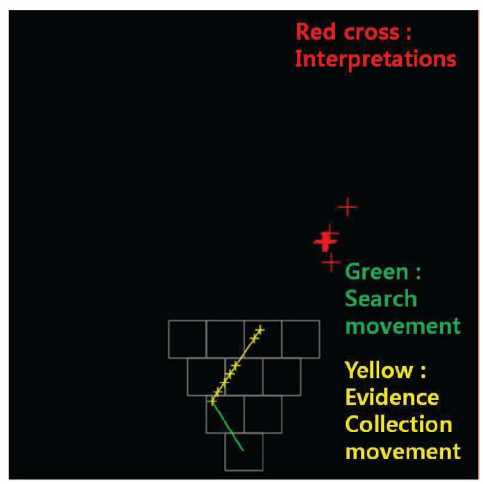


(a) Robot movement & Interpretations position

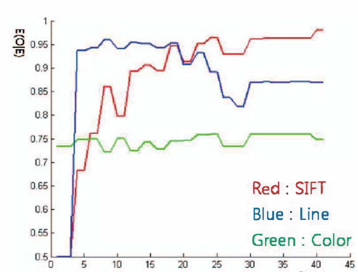


(b)Probability of each feature

Fig. 17. Occlusion Experiment Result



(a) Robot movement & Interpretations position



(b)Probability of each feature

Fig. 18. Diagonal position Experiment Result