Context-Based Scene Recognition Using Bayesian Networks with Scale-Invariant Feature Transform

Seung-Bin Im and Sung-Bae Cho

Dept. of Computer Science, Yonsei University
134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Korea
envymask@sclab.yonsei.ac.kr,
sbcho@cs.yonsei.ac.kr

Abstract. Scene understanding is an important problem in intelligent robotics. Since visual information is uncertain due to several reasons, we need a novel method that has robustness to the uncertainty. Bayesian probabilistic approach is robust to manage the uncertainty, and powerful to model high-level contexts like the relationship between places and objects. In this paper, we propose a context-based Bayesian method with SIFT for scene understanding. At first, image pre-processing extracts features from vision information and objects-existence information is extracted by SIFT that is rotation and scale invariant. This information is provided to Bayesian networks for robust inference in scene understanding. Experiments in complex real environments show that the proposed method is useful.

1 Introduction

Scene understanding is the highest-level operation in computer vision, and it is a very difficult and largely unsolved problem. For robust understanding, we must extract and infer meaningful information from image. Since a scene consists in several visual contexts, we have to recognize these contextual cues and understand their relationships. Therefore, it might be a good approach to start with extracting basic contexts like “where I am” or “what objects exist” in the scene for robust understanding. If we successfully extract these meaningful cues, we can provide them to higher level context understanding.

High-level context, like the correlations between places and objects or between activities and objects, is a key element to solve image understanding problem. For example, a beam-projector usually exists in a seminar room and a washing stand exists in a toilet. This contextual information helps to disambiguate the identity of the object and place despite the lack of sufficient information. Contextual scene recognition is based on common knowledge such as how scenes and objects are organized.

Visual information is powerful and crucial, whereas it is uncertain due to motion blur, irregular camera angle, bad lighting condition, etc. To overcome it, we need a sophisticated method that is robust to uncertainty. Bayesian network (BN) might be suitable for modeling in the domain of image understanding, since probabilistic

J. Blanc-Talon et al. (Eds.): ACIVS 2006, LNCS 4179, pp. 1080 – 1087, 2006.

approach has the characteristic that is robust to inference in various directions and operable to uncertain data [1].

Probabilistic approach has attracted significant attention in the area of vision-based scene understanding. Torralba *et al.* proposed a method to recognize the place using hidden Markov model with global vectors collected from images and use them as context information to decide the detection priorities [2]. This approach is useful to make detection more efficient but the errors are inherited from the place recognition systems. Marengoni *et al.* tried to add the reasoning system to Ascender I which is the system to analyze aerial images for detecting buildings. They use hierarchical Bayesian networks and utility theory to select proper visual operator in the given context, and they could reduce computational complexity [3]. J. Luo, *et al.* proposed that Bayesian framework for image understanding [4]. In this approach, they used low-level features and high-level symbolic information for analyzing photo images.

In the meantime, there are many studies for solving object recognition problem. T. M. Strat and M. A. Fischler assumed that objects were defined by small number of shape models and local features [5]. D. G. Lowe proposed Scale-Invariant Feature Transform (SIFT) that extracts local feature vectors that are robust to image rotation and variation of scale [6]. SIFT shows good performance in extracting objects-existence but performance deteriorates if object has scanty texture element. Because performance of the object recognition algorithms is subject to low-level feature extraction results, we need a method that not only adopts low-level features but also uses high-level contexts.

In this paper, we propose a context based image understanding methodology based on Bayesian belief networks. The experiments in real university environment showed that our Bayesian approach using visual context based low level feature and high level object context which extracted by SIFT is effective.

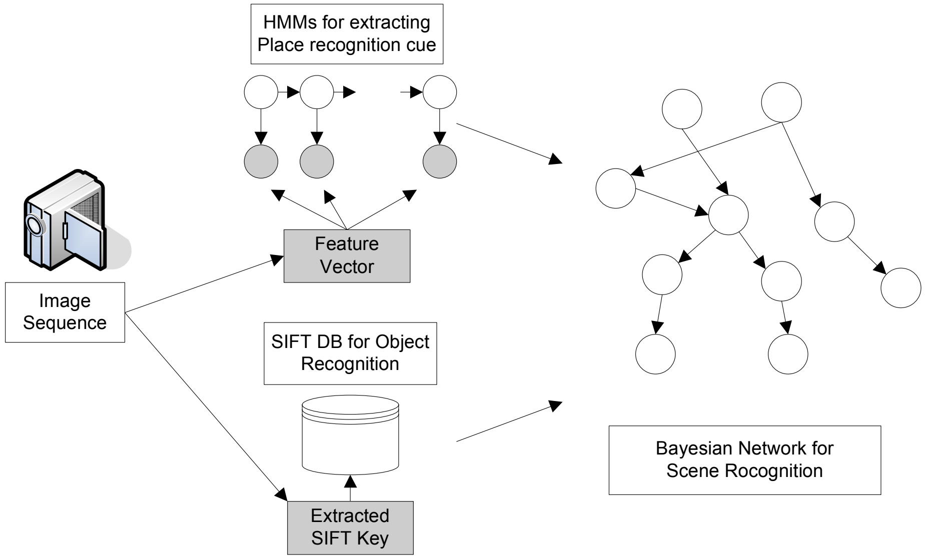

Fig. 1. An overview of Bayesian scene recognition

2 Context-Based Scene Recognition

In this section we describe the recognition of places and objects based on context. At first, we explain global feature extraction and HMMs learning, and describe object recognition with SIFT. Finally, context-based Bayesian network inference will be illustrated. The overview of the proposed method is shown in Fig 1.

2.1 Visual Context-Based Low-Level Feature Extraction

It would be better to use features that are related to functional constraints, which suggests to examine the textural properties of the image and their spatial layout [2]. To compute texture feature, a steerable pyramid is used with 6 orientations and 4 scales applied to the gray-scale image. The local representation of an image at time t is as follows:

$$v_t^L(x) = \{v_t, k(x)\}_{k=1,N} \text{ , } where\ N = 24 \tag{1}$$

It is desirable to capture global image properties, while keeping some spatial information. Therefore, we take the mean value of the magnitude of the local features averaged over large spatial regions:

$$m_t(x) = \sum_{x'} | v_t^L(x') | w(x'-x) \text{ , } where\ \ w(x)\ is\ the\ averaging\ window \tag{2}$$

The resulting representation is down-sampled to have a spatial resolution of 4x4 pixels, leading to the size of m_t as 384(4 x 4 x 24), whose dimension is reduced by PCA (80 PCs).

Then, we have to compute the most likely location of the visual features acquired at time t . Let the place be denoted as $Q_t \in \{1,...,N_p\}$ where $N_p = 5$. Hidden Markov model (HMM) is used to get place probability as follows:

$$\begin{aligned} P(Q_t = q \mid v_{1:t}^G) &\propto p(v_{1:t}^G \mid Q_t = q) P(Q_t = q \mid v_{1:t-1}^G) \\ &= p(v_{1:t}^G \mid Q_t = q) \sum_{q'} A(q', q) P(Q_{t-1} = q' \mid v_{1:t-1}^G) \text{ ,} \end{aligned} \tag{3}$$

where $A(q', q)$ is the topological transition matrix. The transition matrix is simply learned from labeled sequence data by counting the number of transitions from location i to location j .

We use a simple layered approach with HMM and Bayesian networks. This presents several advantages that are relevant to modeling high dimensional visual information: learning each level independently with less computation, and although environment changes, only first layer requires new learning with the remaining unchanged [7]. The HMM is for extracting place recognition and BNs are for high-level inference.

2.2 High-Level Context Extraction with SIFT

Scale-Invariant Feature Transform (SIFT) is used to compute high-level object existence information. Since visual information is uncertain, we need a method that has robustness to scale or camera angle change. It was shown that under a variety of reasonable assumptions the only possible scale-space kernel was the Gaussian function [6]. Therefore, the scale space of an image is defined as a function, $L(x, y, \sigma)$ that is produced by the convolution of a variable-scale Gaussian, $G(x, y, \sigma)$, with an input image, $I(x, y)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y), \tag{4}$$

where * is the convolution operation in x and y, and

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \tag{5}$$

To efficiently detect stable key-point locations in scale space, scale-space extrema in the difference-of-Gaussian function are convolved with the image, $D(x, y, \sigma)$, which can be computed from the difference of two nearby scales separated by a constant multiplicative factor k:

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned} \tag{6}$$

Extracted key-points are examined in each scene image, and the algorithm decides that the object exists if match score is larger than a threshold.

In this paper, SIFT features of each object are extracted from a set of reference images and stored in an XML database. Each reference image is manually extracted from the training sequence set.

2.3 Context-Based Bayesian Network Inference

A Bayesian network is a graphical structure that allows us to represent and reason in an uncertain domain. The nodes in a Bayesian network represent a set of random variables from the domain. A set of directed arcs connect pairs of nodes, representing the direct dependencies between variables. Assuming discrete variables, the strength of the relationship between variables is quantified by conditional probability distributions associated with each node [8].

Consider a BN containing n nodes, Y_1 to Y_n, taken in that order. The joint probability for any desired assignment of values $< y_1, \dots, y_n >$ to the tuple of network variables $< Y_1, \dots, Y_n >$ can be computed by the following equation:

$$p(y_1, y_2, \dots, y_n) = \prod_i P(y_i \mid Parents(Y_i)) \tag{7}$$

where $Parents(Y_i)$ denotes the set of immediate predecessors of Y_i in the network.

BN used in this paper consists of 4 types of nodes: (1) ‘PCA Node’ for inserting global feature information of current place, (2) ‘Object Node’ representing object existence and correlation between object and place, and (3) ‘Current Place Node’ representing the probability of each place.

Let the place be denoted $Q_t \in \{1,...,N_p\}$ where $N_p = 5$, and object existence is denoted by $O_{t,i} \in \{1,...,N_{object}\}$ where $N_{object} = 14$. Place recognition can be computed by the following equation:

$$Current\ Place = \arg\max P(Q_t = q \mid v^G_{1:t}, O_{t,i},...,O_{t,N_{object}}) \tag{8}$$

The BNs are manually constructed by expert, and nodes that have low dependency are not connected to reduce computational complexity. Fig. 2 shows a BN that is actually used in experiments.

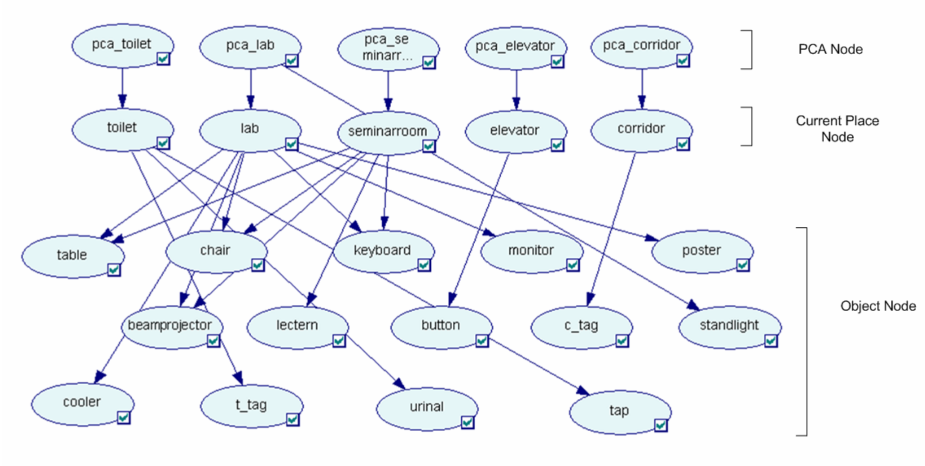

Fig. 2. A BN manually constructed for place and object recognition

3 Experimental Results

To collect input images, a USB mobile camera with notebook PC was used in the experiments. The camera was set to capture 4 images per second at a resolution of 320x240 pixels. The camera was set on a cap at the height of human sight, and the images were captured during user visits 5 different locations. The locations were visited in a fairly random order. We gathered 5 sequence data sets (one for training, others for testing) by the camera in the campus indoor environments. The sequences gathered contain many low quality images, due to motion blur, low-contrast and non-informative views, etc, but experimental results show that the proposed method overcomes these uncertainties.

Fig. 3 shows an experimental result that is the one of sequences that were used in our movements. The x-axis shows the flow of time and a solid line is the true places. Dots represent the probability of each inference result. The proposed method successfully recognized the entire image sequences in general. However, during t = 0 to 100, in '*Elevator*', the proposed method made several false recognitions, because of low-contrast and strong day light that passed through the nearby window. Due to scattered reflection, toilet and corridor also caused several false recognitions (t = 320 to 500).

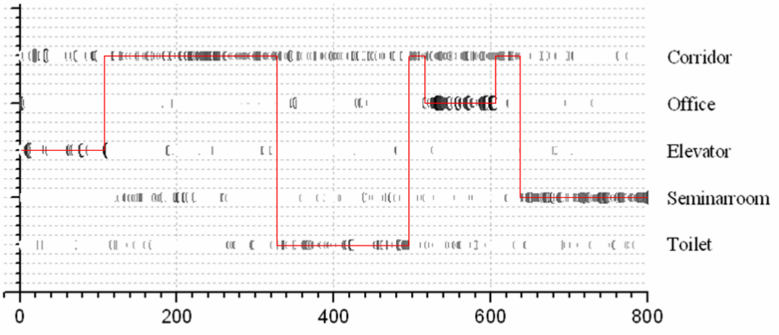

Fig. 3. One of the testing sequence result

Fig. 4 shows overall place recognition performance of the proposed method. The square dots show the place recognition results that used extracted low-level features only and diamond dots show the results of the method that used the BN with SIFT. It can be easily confirmed that the proposed method produces better performance. The hit rate of the proposed method increased 7.11% compared to the method that did not use BN. *Laboratory* shows highly increased recognition result since objects recognition performance by SIFT is good. On the other hand, *elevator* shows bad performance and smaller increase than other locations, because there is no particular object in elevator except elevator buttons, and bad light condition causes worse performance. In *toilet*, lack of the object existence information caused by diffused reflection made low recognition rate.

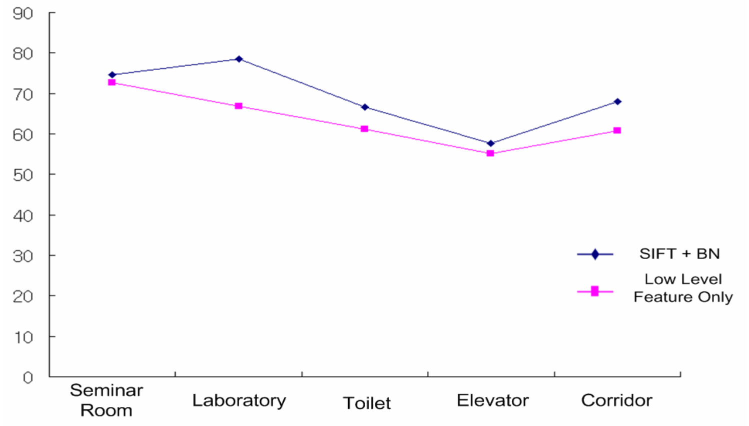

Fig. 4. Overall performance of each place recognition results

Fig. 5 shows the results of the SIFT object recognition. Objects with low texture features caused bad recognition results in the cases of *tap* and *urinal*. It can be easily confirmed that sufficient textual information makes good recognition result for the instances of the *keyboard* and *poster*. Fig. 6 shows the object recognition results of the proposed method. If the inferred objects-existence probability is larger than 75% or SIFT detects the object, the proposed method decides that object exists. Overall recognition score shows better results and recognition performance of objects that were not recognized by SIFT is increased especially (*monitor*, *urinal*). In addition, occluded objects were detected by Bayesian inference. However, it is a defect that false detection rate is increased in some objects.

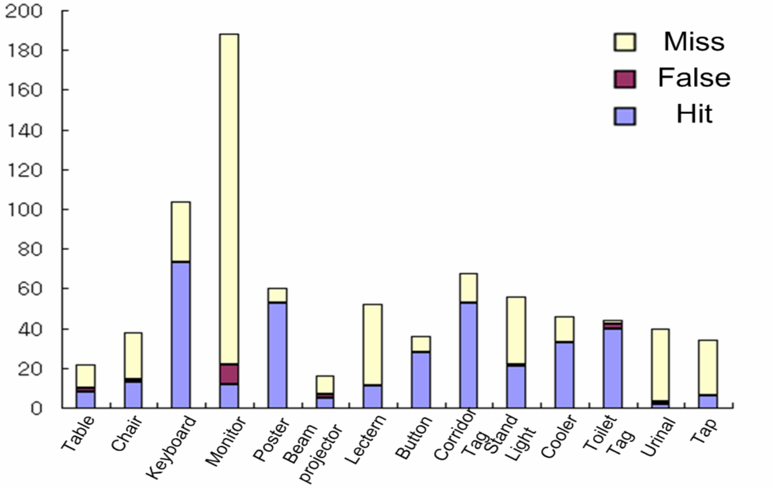

Fig. 5. Objects recognition results by SIFT

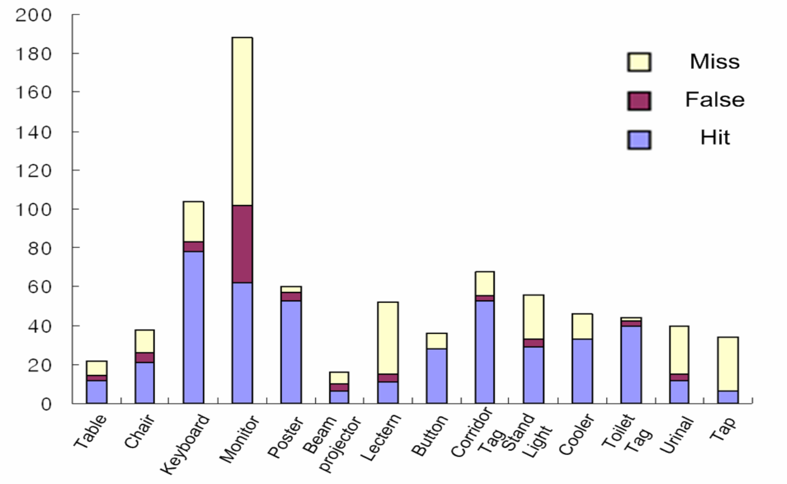

Fig. 6. Objects recognition results by the proposed method

4 Conclusions and Future Works

We have verified that the context-based Bayesian inference for scene recognition shows good performance in the complex real domains. Even though the global feature information extracted is the same, the proposed method could produce correct result using contextual information: relationship between object and place. But SIFT algorithm showed low performance when objects had insufficient textual features, and this lack of the information caused to the low performance of scene understanding. To overcome it, we need a method that disjoints objects with ontology concept, and extracts SIFT key-points in each component. Besides, we could easily adopt more robust object recognition algorithm to our method.

In the future works, we are under going to use the dynamic Bayesian network that represents previous state in scene understanding. Also, the application of the proposed method to real robot will be conducted.

Acknowledgments. This research was supported by the Ministry of Information and Communication, Korea under the Information Technology Research Center support program supervised by the Institute of Information Technology Assessment, IITA-2005-(C1090-0501-0019).

References

1. P. Korpipaa, M. Koskinen, J. Peltola, S. Mäkelä, and T. Seppänen “Bayesian approach to sensor-based context awareness,” *Personal and Ubiquitous Computing Archive*, vol. 7, no. 4, pp. 113-124, 2003.
2. A. Torralba, K.P. Mutphy, W. T. Freeman and M. A. Rubin, “Context-based vision system for place and object recognition,” *IEEE Int. Conf. Computer Vision*, vol. 1, no. 1, pp. 273-280, 2003.
3. M. Marengoni, A. Hanson, S. Zilberstein and E. Riseman, “Decision making and uncertainty management in a 3D reconstruction system,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 852-858, 2003.
4. J. Luo, A.E. Savakis, A. Singhal, “A Bayesian network-based framework for semantic image understanding”, *Pattern Recognition*, vol. 38, no. 6, pp. 919-934, 2005.
5. T.M. Strat and M.A. Fischler, “Context-based vision: Recognizing objects using information from both 2-D and 3-D imagery,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 10, pp. 1050-1065, 1991.
6. D.G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Intl. J. Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.
7. N. Oliver, A. Garg and E. Horvitz, “Layered representations for learning and inferring office activity from multiple sensory channels,” *Computer Vision and Image Understanding*, vol. 96, no. 2, pp. 163-180, 2004.
8. R.E. Neapolitan, *Learning Bayesian Network*, Prentice hall series in Artificial Intelligence, 2003.
9. J. Portilla, and E.P. Simoncelli, “A parametric texture model based on joint statistics of complex wavelets coefficients,” *Intl. J Computer Vision*, vol. 40, no. 1, pp. 49-71, 2000.
10. G.F. Cooper and E. Herskovits, “A Bayesian method for the induction of probabilistic networks from data,” *Machine Learning*, vol. 9, no. 4, pp. 309-347, 1992.