Vision based place recognition using Bayesian inference with feedback of image retrieval

Hu Yi, Chang Woo Lee
Dept. of Computer Information Science, Kusan National University
e-mail : {sharpoo7,leecw}@kunsan.ac.kr

Abstract

In this paper we present a vision based place recognition method which uses Bayesian method with feed back of image retrieval. Both Bayesian method and image retrieval method are based on interest features that are invariant to many image transformations. The interest features are detected using Harris-Laplacian detector and then descriptors are generated from the image patches centered at the features' position in the same manner of SIFT. The Bayesian method contains two stages: learning and recognition. The image retrieval result is fed back to the Bayesian recognition to achieve robust and confidence. The experimental results show the effectiveness of our method.

1. Introduction

Fast and accurate place recognition is a basic but important ability for a robot to achieve other advanced tasks such as auto navigation. In this area, vision based methods are commonly proposed and the omni camera are used [1], [2]. In vision based approaches, some authors have used color information [1] but the most popular way is based on interest features which are exploited from the environment and are shown to be suitable for the recognition tasks [2], [3], [4], [5], [6], [7]. An interest feature with high repeatability and invariant to most image transformations is important to achieve high accurate recognition in feature based approach. Evaluations [8], [9] have shown that the Harris-Laplacian detector [10], [11] attains best performance among the feature detectors, and the SIFT is the state of the art which contains both detector and descriptor [12], [13]. To recognize a place based on interest features, some authors use probabilistic methods [4], [5], [17] and image matching methods [1], [2], [3]. And some authors use the forest to tree [14] or top down method [15] to recognize the places. In this paper we present a vision based place recognition method which uses Bayesian method with feed back of image retrieval. The Bayesian method contains two stages: learning and recognition. In learning stage, a large amount of interest features are extracted from the sample images and then are clustered to generate a dictionary. For each place, a feature distribution histogram is generated with respect to the dictionary. In recognition stage, the features extracted from the frames captured by the camera are treated as observed data for Bayesian classification. The features used in Bayesian recognition stage are used in image retrieval and the result is fed back to the Bayesian recognition to achieve robust and confidence.

In Bayesian method, a similar method is proposed in [16], the author compared the SVM with Bayesian method to categorize objects using affine Harris feature. But neither SVM nor Bayesian method singly used can achieve very high accuracy. In [5], the author also used the Bayesian method to recognize the place. In that method each image in the database is represented by a set of SIFT features. For each query image and its associated keypoints a set of corresponding keypoints between query image and each image in the database is calculated using a Euclidian distance measure as described in [12]. Then the conditional probability is calculated for the Bayesian filter using the normalized correspondence set. This method requires a database of feature sets represent the images. To calculate the conditional probability the correspondence between the query image and all the feature sets in the database must be calculated which will take much compute time. And most important, the recognition rate in that paper is not very high.

This paper is organized as follows: Section 2 introduces the interest feature used in our method. Section 3 presents the place recognition method. In Section 4 the experiment results are shown to evaluate the performance of the method. Conclusions and future work is discussed in section 5.

2. Interest features

In this paper, an interest feature which combines Harris-Laplacian detector and SIFT descriptor is used. Firstly, a scale space [18], [19] is built by convolving image with different scale of Gaussian kernels. Denote $\sigma 0$ as the initial scale of the scale-space and σn as the scale of the nth image, $\sigma_n = s^n * \sigma_0$, where s is a scale factor between successive levels of the scale-space. Then we run the Harris corner detector [20] in each image of the scale space to detect candidate feature points. The Harris corner detector calculates the corner response R for all the pixels in images and choose the points as feature points which meet the following two conditions: (1) R(x, y) attains a local maxima from its neighbors. (2) R(x, y) is lager then a certain threshold. From the detected Harris corner points, accept the points as interest points for which the normalized Laplacian

of Gaussian (LoG) attains maxima in neighbor scales and lager than a threshold:

$$LoG(x,y,\sigma_n) > s^{-2}LoG(x,y,\sigma_{n-1})$$
$$\text{AND } LoG(x,y,\sigma_n) > s^{2}LoG(x,y,\sigma_{n+1})$$
$$\text{AND } LoG(x,y,\sigma_n) > T_2 \quad (1)$$

, where LoG denotes the Laplacian of Gaussian function, s denotes the scale ratio between scale σ_n and σ_{n-1} ($\sigma_n = s*\sigma_{n-1}$).

The Laplacian of Gaussian (LoG):

$$LoG(x,y,\sigma_n) = \left| I_{xx}(x,y,\sigma_n) + I_{yy}(x,y,\sigma_n) \right| \quad (2)$$

, where x, y denotes the position of the interest point in the image and σ_n denotes the scale of the image where the interest point is found in the scale space.

The descriptor of the points is made in the same manner of SIFT. At First, dominant orientation is assigned to each interest point based on local image gradient directions and the orientations within 80% of the dominant orientation are also accepted to create multi interest points. Then image patches centered at these points are used to generate the descriptors based on the image gradients at the region around the point location. See details of SIFT in [12] [13].

3. The proposed place recognition method

The Bayesian recognition method contains two main stages: learning and recognition. In learning stage, a video clip is captured for each place and then interest features are detected and extracted from the clips. Thus, for each place we get a set of interest features as learning data. All these sets of features are gathered together and then a k-means clustering method is employed to cluster them to k clusters as a dictionary. This dictionary is saved and will be used in both learning and recognition stage. In our system, k is set to 1000 based on the experimental result and also suggested by other author [16]. An approximate probability distribution histogram is made to estimate the probability distribution of the interest features. For each place, we have collected a corresponding set of interest features for learning. In each feature set corresponding to a certain place, count the number of features that are assigned to a certain cluster which means the feature is nearest to the cluster's center. For example, we denote the k centers as C = { C_1, C_2, ..., C_k }, and the features of place ω_i are distributed as: $P(C_j \mid \omega_i) = n$. Thus, for each place, a histogram is generated of which Cj means a cluster and the value of n means the number of features which are nearest to the cluster using certain distance measurement (In our system, sum of square distance is used). And the histogram is normalized to attain the approximate probability distribution histogram. To avoid the value of being zero, a Laplacian smoothing method is applied instead of standard normalization.

$$P(c_j \mid \omega_i) = \frac{c_j + 1}{|\omega_i| + k} \quad (3)$$

, where $P(c_j|\omega_i)$ denotes the value of a bin in the histogram whose index is i, and k is the number of clusters. To avoid too large value of a bin, we restrict the histogram bins' value to smaller than a certain threshold T (In experiment we use T=0.3). Then normalize the histogram again. In recognition stage, a naive Bayesian classifier is employed to achieve the recognition task. In an arbitrary place, frames are captured by a camera and interest features are obtained from the frames. These interest features are observed data used for classification. The Bayesian rule:

$$P(\omega \mid X) \propto P(\omega) \cdot \prod_{j=1}^{|X|} P(X_j \mid \omega) \quad (4)$$

, where X denotes the observed interest features, ω denotes the a category. To estimate the posterior probability of a category ω_i from X, every interest feature x_j of X is assigned to a nearest cluster to find the approximate probability which is a bin in the histogram corresponding to ω_j. Suppose that there are N(j) features in X which is nearest to center j, then P(cj| ω) in histogram of ω will be multiplied by N(j) times. Term (4) can be rewritten as:

$$P(\omega_i \mid X) \propto P(\omega_i) \cdot \prod_{j=1}^{k} P(c_j \mid \omega_i)^{N(j)} \quad (5)$$

, where $P(c_j|\omega_i)$ is from the histogram corresponding to ω_i, since from learning stage approximate probability histograms have been made for each place ω_i. After computing all $P(\omega_i|X)$ for all places, the maximum $P(\omega_i|X)$ is selected out and the corresponding ω_i is most probable place where X from. To avoid $P(\omega_i|X)$ of being too small, a logarithmic measurement is used. The final classifier:

$$Classify(X) = \arg\max_i \{\log[P(\omega_i \mid X)]\} \quad (6)$$

Single Bayesian method is insufficient for the place recognition task. An image retrieval method is supplemented to make the recognition result more accuracy. For each place, a set of images are collected to represent the place. Then interest features are obtained from each image to build a feature database. Images are matched by matching the feature sets of them. Considering the high repeatability of the feature detector, we assume that the number of features obtained from two similar views will be similar. When matching two images, the number of features will be checked at first to accept the image pair of which the numbers of features are not too different.

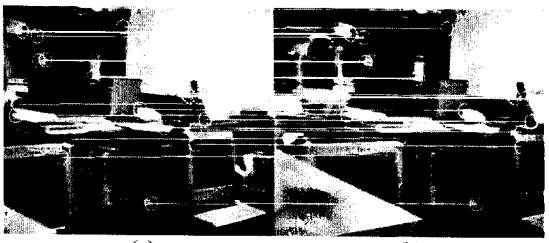
$$Num(i) > Num(j) \cdot r \wedge Num(j) > Num(i) \cdot r \quad (7)$$

, where Num(x) denotes the number of features obtained from image x. r is a ratio and is set to 0.7 in our system.

To determine feature matches, the following restrictions are used:

1. *Reject the match pair if the square distance is less than a threshold T3. (T3 = 25000, in our system)*
2. *Reject the match pair if the square distance of nearest neighbor is larger than 0.8 times the second nearest neighbor [12].*

After above restrictions, if the number of matched features is large then a percentage P of the average number of two i mages, then the two images is accepted as candidate matched pair. In experiments, P is set to 0.1. Fig.1 shows the result of matching two example images. There are about 20% feature s are correctly matched.



(a) (b)

Fig.1 Result of image matching by features. 97 features are detected and from image (a) and 110 features are detected from image (b). 21 features are well matched.

When the Bayesian recognition result attempt to chan ge which means place transition, the image retrieval method will be activated for both last place and current place and feedback the result to global recognition to co nfirm whether the transition is correct. It considers the f rames which global recognition has just analyzed and ch oose the place where more images are matched. If there is no image matches in the last place and current place, Bayesian recognition result will not change.

4. Experimental Results

We present performance of the proposed method tested over several places of indoor environment. The learning data is a video clip captured over 6 places contains 7760 frames.

Since the Bayesian classifier takes a set of features as observed data, the number of features for classification each time is considered. Fig.2 shows the result of Bayesian classification with different numbers of features. The posterior probabilities of the places are more distinguishable when the number of features increases.

With more features, the result of Bayesian classifier will return more trusty result. At the same time, more features need more frames and it will make more time to recognize a place. Another drawback of using large number of features is the ambiguity when frames from two neighbor places. Imagine that when the robot transit from place A to place B, half of features are from A and another half of features are from place B, then it's difficult to recognize. The number of features is chosen from the experimental result. Fig.3 shows the result of Bayesian place recognition using different number of features. The system achieves peak correct rate with 300 features. This number will be used in all following experiments.

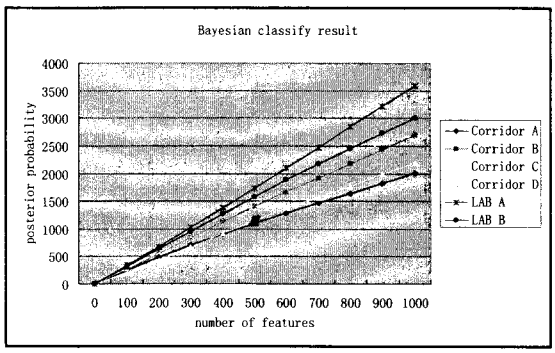


Fig.2 The result of Bayesian classification with different numbers of features.

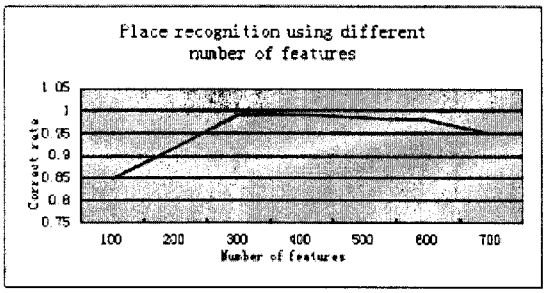


Fig.3 The result of place recognition with different number of features.

For image retrieval, 20 images are collected for each place. Fig.4 shows the result of our place recognition method. The recognition attains nearly full correct rate.

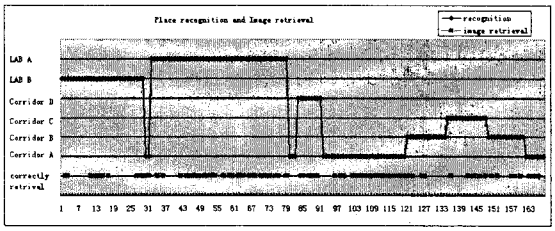


Fig.4 The result of hierarchical place recognition. (The path of the camera: Lab B -> Corridor A -> Lab B -> Corridor A -> Corridor C -> Corridor A -> Corridor B -> Corridor C -> Corridor B -> Corridor A).The blue dots in the figure represents the recognition of hierarchical place recognition result and the purple dots means that images are correctly retrieved when from the database.

5. Conclusions and Further Works

In this paper, a hierarchical place recognition method b ased on interest features is proposed. The experimental r esults show that this method achieves accurate place recognition and is very fast. In the experiment, considering the large size of learning data for clustering, which will take days time, only six places are tested. In the feature work, faster method for probability distribution estimatio n will be considered and more places will be tested to prove the performance and effectiveness of the recognition

method. Other future works such as comparison of different interest features, different clustering method and different distance measurements will be considered. Based on the current work, more deep recognition such as object recognition will be considered.

References

[1] Ulrich, I., Nourbakhsh, I.: Appearance-based place recognition for topological localization. In: IEEE International Conference on Robotics and Automation. Vol. 2 (2000) 1023--1029

[2] Andreasson, H. and Duckett, T.: Topological localization for mobile robots using omnidirectional vision and local features. In: Proceedings of the 5th IFAC Symposium on Intelligent Autonomous Vehicles, Lisbon, Portugal (2004)

[3] Wolf, J., Burgard, W., and Burkhardt, H.: Robust Vision-based Localization for Mobile Robots using an Image Retrieval System Based on Invariant Features. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2002)

[4] Kosecka, J. and Li, L.: Vision based topological Markov localization. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2004)

[5] Ledwich, L. and Williams, S.: Reduced SIFT features for image retrieval and indoor localization. In: Australian Conference on Robotics and Automation (ACRA) (2004)

[6] Dudek, G. and Jugessur, D.: Robust place recognition using local appearance based methods. In: IEEE International Conference on Robotics and Automation, San Francisco, CA, USA, April (2000) 1030-1035

[7] Se, S., Lowe, D. and Little, J.: Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. In: International Journal of Robotics Research, Vol. 21, No. 8 (2002) 735-758

[8] Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of interest point detectors. In: International Journal of Computer Vision Vol. 37, No. 2 (2000) 151-172

[9] Mikolajczyk, K. and Schmid, C.: A performance evaluation of local descriptors. In: International Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 2 (2003) 257-263

[10] Mikolajczyk, K. and Schmid, C.: Indexing based on scale invariant interest points. In: Proceedings of the International Conference on Computer Vision, Vancouver, Canada (2001) 525-531

[11] Mikolajczyk, K. and Schmid, C.: Scale and affine invariant interest point detectors. In: International Journal of Computer Vision, Vol. 60, No. 1 (2004) 62-86

[12] Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints. In: International Journal of Computer Vision Vol. 60, No. 2 (2004) 91-110

[13] Lowe, D.: Object Recognition from Local Scale-Invariant Features. In: Proceedings of the International Conference on Computer Vision, Corfu, Greece (1999) 1150-1157

[14] Murphy, K., Torralba, A. and Freeman, W.: Using the forest to see the tree: a graphical model relating features, objects and the scenes. In: NIPS (2003)

[15] Torralba, A. Murphy, K., Freeman, W. and Rubin, M.: Context-based vision system for place and object recognition. In: Proceedings of 9th IEEE International Conference on Computer Vision,Vol. 1. Nice, France (2003) 273-280

[16] Dance, C., Willamowski, J., Fan, L., Bray, C. and Csurka, G.: Visual categorization with bags of keypoints. In: ECCV International Workshop on Statistical Learning in Computer Vision (2004)

[17] Keysers, D., Motter, M., Deselaers, T. and Ney, H.: Training and recognition of complex scenes using a holistic statisticalmodel. In: DAGM 2003, Pattern Recognition, 25th DAGM Symposium, Magdeburg, Germany (2003)

[18] Lindeberg, T.: Feature detection with automatic scale selection. In: International Journal of Computer Vision,Vol. 30, No. 2 (1998) 77-116

[19] Lindeberg, T.: Scale-space theory: A basic tool for analysing structures at different scales. In: Journal of Applied Statistics, Vol. 21, No. 2 (1994) 225-270

[20] Harris, C., Stephens, M.: A combined corner and edge detector. In: Proceedings of the Fourth Alvey Vision Conference. (1988) 147-151