Research Journal of Applied Sciences, Engineering and Technology 7(14): 2806-2812, 2014
DOI:10.19026/rjaset.7.603
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: May 01, 2012 Accepted: May 26, 2012 Published: April 12, 2014

Research Article

A Spatial Visual Words of Discrete Image Scene for Indoor Localization

Abbas M. Ali, Md Jan Nordin and Azizi Abdullah
Center for Artificial Intelligence Technology, Faculty of Information Science and Technology,
Universiti Kebagsaan Malaysia, Bangi, Selangor, Malaysia

Abstract: One of the fundamental problems in accurate indoor place recognition is the presence of similar scene images in different places in the environmental space of the mobile robot, such as the presence of computer or office table in many rooms. This problem causes bewilderment and confusion among different places. To overcome this, the local features of these image scenes should be represented in more discriminate and more robust way. However to perform this, the spatial relation of the local features should be considered. This study introduces a novel approach for place recognition based on correlation degree for the entropy of covariance feature vectors. In fact, these feature vectors are being extracted from the minimum distance of SIFT grid features of the image scene and optimized K entries from the codebook which is constructed by K means. The Entropy of Covariance features (ECV) issued to represent the scene image in order to remove the confusion of similar images that are related to different places. The conclusion observed from the acquired results showed that this approach has a stable manner due to its reliability in the place recognition for the robot localization and outperforms the other approaches. Finally, the proposed ECV approach gives an intelligent way for the robot localization through the correlation of entropy covariance feature vectors for the scene images.

Keywords: Entropy covariance features vectors, grid, place recognition, SIFT K-means

INTRODUCTION

Place recognition is one of the basic issues in mobile robotics based localization through the environmental navigation. One of the fundamental problems in the visual place recognition is the confusion of matching visual scene image with the stored database images. This problem is caused by instability of local features representation. Machine learning is used to improve the localization process for known or unknown environments. This led the process to have two modes, supervised mode like (Booij *et al*., 2009; Wnuk *et al*., 2004; Oscar *et al*., 2007; Miro *et al*., 2006) and unsupervised mode, like (Abdullah *et al*., 2010). The most common tools used in machine learning is the K-means clustering technique to cluster all probabilistic features in the scene images in order to construct the codebook. Several works used clustering technique, where the image local features in a training set are quantized into a "vocabulary" of visual words (Ho and Newman, 2007; Cummins and Newman, 2009; Schindler *et al*., 2007). Clustering technique may reduce the dimensionality of features and the noise by the quantization of local features into the visual words. The process of quantizing the features is quite similar with the Bag of Words (BOW) model as in Uijlings *et al*. (2009). However, these visual words do not possess spatial relations. The BOW model is employed to get more accurate features for describing the scene image in place recognition.

In Cummins and Newman (2009), they used BOW to describe an appearance for Simultaneous Localization and Mapping (SLAM) system, which was used for a large scale rout of images. In Schindler *et al*. (2007) an informative features was proposed to be added to each location and vocabulary trees (Nister and Stewenius, 2006) for recognized location in the database. In contrast, (Jan *et al*., 2010) measured only the statistics of mismatched features and that required only negative training data in the form of highly ranked mismatched images for a particular location. In Matej *et al*. (2002), an incremental eigen space model was proposed to represent the panoramic scene images, which was taken from different locations, for the sake of incremental learning without the need to store all the input data. The study in Iwan and Illah (2000) was based on color histograms for images taken from the omnidirectional sensor, these histograms were used for appearance based localization. Recently, most works in this area are focusing on large-scale navigation environments. For example, in Murillo and Kosecka (2009) a global descriptor for portions of panoramic images was used for similar measurements to match images for a large scale outdoor Street View dataset. In

Corresponding Author: Abbas M. Ali, Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebagsaan Malaysia, Bangi, Selangor, Malaysia

Jana *et al*. (2003) qualitative topological localization established by segmentation of temporally adjacent views relied on similarity measurement for global appearance. Local scale-invariant key-points were used as in Jana *et al*. (2005) and spatial relation between locations was modeled using Hidden Markov Models (HMM). In Sivic and Zisserman (2003), the Support Vector Machines (SVM) was used to evaluate the place recognition in long-term appearance variations. The performance of the covariance proved by Forstner and Moonen (1999) and Oncel *et al*. (2006) used covariance features with integral images, so the dimensionality is much smaller and gets faster computational times. Most of the implementations need spatial features, which arises as the robot is navigated in the places which are similar, for example two offices which are furnished in a similar manner. In feature based Robot navigations, Land Marks are commonly used to find the correspondence between the current scene and the database. In Jinjun *et al*. (2010) the covariance is also used with SVM for classification purposes called Locality-constrained Linear Coding. In General, the covariance implementation results of the previous studies showed that it has a promising result for the recognition process.

The main contribution of this study is that: using the entropy of covariance features to give spatial relation for the visual words to decrease the confusion problem for visual places recognition in large indoor navigation processes. The entropy in spatial relation of features is used in many applications and was proved for recognition by Sungho *et al*. (2007).

METHODOLOGY

Clustering image features is a process of learning visual recognition for some types of structural image contents. Each image I_j contains a set of features $\{f_1, f_2, \ldots. f_m\}$ and each f_i is a 128 size element. To organize all these features into K clusters $C = (C_1 \ldots C_k)$, the features that are close to each other's will be grouped together (Sivic and Zisserman, 2003), as in (1):

$$KCL(K) = \sum_{i=1}^{n} \min_{1 \le j \le k} \left(f_{i-} x'_j\right)^p \quad (1)$$

where, K is the number of clustering means of features, p is the measurement of the distance between these features; and $x'_1, x'_2, \ldots x'_k$ are the means. In this study, SIFT grid approach is used to extract the local features *fs* for the images of 30×30 grid block. The MATLAB code used for this purpose is Lazebnik *et al*. (2006).

The local features for any selected image is represented by distance for these features from the centroid c of the codebook B, which is represented by a distance table containing m distance vectors of size (128) from each centroid c in B as in Eq. (2):

$$Dt(c) = dist(c, x_{1..m}) \quad (2)$$

The Covariance (COD) of Dt in Eq. (3) gives the covariance distances of all features related to the selected images. sb is the row size of the matrix Dt:

$$\text{COD} = diag\left(\frac{1}{Sb-1} Dt' \times Dt\right) \; sb \neq 1 \quad (3)$$

The Minimum Distance (MDT) for the table Dt in Eq. (4), produces a row of minimum value for each column in the table. The size of this row is the number of centroid c in the code Book (B), informed as sb:

$$d = MDT(c) = \min_i (Dt_i) \quad (4)$$

The covariance of minimum distance for each image will be expressed as:

$$cov(d) = \frac{1}{Sb-1} d * d' \; sb \neq 1 \quad (5)$$

The eigen values Er and eigen vectors Ev are calculated from the constructed covariance matrix as in Sebastien (2011) and used in Eq. (6) to give the covariance matrix (T). The result is optimized by multiplication of exponential entropy for (T) added with the mean of the minimum distance feature vector (X), then their sum is multiplied by exponential of the 1/trace (T) to filter the body features vector, As in Eq. (7):

$$T = Ev * diag\left(\sqrt{\frac{1}{diag(Er)+0.1}} * Ev'\right) \quad (6)$$

$$ef = x * T * e^{(entropy(T)+mean(X))} * e^{\left(\frac{1}{trace(T)}\right)} \quad (7)$$

The size of entropy of covariance feature vector (ef) is the same size as d. To speed up this calculation for the Er and Ev, the minimum distance d is subdivided into n parts to calculate the covariance for each part separately as in Fig. 1.

```
Im=Read Image
sf=Sift Feature(Im)
d= calculate_ minimum distance ( sf , code-book)
sb=size(d);
n=better division (sb)
k=1;
for  i=n:sb step n
     part_of_d k= di....i+n-1
Tk =calculate the covariance (part_of_dk as in (5, 6))
   efk=calculate the entropy of covariance feature(Tk as in (7))
   k=k+1;
End for
ef= {ef1,ef2,....efm};
```

Fig. 1: The pseudo code for speedup ef

Classification and correlation: Recognition process in this study simply uses the calculation of the Entropy for the Covariance of the minimum distance (ECV) generated from the query scene image as in Eq. (7). To examine the similarity of two images like x and y, the correlation between the two entropy feature vectors *ef*1 and *ef*2 is calculated as in Eq. (8):

$$corr(ef1, ef2) = \frac{cov(ef1, ef2)}{std(ef1) * std(ef2)} \quad (8)$$

where, the correlation coefficient is Pearson's coefficient for the two variables *ef*1 and *ef*2, that varies between -1 and +1.

The results for all correlation values are sorted; then, the maximum values are taken to be the best matching visual places. This approach is also called as a k Nearest Neighbor (k-NN). The average precision can be calculated as in Azizi (2010), where the Precision (P) of the first N retrieved images for the query Q is defined as:

$$p(Q, N) = \frac{|\{Ir | Rank(Q, Ir) \leq N \text{ such that } Ir \in g(Q)\}|}{N} \quad (9)$$

where, Ir is the retrieved image and g (Q) represents the group category for the query image.

EXPERIMENTS AND RESULTS

Two types of experiments have been conducted to check the accuracy performance of ECV.

First experiment is to test the accuracy of the proposed approach, through working on the data set of IDOL (Pronobis *et al*., 2009).

The SIFT features were extracted using SIFT grid algorithm for each image. The size of each frame image was 230×340. Figure 2 The ECV features vectors extracted using cluster number 260, it was used to express different places of the environmental navigation namely a one-person office, a corridor, a two-person office, a kitchen and a printer area. To demonstrate the accuracy performance of ECV, the algorithm implemented on various illumination condition groups (sunny, cloudy, night) for IDOL dataset each group divided into two parts such as train and test images, each parts were divided into 16 subgroups. Five different running tests were used. In addition to these experiments, mixed groups have also been used. Then the performances were reported using the average of the obtained classification results. Table 1 shows the experiment results for HBOF, MDT and ECV approach implementation on one IDOL data set using K-NN and linear SVM for WEKA software (Waikato, 2011), to classify the images corresponding to their places.

Fig. 2: SIFT features extracted using SIFT grid algorithm

Table 1: A comparison of some approaches

Classes	K-NN	SVM
HBOF	84.0784±0.1937	85.2723±0.4176
MDT	92.6356±0.2676	90.8000±0.2333
ECV	97.8509±0.2859	92.8078±0.1682

The performance of the proposed approach using k-NN is more accurate than SVM. This doesn't mean that k-NN is better than SVM, since the theoretical background for the two methods is known; therefore k-NN is adapted in the second experiment for navigation process. The accuracy performance under various illumination conditions (sunny, cloudy and night) is about 97%, depending on the specific environment difficulties. Figure 3, Shows random selections of images for testing the retrieval of the best similar 5 images according to the highest correlation values.

Indoor experiment: In this section, two more experiments were conducted. First, a simulation of navigation for the whole IDOL dataset using ECV approach was used, to check the accuracy performance for the robot navigation. This is done by using pre-stored images as landmarks from the dataset with their locations and then by giving each place its color to know the error of confusingly recognized places. Figure 4, Shows the results of simulation. Each color indicates a specific group in the dataset. The wrong correlation leads to the confusion of place recognition, which leads to give the wrong color in the topological map.

The Second experiment has done on a large room with five sections. Figure 5a shows a topological map for the room which has been tested and the pathways required from the system to navigate through. Figure 6 shows a set of random query images tested using

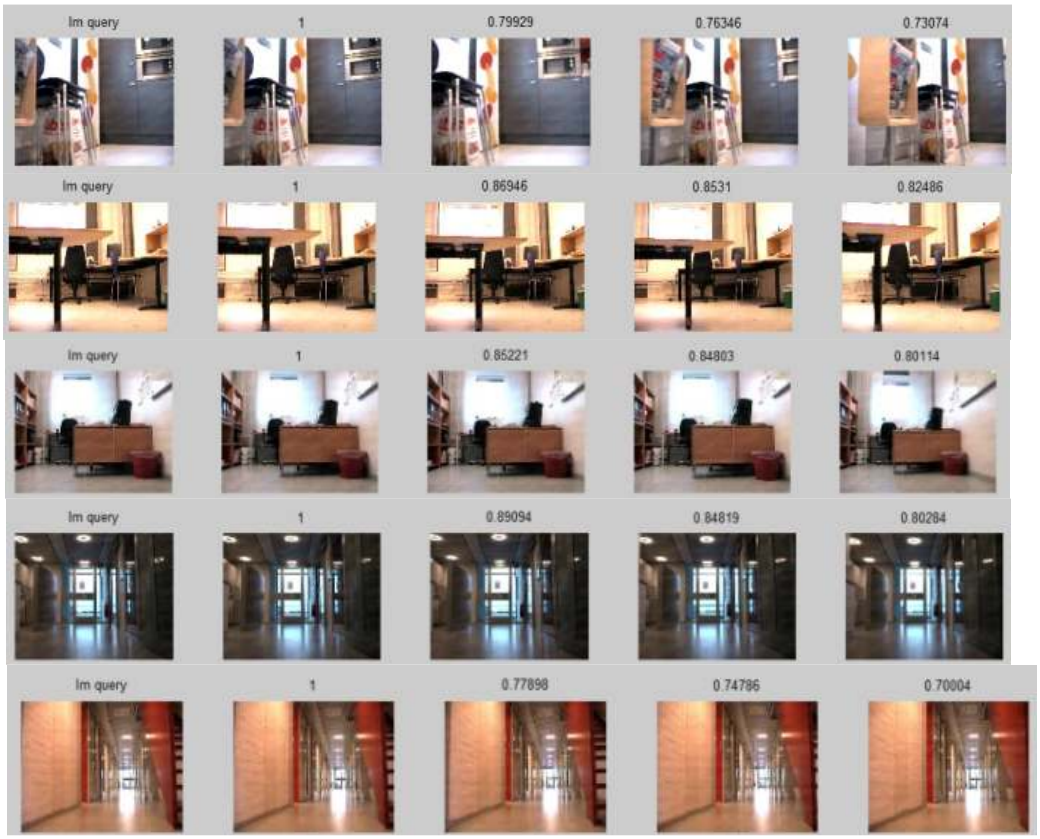
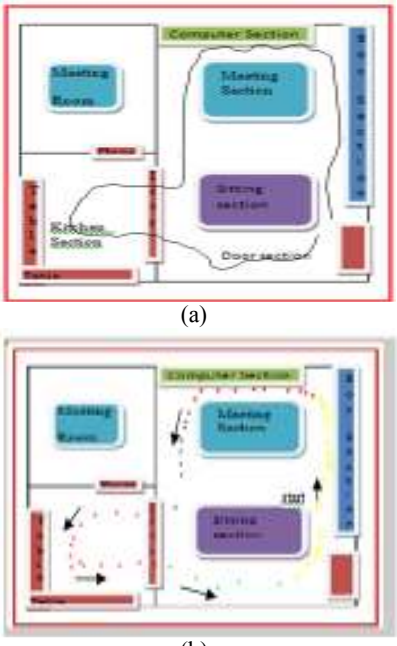


Fig. 3: Random query images and image retrieving

Fig. 4: Idol dataset groups recognition

correlation of covariance features, where each image entitled by the value of the correlation, if the title is 1, it means that the image is found and if it's less than 1 means that the retrieved image is similar to the query image.

A hand held camera navigation using ECV approach was used to verify the results, as shown in Fig. 5b. A set of land marks, were used to recognize the place for localization.



(a)

(b)

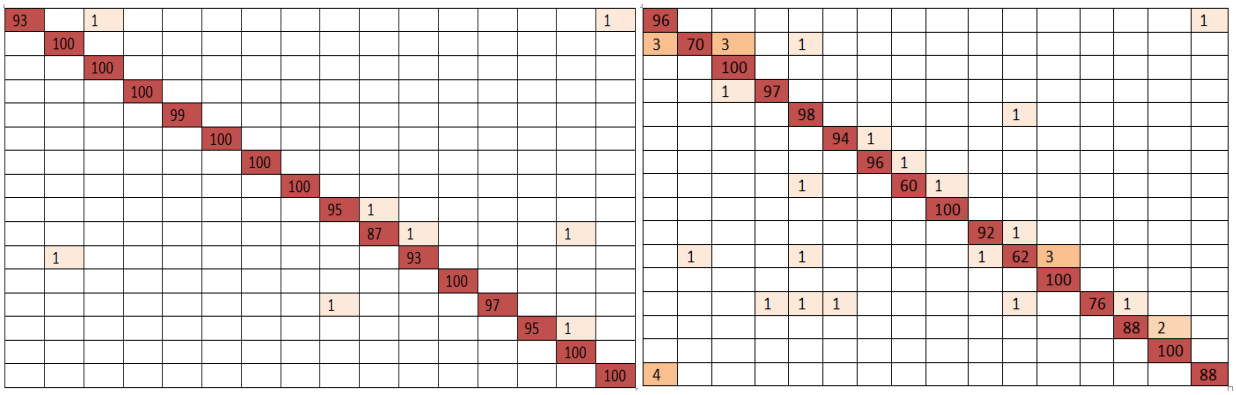
Fig. 5: (a) Topological map, the black pathway is required path for navigation, (b) the ECV recognition of places

Fig. 6: Random query images and image retrieving

DISCUSSION

The place recognition is done through a sequence of image scenes converted to visual words; the ECV for these visual words gives a spatial relation between them. The correlation between these ECV features gives an indication that to which extent does these two images were related to each other. The decision making for localization is done according to the correlation values related to correlation values between the current image scene and all of the stored landmarks. The maximum values of k nearest neighbor are used to select current localization place for the robot. The algorithm also can be used as an auditory advising for the blind person navigating through indoor environment. Place recognition based on ECV gives reliability and accurate perception for the global localization and it reduces the confusion for place recognition.

The system consistently shows that the on-line performances are more than 97% for environmental recognition, the image for the landmarks stored as entropy of covariance features was extracted from the quantized SIFT features according to the codebook. The more number of landmarks will give more accurate localization process within the navigated environment. Figure 7, shows confusion matrix for idol dataset which is the confused place recognition through landmarks recognition, that effects on the localization process. The landmarks were selected in an accurate way to be discriminated from each other, in such a way that gives accurate localization for the robot within the topological map.



(a) ECV (b) MDT

(C) HBOF

Fig. 7: Confusion matrix for IDOL, the rows and columns in the table are places group matching

CONCLUSION

One important issue in the robot localization is accurate place recognition in the environment to give accurate mapping. The problem of confusion for the similar place recognitions is a challenging issue in computer vision. Accurate spatial representation for the visual words may give a good solution for this issue. The study proposed a novel approach using correlation of Entropy Covariance minimum distance (ECV) for place recognition. ECV has been compared with some approaches using the same dataset to evaluate and measure the accuracy. The experimental results show that the proposed method can be better than the other methods. It is an establishment of an algorithm to conceptualize the environment; using spatial relations of clustered SIFT features in navigation and localization techniques.

REFERENCES

Abdullah, A., R.C. Veltkamp and M.A. Wiering, 2010. Fixed partitioning and salient with mpeg-7 cluster correlograms for image categorization. Pattern Recogn., 43(3): 650-662.

Azizi, A., 2010. Supervised learning algorithms for visual object categorization. Wiskunde en Informatica Proefschriften, Universiteit Utrecht.

Booij, O., Z. Zivkovic and B. Krose, 2009. Efficient data association for view based SLAM using connected dominating sets. Robot. Auton. Syst., 57(12): 1225-1234.

Cummins, M. and P. Newman, 2009. Highly scalable appearance-only SLAM-FAB-MAP 2.0. Proceedings of Robotics: Science and Systems. Seattle, USA.

Forstner, W. and B. Moonen, 1999. A metric for covariance matrices. Technical Report. Department of Geodesy and Geoinformatics, Stuttgart University.

Ho, K.L. and P. Newman, 2007. Detecting loop closure with scene sequences. Int. J. Comput. Vision, 74(3): 261-286.

Iwan, U. and N. Illah, 2000. Appearance-based place recognition for topological localization. Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'00). San Francisco, CA, USA.

Jan, K., S. Josef and P. Tomas, 2010. Avoiding confusing features in place recognition. Proceedings of the European Conference on Computer Vision.

Jana, K., L. Fayin and Y. Xialong, 2005. Global localization and relative positioning based on scale-invariant key points. Robot. Auton. Syst., 52: 27-38.

Jana, K., Z. Liang, B. Philip and D. Zoran, 2003. Qualitative image based localization in indoors environments. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'03). Madison, WI, USA.

Jinjun, W., Y. Jianchao, Y. Kai, L. Fengjun, H. Thomas and G. Yihong, 2010. Locality-constrained linear coding for image classification. Proceeding of IEEE Conference on Computer Vision and Pattern Classification.

Lazebnik, S., C. Schmid and J. Ponce, 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2: 2169-2178.

Matej, A., J. Matjaž, L. Aleš and R. Mobile, 2002. Localization using an incremental eigenspace model. Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'02), pp: 1025-1030.

Miro, J.V., W.Z. Zhou and G. Dissanayake, 2006. Towards visionbased navigation in large indoor environments. Proceeding of the, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS, 2006), pp: 2096-2102.

Murillo, A.C. and J. Kosecka, 2009. Experiments in place recognition using gist panoramas. Proceeding of IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops). Kyoto, pp: 2196-2203.

Nister, D. and H. Stewenius, 2006. Scalable recognition with a vocabulary tree. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06), 2: 2161-2168.

Oncel, T., P. Fatih and M. Peter, 2006. Region covariance: A fast descriptor for detection and classification. Proceedings of the 9th European Conference on Computer Vision (ECCV '06), Part II: 589-600.

Oscar, M.M., G. Arturo, B. Monica and R. Oscar, 2007. Interest point detectors for visual SLAM. Proceeding of the Current Topics in Artificial Intelligence: 12th Conference of the Spanish Association for Artificial Intelligence (CAEPIA), pp: 170-179.

Pronobis, A., B. Caputo, P. Jensfelt and H.I. Christensen, 2009. A realistic benchmark for visual indoor place recognition. Robot. Auton. Syst., 58(1): 81-96.

Schindler, G., M. Brown and R. Szeliski, 2007. City-scale location recognition. Proceeding of IEEE Conference on Computer Vision and Pattern Recognition, pp: 1-7.

Sebastien, P., 2011. Scenes/Objects Classification Toolbox. Scenes recognition toolbox for vision systems, Dec. 21, 2010, (Updated Apr. 25, 2011). Retrieved from: www.mathworks.com/... scenesobjects-classification-toolbox/...toolbox/...

Sivic, J. and A. Zisserman, 2003. Video google: A text retrieval approach to object matching in videos. Proceeding of the 9th IEEE International Conference on Computer Vision (ICCV '03). UK, Oct. 3-16, 2: 1470-1477.

Sungho, K., K. In-So and L. Chil-Woo, 2007. Visual categorization robust to large intra-class variations using entropy-guided codebook. IEEE International Conference on Robotics and Automation. Roma, pp: 3793-3798.

Uijlings, J.R.R., A.W.M. Smeulders and R.J.H. Scha, 2009. Real-time bag of words. Proceeding of ACM International Conference on Image and Video.

Waikato, S., 2011, Waikato Environment for Knowledge Analysis. Version 3.6, Waikato University, Hamilton, New Zealand.

Wnuk, K., F. Dang and Z. Dodds, 2004. Dense 3D mapping with monocular vision. Proceeding of International Conference on Autonomous Robots and Agents.