

Full Length Research Paper

Place recognition using semantic concepts of visual words

V. Rostami^{1*}, Abd Rahman Ramli², Khairulmizam Samsudin² and M. Iqbal Saripan²

¹Institute of Advanced Technology (ITMA) at Universiti Putra Malaysia.

²Faculty of Engineering, Universiti Putra Malaysia 43400 Serdang, Selangor, Malaysia.

Accepted 27 July, 2011

Applying the 'bag-of-visual-words' has recently become popular for image understanding. Although, using the histogram of visual words suffers the problem when the patches of an image faced with similar appearance corresponding to differentiate semantic concepts and vice versa. Due to varying views and dynamic objects, this problem is more complicated in the mobile robot applications such as global localization and place recognition systems. This paper presents a supervised learning framework for place recognition using the semantic concepts of visual words. Specifically, the k-mean algorithm is firstly applied to quantize the low-level visual features as bag-of-visual-words (BOVW). And then the visual latent semantic analysis (VLSA) is introduced to obtain semantic concepts of these words from the correlation of the image patches. Once obtained the semantic concepts, the corresponding of these concepts in a query image are formed as a vector of similarity density, which it can be exploited in the place recognition using the support vector machine (SVM) classifier. Experiments on synthesis and challenging indoor datasets reveal that the average recognition performance in two different datasets is improved from 77.54 to 90.92% using the histogram of BOVW and the proposed method respectively.

Key words: Place understanding, image classification, robot vision, semantic analyzing.

INTRODUCTION

Typically indoor environments are divided into places according to their functions like offices, kitchens or seminar rooms. Using this semantic information facilitates a mobile robot to perform more efficiently a variety of tasks such as localization, path planning or navigation. Therefore, quantification and formalization of the intuitive of relatedness between images in different views, has been a major challenge in this issue. In recent years, most of the research is focused on the concept of images (Leow and Lai, 1996; Liu et al., 2010) instead of the context (Landauer et al., 1998; Deerwester et al., 1990; Wolfe and Goldman, 2003; Oliva and Torralba, 2007) (for example feature, texture) for image description. One of

the problems afforded by these specialists is that specifying a measure to quantify the terms (that is a group of elements to express a concept) in the images. Early efforts at place recognition have much paid attention to modeling a place and a scene using the statistical information of images which is inspired on human perception. Subsequent literature on image descriptors approaches (Lowe, 2004; Mikolajczyk and Schmid, 2005; Wu and Rehg, 2010), it seems that the relation of image features and semantic information about places would be more proper than using directly the image features to distinguish between images for recognizing the places (Pronobis et al., 2010). In this case, all images which they have relation between their features tend to co-occur more often than the frequency appearance of their features. Suchlike, when people read a text, the relations between the words contribute to their understanding. Related pairs of words may join together

*Corresponding author. E-mail: vh_rostami@ieee.org. Tel: +60172496073.

to form larger groups of related words that can extend freely over sentence boundaries. These term groups contribute to the meaning of the whole text (Wolfe and Goldman, 2003; Oliva and Torralba, 2007). This example tells us that the relations of terms (for example words or visual features) guide human to understand a text or images from term meaning to information or knowledge.

The main goal of this paper is to find a model for places based on relationships between terms of the bag-of-visual-words (BOVW) and concepts of places. The matrix of observed co-occurrence (Wolfe and Goldman, 2003) is applied to the visual word terms of images to estimate the parameters of that underlying conceptual model for places. In this way, it can be estimated what the observed occurrences should have been. This approach exploits the local SIFT (scale invariant feature transform) features of images as visual words and latent semantic analysis (LSA) (Landauer et al., 1998) technique to build a semantic model of places. LSA is an information retrieval technique which is typically called latent semantic indexing (LSI) that analyzes the distribution of terms over a set of elements (Wolfe and Goldman, 2003; Oliva and Torralba, 2007). In this work, the BOVW distribution is considered to represent the places using terms to concept matrix (TCM) instead of terms document matrix (TDM).

RELATED WORKS

Feature extraction

The feature extraction step intends to provide a representation of an image in a plausible feature vector which includes distinctive visual features or image patches descriptor. It should be pointed out that the feature extraction methods encounter with the general issues such as dynamic background, occlusion, viewpoint and light condition changes. The visual features are commonly obtained by one of the following two types: global or local features. Global visual features are extracted from whole of image pixels such as color (Deerwester et al., 1990; Landauer et al., 1998), segmentation (Sojodishijani et al., 2010), texture (Wolfe and Goldman, 2003) or a combination of both (Siagian and Itti, 2007; Pronobis et al., 2010). These types of visual features are limited to recognize places (for example rooms, offices and cluttered scenes) because it is harder to infer a change in different views, even when the robot changes its position. Local features are computed over a partial region of image where that should be a local salient point. A well-known approach for local visual features is scale-invariant feature transform (SIFT) (Lowe, 2004). This approach has been utilized in recent years for object recognition and land mark detection. The SIFT is commonly used in place understanding because the local processing can

overcome to the occlusion and some viewpoint changes (Se et al., 2005; Pronobis et al., 2010; Liu et al., 2010).

The local feature extraction algorithms generally consist of two phase: detecting the interest points and making the descriptors. First phase identifies a set of characteristics of points in the image. That could be re-detected in various transformations and illuminations conditions in which these points are termed key-points. The second phase describes the detected key-points as local patch considering invariant to image rotation, image transform and illumination changes.

Latent semantic analyzing

In image matching, the most measures have been given as a function of the frequency of occurrence in a sequence of images. All features are not same discriminative measure to distinguish the images in a same rank. That means, features with high frequency are indicated that they are more common and likewise features with low frequency are too rare. Therefore, both of them do not contribute significantly to the content of places. Consequently, the highest ability to distinguish the content is based on features with medium frequency (Wolfe and Goldman, 2003). Hence all features below a lower and above an upper than threshold are excluded. Latent semantic analysis (LSA) is a natural language processing technique proposed by Deerwester et al. (1990). They addressed a general method for creating similarity measurement by finding the relation between the combination of documents and words from a vocabulary. The LSA is closely concerned to neural net models in which singular value decomposition (SVD) is used (Landauer et al., 1998). A matrix decomposition technique is closely related to factor analysis of documents. That is applicable to documents analyzed such as corpora (Wolfe and Goldman, 2003; Deerwester et al., 1990) source code (Kuhn et al., 2007) which are volumes of relevant experienced by people. LSA enhances input vectors semantically to overcome problems with similarity/synonym and discriminatory/polysemy measurements. Similarity denotes that multiple features have same meaning and polysemy denotes a single feature as a word in documents has multiple meanings (Landauer et al., 1998; Coles, 2001).

As a practical approaches that were reported in Wolfe and Goldman (2003) and Deerwester et al. (1990) for the characterization of word meaning, they applied LSA for computing the reliance of passages to its words in which were correlated with human cognitive phenomena. Delponte et al. (2006) proposed an improved SVD matching between pair of images for same scene and Khadem et al. (2010) proposed semantic clustering the visual words for scene recognition.

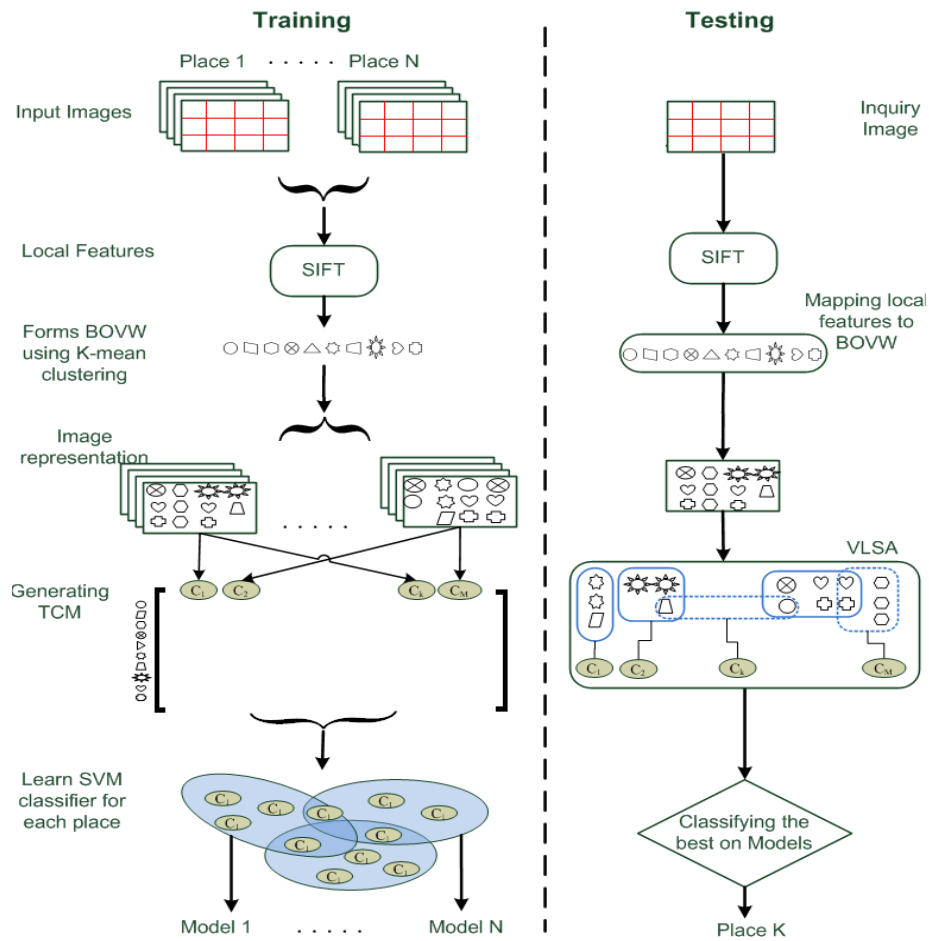


Figure 1. A schematically presentation of proposed frame work. The left of dash line indicates to the BOVW and TCM construction and the right side is recognition path way.

Singular vector decomposition (SVD)

SVD is based on linear algebra theorem (Wall et al., 2003; Delponte et al., 2006) that says a rectangular matrix A can be divided into the product of three matrices as follows:

$$A = U \times S \times D^T \tag{1}$$

Where, the singular values (diagonal matrix $S_{m \times n}$) are Eigen-values from $A^T \times A$ or $A \times A^T$, and Eigen-vector of $A \times A^T$ are termed ‘left singular vectors’ (orthogonal matrix $U_{m \times m}$) while transpose of an orthogonal matrix Eigen-vectors of $A^T \times A$ are ‘right singular vectors’ (orthogonal matrix $D_{n \times n}$). In Equation 1 (if A with dimension $m \times n$ and $m > n$) the off-diagonal entries of S are all 0’s and diagonal elements are satisfied as follows:

$$\delta_1 \geq \delta_2 \geq \delta_3 \geq \dots \delta_k \geq 0 \tag{2}$$

It reveals that the rank of A is equal to k , which is the

number of nonzero singular values as well as the amplitude of the singular values provides a measure of how close A is to a matrix of lower rank k . Using SVD on natural data utilizes a distribution of singular values that follows the power law theory (Coles, 2001): in natural data (for example corpora, news and natural images), there are a few terms with large values and a long tail with very small values. Therefore, even if the dimension of data matrix goes into millions, there are typically only about 200-500 relevant singular values (Deerwester et al., 1990; Delponte et al., 2006).

PROPOSED METHODS

Since the performance of place recognition approaches using the image features or visual words directly affects to the problems in terms of huge intra-class variation and different parts or viewpoints in the same location. These approaches have been gaining attention to generate semantically and/or conceptual image descriptors in recent literature (Khadem et al., 2010; Delponte et al., 2006; Kuhn et al., 2007). Figure 1 shows schematically the basic architecture of the proposed method for training (red arrows) and

recognition (blue arrows) path way. At the first stage, the local features using SIFT (Lowe, 2004) are extracted from each regular cells of input images. The second stage is to find a way to quantize the primary features of patches to terms of visual words. Once the visual words terms are obtained, the next stage is to find concepts of the patches by exploiting the LSA technique and visual word terms to concept matrix (TCM). Finally, the place recognition task via SVM classifier identifies an input image where it is taken at.

Embedding the terms into concept matrix

As shown in Figure 1 an input image is gridded to $w \times w$ cells then each cell is assigned to a visual word, so the aggregation of cells indicates the concepts of the image. For this aim, first a bag of visual words (BOVW) is built by K-mean clustering algorithm over the SIFT features which are sampled from all of the training images. In this step, according to empirical finding based on (Teynor and Burkhardt, 2007) we obtained a set of 50 K sampled features by random sampling and cluster them to $K = 400$ words. After obtaining the BOVW, each local feature in an input image is further quantized into one word in the BOVW according to the nearest neighbor regulation. The terms of an input image are obtained by selecting the index of words for all cells into a quantized terms vector by using visual latent semantic analysis (VLSA). Those are likewise using words in document analyzing.

Visual LSA algorithm

As we know a sequence of words in a paragraph follows a concept in our mind; identically, a sequence of images which are taken from a place can label a name in our mind. Human's brain saves key-points and relation of them not all of images for remembering the places. Consequently, we extract the latent semantic of images which they are taken in a place by applying LSA (Coles, 2001). Here, terms refer to the visual words and document refers to the concepts which came out from video images. In this case, word-document matrix is called term to concept matrix in which a set of concepts describes a place using the relation of its landmarks. Algorithm 1 shows an unsupervised algorithm that is executed for any input query terms in two phases: computing the similarity values of existing concepts (steps 1 to 3) and updating concepts (step 4). First phase is only for recognition while both phase runs for training the sampled images. First phase computes the similarity values between the input query terms and term to concept matrix (TCM). This matrix will be updated if there is a concept for input image, else it will be added as a new concept.

The algorithm in Algorithm 1 returns a vector of similarity coefficients concept. In this vector, element i^{th} contains a scalar value which implies to the semantic similarity measure between the query image and the concept i^{th} . Hence, the high value of this vector indicates the high conceptual similarity. This vector can be fed to the classifier to recognize the place of input image.

Algorithm 1. Generalization in pseudo-code the semantic analyzing algorithm of visual words.

Algorithm VLSA(Input TCM, q, k)

/* Compute the Eigen-values of all clusters in the k-dimensional space by : */

Step 1:

(U, S, D) \leftarrow compute SVD of TCM with rank k;

/* Compute the co-ordinate of the input query vector q */

Step 2:

$Q \leftarrow q^T \times U \times S$;

/*Compute the similarity coefficients between the input query vector

and the i^{th} cluster for all of clusters that denotes by D. */

Step 3:

for j = 1 to n

$$\text{Concept}(j) \leftarrow \frac{\sum_i Q_i \times D_{ij}}{\text{norm}(Q) \times \text{norm}(D_j)}$$

End

If testing phase mod

return **Concept**

End if

/* Update the TCM */

Step 4:

If max(Concept) ≤ 0

Add q^T to TCM as a new concept

else

Step 5:

J \leftarrow ArgMax (Concept);

for i = 1 to m

$$\text{TCM}_{ij} \leftarrow \text{TCM}_{ij} + q_i / \text{norm}(q)$$

End

End if

Return Concept , TCM

End of function

Classification: Multi-class SVM

The choice of appropriate classifier is another key factor for a recognition system. Since the state-of-the-art performance in the visual recognition systems (Pronobis et al., 2010; Liu et al., 2010) multi-class SVM classifier using one versus one strategy is applied. It focuses on structural risk minimization by maximizing the decision margin. In this classifier the kernel function plays an important role to achieve high performance of SVMs. Having the conceptual image representation, the radial basis function (RBF) is applied as a kernel as $K(x_i, x_j) = \exp(-g \|x_i - x_j\|^2)$, $g > 0$. In the training process, it is crucial to find the right parameters. Tradeoff cross validation on training set with a grid search varying (c, g) is performed to find the best parameters with highest accuracy. Within the optimal parameters, it then will be applied to assign a query image to the label of its category (Pronobis et al., 2010; Chang and Lin, 2001). The inputs of classifier are a set of labeled training examples $\langle x_i, y_i \rangle$, $i = 1, \dots, N$, where each x_i is the concepts of query image i which is obtained by the VLSA algorithm and $y_i \in \{\omega_1, \dots, \omega_c\}$ is a label to indicate the place number.

EXPERIMENTAL RESULTS AND DISCUSSION

The experiments are performed several times using all of possible permutations of the training and test sequences. First the discriminancy of classes based on LSA using synthesis data was accomplished. Then the method was run on video images with two diverse environments STL (sensor technology lab) and INDECs (indoor environment under changing conditions) database (Pronobis and

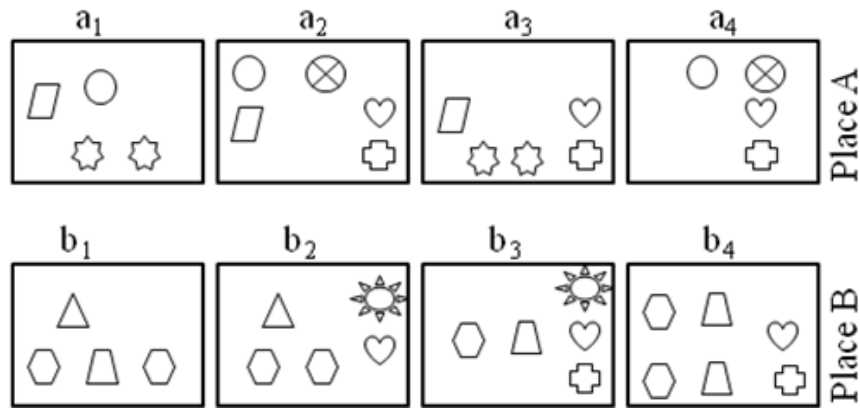


Figure 2. A simple dataset consisting of 8 synthesis images for two places A and B. The visual feature words of images are represented with 10 non-homological small icons.

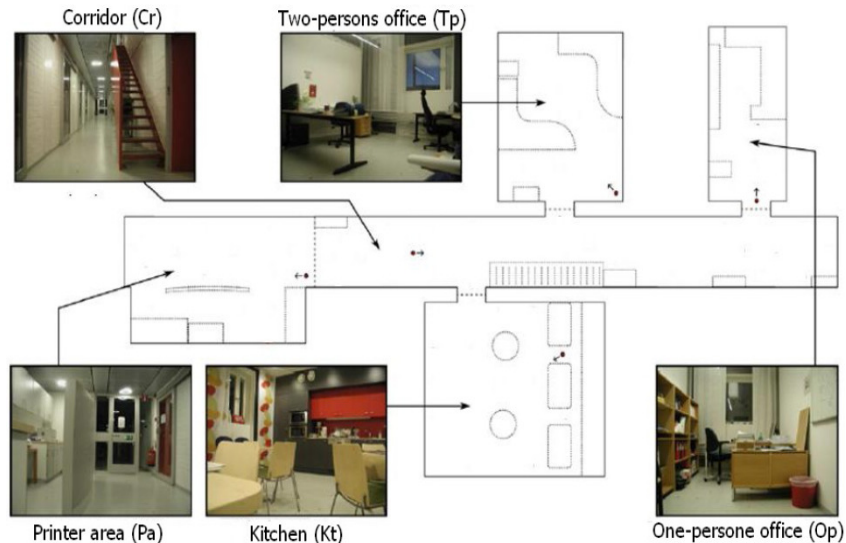


Figure 3. A general map of the part of the office environment that was imaged during acquisition of the INDECS databases.

Caputo, 2005). The map and snapshot of places where images are taken from arrowed points are shown in Figures 2 and 3. In all experiments video images with 240 x 240 pixels and 24 bits RGB color were exploited which are popular sizes in robotic applications.

Experiments on conceptual discriminancy

This experiment explains the correlations of images within class and between classes using synthesis data for two places A and B. As shown in Figure 2, a small example (included 8 images with 10 BOVW) that gives the favorite terms to analyze the latent semantic concept. This dataset can be described by terms to concept matrix

(TCM_{mn}) where m is number of visual word terms and n is number of images samples, here each image indicates one concept. As shown in Table 1, each cell of this matrix indicates the frequency appearance which a visual feature occurs in an image. In the Table 1, correlations among the images were generally 0.24 within place A, 0.54 within place B and -0.29 between A and B. These values indicate the discriminative power of selected visual features. In the reconstructed $\hat{TCM} = U_{m \times k} S_{k \times k} D_{k \times n}$ using two singular values ($k = 2$), the average of correlations were dramatically increased from 0.24 to 0.96 for place A and from 0.54 to 0.98 for place B. This happening is not because of the similarity between the images in rows of TCM while the \hat{TCM} constructed with

Table 1. The occurrences of words in images (TCM).













| BoW | a1 | a2 | a3 | a4 | b1 | b2 | b3 | b4 |
|---|----|----|----|----|----|----|----|----|
|  | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
|  | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
|  | 0 | 0 | 0 | 0 | 2 | 2 | 1 | 2 |
|  | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
|  | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
|  | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
|  | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 2 |
|  | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
|  | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
|  | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |

Table 2. Inter-correlation among vectors of image sample using the original full dimensional source data from Table 1.

| | a1 | a2 | a3 | a4 | b1 | b2 | b3 |
|----|-------|-------|-------|-------|------|------|------|
| a2 | 0.00 | | | | | | |
| a3 | 0.67 | 0.15 | | | | | |
| a4 | -0.18 | 0.82 | 0.00 | | | | |
| b1 | -0.36 | -0.60 | -0.45 | -0.49 | | | |
| b2 | -0.45 | -0.45 | -0.33 | -0.30 | 0.67 | | |
| b3 | -0.60 | -0.20 | -0.15 | 0.00 | 0.30 | 0.45 | |
| b4 | -0.45 | -0.25 | -0.19 | -0.10 | 0.68 | 0.37 | 0.75 |

the relations and coincidence of occurrence of visual words. However the average of correlations between A and B was decreased from -0.29 to -0.45, it is not as much as within class improvement. This is because of appearing some visual feature words in both places with same occurrences such as   appearing in a2, a3, a4, b3 and b4 (Tables 2 and 3). However, LSA improved similarity distance more than 0.95 within two classes as well as discrimination power -0.45 between classes. It can also reduce the dimension of feature space when singular values are set to a proper threshold.

In practical we eliminated vectors where their singular values are less than 0.1 of the first singular value. Compared with classifiers based on K^2 and Euclidian distances, LSA deals with coincident occurrences of features words or objects in the scenes of the places. To represent the effects of coincident occurrences of feature words for decision making of classifiers, similarity values between all images and a query image in class B such that $Q_1 = (0\ 0\ 0\ 1\ 0\ 0\ 0\ 1\ 1\ 1)^T$ is computed using the LSA algorithm (Algorithm 1) and K^2 distance. According to Table 4 and 1 NN (one nearest neighborhood) classifier, the query image is match to the b3 in place B with

similarity value 0.687 using LSA and is mismatched to the a4 in place A with similarity value 0.75 using histogram of BOVW with K^2 distance.

Experiments on STL and INDECS datasets

The STL dataset consist of several sets of pictures taken from the rooms and corridors in STL by exploring Active Media Pioneer robot. Both INDECS and STL databases consist of five indoor places where the pictures are taken from many view points under different illumination and weather conditions at different time. Therefore, some activities such as appearing people and moving furniture are changed in the rooms over the times. These experiments follow the classification accuracy using the proposed method and histogram of weighted visual words (Cai et al., 2010) in term of changing viewpoints and illumination. All experiments have been done over again ten times under various environmental conditions of places. The average result per trial was recorded to present in experimental results. The experimental results on the indoor places are reported in confusion tables.

Table 3. Inter-correlation among vectors of the reconstructed images where $k = 2$ in the LSA algorithm.

| | a1 | a2 | a3 | a4 | b1 | b2 | b3 |
|----|-------|-------|-------|-------|------|------|------|
| a2 | 0.94 | | | | | | |
| a3 | 0.98 | 0.99 | | | | | |
| a4 | 0.88 | 0.99 | 0.95 | | | | |
| b1 | -0.82 | -0.59 | -0.70 | -0.46 | | | |
| b2 | -0.75 | -0.49 | -0.62 | -0.36 | 0.99 | | |
| b3 | -0.57 | -0.27 | -0.41 | -0.12 | 0.94 | 0.97 | |
| b4 | -0.69 | -0.40 | -0.54 | -0.26 | 0.98 | 1.00 | 0.99 |

Table 4. Comparing the discrimination of LSA and K^2 distance. Each column indicates the similarity measure between Q_i and images.

| | a1 | a2 | a3 | a4 | b1 | b2 | b3 | b4 |
|---------|--------|-------|-------|-------|--------|-------|-------|-------|
| LSA | -0.254 | 0.314 | 0.147 | 0.497 | -0.608 | 0.397 | 0.687 | 0.149 |
| $1-K^2$ | 0.000 | 0.667 | 0.444 | 0.750 | 0.000 | 0.444 | 0.667 | 0.400 |

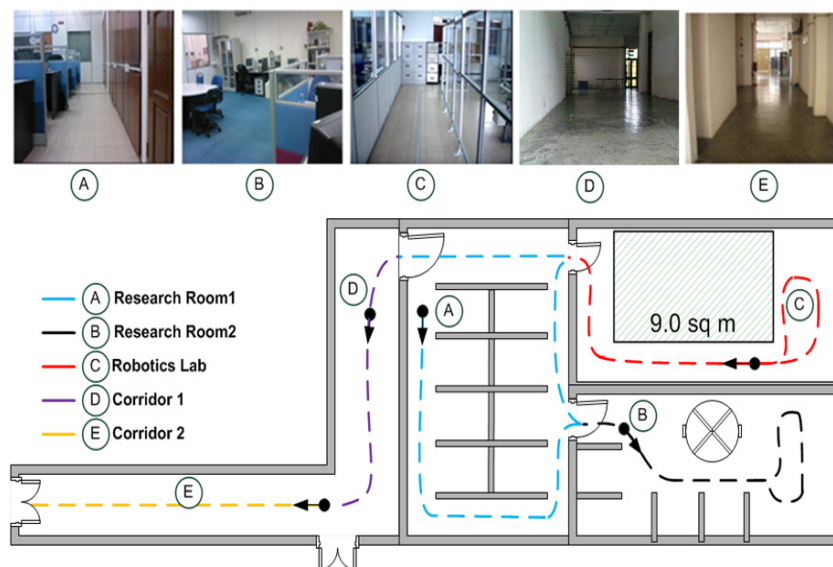


Figure 4. Five different types of indoor environments (in STL) followed by the robot at each place.

Labels of column and rows in these tables indicate the place names which are marked in Figures 3 and 4. The individual actual place is in the i^{th} row, and j^{th} column is the average percentage of predicted inquiry images from learned places. Tables 5 and 6 shows that the system can classify the places of robot during the testing phase with INDECS dataset, totally 75.16% using weighted BOVW and 90.52% using the VLSA algorithm. Boundaries between the five rooms were marked with dashed lines. The pictures are taken from the database

and show the interiors of the five rooms. The small arrows were used to indicate the viewpoints at which the presented pictures were taken (Pronobis and Caputo, 2005). The example pictures are taken from arrowed dots viewpoints.

Rooms and corridors are marked with different colored paths in which arrows indicate the direction of driving of the robot. The Tables 7 and 8 show that the system can classify the places of robot during the testing phase with STL dataset which its map is presented in Figure 4,

Table 5. Confusion matrix for INDES dataset using weighted BOVW.

| | | Predicted places by algorithm | | | | |
|---------------|----|-------------------------------|--------|--------|--------|--------|
| | | CR (%) | PA (%) | KT (%) | OP (%) | TP (%) |
| Actual places | CR | 92.60 | 3.80 | 1.20 | 1.00 | 1.40 |
| | PA | 12.50 | 59.70 | 8.20 | 10.20 | 9.40 |
| | KT | 6.80 | 1.90 | 72.80 | 9.90 | 8.60 |
| | OP | 5.80 | 2.90 | 6.80 | 79.10 | 6.40 |
| | TP | 4.30 | 2.70 | 8.80 | 12.60 | 71.60 |

Table 6. Confusion matrix for INDES dataset using VLSA algorithm.

| | | Predicted places by algorithm | | | | |
|---------------|----|-------------------------------|--------|--------|--------|--------|
| | | CR (%) | PA (%) | KT (%) | OP (%) | TP (%) |
| Actual places | CR | 93.20 | 4.80 | 0.20 | 0.70 | 1.10 |
| | PA | 7.50 | 87.40 | 1.20 | 1.50 | 2.40 |
| | KT | 1.80 | 1.90 | 92.30 | 2.90 | 1.10 |
| | OP | 1.90 | 1.30 | 3.60 | 90.10 | 3.10 |
| | TP | 2.50 | 1.60 | 2.70 | 3.60 | 89.60 |

Table 7. Confusion matrix for STL dataset using weighted BOVW.

| | | Predicted places by algorithm | | | | |
|----------------|---|-------------------------------|-------|-------|-------|-------|
| | | A (%) | B (%) | C (%) | D (%) | E (%) |
| Actual segment | A | 90.20 | 3.40 | 2.10 | 2.70 | 1.60 |
| | B | 13.20 | 67.30 | 7.80 | 10.40 | 1.30 |
| | C | 2.40 | 6.70 | 85.30 | 2.10 | 3.50 |
| | D | 3.20 | 4.60 | 5.30 | 75.60 | 11.30 |
| | E | 2.10 | 3.80 | 5.70 | 7.20 | 81.20 |

Table 8. Confusion matrix for STL dataset using VLSA algorithm.

| | | Predicted places by algorithm | | | | |
|----------------|---|-------------------------------|-------|-------|-------|-------|
| | | A (%) | B (%) | C (%) | D (%) | E (%) |
| Actual segment | A | 94.50 | 2.50 | 0.80 | 1.70 | 0.50 |
| | B | 4.80 | 86.70 | 4.10 | 3.20 | 1.20 |
| | C | 1.40 | 2.80 | 94.10 | 1.10 | 0.60 |
| | D | 0.80 | 1.10 | 2.80 | 89.70 | 5.60 |
| | E | 0.40 | 1.20 | 1.30 | 5.20 | 91.90 |

totally 79.92% using weighted BOVW and 91.38% using the VLSA algorithm.

CONCLUSION

This article presented a hierarchy model for place

understanding base on 'latent semantic analysis'. In this model, first SIFT feature extraction technique was applied to detect and describe the low level information of pixels into analogical feature vectors. Then, these vectors were mapped to BOVW which was obtained k-mean clustering algorithm. LSA technique was used to extract concept of an input image by employing the SVD on the TCM which

was represented the appearances of visual words. The major support of these claims has come from using LSA to derive measures of the similarity between the visual words and concept of images which are taken at the same place. The improvement of discriminative power from -0.24 to -0.54 in the first experiments has shown that the meaning similarities are derived closely human's visual matching. Using direct camera needs to rotate the robot and to correlate images in different views of a place. Furthermore, semantic knowledge of lines during complex goal-driven will be furthered for developing place understanding approaches. Therefore, this research can be continued in the future works by involving salient line segments and indexing images semantically for place understanding by using panoramic vision.

REFERENCES

- s H, Yan, F, Mikolajczyk K (2010). Learning weights for codebook in image classification and retrieval. In *proc. cvpr'10*. IEEE, pp. 2320–2327.
- Chang CC, Lin CJ (2001). Libsvm: a library for support vector machines. Available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- Coles S (2001). An introduction to statistical modeling of extreme values. Springer: London. MSOR Connections, 2: 2.
- Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990). Indexing by latent semantic analysis. *J. Am. society inf. Sci.*, 41: 391-407.
- Delponte E, Isgr, F, Odone F, Verri A (2006). Svd-matching using sift features. *Graphical models. Special Issue on the Vision, Video and Graphics Conference*. 68: 415-431.
- khadem BS, Farahzadeh E, Rajan D, Sluzek A (2010). Embedding visual words into concept space for action and scene recognition. *Proceedings of the British machine vision conference*. BMVA Press. 15:1-15.
- Kuhn A, Ducasse S, Girba T (2007). Semantic clustering: Identifying topics in source code. *12th Working Conference on Reverse Engineering on Information and software technology*, 49: 230-243.
- Landauer TK, Foltz PW, Laham D (1998). An introduction to latent semantic analysis. *Discourse processes*. 25(2): 259-284.
- Leow WK, Lai SY (1996). Invariant matching of texture for content-based image retrieval. *Workshop on Texture Analysis in Machine Vision*. pp. 133–156.
- Liu S, Xu D, Feng S (2010). Discriminating semantic visual words for scene classification. *leice transactions on information and systems*, 93(6): 1580–1588.
- Lowe DG (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 17(5): 590-603.
- Mikolajczyk K, Schmid C (2005). A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*. Pp. 1615–1630.
- Oliva A, Torralba A (2007). The role of context in object recognition. *Trends in cognitive sciences*, 11(12): 520-527.
- Pronobis A, Caputo B (2005). The kth-index database. *Kungliga tekniska hoegskolan, cvap, tech. rep. cvap297*. Available at <http://cogvis.nada.kth.se/INDECS/>.
- Pronobis A, Caputo B, Jensfelt P, Christensen HI (2010). A realistic benchmark for visual indoor place recognition. *Robotics and autonomous systems*. 58: 81–96.
- Se S, Lowe DG, Little JJ (2005). Vision-based global localization and mapping for mobile robots. *IEEE transactions on Robotics*, 21(3): 364-375.
- Siagian C, Itti L (2007). Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE transactions on pattern analysis and machine intelligence*. Pp. 300-312.
- Sojodishijani O, Rostami V, Ramli AR (2010). A video-rate color image segmentation using adaptive and statistical membership function. *Sci. Res. Essays*, 5(24): 3914-3925.
- Teynor RA, Burkhardt H (2007). Fast codebook generation by sequential data analysis for object classification. I: 610–620.
- Wall M, Rechtsteiner A, Rocha L (2003). Singular value decomposition and principal component analysis. *A practical approach to microarray data analysis*, pp. 91–109.
- Wolfe MBW, Goldman SR (2003). Use of latent semantic analysis for predicting psychological phenomena: Two issues and proposed solutions. *Behav. Res. methods*, 35(1): 22-31.
- Wu J, Rehg JM (2010). Centrist: A visual descriptor for scene categorization. *IEEE transactions on pattern analysis and machine intelligence*.