Vision-based Global Localization Using a Visual Vocabulary*

Junqiu Wang
National Laboratory on Machine Perception
Peking University
Beijing 100871, China
jerywang@public3.bta.net.cn

Roberto Cipolla
Department of Engineering
University of Cambridge
Cambridge, CB2 1PZ, UK
cipolla@eng.cam.ac.uk

Hongbin Zha
National Laboratory on Machine Perception
Peking University
Beijing 100871, China
zha@cis.pku.edu.cn

***Abstract*— This paper presents a novel coarse-to-fine global localization approach that is inspired by object recognition and text retrieval techniques. Harris-Laplace interest points characterized by SIFT descriptors are used as natural landmarks. These descriptors are indexed into two databases: an inverted index and a location database. The inverted index is built based on a visual vocabulary learned from the feature descriptors. In the location database, each location is directly represented by a set of scale invariant descriptors. The localization process consists of two stages: coarse localization and fine localization. Coarse localization from the inverted index is fast but not accurate enough; whereas localization from the location database using voting algorithm is relatively slow but more accurate. The combination of coarse and fine stages makes fast and reliable localization possible. In addition, if necessary, the localization result can be verified by epipolar geometry between the representative view in database and the view to be localized. Experimental results show that our approach is efficient and reliable.**

***Index Terms*— Vision-based localization, visual vocabulary, scale invariant features, Mobile robots.**

I. Introduction and Related Work

Mobile robot localization, which estimates a robot's position relative to its environment, is a prerequisite for robot autonomous navigation. The two key problems of mobile robot localization are global localization and local position tracking [19]. Global localization aims to determine the robot's position in an a priori or previously learned map without any other information than that the robot is somewhere on the map. Given the initial robot pose, local tracking is the problem of keeping track of that position over time. Global localization gives mobile robots capabilities to deal with initialization and recovery from ”kidnaps” [15]. In this paper, we describe a complete vision-based global localization system.

Vision-based global localization using natural landmarks are highly desirable for a wide range of applications. Different from other sensors such as sonar sensors and range finders, visual sensors are passive and do not emit energy into the environment. Moreover, it is often possible to recover the robot's current position with a single image captured by a camera mounted on the robot because of the richness of the visual information [4], [7], [15], [22]. Both global and local visual information have been used in many localization systems. Global visual information such as collections of views [24] or color histograms [17] are simple features which are easy to detect. However, they are sensitive to illumination changes. Torralba et al. use texture features through wavelet image decomposition [21]. Their system can recognize specific places and new places. It is difficult, however, for this system to recover exact relative position of a robot.

The difficulty in vision-based localization is how to determine the identity of an environment in the presence of illumination and viewpoint changes and occlusion. Vision-based localization is similar to object recognition in this aspect. In recent years, great progress has been made in the use of invariant features for object recognition and matching. Schmid and Mohr propose a local feature detector for general image recognition problems [14]. Mikolajczyk and Schmid extend this idea to the Harris-Laplace detector which detects Harris interest points at several scales and then selects the right scale by computing the maximum Laplace function [10]. Several local descriptors are available for interest point description. SIFT proposed by Lowe is more powerful than others because it is designed to be invariant to a shift of a few pixels in the interest region position, and this error is one that often happens [9]. Wang, Cipolla and Zha have proposed a localization strategy based on the Harris-Laplace interest point detector and the SIFT descriptor [22]. In their system, each location is represented by a set of interest points that can be reliably detected in images. This system is robust in the environments where occlusion and outliers exist. Košecká and Yang also characterize scale-invariant key points by SIFT descriptor in their localization system [7]. These localization systems have to match a new view to database by nearest neighbor search which is not efficient enough for robot localization. Thus, it is necessary to develop techniques for more efficient localization.

The *Vector Space Model*(VSM), which has been successfully used in text retrieval, is employed in this work to accelerate the localization process. In the VSM, a collection of documents is represented by an inverted index. In this index, each document is a vector and each dimension of the vector represents a count of the occurrence for a word [12]. The documents for retrieval are parsed into

*This work was sponsored in part by the National Science Foundation of China under grant No. 6017525.

words based on a vocabulary; then different weights are assigned to each term according to the frequency of the term in the document. A visual vocabulary is constructed for realizing these ideas in our localization system. We use the *k*-means algorithm [5] to learn a visual vocabulary where each term is a cluster of descriptors with similar appearance. An inverted index is built based on this visual vocabulary. Text retrieval techniques have been used in image retrieval [18] and video retrieval [16]. They use different feature detectors that are slow and not suitable for localization.

The Epipolar geometry is employed in this work to verify the localization result by discarding the outliers. The epipolar geometry is the intrinsic projective geometry between two views, which is contained in the fundamental matrix. The fundamental matrix is computed from correspondences of points and used to find the outliers that are not correct correspondences. When the result of the location recognition is ambiguous, there might be two or even three locations getting almost the same number of point correspondences. Some of the correspondences are outliers.

A. Overview

The global localization strategy in this paper is coarse-to-fine. The flowchart of this approach is given in Fig. 1.

First of all, representative images are captured in the first exploration. Next, scale invariant interest points are detected by the Harris-Laplace detector [10]. The Harris-Laplace detector is built in a multi-scale framework, which makes these interest points robust to scale changes (section II-A). Local features are described by the SIFT descriptor [9] (section II-B). Feature and description are computed on monochrome version of images; color information is not used in this work. A visual vocabulary is learned from these descriptors using the *k*-means algorithm (section III-A). The detected features will be indexed into two location databases: an inverted index (section III-C) and a location database (IV-A). All of the above is done offline.

When a mobile robot roams in this building, it obtains its location by retrieving in the inverted index. The coarse localization results are taken as candidates for fine localization. Each candidate in the location database is matched with the image for localization, and the correct location is the one getting the largest number of votes. In the case when the localization result is still ambiguous, epipolar geometry constraints are employed to verify the result (section V).

II. Scale Invariant Feature Detection and Description

Scale invariant features used in this work are detected by the Harris-Laplace detector and described by the SIFT descriptor.

A. Scale Invariant Feature Detection

The Harris-Laplace detector can detect scale invariant features [10]. The first step of this method is to compute

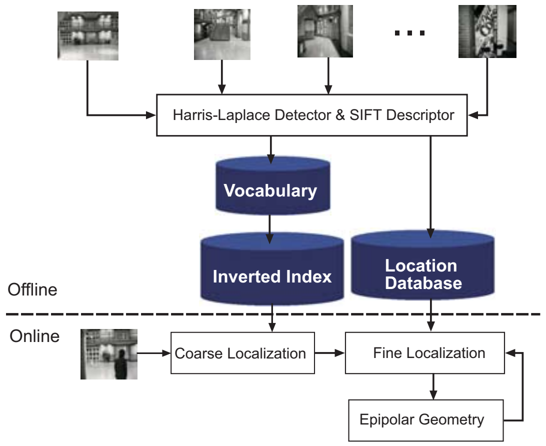

Fig. 1. Flowchart of our localization system

interest points (Harris points) at different scales. Then the points with a local maximal measure (the Laplacian) will be selected as Harris-Laplace interest points.

Harris interest points that are invariant to rotation changes can be detected reliably in images. Harris interest point detection is based on the Harris function [10] as given by:

$$det(C) - \alpha \cdot trace^2(C) > T_H \qquad (1)$$

where

$$C(\mathbf{x}, \sigma_I, \sigma_D) = \sigma_D^2 G(\mathbf{x}, \sigma_I) * \begin{pmatrix} L_x^2(\mathbf{x}, \sigma_D) & L_x L_y(\mathbf{x}, \sigma_D) \\ L_x L_y(\mathbf{x}, \sigma_D) & L_y^2(\mathbf{x}, \sigma_D) \end{pmatrix} \qquad (2)$$

σ_D is the derivation scale, σ_I the integration scale, G the Gaussian and L the image smoothed by a Gaussian.

In Eq.(1), α is the coefficient of the Harris function and T_H is the threshold of the Harris function. In our system, α and T_H are chosen empirically to ensure the stability of features. In this work, α is set to 0.06 and T_H is set to 1500.

The Harris function used here has been normalized for feature detection in different scales. The normalization makes the value of the Harris function comparable between different scales. If the Harris function is greater than the threshold (T_H), the point is defined as a Harris interest point.

Harris interest points can be detected at different scales. According to [8], local extrema over scales of normalized derivatives indicates the presence of the characteristic local structures. The characteristic scale can be found by searching for a local maximum over all scales. There are several derivative based functions (Laplacian, Difference-of-Gaussian and Harris function) that can compute a scale representation of a feature. The Laplacian is used in the Harris-Laplace interest point detection due to its high detection rate [10]. It is defined by

$$|\sigma^2(L_{xx}(\mathbf{x}, \sigma) + L_{yy}(\mathbf{x}, \sigma))| \qquad (3)$$

Harris-Laplace interest points can be detected by comparing the Laplacians at different scales. The scale of the

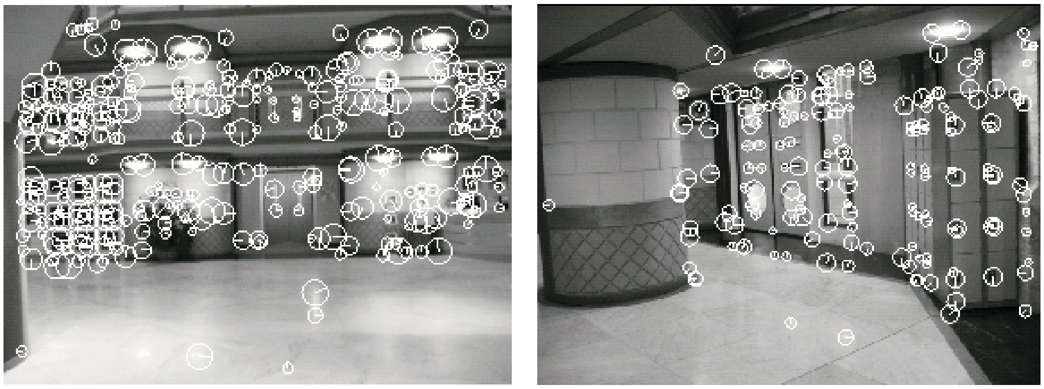

Fig. 2. Interest points detected by Harris-Laplace detector. The centers of the circles are Harris-Laplace interest points. The radiuses of the circles indicate the characteristic scale of interest points.

point with maximum Laplacian is taken as the characteristic scale of this interest point.

The accuracy of detected interest point is at pixel level, which is not good enough for pose recovery. Parabola interpolation is used in this work to get the precise locations of interest points to sub-pixel level.

B. Feature Description

The output of the Harris-Laplace detector is scale invariant features of different sizes (Fig. 2). These features need to be described for indexing. Many techniques have been proposed to describe interest points. Inspired by response properties of complex neurons in visual cortex, Lowe proposed the SIFT (scale invariant transformation feature) descriptor [9]. In such a description, multiple orientation planes represent a number of gradient orientations. Each orientation plane contains only the gradients corresponding to one orientation. This description proved to be robust against feature location (coordinates in image) errors and small geometric distortions [11]. In this paper, the SIFT descriptor is used to represent interest points.

The orientation of an interest point is computed by finding the maximum gradient direction in the neighborhood of the point. This ensures that the description is rotation invariant. The point region is normalized with the mean and the standard deviation of gradients within the point neighborhood, and thus this description is also robust against illumination changes. The SIFT is sampled over a 4×4 grid in the neighborhood of an interest point. The descriptor we get is of dimension 128.

III. Location Retrieval from Inverted Index

The representative images of the locations are indexed into the inverted index. Local features detected in these images are described by the SIFT descriptor that is an 128-dimension vector. The visual vocabulary is learned from these features by using the *k*-means algorithm. Based on this visual vocabulary, these descriptors are weighted and indexed into the inverted index.

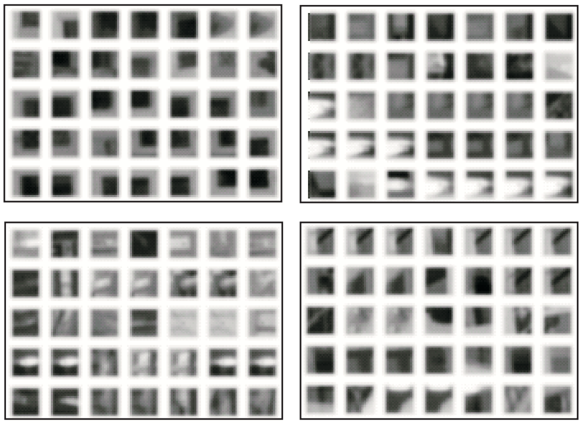

Fig. 3. Four sample terms of the visual vocabulary. These features have different orientations and have not been normalized.

A. Visual Vocabulary Constructing

Construction of a visual vocabulary is realized by clustering similar SIFT descriptors into terms that can be used for indexing. The *k*-means algorithm aims to group similar data objects into clusters. A cluster is a collection of data objects that are similar to one another within the same cluster and dissimilar from objects in different clusters. In the vocabulary computed by the *k*-means algorithm, each term represents a collection of descriptors that have similar appearance.

Lloyd algorithm is a popular *k*-means algorithm. Given the number of cluster, it is implemented in 4 steps: (a) Randomly assign *k* centers as seed points in the data objects; (b) Partition the data objects into *k* subsets by assigning each data object into the cluster with the nearest seed point; (c) Compute the centroids(center of the cluster) of the clusters of the current partition; (d) Take the computed centroids as the new seed points and go back to step (b), stop when no more new assignment occurs. However, Lloyd's algorithm may get stuck in locally minimal solutions that are far from the optimal [5]. For this reason it is necessary to consider heuristics based on local search, in which centers are swapped in and out of an existing solution. A hybrid algorithm which combines these two approaches (Lloyd's algorithm and local search) is used to learn the visual vocabulary [5].

More than 10,000 interest points are detected on the representative images and described by the SIFT descriptor. These vectors are input into the *k*-means algorithm as data objects. *k*-means algorithm learned a visual vocabulary that is composed of 320 terms.

Fig. 3 shows samples of terms learned from SIFT features. Descriptors with similar appearance are clustered into one term.

B. Inverted Index Building

The inverted index in this work employs the VSM in which the representative image of each location is expressed as a vector $\mathbf{d}_j$:

$$\mathbf{d}_j = (w_{1,j}, w_{2,j}, \cdots, w_{n_t,j}) \tag{4}$$

The components of each vector include all the possible terms (t_1–t_{n_t}) in the visual vocabulary. Each index term has an associated weight $w_{t,j}$ that indicates the importance of the index term for the identification. There are several methods available to compute the values of the weights $w_{i,j}$. This work adopts the method combining two factors: the importance of each index term in the representative view of location and the importance of the index term in the whole collection of locations.

$$w_{i,j} = tf_i \times idf_i \tag{5}$$

Importance of the index term in the representative view of location is denoted as *term frequency* (*tf*). It can be measured by the number of times that the term appears in the location.

$$if_i = \frac{n_{id}}{n_d} \tag{6}$$

where n_{id} is the number of occurrence of term *i* in the location *j*, n_d is the total number of terms in the location *j*.

The importance of the index term in the collection is denoted as *inverse document frequency* (*idf*). An index term that appears in every location in the collection is not very useful. However, a term that occurs only in a few locations may indicate that these few locations could be relevant to a query view that uses this term. In other words, the importance of an index term in the collection is quantified by the inverse of the frequency that this term appears in the locations in the index. it is computed by

$$idf_i = log(\frac{N}{n_i}) \tag{7}$$

where *N* is the number of locations in the index and n_i is the number of locations that contain the term *i*.

C. Indexing of SIFT Orientation

A consistent orientation is assigned to each SIFT descriptor based on local properties of the interest region. The descriptor is represented relative to this orientation and therefore achieves invariance to image rotation. Orientation is very helpful information in matching images. The descriptors are not directly indexed into the inverted index using above algorithm. We propose a method that makes orientation information usable in the first stage of localization.

The SIFT descriptors to be indexed are projected onto four directions. Indexing weights $w_{t,j}$ are accumulated in four bins: $w_{t,j,0}$, $w_{t,j,\frac{\pi}{2}}$, $w_{t,j,\pi}$, and $w_{t,j,3\frac{\pi}{2}}$. The representing vector $\mathbf{d}_j$ is expanded to

$$\mathbf{d}_j = (\mathbf{d}_{j,0}, \mathbf{d}_{j,\frac{\pi}{2}}, \mathbf{d}_{j,\pi}, \mathbf{d}_{j,\frac{3\pi}{2}}) \tag{8}$$

Using orientation information of the descriptor increases the correct ratio of location retrieval from inverted index. The benefit of using orientation information is shown in section VI. In addition, this method is robust to in plane rotation, which is also shown in section VI.

D. Coarse Localization

In the coarse localization stage, the VSM evaluates the degree of similarity of representative view with regard to the query view as the correlation between the two vectors $\mathbf{d}_j$ and $\mathbf{q}$. The query view is also a vector:

$$\mathbf{q} = (w_{1,q}, w_{2,j}, \cdots, w_{t,q}) \tag{9}$$

The VSM assumes that the similarity value is an indication of the relevance of the location to the given query. Thus the VSM ranks the retrieved locations by the similarity value. In this work, the cosine of the angle between the two vectors is employed to measure the similarity between the query view and representative view *j* in the inverted index:

$$s_j = \frac{\mathbf{d}_j^T \mathbf{q}}{||\mathbf{d}_j|| ||\mathbf{q}||} \tag{10}$$

To compromise the accuracy and efficiency of the localization system, the locations whose similarities rank top five will be taken as the input of next stage.

Using the inverted index increases the efficiency of localization. Details will be shown in section VI.

IV. Fine Localization

A robot captures a set of images when exploring a building at the first time. Scale-invariant interest points detected in these images are indexed into the location database. This database is the foundation of fine localization.

A. Database Building

The location database *M* contains a set of locations *L*. Each location can be defined by a set of vectors *V* of scale invariant interest point description. Each vector contains the coordinates (*u*,*v*), orientation α and value of the SIFT descriptor $SIFT_{128}$.

$$M = \{L^i | i = 1, 2, 3...m\} \tag{11}$$

$$L^i = \{V_j^i | j = 1, 2, 3...n\} \tag{12}$$

$$V_j^i = (u, v, \alpha, SIFT_{128})_j^i \tag{13}$$

During the database building process, each vector is added into the database with a link to the location where the corresponding representative image is captured.

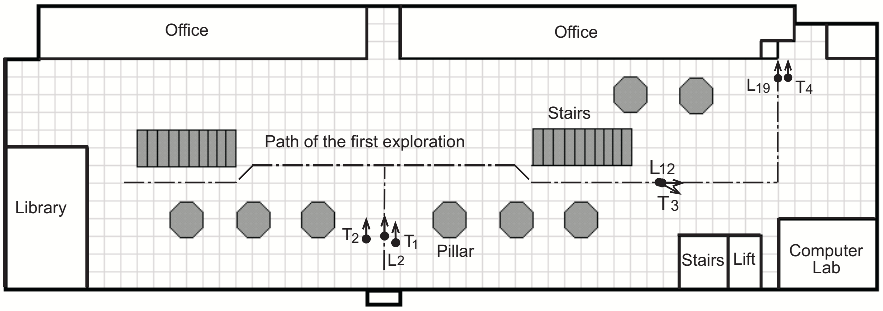

Fig. 4. Layout of the ground floor. L_2, L_{12}, and L_{19} are locations in database (Other locations are not shown in this figure). T_1, T_2, and T_3 are sites where images for localization are taken.

B. Fine Localization

Localization at this stage is carried out based on the results of coarse localization. The top five candidates computed by coarse localization are considered for location recognition.

Fine localization is realized by using a voting scheme. The new view for localization is represented by L_q.

$$L^q = \{V_j^q | j = 1, 2, 3...n\} \tag{14}$$

The Euclidean distances between a SIFT descriptor in L^q and those in an L^i are computed. The nearest neighbor of this descriptor in L^i is found by comparing all the Euclidean distances with high discrimination capability. A SIFT descriptor whose nearest neighbor is at least 0.7 times closer than the second nearest neighbor are considered as a possible vote. The votes for each location in the database are accumulated. The location that gets the largest number of votes is the most likely location.

V. Verification

In most cases, localization system gets correct location after the above two-stage localization. Nevertheless, it is possible that the result of the location recognition is ambiguous. There might be two or even three locations getting almost the same number of votes. In Fig. 7, the first and the second location get 21 votes, the third location gets 18 votes. Under this circumstance, it is difficult to decide which location is the correct one.

It is well known that the relationship between images taken at different viewpoints is determined by epipolar geometry. Therefore, epipolar geometry constraints are used here to verify the voting result. Fundamental matrix is estimated by using RANSAC algorithm [20]. This algorithm is robust to outliers. If a vote (correspondence between the interest point in image captured and features in database) is accepted by using the fundamental matrix, it is an inlier. Otherwise it is an outlier. The inliers of each location are kept and the outliers are discarded. Only the inliers are counted. The location that has the largest number of inliers is the correct location. In Fig. 7, localization system can now decide that the third location is the correct one because it has 16 inliers.

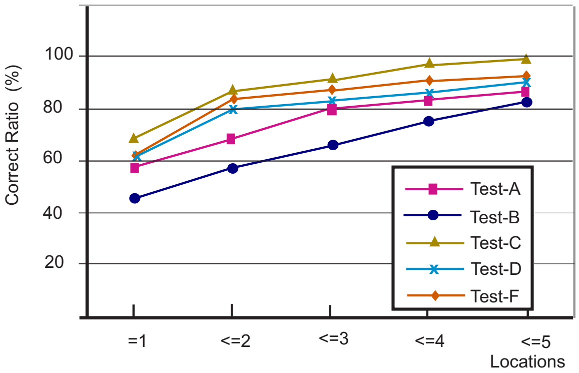

Fig. 5. The correct ratio of coarse localization. *y* (Vertical axis) is the correct ratio. The correct location is ranked among the first *x* (Horizontal axis) of retrieved locations. Test-C and Test-D have better performance than Test-A and Test-B. This is due to the employment of orientation information.

In this work, the verification will be carried out only under the condition that the votes that the second possible location gets are more than 80% of those that the first possible location gets.

VI. Experiments

The global localization strategy described above has been implemented and tested in an indoor environments. The environment model is obtained in the first exploration stage. These images were captured, using a camera, at different locations in the ground floor of a building. Fig. 4 is the sketch of the ground floor. Most images are taken at an interval of 2 meters. The visual vocabulary is learned from the SIFT descriptors of Harris-Laplace interest points. The first database contains 34 representative images.

Three image sequences are captured for testing of our approach. The first one (Sequence-I) is captured roughly along the path of the first exploration by a camcorder. The second one (Sequence-II) is captured in a path that deviates from the one of exploration (about 0.5 meter from the first exploration path). The third image sequence is captured with different viewpoints or illumination conditions. (Sequence-III).

Four experiments are carried out base on Sequence-I and Sequence-II. First, representative images are indexed

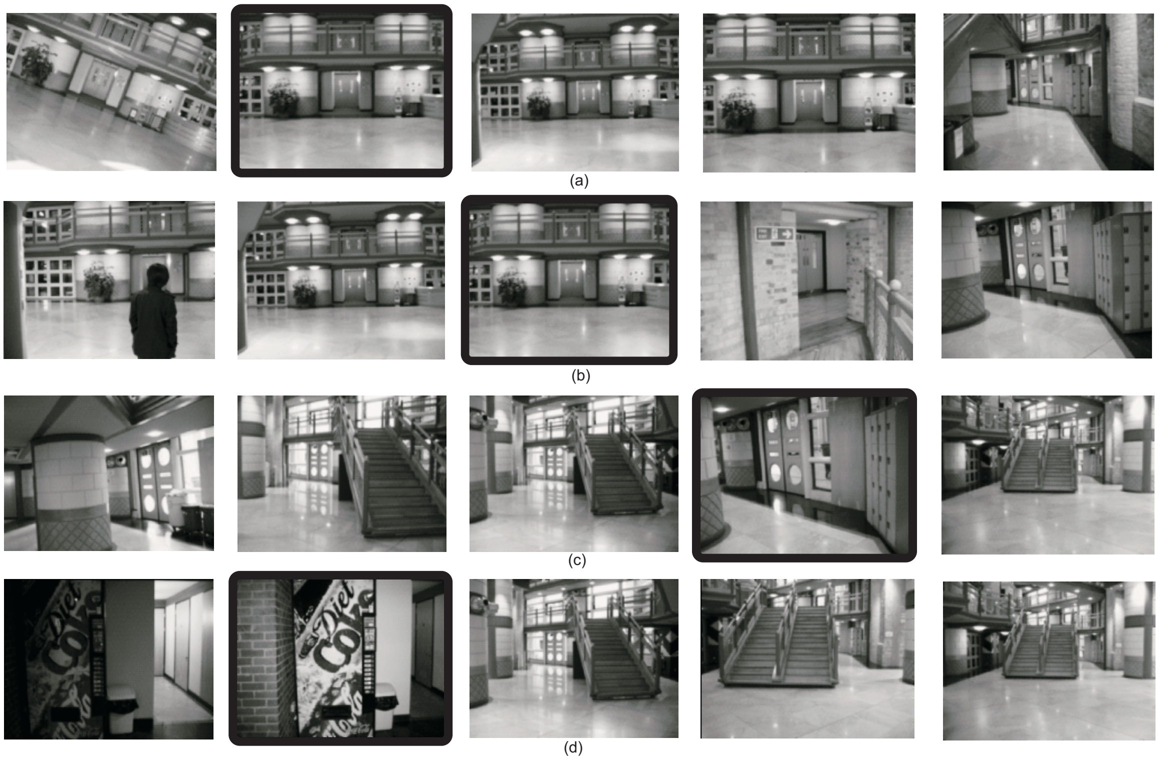

Fig. 6. Localization results. In each row, first one is the image for localization, others are coarse localization results with descending order of matches. The correct locations (denoted by black frames) are found after fine localization. (a) The image with in-plane rotation is ranked at the first in the coarse localization; (b) The image with translation is ranked at the second in the coarse localization; (c) The image with rotation is ranked at the third in the coarse localization; (d) The image with illumination change is ranked at the first in the coarse localization.

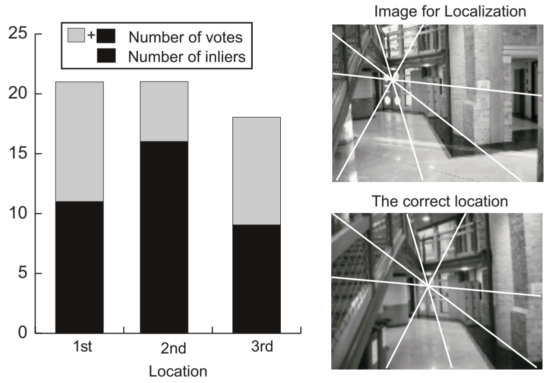

Fig. 7. Result of localization and verification. Only three locations with the largest number of votes are displayed. The third Location is correctly found after using epipolar geometry constraints. Epipolar lines are drawn on these images.

into an inverted index and a database without orientation information. Using this index, Test-A tests Sequence-I, and Test-B tests Sequence-II. Then orientation information is indexed into another database and another inverted index. Using this index, Test-C tests Sequence-I, Test-D tests Sequence-II. The correct ratios of the coarse localization are shown in Fig. 5. It is clear that employment of orientation information increases the correct ratio. The employment of orientation information does not have much effect on in-plane rotation. In Fig 6(a), there are 30 degrees in-plane rotation, and the location is correctly found and ranked the first during the coarse localization stage.

Test-E tests Sequence-III. The localization result shows that our system is robust to viewpoint and illumination changes. We get correct location from the database when the image for localization (T_2 in Fig. 4) is taken at one meters away from the location (L_2 in Fig. 4) in the database (Fig. 6(b)). An image taken at a different viewpoint was correctly retrieved from the database. Localization was accurate when the pan angle between the image in the database (taken at L_{12} Fig. 4) and the captured image (taken at T_3 in Fig. 4) is 20 degrees (Fig. 6(c)). An image taken under very bad illumination condition (taken at T_4 in Fig. 4) was also correctly found in the database (Fig. 6(d)). It demonstrates the advantage of using scale invariant features.

Scalability To test the scalability of our method, the locations are extended to 127. More locations(93) were explored and representative images were captured in the first and the second floor of the same building. These locations were indexed into an inverted index and a location database using the same visual vocabulary. The Test-F uses test Sequence-I and Sequence-II based on the inverted index and the location database that contains 127 locations . The result is shown in Fig. 5.

Computation Time The time for localization is shown in Table I. All of the experiments are carried out on an 1.4GHz laptop. To compare the computation time using our

TABLE I

COMPARISON OF AVERAGE TIMES USED IN LOCALIZATION PROCESS (SECONDS)

	Test-A&B	Test-C&D	Test-E	Direct-A	Direct-B
Coarse	0.11	0.13	0.14		
Fine	0.12	0.12	0.12	1.06	2.58
Total	0.23	0.25	0.26	1.06	2.58

approach with the one that directly uses fine localization, two more tests directly using fine localization method were carried out: Direct-A retrieves location from the database that contains 34 locations and Direct-B gets location from the database that contains 128 locations. It is clear that our approach is more efficient than the method that directly uses fine localization [22].

Computation time of Test-F (127 locations) is almost the same as the time for Test-C and Test-D (34 locations). This is due to the fact that most of the time is spent on matching SIFT features to visual terms.

VII. CONCLUSIONS AND FUTURE WORK

We have discussed a global localization approach based on a visual vocabulary. Object recognition and text retrieval techniques are successfully employed in this work. The coarse-to-fine strategy leads to fast and reliable global localization. The employment of orientation information increases the correct ratio of coarse localization. Our approach is robust against illumination and viewpoint changes.

Our work is a possible solution to deal with initialization and recovery from kidnap problems of SLAM system. We will integrate this approach into a SLAM system.

REFERENCES

[1] A. J. Davidson and D. Murray, ”Simultaneous localization and map building using active vision”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(7) pp. 865-880, 2002.

[2] A. J. Davison, ”Real-time simultaneous localization and mapping with a single camera”, in *Proc. of IEEE Int’l. Conf. on Computer Vision and Pattern Recognition*, pp. 1403-1410, 2003.

[3] I. Gordon and D. G. Lowe, ”Scene modelling recognition and tracking with invariant image features” in *Proc. of Int’l. Symposium on Mixed and Augmented Reality*, 2004.

[4] B. Johansson and R. Cipolla, ”A system for automatic pose-estimation from a single image in a city scene”, in *IASTED Int’l. Conf. Signal Processing, Pattern Recognition and Applications*, 2002.

[5] T. Kanungo, D.M. Mount, N. Netanyahu, C.D. Piatko, R. Silverman, and A. Y. Wu. ”An efficient *k*-means clustering algorithm: analysis and implementation”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(7), pp. 881-892, 2002.

[6] J. Košecká, L. Zhou, P. Barber, Z. Duric, ”Qualitative image based localization in indoors environments”, in *Proc. of IEEE Int’l. Conf. on Computer Vision and Pattern Recognition*, pp. II-3 - II-8, 2003.

[7] J. Košecká and Fayin Li. ”Vision based topological Markov localization”, in *Proc. of IEEE Int’l. Conf. on Robotics and Automation*, pp. 1481-1486, 2004.

[8] T. Lindeberg, ”Feature detection with automatic scale selection”, *Int’l. Journal of Computer Vision*, 30(2): pp. 79-116, 1998.

[9] D. G. Lowe, ”Object recognition from local scale-invariant features”, in *Proc. of Int’l. Conf. on Computer Vision*, pp. 1150-1157, 1999.

[10] K. Mikoljczyk and C. Schmid, ”Indexing based on scale-invariant features”, in *Proc. of Int’l. Conf. on Computer Vision*, pp. 525-531. 2001.

[11] K. Mikoljczyk and C. Schmid. ”A performance evaluation of local descriptors”, in *Proc. of IEEE Int’l. Conf. on Computer Vision and Pattern Recognition*, pp. 1403-1410, 2003.

[12] T. Salton, B. Gerard and A. Buckley, F. Chris. ”Term Weighting Approaches in Automatic Text Retrieval”, *Information Processing and Management* Vol.32 (4), pp. 431-443, 1996.

[13] S. Se, D. Lowe and J. Little, ”Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks”, *Int’l. Journal of Robotics Research*, Vol. 21, No. 8, August 2002, pp. 735-758.

[14] C. Shmid and R. Mohr. ”Local greyvalue invariants for image retrieval”. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(5): pp. 530-534, 1997.

[15] R. Sims and G. Dudek. ”Learning environmental features for pose estimation”, *Image and Vision Computing*, 19(2001) pp.733-739.

[16] J. Sivic and A. Zisserman. ”Video Google: a text retrieval approach to object matching in videos”, in *Proc. of Int’l. Conf. on Computer Vision*. pp. 1470-1477, 2003.

[17] T. Starner, B. Schiele, and A. Pentland, ”Visual contextual awareness in wearable computing”, in *Proc. of Int’l. Symposium on Wearable Computing*, pp. 50-57, 1998.

[18] D.M. Squire, W. Müller, H. Müller, and T. Pun. ”Content-based query of image databases: inspirations from text retrieval”, *Pattern Recognition Letters*, 21:1193-1198, 2000.

[19] S. Thrun, D. Fox, W. Burgard, and F. Dellaert. ”Robust Monte Carlo localization for mobile robots”, *Artificial Intelligence*, Vol:128, Issue:1-2, pp. 99-141, 2001.

[20] P. H. S. Torr and A Zisserman, ”Mlesac: A new robust estimator with application to estimating image geometry”, *Computer Vision and Image Understanding*, 78(1):138-156, 2000.

[21] A. Torralba, K. P. Murphy, W. T. Freeman and M. Rubin, ”Context-based vision system for place and object recognition”, in *Proc. of Int’l. Conf. on Computer Vision*, pp. 273-280, 2003

[22] J. Wang, R. Cipolla, and H. Zha. ”Image-based localization and pose recovery using scale invariant features”, in *Proc. of IEEE Int’l. Conf. on Robotics and Biomimetics*, 2004.

[23] I. H. Witten, A. Moffat, and T. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishers, 1999.

[24] J. Wolf, W. Burgard, and H. Burkhardt, ”Robust vision-based localization for mobile robots using an image retrieval system based on invariant features”, in *Proc. of IEEE Int’l. Conf. on Robotics and Automation*, pp. 359-365, 2002.