Coarse-to-Fine Vision-Based Localization by Indexing Scale-Invariant Features

Junqiu Wang, *Student Member, IEEE*, Hongbin Zha, and Roberto Cipolla, *Member, IEEE*

***Abstract*—This paper presents a novel coarse-to-fine global localization approach inspired by object recognition and text retrieval techniques. Harris–Laplace interest points characterized by scale-invariant transformation feature descriptors are used as natural landmarks. They are indexed into two databases: a location vector space model (LVSM) and a location database. The localization process consists of two stages: coarse localization and fine localization. Coarse localization from the LVSM is fast, but not accurate enough, whereas localization from the location database using a voting algorithm is relatively slow, but more accurate. The integration of coarse and fine stages makes fast and reliable localization possible. If necessary, the localization result can be verified by epipolar geometry between the representative view in the database and the view to be localized. In addition, the localization system recovers the position of the camera by essential matrix decomposition. The localization system has been tested in indoor and outdoor environments. The results show that our approach is efficient and reliable.**

***Index Terms*—Coarse-to-fine localization, scale-invariant features, vector space model, visual vocabulary.**

I. Introduction

MOBILE robot localization aims to estimate a robot's position relative to its environment. It is a prerequisite for robot autonomous navigation. Two key problems of mobile robot localization are global localization and local tracking [25]. 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 position, 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" [20]. Vision-based global localization using natural landmarks is highly desirable for a wide range of applications. Different from other sensors such as sonar sensors and range finders, visual sensors do not suffer from the reflection problem. Moreover, other tasks such as object recognition can be integrated into one vision system using context information [23], [27].

The difficulty with vision-based localization is how to determine the identity of views of an environment in the presence of viewpoint changes, different illumination and occlusion. Visual features invariant to viewpoints and illumination changes are critical for an effective localization system using visual landmarks. In recent years, great progress has been made in the use of invariant features for object recognition and image matching. Schmid and Mohr propose a rotation-invariant feature detector to solve general image recognition problems [19]. Mikolajczyk and Schmid extend this idea to the Harris–Laplace detector which can detect scale-invariant features [12]. Lowe proposes another scale-invariant feature detector which finds local scale space maxima of Difference-of-Gaussian (DoG) [10]. Perspective or affine invariance of the visual feature set is ideal for a localization system. However, the computation of affine-invariant features usually is very expensive [14]. The detection of perspective-invariant features is difficult especially in a cluttered scene. Recently, a projective-invariant feature detector has been proposed based on the cross-ratio invariance [16]. The feature extraction consists of detection of straight line using the computationally expensive Hough transform and construction of a cross-ratio histogram. This detector does not provide good performance especially in a cluttered scene. The Harris–Laplace feature detector is selected in this work for its efficiency and flexibility [14]. Feature descriptors are also important for image matching. We use the scale-invariant transformation feature (SIFT) descriptor proposed by Lowe [10].

The matching of one image to many is slow especially when a lot of images are used to represent a large environment. 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 file* [29]. In this file, each document is a vector and each dimension of the vector represents a count of the occurrence for a term [17], [29]. The documents for retrieval are parsed into terms 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 to realize these ideas in our localization system. A *location vector space model* (LVSM) is built using this visual vocabulary. Localization based on the LVSM is fast, but not accurate enough; whereas localization from the location database is relatively slow, but more accurate. We propose a coarse-to-fine framework to enhance the efficiency and accuracy of localization. The localization process includes two stages: coarse localization and fine localization. The coarse localization results ranked in the top list are taken as candidates for the fine localization stage. The integration of

Manuscript received December 14, 2004; revised May 25, 2005. This work was supported in part by the National Science Foundation of China under Grant 6017525, Grant 60333010, and by the National Key Basic Research Project of China 2004CB318005. This paper was recommended by Associate Editor P. Bhattacharya.

J. Wang and H. Zha are with the National Laboratory on Machine Perception, Peking University, Beijing, 100871, China (e-mail: jerywang@public3.bta.net.cn; zha@cis.pku.edu.cn).

R. Cipolla is with the Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, U.K. (e-mail: cipolla@eng.cam.ac.uk).

Digital Object Identifier 10.1109/TSMCB.2005.859085

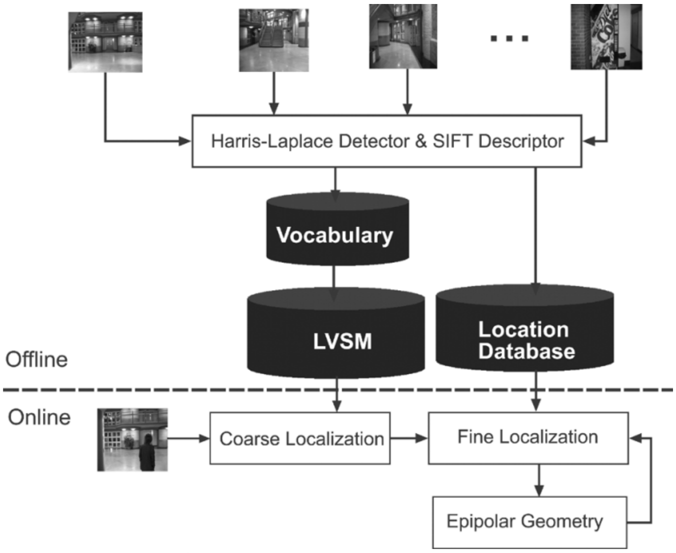

Fig. 1. Flowchart of our localization system.

coarse and fine stages realizes a fast and reliable localization system.

A very important issue in a vision-based localization system is how to represent the world model [4], [24]. Metric and topological models have been used widely for environment representation. Metric approaches represent an environment by evenly spaced grids whereas topological methods describe it by graphs. Compared to a topological map, two-dimensional (2-D) or even three-dimensional (3-D) metric maps (e.g. [2]) have a few disadvantages [24]. First, the detection and recognition is computationally expensive and memory consuming. Secondly, path planning is not convenient especially in large environments. Finally, the major problem is the cumulative error built-up. Topological representations are more robust. The construction of such a topological map is easy because of the employment of scale-invariant features. The topological approach is adopted here to describe environments.

A. Overview

The main contribution of this work is the coarse-to-fine localization framework that leads to a reliable and efficient system. Other contributions include the LVSM, the term selection method for the visual vocabulary, the indexing of feature orientation information, and the verification of localization results.

In Section II, related work will be addressed. The coarse-to-fine localization system will be introduced (Fig. 1) in the following sections.

In the first exploration, representative images are captured. Scale-invariant interest points are detected by the Harris–Laplace detector [12]. The Harris–Laplace detector is built in a multiscale framework, which makes these interest points robust to scale changes (Section III-A). Local features are described by the SIFT descriptor (Section III-B). Feature and description are computed on the 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 IV-A). The detected features will be indexed into two databases: an LVSM (Section IV-C) and a location database (Section V-A). All of the above are done offline.

When a mobile robot roams in the environment, it obtains its location by retrieval from the LVSM. The coarse localization results are taken as the candidates for the following 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 where the localization result is still ambiguous, epipolar geometry constraints are employed to verify the result (Section VI). The epipolor geometry is used to recover the robot's accurate position relative to the location in the database (Section VII).

II. Related Work

Vision-based localization is one of the active research areas in robotics. It is not possible to cover the history in this paper. For a complete review, refer to the excellent work by DeSouza and Kak [3].

Many previous works use local visual features such as vertical lines (indoor) or road boundaries (outdoor) for controlling robot motion [15]. Such features are simple and can be easily detected. However, these features are not always available. Furthermore, localization based on these features usually can only deal with local tracking.

Localization by using object recognition techniques is promising because it uses natural visual features. Se *et al.* use scale-invariant visual marks to deal with mobile robot localization based on the local feature detector and the SIFT descriptor proposed by Lowe [18]. They use Triclops, a vision system that has three cameras. Their system can deal with small areas such as a room. Our approach can work in a much larger environment. In addition, only one camera is used in this work for localization. Wang, Cipolla and Zha have proposed a localization strategy based on the Harris–Laplace interest point detector and the SIFT descriptor [28]. 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. Kosecka and Yang also characterize scale-invariant key points by the SIFT descriptor in their localization system [9]. These localization systems have to match a new view to the views in the database by nearest neighbor search which is not efficient enough for robot localization. Katsura *et al.* developed an outdoor localization system based on segmentation of the images [8]. Their system can obtain the location by matching areas of trees, sky, and buildings. However, occlusions bring much trouble to the matching. Moreover, it cannot recover the relative position of the robot.

In this work, VSM and other techniques from the text retrieval literature are used to accelerate the localization process. Although term weighting and inverted file have been used in image [22] and video retrieval [21], none of their systems can be extended to a localization system because they use different feature detectors, which are slow and not suitable for localization. To the best of our knowledge, the term selection method proposed in this work has not been used in image, video retrieval and localization.

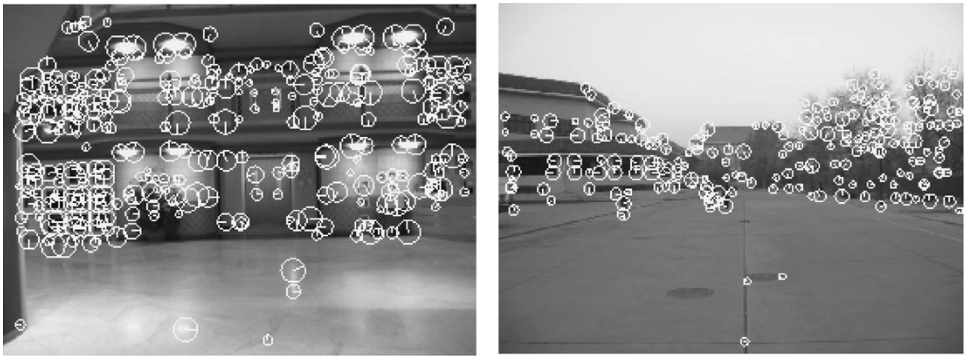

Fig. 2. Interest points detected by the Harris–Laplace detector. Centers of the circles are the Harris–Laplace interest points. The radii of the circles indicate the characteristic scale of the interest points.

III. 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. It detects Harris interest points at several scales and then selects the right scale by finding the maximum of the Laplacian function [12]. In our implementation, Harris interest points are detected at four scales with the initial scale 1.2. Harris–Laplace interest points are detected based on scale selection. Accordingly, our detector can not deal with scale changes more than 2. However, our approach can detect reliable local features and the scale change is enough for a localization system. At the same time, the feature detection time is significantly reduced. The detection takes around 0.6 seconds in an image (640 × 480).

B. Feature Description

The output of the Harris–Laplace detector is scale-invariant features of different sizes (Fig. 2). A reliable feature description is critical for feature matching. The SIFT descriptor is selected here according to [13], in which many descriptors are evaluated and the SIFT is the best one with respect to the overall performance. In the SIFT descriptor, gradient direction histograms are built in the local area. Multiple orientation planes represent the number of gradient orientations. The SIFT is sampled over a 4 × 4 grid in the neighborhood of an interest point. The size of each grid is determined by the scale of the interest point. The descriptor we get is a 128-dimension vector.

IV. Location Retrieval From the LVSM

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

A. Visual Vocabulary Construction

Construction of a visual vocabulary is to build a “code book” for the indexing of local features. This is realized by clustering similar SIFT descriptors into terms that can be used for indexing. The k-means algorithm is used in this work to group similar data objects into clusters. The centroid of each cluster is taken as terms of the visual vocabulary.

The Lloyd algorithm is a simple implementation of the k-means. However, it may get stuck in locally minimal solutions that are far from optimal [7]. 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 here in the learning of the visual vocabulary [7].

According to our experiments, the input of the vocabulary learning algorithm should have enough variety. Otherwise the learning results tend to construct a vocabulary in which most terms only have one or two features. In our implementation, 11732 interest points are detected in 212 images captured in indoor and outdoor environments. The descriptors of these features are input for the k-means algorithm as data objects. The k-means algorithm is run several times with different values of k. The vocabulary with the best performance is used. We use the one in which k is set to 1024. The output of the k-means is 1024 centroids representing different terms.

The contribution of these terms to localization is different. Some terms are not very discriminative because they appear in almost every representative image; some terms which appear only once or twice have little impact on the retrieval performance. The other terms with appropriate frequency of appearance usually contribute more to the retrieval process. In a text corps, term frequency distributions tend to follow Zipf’s law [11]

$$c \approx f \times r \tag{1}$$

where f is the frequency of a term and r is its rank. For the terms providing a high contribution, the product of f and r is approximately a constant c. Similar distribution is observed in Fig. 3 where occurrence of each term is counted. The top 23 terms are put into the stop list because they appear too frequently; 681 terms appearing only once or twice are also discarded (Fig. 3). The remaining 320 terms constitute the visual vocabulary for the retrieval. We also use a vocabulary including all the 1024 terms to evaluate the effectiveness of the term selection. Experimental results show that the above selection method brings higher correct ratio in the coarse localization.

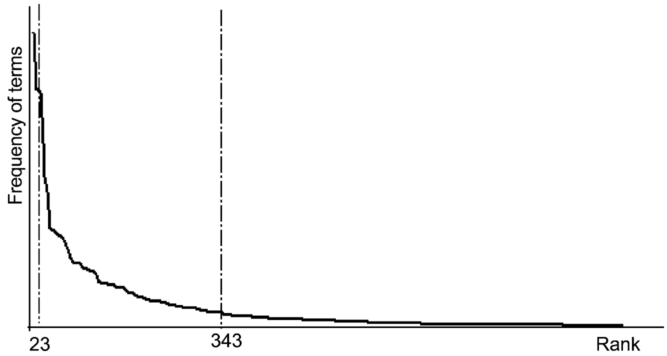

Fig. 3. Term selection using the Zipf's law.

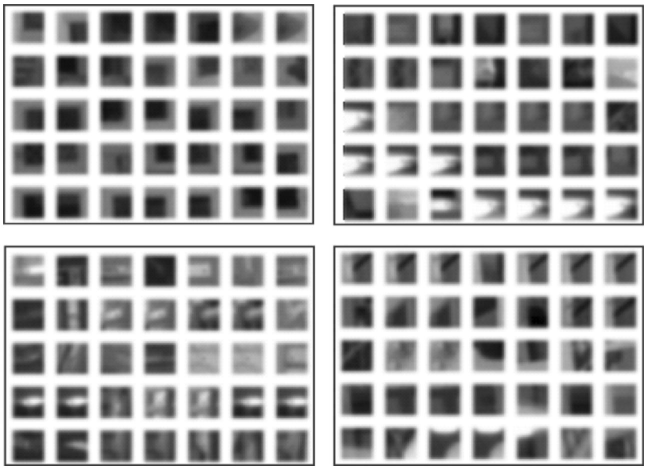

Fig. 4. Four sample terms of the visual vocabulary. These features have different orientations and have not been normalized.

Fig. 4 shows samples of the terms learned from SIFT features. Descriptors with similar appearance are clustered into one term.

B. LVSM Building

In the LVSM, the representative image of each location is expressed as a vector $\mathbf{d}_j$

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

The components of each vector include all the possible terms $\{t_1, t_2, \ldots, t_{n_t}\}$ 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

$$w_{i,j} = tf_i \times idf_i. \tag{3}$$

The 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 the term appears in the location

$$tf_i = \frac{n_{ij}}{n_j} \tag{4}$$

where n_{ij} is the number of occurrences of term i in the location j, and n_j is the total number of terms in the location j.

The importance of the index term in the location collection is denoted as *inverse document frequency (idf)*. An index term that appears in every location in the collection is not very discriminative, whereas 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 can be quantified by the logarithm of the inverse of the frequency with which this term appears in the locations in the LVSM. It is computed by

$$idf_i = \log\left(\frac{N}{N_i}\right) \tag{5}$$

where N is the number of locations in the LVSM 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 point. 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 LVSM using the 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 in four directions. Indexing weights $w_{t,j}$ are accumulated in four bins: $w_{t,j,0}$, $w_{t,j,(\pi/2)}$, $w_{t,j,\pi}$, and $w_{t,j,3(\pi/2)}$. The vector $\mathbf{d}_j$ is expanded to

$$\mathbf{o}_j = \left(\mathbf{d}_{j,0}, \mathbf{d}_{j,\frac{\pi}{2}}, \mathbf{d}_{j,\pi}, \mathbf{d}_{j,\frac{3\pi}{2}}\right). \tag{6}$$

Using the orientation information of the descriptor increases the correct ratio of location retrieval from LVSM. The benefit of using orientation information is shown in Section VIII. In addition, this method is robust for in-plane rotation, which is also shown in Section VIII.

D. Coarse Localization

In this stage, the degree of similarity of a representative view with regard to the query view is evaluated by computing the correlation between the two vectors $\mathbf{d}_j$ and $\mathbf{q}$ (or $\mathbf{o}_j$ and $\mathbf{r}$). The query view is also a vector

$$\mathbf{q} = (w_{1,q}, w_{2,q}, \cdots, w_{t,q}). \tag{7}$$

The tf of $\mathbf{q}$ is computed by using (4) and the idf of each term uses the same value of (5). The $\mathbf{q}$ is also extended to include orientation information

$$\mathbf{r} = \left(\mathbf{q}_0, \mathbf{q}_{\frac{\pi}{2}}, \mathbf{q}_\pi, \mathbf{q}_{\frac{3\pi}{2}}\right). \tag{8}$$

It is assumed that the similarity value is an indication of the relevance of the location to the given query. Thus, the system ranks the retrieved locations by the similarity value. In this work, the co-sine of the angle between the two vectors is employed to measure the similarity between the query view and the representative view j in the LVSM

$$s_j = \frac{\mathbf{o}_j \cdot \mathbf{r}}{|\mathbf{o}_j||\mathbf{r}|}. \tag{9}$$

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

Using the LVSM increases the efficiency of localization. The details of the performance will be shown in Section VIII.

V. Fine Localization

The coarse localization results are ranked and the top five locations are input of the fine localization stage.

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}:

$$M = \{L^i | i = 1, 2, 3 \ldots m\} \tag{10}$$

$$L^i = \left\{V_j^i | j = 1, 2, 3 \ldots n\right\} \tag{11}$$

$$V_j^i = (u, v, \alpha, \mathrm{SIFT}_{128})_j^i. \tag{12}$$

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.

B. Fine Localization

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

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

$$L^q = \left\{V_j^q | j = 1, 2, 3 \ldots n\right\}. \tag{13}$$

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.

VI. Verification

It is well known that the relationship between images taken at different viewpoints is determined by epipolar geometry. The epipolar geometry is the intrinsic projective geometry between two views. A fundamental matrix contains this intrinsic geometry. Fundamental matrix is estimated from correspondences of points by using the RANSAC algorithm that is robust to outliers. The error function in the estimation is a negative likelihood function [26].

Epipolar geometry is employed in this work to verify the localization result by discarding the outliers. In most cases, the localization system gets the 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. If a vote (correspondence between the interest point in image captured and features in the database) is accepted by using the fundamental matrix, it is a correct correspondence. Otherwise it is an outlier. The correct location can be found by discarding outliers.

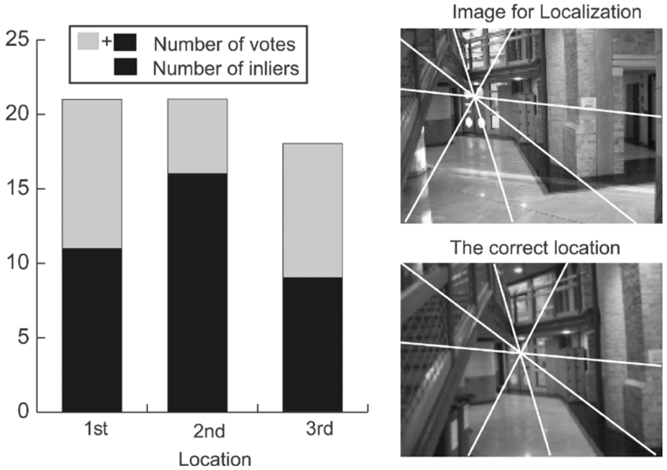

Fig. 5. Result of localization and verification. Only three locations with the largest number of votes are displayed. The third Location is correctly found after using epipolar geometry constraints. Epipolar lines are drawn on these images.

In Fig. 5, 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. Only the correct matches are counted base on the fundamental matrix. The location that has the largest number of correct correspondences is the correct location. In Fig. 5, the localization system can now decide that the second location is the correct one because it has 16 correct matches.

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.

VII. Relative Pose Estimation

The relative pose of the robot is recovered after the global localization stage. Interest points detected in the captured image are matched with the features in the database that represent this location. Many correspondences between interest points are found to compute relative pose with respect to the reference view by decomposing the essential matrix.

We can get two camera internal parameter matrix $\mathbf{K}_1$ and $\mathbf{K}_2$ using a camera calibration method, where $\mathbf{K}_1$ are the internal parameters of the camera used in the first explorations, and $\mathbf{K}_2$ are the internal parameters of the camera used in the exploration. The camera internal parameters do not change during the navigation phase.

Based on the fundamental matrix $\mathbf{F}$ and the camera internal parameters $\mathbf{K}_1$, $\mathbf{K}_2$, the essential matrix $\mathbf{E}$ is computed [1]

$$\mathbf{E} = \mathbf{K}_1^T \mathbf{F} \mathbf{K}_2. \tag{14}$$

The essential matrix can be decomposed into rotation $\mathbf{R}$ and translation $\mathbf{t}$ [1]

$$\mathbf{E} = [\mathbf{t}]_x \mathbf{R} \tag{15}$$

where $[\mathbf{t}]_x$ denotes the cross product matrix associated with the translation vector.

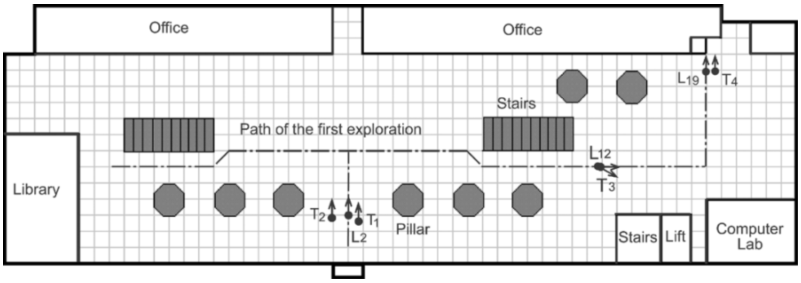

Fig. 6. Layout of the ground floor. L_2, L_{12}, and L_{19} are locations in database (Other locations are not shown in this figure). T_1, T_2, and T_3 are sites where the images for localization are taken.

The essential matrix has two equal singular values and one zero singular value [1], [5]. We can compute the rotation [(18) and (19)] and the translation [(21) and (22)] based on singular value decomposition of the essential matrix [5]:

$$\mathbf{E} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{16}$$

where

$$\mathbf{D} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{17}$$

$$\mathbf{R} = \mathbf{U}\mathbf{W}\mathbf{V}^T \tag{18}$$

or

$$\mathbf{R} = \mathbf{U}\mathbf{W}^T\mathbf{V}^T \tag{19}$$

where

$$\mathbf{W} = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{20}$$

$$\mathbf{t} = \mathbf{U}(0 \quad 0 \quad 1)^T \tag{21}$$

or

$$\mathbf{t} = -\mathbf{U}(0 \quad 0 \quad 1)^T. \tag{22}$$

The solution of the pose is one of following four possible matrices [5]:

$$\left(\mathbf{U}\mathbf{W}\mathbf{V}^T | \mathbf{U}(0 \quad 0 \quad 1)^T\right) \tag{23}$$

$$\left(\mathbf{U}\mathbf{W}\mathbf{V}^T | -\mathbf{U}(0 \quad 0 \quad 1)^T\right) \tag{24}$$

$$\left(\mathbf{U}\mathbf{W}^T\mathbf{V}^T | \mathbf{U}(0 \quad 0 \quad 1)^T\right) \tag{25}$$

$$\left(\mathbf{U}\mathbf{W}^T\mathbf{V}^T | -\mathbf{U}(0 \quad 0 \quad 1)^T\right). \tag{26}$$

The points in the images must lie in front of both cameras. The right solution can be computed by checking whether the point lies before the camera [5].

VIII. Experiments

The global localization strategy described above has been implemented and tested in indoor and outdoor environments. All of these tests are conducted on a 1.4-GHz laptop with 128-M memory.

The size of the images is 640 × 480. The cameras are calibrated and the internal parameters are known. Cameras with different internal parameters can be used in the exploration and localization stages. However, the internal parameters do not change during the localization stage.

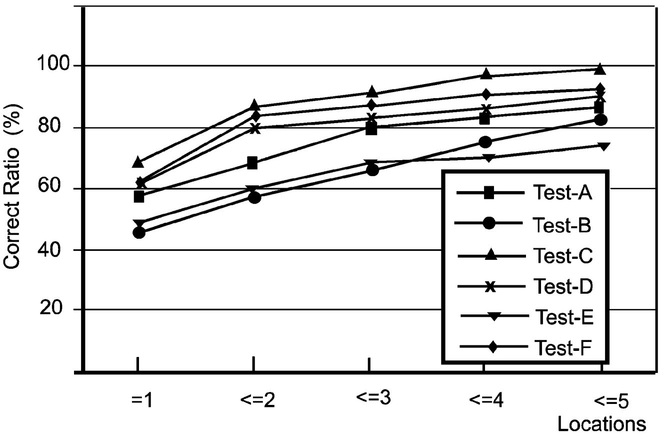

Fig. 7. Correct ratio of coarse localization. y (Vertical axis) is the correct ratio. The correct location is ranked among the first x (Horizontal axis) of retrieved locations. Test-C and Test-D have better performance than Test-A and Test-B. This is due to the employment of orientation information.

The localization result is taken as correct when the following conditions are met: 1) the correct location is retrieved in the first five results in the coarse localization stage; 2) the correct one is found in the fine localization stage; and 3) there are more than eight correct matches which make the recovery of relative pose possible.

A. Indoor Experiments

The indoor 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. 6 is a sketch of the ground floor. Most images are taken at an interval of 2 m. 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 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 m from the first exploration path). The third image sequence is captured with different viewpoints or under different illumination conditions. (Sequence-III).

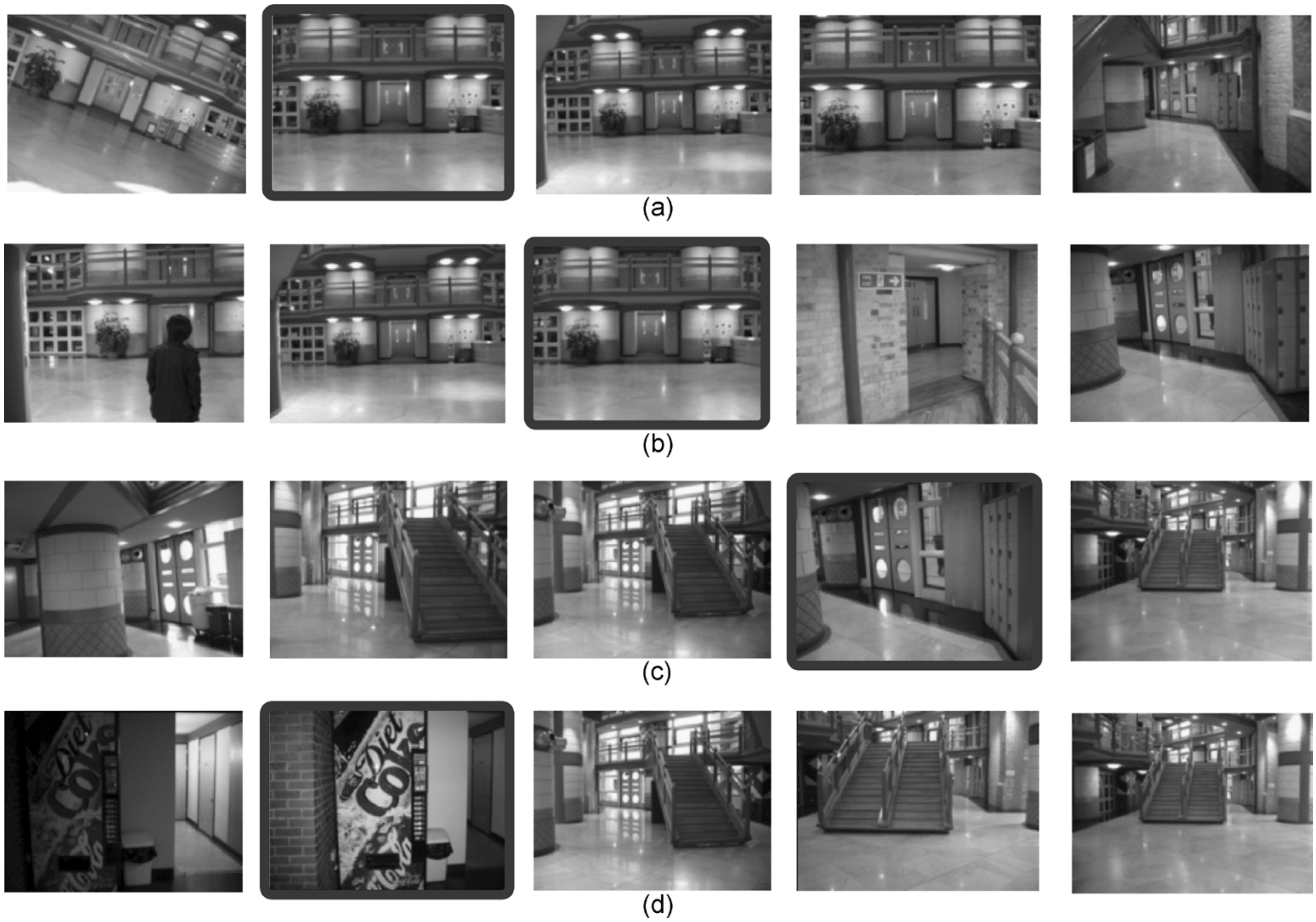

Fig. 8. Localization results. In each row, the first image is the image for localization, others are coarse localization results with descending order of matches. The correct locations (denoted by black frames) are found after fine localization. (a) Image with in-plane rotation is ranked at the first in the coarse localization. (b) Image with translation is ranked at the second in the coarse localization. (c) Image with rotation is ranked at the third in the coarse localization. (d) Image with illumination change is ranked at the first in the coarse localization.

TABLE I
COMPARISON OF AVERAGE TIMES USED IN LOCALIZATION PROCESS (SECONDS)

	Term	Coarse	Fine	Total
Test-A&B	0.09	0.009	0.12	0.23
Test-C, D, E	0.09	0.011	0.12	0.236
Test-F	0.09	0.013	0.12	0.245
Direct-A			1.06	1.06
Direct-B			2.58	2.58

Four experiments are carried out based on Sequence-I and Sequence-II. First, the representative images are indexed into a LVSM 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 the second database and the second LVSM. Using the later index, Test-C tests Sequence-I, Test-D tests Sequence-II. The correct ratios of the coarse localization are shown in Fig. 7. It is clear that the employment of orientation information increases the correct ratio. The employment of orientation information does not have much effect on in-plane rotation. In Fig. 8(a), there are 30 degrees in-plane rotation, and the location ranked the first during the coarse localization stage is correctly found.

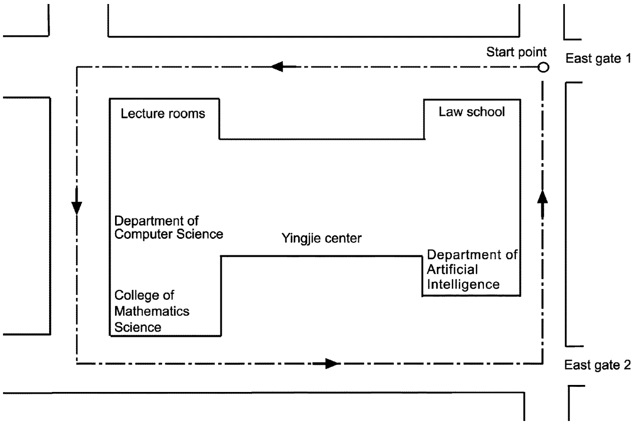
Start point
East gate 1
Lecture rooms
Law school
Department of Computer Science
Yingjie center
Department of Artificial Intelligence
College of Mathematics Science
East gate 2

Fig. 9. Layout of the outdoor environment in a campus.

Test-E tests Sequence-III. We get correct location from the database when the image for localization (T_2 in Fig. 6) is taken at one meter away from the location (L_2 in Fig. 6) in the database [Fig. 8(b)]. An image taken at a different viewpoint is correctly retrieved from the database. Localization is accurate when the pan angle between the image in the database (taken at L_{12} Fig. 6) and the captured image (taken at T_3 in Fig. 6) is 20 degrees [Fig. 8(c)]. An image taken under very bad illumination

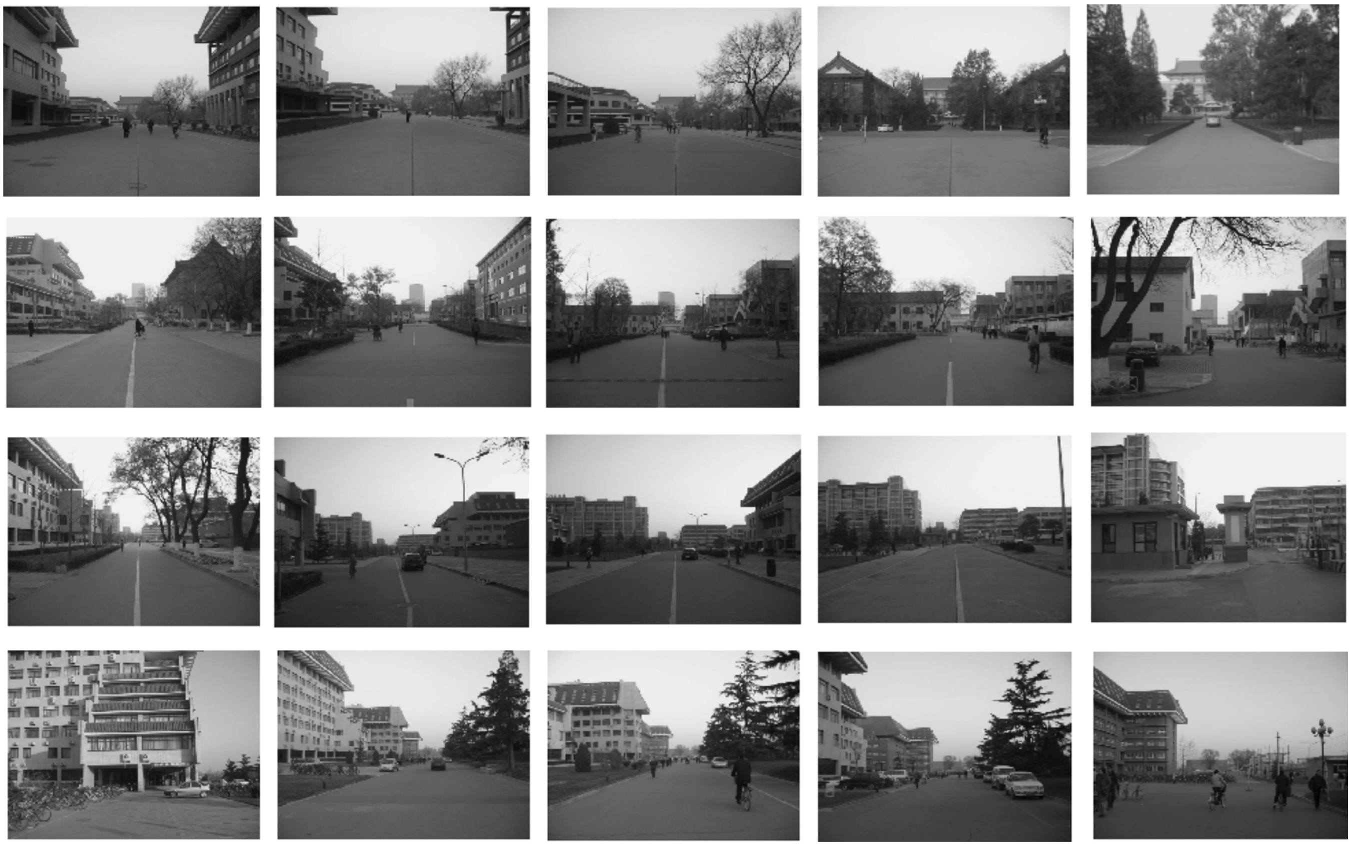

Fig. 10. Examples of the representative images used in the outdoor environment model building stage.

condition (taken at T_4 in Fig. 6) was also correctly found in the database [Fig. 8(d)]. The localization result shows that our system is robust against viewpoint and illumination changes. It demonstrates the advantage of using scale-invariant features.

Scalability: To test the scalability of our method, the number of locations is increased to 127 and 93 more locations were explored. Representative images were captured in the first and the second floor of the same building. These locations are indexed into a LVSM and a location database using the same visual vocabulary. Test-F uses Sequence-I and Sequence-II based on the LVSM and the location database that contains 127 locations. The result is shown in Fig. 7.

Computation Time: It takes 0.58 ± 0.14 seconds to detect and describe Harris–Laplace interest points in an image (640 × 480) by using the Harris–Laplace detector and the SIFT descriptor.

The computation determination of the location includes term assignment, coarse localization from the LVSM and fine localization from the database. The time for term assignment that is the most time consuming process depends on the number of features detected on the input image and the number of terms in the vocabulary. The time for coarse localization is linear to the number of locations. The time for fine localization depends on the number of features in the input image.

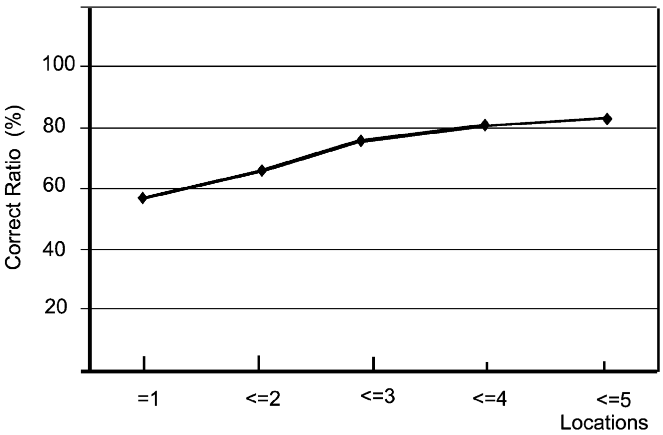

Fig. 11. Correct ratio of the coarse localization in the outdoor experiments. y (vertical axis) is the correct ratio. The correct location is ranked among the first x (horizontal axis) of the retrieved locations.

The time for localization is shown in Table I. To compare the computation time using our approach with the one that directly does fine localization, two more tests directly using the fine localization method are carried out: Direct-A retrieves the location from the database that contains 34 locations and Direct-B retrieves the location from the database that contains 128 locations. It is clear that our approach is more efficient than the method that directly uses the fine localization [28]. The fine localization time and the term assignment time do not change when the location number changes. Therefore, the advantage of our approach will become even more evident if the number of locations increases.

The 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 the SIFT features to the visual terms in the visual vocabulary.

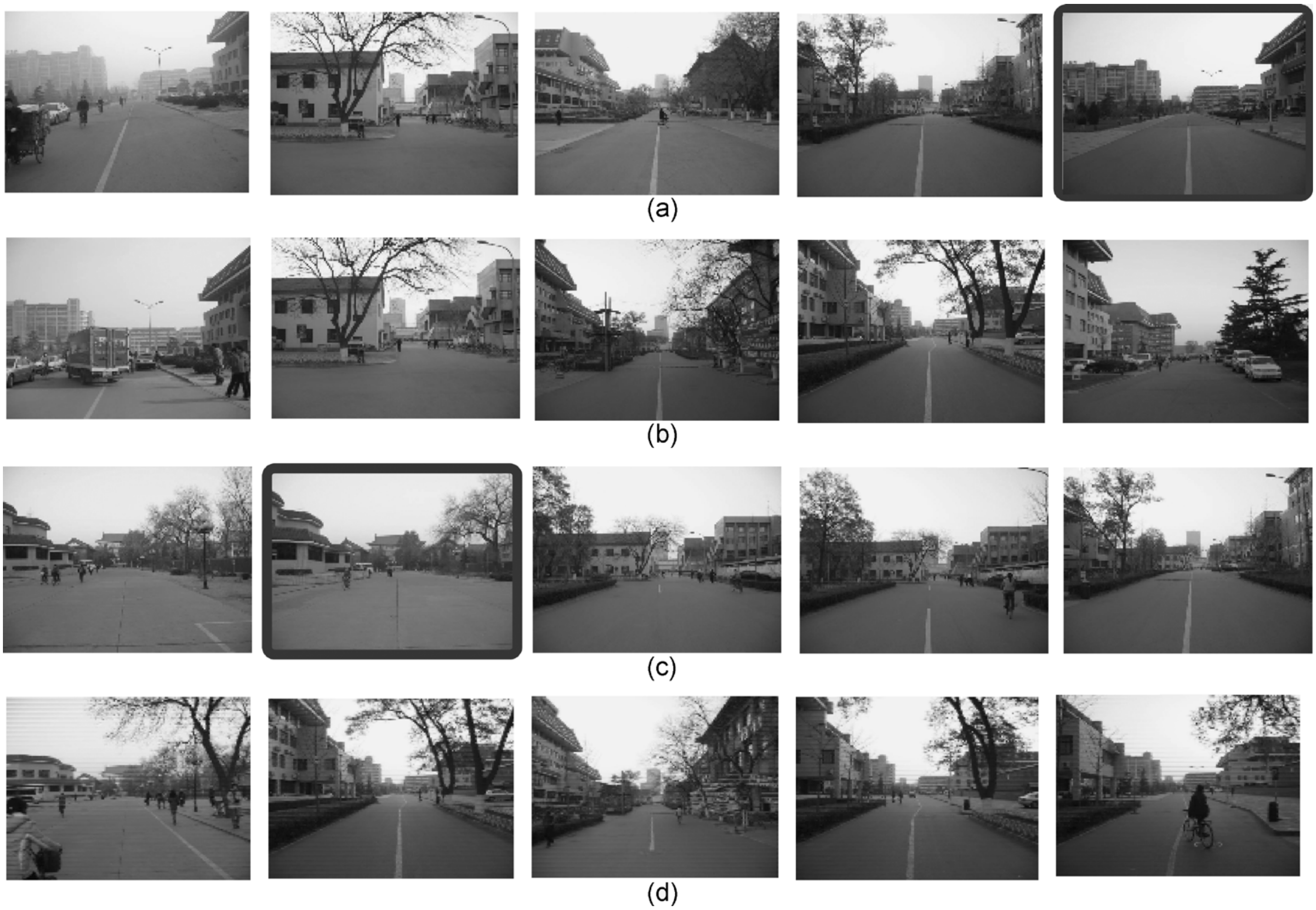

Fig. 12. Examples of localization results. The first image in each row is the one for localization. Localization results are correct in the first row and the third row. The results are wrong in the second row and the fourth row because the correct locations are not retrieved within the top five.

B. Outdoor Experiments

The outdoor localization experiments are carried out in a campus (Fig. 9). At the environment model building stage, 124 images are captured along a route at around every 7 m (Fig. 10). These images are indexed into a LVSM and a database by using the visual vocabulary constructed in indoor experiments. The test set consists of 215 images, which are different from the images for indexing. These images are taken randomly along the route but within 2-m deviation from the first exploration path, at different viewpoints, under different weather conditions. The correct ratio of the coarse localization is described in Fig. 11. Considering the complexity of the outdoor environment, the localization result is rather good. The computation time for localization is similar to the indoor experiments.

In Fig. 12, there are two cases that the localization fails: one is because the image for localization has too much occlusion [Fig. 12(b)]; the other is because the image for localization is taken from a very different viewpoint [Fig. 12(d)]. In such a very different viewpoint, the features resulting from trees change very much. These features become outliers and cannot be dealt with by the Harris–Laplace detector and the SIFT descriptor.

IX. Conclusions and Future Work

We have described a vision-based coarse-to-fine localization framework. The introduction of the LVSM makes the localization process fast. The visual vocabulary is built by learning the descriptors and a term selection method is presented. The employment of orientation information increases the correct ratio of coarse localization. Our approach is robust against illumination and viewpoint changes. Epipolar geometry is used to verify the localization results.

Our work is a possible solution to the initialization and kidnap problems of the SLAM system. We will integrate this approach into a SLAM system which can work in a large environment.

Our localization system can be further improved if we consider the history of movement in the localization process [24], [25], [27]. We are also working on context-based localization by using the Hidden Markov Model.

Acknowledgment

The authors thank the anonymous reviewers for their helpful comments. The first author, J. Wang, would like to thank Dr. J. Barrett for his careful reading of this paper and for advice that improved its readability.

References

[1] R. Cipolla and P. J. Giblin, *Visual Motion of Curves and Surfaces*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[2] A. J. Davidson and D. Murray, "Simultaneous localization and map building using active vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 865–880, Jul. 2002.

[3] G. N. DeSouza and A. C. Kak, "Vision for mobile robot navigation: a survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 237–267, Feb. 2002.

[4] T. Goedemé, M. Nuttin, T. Tuytelaars1, and L. Van Gool, "Markerless computer vision based localization using automatically generated topological maps," in *Proc. European Navigation Conf.*, 2004.

[5] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*: Cambridge Univ. Press, 2001.

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

[7] 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. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002.

[8] H. Katsura, J. Miura, M. Hild, and Y. Shirai, "A view-based outdoor navigation using object recognition robust to changes of weather and seasons," in *Proc. IEEE Int. Conf. Intelligent Robots and Systems*, 2003, pp. 2974–2979.

[9] J. Kosecka and F. Li, "Vision based topological Markov localization," in *Proc. IEEE Intl. Conf. Robotics and Automation*, 2004, pp. 1481–1486.

[10] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. Int. Conf. Computer Vision*, 1999, pp. 1150–1157.

[11] H. P. Luhn, "The automatic creation of literature," *IBM J. Res. and Develop.*, vol. 2, no. 2, pp. 159–165, 1958.

[12] K. Mikoljczyk and C. Schmid, "Indexing based on scale-invariant features," in *Proc. Int. Conf. Computer Vision*, 2001, pp. 525–531.

[13] ——, "A performance evaluation of local descriptors," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2003, pp. 1403–1410.

[14] K. Mikoljczyk, T. Tuyelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *Int. J. Comput. Vis.*, to be published.

[15] T. Ohno, A. Ohya, and S. Yuta, "Autonomous navigation for mobile robots referring pre-recorded image sequence," in *Proc. IEEE Int. Conf. Intelligent Robots and Systems*, vol. 2, Nov. 1996, pp. 672–679.

[16] S. Rajashekhar, S. Chaudhuri, and V. P. Namboodiri, "Image retrieval based on projective invariance," in *Proc. Int. Conf. Image Processing*, Singapore, Oct. 2004.

[17] T. Salton, B. Gerard, A. Buckley, and F. Chris, "Term weighting approaches in automatic text retrieval," *Inform. Process. and Manag.*, vol. 32, no. 4, pp. 431–443, 1996.

[18] S. Se, D. Lowe, and J. Little, "Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks," *Int. J. Robot. Res.*, vol. 21, no. 8, pp. 735–758, Aug. 2002.

[19] C. Shmid and R. Mohr, "Local greyvalue invariants for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 5, pp. 530–534, May 1997.

[20] R. Sims and G. Dudek, "Learning environmental features for pose estimation," *Image and Vis. Comput.*, vol. 19, pp. 733–739, 2001.

[21] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Proc. Int. Conf. Computer Vision*, 2003, pp. 1470–1477.

[22] D. M. Squire, W. Müller, H. Müller, and T. Pun, "Content-based query of image databases: inspirations from text retrieval," *Pattern Recognit. Lett.*, vol. 21, pp. 1193–1198, 2000.

[23] T. Starner, B. Schiele, and A. Pentland, "Visual contextual awareness in wearable computing," in *Proc. Int. Symp. Wearable Computing*, 1998, pp. 50–57.

[24] S. Thrun, "Learning metric-topological maps for indoor mobile robot navigation," *Artif. Intell.*, vol. 99, pp. 21–71, 1999.

[25] S. Thrun, D. Foxy, W. Burgardz, and F. Dellaert, "Robust Monte Carlo localization for mobile robots," *Artif. Intell.*, vol. 128, no. 1–2, pp. 99–141, 2001.

[26] P. H. S. Torr and A. Zisserman, "Mlesac: a new robust estimator with application to estimating image geometry," *Comput. Vis. and Image Understand.*, vol. 78, no. 1, pp. 138–156, 2000.

[27] A. Torralba, K. P. Murphy, W. T. Freeman, and M. Rubin, "Context-based vision system for place and object recognition," in *Proc. Int. Conf. Computer Vision*, 2003, pp. 273–280.

[28] J. Wang, R. Cipolla, and H. Zha, "Image-based localization and pose recovery using scale invariant features," in *Proc. IEEE Int. Conf. Robotics and Biomimetics*, 2004.

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

[30] 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. IEEE Int. Conf. Robotics and Automation*, 2002, pp. 359–365.

Junqiu Wang (S'04) received the B.E. and M.S. degrees from Beijing Institute of Technology, Beijing, China, in 1992 and 1995, respectively.

Since 2002, he has been with National Laboratory on Machine Perception, Peking University, Beijing. His research interests are in computer vision and robotics, including visual tracking, visual servoing, active vision, vision-based localization, and content-based image retrieval.

Hongbin Zha received the B.E. degree in electrical engineering from Hefei University of Technology, Hefei, China, in 1983, and the M.S. and Ph.D. degrees in electrical engineering from Kyushu University, Fukuoka, Japan, in 1987 and 1990, respectively.

After working as a Research Associate in the Department of Control Engineering and Science, Kyushu Institute of Technology, Japan, he joined Kyushu University in 1991 as an Associate Professor. He was also a Visiting Professor in the Centre for Vision, Speech and Signal Processing, Surrey University, Guildford, U.K., in 1999. Since 2000, he has been a Professor at the Center for Information Science, Peking University, Beijing, China. His research interests include computer vision, 3-D geometric modeling, digital museums, and robotics. He has published over 140 technical publications in various journals, books, and international conference proceedings.

Dr. Zha received the Franklin V. Taylor Award from the IEEE Systems, Man, and Cybernetics Society in 1999.

Roberto Cipolla (M'95) received the B.A. degree in engineering from the University of Cambridge, Cambridge, U.K., in 1984, the M.S.E.E. degree in electrical engineering from the University of Pennsylvania in 1985, the M.Eng. degree in robotics from the University of Electro-Communications, Tokyo, Japan, in 1988, and the D.Phil. degree in computer vision from the University of Oxford, Oxford, U.K., in 1991.

From 1985 to 1988, he studied and worked in Japan at the Osaka University of Foreign Studies, Osaka, Japan (Japanese language). During 1991–1992, he was a Toshiba Fellow and Engineer at the Toshiba Corporation Research and Development Center, Kawasaki, Japan. He joined the Department of Engineering, University of Cambridge, in 1992 as a Lecturer and a Fellow of Jesus College. He became a Reader in 1997 and a Professor of information engineering in 2000. His research interests are in computer vision and robotics and include recovery of motion and 3-D shape of visible surfaces from image sequences, visual tracking, and navigation, robot hand-eye coordination, algebraic and geometric invariants for object recognition and perceptual grouping, and novel man-machine interfaces using visual gestures and visual inspection. He is the author of three books, editor of six volumes, and coauthor of more than 200 papers.