Image-based Localization and Pose Recovery Using Scale Invariant Features

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

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

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

Abstract— **In this paper, we propose a vision based mobile robot localization strategy. Local scale-invariant features are used as natural landmarks in unstructured and unmodified environment. The local characteristics of the features we use prove to be robust to occlusion and outliers. In addition, the invariance of the features to viewpoint change makes them suitable landmarks for mobile robot localization. Scale-invariant features detected in the first exploration are indexed into a location database. Indexing and voting allow efficient recognition of global localization. The localization result is verified by epipolar geometry between the representative view in database and the view to be localized, thus the probability of false localization will be decreased. The localization system can recover the pose of the camera mounted on the robot by essential matrix decomposition. Then the position of the robot can be computed easily. Both calibrated and un-calibrated cases are discussed and relative position estimation based on calibrated camera turns out to be the better choice. Experimental results show that our approach is effective and reliable in the case of illumination changes, similarity transformations and extraneous features.**

I. Introduction and Related Work

Mobile robot localization aims to estimate a robot's pose relative to its environment. Since localization gives mobile robot autonomous capability, it plays a pivotal role in mobile robot systems. For birds and insects, scene based navigation strategy is prevalent [1] [17]. Inspired by this phenomenon, visual information has been used to deal with localization and navigation problems. Other sensors such as sonar and laser range finders have also been used for localization. Sonar is fast and cheap but usually very crude. Laser scanning system is active, accurate but slow [3]. In recent years, visual sensor is becoming cheap and reliable. Localization based on visual information attracts more and more attention.

The existing vision based localization approaches can be classified depending on the type of visual information they attempt to use. Many localization systems use global visual information such as collection of views [9] or color histograms [10]. They are sensible to illumination change. Torralba et al. use texture features through wavelet image decomposition [11]. Their system can recognize specific places and new places. It is difficult, however, for the system to recover exact relative position of a robot. Different from the systems using global visual information, we use interest points that can be reliably detected in images. Consequently, our system is robust in the environments where occlusion and extraneous feature exist. Se et al. have used scale invariant visual marks to deal with mobile robot localization [3]. They used Triclops, a vision system that has three cameras. In this work, only one camera is enough for localization [14].

Two key problems in mobile robotics are global position estimation and local position recovery. Global position estimation is to determine the robot's position in an a priori or previously learned map. In our approach, many sets of features in location database represent regions in space (locations). Global position estimation (qualitative localization) is enabled by recognizing locations, which correspond to the regions in the robot's current space that are similar in appearance. After that, many correspondences between features in the representative image and those in the captured image can be found during the localization process. Based on the correspondences, relative pose is recovered. The mobile robot can then decide what to do next step.

If map is a priori unavailable, many applications will allow such a map to be built over time as the robot explores the environment [8]. Vision-based simultaneous localization and map building system (SLAM) can track features and maintaining pose recursively. However, it is only applicable to small scale and feature rich environments up to now. In contrast, our approach can do localization inside a whole building.

A. Overview

The approach described in this paper uses local image features. The flowchart of this approach is in Fig. 1.

First of all, representative images are captured in the first exploration. Next, scale invariant interest points are detected by Harris-Laplace detector. The Harris-Laplace detector is built under multi-scale framework, which makes these interest points robust to scale changes (section II-A). After image feature detection, we describe them by SIFT descriptor (section II-B). Feature detection and description are computed on monochrome version of images; color information is not used in this work. Then, the detected features will be indexed into a location database (section II-C). When mobile robot roams in this building again, it can recognize location by matching image it gets with features in database. In case localization

0-7803-8641-8/04/$20.00 ©2004 IEEE

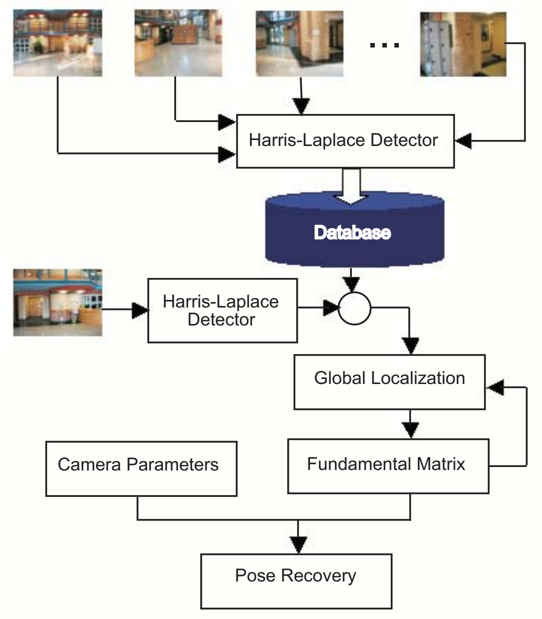

Fig. 1. Flowchart of localization system

result is vague, fundamental matrix estimation is used to verify the result (section III). Further more, it can recover its accurate pose relative to the location in the database (section IV).

II. Database Building

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

A. *Scale Invariant Feature Detection*

Localization in this paper is based on local features that are invariant to scale change. Mikolajczyk and Schmid propose a method for detecting interest point [5]. The first step of this method is to compute interest points (Harris points) at different scales. Then 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 Harris function (Equation 1) [5].

$$det(C) - k \cdot trace(C) > Threshold\_Harris \qquad (1)$$

Where

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

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

In Equation 1, *k* is the coefficient of the Harris function. In this work, it is set to 0.06. *Threshold_Harris* is threshold of the Harris function. It is set to 1500.

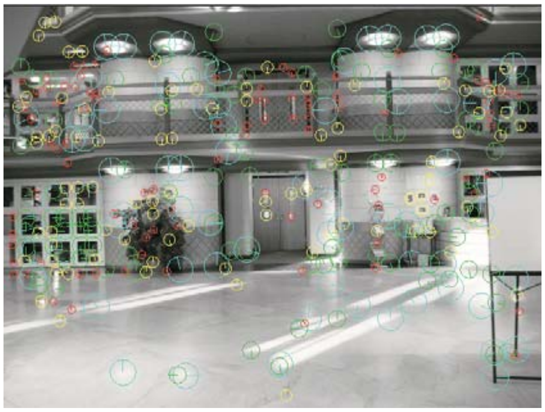

Fig. 2. Interest points detected by Harris-Laplace detector. The centers of the circles are Harris-Laplace interest points. The radiuses of the circles indicate the characteristic scale of interest points.

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 (*Threshold_Harris*), the point is defined as a Harris interest point.

Harris interest points can be detected at different scales. According to [7], local extrema over scale of normalized derivatives indicates the presence of 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. Laplacian (Equation 3) is used in Harris-Laplace interest point detection due to its high detection rate [5].

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

Harris-Laplace interest points can be detected by comparing Laplacian at different scales. The scale of the point with maximum laplacian is taken as the characteristic scale of this interest point.

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

B. *Feature Description*

The output of Harris-Laplace detector is scale invariant points of different size (Fig. 2). These points need to be described for indexing. Many different techniques have been proposed to describe interest point. Inspired by response properties of complex neurons in visual cortex, Lowe proposed SIFT (scale invariant transformation feature) descriptor [6]. 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 is proved to be robust against feature location (coordinates in image) errors and small geometric distortions. In this paper, SIFT descriptor is used to represent interest point. Orientation of point is computed by finding the maximum gradient direction in the point neighborhood. 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, thus this description is also robust under illumination changes. SIFT is sampled over 4x4 grid in the neighborhood of an interest point. The descriptor we get is of dimension 128.

C. *Database Building*

A 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 SIFT descriptor $SIFT_{128}$.

$$M = \{L^i | i = 1, 2, 3...m\} \tag{4}$$

$$L^i = \{V_j^i | j = 1, 2, 3...n\} \tag{5}$$

$$V_j^i = (u, v, \alpha, SIFT_{128})_j^i \tag{6}$$

During the database building process, each vector is added into the database with a link to the location for which it has been computed.

III. Localization and Verification

When the robot roams again in the environment, image is captured. Given a single view, localization system can find the location that that is the closest to the input image from the database. Localization is realized by using voting scheme. Each input descriptor *V* is compared with the descriptors in the database. Then the Euclidean distance between the input descriptor and the descriptors in location database is computed. If the Euclidean distance is below a threshold, the corresponding location gets a vote. The vote for each location in the database can be easily accumulated. The location that gets the largest number of votes is the most likely location.

In most cases, localization system gets correct location directly. Nevertheless, if the image captured has a very different viewpoint from any locations in the location database, there might be two or even three locations getting almost the same number of votes. In Fig. 3, location 1, 18 get 30 votes, 6 gets 28 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 is used here to verify the voting result. Fundamental matrix **F** is estimated by using RANSAC algorithm [13]. This algorithm is robust to outliers. If a vote (correspondence between interest point in image captured and features in database) is accepted by fundamental matrix, this 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. 3, localization system can now decide that location 1 is the location we need because it has 22 inliers.

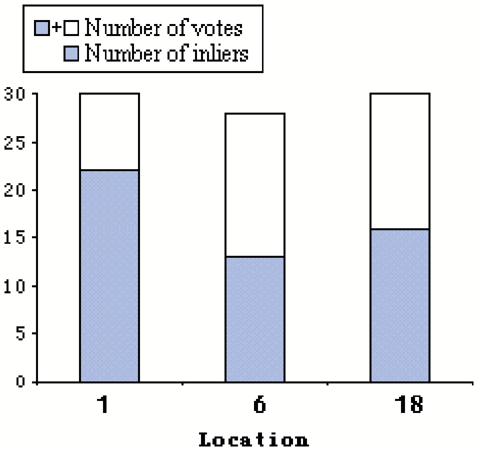

Fig. 3. Result of localization and verification. Only three locations with the largest number of votes are displayed. Location 1 is correctly found after using epipolar geometry constraints.

IV. 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 essential matrix.

We can get camera internal parameter matrix $\mathbf{K}_1$ and $\mathbf{K}_2$ using camera calibration method, where $\mathbf{K}_1$ is the internal parameters of the camera used in the very first exploration, and $\mathbf{K}_2$ is the internal parameters of the camera used in following exploration.

Based on fundamental matrix **F** and camera internal parameters $\mathbf{K}_1$, $\mathbf{K}_2$, essential matrix **E** is computed [15]:

$$\mathbf{E} = \mathbf{K}_1^T \mathbf{F} \mathbf{K}_2 \tag{7}$$

Essential matrix can be decomposed into rotation **R** and translation **t** [15]:

$$\mathbf{E} = [\mathbf{t}]_x \mathbf{R} \tag{8}$$

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

Essential matrix has two equal singular values and one zero singular value [15][16]. We can compute the rotation (Equation. 11 and 12) and translation (Equation. 14 and 15) based on singular value decomposition of essential matrix [16].

$$\mathbf{E} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{9}$$

Where

$$\mathbf{D} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{10}$$

$$\mathbf{R} = \mathbf{U}\mathbf{W}\mathbf{V}^T \tag{11}$$

or

$$\mathbf{R} = \mathbf{U}\mathbf{W}^T\mathbf{V}^T \tag{12}$$

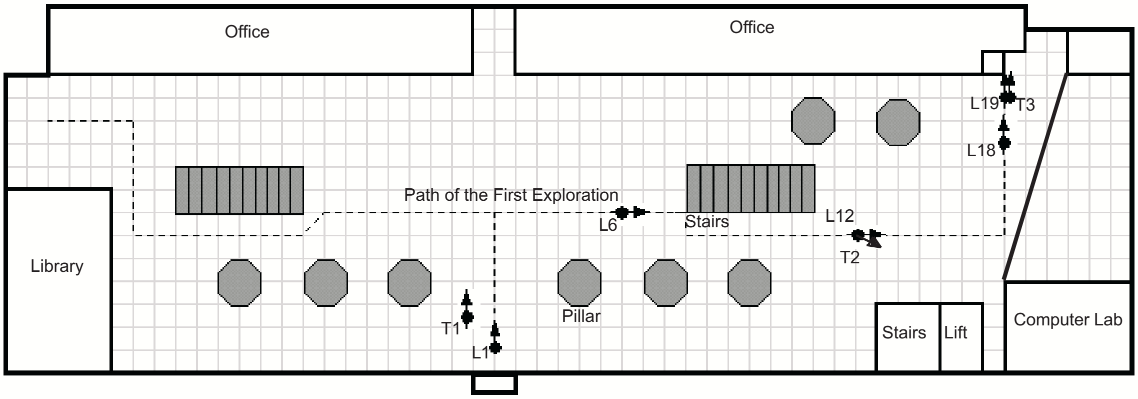

Fig. 4. Layout of the ground floor. L1, L6, L12, L18 and L19 are locations in database (Other locations are not shown in this figure). T1, T2 and T3 are sites where images for localization are taken.

Where

$$\mathbf{W} = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{13}$$

$$\mathbf{t} = \mathbf{U}\begin{pmatrix} 0 & 0 & 1 \end{pmatrix}^T \tag{14}$$

or

$$\mathbf{t} = -\mathbf{U}\begin{pmatrix} 0 & 0 & 1 \end{pmatrix}^T \tag{15}$$

The solution of the pose is one of following four possible matrices [16].

$$(\mathbf{UWV}^T|\mathbf{U}\begin{pmatrix} 0 & 0 & 1 \end{pmatrix}^T) \tag{16}$$

$$(\mathbf{UWV}^T|-\mathbf{U}\begin{pmatrix} 0 & 0 & 1 \end{pmatrix}^T) \tag{17}$$

$$(\mathbf{UW}^T\mathbf{V}^T|\mathbf{U}\begin{pmatrix} 0 & 0 & 1 \end{pmatrix}^T) \tag{18}$$

$$(\mathbf{UW}^T\mathbf{V}^T|-\mathbf{U}\begin{pmatrix} 0 & 0 & 1 \end{pmatrix}^T) \tag{19}$$

The points in images must lie in front of of both cameras. The right solution can be computed by checking whether the point lies before the camera [16].

Different solutions have been given by many researchers to recover pose without camera calibration. We have tried method in [12]. However, the pose recovery result based on this method is too bad to be used (The solution of the pose is totally different from the ground truth). Since camera calibration is much easier than before, Pose recovery with calibration is good enough to deal with most problems. Other localization systems using one camera compute motion between cameras based on known camera internal parameters [1][14].

V. Experiments

The system described above has been implemented and tested in a building. The database used throughout the experiments contains 37 images. These images were captured, using a camera, at different locations in the ground floor of the building. Fig. 4 is the sketch of the ground floor. Most images were taken at an interval of 2 meters. Thus the image sequence was sparse and easy to be indexed into database. In addition, localization was fast when the database was small.

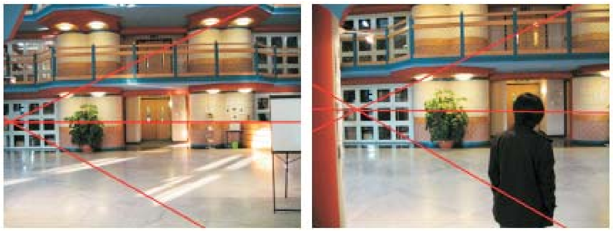

Fig. 5. Left, location gotten from database (L1 in Fig. 4). Right, image for localization (T1 in Fig. 4). Localization system gets correct location despite the distance between them is 2 meters.

We captured 30 images randomly under different conditions such as viewpoint and illumination. These images were used to test the global localization and pose recovery. The result shows that our system is robust to those changes. We got correct location from the database when the distance between them was two meters (Fig. 5). Image taken at different viewpoint was also correctly retrieved from the database. Localization was accurate when the angle between image in database and captured image is 20 degrees (Fig. 6). The image taken under very bad illumination condition was also correctly found in the database (Fig. 7). It demonstrates the advantage of using scale invariant features. Epipolar lines are plotted on images in Fig. 5 to Fig. 7.

In each localization, it took 0.5 second to detect interest points in image and 0.8∼1.2 second to find the correspondent location in the database (on 1.4GHz laptop). The speed is not fast enough. However, our database that uses only 37 images to cover the ground floor. It is a ”sparse” database. In contrast, several hundreds images are used in [9]. The robot can walk for more than 1 meter without doing localization again. In addition, the interest point detection might speed up by using new algorithms we are developing.

VI. Conclusions and Future Work

We have discussed mobile robot localization approach based on visual information. As far as we know, this is the first time that Harris-Laplace interest point detector is used in a

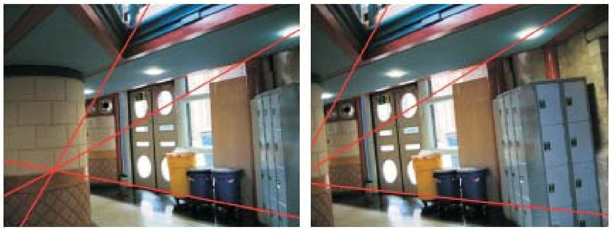

Fig. 6. Left, location retrieved from database (L12 in Fig. 4). Right, image for localization (T2 in Fig. 4). Localization is correct although the angle between them is 20 degrees.

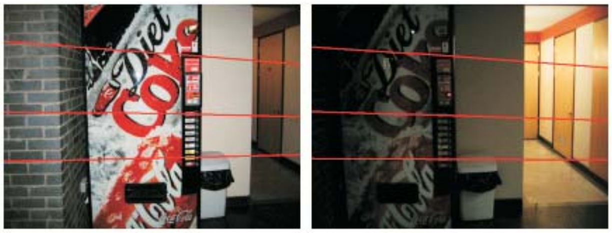

Fig. 7. Left, location retrieved from database (L19 in Fig. 4). Left, image for localization (T3 in Fig. 4). Localization is correct although there are illumination and viewpoint change.

localization system. In the current stage the experiments have been carried out using purely local information. The interest point is characterized by SIFT descriptor. The experiments demonstrate promising performance. It is robust against illumination and viewpoint change. Location can be found correctly.

Odometry is a very important input in many existing systems. However, odometry is not very accurate especially after robot has walked for a long distance. Error will probably be accumulated and localization is no longer precise. Our method does not depend on odometry. It is not only more accurate but also more flexible.

Our work up to now is based on images captured by a common camera. We will use video automatically captured during an exploration. The image sequences in the video will be classified and clustered.

Since the speed of our system is not fast enough, we are working on improving the performance of Harris-Laplace detector. At the same time, we are doing research on a new indexing method based on feature clustering algorithms.

REFERENCES

[1] A. J. Davison, ”Real-time simultaneous localization and mapping with a single camera”, In *Proc.of the IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2003.

[2] J. Borenstein, H.R. Everett, and L. Feng, ”’Where am I?’ Sensors and Methods for Mobile Robot Positioning” Technical Report, The University of Michigan.

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

[4] J. Kosecka, L. Zhou, P. Barber, Z. Duric, ”Qualitative Image Based Localization in Indoors Environments”, In *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2003

[5] K. Mikoljczyk and C. Schmid, ”Indexing based on scale-invariant features”, In *Proc. of the IEEE Int. Conf. on Computer Vision*, pages 525-531. 2001.

[6] D. G. Lowe, ”Object recognition from local scale-invariant features”, In *Proc. of the IEEE Int. Conf. on Computer Vision*, pages 1150-1157, 1999.

[7] T. Lindeberg, ”Feature detection with automatic scale selection”, *International Journal of Computer Vision*, 30(2): 79-116, 1998.

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

[9] 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 the IEEE Int. Conf. On Robotics and Automation*, 2002.

[10] T. Starner, B. Schiele, and A. Pentland, ”Visual contextual awareness in wearable computing”, In *Intl. Symposium on wearable Computing*, pages 50-57, 1998

[11] A. Torralba, K. P. Murphy, W. T. Freeman and M. Rubin, ” Context-based vision system for place and object recognition”, In *Proc. of the IEEE Int. Conf. On Computer Vision*, 2003

[12] R. I. Hartley, ”Estimation of relative camera positions for uncalibrated cameras”, In *Proc. of the IEEE Int. Conf. On Computer Vision*. 1992.

[13] 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.

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

[15] R. Cipolla and P.J. Giblin, *Visual Motion of Curves and Surfaces*, Cambridge University Press, 2000.

[16] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2001.

[17] P. R. Ehrlich, D. S. Dobkin and D. Wheye, ”Flying in vee formation”, *http://www.stanfordalumni.org/birdsite/text/essays/Flying_in_Vee.html*.