Global Localization and Relative Pose Estimation Based on Scale-Invariant Features

Jana Košecká and Xiaolong Yang
Computer Science Department, George Mason University
Fairfax, VA 22030, USA

Abstract

The capability of maintaining the pose of the mobile robot is central for basic navigation and map building tasks. In this paper we describe a vision-based hybrid localization scheme based on scale-invariant keypoints. In the first stage the topological localization is accomplished by matching the keypoints detected in the current view with the database of model views. Once the best match has been found, the relative pose between the model view and the current image is recovered. We demonstrate the efficiency of the location recognition approach and present a closed form solution to the relative pose recovery for the case of planar motion and unknown focal length of the camera. The approach is demonstrated on several examples of indoors environments.

1. Introduction and Related Work

The existing techniques for vision-based localization and map building vary depending on the representation of the environment and means of localization. The methods for continuous pose maintenance typically recover the pose of the robot and structure of the environment in a recursive setting using tracked point features [5]. Such methods have been applied successfully in smaller scale environments. Using more descriptive scale-invariant features, the tasks of map building, pose maintenance and global localization were demonstrated by [14], using trinocular stereo sensor. Approaches for localization by means of recognition vary in the choice of features and means of determining an environment model. Commonly used representations are responses to a banks of filters [17], multi-dimensional histograms [12, 7], local Fourier-transforms [15] and affine invariant feature descriptors [8]. These representations in the context of mobile robot navigation were most commonly obtained via principal component analysis (PCA) or various clustering techniques [1, 6]. Alternative biologically inspired method mimicking behavior of bees was presented in [2].

Our approach is motivated by the recent advances in object recognition using local scale invariant features proposed by [8] and adopts the strategy for localization by means of location recognition. The image sequence acquired by a robot during exploration is first partitioned to individual locations, while recording the neighborhood relationships between them. Each location is represented by a set of model views and their associated scale-invariant features. In the first topological localization stage, the current view is classified as belonging to one of the locations. Once the most likely location view is determined, we compute the relative pose between the current view and the representative view of the location. The scale invariant features are sufficiently discriminant for successful location recognition and can handle larger displacements between the model views and the test views.

2. Scale-Invariant Features

The use of the local feature detectors in the context of object recognition has been demonstrated successfully by several researchers in the past [13, 11]. In this paper we examine the effectiveness of scale-invariant (SIFT) features proposed by [8]. These features have been shown to be stable across wide variations in viewpoint and scale and can be localized efficiently. They correspond to stable points in the scale space and can be detected by searching for peaks in the image $D(x, y, \sigma)$, where $D(x, y, \sigma) = G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y)$ is obtained by taking differences of two neighboring images in the scale space build with Gaussian kernel $G(x, y, \sigma)$. In the second stage the detected peaks with low contrast or poor localization along the edge are discarded. More detailed discussion about enforcing the separation between the features, sampling of the scale space and improvement in feature localization can be found in [8, 4]. Once the location and scale have been assigned to candidate keypoints, the dominant orientation is computed by determining peaks in the orientation histogram of its local neigh-

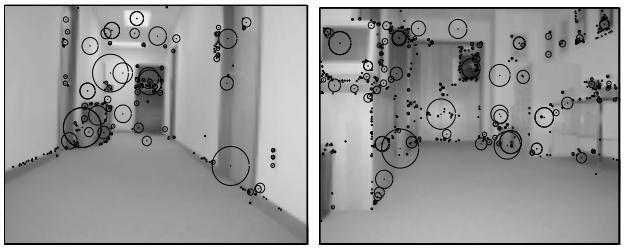

Figure 1. Examples of scale invariant keypoints. The circle center represents keypoint's location and the radius is proportional to it's scale.

bourhood weighted by the gradient magnitude. The keypoint descriptor is then formed by computing local orientation histograms (with 8 bin resolution) over 4×4 grid overlayed over 16×16 neighbourhood of the point. This yields 128 dimensional feature vector which is normalized to unit length in order to reduce the sensitivity to image contrast and brightness changes in the matching stage. Figure 1 shows the keypoints found in the example images in our environment. In man-made indoors environments, the number of features detected varies between 10 to 1000, in an image of size 480×640.

3. Location Recognition

The model of the environment, which we will use to test our localization method is obtained in the exploration stage. Given a temporally sub-sampled sequence acquired during the exploration (images were taken approximately every 2-3 meters), we partition the sequence into 18 different locations. Different locations in our model correspond to hallways, sections of corridors and meeting rooms approached at different headings. The number of views per location vary between 8 to 20 depending on the appearance variation within the location. The transitions between the locations occur either at places where navigation decisions have to be made or when the appearance of the location changes suddenly. In order to obtain more compact representation of the environment, we next choose the representative views for each location. In this stage we experimented with different strategies varying the number of representative views by (1) choosing a single view taken in the middle of the location, (2) choosing views taken in the middle and at the end of each location or (3) evenly sampling the views belonging to individual location yielding up to 4 views per location. Table 1 shows the location recognition results as a function of number of representative views on one training sequence and two test sequences. Given a new image, each detected SIFT keypoint is matched against the database of keypoints choosing the nearest neighbor based on Euclidean distance between two descriptors. In the subsequent voting scheme we determine the location whose keypoints were most frequently classified as nearest neighbors; such location is the most likely location where the current view came from. We only consider point matches whose nearest neighbor is at least 0.5 times closer then the second nearest neighbor. Despite the large number of representative views (up to 4), relatively poor performance on the second test sequence (134 images) was due to several changes in the environment between the training and testing stage. In 5 out of 18 locations several objects were moved or misplaced. This sensitivity to dynamic changes is not surprising, since the most discriminative SIFT features often belong to objects some of which are not inherent to particular locations. The recognition rate in such case can be improved by selecting larger number of representative views and/or exploiting the knowledge of environments' topology captured by the neighborhood relationships between individual locations. The Hidden Markov Model which explicitly exploits the spatial relationships algorithm is described in more detail in [18]. Next we describe how to recover the relative displacement between the current view and the closest view retrieved from the database.

sequence	NO.1 (250)	NO.2 (134)	NO.3 (130)
one view	84%	46%	44%
two views	97.6%	68%	66%
four views	100%	82%	83%

Table 1. Recognition rate in % of correctly classified views.

4. Pose estimation and match refinement

The current view and the matched model view are related by a rigid body displacement $g = (R, T)$ represented by a rotation $R \in SO(3)$ and translation $T = [t_x, t_y, t_z]^T \in \mathbb{R}^3$. Provided that the camera is calibrated, g can be estimated from the epipolar geometry between the two views. This recovery problem can be further simplified taking into account the fact that the motion of the robot is restricted to a plane. Here we outline an algorithm for this special case and demonstrate how to recover the displacement in case of unknown focal length. The case of general motion and unknown focal length was studied by [16] and the solution for the case of planar motion case has been proposed by [3] in the context of uncalibrated stereo. Here we demonstrate a slightly different, more concise solution to the problem. Consider the perspective camera projection model, where 3D coordinates of point $\mathbf{X} = [X, Y, Z]^T$ are related to their

image projections $\mathbf{x} = [x, y, 1]^T$ by an unknown scale λ; $\lambda\mathbf{x} = \mathbf{X}$. In case the camera is calibrated the two views of the scene are related by $\lambda_2\mathbf{x}_2 = R\lambda_1\mathbf{x}_1 + T$, where $(R, T) \in SE(3)$ is a rigid body transformation and λ_1 and λ_2 are the unknown depths with respect to individual camera frames. After elimination of the unknown scales from the above equation, the relationship between the two views is captured by so-called epipolar constraint

$$\mathbf{x}_2^T \widehat{T} R \mathbf{x}_1 = \mathbf{x}_2^T E \mathbf{x}_1 = 0, \tag{1}$$

where $E = \widehat{T}R$ is the essential matrix[1] In case of planar motion, assuming translation in $x - z$ plane and rotation around $y-$axis by an angle θ, the essential matrix has the following sparse form

$$E = \begin{bmatrix} 0 & -t_z & 0 \\ t_z c\theta + t_1 s\theta & 0 & t_z s\theta - t_1 c\theta \\ 0 & t_x & 0 \end{bmatrix} \tag{2}$$

where $s\theta(c\theta)$ denote $\sin\theta(\cos\theta)$ respectively. Given at least four point correspondences, the elements of the essential matrix $[e_1, e_2, e_3, e_4]^T$ can be obtained as a least squares solution of a system of homogeneous equations of the form (1). Once the essential matrix E has been recovered, the four different solutions for θ and $T = \pm[t_x, 0, t_z]$ can be obtained (using basic trigonometry) directly from the parametrization of the essential matrix (2). The physically correct solution is then obtained using the positive depth constraint. In the case of unknown focal length the two views are related by so called fundamental matrix F

$$\tilde{\mathbf{x}}_2^T F \tilde{\mathbf{x}}_1 = 0 \text{ with } \mathbf{x} = K^{-1}\tilde{\mathbf{x}}. \tag{3}$$

The fundamental matrix F is in this special planar, partially calibrated case related to the essential matrix E as follows

$$F = K^{-T} E K^{-1} \text{ with } K = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4}$$

where f is the unknown focal length. The remaining intrinsic parameters are assumed to be known. In the planar motion case the matrix $F = [0, f_1, 0; f_2, 0, f_3; 0, f_4, 0]$ can be recovered from the homogeneous constraints of the form (3) given a minimum of four matched points. The extraction of the unknown motion parameters and the focal length f however is not straightforward, since the translation and the focal length appear in the parametrization of the matrix F in a multiplicative way. We propose to use additional constraints provided by so-called Kruppa's equations [10]. It can be easily verified that a fundamental matrix F between the two views and the unknown intrinsic parameter matrix K satisfy the following constraint

$$FKK^TF^T = \lambda^2 \widehat{e} K K^T \widehat{e}^T \tag{5}$$

1 $\widehat{T}$ denotes a 3×3 skew symmetric matrix associated with vector T.

where $e = \frac{KT}{\|KT\|}$ is the epipole and λ is the unknown scale of the fundamental matrix. In our previous work [10] we have shown that for the special case of planar motion the above equation is satisfied if and only if $\lambda = 1$. Since F and $e = [-f_1, 0, f_4]^T$ can be estimated, the renormalized equation (5) yields following useful constraint on intrinsic parameters K

$$FKK^TF^T = \widehat{e} K K^T \widehat{e}^T. \tag{6}$$

Given the planar motion case, the middle entries of matrices on the left and right side of equation (6) yield a constraint on the focal length and the entries of the fundamental matrix

$$f_2^2 f^2 + f_3^2 = f_4^2 f^2 + f_1^2.$$

The solution for the focal length can then be directly obtained from the above equation as

$$f = \sqrt{\frac{f_1^2 - f_3^2}{f_2^2 - f_4^2}}. \tag{7}$$

Once f is computed, the relative displacement between the views can be obtained by the method outlined for the calibrated case. Additional care has to be taken in assuring that the detected matches do not come from a degenerate configuration. We have used RANSAC algorithm for the robust estimation of the pose between two views, with slightly modified sampling strategy. Figure 2 shows two examples of relative positioning with respect to two different representative views. The focal length estimates obtained for these examples are $f = 624.33$ and $f = 545.30$. The relative camera pose for individual views is represented in the figure by coordinate frame. Although we do not have ground truth measurements for these experiments the recovered motions are consistent with the changes in visual appearance between the views along the two test paths. More detailed experiments evaluating the sensitivity of the method can be found in [18].

5. Conclusions and Future Work

We have demonstrated the suitability and the discrimination capability of scale-invariant SIFT features in the context of location recognition task. Although the matching and location recognition methods can be accomplished using an efficient and simple voting scheme, the recognition rate is not surprisingly affected by small dynamic changes in the environment. This can be partially resolved by incorporating more model views, using alternative more global descriptors and exploiting the topology of the environment. We also presented a method for computing relative displacement between the current view and the model view which enables metric localization with a location and can be used for relative positioning tasks. We are in the process

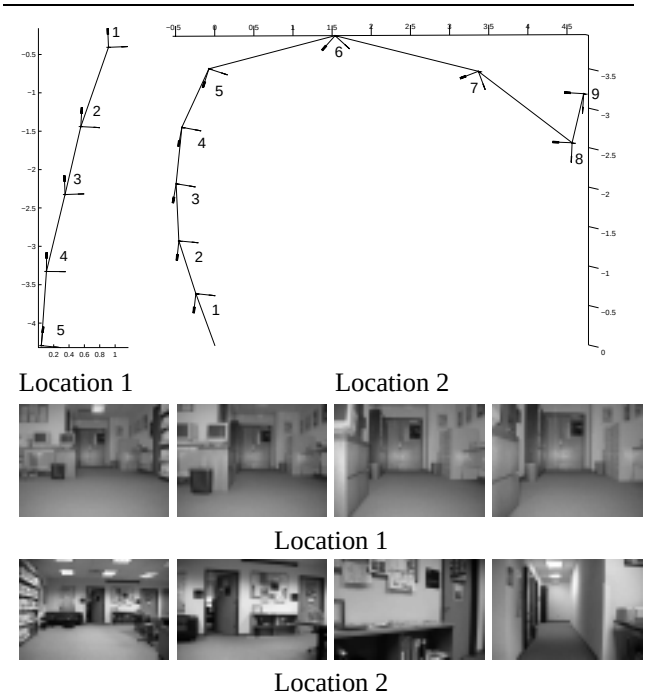

Figure 2. Relative positioning experiments with respect to the representative views. Bottom: Query views along the path between the first view and the representative view for two different locations. Top: Recovered motions for two locations.

of carrying out more extensive experiments and incorporating the above techniques on the on-board navigation system of a mobile robot.

6. Acknowledgements

The authors would like to thank D. Lowe for making available the code for detection of SIFT features. This work is supported by NSF grant IIS-0118732 and George Mason University Provost Scholarship fund.

References

[1] M. Artac, M. Jogan, and A. Leonardis. Mobile robot localization using an incremental eigenspace model. In *IEEE Conference of Robotics and Automation*, pages 1025 – 1030, 2002.

[2] G. Bianco and A. Zelinsky. Biologically-inspired visual landmarks learning and navigation for mobile robots. In *IEEE/RSJ International Conference on Intelligent Robots and System*, 1999.

[3] M. Brooks, L. D. Agapito, D. Huyng, and L. Baumela. Towards robust metric reconstruction via a dynamic uncalibrated stereo. *Image and Vision Computing*, 16(14):989–1002, 1998.

[4] M. Brown and D. Lowe. Invariant features from interest point groups. In *In Proceedings of BMVC, Cardiff, Wales*, pages 656–665, 2002.

[5] A. Davidson and D. Murrray. Simultaneous localization and map building using active vision. *IEEE Transactions on PAMI*, 24(7):865–880, 2002.

[6] J. Gaspar, N. Winters, and J. Santos-Victor. Vision-based navigation and environmental representations with an omnidirectional camera. *IEEE Transactions on Robotics and Automation*, pages 777–789, December 2000.

[7] J. Košecká, L. Zhou, P. Barber, and Z. Duric. Qualitative image based localization in indoors environments. In *IEEE Proceedings of CVPR*, pages 3–8, 2003.

[8] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, page to appear, 2004.

[9] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry. *Invitation to 3D Vision: From Images to Geometric Models*. Springer-Verlag, 2003.

[10] Y. Ma, R. Vidal, J. Košecká, and S. Sastry. Kruppa's equation revisted: Degeneracy and renormalization. In *ECCV*, 2000.

[11] R. Nelson and A. Selinger. Large scale tests of a keyed, apperance based 3d object recognition system. *Vision Research*, 38(15):2469–99, 1998.

[12] B. Schiele and J. L. Crowley. Object recognition using multidimensional receptive field histograms. *International Journal of Computer Vision*, 2000.

[13] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on PAMI*, 19(5):530–534, 1997.

[14] S. Se, D. Lowe, and J. Little. Global localization using distinctive visual features. In *Proc. of International Conference on Robots and Systems*, pages 153–158, 2002.

[15] R. Sims and G. Dudek. Learning envirmental features for pose estimation. *Image and Vision Computing*, 19(11):733–739, 2001.

[16] P. Sturm. On focal length calibration from two views. In *Conference on CVPR, Kauai*, volume 2, pages 145–150, 2001.

[17] A. Torralba and P. Sinha. Recognizing indoor scenes. *MIT AI Memo*, 2001.

[18] X. Xang and J. Košecká. Experiments in location recognition using scale-invariant sift features. Technical Report GMU-TR-2004-2, George Mason University, 2004.