Estimating Geospatial Trajectory of a Moving Camera

Asaad Hakeem[1], Roberto Vezzani[2], Mubarak Shah[1], Rita Cucchiara[2]
[1]School of Electrical Engineering and Computer Science, University of Central Florida,
Orlando, Florida 32816, USA. {ahakeem,shah}@cs.ucf.edu
[2]Dipartimento di Ingegneria dell'Informazione, Universit di Modena e Reggio Emilia,
Modena 41100, Italy. {vezzani.roberto,cucchiara.rita}@unimore.it

Abstract

This paper proposes a novel method for estimating the geospatial trajectory of a moving camera. The proposed method uses a set of reference images with known GPS (global positioning system) locations to recover the trajectory of a moving camera using geometric constraints. The proposed method has three main steps. First, scale invariant features transform (SIFT) are detected and matched between the reference images and the video frames to calculate a weighted adjacency matrix (WAM) based on the number of SIFT matches. Second, using the estimated WAM, the maximum matching reference image is selected for the current video frame, which is then used to estimate the relative position (rotation and translation) of the video frame using the fundamental matrix constraint. The relative position is recovered upto a scale factor and a triangulation among the video frame and two reference images is performed to resolve the scale ambiguity. Third, an outlier rejection and trajectory smoothing (using b-spline) post processing step is employed. This is because the estimated camera locations may be noisy due to bad point correspondence or degenerate estimates of fundamental matrices. Results of recovering camera trajectory are reported for real sequences.

1. Introduction

GPS was first introduced by the US Department of Defense (DoD) about 15 years ago for military personnel and vehicle tracking around the world. Since then the GPS technology has been widely used in the areas of autonomous navigation and localization of vehicles and robots. Recently, commercial applications have employed GPS data with georeferenced maps to recover the map of a city address or to provide directions between different city locations. In this paper, we address two issues: Localizing the geospatial position and estimating the trajectory of a moving camera based on the captured sequence of images. Geospatial position is localized using maximum SIFT matches between the reference images (with known GPS) and video frames while a camera trajectory is estimated using the geometric constraint between maximum matching reference images and video frames.

In literature, a variety of methods have been proposed for motion recovery and measurement of robot trajectory (odometry) using visual inputs. Structure from Motion (SFM) is the most common approach to solve problems such as automatic environment reconstruction, autonomous robot navigation and self-localization. These approaches employ a 3D reconstruction of the environment during learning phase or directly use the test video. Thus, the actual camera position is obtained by 3D matching of the current view with the learned environment map. Methods that use such techniques for recovering geospatial location include [6, 4, 7]. The most recent work using this approach was proposed by Royer et al. in [13] for mobile robot navigation. The robot is first manually guided on a learning path. Later, a map of the environment and its 3D reconstruction is performed off-line. Using this 3D reconstruction, the robot is able to recover its pose with respect to the 3D environment model. An approach for Simultaneous Localization And Mapping (SLAM) was proposed in [3], which is based on the Extended Kalman Filter (EKF). It assumes a robot moving in a world with stationary landmarks (distinctive physical features) that can be observed by some sensor. The positions of the landmarks along with the robot's position at a particular time are considered to be the system state. The problem consists of estimating the new state (robot and landmark positions) at the next time instance, given, the last movement made by the robot and new observations provided by the sensory subsystem. The typical sensors used for measuring the distance and orientation of landmarks with respect to the robot are sonar rings [12], and more frequently, the laser scanners [3]. Davison proposed two methods that do not require specialized hardware (laser scanners or sonar rings) for measuring distance and orientation of landmarks. His solutions are based on an active binocular head [1] and a single camera [2] that estimate the distance and orientation of visual landmarks.

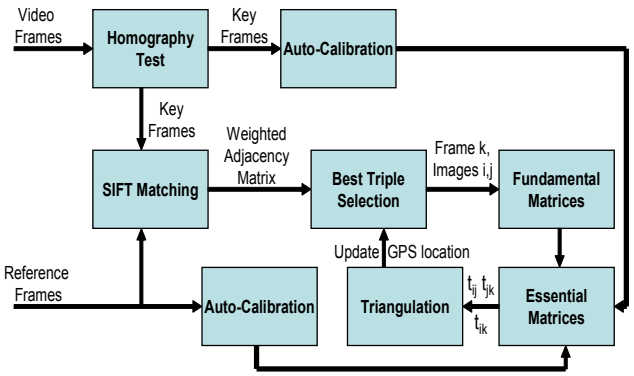

Figure 1. Overall geospatial localization framework.

A disadvantage of these solutions is the need for an initial manual calibration: both for the position and orientation of the robot with respect to a predefined target of known size.

SFM and SLAM based approaches have two major disadvantages. Firstly, the task is computationally very expensive and is unnecessary to recover the trajectory of the camera. Secondly, the 3D reconstruction of the environment may fail at certain instances where distinctive features cannot be computed e.g. images with trees only or areas with sparse buildings. Since these methods rely on 3D environment reconstruction and use a matching algorithm for pose recovery, these methods may not fully recover the complete video trajectory. What is novel in our approach is that we do not require a 3D reconstruction of the environment for recovering the camera trajectory. Instead, we require a set of reference images with known GPS locations for geospatial localization of the novel video data. Furthermore, we use sub-sampled video frames for localization and interpolate the trajectory by spline fitting in order to obtain a smooth camera trajectory. Thus, the advantages of our method are twofold. First, our method does not require all the video frames to have distinctive features for geospatial localization. Second, our method is computationally less expensive since we do not require 3D reconstruction and matching of the environment for localization.

2. System Overview

Our goal is to compute the geospatial localization of the novel video frames $\{V_t, t = 1..M\}$, given a set of reference images $\{I_p, p = 1..N\}$ with known GPS. We assume that some video frames have overlapping field of view with the reference images. Further, the camera does not zoom while capturing the reference images and video frames i.e. constant intrinsic parameters. These assumptions allow the auto-calibration of the capturing devices.

The overall geospatial localization framework is given in Figure 1. Given a set of reference images, we recover the camera intrinsic parameters using a method proposed by Luong et al. in [10]. Similarly, we recover the video camera's intrinsic parameters using the video frames. Next, we perform homography tests between the sequence of video frames to select keyframes from the video. Further, SIFT features are detected and matched between the reference images and the video keyframes (see Section 2.1). Using the maximum matching reference images and video frames, we estimate the fundamental and essential matrices between image I_p and frame V_t pairs to recover the camera pose (see Section 2.2). Finally, we apply triangulation to recover scale ambiguity in the estimated camera pose and obtain the geospatial localization of the video keyframe (see Section 2.3). We repeat these steps for all the video keyframes and apply trajectory smoothing by fitting splines to the estimated GPS locations. The following sections detail these steps.

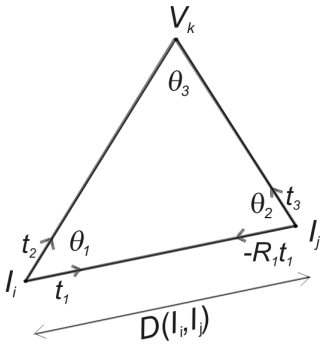

Figure 2. Triangulation between two reference images and video frame used to resolve scale ambiguity for GPS location estimation.

2.1. Estimating the Weighted Adjacency Matrix

In order to obtain geospatial localization using the fundamental matrix constraint, we require feature point correspondence between reference images I_p and video frames V_t. There are several methods to obtain point correspondence between images including Harris corner detector [5], Scale and affine invariant point detector [11], and SIFT [8]. We empirically evaluated all the three point correspondence methods and found SIFT to be the most robust matching method across a substantial range of affine distortion, change in viewpoint, addition of noise, and change in illumination. The SIFT features are highly distinctive, and each feature point is represented by 128 dimensional feature vector.

A match is found for a feature in frame V_t to a feature in I_p by estimating the ratio of the smallest to second smallest Euclidean distance between the feature vectors. In [8], the authors reject all the matches which have a distance ratio greater than 0.8, which eliminates 90% of the false matches discarding less than 5% of the correct matches. In our application, we set this threshold to 0.4, in order to have less number of very reliable matches. The above matching scheme may result in multiple feature points in frame V_t matching with the same feature in image I_p. We obtain a one-to-one correspondence by maximum matching of a bipartite graph. The bipartite graph construction is obtained

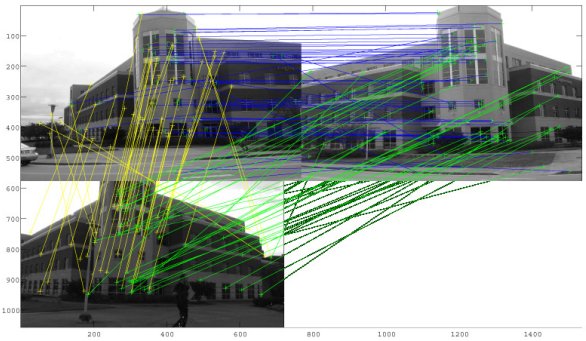

Figure 3. Example of best triplet selection and matching for video frame 270 of Engineering building sequence. The two reference images are in top row while the video frame is in bottom row.

by treating the two feature point sets as nodes in bipartitions (I_p and V_t) and the distance ratio as the weight on the edges between the bipartition. We obtain a weighted adjacency matrix $W(I_p, V_t)$ by finding the point correspondence between all pairs of I_p and V_t. Each entry in the matrix corresponds to the number of matching features between I_p and V_t. The corresponding set of matching point locations is stored in a matching matrix $M(I_p, V_t)$ such that:

$$M(I_p, V_t) = \{(x_1, y_1, x_2, y_2); (x_1, y_1) \in I_p \wedge (x_2, y_2) \in V_t\}$$

2.2. Pose Recovery of the Video Frames

Given a video frame V_t, we want to recover the position $P_{V_t} = \{X_t, Y_t, Z_t\}$ of its camera optical center with respect to a reference image I_p in the world coordinate system. Since the GPS location of the reference images I_p are given in terms of longitude and latitude, we apply spherical to cartesian conversion to obtain $P_{I_p} = \{X_p, Y_p, Z_p\}$. Later, we find the maximum matching reference image I_p (using $W(I_p, V_t)$). Furthermore, we utilize the set of corresponding points $M(I_p, V_t)$ to estimate the fundamental matrix F_t^p between images V_t and I_p using the constraint:

$$[x_1\, y_1\, 1] \cdot F_t^p \cdot [x_2\, y_2\, 1]^T = 0;\ \forall (x_1, y_1, x_2, y_2) \in M(I_p, V_t)$$

Due to noise in feature point location and incorrect point correspondence, the estimation of the fundamental matrix using the above linear constraint is erroneous. In order to obtain a robust estimate, we use RANSAC based fundamental estimation technique proposed by Torr et al. in [14]. Using the fundamental matrix and calibration matrices K_t and K_p (obtained using auto-calibration), we can estimate the essential matrix E_t^p by:

$$E_t^p = K_t \cdot F_t^p \cdot K_p.$$

The rotation R and translation t between V_t and I_p can be recovered from the essential matrix by using methods proposed in [9]. The translation vector t thus obtained, is recovered up to a scale factor and this ambiguity is resolved using triangulation which is described in the next section.

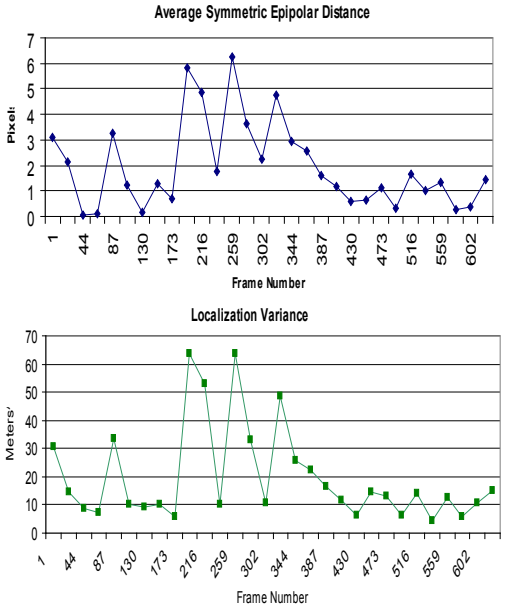

Figure 4. Comparison of the average symmetric epipolar distance with the localization variance for the theater sequence. Higher localization variance is due to higher average symmetric epipolar distance which depicts that the fundamental matrix estimate is erroneous due to noise.

2.3. Resolving Scale Ambiguity

The triangulation scheme is employed to recover the scale ambiguity and obtain $P_{V_t} = \{X_t, Y_t, Z_t\}$ i.e. the position of camera center for the video frame V_t. The triangulation method requires an image triplet (two reference images I_i, I_j and a video frame V_k), and the construction is depicted in Figure 2. Thus, for each video frame V_t, two reference images are selected from the entire set such that the obtained triple has the maximum matching feature points with the video frame. The two reference images should also have different GPS locations, otherwise, the resulting triangulation construction is degenerate. More formally, given a video frame V_k, the best triplet $BT(V_k)$ is obtained using:

$$BT(V_k) = (V_k, I_i, I_j)\ ;\ \text{such that}$$
$$(i,j) = \underset{i,j=1..N}{argmax} \left(min \left(\left\{ \begin{array}{c} W(V_k, I_i), \\ W(V_k, I_j), \\ W(I_i, I_j) \end{array} \right\} \right) \right)$$

Given the rotations and translations between each pair of camera coordinate system (R_i^j, t_i^j), (R_i^k, t_i^k), (R_j^k, t_j^k), we can compute the three internal angles (for the triangle) using:

$$\begin{aligned} \theta_1 &= \cos^{-1} \frac{(dot(t_1, t_2))}{norm(t_1) norm(t_2)} \\ \theta_2 &= \cos^{-1} \frac{(dot(-t_1, R_1 t_3))}{norm(t_1) norm(t_3)} \\ \theta_3 &= 180 - \theta_1 - \theta_2 \end{aligned}$$

where $norm(x)$ is the magnitude of x. The scale factor is recovered through distance $D(I_i, I_j)$ obtained from the

Figure 5. Examples of reference images with known GPS locations used in our experiments.

two GPS locations of the reference images. The location of camera center for video frame V_k is obtained using trigonometric identities. If any angle is 0 degrees (collinear image locations), then the current triplet is discarded and a new triplet is used for triangulation. The recovered camera center is converted from cartesian to spherical coordinates to obtain GPS location in latitude and longitude. Figure 3 shows an example of best triplet selection for a video frame in the engineering building sequence.

2.4. Trajectory Smoothing

The fundamental matrix estimates are highly sensitive to noise in feature correspondence. Thus, numerical errors, insufficient feature point correspondence and noise could result in incorrect estimation of GPS locations for the video frames. Performing multiple estimations of the fundamental matrix (using RANSAC), we obtain a spatial distribution of GPS locations for each video frame. If the point correspondence is reliable, the variance in the spatial distribution of GPS location will be minimal. Therefore, we discard GPS estimates of the video frames with high variance and reduce GPS estimation error for the video trajectory. Finally, we use the remaining GPS locations as control points on a b-spline and interpolate the rest of the trajectory by curve fitting.

3. Results and Discussion

The accuracy of our geospatial localization was tested against the ground truth GPS information obtained using a Garmin GPSMAP 76S unit that has an accuracy of 3 meters. We captured over 300 reference images (some examples shown in Figure 5) using a Nikon D2X camera at 4 MP (mega pixels per image) at various locations on our campus. The video sequences were captured using a Sony HDR FX1 camera at HD quality. We empirically evaluated our method to recover the geospatial trajectory of a moving camera on four sequences totalling over 2500 video frames. The summary of results for the four sequences is given in Table 1.

Given sets of reference images and video frames we recover the respective camera intrinsic parameters using a method proposed by Luong et al. in [10]. Next, we perform

Table 1. MEAN LOCALIZATION ERROR FOR DIFFERENT VIDEOS

Sequence	Total Frames	# of Keyframes	Avg. Est. Error
Public Affairs	405	42	4.27 meters
Engineering II	325	33	3.93 meters
Theater	645	61	3.55 meters
Health Center	1187	82	5.12 meters

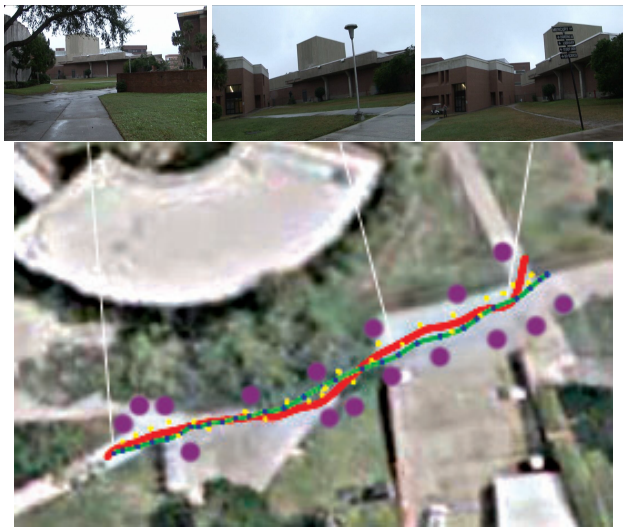

Figure 6. Top: Video frames for the theater sequence. Bottom: Video trajectory (green) obtained using our method is compared with ground truth trajectory (red). The blue points depict control points for the spline, while the yellow points are GPS estimates with localization variance greater than the threshold. The length of trajectory was approximately 35 meters and the reference images are marked as purple dots.

homography tests between the sequence of video frames to select keyframes from the video. That is, we keep only those keyframes that do not fit a homography with the previously selected keyframe. Using our algorithm we estimate all the GPS locations and apply trajectory smoothing as a post processing step. This is because the fundamental matrix estimates are highly sensitive to noise in feature correspondence due to numerical errors or incorrect point correspondence. Performing multiple estimations of the fundamental matrix (100 estimates using RANSAC), we obtain a spatial distribution of GPS locations for each video frame. The GPS estimates of video frames with high variance in the spatial distribution for multiple runs will have a high symmetric epipolar distance (distance between the epipolar lines and corresponding points). This is empirically evaluated and shown in Figure 4 for the theater sequence. We thus retain the mean of the GPS estimates with low variance, discard GPS estimates of the video frames with high variance (threshold learnt using a small training sequence) and reduce GPS estimation error for the video trajectory by fitting splines on the reliable GPS estimates.

Figure 6 shows a video trajectory obtained using our method and a comparison with the ground truth for the the-

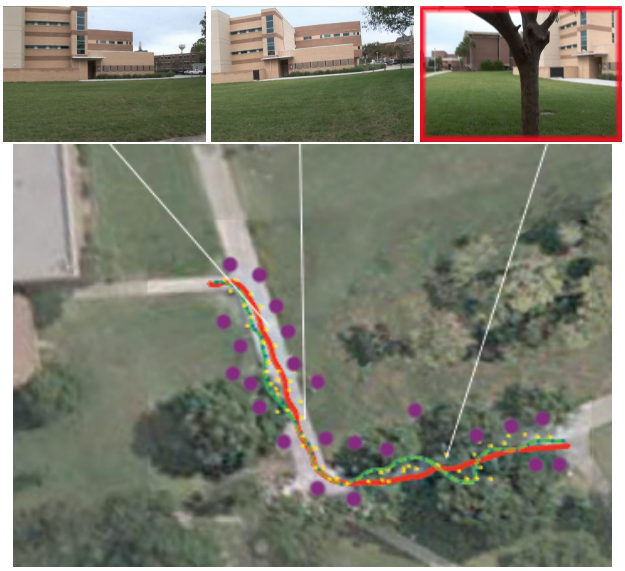

Figure 7. Top: Video frames for the health center sequence. Bottom: Video trajectory (green) obtained using our method is compared with ground truth trajectory (red). The blue points depict control points for the spline, while the yellow points are GPS estimates with localization variance greater than the threshold. The length of trajectory was approximately 65 meters and the reference images are marked as purple dots.

ater sequence. The average localization estimation error for this sequence is 3.55 meters with a standard deviation of 2.12 meters. Figure 7 shows a video trajectory obtained using our method and a comparison with the ground truth for the health center sequence. The average localization estimation error for this sequence is 5.12 meters with a standard deviation of 5.36 meters. The reason for higher localization estimation error is due to the presence of trees and non-distinctive features (grass) in the middle of the sequence. Though the average estimation error for our method is about 5 meters, this sequence cannot be used for SFM based methods that rely on distinctive features throughout the sequence so that a 3D reconstruction and matching can be used for pose recovery and localization. This is the major advantage of our method and is more generally applicable to real sequences for geospatial localization.

In order to test the robustness of our GPS estimation method, we conducted experiments on the golden datasets (test4 and final5) of the ICCV Contest 2005. We estimate the GPS locations for each image with an unknown GPS location using the best triplet and triangulation method (described in Section 2). The summary of our results is provided in Table 2. Further, visual comparison of our results with the ground truth is given in Figures 8 and 9 for test4 and final5 respectively. It is evident from our results that our method has an average estimation error of 3.05 meters and 6.08 meters for datasets test4 and final5 respectively. While the average score for test4 and final5 using the histogram of

Figure 8. Top: Images with unknown GPS location for Test4. Bottom: GPS locations (blue) obtained using our method is compared with ground truth GPS (red). The white lines show the distance between the estimated and actual GPS locations.

error method used in the ICCV Contest is 4.2 and 3.5 respectively. The best scores in the contest for these datasets were 5.0 and 3.5 respectively. This corroborates the fact that our method performs well for estimating GPS locations for standard datasets.

Table 2. MEAN LOCALIZATION ERROR FOR ICCV CONTEST DATASET

Dataset	Knowns	Unknowns	Avg. Score	Avg. Est. Error
Test4	9	20	4.2	3.05 meters
Final5	16	22	3.5	6.08 meters

4. Conclusions

The paper proposed a novel method for estimating the geospatial trajectory of a moving camera. We used the video data and a set of reference images, captured from known GPS locations, to recover the trajectory of the moving camera. The novelty of our approach is that we do not require a 3D reconstruction of the environment for recovering the camera trajectory. Instead, we require a set of reference images with known GPS locations for geospatial localization of the novel video data. Additionally, we use sub-sampled video frames for localization and interpolate the trajectory by spline fitting in order to obtain a smooth camera trajectory. The advantage of our method is twofold: First, our method does not require all the video frames to have distinctive features for geospatial localization. Second, our method is computationally inexpensive since we do not require 3D reconstruction and matching of the environment for localization.

Acknowledgement

This material is based upon the work funded in part by the U.S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government.

Figure 9. Top: Images with unknown GPS location for Final5. Bottom: GPS locations (blue) obtained using our method is compared with ground truth GPS (red). The white lines show the distance between the estimated and actual GPS locations.

References

[1] A. J. Davison and N. Kita. "3D Simultaneous Localisation and Map-Building Using Active Vision for a Robot Moving on Undulated Terrain". In *Proc. of Computer Vision and Pattern Recognition*, pp.384-391, 2002. 1

[2] A. J. Davison. "Real-Time Simultaneous Localisation and Mapping with a Single Camera". In *Proc. of International Conference on Computer Vision*, pp.1403-1410, 2003. 1

[3] G. Dissanayake, P. Newman, S. Clark, H. F. Durrant-Whyte and M. Csorba. "A Solution to the Simultaneous Localization and Map Building (SLAM) Problem". In *IEEE Transaction on Robotics and Automation*, vol.17(3), pp.229-241, 2001. 1

[4] M. A. Garcia and A. Solanas. "3D simultaneous localization and modeling from stereo vision". In *Proc. of IEEE Int'l Conf. on Robotics and Automation*, pp.847-853, 2004. 1

[5] C. Harris and M. Stephens. "A combined corner and edge detector". In *Proc. of the Alvey Vision Conference*, pp.147-151, 1988. 2

[6] K. Kidono, J. Miura and Y. Shirai. "Autonomous Visual Navigation of a Mobile Robot Using a Human-Guided Experience". In *Robotics and Autonomous Systems*, vol.40(2), pp.124-132, 2002. 1

[7] A. Levin and R. Szeliski. "Visual Odometry and Map Correlation". In *Proc. of the Computer Vision and Pattern Recognition*, pp.611-618, 2004. 1

[8] D. G. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints". In *International Journal of Computer Vision*, vol.60(2), pp.91-110, 2004. 2

[9] Q. T. Luong and O. D. Faugeras. "The fundamental matrix: theory, algorithms, and stability analysis". In *International Journal of Computer Vision*, vol.17(1), pp.43-76, 1996. 3

[10] Q. T. Luong and O. D. Faugeras. "Self-Calibration of a Moving Camera from Point Correspondences and Fundamental Matrices". In *International Journal of Computer Vision*, vol.22(3), pp.261-289, 1997. 2, 4

[11] K. Mikolajczyk and C. Schmid. "Scale and affine invariant interest point detectors". In *International Journal of Computer Vision*, vol.60(1), pp.63-86, 2004. 2

[12] P. Newman, J. Leonard, J.D. Tardos and J. Neira. "Experimental Validation of Real-Time Concurrent Mapping and Localization". In *Proc. of the IEEE Int. Conf. on Robotics and Automation*, pp.1802-1809, 2002. 1

[13] E. Royer, M. Lhuillier, M. Dhome and T. Chateau. "Localization in Urban Environments: Monocular Vision Compared to a Differential GPS Sensor". In *Proc. of Computer Vision and Pattern Recognition*, pp.114-121, 2005. 1

[14] P. H. S. Torr and D. W. Murray. "Outlier Detection and Motion Segmentation". In *Sensor Fusion VI*, SPIE vol.2059, Boston, pp.432-443, 1993. 3