Design and Implementation of a Monocular Vision Sensor Model for Mobile Robot Localization

Matthew N. Dailey[1] Akash Dev Nakarmi[2] Rassarin Chinnachodteeranun[1]
Thavida Maneewarn[3] Manukid Parnichkun[4]

[1] Computer Science and Information Management, Asian Institute of Technology
[2] Information and Communication Technologies, Asian Institute of Technology
[3] Institute of Field Robotics, King Mongkut's University of Technology Thonburi
[4] Mechatronics, Asian Institute of Technology

Abstract. General availability of inexpensive, all-terrain, lightweight autonomous mobile robots would drive innovation in many application areas such as security, remote sensing, inspection, and landmine detection. Many of these tasks require accurate 6-DOF robot localization, and cameras are potentially ideal for this purpose, but most current techniques use stereo vision sensors, which are expensive or difficult to calibrate. Towards flexible and low cost vision-based robot localization, we have designed, implemented, and evaluated a prototype system incorporating SIFT feature detection and factorization-based methods for sparse Euclidean reconstruction from three uncalibrated views. We test the system on synthetic data, in a virtual world, and in a real-world outdoor environment, using three different reconstruction algorithms. We find that of the algorithms tested, Han and Kanade's algorithm achieves the best reconstruction. We conclude that SIFT combined with factorization-based structure from motion is suitable for single-camera mobile robot localization.

1 Introduction

Many mobile robot applications require precise 6 degree of freedom estimates of the robot's position in the world. Chief among these are what we might call "coverage" applications like landmine detection, home cleaning, lawn mowing, airport runway crack detection, and so on, in which we must ensure that every point on a given surface in the workspace is covered by the robot's sensors and/or actuators.

We are interested in the development of inexpensive, all-terrain, lightweight robots for coverage applications, especially for cases in large-scale environments in which the environment map is not known a priori. Due to their light weight, low cost, and the rich information they provide, cameras are ideal sensors for mobile robot localization and mapping in these applications.

Recently, there have been several applications of vision sensors to the problem localization and mapping in large-scale environments [1–4] but these systems use binocular or trinocular stereo. To minimize the cost of the robot, it should be possible to perform localization and mapping using a single camera. While there has been progress in this direction [5, 6], the systems are not designed to scale to large environments. The largest-scale experiments to date have employed variants of the Rao-Blackewellized particle filter [7] such as FastSLAM [8] that require Gaussian landmark measurements in robot (or camera) coordinates.

Since a single camera sensor model would normally provide a cone-shaped measurement uncertainty for 2D point features, to obtain Gaussian measurements of a feature location using a single camera, we would need views of that feature from more than one camera location. With known cameras and known extrinsic parameters, it is possible to obtain a Euclidean 3D reconstruction from two views; with unknown extrinsic parameters, we require at least three views, and we can only reconstruct up to an unknown scale factor.

For outdoor robots, it is feasible to use a 3-view uncalibrated method combined with a rough absolute measurement device such as a wheel encoder or an inexpensive GPS receiver to resolve the scale ambiguity. Once mapping is bootstrapped and the global scale is known, previously stored landmarks can in principle be used to resolve the scale ambiguity for new measurements.

In this paper, we therefore focus on the problem of designing a sensor to bootstrap the mapping process via Euclidean 3D reconstruction from three intitial monocular views of an unknown environment. We compare three candidate solutions [9–11] from the literature using SIFT [12] for the 2D interest point detector. We conduct experiments with a synthetic object, in a simulated world, and on real imagery. Of the three Euclidean reconstruction methods, we find that Han and Kanade's algorithm [13, 10] achieves the best reconstruction. We conclude that SIFT combined with factorization-based structure from motion is a suitable design for bootstrapping a single-camera mobile robot localization system.

2 Experimental Methods

2.1 Data collection

As a testbed for experimentation, we collected three types of data:

- Synthetic 3D data
- Images rendered from a VRML model
- Images acquired in a real-world outdoor environment

The synthetic data were 9 3D points drawn at random from a 1m-radius sphere with center placed 10m from a virtual camera with a horizontal field of view of 68 degrees.

The VRML simulation data consisted of a sequence of 44 images taken approximately 25cm apart with small random fluctuations in camera rotation within a 3D model of Housestead's Fort, a Roman garrison from the 3rd century A.D. on Hadrian's Wall in Britain [14]. The virtual camera had a resolution

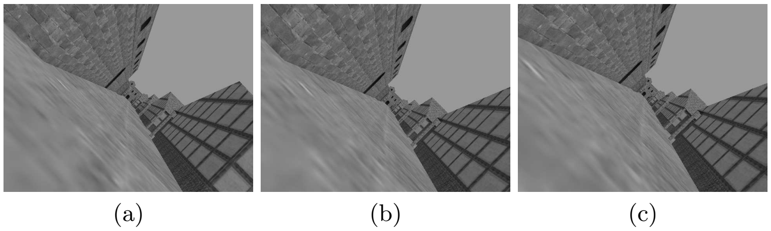


Fig. 1. Sample simulation image set. (a) Initial position. (b) Second position, after approximate 0.25m forward motion. (c) Third position, after another approximate 0.25m forward motion.

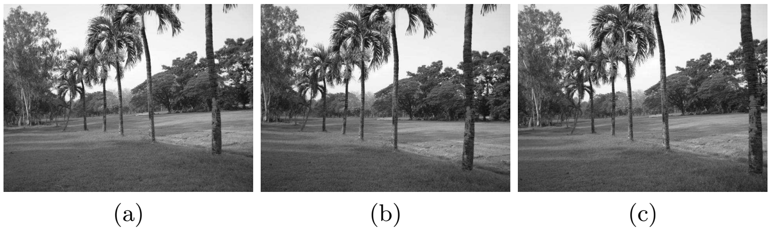


Fig. 2. Sample outdoor imagery after correction for camera distortion. (a) Initial position. (b) Second position, after approximate 0.25m forward motion. (c) Third position, after another approximate 0.25m forward motion.

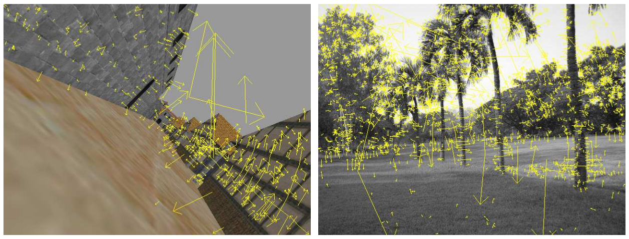
of 640 × 480 pixels and a horizontal field of view of 70 degrees. A sample image "triplet" from the VRML simulation is shown in Fig. 1. This environment is an interesting testbed for mobile robot localization algorithms because it is highly textured, creating a large number of feature points, and the textures are highly repetitive in many places, leading to many ambiguities for correspondence algorithms. The random camera rotations preclude the use of flat-earth or three-degree-of-freedom assumptions.

The real-world outdoor data consisted of 21 images acquired with a Sony DSC-200 digital camera on a golf course on the Asian Institute of Technology campus. We captured the images from positions approximately 50cm from the ground, approximately 25cm apart, with small random fluctuations in camera rotation. We calibrated the digital camera using the CalTech camera calibration toolbox [15] then undistorted each image according to an idealized 640 × 480 pinhole camera with with horizontal field of view of 57.9 degrees. A sample undistorted image triplet is shown in Fig. 2.

In all three environments, we group the image sequences into triplets prior to obtaining point correspondences and performing 3D reconstruction.

2.2 Point correspondences

Once an image triplet has been acquired, the first step is to obtain point correspondences between the three images. Our method first extracts 2D SIFT [12]



(a) (b)

Fig. 3. SIFT features. Direction of each arrow indicate the keypoint's orientation; length of each arrow indicates the keypoint's scale. (a) VRML simulation image. (b) An undistorted outdoor image.

features, obtains a tentative set of correspondences with Beis and Lowe's best bin first (BBF) algorithm [16], then eliminates outliers using fundamental matrix estimation [17] wrapped inside RANSAC [18].

The scale invariant feature transform (SIFT) [12] extracts 2D points along with descriptors that are meant to be invariant with respect to translation, rotation, and scale. The idea is to perform difference of Gaussian (DoG) filtering at multiple scales then find the local extrema of the DoG filter's response across x, y, and scale. SIFT then eliminates extrema with low contrast, simple texture, or uncertain orientation. Finally, for each extremum surviving the previous steps, the algorithm assigns a feature descriptor consisting of relative gradient orientations and magnitudes in the surrounding region. The SIFT keypoints for one each of our simulation and golf course images are shown in Fig. 3.

Now, given descriptors $\{\boldsymbol{k}_{i,j}\}$ representing keypoint j in image i, we need to find a set of tentative correspondences $C = \{< \boldsymbol{k}_{1,i}, \boldsymbol{k}_{2,j}, \boldsymbol{k}_{3,k} >\}$ between keypoints. Ideally, for each keypoint in image 1, we would like to compute the "nearest neighbor" keypoint in image 2 in terms of Euclidean distance, and similarly for the nearest neighbors of the image 2 keypoints in image 3. Our tentative correspondence for keypoint i in image 1 would be the nearest neighbor $\boldsymbol{k}_{2,j}$ of $\boldsymbol{k}_{1,i}$ and the nearest neighbor $\boldsymbol{k}_{3,k}$ of $\boldsymbol{k}_{2,j}$. However, the best known algorithms for exact nearest neighbor search in high dimensional spaces (SIFT keypoint descriptors have 128 elements) require a prohibitive amount of compute time, so we use Beis and Lowe's approximate nearest neighbor algorithm based on the k-d tree [16]. The idea is to perform standard k-d tree search but rather than searching all bins of the tree that might contain the nearest neighbor, BBF maintains a priority queue containing a limited number of candidate bins that will *most likely* contain the nearest neighbor.

After applying BBF to find the approximate set of nearest neighbors of the image 1 keypoints in image 2 and the approximate set of nearest neighbors of the image 2 keypoints in image 3, we eliminate any keypoints matched to more than one keypoint in another image, any keypoints not matched across all three images, and any keypoint for which the nearest neighbor match is not much closer than the second nearest neighbor. The result is a candidate set of unique correspondences across all three images in the given triplet.

Of course, SIFT keypoint descriptors do not capture spatial relationships with other keypoints. This means that the set of nearest-neighbor keypoints are not necessarily spatially consistent. In fact, we find that correspondences obtained as just described, even for small camera motions between subsequent images, are often less than 50% correct. To eliminate inconsistent correspondences, we apply RANSAC-based [18] fundamental matrix estimation as recommended by Hartley and Zisserman [17] and implemented by OpenCV [19]. We perform outlier elimination for each image pair separately. For a given set of point correspondences between two images, the approach repeatedly selects a random sample of 7 correspondences, uses a 7-point algorithm for fundamental matrix estimation, then calculates the the number of "inlier" correspondences for which the putative matching point is within a fixed distance of the predicted epipolar line. The fundamental matrix with the most inliers is retained, used to reject outliers, then a final fundamental matrix is estimated from all of the inliers using a normalized 8-point algorithm.

After eliminating the outlying correspondences in each pair of images, we form a set of high-confidence keypoints matched across all three images. These correspondences are used as input to the 3D Euclidean reconstruction process.

2.3 Euclidean reconstruction

Once we have obtained a reliable set of corresponding points across three frames, it is possible to obtain a Euclidean 3D reconstruction up to an unknown scale factor. In our experiments, we apply three different approaches for Euclidean reconstruction:

- Christy and Horaud [9]
- Han and Kanade [13, 10]
- Tang and Hung [11]

Christy and Horaud's method [9] iteratively performs a Euclidean 3D reconstruction of the corresponding points under the assumption of a paraperspective projection then applies to each image measurement an adjustment by a "paraperspective correction" reflecting the difference between perspective and paraperspective projections of the 3D point. At convergence (the algorithm is not guaranteed to converge), the affine 3D reconstruction from the paraperspective-corrected image points is also a perspective 3D reconstruction of the original image points. The affine reconstruction step first factors the image measurements into motion (projection) and shape (3D point) components using Tomasi and Kanade's method [20]. This gives a projective reconstruction which is then upgraded to a Euclidean reconstruction by applying constraints requiring that the projections must be rigid transformations rather than arbitrary projections.

Han and Kanade's approach [13, 10], rather than affine reconstruction, employs iterative projective reconstruction. The method begins by assuming the projective depth of every point is 1. It then iteratively factors the image measurements scaled by their projective depths into shape and motion components then recomputes the projective depths. After convergence of the projective reconstruction step, Han and Kanade apply metric constraints similar to Christy and Horaud's, but also allow for uncertainty about the camera's intrinsic parameters.

Tang and Hung [11] factor the same scaled measurement matrix as Han and Kanade into separate shape and motion components, but they also allow for points to be missing in some views and guarantee convergence by explicitly minimizing the same objective function (with respect to different parameters) in every step of each iteration. To upgrade from a projective reconstruction to a Euclidean reconstruction, Tang and Hung use the same method as Han and Kanade.

2.4 Reconstruction quality measurement

Each of the three types of data we collected (synthetic, simulation, and real-world) requires a different method for measuring reconstruction quality.

For all three data sources, we report reprojection error, i.e., the RMS error between the actual observed 2D points and the projections of the corresponding reconstructed 3D points.

For the synthetic data, we know the ground truth for the 3D points but the reconstruction is only known up to a similarity ambiguity. In this case we find a least-squares estimate of a projective transformation between the ground truth data and the reconstruction, then report the RMS error for the transformed 3D reconstructions of the points. This only measures the quality of the projective reconstruction, not the Euclidean reconstruction of the 3D data.

3 Results

Here we describe our experiments applying the three Euclidean reconstruction algorithms to each of three types of 2D data. The data sources have already been described. We generated the synthetic data in Gnu Octave. The simulated images of Housestead's Fort were rendered with OpenGL from a VRML 1.0 model. We used Hess' implementation of SIFT and BBF [21]. For RANSAC with fundamental matrix estimation, we used OpenCV [19]. The maximum distance from an inlier correspondence and its predicted epipolar line was 0.5 pixels.

3.1 Synthetic data

Table 3.1 shows the performance of each reconstruction algorithm in terms of reprojection error and 3D point reconstruction error for the 9 points we randomly generated on a 1m-radius sphere. The Han and Kanade method performed best in terms of reprojection of the Euclidean reconstruction. All algorithms performed very well in terms of projective reconstruction, with error less than 0.1% of the distance to the camera.

Table 1. Results for synthetic data experiment

Method	2D RMS (pixels)	3D RMS (meters)
Christy and Horaud [9]	0.6660	0.0081
Han and Kanade [10]	0.0566	0.0081
Tang and Hung [11]	0.3387	0.0072

3.2 Simulation data

On our simulation images, SIFT detected an average of 439 feature points per image. After BBF search and RANSAC we obtained an average of 29 matching keypoints across all three images in each triplet.

Table 3.2 shows the performance of each reconstruction algorithm in terms of reprojection error over all 42 image triplets in the simulation sequence. (Recall that we do not have ground truth for the 3D points in the simulation.) The Han and Kanade method performed best.

Table 2. Results for simulation data experiment

Method	2D RMS (pixels)
Christy and Horaud [9]	2.3738
Han and Kanade [10]	1.0297
Tang and Hung [11]	7.0835

3.3 Outdoor data

On our undistorted outdoor images, SIFT detected an average of 2290 feature points. After BBF search and RANSAC we obtained an average of 122 matching keypoints across all three images in each triplet.

Table 3.3 shows the performance of each reconstruction algorithm in terms of reprojection error, over all 19 image triplets in the image sequence. (Recall that we do not have ground truth for the 3D points in the real-world data.) The Han and Kanade method again performed best.

Table 3. Results for outdoor data experiment

Method	2D RMS (pixels)
Christy and Horaud [9]	2.2060
Han and Kanade [10]	1.1813
Tang and Hung [11]	1.8826

4 Discussion and Conclusion

One way to minimize the cost of robots using vision-based localization and mapping is to eliminate all sensors except for a single camera. However, bootstrapping 3D localization and mapping without any a priori knowledge requires 2D correspondences across 3 views, a good Euclidean reconstruction algorithm, and some way to resolve the scale ambiguity.

In this paper, we have shown that Han and Kanade's reconstruction approach, combined with SIFT 2D feature extraction, is well-suited to the 3D map bootstrapping task. With perfect correspondences, the method achieves excellent results, in terms of both reprojection error and 3D projective reconstruction error. As we increase the realism of the test, moving from synthetic data to a VRML simulation to real-world outdoor imagery, the error increases, as expected, but Han and Kanade's approach still outperforms two other methods on the same data.

Techniques like those demonstrated in this paper will be extremely useful in the constructon of low-cost autonomous mobile robots able to perceive and act in highly uncertain environments.

In future work, we plan to deploy the reconstruction technique on a prototype teleoperated landmine detection robot and explore methods for growing an established map once the scale ambiguity is resolved.

Acknowledgments

This research was supported by Thailand Research Fund grant MRG4780209 to MND. ADN was supported by a scholarship from the government of Finland and a fellowship from AIT. RC was supported by a scholarship from the Royal Thai Government.

References

1. Se, S., Lowe, D., Little, J.: Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. The International Journal of Robotics Research **21**(8) (2002) 735–758
2. Sim, R., Elinas, P., Griffin, M., Little, J.: Vision-based SLAM using the Rao-Blackwellised particle filter. In: IJCAI Workshop on Reasoning with Uncertainty in Robotics (RUR). (2005)
3. Se, S., Barfoot, T., Jasiobedzki, P.: Visual motion estimation and terrain modeling for planetary rovers. In: Proceedings of the ISAIRAS 2005 Conference. (2005)
4. Dailey, M.N., Parnichkun, M.: Simultaneous localization and mapping with stereo vision. In: International Conference on Automation, Robotics, and Computer Vision (ICARCV). (2006)
5. Davison, A.: Real-time simultaneous localisation and mapping with a single camera. In: Proceedings of the International Conference on Computer Vision (ICCV). (2003) 1403–1410

6. Davison, A., Cid, Y., Kita, N.: Real-time 3D SLAM with wide-angle vision. In: Proceedings of the IFAC Symposium on Intelligent Autonomous Vehicles. (2004)
7. Murphy, K.: Bayesian map learning in dynamic environments. In: Advances in Neural Information Processing Systems (NIPS). (1999)
8. Thrun, S., Montemerlo, M., Koller, D., Wegbreit, B., Nieto, J., Nebot, E.: Fast-SLAM: An efficient solution to the simultaneous localization and mapping problem with unknown data association. Journal of Machine Learning Research (2004) To appear.
9. Christy, S., Horaud, R.: Euclidean shape and motion from multiple perspective views by affine iterations. Pattern Analysis and Machine Intelligence **18**(11) (1996) 1098–1104
10. Han, M., Kanade, T.: Multiple motion scene reconstruction from uncalibrated views. In: International Conference on Computer Vision (ICCV). Volume 1. (2001) 163–170
11. Tang, W.K., Hung, Y.S.: A subspace method for projective reconstruction from multiple images with missing data. Image and Vision Computing **24**(5) (2006) 515–524
12. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110
13. Han, M., Kanade, T.: Scene reconstruction from multiple uncalibrated views. Technical Report CMU-RI-TR-00-09, The Robotics Institute, Carnegie Mellon University (2000)
14. British Broadcasting Corporation: Housestead's Fort (3D model) (2004) VRML model available at `http://www.bbc.co.uk/history/interactive/virtual_tours/roman_housesteads/fort.wrl`, last retrieved May 2007.
15. Bouguet, J.Y.: Camera calibration toolbox for Matlab (2007) Matlab software available at `http://www.vision.caltech.edu/bouguetj/calib_doc/`, last retrieved May 2007.
16. Beis, J.S., Lowe, D.G.: Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In: Computer Vision and Pattern Recognition (CVPR), IEEE (1997)
17. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. 2nd edn. Cambridge University Press (2003)
18. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24** (1981) 381–395
19. Intel Corporation: OpenCV Computer Vision Library, version 1.0.0 (2006) C++ software available at `http://sourceforge.net/projects/opencv/`, last retrieved May 2007.
20. Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: A factorization method. International Journal of Computer Vision **9**(2) (1992) 137–154
21. Hess, R.: SIFT feature detector (2006) C software available at `http://web.engr.oregonstate.edu/~hess/index.html`, last retrieved May 2007.