Selecting Stable Image Features for Robot Localization Using Stereo*

James J. Little, Jiping Lu, and Don R. Murray
Department of Computer Science
The University of British Columbia
Vancouver, BC Canada, V6T 1Z4

Abstract

To navigate and recognize where it is, a mobile robot must be able to identify its current location. In an unknown initial position, a robot needs to refer to its environment to determine its location in an external coordinate system. Even with a known initial position, drift in odometry causes the estimated position to deviate from the correct position, requiring correction.

We show how to find landmarks without models. We use dense stereo data from our mobile robot's trinocular system to discover image regions that will be stable over widely differing viewpoints. We find image brightness "corners" in images and select those that do not straddle depth discontinuities in the stereo depth data. Selecting corners only in regions of nearly planar stereo data results in landmarks that can be seen in images taken from different viewpoints.

1 Introduction

Localization is the problem of identifying the current location of a mobile robot. Given a known initial position and perfect odometry, the problem is a simple matter of calculation. However, in an unknown initial position, a robot needs to refer to its environment to determine its location in an external coordinate system. Even with a known initial position, drift in odometry causes the estimated position to deviate from the correct position, requiring correction.

One solution is to supply landmarks, distinctive locations with identifiable appearances, and then provide a sensing method to identify the landmarks and localize the robot with respect to them. This engineering solution is prohibitively expensive except where its cost can justified, as in hospitals or factories[Nickerson93].

One can also search the local environment for visually distinctive locations and record these locations and a method to identify them (usually their visual appearance) as landmarks. This obviates the need to instrument the world and is preferable. The robot can acquire images and derive image descriptions that permit it to localize itself[OHD97]. One limitation of such an approach is that the appearance of surfaces varies significantly with differing viewpoints and is invariant only under restrictive assumptions. Other systems concentrate on known types of visual events, finding vertical lines associated with doors, or using range sensing, such as laser stripe systems, sonar, or active stereo vision to find locally salient geometric locations, such as corners, door joists, or pillar/floor junctions. We plan to learn the useful, stable, observable points (surface patches) from our 3D sensing and use them as landmarks.

Our current visually guided robot, *José*, maps its environment using real-time trinocular stereo from a camera fixed in the direction of motion. *José* integrates in a simple manner the recent stereo data with previous data in a 2D occupancy grid, used for navigation and obstacle avoidance. Features found in such maps are often used as landmarks for navigation. In our work, we combine the dense 2D occupancy grid with sparse 3D landmarks. Geometric information derived from stereo varies predictably with viewpoint, and can be a rich descriptor of the scene.

A common solution for defining both the structure of the environment and distinctive locations is to track "corners", local 2D image features, over long sequences of image frames[ST94], as the corners undergo incremental motion in the image. The 3D location of the corner points and the motion of the sensor can be determined[TK91].

On our mobile robot, maintaining corner points in the field of view conflicts with the requirements of navigation and obstacle avoidance. But we want to engage in ongoing activities whose main point is not tracking,

*This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Networks of Centres of Excellence Institute for Robotics and Intelligent Systems.

and which have high computational requirements. Alternatively, the robot could store the images for later processing, but such storage is prohibitive.

We must then process images acquired under large relative motion. In such images, the structure of the scene and the relative motion can be recovered. [McR96] uses differential invariants, local image descriptors composed of derivatives of the smoothed local brightness function, to describe image points so that the disturbance due to variation over viewpoint is minimized. However, accidental alignments[Low87] cause the projection of edges in the scene to create corners that are not stable under varying viewpoint.

Stereo depth data provides geometric information that lets us eliminate corners from consideration that arise from accidental alignments. We retain only those corners that lie on a relatively flat (almost planar) surface. Moreover, the local surface normal is estimated from the stereo data. With the 3D position information, the normal lets us predict the appearance of the corner point from another viewpoint, and lets us easily match it with corner points in the new image. The resulting system should provide distinctive landmarks visible over a wide range of viewpoints.

Landmarks can provide the skeleton of a topological basis for coordinate system in which several robots or sensors can operate. When two sensors observe a sufficient set of landmarks, they can derive, using their own interior camera orientation, their exterior orientation. This permits them to describe their observations in a common coordinate system.

Two such robots might identify a location to each other by transmitting the local subimage around its point of attention. The other could reconstruct the image patch as seen from its point of view, using view interpolation[SD95]. In our robots, this is greatly simplified since they have access to surface structure as well as appearance, using stereo. The robots can then predict the local surface appearance, given its position and orientation from stereo.

The system described here is a first step towards an automatic landmark acquisition system based on stereo depth data and brightness image descriptions. Our initial goal was to use purely geometric information derived from stereo, but our stereo algorithm, described later, lacks the resolution to provide such geometric features. The present system takes advantage of the strengths of the dense stereo data it does produce. A contour-based system [Lan98] may provide precise data permitting a purely geometric system in the future.

Figure 1: *José*, the mobile robot

1.1 Landmark-based Localization

Borthwick and Durrant-Whyte [BDW94] base their system on detecting corner and edge features in 2D, with no a priori knowledge of features or map necessary. It uses Extended Kalman Filtering [MKS89] to estimate both feature locations and robot position. The features assume a rectilinear world.

Weckesser *et al.* [WDEH95] use *a priori* landmarks at known positions in the global coordinate frame and particular models for landmarks (such as door jambs). Their system uses active stereo cameras and can effectively solve for pose of robot with respect to landmarks. The drawback is the need for models and prior knowledge of world.

Thrun and Bucken [TB96] have a system based on sensing regular landmarks (ie: not distinctive landmarks) (e.g., overhead lights, doorways). It uses a Bayesian approach implemented in neural nets, and learns which landmarks are most salient.

In the next section we describe our mobile robot and its trinocular stereo system. We explain how we determine the local surface geometry in the scene and we show how we find "corners" in images.

2 Architecture

2.1 Mobile robot: José

We used a Real World Interfaces (RWI)[1] B-14 mobile robot to conduct our experiments in vision-based robot navigation. *José* is equipped with a Pentium^{TM} PC running the Linux operating system as its onboard processing. The use of a Unix operating system allows a multi-threaded, flexible software architecture, with short development time, compared to an embedded solution. This robot is a significant improvement over *Spinoza*, our other mobile robot that was reported in [TSM+97]. *Spinoza* used embedded processors exclusively and, although it a powerful system, it proved to be a difficult development environment. However, with *José*, we have successfully built on the stereo vision and mobile robot navigation pioneered on *Spinoza*.

José is equipped with an Aironet ethernet radio modem. This allows communication to a host computer also equipped with a radio modem, just as over an ethernet network. Communication is achieved through the Unix socket communication primitives. *José* is also equipped with a Matrox Meteor RGB frame grabber board connected to a Triclops trinocular stereo vision camera module. This provides the stereo images used in navigating through the environment.

2.2 Trinocular Stereo Vision Rig

The stereo vision module used is the Triclops stereo vision system that was developed at the UBC Laboratory for Computational Intelligence (LCI) and is being marketed by Point Grey Research, Inc.[2] The stereo vision module has 3 identical wide angle (90° degrees field-of-view) cameras. The cameras are autogain, which makes them robust to changing lighting conditions.

The system is calibrated using Tsai's approach [Tsa87]. Correction for lens distortion, as well as misalignment of the cameras, is performed in software to yield three corrected images. These corrected images conform to a pinhole camera model, and have square pixels. The camera coordinate frames are co-planar and aligned so that the the epipolar lines of the camera pairs lie along the rows and columns of the images. This simplifies the stereo matching algorithm and improves its computational speed. The real-time dense stereo operates at approximately 10Hz.

[1] *www.rwii.com*
[2] *www.ptgrey.com*

Figure 2: An occupancy grid map built using dense stereo vision.

Since the camera lenses are fixed (with no possible changes in zoom or focus) and the cameras are bolted to a stiff back plane, the camera module will stay in calibration unless subjected to severe shocks. The final corrected images have an accuracy of 0.5 pixels RMS error for 640x480 images.

3 Method

3.1 Trinocular Stereo Approach

The trinocular stereo approach is based on the multi-baseline stereo developed by Okutomi and Kanade [OK93]. For each pixel in the reference image, the correlation values of the two image pairs (left/right and top/bottom) are summed to yield a combined score. The correlation measure used is the sum of absolute differences (SAD). The disparity data are interpolated to subpixel accuracy.

3.2 2D Navigation

We have developed an occupancy grid approach to robot navigation using stereo vision [MJ97]. This allows the robot to make 2D maps and safely navigate in an controlled indoor environment. Figure 2 shows an occupancy grid derived from stereo data.

Although this approach has been quite successful, it lacks consistency with an external coordinate frame. The map is only locally accurate, as over time the odometry errors become considerable. It is now apparent that in order to develop a robust and re-usable

left right

Figure 3: Left and right images taken with the Triclops system on *José*.

mapping system, some form of self-localization must be implemented to allow the robot to correct its position within an external coordinate frame.

3.3 Finding Corners

The corner-finding methods of [HP88] deliver corners that can be tracked over a sequence of images. A corner is detected by filtering the image with a simple mask that detects brightness variation in two (nearly) orthogonal directions.

We are using KLT, an implementation of the Kanade-Lucas-Tomasi feature tracker, provided by Stan Birchfield at Stanford University. It is based on the early work of Lucas and Kanade [LK81], and refined by Shi and Tomasi[ST94]. However, we do not use the tracker, only the corner locator. Its notion of a corner is essentially one where the local eigenvalues of the image Hessian are both significantly non-zero.

3.4 Fitting Planes

We convert the disparities, using calibrated measures, into depth measurements. At each point in the resulting depth image, we fit a plane using all valid depth measurements in surrounding square region (the size is determined by a system parameter). The fit finds $\vec{n}$ and c to minimize the sum of the normal errors (the distance of points in the surface normal direction of the fit plane), a total least-square problem [Gv79],

$$min \sum \vec{n} \cdot \vec{p} - c$$

where $\vec{n}$ is the unit surface normal, p a surface point, and c is the distance of the plane from the origin. Consequently, our fit is coordinate system independent. Thus we can compare results from differing frames.

The adequacy of the fit is measured by the sum of the absolute value of the normal error relative to the fit plane. Any points where the total error is large are rejected. This allows us to eliminate locations in the image where there are depth discontinuities.

In the interest of efficiency, we can restrict the computation of the planar fit to corner locations.

Figure 4: Disparity result of trinocular stereo: brighter points are nearer.

4 Experiments

We have taken a series of rectified trinocular stereo images using *José*'s Triclops cameras. Figure 3 shows the left and right images taken by our mobile robot; we have omitted the third, top, image used in the trinocular stereo. We process the images to produce subpixel interpolated disparity maps. Figure 4 displays the stereo disparity data before conversion to depth data. We fit local planes to the depth data over 13×13 regions, and determine the goodness of the local fit using the sum of absolute normal error from the fitted plane. Figure 5 shows the total absolute normal error over the planar fit regions.

We find corners (Figure 6) using by the KLT system. We choose only those corner points where the local error is in the lowest 5%. These are almost planar regions with little depth change over the region. Only a small set of corners survive, which are good candidates for landmarks. Figure 7 shows the reference (right) image with only the corners in regions of good planar fit.

Figure 8 shows the reference (right) image for the second images with only the corners in regions of good planar fit. A substantial fraction of the corners have been eliminated by rejecting regions that straddle depth discontinuities. We processed ten images in similar fashion. For each image we visually

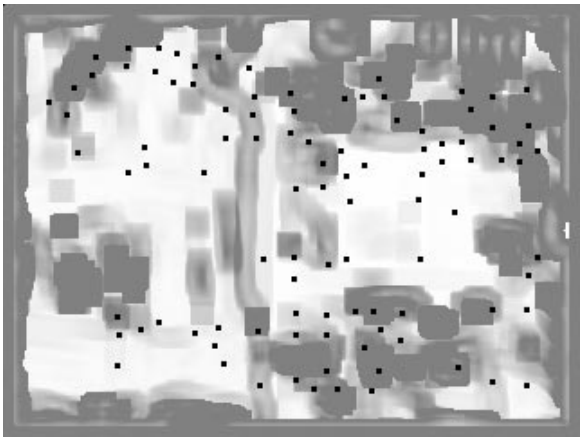

Figure 5: Fitting error (normal) for the planar fits; brighter is lower error, i.e., more like a plane. Corners are shown in black.

Figure 6: Corners (white) found by KLT superimposed on the right image.

Figure 7: Right image with corners (white) in planar regions.

Figure 8: A second image (right) with corners in planar regions.

Sequence	Without stereo	Using stereo
1	40	83.3
2	30	85.7
3	19	73.3
4	20	83.3
5	40	45.0
6	30	65.0
7	39	100.0
8	39	100.0
9	28	52.0
10	24	50.0

Table 1: Percent of corners on planar surfaces.

determined which corners were "stable", i.e., lying on planar patches. Table 1 shows the percentage of stable corners for the sequences. Without stereo, the average number of stable corners is 30.9%. Using stereo to eliminate corners formed by accidental alignments increases the number to 73.8%. A standard method for determining pose from point matches[ML96] can handle these better point features with little difficulty.

5 Summary and Conclusions

Dense stereo data can be analyzed to determine scene locations that are locally almost planar. We reject image points where the surface is not planar, so we can with confidence assume that the brightness corners result from surface markings on a nearly planar surface, not the accidental projection of brightness edges separated in space. These selected corners are easier to identify from widely differing viewpoints.

We show initial steps toward building a stereo landmark system that uses stereo data and corners. Selecting corners only in regions of nearly planar stereo data results in landmarks that can be seen in images taken from different viewpoints. The task of matching the corners and then determining pose is greatly simplified. Geometric information from surface orientation aids matching.

The long-range goal of this work is to accumulate such reliable scene points as landmarks, together with descriptions of the local image brightness and local scene geometry. Because of the local scene geometry, these landmarks will be reliable. By comparing the landmarks for distinctiveness, we can select a subset that work well for localization.

References

[BDW94] Stephen Borthwick and Durrant-Whyte. Simultaneous localisation and map building for autonomous guided vehicles. In *IROS-94*, pages 761–768, 1994.

[Nickerson93] S. B. Nickerson *et al.* Ark: Autonomous navigation of a mobile robot in a known environment. In *IAS-3*, pages 288–293, 1993.

[Gv79] G. H. Golub and C. F. van Loan. Total least squares. In A. Dold and B. Eckmann, editors, *Lexture Notes in Mathematics*, pages 69–76. Springer-Verlag, 1979.

[HP88] C. G. Harris and J. M. Pike. 3d positional integration from image sequences. *Image and Vision Computing*, 6:87–90, 1988.

[Lan98] Jochen Lang. Contour stereo matching with a constraint satisfaction approach. In *VI-98*, 1998.

[LK81] B. D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. 7th Int. Joint Conf. on Artificial Intelligence*, pages 674–679, Vancouver, 1981.

[Low87] D. G. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31:355–395, 1987.

[McR96] Daniel McReynolds. *Rigidity checking for matching 3D point correspondences under perspective matching.* PhD thesis, University of British Columbia, October 1996.

[MJ97] Don Murray and Cullen Jennings. Stereo vision based mapping for a mobile robot. In *Proc. IEEE Conf. on Robotics and Automation, 1997*, May 1997.

[MKS89] L. Matthies, T. Kanade, and R. Szeliski. Kalman filter-based algorithms for estimating depth from image sequences. *International Journal of Computer Vision*, 3:209–236, 1989.

[ML96] Daniel M. McReynolds and David G. Lowe. Rigidity checking of 3D point correspondences under perspective projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(12):1174–1185, December 1996.

[OHD97] S. Oore, G.E. Hinton, and G. Dudek. A mobile robot that learns its place. *Neural Computation*, 9(3):683–699, April 1997.

[OK93] M. Okutomi and T. Kanade. A multiple-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(4):353–363, 1993.

[SD95] Steven M. Seitz and Charles R. Dyer. Physically-valid view synthesis by image interpolation. In *IEEE Workshop on Representation of Visual Scenes*, 1995.

[ST94] Jianbo Shi and Carlo Tomasi. Good features to track. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition, 1994*, pages 593–600, 1994.

[TB96] Sebastian Thrun and Arno Bucken. Learning maps for indoor mobile robot navigation. Technical report, CMU-CS-96-121, April 1996.

[TK91] C. Tomasi and T. Kanade. Factoring Image Sequences into Shape and Motion. In *Proc. IEEE Workshop on Visual Motion, 1991*, pages 21–28. IEEE, 1991.

[Tsa87] R. Y. Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Journal of Robotics and Automation*, 3:323–344, 1987.

[TSM+97] Vladimir Tucakov, Michael Sahota, Don Murray, Alan Mackworth, Jim Little, Stewart Kingdon, Cullen Jennings, and Rod Barman. Spinoza: A stereoscopic visually guided mobile robot. In *Proceedings of the Thirteenth Annual Hawaii International Co nference of System Sciences*, pages 188–197, January 1997.

[WDEH95] P. Weckesser, R. Dillmann, M. Elbs, and S. Hampel. Multiple sensorprocessing for high-precision navigation and environmental modeling with a mobile robot. In *IROS-95*, 1995.