Proceedings of the 5th WSEAS Int. Conf. on Signal Processing, Computational Geometry & Artificial Vision, Malta, September 15-17, 2005 (pp176-181)

Mobile Robot Localisation with Stereo Vision

ALDO CUMANI and ANTONIO GUIDUCCI
Istituto Elettrotecnico Nazionale Galileo Ferraris
str. delle Cacce, 91 - I-10135 Torino
ITALY

Abstract: This work considers the problem of reducing the accumulated pose error in a grid-based SLAM system using a stereo vision sensor. It is shown that by periodically estimating the heading change by vision it is possible to recover most of the heading error with respect to dead reckoning, while 2D positional error can be efficiently recovered by map correlation. Experimental results confirm the validity of the approach.

Key-Words: Robot localisation, Mapping, Stereo vision, SLAM

1 Introduction

A critical factor for autonomous robot navigation in a partially or totally unknown environment is its ability to use suitable sensing devices for tracking its pose relative to the environment, while incrementally building a map of the environment itself. This *Simultaneous Localisation And Mapping* (SLAM) problem has therefore been a highly active research topic for more than a decade.

Existing approaches to SLAM differ either in the kind of sensing devices used (sonars, laser scanners, vision systems), in the way they treat sensor data, and finally in the way they represent the acquired knowledge. With regard to the latter, 2D occupancy grids [1, 2, 3] have been quite popular, at least for indoor applications, as they provide an easy way for fusing data from very different sensors.

As concerns the problem of translating sensor readings into pose data, most proposals are based on the Extended Kalman Filter (EKF) (e.g. [4, 5, 6, 7]). These approaches reformulate the problem in terms of estimating the state of the system (robot and landmark positions) given the robot's control inputs and sensor observations, which are assumed affected by Gaussian noises. Other stochastic approaches do not make the latter assumption but directly estimate distributions conditioned on sensor data, as e.g. the particle system proposed in [8].

Both EKF and particle-system based SLAM, however, need a model of robot motion and of sensor measurement. In contrast, there are approaches that can estimate directly the robot's egomotion from sensory data. This is the case e.g. when using vision sensors. Indeed, even using a single onboard camera, it has been shown [9] that localisation and map building can be achieved by standard Structure-from-Motion methods.

In any case, it must be remarked that the SLAM concept implies that map building should be *incremental*. However, estimates are typically affected by errors, and the latter have the bad habit of accumulating; this leads to the loop-closure problem, where the robot, on return to a previously visited place, thinks it's somewhere else. Though solutions to this problem have been proposed, it is nevertheless worth investigating methods able to reduce the localisation error at each step, so automatically reducing the accumulated global error as well.

In this paper we present some results from a SLAM algorithm which uses a stereo head, mounted on a pan-tilt unit, as sensor, and an occupancy grid to store the acquired map. At more or less regular intervals along its trajectory, the robot stops and "looks around", i.e. acquires a set of stereo pair images, covering a field of view of more than 180°, by panning its head. Point and line features extracted in the left and right image of each pair are then matched and their 3D estimated positions are used to build a local occupancy grid map.

Local maps are then merged into a global map after being registered to the global reference frame, using the current estimate of the robot pose. In this work we assume that an odometric estimate of robot pose is available; however, the long-term accuracy of the latter is often not sufficient. A possible way of tackling this problem consists in correcting the odometric estimate of the change in robot pose by cross-correlating the local map with the current global one [10, 11]. This method can yield very good results, but is quite computation-intensive, especially since the search space is three-dimensional (x, y, θ) and the search in θ requires rotation of the local map. Some speed-ups have been proposed [11], but in our experience they are not reliable enough.

Our proposal is to restrict correlation search to (x, y), while the heading change is estimated by registration of 3D point clouds obtained by vision during the robot motion. Indeed, 3D registration could be used on its own, as proposed in [12]; however, while this method yields a rather good estimate of rotation, the translation accuracy is often quite inferior to that obtainable by map correlation.

Combining the visual heading estimate with the translation estimate from map correlation, yields a good compromise between speed and accuracy, as confirmed by the experimental results in Sec. 5.

2 The algorithm

As said above, at regular intervals along its path the robot stops and acquires a set of stereo pair images, by panning its head (*panning stop*). Features extracted in the left and right image of each pair are then matched and their 3D estimated positions are used to build a local occupancy grid map. The reason for this behaviour is that the field of view of the stereo rig is rather limited, and using a single stereo pair cannot yield a local map with enough structure.

We assume that the robot starts with a panning sequence as described above, and the so obtained local map becomes the starting global map. The world reference frame (WF) is defined as the robot frame at its starting pose.

After each panning stop, a stereo pair is acquired, features are detected and left-right matched, and a 3D cloud of points is reconstructed and transformed to the WF using the current pose estimate. The robot then moves along its planned path towards the next panning stop. Along the path, stereo pairs are acquired at maximum allowable speed, and features are tracked. Every m (say $m = 3$) frames (i.e. at *key frames*), a 3D reconstruction is performed, and the so obtained point cloud is registered against the previous one, so allowing to get a visual estimate of the change in robot heading (which will generally be different from the one predicted by dead reckoning). After that, new features are also extracted and left-right matched; this prevents the number of tracked features to become too low, especially in the case of fast heading changes, which would prevent getting a reliable estimate of robot motion.

At the next panning position, the robot pose estimate is updated by taking into account both the odometric data and the accumulated visual corrections computed as above. A panoramic set of stereo pairs is grabbed, and 3D reconstructed points/lines from this set, after transformation into the WF, are used to build a local map. The latter is then registered against the global one by correlation in x and y, and used to upgrade the global map.

3 Features

3.1 Point features

The current implementation uses Shi-Tomasi features [13], i.e. small textured image patches, whose centers yield pointwise measurements. A significant advantage of Shi-Tomasi features is that their definition implicitly provides an efficient frame-to-frame tracking algorithm, provided that the image-plane displacement be small, i.e. well within the size of the patch. Using the same algorithm for stereo matching, where the displacement (disparity) may be rather large, especially for near objects, is still possible, provided that some coarse initial estimate of disparity be available. In our implementation, such an estimate is provided by a standard stereo correlation algorithm. Notice that the latter may be run on lower resolution images, so substantially lowering computational load.

Matched point pairs are then backprojected to a 3D point estimate, in the camera reference, using the method in [14].

3.2 Line features

Edge contours are extracted in both the left and right images using a standard second-directional-derivative method [15]. Contour lines are then segmented into quasi-rectilinear pieces, augmented with photometric attributes (namely, the average luminances on either side of the segment). Since we use a binocular stereo pair with horizontal baseline, horizontal or nearly horizontal segments cannot give reliable depth information, and are therefore discarded by imposing a threshold on the angle of the segment with the image y axis; the remaining segments are then left-right matched. This latter step is accomplished by first ordering the segments, in the left and right images, by increasing x coordinate of their midpoint and then performing, for each segment in the left image, a search for segments in the right image within the x range delimited by the minimum and maximum allowed disparity. For each putative match a score is computed, which takes into account both the similarity of photometric segment parameters and the fraction of overlap in the y direction, and the best scoring match is kept.

The same matching procedure is repeated in the opposite direction (right to left), and only consistent matches are kept. The y-overlapping portions of matched pairs are then backprojected so obtaining 3D segment estimates in the camera frame.

3.3 Tracking and motion estimation

The features detected at a key frame are tracked along the sequence, separately for left and right image features, up to the next key frame. At this point, a new 3D reconstruction is made from the tracked left/right features, and registered against the previous one in order to get an estimate of the robot motion between the two key frames.

At present, only point features are tracked and reconstructed (though both point and line features contribute to the map). As said above, the frame-to-frame tracking algorithm expects limited feature displacements between subsequent frames. This is seldom the case, especially when the robot is rotating. However, since each feature has attached to it an estimate of the corresponding 3D position relative to the robot, combining the latter with the known planned robot motion the image position of the feature in the new image can be predicted with sufficient accuracy to allow reliable tracking.

At this point, we have a set of N features $\mathcal{F}_i$, left-right matched and tracked from key frame k to the next one $k+1$, to which are attached pairs of 3D position estimates, namely $\mathbf{X'}_i$ from the initial reconstruction at key frame k and $\mathbf{X''}_i$ from the last one. An estimate of robot motion from k to $k+1$ is then obtained as the rototranslation $(\mathsf{R}_k, \mathbf{t}_k)$ that minimises a suitable fitting criterion

$$J = \sum_{i=1}^{N} f_i(\|\mathbf{d}_i\|^2)$$

with

$$\mathbf{d}_i = \mathbf{X}''_i - (\mathsf{R}\mathbf{X}'_i + \mathbf{t})$$

and the vertical component of the rotation R yields the visual estimate of heading change. With regard to the choice of fitting criterion, we are currently using a Lorentzian cost function, i.e.

$$f_i(\|\mathbf{d}_i\|^2) = \log(1 + \frac{\|\mathbf{d}_i\|^2}{\sigma_i^2})$$

which makes the estimate more robust against outliers. The σ_i allow to take into account the different accuracy of point estimates, e.g. as a function of the distance from the sensor. However, a more sound approach, which is currently under study, would be to perform a full bundle adjustment of point coordinates and relative robot pose, using as fitting criterion the image-plane backprojection error in the four images of the two stereo pairs.

4 Map building and updating

2D occupancy grids [1, 2, 3] are 2D metric maps of the robot's environment, where each grid cell contains a value representing the robot's subjective belief whether or not it can move to the center of the cell. Occupancy grids are a popular way of representing acquired geometrical evidence about the environment, as they allow easy integration of measurements from different sensor types.

Since vision yields full 3D measurements, however, it is possible to build a layered map, where each layer corresponds to some range of height above the ground plane. The map building approach used in our test is similar to the FLOG (Fuzzy Logic-based Occupancy Grid) approach proposed in [16]. For each grid cell (x, y) in layer l several fuzzy set membership functions are defined, namely $\mu_E(x, y, l)$ for the *empty* fuzzy set E_l, $\mu_O(x, y, l)$ for the *occupied* set O_l, plus a *confidence* measure $\mu_K(x, y, l)$. Adding a measurement (3D point) to the map is performed by suitably modifying the membership functions of cells traversed by the rays going from the stereo head to the estimated 3D points. As concerns line features, each estimated 3D segment is subdivided into parts of predetermined image-plane length, and the midpoint of each part is treated as a point measurement.

Depending upon the purpose, a synthetic map M can be defined as a suitable combination of E_l, O_l and K_l [17]. In particular, for the purpose of correlating local/global maps, we adopt the following definition:

$$M = \cup_l (O_l \cap K_l)$$

Each local map is built after transforming the 3D points into the WF, using the best current estimate of the robot pose, which incorporates the visual heading correction. A new estimate of robot pose is then obtained by searching for a maximum of the correlation between the membership functions $\mu_{M_L}(x, y)$ and $\mu_{M_G}(x, y)$ of the local and global synthetic maps M_L and M_G as defined by the equation above.

At this point, the global map is updated by weighted averaging with the registered local one as in [16].

5 Experimental results

This section presents some results obtained by processing sequences of images acquired with our ActivMedia Pioneer 3-DX robot, equipped with a Videre Design STH-MDCS stereo head (Fig. 1). The latter is a low-cost commercial product nominally capable of yielding pairs of 1280×960 colour images at 7.5 fps, or lower resolution images (640×480 and 320×240) at higher speeds, up to 30 fps. A serious limitation of this device is its small stereo baseline (88 mm, non-adjustable).

In the experiment described here, the robot wandered through a large (about 14m$\times$9m) laboratory

room with several desks and various instruments (see Figs. 2 and 3). Along its trajectory (about 60m total), the robot stopped for looking around either after covering a path length of 1 m, or after a heading change of 45°. About 7000 stereo pairs were acquired at 640×480 resolution.

Fig. 4a shows the final global map built using for the robot pose the raw odometric data, without applying the correction proposed above. Comparing the latter with the CAD model of the environment shown in Fig. 2, it is evident that the odometric trajectory accumulates a rather large heading error (about 20°). As can be seen in Fig. 4b, not much is gained by correcting only the translation estimate by map correlation.

By contrast, Fig. 5a shows the result of the proposed approach. Fig. 5b shows that the result is not significantly improved by performing also an angular search for best correlation around the visual heading estimate. Fig. 6 shows the cumulative heading correction for the two latter cases, i.e. with and without angular search.

It must be remarked that all correlation computations are done with the robot stopped, while the visual heading algorithm runs during the robot's movement. Therefore, eliminating the need for angular correlation search in the registration step greatly reduces dead times.

Figure 1: The mobile robot with stereo head.

6 Concluding remarks

In this work we have considered the problem of reducing the accumulated pose error in a grid-based SLAM system using a stereo vision sensor. It has been shown that periodically estimating the heading change by vision it is possible to recover most of the heading error with respect to dead reckoning, while 2D positional error can be efficiently recovered by map correlation. Further work on this topic is needed, particularly for improving the visual heading algorithm, e.g. by using bundle adjustment techniques.

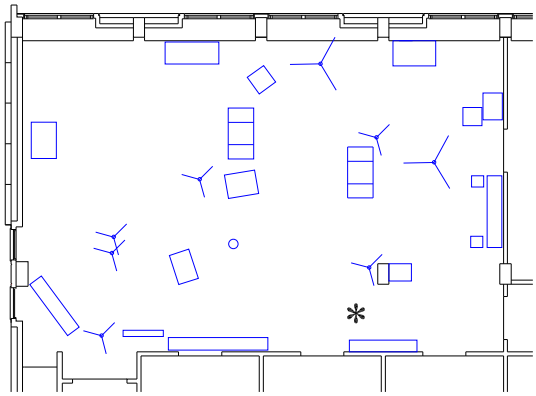

Figure 2: Approximate CAD model of the environment. The asterisk marks the point of view of the panorama in Fig. 3.

References:

[1] H. P. Moravec, "Sensor fusion in certainty grids for mobile robots," *AI Magazine*, vol. 9, no. 2, pp. 61–74, 1988.

[2] M. C. Martin and H. P. Moravec, "Robot evidence grids," Tech. Rep. CMU-RI-TR-96-06, Carnegie Mellon University, 1996.

[3] S. Thrun, "Learning metric-topological maps for indoor mobile robot navigation," *Artificial Intelligence*, vol. 99, no. 1, p. 21, 1998.

[4] R. Smith, M. Self, and P. Cheeseman, "Estimating uncertain spatial relationships in robotics," in *Autonomous Robot Vehicles* (I. Cox and G. Wilfong, eds.), pp. 167–193, Springer-Verlag, 1990.

[5] J. A. Castellanos, J. M. M. Montiel, J. Neira, and J. D. Tardós, "The SPmap: a probabilistic framework for simultaneous localization and map building," *IEEE Trans. Robotics and Automation*, vol. 15, no. 5, pp. 948–953, 1999.

[6] J. J. Leonard and H. J. S. Feder, "A computationally efficient method for large-scale concurrent mapping and localization," in *Proceedings of the 9th International Symposium on Robotics Research*, 1999.

[7] A. J. Davison, "Real-time simultaneous localisation and mapping with a single camera," in *Proceedings of the 9th International Conference on Computer Vision, Nice*, 2003.

Figure 3: Panorama of the robot environment.

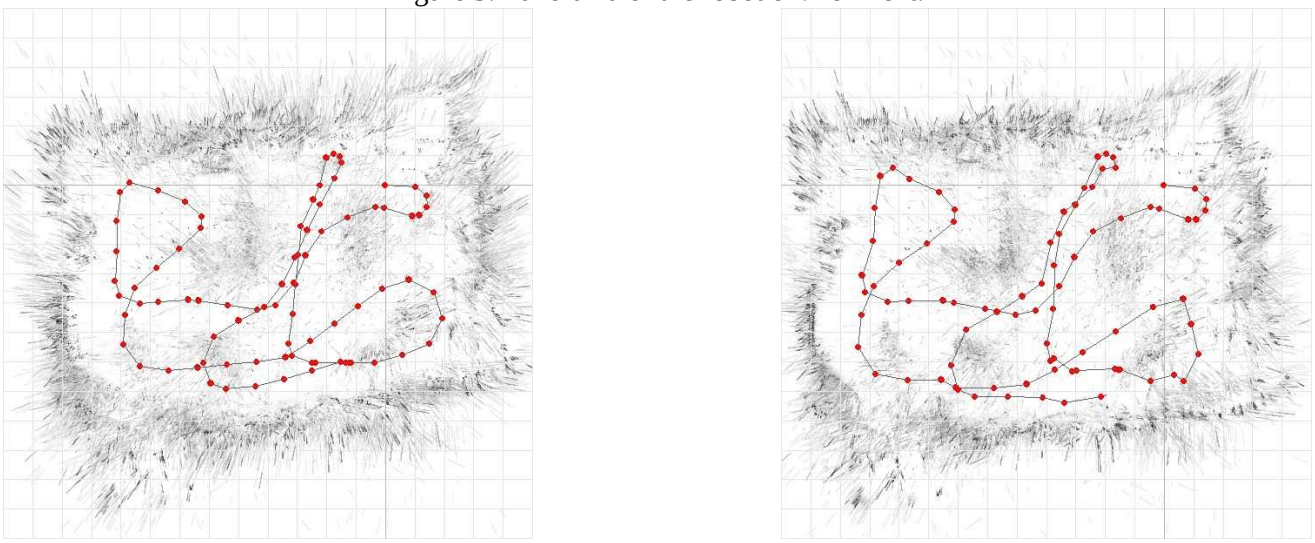

Figure 4: Map obtained: (left) by dead reckoning, (right) by correlation registration in x and y only.

[8] S. Thrun, W. Burgard, and D. Fox, "A real-time algorithm for mobile robot mapping with applications to multi-robot and 3D mapping," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2000.

[9] A. Cumani, S. Denasi, A. Guiducci, and G. Quaglia, "Integration of visual cues for mobile robot localization and map building," in *Measurement and Control in Robotics* (M. A. Armada, P. Gonzales de Santos, and S. Tachi, eds.), Instituto de Automatica Industrial, Madrid, 2003.

[10] S. Thrun, "Learning maps for indoor mobile robot navigation," *Artificial Intelligence*, 1997.

[11] A. C. Schultz and W. Adams, "Continuous localization using evidence grids.," in *ICRA*, pp. 2833–2839, 1998.

[12] A. Cumani, S. Denasi, A. Guiducci, and G. Quaglia, "Robot localisation and mapping with stereo vision," *WSEAS Transactions on Circuits and Systems*, vol. 3, no. 10, pp. 2116–2121, 2004.

[13] J. Shi and C. Tomasi, "Good features to track," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 593–600, 1994.

[14] R. Hartley and P. Sturm, "Triangulation," in *Proc. ARPA Image Understanding Workshop, Monterey*, pp. 957–966, 1994.

[15] P. Grattoni and A. Guiducci, "Contour coding for image description," *Pattern Recognition Letters*, vol. 11, no. 2, pp. 95–105, 1990.

[16] H. Hirschmuller, "Real-time map building from a stereo camera under unconstrained 3d motion," Faculty Research Conference, De Montfort University, Leicester, 2003.

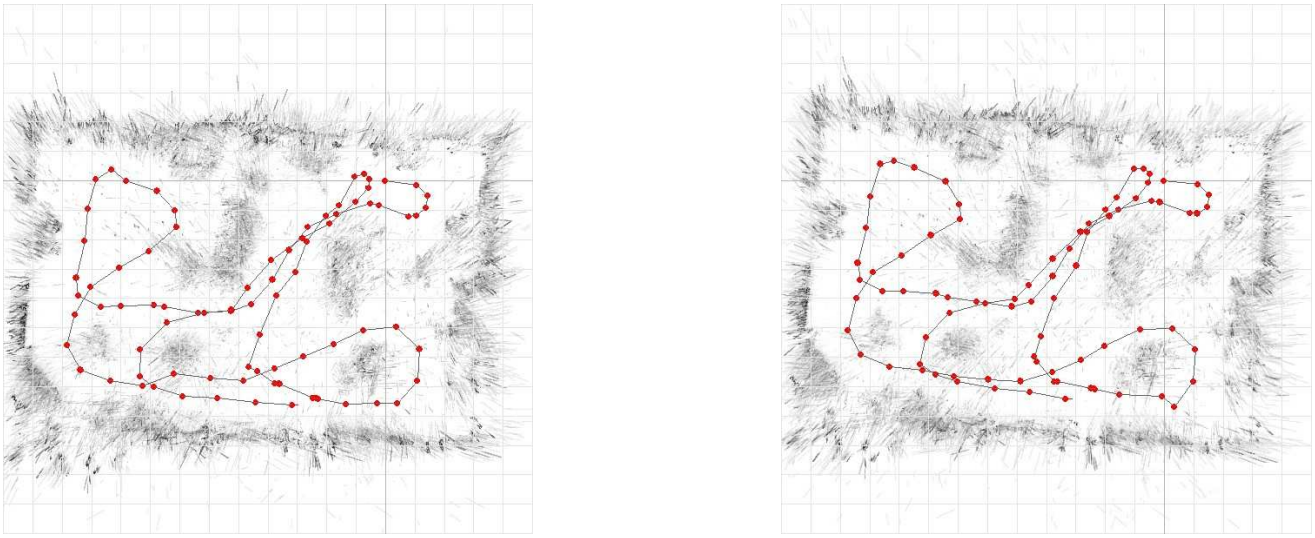

Figure 5: Map obtained: (left) by correlation registration in x and y, with the proposed visual heading correction, (right) with additional heading correction by angular correlation search.

[17] G. Oriolo, G. Ulivi, and M. Vendittelli, "Real-time map building and navigation for autonomous robots in unknown environments," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 28, no. 3, pp. 316–333, 1998.

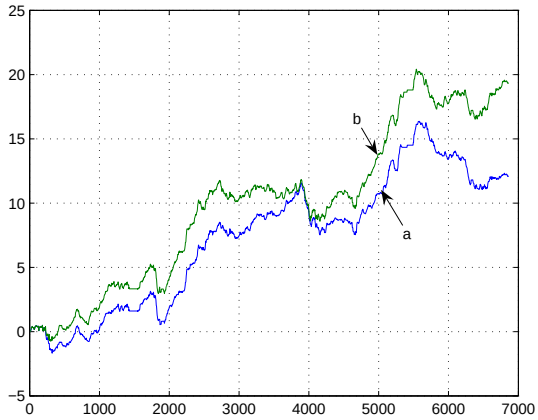

Figure 6: Cumulative heading correction without (a) and with (b) angular search.