Robot Localisation and Mapping with Stereo Vision

A. CUMANI, S. DENASI, A. GUIDUCCI, G. QUAGLIA
Istituto Elettrotecnico Nazionale Galileo Ferraris
str. delle Cacce, 91 - I-10135 Torino
ITALY

Abstract: This paper presents an approach to Simultaneous Localization and Mapping (SLAM) based on stereo vision. Standard stereo techniques are used to estimate 3D scene structure (point clouds). Point clouds at subsequent times are registered to estimate robot motion, and used to build a global environment map. Preliminary experimental results are also presented and discussed.

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

1 Introduction

Simultaneous Localisation And Mapping (SLAM), also known as Concurrent Mapping and Localisation (CML), is the process by which an autonomous mobile robot can track its pose relative to its environment, while incrementally building a map of the environment itself. SLAM is clearly a critical factor for successful navigation in a partially or totally unknown environment, and has therefore been a highly active research topic for more than a decade.

Many of the existing approaches to SLAM are based on the Extended Kalman Filter (EKF) (e.g. [1, 2, 3, 4]). These approaches assume that the robot is moving within a stationary environment, providing some distinctive landmarks whose position relative to the robot can be measured by some sensor (typically, sonar rings or laser scanners). The problem is then reformulated in terms of estimating the state of the system (robot and landmark positions) given the robot's control inputs and sensor observations.

The EKF approach implies a stochastic model with independent Gaussian noises. Other probabilistic approaches do not make such assumption, e.g. the particle system proposed in [5], where the robot pose is estimated by maximizing its probability conditioned on past sensor data.

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 [6] that localisation and map building can be achieved by standard Structure-from-Motion methods.

Monocular vision has, however, the drawback that it cannot estimate absolute scale, a problem that can only be overcome either by knowing the size/position of some recognizable predetermined landmarks, or relying on odometry. By contrast, using a calibrated stereo head the robot can determine metric 3D structure, including absolute scale, from vision data alone.

This paper presents some results from a stereo-vision based approach to SLAM we are currently developing. From each image pair of the sequence acquired while the robot is moving, features are extracted to serve as visual landmarks. Such features are both left-right matched and tracked along the sequence. Stereo matches allow to determine 3D structure as a point cloud. Clouds generated at different times are then recursively registered in order to obtain an estimate of robot pose and of 3D point coordinates in a global reference. The latter are used to build a 2D occupancy grid map. The paper also presents some preliminary results obtained with our ActivMedia P3-DX mobile robot equipped with a Videre Design STH-MDCS stereo head.

2 SLAM by stereo vision

Our approach can be summarised as follows:

- extraction of features from the image pairs and left-right matching, allowing to determine local 3D structure as a cloud of 3D points;
- tracking of the features along the sequence and registration of point clouds from subsequent image pairs into a same global reference, allowing to determine robot motion;
- building of a global occupancy grid map from the visual measurements;
- possible correction of accumulated errors by comparison of submaps.

2.1 Features

From each frame of the image sequence acquired while the robot is moving, various kinds of features can be extracted to serve as visual landmarks: corners, edge segments, textured patches etc. Our current implementation uses Shi-Tomasi features [7], i.e. small textured image patches, whose centers yield pointwise measurements useful for motion/structure estimation. A significant advantage of Shi-Tomasi features is that their definition implicitly provides an efficient frame-to-frame tracking algorithm; other approaches may require independent feature extraction from each image and a costly search for matching pairs. The same algorithm can be used for left-right matching as well, and moreover, since the tracking algorithm allows for affine distortion, such features can be successfully followed over large relative displacements in the image.

Features are extracted in the first frame of the sequence, and thereafter at more or less regular intervals; the frames from which new features are extracted are called *key frames*. The spacing between key frames is chosen as a compromise between the required frequency of pose/map updating on one side, and the computational load on the other, since most computations are done at key frames.

2.2 Stereo

At every key frame, new features are extracted from the left image, possibly keeping old features tracked from the previous frame, and matched against the right image. Shi-Tomasi features provide excellent localisation accuracy, provided that the centre of the matched area lies within the window defining the template feature. With stereo, this may not be the case for features corresponding to near objects, which exhibit a larger stereo disparity. For this reason, a rough estimate of feature disparity is computed from a dense disparity map over the whole image, obtained by a standard correlation method; this coarse estimate is used to predict the right image position of each feature so that the Shi-Tomasi algorithm can be safely applied. It must be noticed that, since only a rough estimate of disparity is needed, the map can be computed on a lower (e.g. half) resolution image, so substantially reducing computational load.

From stereo matches, an estimate of the 3D positions of the features is then obtained by triangulation, using the approach by [8], which minimises the image plane error between observed features and back-projected points. The result of this step is a cloud of 3D points in the reference frame of the stereo head.

2.3 Tracking and registration

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.

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

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

With regard to the choice of fitting criterion, it must first be observed that the errors $\mathbf{d}_i$ cannot be equally weighted, as estimates of points farther away have much greater uncertainty (the variance grows roughly with the square of distance from the stereo head). Moreover, the unavoidable presence of many *outliers* in the sample (e.g. false matches) makes the sample deviate considerably from the Gaussian error assumption that could justify the use of a standard sum-of-squares cost function. We have empirically found, instead, that satisfactory results can be obtained using a Lorentzian cost, i.e.

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

where the σ_i take into account the aforementioned dependence of the uncertainty on the distance from the sensor.

2.4 Map building

When the robot motion is constrained to be planar, as in a typical indoor environment, the 3D measures obtained from the vision algorithm can be used to build

a 2D occupancy grid map [9, 10, 11]. The latter is a 2D metric map 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. Grid maps are a convenient way of representing the global structure of the environment; matching a local map with the previously built global map allows an easy estimation of the robot location. Moreover, grid maps allow easy integration of measurements from different sensor types.

Since vision yields full 3D measurements, however, it is possible to build a 3D grid map by layering 2D maps, where each layer corresponds to some range of height above the ground plane. The map building approach used in our test follows the FLOG (Fuzzy Logic-based Occupancy Grid) approach proposed in [12]. In this approach, for each grid cell C several fuzzy set membership functions are defined, namely $\mu_E(C)$ for the *empty* fuzzy set E, $\mu_O(C)$ for the *occupied* set O, plus a *confidence* measure $\mu_K(C)$. Map updating is performed by suitably modifying the membership functions of cells traversed by the rays going from the stereo head to the estimated 3D points. A map value for each cell is then obtained by a suitable combination of μ_E, μ_O and μ_K.

Map updating is performed at every key frame, by first building a "local" occupancy map, i.e. a map which, although built in the global reference frame, only uses point data accumulated since the last update, and then adding it to the global map.

3 Experimental results

3.1 Simulated navigation

We did some preliminary experiments by simulating robot navigation with our Samsung AW1 robotic arm. The latter is a lightweight 6-DOF manipulator, whose end effector carried a Sony XC55 progressive camera yielding non-interlaced B/W images at a resolution of 640×480, 30 fps. Stereo was simulated by acquiring, at each planned position, a pair of images laterally shifted. Due to the limited range of arm motion, both the stereo baseline and the environment size were suitably scaled down, namely using 20mm for the baseline and keeping the camera at a distance of the order of 1 m from the observed objects.

Fig. 1 shows a frame (left image) from a sequence, and Fig. 2 the corresponding disparity map. Fig. 3 displays a top view of the final estimated trajectory and point cloud. Robot pose is indicated by a dot for position and a segment for heading. Fig. 4 shows the corresponding final occupancy map. These figures are to be compared against the model of the scene shown in Fig. 5 together with the planned trajectory (note that the reconstruction reference frame is aligned with the initial robot position, in the left hand corner of Fig. 5).

Figure 1: A (left) image from the simulated navigation.

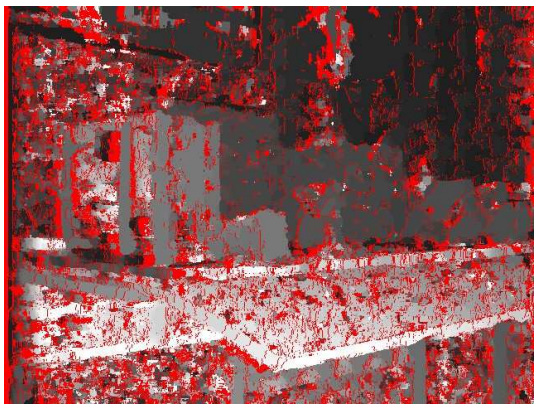

Figure 2: Dense disparity map.

3.2 Mobile robot

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. 6). 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). Since the error in distance estimation increases quadratically with the ratio distance/baseline,

In the experiment described here, the robot was programmed to follow a trajectory, comprising several arcs and a short straight segment, through a large (about 14m$\times$9m) laboratory room with several desks and various instruments (see Fig. 7). About 1500

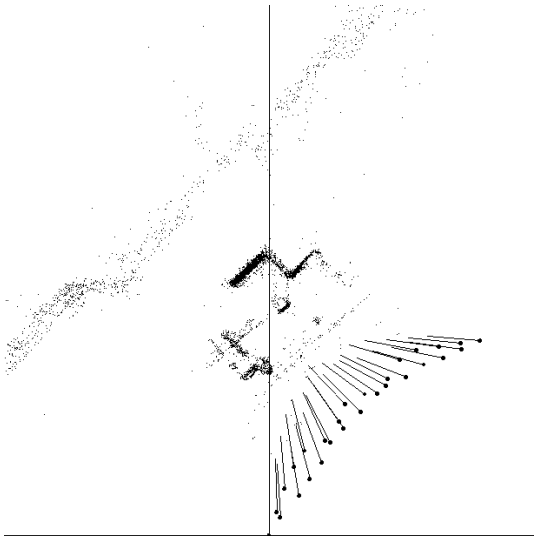

Figure 3: Top view of the estimated trajectory and point cloud. Robot pose is indicated by a dot for position and a segment for heading.

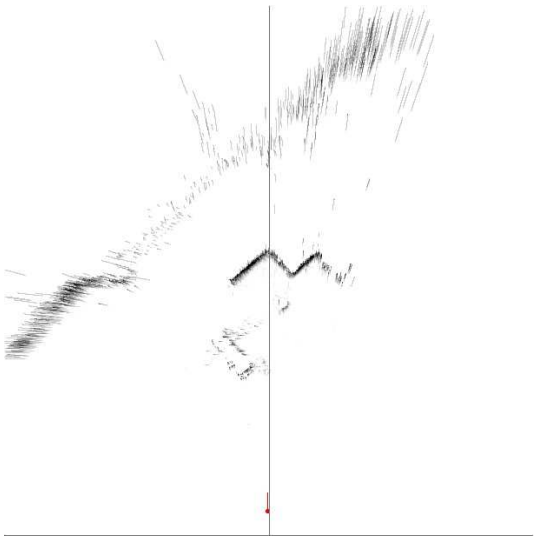

Figure 4: Final occupancy map.

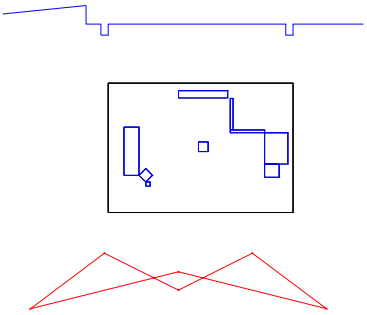

Figure 5: Model of the scene and planned trajectory.

stereo pairs were acquired at 640×480 resolution and a frame rate of about 7.5 fps.

Fig. 8 shows the final trajectory estimate and point cloud, while Fig. 9 displays the final global occupancy map. Note that many of the points visible in Fig. 8 at the center of the room are actually either reflections from the shiny floor, or feature points from the ceiling, and do not contribute to the map (which is restricted to points between 0m and 2m height).

From these pictures it can be observed that stereovision based SLAM compares favorably with methods based on different kinds of sensors. As all incremental methods, however, also this one is affected by the loop-closing problem caused by error accumulation, as indicated by the doubling of the upper wall in Figs. 8 and 9. In a previous work [6] the authors proposed the use of reference visual landmarks, stored by the robot during its navigation, which in case of loop closing can be used to estimate the accumulated error and to modify the estimated trajectory and map by back-propagating the correction.

However, the estimated shape of the environment appears rather good for what concerns wall angles, indicating a good accuracy in the estimate of changes in the robot heading. Indeed, the most critical part in the trajectory estimation lies in the determination of the robot displacement; the latter depends upon the estimates of feature distances, which are inaccurate at larger distances. This inaccuracy can only be reduced by increasing the size of the stereo base, which is impractical beyond a certain limit (and was not even possible in our setup), and/or increasing the accuracy of the calibration of the stereo head (which also implies the use of better and more costly lenses).

Another possible way of reducing the accumulated error in the estimate of robot motion, that we are currently implementing and testing, consists in correcting

Figure 6: The mobile robot with stereo head.

Figure 7: A (left) image from the actual navigation sequence.

the estimated robot displacement by registering the local map against the stored global map, and adding the former to the global map only after this registration step. Due to the limited camera field of view, the local map may not have enough structure for a reliable registration; this drawback can be overcome, if the stereo head is mounted on a pan-tilt unit, by "looking around", i.e. by combining several views of the environment from the same robot position into a more informative local map.

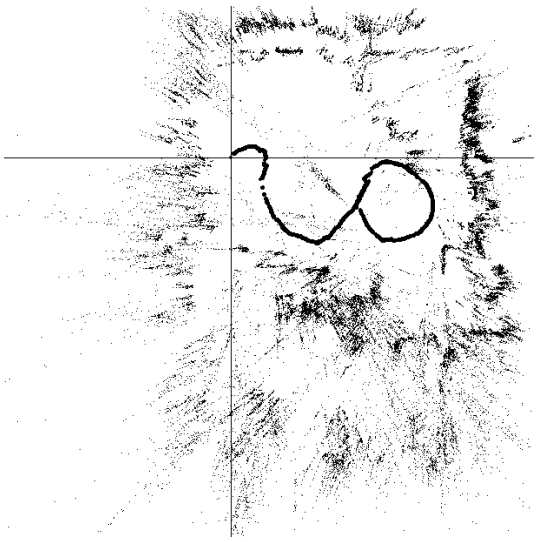

Figure 8: Top view of the estimated trajectory and point cloud for the navigation sequence.

4 Conclusion

The above results indicate that stereo vision can be used profitably for SLAM. With respect to other kinds of sensors, a stereo head has the advantage of providing direct 3D measurements without the need to make an explicit model of the robot and its environment, and without relying on accurate odometry. Moreover, vision yields a much greater quantity of information about the environment, that can be used e.g. to recognize previously visited places in order to disambiguate localisation. On the other hand, the accuracy obtainable using off-the shelf components is not as satisfactory as one could expect, and the computational burden is still rather large even using up-to-date processing technology. More work is therefore needed in this respect.

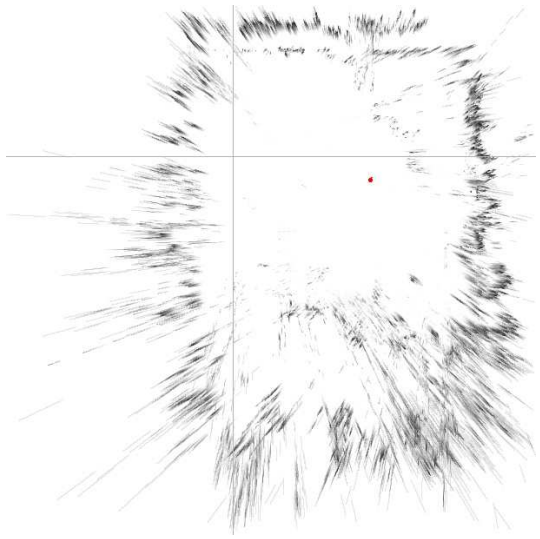

Figure 9: Final occupancy map for the actual navigation sequence.

Acknowledgment

The authors wish to thank Fabrizio Abrate for his invaluable help in programming robot motion and image acquisition.

References:

[1] 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.

[2] 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.

[3] J. J. Leonard and H. J. S. Feder, "A computationally efficient method for large-scale concur-

rent mapping and localization," in *Proceedings of the 9th International Symposium on Robotics Research*, 1999.

[4] 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.

[5] 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.

[6] 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.

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

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

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

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

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

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