Towards Monocular On-Line 3D Reconstruction

Pekka Paalanen*, Ville Kyrki**, and Joni-Kristian Kamarainen

Machine Vision and Pattern Recognition Research Group
Lappeenranta University of Technology, Finland

Abstract. The use of visual sensing for action generation in unknown environments is an attractive option due to the great representation power of vision, but it is challenging for two reasons. The representations used in vision are often not well suitable for planning, thus requiring complex learning approaches. Furthermore, an active agent needs to make decisions on-line, without the delay of off-line processing.
This paper proposes to combine monocular visual SLAM with dense visual reconstruction techniques in order to build geometrically correct three-dimensional models, which can be used for action generation, such as path or grasp planning in a robotic system. We propose a vision only monocular solution which will run on-line on commodity hardware. The problem is very challenging and currently no complete solutions exist, though similar off-line systems are quite mature.

1 Introduction

The use of visual sensing for action generation in uncertain environments is continuously increasing in robotics, as vision is a powerful medium, able to convey a large variety of information. However, the use of vision in generating agent actions is challenging, especially for two reasons: First, the visual representations are often appearance based and sparse. For example, in visual SLAM the environment is represented as a sparse set of landmarks, which is useful for localisation but is difficult to use for path planning. Second challenge in the use of vision is the real-time requirement.

The visual representations used for action planning can in general be divided into appearance-based and geometric. The first pure appearance-based approaches have been recently proposed (e.g., [1]), but their use is still in its infancy, since the mapping from appearance to action is usually very complex. In contrast, geometric approaches with 3D reconstructions can often be readily used with existing planning algorithms for both navigation and manipulation of objects. This paper concentrates on the latter approach, geometrically correct 3D reconstructions. Specifically, the building of 3D reconstructions is considered, as the planning using 3D models is quite well known in the literature.

In this work, we propose an on-line dense 3D reconstruction system for unstructured scenes using monocular vision, especially on commodity hardware.

* paalanen@lut.fi, supported by East Finland Graduate School ECSE.
** The support from Academy of Finland is gratefully acknowledged.

Visual reconstruction can be seen to consist of two distinct components. First, the trajectory of the camera is estimated, usually through sparse image correspondences. Second, dense correspondences are estimated to produce a textured 3D model. We propose a system which performs these both steps in near real-time, using visual SLAM for the on-line estimation of the camera trajectory. For the dense correspondences, we use the estimated camera poses to determine the epipolar geometry, which allows the building of disparity and depth maps. Every well-behaved pixel in the depth map contributes a point into the 3D model. The 3D reconstruction therefore is a coloured semi-dense point cloud.

To the authors' knowledge, the proposed system is the first to perform near real-time dense 3D reconstruction of an unstructured environment employing only monocular visual SLAM. The current implementation is not strictly real-time or parallelised, but shows that real-time operation with adequate reconstruction quality is possible with current commodity hardware and algorithms.

1.1 Related Work

Off-line systems for reconstructing a textured 3D-model are quite mature and accurate (e.g. [2, 3]), but deploy methods that cannot be applied on-line. For instance, bundle adjustment is a popular method for iteratively estimating scene geometry and camera parameters in all images simultaneously. An on-line method needs to, however, process each image at a time in the order they are taken and the complexity should not depend on the number of images or the length of the video sequence. A bundle adjustment based approach with a separate real-time 3D tracking was introduced in [4] which builds a bridge between accurate and iterative off-line methods and on-line active vision.

Few works appear in the literature on the subject of real-time visual reconstruction, apparently mostly due to the difficulty of on-line camera estimation. As alternative approaches to monocular vision, real-time visual reconstruction has been proposed based on stereo (e.g. [5]), artificial markers (e.g. [6]), active illumination (e.g. structured light [7]), or additional sensors (e.g. [8, 9]).

Simplification of the reconstructed point clouds into geometric models and the subsequent planning is not considered in this paper. There are several approaches for the simplification problem, with probably the most impressive results from fitting superquadrics (e.g. [10]). Lately, a simpler approach of minimum volume bounding box decompositions [11] was introduced with grasp planning from stereo data as an application.

2 Dense Visual 3D Reconstruction

The core of the 3D reconstruction system is EKF-based visual SLAM running at framerate, similar to the MonoSLAM [12]. The process model is a zeroth order model assuming stationary camera with Gaussian noise in the pose. The pose is parametrised by a Cartesian 3-vector for location and a quaternion for orientation. Landmark locations are parametrised by inverse depth as in [13].

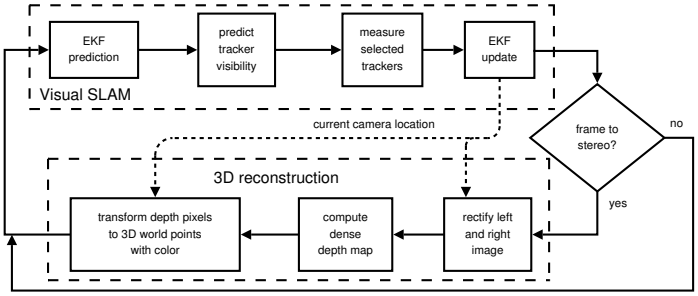

Fig. 1. Structure of the 3D reconstruction system using visual SLAM. SLAM and reconstruction processes could run in parallel. SLAM is independent of reconstruction. *Left* and *right image* refer to previous and current frames selected for dense stereo.

Measurements are provided by interest point trackers, which use RGB template matching with SSD-score to find translation, and adapt to the the predicted scale and rotation around the optical axis. Trackers are initialised by the Harris corner detector. The 3D reconstruction depends on the camera pose estimate from the visual SLAM. The pose is used to select video frames for stereo, for rectification of the stereo pair, and directly as pose of 3D point clouds without further registration. The general process structure is shown in Fig. 1.

We use the iterated EKF (IEKF) [14] for the measurement update step in order to minimise the linearisation errors of the measurement model (perspective projection). The problem of outlier measurements has been solved using the recursive branch and bound search [15] for the maximal joint compatibility proposed by Neira and Tardos [16].

An essential part of the system in Fig. 1 is the frame-to-stereo decision. The selected video frames should be of good quality and provide sufficient baseline. The camera should be well localised to the scene. The proper localisation uncertainty measure is the camera pose estimate given the measurements and the map uncertainties. A simplified approach, as proposed here, is to inspect the dynamic uncertainties as is.

Each new measurement, even from a poorly localised (in the SLAM map) landmark improves the camera pose estimate. Based on this, the following uncertainty indicator U is proposed:

$$U = \sqrt{\left(\sum_i \frac{1}{\operatorname{Tr} \boldsymbol{Q}_{3,i}} \right)^{-1}} \quad , \tag{1}$$

where $\boldsymbol{Q}_{3,i}$ is the Cartesian covariance matrix of the ith successfully measured landmark, and the sum goes over all measured landmarks. The trace of $\boldsymbol{Q}_{3,i}$ (total variance, TVAR) measures uncertainty, and the smaller the uncertainty, the higher weight the landmark has in estimation of the camera pose. Sum over inverses of TVARs accumulates the amount of "certainty", and inverse of the sum is again "uncertainty" in the variance domain. Square root gives an indicator in

the standard deviation domain of the uncertainty of the camera pose estimate. U is fast to compute and problems of division by near-zero are avoided by the fact that no measurement, and hence no landmark, has a near-zero uncertainty.

The notation for the decision rules is defined as follows. Time index t refers to the current video frame or view. The most recent selected left view is marked with time index p. Before acquiring a video frame t, the SLAM predicts which known landmarks in the SLAM 3D map could be visible, denoted as the set V_t. After measuring landmarks, the set $M_t \subseteq V_t$ contains the successfully measured landmarks. The size of a set is denoted $|\cdot|$. The baseline (according to SLAM) between two instants p and t is the translation vector $\hat{\boldsymbol{b}}_{p,t}$ from the camera optical centre p to t.

We propose the following rules for selecting video frames:

1. $|M_t| \geq T_{\text{meas}} = 5$.
2. Uncertainty $U < T_{\text{U}} = 0.02$ (1).
3. If the left view p is not set, set $p \leftarrow t$ and skip stereo.
4. $|M_p \cap V_t| \geq T_{\text{pcomm}} = 5$. If not, set $p \leftarrow t$ and skip stereo.
5. $\|\hat{\boldsymbol{b}}_{p,t}\| \geq T_{\text{base}} = 0.09$ SLAM map units ($\approx 5^\circ$ parallax angle).
6. $|M_p \cap M_t| \geq T_{\text{corr}} = 4$

If any condition fails, the right view is not assigned and the stereo reconstruction is skipped. Otherwise, the current view t is selected as the right view, a disparity map is computed ([17, 18] with Birchfield-Tomasi measure), and then the right view becomes the new left view ($p \leftarrow t$).

Rule 1 guarantees the minimum number of successful measurements, which is related to overall image quality. Rule 2 limits the camera pose uncertainty and also assures map scale stability. Rule 3 simply initialises the left view, if necessary. Rule 4 is the baseline accumulation abort rule, which guarantees that in the near future it is still possible to receive a view with sufficient overlap with the currently selected left view. Otherwise it restarts the baseline accumulation by resetting the left view. Rule 5 enforces sufficient baseline length. Finally, Rule 6 requires some number of landmarks to be observed in both views to guarantee overlapping views and to aid in the rectification step of the stereo algorithm. The threshold values have been found experimentally.

3 Examples

We showcase the system with a difficult video of a cluttered desk. Camera trajectory contains mostly translation, making a sweep from left to right and back. The furthest views have no overlap. The camera is a Logitech Quickcam Pro 4000, running at 30 Hz and 320x240 resolution. Length of the video is 471 frames or 15.7 seconds. The camera is moved by a free hand and shakes noticeably, and the intrinsics have been pre-calibrated. Progress of the 3D reconstruction is shown in Fig. 2, including two novel views of the completed model with 124799 points. Related videos can be found at `http://www.it.lut.fi/project/rtmosaic/`. With an AMD Athlon64 3500+ CPU and Nvidia Geforce 6600 graphics card

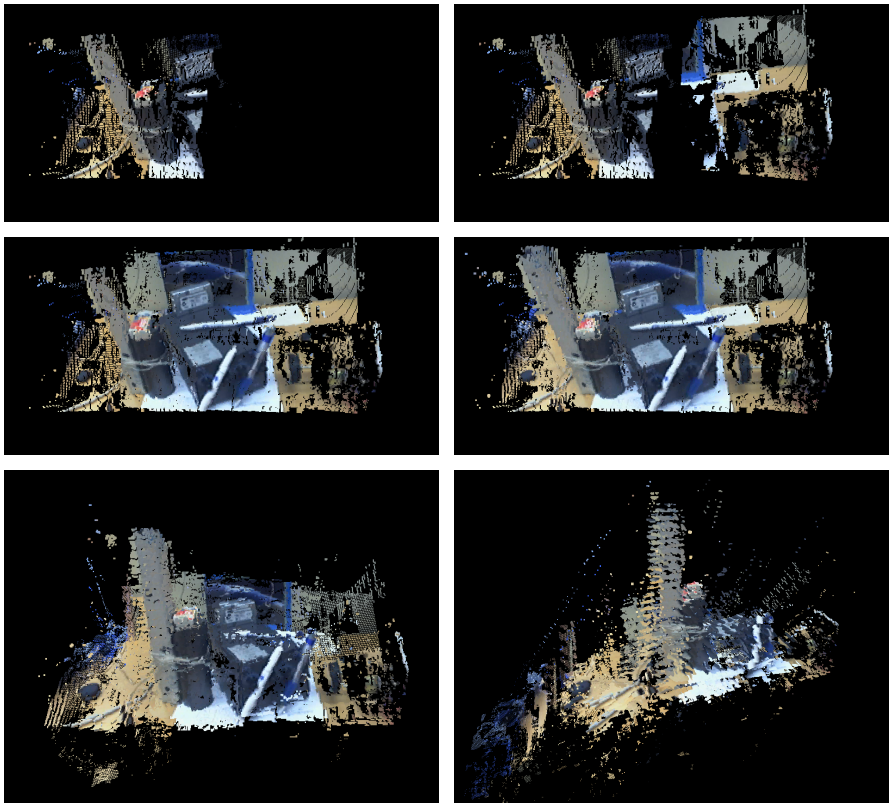

Fig. 2. The 3D model after each stereo processing cycle at $t = \{78, 362, 383, 437\}$ and in the last row from two novel angles. The bottom right image shows the discrete nature of the disparity map. No pruning is applied to the joined point clouds.

(for visualisation only), our current implementation takes on average 135% of the available time per frame with this specific 30 Hz video sequence. The bottleneck is the stereo implementation.

4 Summary

This paper presents a rudimentary system capable of nearly real-time 3D reconstruction using visual SLAM. Its input is a monocular video sequence of a static scene and output is a semi-dense coloured point cloud in three dimensions. The camera is assumed to be calibrated, but no restrictions are explicitly set on the scene structure. Visual SLAM provides an estimate of camera pose for each video frame. A stereo correspondence algorithm creates 3D points from selected video frame pairs. Thus, any video sequence can be used to produce on-line a 3D model, usable for action generation, e.g. grasp planning or navigation. Our future plans include evaluation of the system in grasping in unstructured environments.

Most of the individual algorithms used are either published state-of-the-art or standard well-known techniques. However, to the authors knowledge, no-one has yet demonstrated a system that combines all these components into a real-time reconstruction application. The novel contribution considers bringing these components into co-operation and especially selecting proper frames from a video stream for the stereo algorithm. Our system is simple, but offers an excellent basis for developing real-time reconstruction based on visual SLAM, and the example proves real-time performance achievable.

References

1. Saxena, A., Driemeyer, J., Ng, A.Y.: Robotic grasping of novel objects using vision. International Journal of Robotics Research **27**(2) (2008) 157–173
2. Pollefeys, M., van Gool, L., Vergauwen, M., et al.: Visual modeling with a hand-held camera. Int. Journal of Computer Vision **59**(3) (Sep 2004) 207–232
3. Goldlucke, B., Ihrke, I., Linz, C., Magnor, M.: Weighted minimal hypersurface reconstruction. IEEE TPAMI **29**(7) (Jul 2007) 1194–1208
4. Klein, G., Murray, D.: Parallel tracking and mapping for small ar workspaces. In: Mixed and Augmented Reality (ISMAR). (2007) 1–10
5. Chang, W.C., Lee, S.A.: Real-time feature-based 3d map reconstruction for stereo visual guidance and control of mobile robots in indoor environments. In: Systems, Man and Cybernetics. Volume 6. (2004) 5386–5391
6. Fudono, K., Sato, T., Yokoya, N.: Interactive 3-d modeling system using a hand-held video camera. In: Proc. of SCIA. LNCS, Springer (2005) 1248–1258
7. Rusinkiewicz, S., Hall-Holt, O., Levoy, M.: Real-time 3d model acquisition. ACM Trans. Graph. **21**(3) (2002) 438–446
8. Pollefeys, M., Nistér, D., Frahm, J.M., et al.: Detailed real-time urban 3d reconstruction from video. Int. Journal of Computer Vision **78**(2) (Jul 2008) 143–167
9. Hogue, A., German, A., Jenkin, M.: Underwater environment reconstruction using stereo and inertial data. In: IEEE Int. Conf. SMC. (2007) 2372–2377
10. Chevalier, L., Jaillet, F., Baskurt, A.: Segmentation and superquadric modeling of 3D objects. Journal of WSCG (2003)
11. Huebner, K., Ruthotto, S., Kragic, D.: Minimum volume bounding box decomposigion for shape approximation in robotic grasping. In: IEEE Int Conf on Robotics and Automation, Pasadena, CA, USA (2008) 1628–1633
12. Davison, A., Reid, I., Molton, N., Stasse, O.: Monoslam: Real-time single camera SLAM. IEEE TPAMI **29**(6) (Jun 2007) 1052–1067
13. Montiel, J., Civera, J., Davison, A.: Unified inverse depth parametrization for monocular SLAM. In: Robotics: Science and Systems, Philadelphia, USA (2006)
14. Bar-Shalom, Y., Li, X.R.: Estimation and Tracking: Principles, Techniques and Software. Artech House (1993)
15. Clemente, L.A., Davison, A.J., Reid, I.D., Neira, J., Tardos, J.D.: Mapping large loops with a single hand-held camera. In: Robotics: Science and Systems. (2007)
16. Neira, J., Tardós, J.D.: Data association in stochastic mapping using the joint compatibility test. IEEE Trans. on Robotics and Automation (2001)
17. Ogale, A., Aloimonos, Y.: Robust contrast invariant stereo correspondence. IEEE Int. Conf. on Robotics and Automation (2005) 819–824
18. Ogale, A., Domke, J.: Openvis3d, open source 3d vision library. Website (2008) Referred Mar 15th, 2008. http://code.google.com/p/openvis3d/.