The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Selecting Good Corners for Structure and Motion Recovery Using a Time-of-Flight Camera

Peter Gemeiner, Peter Jojic and Markus Vincze
Automation and Control Institute
Vienna University of Technology
{gemeiner, jojic, vincze}@acin.tuwien.ac.at

Abstract— **In the robotics and computer vision communities, localization and mapping of an unknown environment is a well studied problem. To tackle this problem in real-time using a single camera, state-of-the-art Simultaneous Localization and Mapping (SLAM) or Structure from Motion (SfM) algorithms can be used. To create the model of the unknown environment, the camera moves and adds to the map from point to point, and assumes that these detected points are unique 3D corners. However, the scene usually contains false 3D corners, lying at e.g. occlusion boundaries. Inserting these points into the map may lead to SLAM failure or to less accurate estimations in SfM. In this work, a corner selection scheme is proposed that exploits the amplitude and depth signals of a Time-of-Flight (ToF) camera. The selection scheme detects false 3D corners based on a 3D cornerness measure. We then prove that the rejection of these corners increases the accuracy with a simulated SfM example and show the results of using our selection scheme with the ToF camera sequences.**

I. Introduction

For mobile robotics or head gears in augmented reality (AR) applications, it is essential to continuously localize and estimate 3D positions of new landmarks in an unknown environment. In robotics, this is required for navigation, in AR to overlay virtual information correctly. The *localization* and *mapping* can be addressed with an incremental probabilistic approach as Bearing-Only SLAM [1], [2], [3] or SfM [4]. The common task of these methods is localization, but the difference between them is that SLAM commonly builds a sparse map and SfM aims to produce a dense map of the unknown scene. The roots of SfM can be found in photogrammetry e.g. [5], and despite involving a non-linear optimization technique it can work in real-time [4], [6]. The main drawback compared to SLAM is that the projected 3D corners have to be tracked from frame to frame, and the motion drift can not be corrected after re-observing the same place.

As input for SLAM, different kinds of sensors (e.g. laser [7], sonars [8]) can be used. One of the most interesting (cost, weight, etc.) and challenging SLAM sensors is a single perspective-projection camera. When observing the environment with a camera, the depth information of new landmarks can not be directly acquired. To recover this depth information, the camera has to move, and observe this landmarks from different viewpoints.

Davison et al. introduced the first real-time Monocular SLAM (MonoSLAM) (recently summarized in [3]) algorithm. The camera motion estimation and incremental map building (from new landmarks) are computed within a standard Extended Kalman Filter (EKF) SLAM framework. An alternative SLAM framework is typically based on FastSLAM-type [9] particle filter algorithms.

Another visual SLAM framework based on a FastSLAM-type particle filter introduced by Eade and Drummond [10] can incorporate hundreds of features in real-time. However, the filter needs to be adapted for closing loops over large trajectories.

Recently, a very interesting real-time parallel tracking and mapping approach was introduced by Klein and Murray in [11]. They proposed splitting the tracking and mapping into two separate tasks, running in parallel threads. The result is a system capable of estimating thousands of features in real-time, but suitable for small workspaces.

When using a perspective-projective camera, landmarks have to be seen from different viewpoints before they can be inserted into the model of the scene. The state-of-the-art model for these new features is the inverse depth parameterization introduced by Civera et al. in [12]. In contrast to the particle filter, feature initialization proposed by Davison et al. in [3] the inverse depth parameterization has the ability to estimate very distant features, possibly lying at infinity.

Above-mentioned Bearing-Only SLAM [3], [9], [10] and SfM [4], [5], [6], [11] approaches can not infer the depth information of the scene from a single view due to the perspective-projective nature of the camera. To initialize new landmarks' positions, different views of the scene are usually needed, and in Bearing-Only SLAM [3] this is known as delayed initialization. However, in [13] an undelayed initialization method is proposed, but only simulated results are introduced. Despite these efforts, landmarks' initialization in Bearing-Only SLAM and SfM is still a challenging task, because if these new landmarks do not represent distinct 3D corners and lie at e.g. occlusion boundaries, it can lead to SLAM [3] failure or SfM [4] inaccuracies.

In this work, a ToF camera capable of measuring the scene amplitude and depth directly has been used. Due to the available depth data, the geometry of the visible scene can be inferred. We present a three-stage selection scheme capable

978-1-4244-3804-4/09/$25.00 ©2009 IEEE

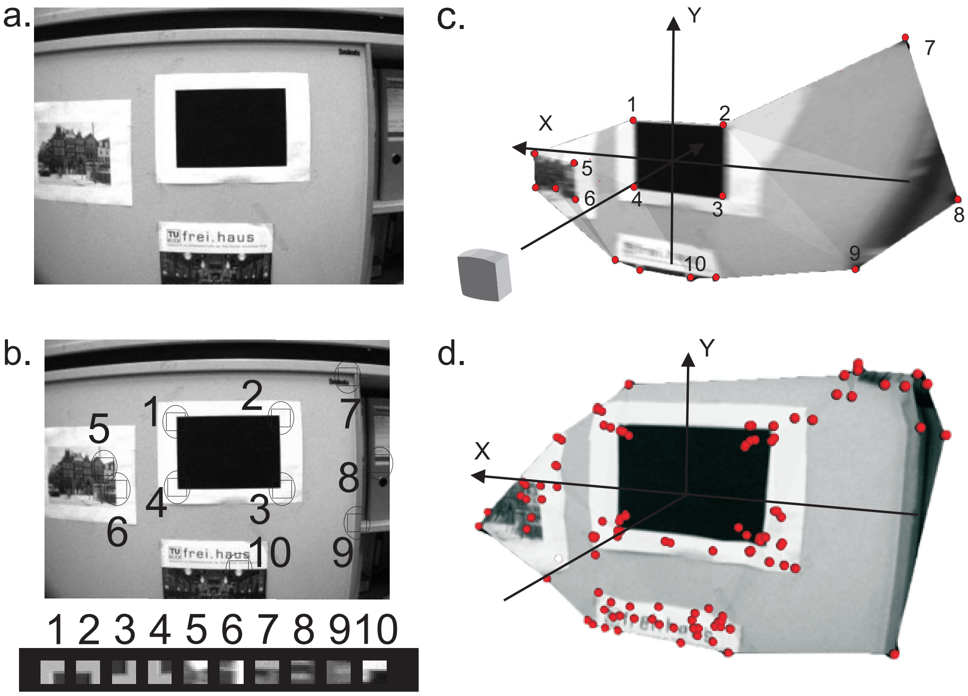

Fig. 1. The only sensor input in MonoSLAM are images from a single camera (a). As the camera moves, new distinctive features are detected (b). The output of MonoSLAM is the camera pose and a sparse, three-dimensional map of these distinctive features (c). An alternative to MonoSLAM sparse mapping is a more dense structure reconstruction using bundle adjustment (d).

of pruning false landmarks lying at occlusion boundaries, for example. The first stage of the presented selection scheme uses a well-known corner detector applied to the amplitude data. The second stage analyzes the geometry of the depth measurements, and decides if the landmark is a valid 3D corner. If necessary, the third stage proves using RANSAC [14] if the corner lies in a plane. To our knowledge, this is the first use of such an approach to locate 3D corners in the scene.

An alternative approach on how to use the depth ToF measurements to initialize new landmarks could be Range-Bearing SLAM, e.g. [15]. This SLAM algorithm only takes advantage of the depth geometric information. However, the contribution of our work focuses on how to combine the amplitude and depth ToF measurements to select good scene corners.

A comprehensive overview of the ToF principle and camera which has been used in this work can be found in [16].

In this paper, the bundle adjustment used in SfM is briefly introduced in Section II. Details related to the proposed corner selection scheme are introduced in Section III. Section IV explains how the false corners influence the structure and motion recovery, analyzes the ToF data sequences, and discusses the experimental results. Section V closes with a discussion and an outlook to the future work.

II. Bundle Adjustment

In this section, a brief overview of SfM recovery algorithm commonly known as bundle adjustment is given. The difference between a Bearing-Only SLAM technique like MonoSLAM and bundle adjustment is that MonoSLAM recovers a rather sparse map of the scene in real-time. An example is displayed in Fig. 1. When the camera moves, new distinctive features can be detected (e.g. corners with numbers 5-10 in Fig. 1-b and Fig. 1-c). The output of MonoSLAM is the camera poses and a sparse map of recovered features, as depicted in Fig. 1-c. Due to the real-time demand, *mapping* in MonoSLAM does not play a crucial role, and should rather support *localization*.

A. The Basic Perspective-Projective Camera Model

The perspective-projective, also known as the pin-hole camera, maps 3D landmarks into a 2D image plane. This mapping, in our case *central projection* P_i [17], [18], has the following form

$$\lambda_{ij} x_{ij} = P_i X_j ,\ i = 1 \ldots m,\ j = 1 \ldots n\ , \tag{1}$$

where X_j are the homogeneous 3D landmarks, x_{ij} are the homogeneous 2D points and λ_{ij} are the projective depths.

B. Iterative Non-linear Optimization

The recovery of an unknown structure and motion can be solved, when this problem is formulated as Maximum Likelihood Estimation [18]

$$L = \max \prod_{i,j \in I} e^{|x_{ij} - g(P_i, X_j)|^2 / 2\sigma^2}\ , \tag{2}$$

where $g(P_i, X_j)$ represents the reprojected landmarks.

Fig. 2. The Swiss Ranger™ SR3000 ToF camera (courtesy of Mesa Imaging - http://www.mesa-imaging.ch).

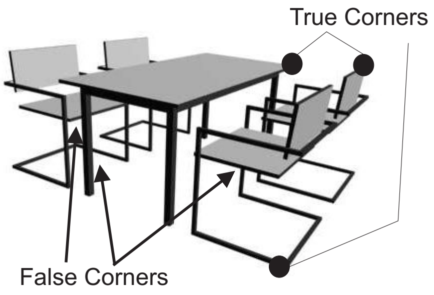

Fig. 3. Our simulated scene contains one table and four chairs. In this bundle adjustment example, twenty-two good corners were used, but in this figure, only three of them are depicted as circles. We used up to three bad corners, which are marked with arrows.

More commonly known is a standard reformulation of this maximization problem. The result of this reformulation is the following equation

$$f = \sum_{i,j \in I} (x_{ij} - g(P_i, X_j))^2 \ , \tag{3}$$

where the problem is to minimize the negative logarithm of the likelihood function L [18].

The 2D reprojections x_{ij} of 3D corners X_j are calculated as

$$x_{ij} = g(K_i, R_i, t_i, X_j) \ . \tag{4}$$

Central projection P_j in (1) encompasses the same camera parameters (K_i, R_i and t_i), where K_i are the intrinsic parameters, R_i is the orientation and t_i is the position.

The task of bundle adjustment is to minimize the deviation between 2D measured and reprojected points in L^2 norm. The formulation is stated here

$$\min_{K_i, R_i, t_i, X_j} \sum_{ij} (x_{ij} - \widehat{x}_{ij})^2 \ . \tag{5}$$

To find a minimum to this formulation, an iterative non-linear optimization method e.g. Levenberg-Marquardt can be used.

For further camera geometry details and bundle adjustment explanations please refer to e.g. [17], [18].

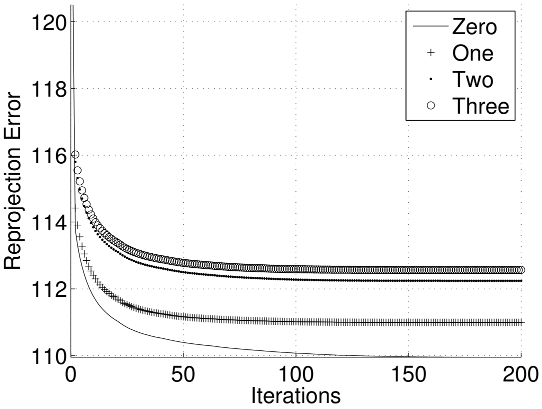

Fig. 4. Simulated bundle adjustment reprojection errors using zero, one, two or three false corners. It can be seen that the addition of a false corner increases the reprojection errors.

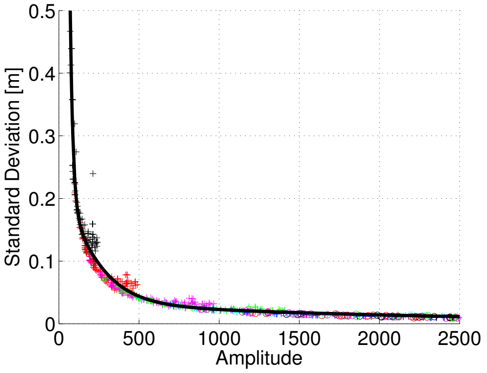

Fig. 5. The standard deviation of depth data as a function of amplitude data. This function has been fitted using the non-linear optimization.

III. Corner Selection Scheme

In the real scenes it is obvious that there are usually features which do not represent true 3D points, e.g. lying at occlusion boundaries, caused by specular highlights or curved surfaces. However, one of the underlying Bearing-Only SLAM or SfM assumptions is that the scene is rigid, which means that features are not expected to change their position within the scene. The map management implemented (e.g. in MonoSLAM) can prune some of these bad corners, but using the depth information a further feature validation can be performed.

Our proposed corner selection scheme contains three stages:

- detect the best 2D corners using the amplitude data,
- compute the 3D cornerness measure at the found 2D corner position using the depth measurements, and
- if the 3D cornerness measure is above a threshold, then fit a plane to the depth data.

In the first stage, `Fast` feature detector [19] is applied to the amplitude data. The best `Fast` features are then found using the Shi and Tomasi [20] cornerness measure.

With the second stage we decide whether it is a good 3D corner. To evaluate this, the standard matrix of image partial

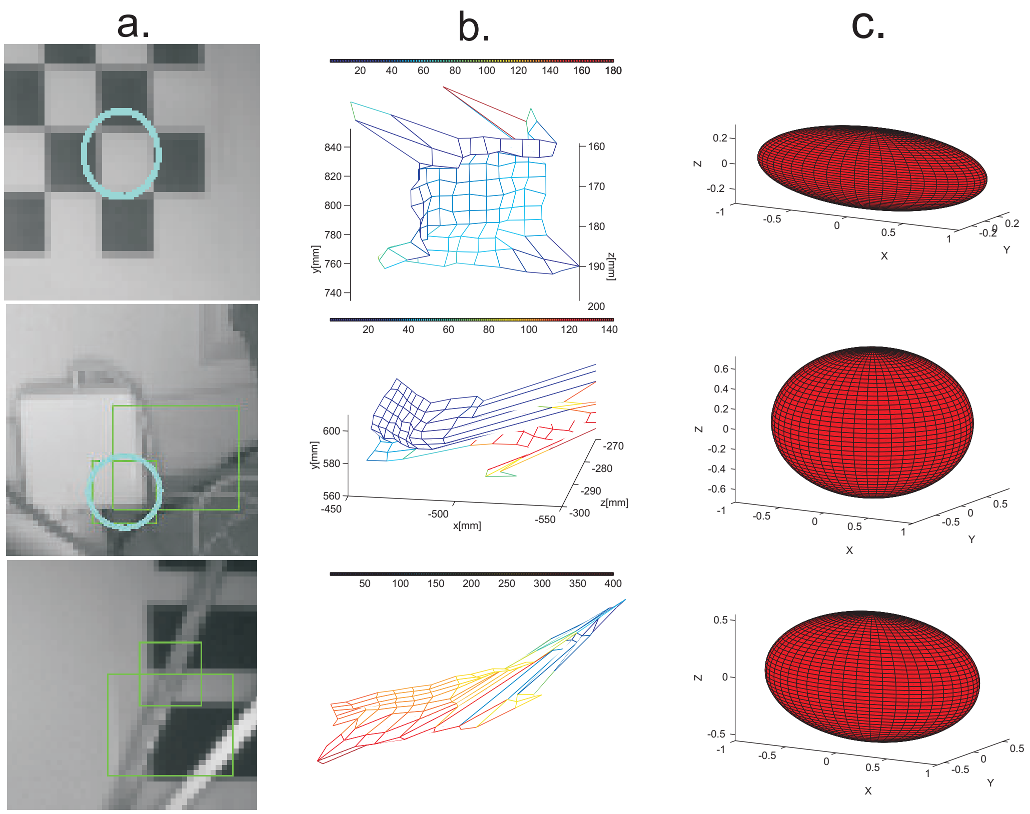

Fig. 6. Detection of new features using the real image and depth information. The upper row is a planar feature, the middle row depicts a good 3D feature, and the bottom row displays a bad 3D corner. The first column displays the amplitude data provided by the camera. The second column depicts measured depth information w.r.t. the world frame. The third column represents the visualization of the eigen decomposition of the 3D feature gradient matrix. The structure tensor visualization of a good 3D corner is expected to have a sphere like structure, as displayed in the last column of the middle row. The other two features did not pass the 3D cornerness measure.

gradients:

$$S_i = \begin{pmatrix} I_x^2 & I_xI_y \\ I_xI_y & I_y^2 \end{pmatrix} \tag{6}$$

is extended to three dimensions as proposed in [21]

$$S_d = \begin{pmatrix} D_x^2 & D_xD_y & D_xD_z \\ D_xD_y & D_y^2 & D_yD_z \\ D_xD_z & D_yD_z & D_z^2 \end{pmatrix} , \tag{7}$$

using the depth instead of the visual information. To compute the 3D cornerness measures, two operators have been compared. Rohr presented in [21] a survey on 3D operators, and we selected the following one:

$$Op(x, y, z) = \frac{det\ S}{trace\ S} \rightarrow \max, \tag{8}$$

because it is related to the Shi and Tomasi [20] 2D cornerness measure. Another 3D cornerness measure presented by Arseneau and Cooperstock in [22] is based on the eigen decomposition of the S_d matrix, where this matrix is called the *structure tensor*. If a corner does not pass this stage, it can still be a good corner, possibly lying in the plane.

In the third stage, a plane is fitted to the depth data at the 2D corner position found using RANSAC to test whether this is a planar feature. We propose fitting the plane using two criteria. The first is the distance of points to the fitted plane, and the second one is the angle between the direction of a new feature and the plane normal vector.

For the robust plane fit, the distance function between the plane P and an array of measured depth points is computed as follows

$$\begin{aligned} \vec{n} &= (P_2 - P_1) \times (P_3 - P_1) \\ d_i &= |(X_i - P_1)\ \vec{n}|,\ i = 1 \dots m \end{aligned}$$

where the plane P is a matrix defined column-wise by three points, $\vec{n}$ is the plane normal vector, X_i are the measured points, and d_i are the distances.

IV. Experimental Results

In this section, we present our experimental setup as well as simulated and real data results in three parts. Firstly, we show on a simulated example that every false corner used in the model of the scene increases the SfM inaccuracies. Secondly, the measurement principle as well as the properties of the ToF camera are briefly explained. Thirdly, the relation between the amplitude and standard deviation in distance information is estimated. Finally, our proposed feature extraction scheme is evaluated.

A. *Simulated Bundle Adjustment Example*

To prove how false corners influence the SfM results, an unknown structure of a simulated scene and virtual camera motion has been recovered. The simulated scene is depicted in Fig. 3.

The evaluation criterion is the reprojection error f as defined in (3). The empirical results are displayed in Fig. 4, where it is obvious that every additional false corner introduces inaccuracies to the reprojection errors.

We think that an additional comparison of how false corners influence estimation results using a SLAM approach is not necessary. As presented in e.g. [23], SfM using Levenberg-Marquardt optimization can be used as ground-truth.

B. *ToF Camera*

In this work, we have been using the state of the art Swiss Ranger$^{\text{TM}}$ SR3000[1] ToF camera (see Fig. 2). The measurement principle of this camera is based on the amplitude-modulated, near infrared (NIR) light, which uses NIR light pulses, and leads to the phase difference distance measurement principle as explained e.g. in [16]. This principle is complex and includes systematic (e.g. distance-related, amplitude-related, fixed pattern noise) and non-systematic (e.g. signal-to-noise ratios, rays reflection, light scattering) errors, which are further described in [16], [24]. Except these errors the camera has a limited field of view (horizontal 47.5°, vertical 39.6°), a small resolution (176x144pix), and a depth range up to 7.5m.

The SR3000 ToF camera is using a perspective-projection camera model [25] for the amplitude data. A model of the distance information has not been used, but the distances have been calibrated for every single pixel separately [16]. The amplitude and distance calibration has been performed by the camera manufacturer.

Before capturing real data sequences, the camera integration time, which influences the measurement precision and read-out speed, has to be adjusted. In this paper, the camera integration time is adjusted to 7200μs, and the read-out speed then equals approximately 21Hz. A faster read-out speed would be suitable to reduce the motion blur. However, a faster read-out leads to a higher signal-to-noise ratio (SNR), and this introduces inaccuracies in feature extraction [26].

The ToF camera is calibrated by the producer, so the intrinsic parameters (optical center and focal length) are known, and only the distortion coefficient needs to be adjusted. The distance-related error calibration is also provided by the producer, and it is done by reducing the measurement offset using an acquired Fixed Pattern Noise matrix [16].

C. *Static Depth Noise Modelling*

To estimate the relation between the amplitude and the standard deviation in depth measurements as proposed by Kahlmann in [16], we perceived multiple measurements of a white wall using several integration times (from 200μs to 9800μs). The obtained relation between amplitude and standard deviation in distance measurement for our camera is depicted in Fig. 5.

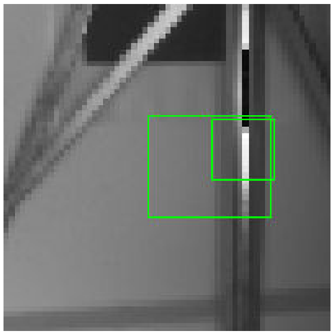

Fig. 7. An example of a ray reflection on a metallic surface, which is one type of the possible non-systematic ToF camera errors.

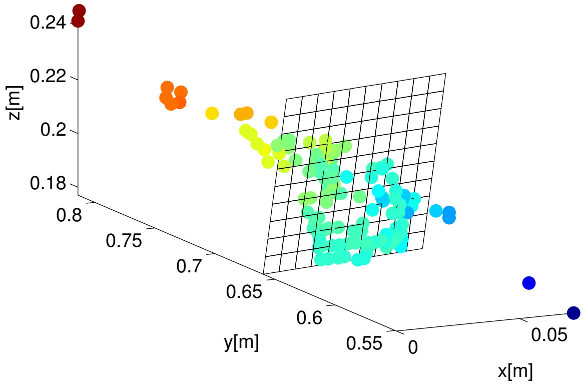

Fig. 8. Plane fitted to the depth data using `RANSAC`.

D. *Feature Extraction*

To prevent inserting bad corners into the model of the scene, we have implemented three validation stages, as explained in Sec. III. This sub-section presents the experimental results using these three stages.

1) 2D Corner Extraction: The first validation stage is used to detect the 2D corner in the amplitude data. As displayed in the first row of Fig. 6(a), this validation stage works well for most of the real corners.

However, we typically find occluded features as e.g. the last row in Fig. 6(a). Using the ToF camera amplitude data, other kind of bad features can also occur, e.g. caused by rays reflection as depicted in Fig. 7. Davison et al. introduced in [3] a map management algorithm, but this usually does not prune these bad features.

2) 3D Corner Extraction: To improve the feature extraction, we have implemented two similar 3D feature cornerness measures, as described in Sec. III. Both of these measures use the depth gradient information, but the one proposed by Rohr in [21] proved to be less depth distinctive.

In this validation stage, only the structure tensor [22] cornerness measures have been used. The visualization of a good 3D corner is expected to be similar to a sphere like

[1] Produced by Mesa Imaging - http://www.mesa-imaging.ch.

structure [22]. A real good corner is depicted in the second row of Fig. 6(c).

Our criterion for a good 3D corner (see the second row of Fig. 6) is that the differences of the three computed structure tensor eigen values have to be lower than a predefined threshold. Using this criterion, planar features and bad 3D corners are typically pruned (see the first and third row of Fig. 6).

3) Planar Corner Extraction: The third stage is needed, when a new real feature is pruned by the 3D cornerness measure, but it can lie in a plane. To evaluate this stage, a robust plane fit using RANSAC has been implemented, and an example of the fitted plane is depicted in Fig. 8.

Using this RANSAC plane fit, planar features are detected and initialized as depicted in the first row of Fig. 6(a).

V. Conclusion

The contribution of this paper is a three-stage corner selection scheme, which enables the rejection of false features lying at occlusion boundaries or curved surfaces, for example. Our corner selection scheme takes advantage of a recently introduced ToF camera, which can measure scene amplitude and depth directly. We showed that our contribution can help SLAM and SfM approaches to select good corners for tracking.

A. Future Work

Current ToF cameras have several drawbacks (e.g. field of view, noisy data) due to the complex measurement principle, but enhanced sensors have already been announced. It is obvious that this camera can ease the parallel localization and mapping task significantly, and we think that it is an interesting sensor for indoor robotics navigation and many future applications.

The fusion of ToF amplitude and depth information for parallel localization and mapping has been addressed in the work of Weingarten in [27]. The conclusion is that the ToF camera is less suited for localization and mapping than the rotating laser scanner due to its noisy data and limited field of view. However, with the advantage of excluding false corner points this conclusion may now be revised.

In an offline 3D pose estimation and mapping approach introduced by May et al. in [24], the *Iterative Closest Point* algorithm has been used to register point clouds from different perspectives, and KLT or SIFT features has been applied to amplitude data. We think that our proposed corner selection scheme could increase the performance of May et al. approach.

Acknowledgments

This research was supported by the European Union project XPERO IST-29427.

References

[1] H. F. Durrant-White and T. Bailey, "Simultaneous localization and mapping: Part I," *IEEE Robotics and Automation Magazine*, vol. 13, no. 2, pp. 99–110, June 2006.

[2] T. Bailey and H. Durrant-Whyte, "Simultaneous localization and mapping (SLAM): part II," *IEEE Robotics and Automation Magazine*, vol. 13, no. 3, pp. 108–117, 2006.

[3] A. J. Davison, I. Reid, N. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1052–1067, June 2007.

[4] D. Nistér, "Preemptive RANSAC for live structure and motion estimation," in *IEEE International Conference on Computer Vision*, vol. 1, 2003, pp. 199–206.

[5] M. Pollefeys, R. Koch, and L. V. Gool, "Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters," *International Journal of Computer Vision*, vol. 32, no. 1, pp. 7–25, 1999.

[6] D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry," in *IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 1, 2004, pp. 652–659.

[7] J. Weingarten and R. Siegwart, "EKF-based 3D SLAM for structured environment reconstruction," in *IEEE International Conference on Intelligent Robots and Systems*, Edmonton, Canada, 2005, pp. 3834–3839.

[8] G. Zunino and H. Christensen, "Simultaneous localization and mapping in domestic environments," in *Multisensor Fusion and Integration for Intelligent Systems*, Baden-Baden, Germany, 2001, pp. 67–72.

[9] M. Pupilli and A. Calway, "Real-time visual SLAM with resilience to erratic motion," in *IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 1, June 2006, pp. 1244–1249.

[10] E. Eade and T. Drummond, "Scalable monocular SLAM," in *IEEE International Conference on Pattern Recognition*, vol. 1, 2006, pp. 469–476.

[11] G. Klein and D. W. Murray, "Parallel tracking and mapping for small AR workspaces," in *ISMAR*, 2007.

[12] J. M. M. Montiel, J. Civera, and A. J. Davison, "Unified inverse depth parametrization for monocular SLAM," in *Proceedings of Robotics: Science and Systems*, Philadelphia, USA, August 2006.

[13] J. Sola, A. Monin, M. Devy, and T. Lemaire, "Undelayed initialization in bearing only SLAM," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, August 2005, pp. 2499–2504.

[14] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, June 1981.

[15] E. Olson, J. Leonard, and S. Teller, "Robust range-only beacon localization," in *Proceedings of Autonomous Underwater Vehicles*, 2004.

[16] T. Kahlmann, *Range Imaging Metrology: Investigation, Calibration and Development*. ETH Zurich, Switzerland: PhD thesis, 2007.

[17] R. Hartley and A. Zisserman, *Multiple View Geometry*. Cambridge University Press, 2003.

[18] N. Guilbert, "Structure from motion," Ph.D. dissertation, LTH, Lund, Sweden, 2004.

[19] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *European Conference on Computer Vision*, vol. 1, May 2006, pp. 430–443.

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

[21] K. Rohr, "On 3D differential operators for detecting point landmarks," *Image and Vision Computing*, vol. 15, no. 3, pp. 219–233, March 1997.

[22] S. Arseneau and J. Cooperstock, "An asymmetrical diffusion framework for junction analysis," in *British Machine Vision Conference*, 2006.

[23] E. Eade and T. Drummond, "Monocular SLAM as a graph of coalesced observations," in *IEEE International Conference on Computer Vision*, 2007.

[24] S. May, D. Dröschel, D. Holz, C. Wiesen, and S. Fuchs, "3D pose estimation and mapping with ToF cameras," in *IEEE International Conference on Intelligent Robots and Systems*, Nice, France, September 2008.

[25] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *IEEE International Conference on Computer Vision*, September 1999, pp. 666–673.

[26] P. Jojic, "Implementing a ToF camera interface for visual simultaneous localization and mapping," Master's thesis, Vienna University of Technology, 2008.

[27] J. Weingarten, *Feature Based 3D SLAM*. EPFL Lausanne, Switzerland: PhD thesis, 2006.