Towards Semantic KinectFusion

Nicola Fioraio, Gregorio Cerri, and Luigi Di Stefano

Dept. of Computer Science and Engineering
University of Bologna, viale Risorgimento, 2, Bologna, Italy
{nicola.fioraio,luigi.distefano}@unibo.it
http://vision.deis.unibo.it

Abstract. In this paper we propose an extension to the KinectFusion approach which enables both SLAM-graph optimization, usually required on large looping routes, as well as discovery of semantic information in the form of object detection and localization. Global optimization is achieved by incorporating the notion of keyframe into a KinectFusion-style approach, thus providing the system with the ability to explore large environments and maintain a globally consistent map. Moreover, we integrate into the system our recent object detection approach based on a new Semantic Bundle Adjustment paradigm, thereby achieving joint detection, tracking and mapping. Although our current implementation is not optimized for real-time operation, the principles and ideas set forth in this paper can be considered a relevant contribution towards a Semantic KinectFusion system.

Keywords: KinectFusion, semantic SLAM, semantic bundle adjustment, object detection.

1 Introduction

In the last decade SLAM (Simultaneous Localization and Mapping) has witnessed impressive progresses [7,14,18,24]. Feature-based approaches [7,14] have been deeply explored, so as to attain both accurate maps and real-time operation. 3D reconstruction, though, turns out typically quite sparse and hence often unsuited to robotic tasks such as motion planning and obstacle avoidance. Recently, novel dense approaches using all image pixels to infer a 3D model of the environment from the video sequence have been proposed [22,17]. Depth cameras, such as the Microsoft Kinect, have then brought in new sensing modalities providing 3D measurements at 30Hz, thus paving the way for a brand-new RGBD-SLAM research topic. RGB-D Mapping, proposed by Henry *et al.* [12], was one of the first RGBD-SLAM frameworks whereby both color and depth information are deployed for tracking. The system also includes loop closure detection and a 3D surface reconstruction engine. Later, Newcombe *et al.* [18] introduced KinectFusion, a real-time tracking algorithm capable of unrivaled accuracy. KinectFusion constantly updates a volumetric representation of the scene consisting of a Truncated Signed Distance Function (TSDF) [6] and tracks incoming frames based on the current 3D reconstruction, achieving real-time operation through efficient GPU implementation. As the main limitation of KinectFusion is the bounded mapped scene volume, Whelan *et al.* [24] proposed to shift the TSDF cube according to camera movements throughout the environment. While the camera moves, the voxels exiting the

A. Petrosino (Ed.): ICIAP 2013, Part II, LNCS 8157, pp. 299–308, 2013.

current volume are triangulated so that a highly detailed 3D reconstruction of large environments can be obtained. However, even though KinectFusion is a low-drift tracking system, mapping large areas is inherently prone to propagation and amplification of pose errors which, afterward, may lead to a poor-quality reconstruction. As highlighted by the authors, possible solutions to address this issue might be reintegration into the moving volume of previously extracted surfaces, as well as handling loop closure to carry out a global optimization.

In this paper we propose first a novel extension to KinectFusion-style system which uses a TSDF volume to build a keyframe representation of the environment. This approach allows for seamless reintegration of already mapped areas and global bundle-adjustment optimization. Moreover, as a second contribution, we enrich the 3D reconstruction with semantic information by means of Semantic Bundle Adjustment [10]: while the camera moves object instances are detected leveraging SLAM and their poses estimated altogether with camera poses. Though other systems allows for semantic SLAM [5,8] or semantic SFM [2], [10] casts the problem as a bundle adjustment style optimization, enabling a fully integrated pipeline. As a result, our system can close seamlessly medium-size loops, delivers accurate 3D reconstructions and peculiarly provides semantic knowledge concerning the mapped environment.

Next section discusses our novel KinectFusion extension, while in Sec. 3 we will show how to integrate the semantic bundle adjustment framework [10] into the proposed KinectFusion extension. Sec. 4 reports both quantitative and qualitative results, in particular by comparing our methods to RGB-D Mapping. Finally, in Sec. 5 some concluding remarks are drawn.

2 Bundle Adjustment by the TSDF

The TSDF representation of a 3D volume consists of a voxel data structure with each element storing the signed distance to the nearest surface. According to Curless *et al.* [6], a depth image is merged into the current volume by back-projection of voxels onto the image plane; then, the resulting pixel gives the position of the surface and the required distance can be computed. Denoted as $[\mathrm{F}_{k-1}(\mathbf{p}), \mathrm{W}_{k-1}(\mathbf{p})]$, respectively, the signed distance value and a corresponding weight at the 3D location $\mathbf{p}$ at time $k-1$, a new measurement $[\mathrm{F}(\mathbf{p}), \mathrm{W}(\mathbf{p})]$, coming from the projection of $\mathbf{p}$ onto the image plane is integrated as:

$$\mathrm{F}_k(\mathbf{p}) = \frac{\mathrm{W}_{k-1}(\mathbf{p})\,\mathrm{F}_{k-1}(\mathbf{p}) + \mathrm{W}(\mathbf{p})\,\mathrm{F}(\mathbf{p})}{\mathrm{W}_{k-1}(\mathbf{p}) + \mathrm{W}(\mathbf{p})} \tag{1}$$

$$\mathrm{W}_k(\mathbf{p}) = \mathrm{W}_{k-1}(\mathbf{p}) + \mathrm{W}(\mathbf{p}) \tag{2}$$

If the signed distance appears to be over a threshold μ from the surface, it must be truncated, as discussed in [6]. In the following, we will refer to the voxels having a non-truncated function value as *surface voxels*.

Recently, the KinectFusion system by Newcombe *et al.* [18] showed how to exploit the TSDF representation not only to achieve detailed surface reconstruction, but also during the camera tracking process. First, a synthetic depth map is rendered by ray-casting each image pixel according to the known camera intrinsic parameters and following the ray to the intersection with the surface. As described in [6,18], this process is sped up by the TSDF, as the surface is implicitly represented by the zero-crossings of the function. Then, each incoming frame is registered to this virtual depth image through an ICP-like scheme [4,3] and the final aligned frame merged with the surface according to the estimated pose.

The continuous process of TSDF ray-casting and updating makes the system robust and produces low-drift mapping, especially for very loopy trajectories. However, the extent of the mappable volume is bounded by the initial TSDF volume so that it is not possible to reconstruct larger environments. While Whelan *et al.* [24] proposed to simply move the TSDF volume, we think that this strategy is inherently unable to maintain a globally consistent map, since it is not clear how previously mapped locations can be taken into account for global error minimization, *e.g.* how to achieve consistent mapping across large looping routes.

Following a strategy common to many successful SLAM algorithms [12,9], we introduce the notion of *keyframe* as a spatial sampling of camera trajectory and globally optimize keyframe poses for consistent mapping. However, unlike previous proposals, keyframe selection and constraining is performed through the TSDF representation of the environment. To achieve this, we add a new field to each TSDF voxel to keep track of a list of keyframes. Then, when a new frame promoted to keyframe is merged into the volume, it is also added to the keyframe list of all its surface voxels, *i.e.* those voxels to whom a non-truncated distance value has been assigned. Accordingly, on one hand for every point close to the estimated surface we always know which keyframes it was captured in. On the other hand, surface voxels with no associated keyframes clearly represent non-mapped space. Therefore, at the end of the merging phase, we can mark a frame as a keyframe if the ratio between non-mapped surface voxels and total number of surface voxels in the TSDF is above a threshold, *i.e.* the frame describes a volume of the explored environment only partially overlapped with the current map.

Every time a new keyframe is detected and merged, a new optimization step is carried out to ensure consistency over long trajectories. Purposely, a pose graph is created and a suitable cost function defined over this graph structure is minimized by the G2O optimizer [15]. For each unknown keyframe pose we add a vertex to the graph and link it to the other vertex poses with edges representing constraints in the form of non-linear least-squares terms:

$$\mathrm{e}_{ab} = \left\|\mathbf{p}_a - \mathrm{T}_a^{-1}\mathrm{T}_b\mathbf{p}_b\right\|^2 \tag{3}$$

with $\mathbf{p}_a, \mathbf{p}_b \in \mathbb{R}^3$ representing two corresponding 3D points from keyframe a and b, $\mathrm{T}_a, \mathrm{T}_b$ the unknown poses of those frames. Although data association may be performed densely, for the sake of computational efficiency we overlay a regular grid onto the image and match only such sampled points [11]. Fig. 1 visually summarizes our novel matching scheme; in particular, for each sampled point we:

1. find its corresponding surface voxel by back-projection;
2. get the keyframe list of that voxel and project the 3D point onto all those image planes;
3. perform a local 2D search and select the nearest 3D point under some conditions.

It is noteworthy to point out that the keyframe list stored in every surface voxel allows for searching only in a well-defined subset of possible keyframes, thus achieving both

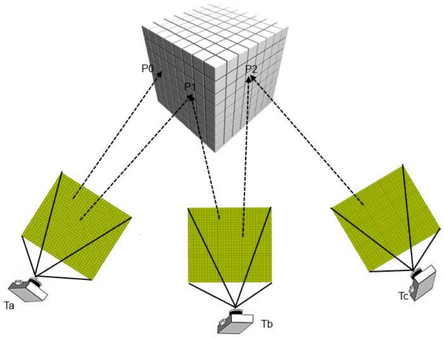


Fig. 1. For each point sampled on keyframes Ta, Tb, Tc we found its matches by projection onto all image planes relative to the keyframe list of the corresponding surface voxel

higher matching quality as well as faster speed. Moreover, the local search in the proximity of the projection ensures to find the nearest 3D point, possibly fulfilling also other requirements, such as e.g. alignment of surface normals [19]. The resulting set of vertex poses and constraints is finally optimized so as to achieve global error minimization.

Once keyframe poses have been jointly optimized, the TSDF no longer represents the mapped environment and it must therefore be recreated to reflect the new estimate of the trajectory. To allow for unbounded camera movements, the new reconstruction is centered at the last keyframe's camera pose. Though definitely more expensive, we claim that such volume shifting strategy is more effective than the proposal in [24] because:

- we do not re-center the active volume only looking at the translation part of the last estimated pose, but quantitatively evaluating the novelty of the brought information;
- we ensure consistent mapping by detecting keyframes to be optimized.

Finally, to reduce the computational effort, the optimization is usually limited to those keyframes having their camera reference frame inside the active volume. This approach was inspired by the Relative Bundle Adjustment framework [21].

3 Semantic KinectFusion

A standard SLAM approach would not deploy any kind of semantic information, such as *e.g.* the presence of known objects within the explored environment (see Fig. 2a), although this may set forth useful constraints concerning camera poses. Likewise, a major nuisance hindering standard object detection approaches in cluttered scenes deals with having to establish upon objects' presence and pose based on a single possibly unfavorable viewpoint, whereas a SLAM framework would in principle enable continuous incremental discovery from several known viewpoints (see Fig. 2b). In essence, unlike the state-of-the-art paradigms in both fields, we envision SLAM and object detection as tightly connected and synergistic rather than disjoint processes. Accordingly, we have recently proposed a novel Semantic Bundle Adjustment framework [10], whereby unknown object poses are explicitly included into the graph as pose vertexes constrained

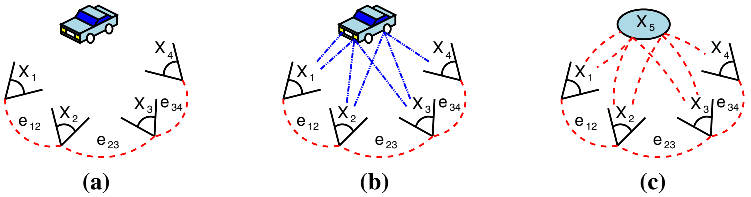


Fig. 2. A standard pose graph (a) ignores any semantic information. Instead, we include into the optimization object matches (b) as graph edges and the object pose as a vertex (c) so as to achieve object detection and improve SLAM.

to frames by a set of verified hypotheses (see Fig. 2c). Our proposal uniquely attain semantically constrained SLAM together with multi-view object detection and 6DOF localization.

When a frame is promoted to keyframe, 3D features, *e.g.* [13,23], are extracted and matched against those stored into a database of object models. Then, a set of candidate hypotheses on objects' presence and poses is drawn by RANSAC-based 6DOF pose estimation [1] and a *validation graph* is created to verify the consistency of each of such hypotheses with respect to the current 3D reconstruction. As detailed in [10], each validation graph is populated with all the vertex poses featuring object matches, their edges, a new vertex pose for the unknown object location and an edge for each matching feature:

$$\mathrm{e}_{fk}^{n} = s_{fk}^{n} \left\| \mathbf{p}_f - \mathrm{T}_f^{-1}\mathrm{T}_k\mathbf{p}_k^n \right\|^2 \tag{4}$$

$\mathbf{p}_k^n$ being the n^{th} 3D point feature on object k, $\mathbf{p}_f$ the matching 3D point feature on frame f, s_{fk}^n the matching score, T_f and T_k the unknown frame and object poses. Moreover, virtual edges are created when different camera frames match the same object feature [10]. Let $\mathbf{p}_{f_0} \in f_0$ and $\mathbf{p}_{f_1} \in f_1$ be two matches for object feature $\mathbf{p}_k^n$, then, under certain assumptions [10], we can add to the graph the virtual constraint:

$$\mathrm{e}_{fo}^{n} = s_{f_0k}^{n} s_{f_1k}^{n} \left\| \mathbf{p}_{f_0} - \mathrm{T}_{f_0}^{-1}\mathrm{T}_{f_1}\mathbf{p}_{f_0} \right\|^2 \tag{5}$$

The validation graph is optimized and a cleaning procedure is performed to detect wrong hypotheses (see [10] for more details). Finally, if the detection is confirmed, all the constraints are included into the global graph (see Sec. 2) and a global semantic optimization is performed.

Fig. 3 summarizes our overall proposal. Each incoming frame is fed to the tracking module (left box), which merges the new measurements with the current TSDF volume and checks the size of the non-mapped area to detect keyframes. If a new keyframe is spawned, 3D features are extracted and matched against models' features (right box), thus yielding hypotheses which have to be verified by the semantic bundle adjustment process. Finally, detected objects and keyframes are jointly optimized (bottom box) and the TSDF volume is updated.

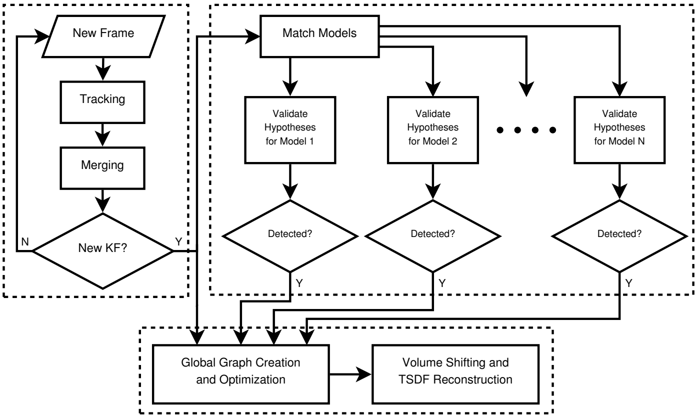

New Frame
Tracking
Merging
New KF?
N
Y
Match Models
Validate Hypotheses for Model 1
Validate Hypotheses for Model 2
Validate Hypotheses for Model N
Detected?
Detected?
Detected?
Y
Y
Y
Global Graph Creation and Optimization
Volume Shifting and TSDF Reconstruction

Fig. 3. A simplified view of our system

4 Results

The proposed approach has been extensively tested, both quantitatively as well as qualitatively. As for quantitative experiments, to the best of our knowledge available RGB-D datasets, proposed either for SLAM or 3D object detection, cannot be used to assess joint SLAM *and* 3D object detection, as the ground truth concerns camera trajectory only (in SLAM datasets) or, alternatively, object identities and poses only (in object detection datasets). Therefore, as proposed in [10], we rely here on a semi-synthetic setup: 3D object models scanned by a Kinect are rendered by ray-casting into SLAM datasets according to ground-truth camera poses. Although object detection turns out easier than in a real environment, we found such semi-synthetic datasets to deliver valuable insights and feedbacks concerning behavior and accuracy of the proposed method.

Hence, we took the video sequences from the RGB-D SLAM dataset [9], using the provided ground truth trajectory for precise model placement and ray-casting. Then, we also captured sequences by a Kinect camera featuring known objects in real environments, so as to provide also qualitative results on truly real data. As for the feature matching step underlying object detection, in all experiments we used the SIFT3D keypoint detector [16,20] and the Color-SHOT feature descriptor [23], using for both algorithms the implementations publicly available in the open-source Point Cloud Library [20].

Tab. 1 and 2 report the results obtained, respectively, with our implementation of RGB-D Mapping and with the extension to KinectFusion described in Sec. 2. As both such systems do not incorporate semantic information into the SLAM process, for their quantitative evaluation we can use the original RGB-D SLAM sequences [9]. As vouched by the Tables, we usually perform better, especially when considering the whole trajectory (*All Frs* column). Indeed, when the sensor comes back to an already mapped place, previous keyframes correctly enters the active TSDF volume thanks to the small drift and we can start tracking with respect to the known map. Instead, RGB-D Mapping looks for loop closure every time a new keyframe is detected, with no limit to their proliferation. Such effect is particularly evident in Fig. 4, which comes from the dataset used for qualitative evaluation (*i.e.* our own dataset without ground-truth

Table 1. Quantitative results with RGB-D Mapping (translation error wrt ground-truth camera poses). *Kfs* denotes the error at keyframes only, *All Frs* the error for all estimated camera poses.

	RGB-D Mapping					
	Mean Err. (m)		*Max Err. (m)*		*RMSE (m)*	
Sequence	Kfs	All Frs	Kfs	All Frs	Kfs	All Frs
FR1 Floor	0.056	0.093	0.1	0.244	0.06	0.104
FR1 360	0.089	0.197	0.211	0.4	0.099	0.208
FR1 Desk	0.069	0.074	0.163	0.181	0.076	0.081
FR1 Desk2	0.111	0.085	0.219	0.262	0.121	0.097

Table 2. Quantitative results with our extension to KinectFusion (translation error wrt ground-truth camera poses). *Kfs* denotes the error at keyframes only, *All Frs* the error for all estimated camera poses.

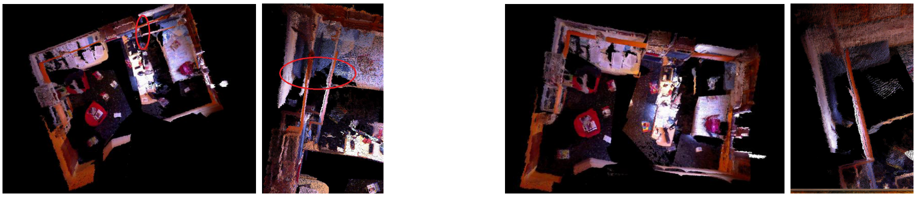
	Our Extension to KinectFusion					
	Mean Err. (m)		*Max Err. (m)*		*RMSE (m)*	
Sequence	Kfs	All Frs	Kfs	All Frs	Kfs	All Frs
FR1 Floor	0.056	0.054	0.093	0.13	0.062	0.058
FR1 360	0.083	0.183	0.159	0.324	0.091	0.192
FR1 Desk	0.04	0.055	0.131	0.167	0.047	0.058
FR1 Desk2	0.096	0.083	0.135	0.210	0.098	0.09

Table 3. Quantitative results with our Semantic KinectFusion (translation error wrt ground-truth camera poses). *Kfs* denotes the error at keyframes only, *All Frs* the error for all estimated camera poses.

	Our Semantic KinectFusion					
	Mean Err. (m)		*Max Err. (m)*		*RMSE (m)*	
Sequence	Kfs	All Frs	Kfs	All Frs	Kfs	All Frs
FR1 Floor	0.054	0.05	0.092	0.106	0.059	0.055
FR1 360	0.066	0.144	0.167	0.249	0.073	0.152
FR1 Desk	0.044	0.069	0.092	0.169	0.048	0.076
FR1 Desk2	0.042	0.085	0.088	0.132	0.046	0.088

information). Here camera motion follows a circular trajectory performing several loops around the room: at the end of the first loop, our technique hangs on to the known map, reducing the localization error and thus not affecting the reconstruction due to no new keyframe being spawned.

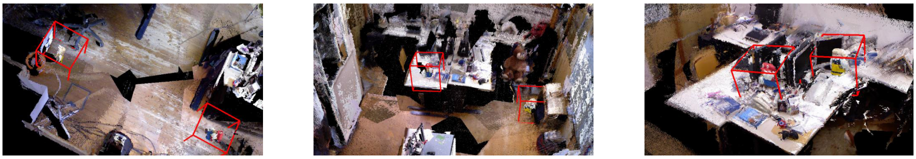
As for our proposed Semantic KinectFusion approach, Tab. 3 shows no reduction of accuracy, and often some improvements due to semantic information effectively



(a) RGB-D Mapping: final reconstruction (left) and a detail (right)

(b) Our extension to KinectFusion: final reconstruction (left) and a detail (right)

Fig. 4. We moved the camera on a circular trajectory and repeatedly acquired the same loop. Our proposal always tracks against the map, RGB-D mapping eventually drifts away.



(a) FR1 Floor sequence

(b) FR1 360 sequence

(c) FR1 Desk sequence

Fig. 5. Semi-synthetic RGB-D SLAM dataset: the semantic framework achieves high quality reconstruction and accurate object detection and localization.

constraining camera poses across many views. Moreover, beside a high quality reconstruction, the semantic framework can precisely localize the objects looked for, as shown in Fig. 5. Finally, as object detection is easier with the semi-synthetic setup than in real settings, we show also qualitative experiments on real data acquired by a Kinect camera in indoor environments. Fig. 6 reports the results obtained adding physically an object belonging to the model database into the same room as in Fig. 4. From one hand, our novel TSDF-enabled optimization (Sec. 2) achieves high mapping accuracy and low drift even on multiple loops; on the other hand, the Semantic Bundle Adjustment framework (Sec. 3) detects and localizes the object in the scene, thereby providing also semantic constraints which help improving the overall global alignment.

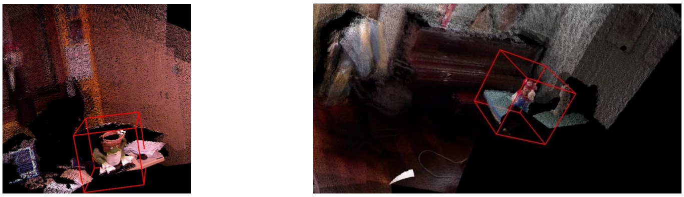


Fig. 6. We moved the camera on a circular trajectory and repeatedly acquired the same loop. TSDF-enabled optimization yields very low drift tracking, while semantic bundle adjustment detects and localizes objects (see the red boxes).

5 Concluding Remarks

A novel TSDF-based technique for bundle adjustment style optimization has been presented. Constraints across views are obtained by means of the TSDF representation of the environment and keyframes, alike, are selected with respect to such a representation. Then, we integrated a state-of-the-art Semantic Bundle Adjustment framework into the system, thus achieving effective joint detection, tracking and mapping, as vouched by both quantitative and qualitative experiments.

Future works will definitely concern real-time implementation on a modern GPU architecture, so as to extend the successful KinectFusion framework with global bundle adjustment and seamless object detection and localization. We also plan to investigate on whether and how effective features for object detection may be extracted directly from the TSDF reconstruction, which holds the potential to provide a more comprehensive and robust description than a single camera frame, as well as on how the TSDF-enabled BA could be extended to objects too. Finally, in this work we chose keyframes among camera frames; however, an interesting alternative we wish to explore deals with raycasting of the scene to get dense, low noise, virtual keyframes.

References

1. Arun, K.S., Huang, T.S., Blostein, S.D.: Least-squares fitting of two 3-d point sets. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 9(5), 698–700 (1987)
2. Bao, S.Y., Savarese, S.: Semantic structure from motion. In: CVPR (2011)
3. Besl, P.J., McKay, H.D.: A method for registration of 3-d shapes. PAMI 14(2), 239–256 (1992)
4. Chen, Y., Medioni, G.: Object modelling by registration of multiple range images. In: Proc. of the IEEE Int'l Conf. on Robotics and Automation, vol. 3, pp. 2724–2729 (April 1991)
5. Civera, J., Gálvez-López, D., Riazuelo, L., Tardós, J.D., Montiel, J.M.M.: Towards semantic SLAM using a monocular camera. In: Proc. of the Int'l Conf. on Intelligent Robot Systems (IROS), pp. 1277–1284 (2011)
6. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1996, pp. 303–312. ACM, New York (1996)
7. Davison, A.J.: Real-time simultaneous localisation and mapping with a single camera. In: IEEE Int'l Conf. on Computer Vision (ICCV), Washington, DC, USA, p. 1403 (2003)
8. Ekvall, S., Jensfelt, P., Kragic, D.: Integrating active mobile robot object recognition and slam in natural environments. In: IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (October 2006)
9. Endres, F., Hess, J., Engelhard, N., Sturm, J., Cremers, D., Burgard, W.: An evaluation of the RGB-D SLAM system. In: IEEE Int'l Conf. on Robotics and Automation (ICRA), St. Paul, MA, USA (May 2012)
10. Fioraio, N., Di Stefano, L.: Joint detection, tracking and mapping by semantic bundle adjustment. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA (2013)
11. Fioraio, N., Konolige, K.: Realtime visual and point cloud slam. In: Proc. of the RGB-D Workshop on Advanced Reasoning with Depth Cameras at Robotics: Science and Systems Conf. (RSS), pp. 27 (2011)

12. Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D.: Rgb-d mapping: Using depth cameras for dense 3d modeling of indoor environments. In: Proc. of Int'l Symp. on Experimental Robotics (ISER) (2010)
13. Johnson, A.: Spin-Images: A Representation for 3-D Surface Matching. Ph.D. thesis, Robotics Institute, Carnegie Mellon University (August 1997)
14. Klein, G., Murray, D.: Parallel tracking and mapping for small ar workspaces. In: IEEE and ACM Int'l Symp. on Mixed and Augmented Reality (ISMAR), pp. 225–234 (November 2007)
15. Kümmerle, R., Grisetti, G., Strasdat, H., Konolige, K., Burgard, W.: g2o: A general framework for graph optimization. In: ICRA, Shanghai, China (May 2011)
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–119 (2004)
17. Newcombe, R., Lovegrove, S., Davison, A.: Dtam: Dense tracking and mapping in real-time. In: IEEE Int'l Conf. on Computer Vision (ICCV), pp. 2320–2327 (November 2011)
18. Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohli, P., Shotton, J., Hodges, S., Fitzgibbon, A.: Kinectfusion: Real-time dense surface mapping and tracking. In: ISMAR, Washington, DC, USA, pp. 127–136 (2011)
19. Rusinkiewicz, S., Levoy, M.: Efficient variants of the ICP algorithm. In: Proceedings Third International Conference on 3-D Digital Imaging and Modeling, pp. 145–152. IEEE Comput. Soc. (2001)
20. Rusu, R.B., Cousins, S.: 3D is here: Point cloud library (PCL). In: IEEE Int'l Conf. on Robotics and Automation (ICRA), Shanghai, China, May 9-13 (2011)
21. Sibley, G., Mei, C., Reid, I., Newman, P.: Adaptive relative bundle adjustment. In: Robotics Science and Systems (RSS), Seattle, USA (June 2009)
22. Stühmer, J., Gumhold, S., Cremers, D.: Real-time dense geometry from a handheld camera. In: Goesele, M., Roth, S., Kuijper, A., Schiele, B., Schindler, K. (eds.) DAGM 2010. LNCS, vol. 6376, pp. 11–20. Springer, Heidelberg (2010)
23. Tombari, F., Salti, S., Di Stefano, L.: A combined texture-shape descriptor for enhanced 3D feature matching. In: 18th IEEE Int'l Conf. on Image Processing (ICIP), September 11-14, pp. 809–812. Brussels, Belgium (2011)
24. Whelan, T., McDonald, J., Kaess, M., Fallon, M., Johannsson, H., Leonard, J.: Kintinuous: Spatially extended Kinect Fusion. In: RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras, Sydney, Australia (July 2012)