15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Direct Methods for 3D Reconstruction and Visual SLAM

Daniel Cremers
Departments of Informatics & Mathematics
Technical University of Munich, Germany
http://vision.in.tum.de

Abstract

The reconstruction of the 3D world from camera images has a tradition of over 100 years. Nevertheless, we have witnessed over the last few years a dramatic boost in performance of reconstruction algorithms. An important innovation underlying this performance boost is the development of direct methods *to estimate the 3D structure and the camera motion. Some of these developments shall be reviewed in the following.*

1 Introduction

More than 100 years ago, researchers in the field of photogrammetry studied the problem of reconstructing the 3D world from multiple photographs. In 1913, the Austrian mathematician Erwin Kruppa [12] proved that given five corresponding point pairs observed in two images, one can recover the camera motion and the 3D point coordinates up to finitely many solutions. This and similar works paved the way for the development of computer vision algorithms to tackle the so-called *structure-and-motion problem* (later often called *visual SLAM*), culminating in algorithms like the linear 8-point algorithm [13] or the 5-point algorithm [15]. The first real-time capable algorithms appeared in the early 2000's [1, 3, 16, 10].

While Kruppa's work was undoubtedly pioneering for the field of image-based 3D reconstruction, it may also have been highly misleading as the key assumptions underlying Kruppa's work are typically not fulfilled in applications of camera-based reconstruction:

- When switching on a camera, we typically do not observe a sparse set of points but rather a sheer endless amount of colors. How to optimally select a sparse subset of points out of these is far from obvious and has been tried with numerous (often heuristically motivated) keypoint detectors. Algorithms which then solely use the respective keypoints will invariably be suboptimal because they do not exploit all available sensor data.

- Even after selecting points in respective images, these points are not in correspondence. Computing this correspondence – essentially the problem of optical flow estimation – is a classical ill-posed problem and one of the nastiest computational challenges in computer vision. While there exist numerous algorithms and strategies such as Ransac to determine this point correspondence, respective solutions are invariably suboptimal. Obviously errors in this point correspondence will propagate to errors in the reconstruction in a more or less benign manner depending on the choice of cost function.

In recent years, we have witnessed a surge of so-called *direct methods* for 3D reconstruction and visual SLAM. Some of these developments shall be sketched in the following.

2 Direct Methods for Dense Geometric Reconstruction

While the estimation of dense correspondence is a difficult computational challenge (respective spatially discrete formulations being NP hard), for a given set of camera locations and orientations one can tackle the problem of dense geometric reconstruction without explicitly computing correspondence: For any voxel in the reconstruction volume $V \subset \mathbb{R}^3$, one can determine a value of photoconsistency $\rho : V \to [0, 1]$, which takes on small values if the projection of that voxel into various cameras gives rise to a consistent color (or local texture) and large values for voxels which give rise to very different color (or local texture). Subsequently, one can compute an optimally photoconsistent surface S by minimizing the photoconsistency-weighted surface area

$$\min_S \int_S \rho(s)\, dA(s), \tag{1}$$

where $dA(s)$ denotes the surface area element located at point $s \in S$. In their pioneering paper [9], Faugeras and Keriven suggested to minimize such functionals by means of level set methods thereby computing a spatially dense locally optimal reconstruction. A more generative formulation (modeling foreground and background colors) which can be interpreted was a silhouette-based dense reconstruction was proposed by Yezzi and Soatto in [22].

While such locally optimal solutions are often highly dependent on the initialization (in particular with respect to topology of the reconstructed object), later works of Kolev and coworkers [11] showed that provably optimal solutions for functionals of form (1) – in combination with suitably defined unary terms, since the optimum of (1) is obviously the empty set – can be computed by means of convex relaxation methods. This framework was later extended to computing provably silhouette-consistent reconstructions [2] and to computing spatio-temporal reconstructions [17]. The key practical implication of these convexifications is that the computed solutions are independent of the choice of initialization and either provably optimal or within a computable bound of the optimum. As evident from the results in Figure 1, these purely camera-based dense reconstructions are at a level of precision where the entire rope of the rope-jumping girl can be reconstructed.

While the computation of spatially dense volumetric reconstruction is typically time-consuming – the reconstructions in Figure 1, for example, took around 3

© 2017 MVA Organization

Figure 1. Reconstruction of spatio-temporal actions from multiple synchronized videos using convex relaxation methods [17]. In contrast to local optimization techniques, the convex relaxation methods provide solutions that are independent of initialization and either provably optimal or within computable bounds of the optimum. Input video data is courtesy of http://4drepository.inrialpes.fr.

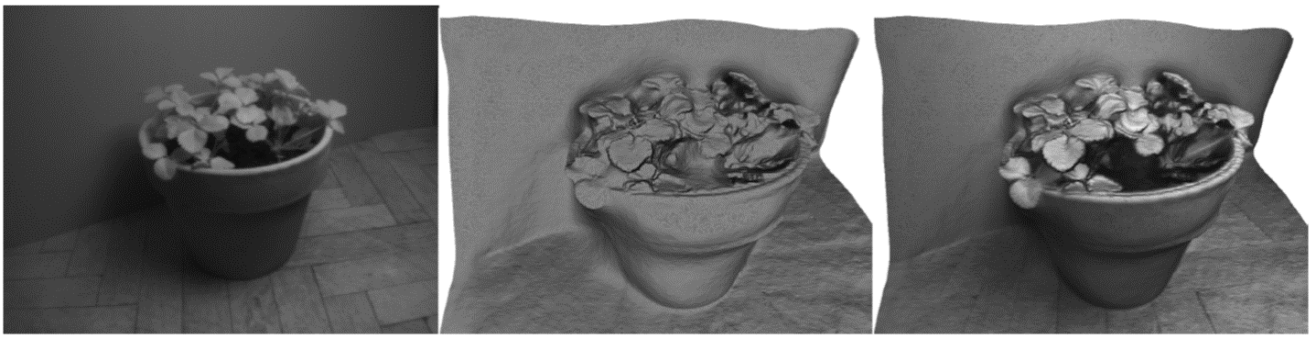

Figure 2. Real-time dense reconstruction from a handheld camera computed using a variational methods [19]. A multiple-view data term is combined with a smoothness regularizer which creates a geometric fill-in effect in locations of lacking observations, for example at object boundaries. The resulting solutions are reminiscent of a soap film settling over the observed 3D scene.

minutes per time step on a high-performance GPU – for many applications such as driver assistance and autonomous robots one may want to compute spatially dense reconstructions in realtime. One of the first dense reconstruction in realtime was proposed by Stuehmer and coworkers who suggested to compute a dense depth map $u : \Omega \to \mathbb{R}$ on the image plane $\Omega \subset \mathbb{R}^2$ by minimizing the functional

$$\min_u \sum_{i=1}^{n} \int \left| I_1(x) - I_i(\pi\, g_i\, u\, x) \right| dx \;+\; \lambda \int |\nabla u|\, dx, \quad (2)$$

where x is given in homogeneous coordinates (such that $u\,x$ is the corresponding 3D coordinate) and $g_i \in SE(3)$ is the rigid body motion between camera 1 and camera i . The total variation regularizer on u creates a geometric fill-in effect in locations of lacking observation (for example at object boundaries) – see Figure 2. Conceptually related approaches were subsequently published in [18] and [21].

3 Direct Tracking and Mapping (DTAM)

The work [18] tackles the joint problem of camera motion estimation and structure reconstruction by means of a direct method, i.e. without computing feature points and correspondence (using the feature point based method PTAM only for initialization). As apparent from equation (2) the central idea underlying these direct methods is the notion of image alignment: Rather than minimizing a geometric error (as done for example in bundle adjustment) the goal is to minimize a photometric error of color or brightness consistency. Yet, rather than in optical flow estimation one does not compute a 2D correspondence field but rather directly the geometry u.

4 Large-Scale Direct (LSD) SLAM

Engel and coworkers [5] proposed a direct method to compute a semi-dense geometry and the camera motion in real-time on a simple laptop CPU (rather than requiring GPUs as in [18]). In a follow-up publication [4], Engel et al. introduced the notion of *pose graph optimization* in order to carry the visual SLAM algorithm to a large scale capability. The key idea of this technique (adopted from laser-based approaches) is to compute (in parallel to the camera motion $\tilde{g}_{ij} \in SE(3)$ estimated by alignment of images i and j) a globally consistent camera trajectory $g_i \in SE(3)$ by minizing the nonlinear least squares problem

$$\min_{g_1,\ldots,g_n} \sum_{i \sim j} d(\tilde{g}_{ij}, g_i \circ g_j), \quad (3)$$

with a suitably chosen metric $d(\cdot,\cdot)$ which imposes consistency of the respective rigid body motions. Figure 3 shows examples of large-scale reconstructions obtained with this method.

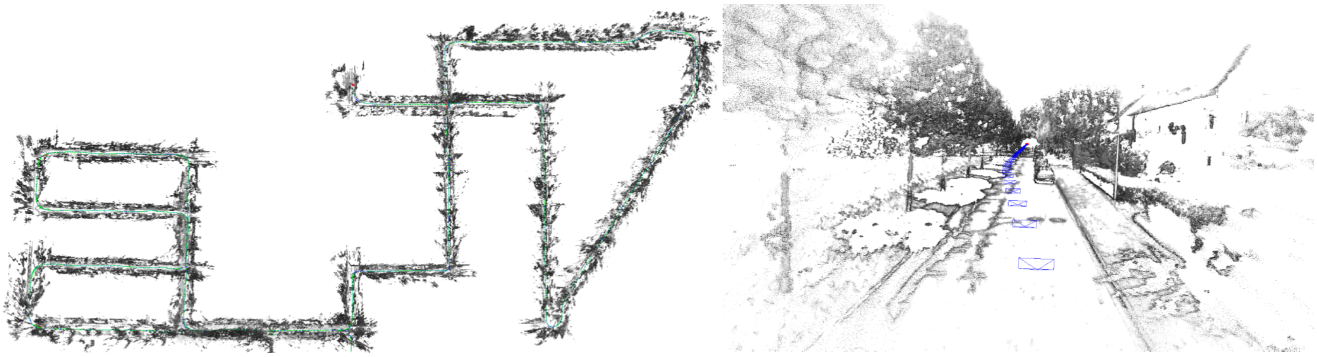

Figure 3. Real-time reconstruction of camera trajectory and environment computed on a laptop CPU using a stereo-version of Large-Scale Direct (LSD) SLAM [7, 20]. In contrast to the monocular version of LSD SLAM [4], the stereo version has several advantages: Firstly, the stereo version can recover reconstructions at scale since the baseline between the two cameras is known. Secondly, the additional information from static stereo facilitates the initialization and increases robustness and accuracy.

5 Direct Sparse Odometry

The main difference between keypoint-based approaches and direct approaches to visual SLAM is that in keypoint-based methods the overall problem is split into two consecutive steps – namely extracting and matching keypoints and (subsequently) solving the structure and motion or SLAM problem. In contrast, direct methods tackle the overall problem in a single step. The cost functions used for the SLAM computation therefore differ: In keypoint-based methods the cost measures a *geometric reprojection error* of respective 3D points projected into respective images. In contrast, in direct methods the cost underlying the SLAM computation corresponds to a *photometric error* of imposing color or brightness consistency across all images – see equation (2). As a consequence, direct methods model directly the image formation process and the sensor measurements (colors).

An immediate advantage of direct methods is that the cost function to be optimized can be designed to accurately encode the image formation process. In fact, if one studies real-world cameras one comes to realize that the brightness of 3D points observed in multiple images is by no means constant, even for Lambertian objects. The brightness measured in a given pixel depends on numerous aspects, including the camera aperture, the vignette and the gamma correction.

Engel and coworkers [6] precisely modeled this brightness transformation and computed a pixel-accurate vignette. In addition the made use of a marginalization strategy to take into account the older image observations. The resulting algorithm called *Direct Sparse Odometry* leads to highly accurate point-clouds and camera tracks with nearly no drift or distortion – see Figure 4. These are be computed in real-time from a handheld moving camera.

A systematic quantitative evaluation of *Direct Sparse Odometry* to *ORB SLAM* [14], a state-of-the-art keypoint based visual SLAM algorithm shows a drastic improvement in robustness and accuracy – see Figure 5.

6 Conclusion

We have reviewed a number of recent developments in real-time structure and motion, often referred to as real-time visual simultaneous localization and mapping (visual SLAM). In particular, we focussed on the development *direct methods*. These differ from the more traditional keypoint based methods in that they do not separate the two steps of point correspondence estimation and SLAM, but rather solve both problems directly by minimizing a photoconsistency error (rather than a geometric reprojection error). Based on this more direct modeling of the sensory measurements (the colors or brightnesses of pixels), they can directly incorporate accurate models of the image formation process, including aspects such as lens attenuation (vignetting), exposure time and gamma correction. As a result, this replaces the classical "brightness constancy assumption" to an "irradiance constancy assumption", i.e. the light emmitted (!) from a given 3D point is assumed to be constant and direction-independent.

This direct modeling of the sensory measurements and the image formation process gives rise to substantial improvements of visual SLAM algorithms, both with respect to accuracy and with respect to robustness. As a consequence, they can be deployed in challenging real-world scenarios to map large-scale environments (entire street passages) and determine a highly-accurate and nearly drift-free camera trajectory. We believe that these direct visual SLAM algorithms will form core ingredients for many technologies such as smart-phone applications, self-driving cars and other robotic systems.

References

[1] Alessandro Chiuso, Paolo Favaro, Hailin Jin, and Stefano Soatto. 3-d motion and structure from 2-d motion causally integrated over time: Implementation. In *European Conference on Computer Vision*, pages 734–750. Springer, 2000.

[2] D. Cremers and K. Kolev. Multiview stereo and silhouette consistency via convex functionals over convex domains. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1161–1174, 2011.

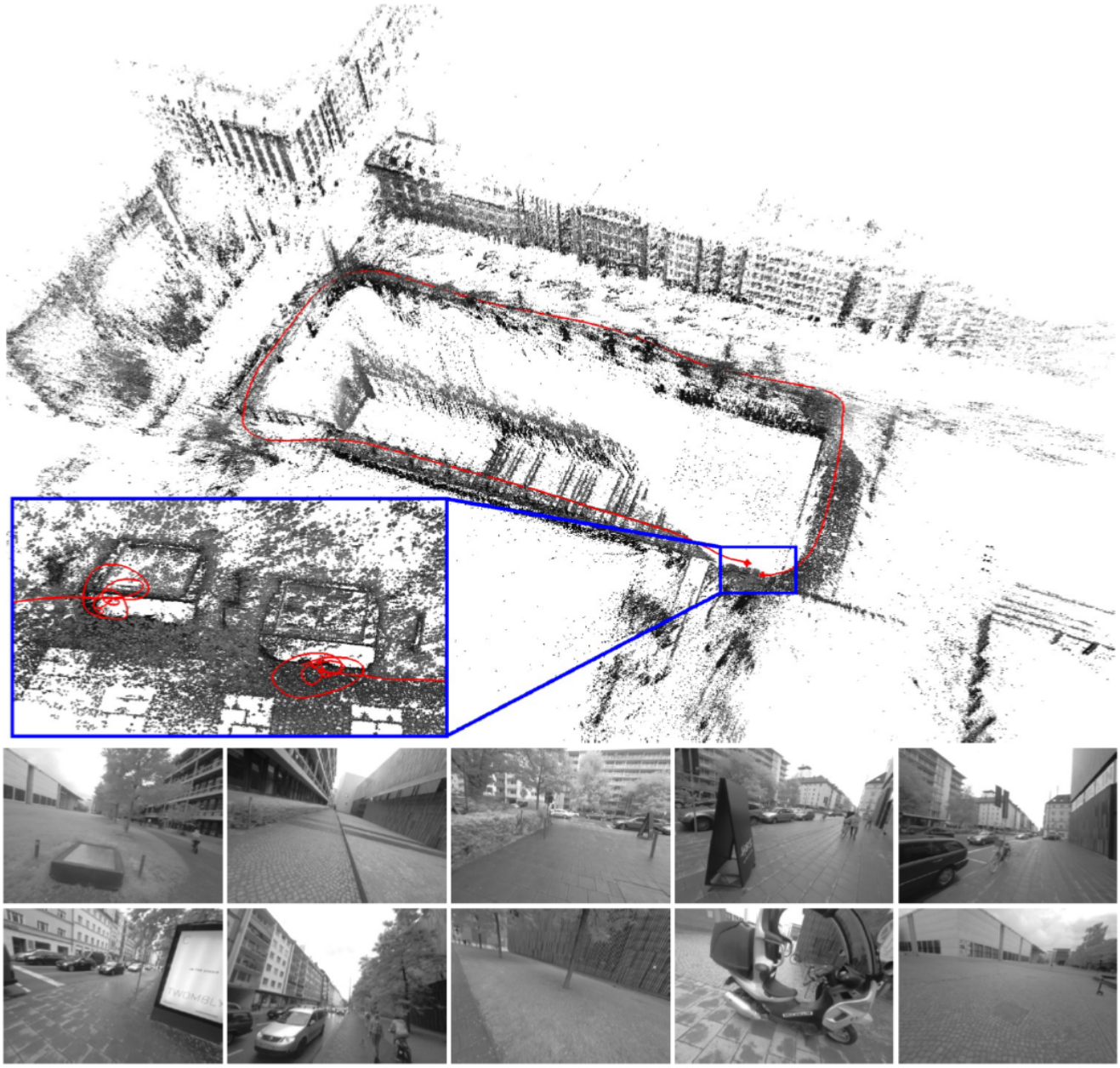

Figure 4. 3D point cloud and camera trajectory computed from a 1:40 minute monocular video recorded from a bicycle using Direct Sparse Odometry (DSO) [6]. Without any pose graph optimization the camera trajectory and point cloud computed with DSO are highly accurate and nearly distortion- or drift-free even on a large scale. The bottom rows show some of the input frames.

[3] Andrew J Davison. Real-time simultaneous localisation and mapping with a single camera. In *ICCV*, volume 3, pages 1403–1410, 2003.

[4] J. Engel, T. Schöps, and D. Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *European Conference on Computer Vision*, September 2014.

[5] J. Engel, J. Sturm, and D. Cremers. Semi-dense visual odometry for a monocular camera. In *IEEE Int. Conf. on Computer Vision (ICCV)*, December 2013.

[6] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. To appear.

[7] Jakob Engel, Jörg Stückler, and Daniel Cremers. Large-scale direct slam with stereo cameras. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, pages 1935–1942. IEEE, 2015.

[8] Jakob Engel, Vladyslav Usenko, and Daniel Cremers. A photometrically calibrated benchmark for monocular visual odometry. *arXiv preprint arXiv:1607.02555*, 2016.

[9] O. Faugeras and R. Keriven. Variational principles, surface evolution, PDE's, level set methods, and the stereo problem. *IEEE Trans. on Image Processing*, 7(3):336–344, March 1998.

[10] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 225–234. IEEE, 2007.

[11] K. Kolev, M. Klodt, T. Brox, and D. Cremers. Con-

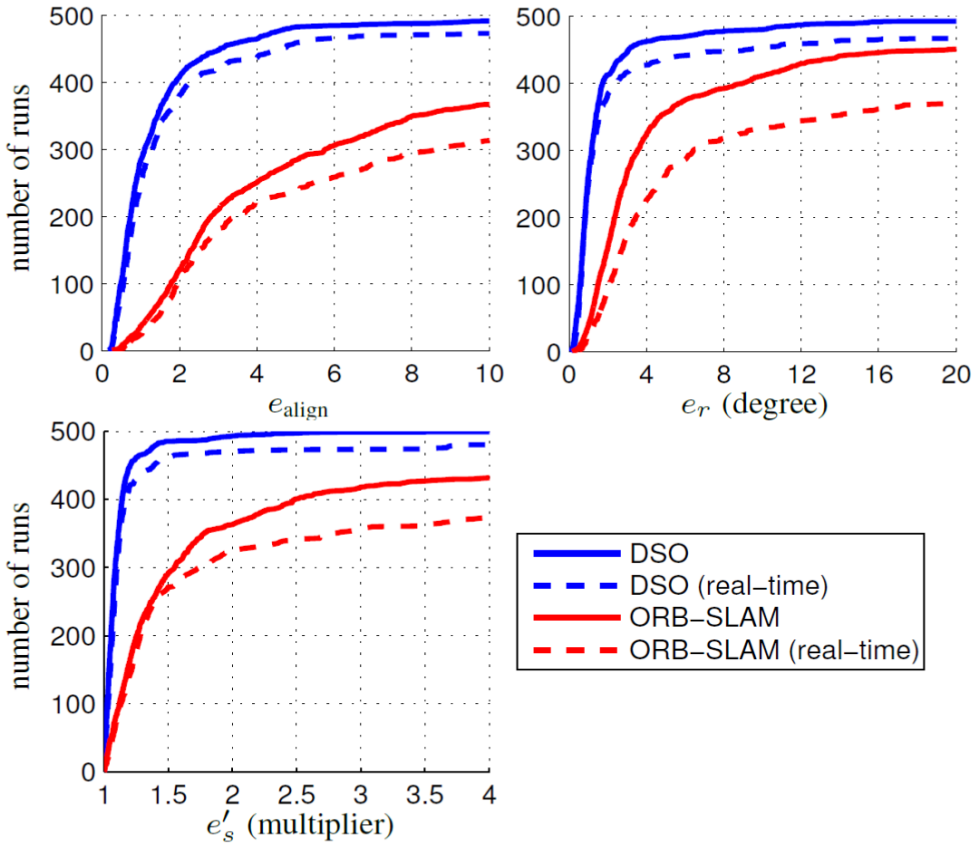

Figure 5. Quantitative comparison of Direct Sparse Odometry (DSO) [6] against the keypoint-based competitor ORB-SLAM [14] evaluated on the TUM-monoVO dataset containing 50 realworld sequences [8]. The plots show the number of runs for which the alignment error and the drifts in rotation and scale are below a given threshold. Clearly, DSO is substantially more accurate, both in terms of accuracy and robustness. For example, while ORB SLAM tracks around 100 sequences with an alignment error below 2, DSO tracks around 400 sequences with an alignment error below 2. Similarly, while for ORB SLAM the best 300 sequences are tracked with a maximal error of 6, DSO merely has a maximal case error near 1.

tinuous global optimization in multview 3d reconstruction. *International Journal of Computer Vision*, 2009.

[12] Erwin Kruppa. *Zur Ermittlung eines Objektes aus zwei Perspektiven mit innerer Orientierung*. Hölder, 1913.

[13] H Christopher Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms, MA Fischler and O. Firschein, eds*, pages 61–62, 1987.

[14] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

[15] David Nistér. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):756–770, 2004.

[16] David Nistér, Oleg Naroditsky, and James Bergen. Visual odometry. In *Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–I. Ieee, 2004.

[17] Martin Ralf Oswald, Jan Stühmer, and Daniel Cremers. Generalized connectivity constraints for spatio-temporal 3d reconstruction. In *European Conference on Computer Vision*, pages 32–46. Springer, 2014.

[18] S. J. Lovegrove R. A. Newcombe and A. J. Davison. DTAM: Dense tracking and mapping in real-time. In *IEEE Int. Conf. on Computer Vision (ICCV)*, 2011.

[19] J. Stühmer, S. Gumhold, and D. Cremers. Real-time dense geometry from a handheld camera. In *Pattern Recognition (Proc. DAGM)*, pages 11–20, Darmstadt, Germany, September 2010.

[20] Vladyslav Usenko, Jakob Engel, Jörg Stückler, and Daniel Cremers. Reconstructing street-scenes in real-time from a driving car. In *3D Vision (3DV), 2015 International Conference on*, pages 607–614. IEEE, 2015.

[21] Andreas Wendel, Michael Maurer, Gottfried Graber, Thomas Pock, and Horst Bischof. Dense reconstruction on-the-fly. In *Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1450–1457. IEEE, 2012.

[22] A. Yezzi and S. Soatto. Stereoscopic segmentation. *Int. J. of Computer Vision*, 53(1):31–43, 2003.