A brief survey of visual odometry for micro aerial vehicles

Conference Paper · October 2016

DOI: 10.1109/IECON.2016.7793198

7 authors, including:

mo Shan
University of California, San Diego
14 PUBLICATIONS **20** CITATIONS
SEE PROFILE

Zhi Gao
National University of Singapore
54 PUBLICATIONS **646** CITATIONS
SEE PROFILE

Feng Lin
National University of Singapore
58 PUBLICATIONS **267** CITATIONS
SEE PROFILE

Ben M Chen
National University of Singapore
417 PUBLICATIONS **6,672** CITATIONS
SEE PROFILE

A Brief Survey of Visual Odometry for Micro Aerial Vehicles

Mo Shan*, Yincai Bi†, Hailong Qin§, Jiaxin Li†, Zhi Gao*, Feng Lin*, Ben M. Chen‡
*Temasek Laboratories, National University of Singapore
†NUS Graduate School for Integrative Sciences & Engineering, National University of Singapore
‡Department of Electrical & Computer Engineering, National University of Singapore
§Department of Mechanical Engineering, National University of Singapore

***Abstract*—Recently, visual odometry (VO) has experienced a rapid growth, which makes it viable for a range of applications. This survey paper attempts to provide a timely and comprehensive review of this field, focusing specifically on micro aerial vehicles (MAVs), with monocular, stereo or RGB-D cameras onboard. In the survey, the milestones in the development of VO will be reviewed, followed by an illustration of its general workflow, the commonly used datasets. The survey is concluded by an overall discussion.**

I. Introduction

VO is essential for the localization and navigation of MAVs in GPS-denied environments. It is aimed at calculating the egomotion of the MAVs by estimating the pose incrementally using onboard cameras. A formal definition of VO is presented in the tutorial [1], [2]. Suppose the MAVs are flying in an environment while taking images with a camera attached rigidly to its body, at discrete time instants, then the camera positions at adjacent time instants are related by the rigid body transformation $T \in \mathbb{R}^{4\times4}$. The objective of VO is to compute the inter frame transformations T and concatenate them to form the full trajectory.

Considering the current approaches of VO, there are feature based, pixel intensity based, or hybrid approaches, based on the differences in the image matching phase. A brief taxonomy of feature based and direct methods could be found in [3]. Despite the variation in implementation, a similar general framework is shared, as shown in Fig. 1. An overview of the VO workflow will be discussed next, and different approaches are compared and contrasted.

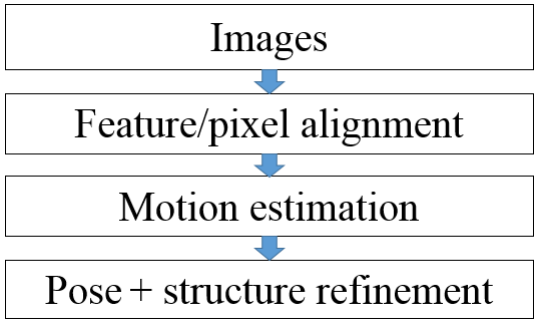

Fig. 1: A block diagram showing the general workflow of VO.

A. Feature based VO

The basic steps for sparse feature based VO includes feature extraction and association, initialization, sparse feature tracking and mapping performed sequentially.

1) Feature extraction and association: To improve the efficiency of matching, the saliency feature are searched instead of directly matching the whole image. The general local feature is defined as an image pattern which has a profound difference compared with its neighborhood. For the developments in recent years, SIFT [4] and SURF [5] are scale invariant. Other new features such as BRIEF [6] are not rotation invariant. The recent ORB [7] feature, which achieves great success in visual SLAM, is a fusion of FAST and BRIEF which uses pyramid to generate multiscele representations. The robustness and efficiency of these features have been extensively studied in [1], it shows that there does not exist perfect feature detectors, and therefore a robust VO framework is needed. To match the detected feature, a feature descriptor such as SIFT and SURF is constructed to describe the region around the detected feature. Using the appearance information around the feature is the most straightforward approach. However, it is not stable because of the illumination and scale variation. Besides feature descriptors, local search like correlation also can be used for matching.
2) Initialization: Estimating the motion and creating the 3D sparse feature map (up to an arbitrary scale) is a critical step in the current VO framework. A good initial map is essential for not only the tracking thread to obtain correspondence, but also propagating the map incrementally to maintain the scale. The classical approach is to use the homography model by assuming a planar scene for initialization [8]. However, the homography only works well in the planar and low parallax scene. Solving the fundamental matrix, in contrast, can work in general condition. A heuristic selection criteria is established to initialize the motion and map by either using homography matrix or fundamental matrix [9].
3) Sparse feature tracking and mapping: Many VO approaches adopt the parallel tracking and mapping framework used in PTAM [10]. The sparse local tracking

and mapping are maintained as two threads, which are used mutually to estimate the egomotion and extend the existing map simultaneously. To improve the robustness, keyframes are maintained so that the optimization is more efficient and the drift is reduced. Once the keyframe is determined, the existing map will be projected on the current keyframe to search for the correspondence. The local bundle adjustment is used to minimize the reprojection error and the new map points are triangulated by connected keyframes [11].

B. Direct VO

Instead of computing the correspondence based on detected features, the direct approach performs the image alignment directly on pixel by using photoconsistency constrains. Compared with sparse feature based approach, the direct approach does not require an engineered feature since it works on the pixels. Nevertheless, GPUs are often used to make the per-pixel depth estimation viable for real time applications. The basic steps for direct VO contains dense tracking and depth map estimation illustrated below.

1) Dense tracking: Similar to sparse feature based approach, the egomotion of the camera can be obtained by minimizing the tracking error. However, the direct VO utilizes the photometric differences or its residual difference as cost function.
2) Depth map estimation: The depth map is initialized by triangulating the first keyframe pair and propagated by projecting points from new keyframes. The created keyframe map can be refined by non keyframes [12].

In addition to the approaches mentioned above, hybrid methods have also been proposed, such as [3]. These approaches do not depend on feature description and correspondence, thus the computational cost is reduced dramatically compared with feature based ones.

The rest of the paper is organized as follows: Section II presents the existing survey papsers; the milestones of the development in VO are illustrated in Section III; Section IV contains datasets for performance evaluation; Section V consists of the discussion.

II. Related works

The tutorial [1], [2] covers the research on VO from 1980 to 2011, which surveys the field from different technical perspectives, ranging from camera modeling, calibration, motion estimation pipelines, to feature matching, optimization. The tutorial also differentiates VO from structure from motion (SFM), which deals with both the camera poses and the structure using unordered images, and refines the structure and poses offline. In contrast, VO is an sequential estimation of camera poses in real time, and the use of local optimization such as bundle adjustment is optional. Furthermore, VO can be regarded as a building block of visual simultaneous localization and mapping (SLAM), which is reviewed in [13], because the recovery of local path is required before loop closure detection.

In addition to the tutorial, there are several surveys on existing VO works, such as [14], which reviews the approaches divided into filtering-based and optimization-based. For the first category, camera is used to obtain measurement and IMU is used for prediction in the EKF framework. For the second category, there are two stages, namely mapping and tracking. During mapping, feature detectors are used to extract features, whose reprojection error is used as a cost function to be optimized to find the coordinates of features. During tracking, the reprojection of features are used to find the changes in position and orientation with an optimization algorithm. The reason to separate these two processes is that tracking is much faster than mapping.

There is also a recent paper on inertial aided VO [15], which reviews the relevant topics briefly. In this field, the two dominant concepts are batch nonlinear optimization as well as recursive filtering. The former minimizes both the reprojection error and camera movement from IMU, whereas the latter only use the camera movement to propagate the state update from visual cue. Though the batch based approaches may produce more accurate results despite the presence of outliers, their computational load is also higher than filtering methods. Another way of categorization is loosely coupled versus tightly coupled. On one hand, loosely coupled approaches estimate the pose first and then fuse with IMU. On the other hand, tightly coupled methods considers the camera pose and IMU jointly, leading to more precise results.

Despite that survey papers for VO already exist, its fast paced development necessitates an updated taxonomy on this topic to facilitate the application of these approaches. Besides the timeliness, this survey also differs from previous papers as it focuses exclusively on MAVs. Meanwhile, this review will accommodate different types of cameras that are commonly used, especially monocular, stereo, as well as RGB-D cameras.

III. State-of-the-art approaches

The intent of this section is to discuss recent works on VO, not to present an exhaustive evaluation, but to draw attention to some key developments. To present these approaches more clearly, they are divided into different categories based on the type of camera used.

A. Monocular odometry

Stephan et al. [16] propose to use camera and IMU to estimate the speed of the MAV. The up-to-scale translation obtained from camera is fused with IMU in EKF to recover the metric scale. Unlike previous approaches, the calibration of camera and IMU is not necessary as it is handled in EKF. The VO depends only on the optical flow between two frames and the IMU data, and thus it runs at a high frequency of 40 Hz. The speed estimate is used as initialization and back up for PTAM. The authors extend their work and propose an approach without relying on SLAM in [17]. The terrain plane is found by regression of the point cloud of the scene, and the plane is assumed to be locally planar. Using the scaled camera to plane distance, the scale estimate in EKF is more

accurate. Specifically, by setting the origin on the terrain plane, the global position of the MAV and its yaw are observable. As a result, the position drift is limited parallel to the plane, and the MAV can perform terrain following in large environments.

Forster et al. [3], [18] have proposed a semi-direct visual odometry (SVO) pipeline that uses two threads for motion estimation and mapping. For initialization, FAST corners are detected evenly in the image and then tracked by optical flow to estimate a homography, assuming a planar scene. The initial map is obtained via triangulation. In the motion estimation thread, the relative pose is found by minimizing the photometric error between corresponding pixels, refined by alignment of feature patches, and optimized by bundle adjustment. In the mapping thread, if the Euclidean distance of the new frame is large, a keyframe is selected. Each feature in the keyframe is initialized using a depth filter to estimated the 3D piont. The depth estimate is updated in the subsequent frames until its uncertainty is small enough to be inserted in the map. This work produces a map where the scale drift is very small. This scale can be recovered by sensor fusion.

Bloesch et a. [19] present a robust visual inertial odometry (ROVIO) framework that uses the pixel intensity directly instead of features. The multilevel patch features are detected and tracked, which is coupled to the EKF. the estimation of landmarks is robocentric and its position is parameterized by a bearing vector and a distance. Consequently, the initialization process could be omitted. It is used for onboard feedback control for take off and landing.

B. Stereo odometry

As the name implies, stereo visual odometry applies a well-calibrated stereo rig to determine the egomotion in all six degrees of freedom that are possible in a 3D world: three for translation and three for rotation. The feature-based stereo VO methods typically proceed through such steps, as summrized in Fig. 1, and typical works includes [20], [21], [22], [23], [24], [25], [26], [27], [28], [29]. By and large, the rotation and translation parameters are obtained via aligning corresponding 3D pont clouds (which is estimated readily by performing triangulation) between consective stereo pairs. Unsurprisingly, most methods of this category share the framework, but with more or less innovations in one or some of these components.

In the steps of feature detection and association which are closely related, a variety of feature detectors have been exploited together with different association strategies, tracking or matching, according to the given hardware conditions and the imposed target of working scenarios. Tracking is to find correspondences of features appeared in previous frame using local search techniques, thus it is suitable to process images taken from nearby viewpoints. For example, such corner feature detectors Moravec, Forstner, KLT, FAST have been popularly used for tracking in the stereo VO.

Matching is to independently detect features in each image and match them based on some similarity metric between their descriptors, resulting in the applicability of dealing with large motion or viewpoint change. For example, SIFT, SURF, CENSURE, and ORB features have been deployed for matching in stereo VO [30], [31]. For a detailed evaluation of feature detectors and descriptors for indoor and outdoor VO, readers can refer to [32] and [33] respectively. To eliminate the outliers in feature association, which could severely deteriorate the final motion estimation results, RANSAC [34] has been established as the standard, and an interesting clique-based inlier detection method is proposed in [20].

To perform pose estimation, there will be three different categories of methods according to whether the feature correspondences are specified in two or three dimensions. Firstly, in the 2D-to-2D category, the camera pose is extracted from the essential matrix using image point correspondences [35]. The second category is to use 3D-to-2D point correspondences to calculate extrinsic matrix, and the typical work includes [36]. Thirdly, the 3D-to-3D point clouds alignment can be conducted to estimate the camera pose, see the work [22], [23], [21], [25]. Moreover, it is worthwhile to mention that the work [37], [38] and [39] exploited trifocal and quadrifocal tensors respectively, which are fed into EKF to get the camera motion.

Several recent works worth particular attention. Kitt et al. [40] propose a robust visual odometry method (LIBVISO1) based on the observation of trifocal tensor between image triples. This method does not need to calculate the 3D scene structure and hence it reduces the computational cost. The trifocal tensors are fed into an EKF to get the relative camera motion. A RANSAC outlier rejection scheme is deployed to reduce long-term drift of the odometry. They also use a feature bucketing technique to ensure that all the features are well distributed in the image. Gieger et al. [41] propose an improved method (LIBVISO2) which outperforms the original approach both in accuracy and runtime. They detect corner features using non-maximum-suppression and calculate the egomotion with an optimization of image projection errors. Furthermore, they demonstrate a real-time 3D reconstruction based on the fusion of dense disparity [42].

Leutenegger et al. [15] present a tightly coupled open keyframe-based visual-inertial (OKVIS) approach, which is demonstrated to benefit in both accuracy and robustness comparing with vision only and loosely coupled methods. A probabilistic derivation of IMU error terms and non-linear optimization are developed to jointly estimate the camera motion. The system employs keyframe technique, and it matches keypoints and rejects outliers using inertial cues. Their work belongs to a leading trend in the MAV state estimation research to shift from filtering techniques to optimization methods.

In addition, some stereo VO works have deployed special visual sensors. In [43], a pair of thermal cameras have been applied to realize localization for UAVs in night-time. The work [44] and [45], belonging to the feature-based and the direct categories respectively, exploit pair of fisheye cameras, and competitive performance is reported.

C. RGB-D odometry

Huang et al. [46] demonstrate one of the initial odometry and mapping systems, termed the fast odometry from vision (FOVIS), based on the first commercialized RGB-D camera named Kinect. They use FAST to detect feature points and reject outliers by building and checking a consistency graph, based on distance preserving under rigid body motion. A keyframe technique is used to eliminate short scale drift, which calculates the camera motion based on the reference frame. This works well in the situation of the MAV hovering because of the constraint on signal-noise-ratio.

Kerl et al. [47] propose a novel dense visual odometry (DVO) approach, which is directly based on the minimization of photometric errors between two consecutive frames. This method is different from feature based methods because it does not need feature detection and matching. Instead, it relies on the photo-consistency assumption and use a probabilistic representation. The pose is optimized by the Maximum A Posteriori estimation. A comparative study of this approach and alternative methods is presented in [48].

Zhang et al. [49], [50], [51], [52] have published several papers on VO using depth from RGB-D camera or laser scanner. The Harris corners are evenly detected across the image, and then they are tracked using optical flow from frame to frame. Based on rigid body transformation, each corner with known depth contributes two equations, whereas the one with unknown depth contribute one equation. Stacking the equations together and solving it using the levenberg-Marquardt (LM) method give the translation and rotation. The depth map is acquired by RGB-D camera or laser scanner, but if the depth of the feature is unknown, the local planar patch containing the feature with three points with known depth is searched. Moreover, if the feature is tracked for a long distance, the depth is obtained through triangulation. The frame to frame motion is refined by bundle adjustment.

Zheng et al. [53], [54], [55] focused on the application of autonomous flight in downgraded environment using RGB-D sensor. In [53], they compare and evaluate several state-of-the-art real-time odometry methods in different challenging scenarios. The investigated methods include feature based VO, direct dense VO, point cloud based odometry. The evaluation scenarios consists of less features or even featureless case. By analysing each method, they comment on the pros and cons of each category regarding the robustness and accuracy. In [54], they further propose a fast odometry method working in downgraded environment based on the range rate constraint equation and photometric error metric. A particle filter is deployed to reduce the drift with the help of a given 3D map.

IV. Benchmark

To evaluate the performance of VO objectively, datasets with known camera intrinsics and the ground truth trajectory are required. This section will present some popular datasets produced by different sensor setups.

A. Monocular camera

Zhang et al. [56] provide two synthetic scenes generated using Blender. One is a vehicle moving in a city, and another is a flying robot hovering in a confined room. Three different lens (perspective, fisheye and catadioptric) are used in each scene while the camera sensors are the same. They use this dataset to evaluate the optimal camera setups for VO (SVO is used in their case). The result is very interesting, which shows that it is advantageous to use a large field of view (FoV) camera for indoor scenes and a smaller FoV camera for urban environments. They also presents the details of extending SVO to work with fisheye as well as catadioptric cameras.

B. Stereo camera

Giger et al. [57] establish the most popular Kitti benchmark for the stereo odometry, mainly intended for autonomous driving. The wide angle stereo cameras are mounted on an autonomous driving car. The stereo odometry dataset consists of 22 stereo sequences with a total length of 39.2km. The ground truth is obtained using a RTK GPS/IMU localization unit with accuracy of less than 5cm. The baseline of the stereo camera is about 54cm, which could be reduced if used for a small sized MAV.

Compared to Kitti dataset, the recent European Robotics Challenge (EuRoC) dataset [58] is more focused on MAVs. The dataset is captured using a global shutter stereo camera, which is the Skybotix VI sensor, and the dataset is specially designed for visual navigation of MAVs. The key feature of this dataset is that it provides synchronized stereo images and IMU data collected onboard a MAV. Several image sequences captured in Vicon room and machine hall respectively are included in the dataset. Furthermore, they are categorized as easy, medium and difficult according to the maneuver level of the MAV. This dataset is the most promising one available, which is widely used in the research community of MAVs.

C. RGB-D camera

Sturm et al. [59] provide a benchmark for the evaluation of RGB-D SLAM methods, which is also applicable to the evaluation of RGB-D odometry methods. The sequences are captured from Microsoft Kinect sensor held by hand or mounted on a ground robot. There are 39 sequences in total, covering an office environment and an industrial hall with static and dynamic objects. The frame rate of RGB and depth image is 30fps and the resolution is 640×480. The ground truth trajectory is provided using a motion capture system at 100hz. This is ranked as one of the most popular RGB-D benchmark at present, compared with other benchmarks, such as corbs [60] for Kinect v2, and ICL-NUIM dataset [61]. For the MAV application, this could be used for algorithms testing, but the movement of flying is quite different from that of the dataset. Actually, a comprehensive RGB-D benchmark for MAVs is still vacant in the community.

V. Discussion

This paper is aimed at bridging the gap between the existing surveys and the newly developed state-of-the-art approaches, by giving an overview of the recent development of VO.

Compared with the monocular visual odometry techniques which can only estimate the motion up to a scalar (such scale factor must be determined from other sensors, such as IMU, air pressure, or other direct measurement), the stereo odometry can resolve such ambiguity by estimating metric depth from stereo pairs estimate. Moreover, stereo VO is expected to be more accurate and robust due to the larger field of view and more data being available. But, in cases where the scene distance is much larger than the stereo baseline, stere vision degenerates and becomes ineffective, we then should resort to monocular VO.

At first, most VO approaches rely only on cameras and make little use of IMU. To reduce the drift, they usually apply local (windowed) bundle adjustment on last n frames to realize optimization of pose parameters. With the advent of high-precision IMU sensor, also due to its complementary nature (providing valuable information about short-term motion and rendering global roll, pitch, and scale observable) and abundant presence, a variety of methods have been proposed to intergrate IMU to achieve improved VO performance in terms of both accuracy, efficiency, and robustness.

In early works, visual-inertial fusion has been approached as a pure sensor-fusion problem: vision is treated as an independent, black-box 6-DoF sensor which is fused with inertial measurements in a filtering framework [62], [63], [64]. This so-called loosely coupled approach allows to use existing vision-only methods (such as PTAM, or LSD-SLAM) without any modifications; and the chosen method can easily be substituted for another one. More recent works follow a tightly coupled approach, treating visual-inertial odometry as one integrated estimation problem, where the term of minimizing photometric error for image alignment and the IMU error term are combined into one cost function. The additional IMU error term ensures convergence even for rapid motion [65]. Within such framework, variants have been proposed, the alignment could be performd on pixels [65], [66], lines [67], patches [45] or certain keypoints [15].

In conclusion, VO has gained significant popularity for MAVs in GPS denied environment. Over the years, several projects has been initiated, such as sFly, which focuses on inertial aided VO. Consequently, it is foreseeable that MAVs will make more significant contribution for societal benefits.

Acknowledgment

The authors would like to thank the members of NUS UAV Research Group for their kind support.

References

[1] D. Scaramuzza and F. Fraundorfer, "Visual odometry. part i: The rst 30 years and fundamentals," *IEEE Robot. Autom. Mag*, vol. 18, p. 8092, 2011.

[2] ——, "Visual odometry: Part ii-matching, robustness, and applications," *IEEE Robotics and Automation Magazine*, vol. 19, no. 2, 2012.

[3] C. Forster, M. Pizzoli, and D. Scaramuzza, "Svo: Fast semi-direct monocular visual odometry," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014, pp. 15–22.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[5] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer vision–ECCV 2006*. Springer, 2006, pp. 404–417.

[6] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," *Computer Vision–ECCV 2010*, pp. 778–792, 2010.

[7] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: an efficient alternative to sift or surf," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2564–2571.

[8] C. Mei, S. Benhimane, E. Malis, and P. Rives, "Efficient homography-based tracking and 3-d reconstruction for single-viewpoint sensors," *Robotics, IEEE Transactions on*, vol. 24, no. 6, pp. 1352–1364, 2008.

[9] R. Mur-Artal, J. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *Robotics, IEEE Transactions on*, vol. 31, no. 5, pp. 1147–1163, 2015.

[10] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*. IEEE, 2007, pp. 225–234.

[11] K. Konolige and W. Garage, "Sparse sparse bundle adjustment." Citeseer.

[12] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 834–849.

[13] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, "Visual simultaneous localization and mapping: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 55–81, 2015.

[14] J. Gui, D. Gu, S. Wang, and H. Hu, "A review of visual inertial odometry from filtering and optimisation perspectives," *Advanced Robotics*, vol. 29, no. 20, pp. 1289–1301, 2015.

[15] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 314–334, 2015.

[16] S. Weiss, M. W. Achtelik, S. Lynen, M. Chli, and R. Siegwart, "Real-time onboard visual-inertial state estimation and self-calibration of mavs in unknown environments," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 957–964.

[17] S. Weiss, R. Brockers, and L. Matthies, "4dof drift free navigation using inertial cues and optical flow," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013, pp. 4180–4186.

[18] M. Faessler, F. Fontana, C. Forster, E. Mueggler, M. Pizzoli, and D. Scaramuzza, "Autonomous, vision-based flight and live dense 3d mapping with a quadrotor micro aerial vehicle," *Journal of Field Robotics*, 2015.

[19] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct ekf-based approach," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 298–304.

[20] A. Howard, "Real-time stereo visual odometry for autonomous ground vehicles," in *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*. IEEE, 2008, pp. 3946–3952.

[21] J. Witt and U. Weltin, "Robust stereo visual odometry using iterative closest multiple lines," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013, pp. 4164–4171.

[22] Y. Cheng, M. W. Maimone, and L. Matthies, "Visual odometry on the mars exploration rovers-a tool to ensure accurate driving and science imaging," *Robotics & Automation Magazine, IEEE*, vol. 13, no. 2, pp. 54–62, 2006.

[23] M. Maimone, Y. Cheng, and L. Matthies, "Two years of visual odometry on the mars exploration rovers," *Journal of Field Robotics*, vol. 24, no. 3, pp. 169–186, 2007.

[24] A. E. Johnson, S. B. Goldberg, Y. Cheng, and L. H. Matthies, "Robust and efficient stereo feature tracking for visual odometry," in *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*. IEEE, 2008, pp. 39–46.

[25] A. Milella and R. Siegwart, "Stereo-based ego-motion estimation using pixel tracking and iterative closest point," in *Computer Vision Systems,*

2006 ICVS'06. IEEE International Conference on. IEEE, 2006, pp. 21–21.

[26] M. Achtelik, A. Bachrach, R. He, S. Prentice, and N. Roy, "Stereo vision and laser odometry for autonomous helicopters in gps-denied indoor environments," in *SPIE Defense, Security, and Sensing*. International Society for Optics and Photonics, 2009, pp. 733 219–733 219.

[27] H. Badino, A. Yamamoto, and T. Kanade, "Visual odometry by multi-frame feature integration," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 222–229.

[28] J. Kelly and G. S. Sukhatme, "An experimental study of aerial stereo visual odometry," in *Proc. Symp. Intelligent Autonomous Vehicles, Toulouse, France*, 2007.

[29] K. Schmid, T. Tomic, F. Ruess, H. Hirschmuller, and M. Suppa, "Stereo vision based indoor/outdoor navigation for flying robots," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013, pp. 3955–3962.

[30] N. M. Suaib, M. H. Marhaban, M. I. Saripan, and S. A. Ahmad, "Performance evaluation of feature detection and feature matching for stereo visual odometry using sift and surf," in *Region 10 Symposium, 2014 IEEE*. IEEE, 2014, pp. 200–203.

[31] Y. Ren, X. Xie, J. Hu, and Z. Li, "A stereo visual odometry based on surf feature and three consecutive frames," in *Smart Cities Conference (ISC2), 2015 IEEE First International*. IEEE, 2015, pp. 1–5.

[32] A. Schmidt, M. Kraft, and A. Kasiński, "An evaluation of image feature detectors and descriptors for robot navigation," in *Computer Vision and Graphics*. Springer, 2010, pp. 251–259.

[33] N. Govender, "Evaluation of feature detection algorithms for structure from motion," 2009.

[34] 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, 1981.

[35] D. Nistér, "An efficient solution to the five-point relative pose problem," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 6, pp. 756–770, 2004.

[36] M. Tomono, "Robust 3d slam with a stereo camera based on an edge-point icp algorithm," in *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*. IEEE, 2009, pp. 4306–4311.

[37] B. Kitt, A. Geiger, and H. Lategahn, "Visual odometry based on stereo image sequences with ransac-based outlier rejection scheme." in *Intelligent Vehicles Symposium*, 2010, pp. 486–492.

[38] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3d reconstruction in real-time," in *Intelligent Vehicles Symposium (IV), 2011 IEEE*. IEEE, 2011, pp. 963–968.

[39] A. I. Comport, E. Malis, and P. Rives, "Accurate quadrifocal tracking for robust 3d visual odometry," in *Robotics and Automation, 2007 IEEE International Conference on*. IEEE, 2007, pp. 40–45.

[40] B. Kitt, A. Geiger, and H. Lategahn, "Visual odometry based on stereo image sequences with ransac-based outlier rejection scheme," in *Intelligent Vehicles Symposium (IV)*, 2010.

[41] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3d reconstruction in real-time," in *Intelligent Vehicles Symposium (IV)*, 2011.

[42] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *Asian Conference on Computer Vision (ACCV)*, 2010.

[43] T. Mouats, N. Aouf, L. Chermak, and M. A. Richardson, "Thermal stereo odometry for uavs," *Sensors Journal, IEEE*, vol. 15, no. 11, pp. 6335–6347, 2015.

[44] S. Shen, Y. Mulgaonkar, N. Michael, and V. Kumar, "Vision-based state estimation and trajectory control towards high-speed flight with a quadrotor." in *Robotics: Science and Systems*, vol. 1. Citeseer, 2013.

[45] L. Heng and B. Choi, "Semi-direct visual odometry for a fisheye-stereo camera," in *submitted to Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. IEEE, 2016.

[46] A. S. Huang, A. Bachrach, P. Henry, M. Krainin, D. Maturana, D. Fox, and N. Roy, "Visual odometry and mapping for autonomous flight using an rgb-d camera," in *International Symposium on Robotics Research (ISRR)*, 2011, pp. 1–16.

[47] C. Kerl, J. Sturm, and D. Cremers, "Robust odometry estimation for rgb-d cameras," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3748–3754.

[48] S. Alexandrov and R. Herpers, "Evaluation of recent approaches to visual odometry from rgb-d images," in *RoboCup 2013: Robot World Cup XVII*. Springer, 2013, pp. 444–455.

[49] J. Zhang, M. Kaess, and S. Singh, "A real-time method for depth enhanced visual odometry," *Autonomous Robots*, pp. 1–13, 2015.

[50] J. Zhang and S. Singh, "Visual-lidar odometry and mapping: Low-drift, robust, and fast."

[51] ——, "Visual–inertial combined odometry system for aerial vehicles," *Journal of Field Robotics*, vol. 32, no. 8, pp. 1043–1055, 2015.

[52] J. Zhang, M. Kaess, and S. Singh, "Real-time depth enhanced monocular odometry," in *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*. IEEE, 2014, pp. 4973–4980.

[53] Z. Fang and S. Scherer, "Experimental study of odometry estimation methods using rgb-d cameras," in *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*. IEEE, 2014, pp. 680–687.

[54] ——, "Real-time onboard 6dof localization of an indoor mav in degraded visual environments using a rgb-d camera," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5253–5259.

[55] Z. Fang, S. Yang, S. Jain, G. Dubey, S. Maeta, S. Roth, S. Scherer, Y. Zhang, and S. Nuske, "Robust autonomous flight in constrained and visually degraded environments."

[56] Z. Zhang, H. Rebecq, C. Forster, and D. Scaramuzza, "Benefit of large field-of-view cameras for visual odometry," in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE.

[57] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[58] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The euroc micro aerial vehicle datasets," 2016.

[59] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*. IEEE, 2012, pp. 573–580.

[60] O. Wasenm, M. Meyer, and D. Stricker, "CoRBS: Comprehensive rgb-d benchmark for slam using kinect v2," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, vol. . IEEE, March 2016, p. . [Online]. Available: http://corbs.dfki.uni-kl.de/

[61] A. Handa, T. Whelan, J. McDonald, and A. Davison, "A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM," in *IEEE Intl. Conf. on Robotics and Automation, ICRA*, Hong Kong, China, May 2014.

[62] L. Meier, P. Tanskanen, F. Fraundorfer, and M. Pollefeys, "Pixhawk: A system for autonomous flight using onboard computer vision," in *Robotics and automation (ICRA), 2011 IEEE international conference on*. IEEE, 2011, pp. 2992–2997.

[63] M. Li and A. I. Mourikis, "Optimization-based estimator design for vision-aided inertial navigation," in *Robotics: Science and Systems*, 2013, pp. 241–248.

[64] J. Kelly, S. Saripalli, and G. Sukhatme, "Combined visual and inertial navigation for an unmanned aerial vehicle," in *Field and Service Robotics*. Springer, 2008, pp. 255–264.

[65] V. Usenko, J. Engel, J. Stückler, and D. Cremers, "Direct visual-inertial odometry with stereo cameras," in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 2016.

[66] J. Engel, J. Stuckler, and D. Cremers, "Large-scale direct slam with stereo cameras," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 1935–1942.

[67] T. Holzmann, F. Fraundorfer, and H. Bischof, "Direct stereo visual odometry based on lines," in *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016, pp. 474–485.