14th International Conference on Intelligent Autonomous Systems (IAS), Shanghai, China, July 2016.

Combining Feature-based and Direct Methods for Semi-dense Real-time Stereo Visual Odometry

Nicola Krombach, David Droeschel, and Sven Behnke

Autonomous Intelligent Systems Group, University of Bonn, Germany
krombach@ais.uni-bonn.de

Abstract. Visual motion estimation is challenging, due to high data rates, fast camera motions, featureless or repetitive environments, uneven lighting, and many other issues. In this work, we propose a two-layer approach for visual odometry with stereo cameras, which runs in real-time and combines feature-based matching with semi-dense direct image alignment. Our method initializes semi-dense depth estimation, which is computationally expensive, from motion that is tracked by a fast but robust feature point-based method. By that, we are not only able to efficiently estimate the pose of the camera with a high frame rate, but also to reconstruct the 3D structure of the environment at image gradients, which is useful, e.g., for mapping and obstacle avoidance. Experiments on datasets captured by a micro aerial vehicle (MAV) show that our approach is faster than state-of-the-art methods without losing accuracy. Moreover, our combined approach achieves promising results on the KITTI dataset, which is very challenging for direct methods, because of the low frame rate in conjunction with fast motion.

1 Introduction

For the autonomous navigation of mobile robots, a robust and fast state estimation is of great importance. Many mobile robots contain cameras since they are inexpensive and lightweight and can be used for a variety of tasks, including visual obstacle detection, 3D scene reconstruction, visual odometry, and even visual simultaneous localization and mapping (SLAM).

Visual odometry (VO) describes estimating the egomotion solely from images, captured by a monocular or stereo camera system. A variety of VO methods exists that can be classified into feature-based and direct methods. Most VO methods are feature-based and work by detecting feature points and matching them between subsequent frames. In contrast, direct VO methods estimate the camera motion by minimizing the photometric error over all pixels. As the minimization over all pixel is computationally more demanding than determining the reprojection error of sparse feature points, direct methods are often slower than feature-based methods. In this work, we propose a novel approach that combines direct image alignment with sparse feature matching for stereo cameras. By combining both approaches, we are able to process images with high frame rate and

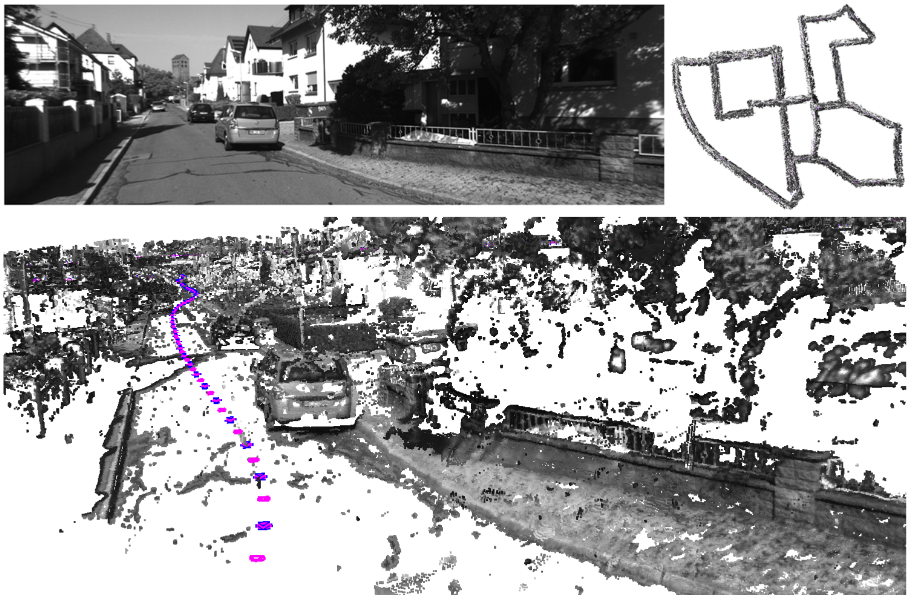

Fig. 1: Semi-dense 3D reconstruction of KITTI 00: Top left: Camera image. Bottom: Semi-dense 3D reconstruction with colored camera trajectory (key frames blue, feature-based tracked frames pink). Top right: Bird's eye view of the complete reconstructed scene.

to fill gaps caused by large motions. Due to the distinctiveness of the tracked features, our method performs well on datasets with low frame rates, which is often a problem for direct methods since they need sufficient image overlap. Our work is based on LSD-SLAM [5], which is a fully direct method for monocular SLAM that is real-time capable on strong CPUs due to its semi-dense approach.

We extend LSD-SLAM to stereo and restrict semi-dense tracking to key frames to achieve a higher frame rate. To estimate the motion between key frames, we employ a feature-based VO method and use the estimated motion as prior for the direct image alignment. Thus, we restrict the search space for direct image alignment and gain real-time performance even on CPUs for mobile applications.

2 Related Work

VO methods estimate egomotion using only images of a single or of multiple cameras. They can be classified into feature-based, direct and hybrid methods.

2.1 Feature-based Methods

The general pipeline for feature-based VO methods can be summarized as follows: Features are detected and either matched or tracked over time. Based on these feature correspondences, the relative motion between two frames is computed. To compensate for drift, many methods make use of pose-graph optimization. Popular feature-based methods are MonoSLAM [4] and Parallel Tracking and Mapping (PTAM) [12]. PTAM is a widely used feature-based monocular SLAM method, which allows robust state estimation in real-time and has been successfully used on MAVs with monocular cameras [23]. Recently, ORB-SLAM [14] has been proposed as a monocular visual SLAM method that tracks ORB features in real-time. When using monocular methods, additional sensors are needed to estimate the absolute scale of a scene. In contrast, stereo or multi-camera VO methods [11, 18, 17] do not suffer from scale ambiguity. Recent feature-based methods also incorporate readings from an inertial measurement unit (IMU) as high-frequency short-term estimates between frames [1, 13].

In our work, we rely on an efficient feature-based library for stereo visual odometry [11], which provides a good trade-off between accuracy and runtime.

2.2 Direct Methods

In contrast to feature-based methods, which abstract images into a sparse set of feature points, direct methods use the entire image information in order to minimize the photometric error. Therefore, these methods are computationally very intense and thus much slower than feature-based methods. Direct approaches exist for stereo, RGB-D and monocular cameras [5, 6, 3, 20, 21]. They often need to use GPUs to achieve real-time performance [16, 19]. By using only pixels with sufficient gradient, LSD-SLAM [5] reduces the computational demand and real-time monocular semi-dense SLAM becomes possible with a strong CPU. This approach has been extended to stereo cameras recently [6].

2.3 Hybrid Methods

For the 3D environment reconstruction, direct methods have the advantage of estimating a dense map, while feature-based methods can only rely on sparse features that have been tracked. Dense direct methods are computationally demanding and are often executed as a final step for estimating a globally consistent dense map after pose tracking with sparse interest-points succeeded. To speed up global optimization, already tracked sparse feature-points can be used as initialization for dense mapping [15]. A recent semi-direct method uses direct motion estimation for initial feature extraction and then continues by using only these features [8].

In contrast to this, we continually combine feature-based and semi-dense direct tracking over time, taking advantage of the fast tracking from the feature-based method and the accurate alignment of image gradients from direct methods. The feature-based tracking result is immediately fed to the direct tracking at runtime as initial guess.

3 Method

Our method is mainly based on the monocular version of LSD-SLAM that we extended to work with stereo cameras. By using stereo cameras instead of a monocular camera, the absolute scale of the scene becomes observable, eliminating scale ambiguity and the need for additional sensors. To ensure a high frame rate, we restrict the semi-dense direct alignment to key frames only and estimate the motion for all other frames by the feature-based method LIBVISO2 [11]. This motion estimate is used as initial estimate for direct alignment of key frames. The semi-dense environment mapping runs in a parallel thread.

3.1 Notation

We follow the notation of Engel et al. [5]. The monochrome stereo images captured at time i are denoted with $I_i^{l/r} : \varOmega \subset \mathbb{R}^2 \to \mathbb{R}$, with image domain $\varOmega$. Each key frame $KF_i = \{I_i^l, I_i^r, D_i, V_i\}$ consists of the left and right stereo images $I_i^{l/r}$, the semi-dense inverse depth map $D_i : \varOmega_{Di} \to \mathbb{R}^+$, and the corresponding variance map $V_i : \varOmega_{Di} \to \mathbb{R}^+$. The inverse of the depth z of a pixel is denoted as $d = z^{-1}$. Camera motions are represented as twist coordinates $\boldsymbol{\xi} \in \mathfrak{se}(3)$ with corresponding transformation matrix $\boldsymbol{T_\xi} \in SE(3)$. A 3D point $\boldsymbol{p} = (p_x, p_y, p_z)^T$ is projected into image coordinates $\boldsymbol{u} = (u_x, u_y, 1)^T$ by the projection function $\pi(\boldsymbol{p}) := \boldsymbol{K}\left(p_x/p_z, p_y/p_z, 1\right)^T$ with intrinsic camera matrix $\boldsymbol{K}$. Thus, the inverse projection function $\pi^{-1}(\boldsymbol{u}, d)$ maps a pixel with corresponding inverse depth to a 3D point $\boldsymbol{p} = \pi^{-1}(\boldsymbol{u}, d) := \left((d^{-1}\boldsymbol{K}^{-1}\boldsymbol{u})^T, 1\right)^T$.

3.2 LSD-SLAM

The processing pipeline of LSD-SLAM [5] consists of the three main components: Tracking, depth map estimation, and global map optimization. Tracking is based on maximizing photo-consistency and thus minimizing the photometric error between the current frame and the most recent key frame using Gauss-Newton optimization:

$$E(\boldsymbol{\xi}) := I_{KF}(\pi(\boldsymbol{p})) - I(\pi(\boldsymbol{T_\xi}\ \boldsymbol{p})) \ , \tag{1}$$

where $\boldsymbol{p}$ is warped from I_{KF} to I by $\boldsymbol{\xi}$. New frames are tracked towards a key frame and the rigid body motion of the camera $\boldsymbol{\xi} \in \mathfrak{se}(3)$ is estimated. In the depth map estimation, tracked frames are then used to refine the existing depth map of the key frame by many small-baseline stereo comparisons. With each new tracked frame, the depth map of the key frame is refined by either creating new depth hypotheses or improving existing ones. New key frames are created when the distance exceeds a certain threshold and are initialized by propagating depth of the previous key frame towards the new frame. Once a key frame is replaced, it is added to the pose-graph for further refinement and loop closing.

3.3 LIBVISO2

LIBVISO2 [11] is a fast feature-based VO library for monocular and stereo cameras. Similar to other feature-based methods, it consists of feature matching over subsequent frames and egomotion estimation by minimizing the reprojection error. Features are extracted by filtering the images with a corner and blob mask and performing non-maximum and non-minimum suppression on the filtered images. Starting from all feature detections in the current left image, candidates are matched in a circular fashion over the previous left image, the previous right image, the current right image, and back to the current left image. If the first and last feature of such a circle match differ, the match is rejected. Based on all found matches, the egomotion is then estimated by minimizing the reprojection error using Gauss-Newton and outliers are removed using RANSAC.

3.4 Semi-dense Alignment of Stereo Key Frames

We build upon the open source release of monocular LSD-SLAM and extend it with stereo functionality. In contrast to monocular visual odometry, stereo allows to compute absolute depth maps and, thus, does not suffer from scale drift. By extending LSD-SLAM to stereo, we combine the existing depth map computation over time with instant stereo depth from the current image pair. While monocular LSD-SLAM uses a random initialization and has to bootstrap over the first frames, we take advantage of using stereo cameras and initialize our method with absolute depth values. We use ELAS [10] to compute the depth map of the initial key frame. The following key frames are registered with their previous key frame by minimizing the photometric error as well as the depth error. While in the monocular case, absolute depth is not observable, with stereo cameras absolute depth is observable for every incoming stereo image pair. This allows us to minimize the depth error in addition to the photometric error. Hence, for direct tracking with stereo, we extend the minimization of the photometric residual r_p to take the depth residual r_d into account:

$$\begin{aligned} r_p(\boldsymbol{p},\boldsymbol{\xi}) &= \left\| I_{KF_i}(\pi(\boldsymbol{p})) - I_j(\pi(\boldsymbol{T}_{\boldsymbol{\xi}}\ \boldsymbol{p})) \right\| , \\ r_d(\boldsymbol{p},\boldsymbol{\xi}) &= \left\| D_{KF_i}(\pi(\boldsymbol{p})) - D_{stereo_j}(\pi(\boldsymbol{T}_{\boldsymbol{\xi}}\ \boldsymbol{p})) \right\| , \end{aligned} \tag{2}$$

where $\boldsymbol{\xi}$ is the camera motion from the i-th key frame to the new j-th frame and D_{stereo_j} is the initial instant stereo depth map of the j-th frame. The minimization is performed using a weighted least squares formulation and solved with the Gauss-Newton method. The residual is formulated as stacked residual $\boldsymbol{r}$ and is weighted with a 2×2 weight matrix $\boldsymbol{W}$:

$$\boldsymbol{r}(\boldsymbol{\xi}) = \sum_{\boldsymbol{p}\in\Omega_{D\,i}} \begin{pmatrix} r_p(\boldsymbol{p},\boldsymbol{\xi}) \\ r_d(\boldsymbol{p},\boldsymbol{\xi}) \end{pmatrix} ; \qquad \boldsymbol{W}(\boldsymbol{\xi}) = \sum_{\boldsymbol{p}\in\Omega_{D\,i}} \begin{pmatrix} w(r_p(\boldsymbol{p},\boldsymbol{\xi})) & 0 \\ 0 & w(r_d(\boldsymbol{p},\boldsymbol{\xi})) \end{pmatrix} , \tag{3}$$

where both residuals are weighted with the Huber norm denoted as $w(\cdot)$.

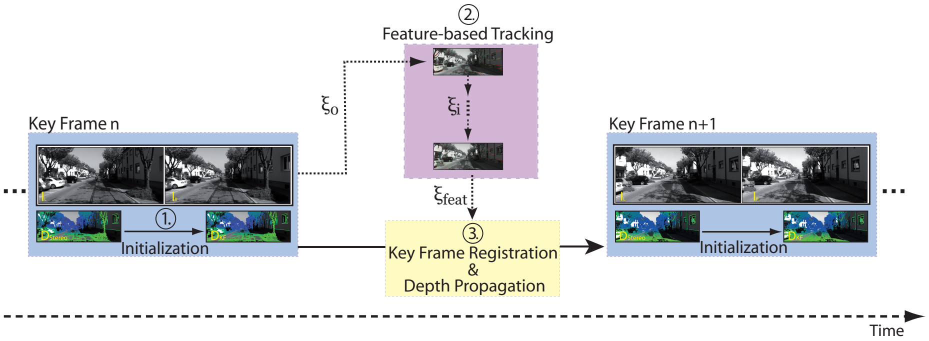

Fig. 2: Overview of our hybrid approach. While direct tracking is only performed on key frames, feature-based tracking is performed for frames in between. The output of the feature-based odometry serves as prior for the direct tracking.

3.5 Hybrid Odometry Estimation

Our idea is to take advantage of the different strengths of both approaches and, thereby, combine fast feature matching with precise semi-dense image alignment for efficient and reliable state estimation. The modular structure of our approach is illustrated in Fig. 2.

We initialize the first key frame with a dense depth map computed by ELAS. Subsequent frames are then tracked towards the key frame incrementally using feature-based LIBVISO2. The relative poses of the tracked frames are concatenated and form the relative pose of the camera to the key frame:

$$\boldsymbol{\xi}_{feat} = \boldsymbol{\xi}_{in} \circ \boldsymbol{\xi}_{in-1} \circ \cdots \circ \boldsymbol{\xi}_{i0} \quad . \tag{4}$$

The current absolute pose of the camera at step j and key frame i can be retrieved by:

$$\boldsymbol{\xi}_{ij} = \boldsymbol{\xi}_{KFi} \circ \boldsymbol{\xi}_{ij-1} \quad . \tag{5}$$

We perform feature-based odometry as long as the motion is sufficiently small. As soon as the motion exceeds the motion threshold ϵ_{motion}, we perform direct registration again and the previous key frame is replaced with the new frame:

$$\epsilon_{motion} = \frac{1}{n} \sum_{k=1}^{n} \sqrt{\left(\boldsymbol{u}_i^k - \boldsymbol{u}_{i-1}^k\right)^2}, \tag{6}$$

where n is the number of matched feature points and $(\boldsymbol{u}_i^k)$ and $(\boldsymbol{u}_{i-1}^k)$ are corresponding feature matches between the current and the previous image. The motion $\boldsymbol{\xi}_{feat}$ serves as initial estimate for the direct registration of the new frame towards the key frame:

$$\boldsymbol{\xi}_{KFi+1} = \boldsymbol{\xi}_{KFi} \circ \boldsymbol{\xi}_{feat} \quad . \tag{7}$$

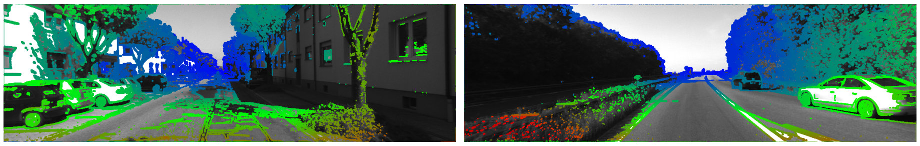

Fig. 3: Computed semi-dense depth maps for KITTI datasets (sequences 00 and 01). Color depicts estimated distance to the sensor.

This allows us to track larger motions faster and more robustly. The depth map of a new key frame is initialized by instant stereo correspondences and then fused with the previous depth map by propagation as described in the next section. Once a new key frame is initialized, we start feature-based matching again.

3.6 Map Update

The depth map of each key frame is updated with instant stereo measurements as well as with propagated depth from the previous key frame. If a new key frame is created, the depth map is computed by instant stereo from the left and right images. For this, we use a simple but fast block matching along epipolar lines. Corresponding pixels are found by minimizing the sum of absolute distances (SAD) error over a 15×15 pixel window. The variance ω for each depth hypothesis is determined as described by Engel et al. [7].

After initializing the depth map with stereo measurements, the depth estimates are refined by propagating depth hypotheses of the old depth map to the new frame:

$$\boldsymbol{p}_{new}(\boldsymbol{p}) = \boldsymbol{R}_{C,KF}\ \boldsymbol{p} + \boldsymbol{t}_{C,KF}, \tag{8}$$

where $\boldsymbol{p}$ is the 3D point in the old key frame. The rotation $\boldsymbol{R}_{C,KF}$ and translation $\boldsymbol{t}_{C,KF}$ describe the coordinate transformation from the key frame coordinate system KF to the candidate coordinate system C. If the residual between the instant and propagated depth is high, the depth value with smaller variance is chosen. Otherwise both estimates—d_{stereo} and d_{prop}—are fused to a new depth estimate d_{new} as a variance-weighted sum:

$$d_{new} = (1 - \omega)\ d_{stereo} + \omega\ d_{prop}\ \ . \tag{9}$$

Fig. 3 shows the resulting semi-dense depth maps for two KITTI sequences.

4 Evaluation

For the evaluation of our hybrid approach, we perform experiments on two challenging stereo datasets: The well-known KITTI-dataset [9] and the EuRoC dataset [2]. The datasets differ in terms of frame rate, apparent motion,

Table 1: ATE Results on KITTI Dataset.

KITTI	**Absolute Trajectory Error RMSE (Median) in m**			
Sequence	Ours	LIBVISO2	ORB-SLAM	S-PTAM
00	**6.15 (5.02)**	29.71 (18.49)	8.30 (6.04)	7.83 (6.30)
01	**61.74 (55.48)**	66.54 (60.46)	335.52 (303.79)	204.65 (157.10)
02	19.47 (15.80)	34.26 (27.36)	**18.66 (15.03)**	20.78 (17.28)
03	**0.67 (0.58)**	1.67 (1.54)	11.91 (9.19)	10.53 (10.41)
04	**0.72 (0.49)**	0.80 (0.66)	2.15 (1.73)	0.98 (0.88)
05	5.78 (4.69)	22.14 (19.07)	4.93 (4.73)	**2.80 (2.24)**
06	4.37 (3.53)	11.54 (10.26)	16.01 (15.56)	**4.00 (4.01)**
07	2.63 (1.77)	4.41 (4.37)	4.30 (3.65)	**1.80 (1.53)**
08	8.75 (7.26)	47.67 (34.84)	38.80 (18.12)	**5.13 (4.26)**
09	**5.55 (4.07)**	89.83 (77.57)	7.46 (6.91)	7.27 (4.61)
10	**1.87 (1.68)**	49.35 (36.00)	8.35 (7.55)	2.08 (1.70)
mean	**10.70 (9.12)**	32.54 (26.42)	41.49 (35.66)	25.74 (20.26)
mean w/o S 01	**5.60 (4.49)**	29.14 (23.02)	12.09 (8.85)	7.85 (6.57)

and stereo baseline. All experiments have been conducted on an Intel Core i7-4702MQ running at 2.2 GHz with 8 GB RAM. The processing is performed on the original image resolution of the rectified images of 1241×376 and 752×480, respectively. We compare the quality of our combined approach in terms of accuracy and runtime to LSD-SLAM [5] and LIBVISO2 [11], as well as to two more state-of-the-art methods: S-PTAM [18] and ORB-SLAM [14]. The results of the referred methods have been obtained using the provided default parameters. As ground truth for all sequences is available, we employ the evaluation metrics by Sturm et al. [22] and measure the absolute trajectory error (ATE) by computing the root mean squared error (RMSE) over the whole trajectory. Additionally, for the monocular systems, the scaling factor is estimated to obtain the absolute scale of the camera trajectory. For an intuitively accessible visualization, trajectories are always shown in bird's eye perspective.

4.1 Accuracy

In terms of accuracy, we achieve similar results as current state-of-the-art stereo methods. As our method is a pure odometry method, it accumulates drift over time, especially at large rotations, where direct alignment of key frames becomes more demanding.

The KITTI benchmark is very challenging for direct methods, as it contains fast motions up to 80 km/h in combination with a low frame rate, which causes inter-frame motions up to 2.8 m per frame.

The results of our evaluation on the KITTI dataset are shown in Table 1. Unfortunately, LSD-SLAM fails on all sequences of the KITTI dataset. This is probably caused by too large inter-frame motion for a pure monocular direct

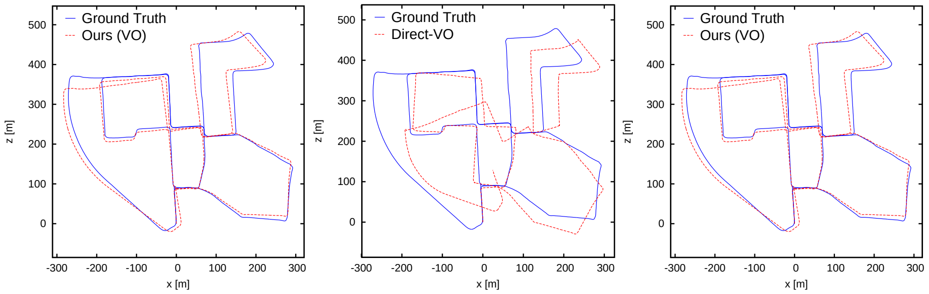

Fig. 4: Comparison of LIBVISO2 (left), direct (middle) and the proposed hybrid odometry (right) on KITTI Sequence 00. Our method accumulates less drift.

method, as sufficient scene overlap is important for successful tracking. Moreover, it can be seen, that all methods lack performance on Sequence 01, resulting in a very high ATE. Sequence 01 contains images from driving on a highway, thus it is hard to find re-occurring feature points in subsequent frames. When averaging over the eleven training sequences, our method ranks first, followed by S-PTAM, ORB-SLAM, and LIBVISO2. However, the bad results from Sequence 01 greatly affect the final average computation, which is why we also show mean values omitting this sequence.

Unfortunately, to our knowledge there is no other publicly available direct method other than LSD-SLAM to compare with. However, as LSD-SLAM fails on the KITTI sequences, we compare our hybrid approach to its fully direct version without feature-based initial estimates. In particular, we compare our hybrid approach to its building blocks—LIBVISO2 and direct stereo tracking—separately. Fig. 4 shows the estimated trajectories of the three methods exemplary on the KITTI 00 sequence. It can be seen that our hybrid approach accumulates less drift than the pure feature-based or direct method. The direct tracking performs worst on this dataset, because it fails at tracking large inter-frame motions and strong rotations without a good initial estimate.

Table 2: ATE Results on EuRoC Dataset.

EuRoC	**Absolute Trajectory Error RMSE (Median) in m**				
Dataset	Ours	LIBVISO2	LSD-SLAM	ORB-SLAM	S-PTAM
V1_01	0.25 (0.18)	0.31 (0.31)	**0.19 (0.10)**	0.79 (0.62)	0.28 (0.19)
V1_02	**0.24 (0.16)**	0.29 (0.27)	0.98 (0.92)	0.98 (0.87)	0.50 (0.35)
V1_03	**0.81 (0.76)**	0.87 (0.64)	X	2.12 (1.38)	1.36 (1.09)
V2_01	**0.22 (0.13)**	0.40 (0.31)	0.45 (0.41)	0.50 (0.42)	2.38 (1.78)
V2_02	**0.31 (0.25)**	1.29 (1.08)	0.51 (0.48)	1.76 (1.39)	4.58 (4.18)
V2_03	**1.13 (0.97)**	1.99 (1.66)	X	X	X
mean	**0.49 (0.41)**	0.85 (0.71)	0.53 (0.48)	1.23 (0.94)	1.82 (1.52)

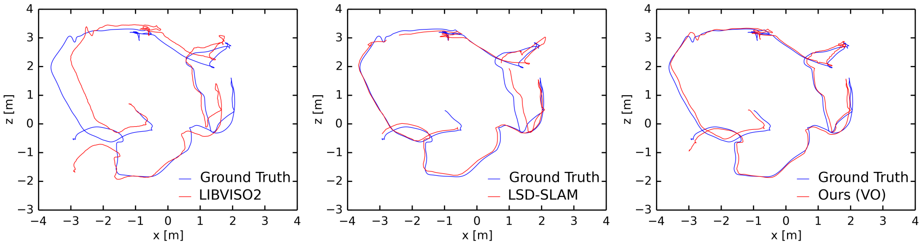

Fig. 5: Comparison of LIBVISO2 (left), Mono LSD-SLAM (middle) and our hybrid odometry (right) on dataset V2_01 with ground truth from a Vicon motion capture system. Our method is much closer to the ground truth.

In addition to the evaluation on the KITTI dataset, we perform further experiments on the recently released EuRoC MAV dataset [2] that contains WVGA stereo images, captured with 20 Hz on an Asctec Firefly hex-rotor. We choose six trajectories with different difficulties from the two Vicon datasets V0 and V1. The data has been collected from flights in a room that is equipped with a Vicon motion capture system, providing 6D ground truth poses. Each dataset contains three trajectories with increasing difficulty: Easy (_01), medium (_02), and difficult (_03). The easy trajectories have good illumination, are feature rich, and show no motion blur, only low optical flow, and low varying scene depth. They capture a static scene. The difficulty increases in the medium trajectories by adding challenging lighting conditions, high optical flow, and medium varying scene depth. However, they still show a static scene and a feature rich environment without motion blur. In contrast, the difficult scene contains areas with only few visual features and more repetitive structures. Moreover, they add motion blur and more challenging lighting conditions. The MAV performs very aggressive flight maneuvers resulting in high optical flow and highly varying scene depth in a non-static scene. The resulting ATE values for this datasets are listed in Table 2. As the difficult datasets V1_03 and V2_03 contain very dynamic movements and fast rotations with an MAV, LSD-SLAM often loses track after a few seconds and is then unable to re-localize for the rest of the trajectory. This is denoted as failure (X). Similarly, S-PTAM and ORB-SLAM lose track for the difficult trajectory V2_03. This dataset shows very challenging conditions with strong motion blur and fast aggressive maneuvers. Moreover, the absence of sufficient visual features makes it hard for the feature-based methods to succeed.

Table 2 shows that our approach outperforms the other methods in terms of accuracy and robustness, and reliably recovers the motion for all test sequences. Additionally, it can be seen, that the results of LIBVISO2 are improved on every trajectory. On average, our hybrid odometry achieves a higher accuracy, with 0.49 m ATE, than LSD-SLAM (0.53 m), ORB-SLAM (1.23 m), and S-PTAM (1.82 m). LSD-SLAM, ORB-SLAM and S-PTAM often suffer from fast motions in combination with rotations, and temporarily lose track. In addition,

Table 3: Average runtimes of all evaluated methods.

Dataset	Method	Tracking	Mapping	Total (VO)
KITTI	Ours	26.5 ms	36.6 ms	63.1 ms
	LSD-SLAM	-	-	-
	ORB-SLAM	30.7 ms	254.0 ms	284.6 ms
	S-PTAM	71.1 ms	5.7 ms	77.4 ms
	LIBVISO2	33.8 ms	-	33.8 ms
EuRoC	Ours	22.6 ms	39.6 ms	62.2 ms
	LSD-SLAM	27.6 ms	85.6 ms	113.2 ms
	ORB-SLAM	17.9 ms	159.2 ms	177.1 ms
	S-PTAM	47.3 ms	1.5 ms	48.8 ms
	LIBVISO2	24.8 ms	-	24.8 ms

we compare results of our hybrid odometry to fully feature-based and direct methods, shown exemplified on dataset V2_01 in Fig. 5. While the results from LIBVISO2 and LSD-SLAM show ATEs of approximately 40 cm, our method yields an error of only 22 cm.

In summary: We achieve similar or better accuracy on different challenging datasets as current state-of-the-art stereo methods. Moreover, our combined approach performs better than both—feature-based and direct—odometries on their own.

4.2 Runtime

We evaluated the average runtime of our method over all datasets on full resolution. As the tracking of new frames and the map building run in distinct threads, timings are given for each part separately in Table 3. On the KITTI dataset, our average runtime for tracking new frames lies below 30 ms. Hence, our approach is significantly faster than recent direct stereo methods [6] and reaches real-time performance. With higher frame rate—as in the EuRoC dataset—we achieve even better results because we need to do direct tracking less often.

4.3 Qualitative Analysis

A major advantage of using a direct approach for tracking is an accurate semi-dense 3D point cloud, which contains every pixel with sufficient gradient. Thus, our odometry not only estimates the current pose of the camera, but also builds a 3D map of the environment, which can be used for additional tasks, like obstacle avoidance. Fig. 6 shows an example of recovered scene depth at near distance and Fig. 1 displays the reconstructed scene of a longer odometry segment from the KITTI dataset 00.

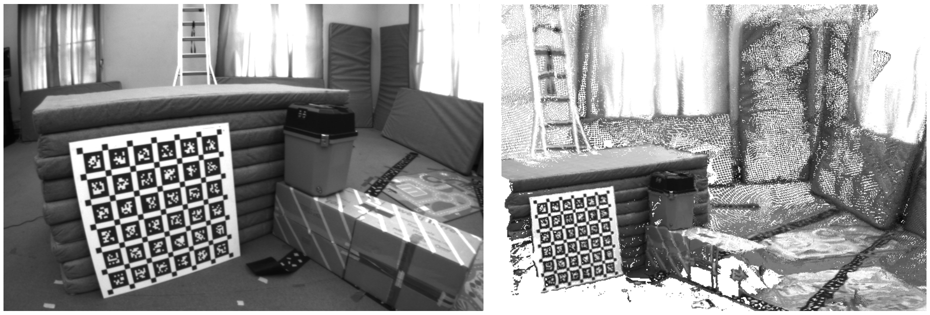

Fig. 6: Left: Camera image. Right: Reconstructed semi-dense 3D point cloud.

5 Conclusions

In this paper, we proposed a novel hybrid visual odometry method that combines feature-based tracking with semi-dense direct image alignment. Our method fuses depth estimates from motion between key frames with instantaneous stereo depth estimates. The performance of our method has been evaluated in terms of accuracy and runtime on two challenging datasets. Our experiments show that for tracking egomotion between image frames, we achieve accuracies similar to the state-of-the-art at high frame rate without the necessity to reduce the image resolution. Due to the feature-based tracking as prior for semi-dense direct alignment, our method is computationally less expensive than direct methods, but still takes advantage of all image points with sufficient gradient for precise keyframe registration. The distinctiveness of the tracked features makes our method also more robust against large inter-frame motion than direct methods.

In future work, we plan to incorporate high frequency IMU readings and to evaluate other feature-based tracking priors, e.g. ORB features. Moreover, since our method accumulates drift over time, we plan to extend our method by a SLAM backend to enhance accuracy and robustness.

Acknowledgement

This work has been supported by the German Federal Ministry for Economic Affairs and Energy (BMWi) in the Autonomics for Industry 4.0 project InventAIRy.

References

1. Achtelik, M., Achtelik, M., Weiss, S., Siegwart, R.: Onboard IMU and monocular vision based control for MAVs in unknown in- and outdoor environments. In: Int. Conf. on Robotics and Automation (ICRA) (2011)
2. Burri, M., Nikolic, J., Gohl, P., Schneider, T., Rehder, J., Omari, S., Achtelik, M.W., Siegwart, R.: The EuRoC micro aerial vehicle datasets. The Int. Journal of Robotics Research (2016)

3. Comport, A., Malis, E., Rives, P.: Accurate quadrifocal tracking for robust 3D visual odometry. In: Int. Conf. on Robotics and Automation (ICRA) (2007)
4. Davison, A., Reid, I., Molton, N., Stasse, O.: Monoslam: Real-time single camera SLAM. Pattern Analysis and Machine Intelligence 29(6), 1052–1067 (2007)
5. Engel, J., Schöps, T., Cremers, D.: LSD-SLAM: Large-scale direct monocular SLAM. In: European Conf. on Computer Vision (ECCV) (2014)
6. Engel, J., Stückler, J., Cremers, D.: Large-scale direct SLAM with stereo cameras. In: Int. Conf. on Intelligent Robots and Systems (IROS) (2015)
7. Engel, J., Sturm, J., Cremers, D.: Semi-dense visual odometry for a monocular camera. In: Int. Conf. on Computer Vision (ICCV) (2013)
8. Forster, C., Pizzoli, M., Scaramuzza, D.: SVO: Fast semi-direct monocular visual odometry. In: Int. Conf. on Robotics and Automation (ICRA) (2014)
9. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. In: Conf. on Computer Vision and Pattern Recognition (CVPR) (2012)
10. Geiger, A., Roser, M., Urtasun, R.: Efficient large-scale stereo matching. In: Asian Conf. on Computer Vision (ACCV) (2010)
11. Geiger, A., Ziegler, J., Stiller, C.: Stereoscan: Dense 3D reconstruction in real-time. In: Intelligent Vehicles Symposium (IV) (2011)
12. Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: Int. Symposium on Mixed and Augmented Reality (ISMAR) (2007)
13. Leutenegger, S., Lynen, S., Bosse, M., Siegwart, R., Furgale, P.: Keyframe-based visual-inertial odometry using nonlinear optimization. Int. Journal of Robotics Research (2014)
14. Mur-Artal, R., Montiel, J., Tardós, J.D.: ORB-SLAM: A versatile and accurate monocular SLAM system. Trans. on Robotics 31(5), 1147–1163 (2015)
15. Mur-Artal, R., Tardós, J.D.: Probabilistic semi-dense mapping from highly accurate feature-based monocular SLAM. In: Robotics: Science and Systems (2015)
16. Newcombe, R.A., Davison, A.: Live dense reconstruction with a single moving camera. In: Conf. on Computer Vision and Pattern Recognition (CVPR) (2010)
17. Nieuwenhuisen, M., Droeschel, D., Schneider, J., Holz, D., Läbe, T., Behnke, S.: Multimodal obstacle detection and collision avoidance for micro aerial vehicles. In: European Conf. on Mobile Robots (ECMR) (2013)
18. Pire, T., Fischer, T., Civera, J., Cristóforis, P.D., Berlles, J.J.: Stereo Parallel Tracking and Mapping for robot localization. In: Int. Conf. on Intelligent Robots and Systems (IROS) (2015)
19. Pizzoli, M., Forster, C., Scaramuzza, D.: REMODE: Probabilistic, monocular dense reconstruction in real time. In: Int. Conf. on Robotics and Automation (ICRA) (2014)
20. Stückler, J., Behnke, S.: Multi-resolution surfel maps for efficient dense 3d modeling and tracking. Journal of Visual Communication and Image Representation 25(1), 137–147 (2014)
21. Stückler, J., Gutt, A., Behnke, S.: Combining the strengths of sparse interest point and dense image registration for rgb-d odometry. In: Int. Symposium on Robotics (ISR) and 8th German Conf. on Robotics (ROBOTIK) (2014)
22. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of RGB-D SLAM systems. In: Int. Conf. on Intelligent Robots and Systems (IROS) (2012)
23. Weiss, S., Scaramuzza, D., Siegwart, R.: Monocular-SLAM–based navigation for autonomous micro helicopters in GPS-denied environments. Journal of Field Robotics 28(6), 854–874 (2011)