ENHANCEMENT STRATEGIES FOR FRAME-TO-FRAME UAS STEREO VISUAL ODOMETRY

J. Kersten *, V. Rodehorst

Faculty of Media, Bauhaus-Universität Weimar, Germany -
(jens.kersten, volker.rodehorst)@uni-weimar.de

Commission III, WG III/3

KEY WORDS: Unmanned Aircraft Systems, Stereo Visual Odometry, Pose Estimation, Drift Reduction

ABSTRACT:

Autonomous navigation of indoor unmanned aircraft systems (UAS) requires accurate pose estimations usually obtained from indirect measurements. Navigation based on inertial measurement units (IMU) is known to be affected by high drift rates. The incorporation of cameras provides complementary information due to the different underlying measurement principle. The scale ambiguity problem for monocular cameras is avoided when a light-weight stereo camera setup is used. However, also frame-to-frame stereo visual odometry (VO) approaches are known to accumulate pose estimation errors over time. Several valuable real-time capable techniques for outlier detection and drift reduction in frame-to-frame VO, for example robust relative orientation estimation using random sample consensus (RANSAC) and bundle adjustment, are available. This study addresses the problem of choosing appropriate VO components. We propose a frame-to-frame stereo VO method based on carefully selected components and parameters. This method is evaluated regarding the impact and value of different outlier detection and drift-reduction strategies, for example keyframe selection and sparse bundle adjustment (SBA), using reference benchmark data as well as own real stereo data. The experimental results demonstrate that our VO method is able to estimate quite accurate trajectories. Feature bucketing and keyframe selection are simple but effective strategies which further improve the VO results. Furthermore, introducing the stereo baseline constraint in pose graph optimization (PGO) leads to significant improvements.

1. INTRODUCTION

In recent years the research for autonomous UAS has focused on indoor navigation without benefit of global navigation satellite systems (GNSS). On-board stabilization and autonomous navigation of multi-rotor UAS in GNSS-denied environments require the handling of fast flight dynamics solely based on indirect measurements. Six degrees of freedom (DOF) UAS ego-motion estimation is usually tackled by the fusion of IMU-outputs with further measurements, e.g. from compass, barometer and ultrasonic, using extended Kalman (Haykin, 2001) or particle filtering (Ristic et al., 2004). Nevertheless, pose estimation based on these sensors is known to be affected by high drift rates. Estimating the ego-motion from the input of a single or multiple cameras is denoted as VO (Nistér et al., 2004). Especially in case of slow motion, IMU measurements tend to drift whereas VO is able to produce better pose estimation results (Carrillo et al., 2012) due to complementary measurement principles. Hence, several systems which incorporate cameras were proposed in the past (Achtelik et al., 2009; Huh et al., 2013; Schmid et al., 2013; Brockers et al., 2014). Besides sensor fusion, SBA (Lourakis, 2010), keyframe selection (Klein and Murray, 2007), tracking of natural landmarks (Tardif et al., 2008) and loop closing techniques like simultaneous localization and mapping (SLAM) (Bailey and Durrant-Whyte, 2006; Durrant-Whyte and Bailey, 2006) are common approaches to reduce VO estimation uncertainties and drift rates.

Motivated by the challenge of choosing the best method or combination of these methods, we examine and directly compare the impact and value of different approaches for VO drift reduction and enhancement. First, we propose a straight-forward frame-to-frame stereo VO chain based on carefully chosen components. Second, we modify and further extend this chain by incorporating well known methods, like keyframe selection and SBA, and measure the impact of each modification. The following strategies as well as their combinations are evaluated: feature bucketing, outlier filtering, keyframe selection, SBA and pose-graph optimization (PGO). Experiments are conducted using the New Tsukuba (Martull et al., 2012) and KITTI (Geiger et al., 2012) benchmark datasets as well as own stereo data.

In the remainder of this paper, VO optimization strategies are reviewed in section 2. A description of our proposed stereo VO method is presented in section 3 and the experimental results are described in section 4, followed by a discussion and concluding remarks in sections 5 and 6.

2. RELATED WORK

VO generally involves the following main steps: (1) image acquisition and optional pre-processing, (2) feature detection, (3) feature matching (or tracking), (4) motion estimation and (5) optional optimization. The most obvious drawback of standard stereo frame-to-frame VO (steps 1-4) is the lack of control mechanisms using available information obtained during processing. For example independently estimated paths of the left and right stereo camera are likely to be different and could further be constrained and enhanced by incorporating the stereo baseline constraint. In this section, relevant aspects and state-of-the-art approaches for the enhancement of frame-to-frame VO are reviewed.

* Corresponding author

2.1 Distribution of Feature Points

In order to avoid unfavorable spatial resection configurations, different strategies with the goal of a uniform-like image point distribution are available. In (Nistér et al., 2006) a non-maxima suppression of Harris points is applied for each 5×5 pixel neighborhood. The number of detected features is limited based on local density instead of a typically used global corner response threshold. Furthermore, features are detected in 10×10 buckets of the image. The authors of (Achtelik et al., 2009) use a feature pruning technique in order to reduce computational costs and enhance the feature tracking. The feature set is reduced by computing the distance between all possible feature pairs, eliminating the feature with the smaller score (Harris corner-response) if the distance is less than a specified threshold. In (Mei et al., 20011) it was demonstrated that especially in case of images containing vegetation their quadtree-based feature distribution approach outperforms the globally thresholded Harris corner response. (Chen and Chiang, 2015) is a further example for subdividing the images into non-overlapping rectangles in which features are detected independently.

2.2 Number of Feature Points

The number of feature points has a major impact on the quality of VO results (Strasdat et al., 2010) and more (reliable) features provide more stable motion-estimation results. In (Nistér et al., 2006) up to 5000 points per image are used. According to a rule of thumb stated in (Fraundorfer and Scaramuzza, 2012), 1000 features is a good number for an image with 640×480 pixel. In (Achtelik et al., 2009) the initial set is reduced to around 150 points. The authors of (Nannen and Oliver, 2012) point out that it is impractical to individually find the best number of points according actual constraints of computational speed and accuracy, average point quality, or average amount of overlap between the images. A setting of 100 points per image quickly emerged as the practical optimum in this study.

2.3 Keyframe Selection

Especially with high image acquisition frame rates or slow motion in relation to the distance to the observed scene, triangulated 3D points from multiple frames tend to have large uncertainties. Based on the standard deviation of triangulated 3D points, significantly wrong pose estimation results together with the corresponding images can be identified and discarded. Another motivation for selecting keyframes is to save computational loads for optimization by reducing the data to a representative sample set of all information. As pointed out by (Scaramuzza and Fraundorfer, 2011), keyframe selection is a very important step in VO and should always be done before updating the motion state. In (Warren, 2015) a frame-striding technique is proposed to actively reduce the number of processed frames based on knowledge related to the behavior of a fixed-wing UAS. Since multi-rotor platforms enable much more complex manoeuvres, this technique is not suitable here. Instead of avoiding degenerated camera motion configurations, as for example in (Pollefeys et al., 2002), or in (Thormählen et al., 2004) a criterion for selecting the keyframe pairing based on the expected estimation error of initial camera motion and object structure is proposed. In (Bellavia et al., 2015) keyframes are selected according to the observation that 3D points related to low temporal flow disparity matches have a higher uncertainty when compared to 3D points with larger temporal disparities.

2.4 Landmark Tracking

According to (Nistér et al., 2006) VO drift cannot be avoided without using landmarks. SLAM-approaches try to estimate a global, consistent robot path including loop-closing or landmark matching. Since triangulated 3D-point-based approaches often suffer from uncertainties in depth, the authors of (Olsen et al., 2003) among others use natural landmarks and reject those who move the most.

Loop closing requires a suitable landmark description as well as an efficient data organization. A global landmark database using histogram of oriented gradients (HOG) descriptors and a hierarchical k-means clustering-based vocabulary tree is utilized in (Zhu et al, 2007). In contrast to traditional structure-from-motion techniques, where features between all frames exhaustively are attempted to match, (Warren, 2015) utilizes the vocabulary-tree-based openFABMAP (Glover et al., 2012) library for loop closing. Recently (Lynen et al., 2015) demonstrated that large-scale, real-time pose estimation and tracking can be performed on mobile platforms by employing map and descriptor compression schemes together with efficient search algorithms.

2.5 Multiple Frame Feature Tracking

Instead of tracking known 3D landmarks, several approaches use information related to the tracked 2D image points in order to reject unstable features. In (Nistér et al., 2006), a sub-track-wise refinement of feature tracks including firewall-based propagation error avoidance was proposed. In (Badino et al., 2013) a strategy of incorporating the whole history of tracked 2D points to obtain a single, improved estimate for the 2D points in the current frame is proposed. The method is able to efficiently reduce ego-motion drift while maintaining high inter-frame motion accuracy. In (Han and Choi, 2014) this approach was further improved for real-time applications using high speed tracking and least squares optimization.

2.6 Utilizing the Baseline Constraint

The relative orientation of two stereo views in absolute orientation (AO) techniques is obtained based on the matching of 3D point clouds (for example in (Carrillo et al., 2012)). In contrast, spatial resection using perspective from *n* points (PnP) methods use 3D-2D point correspondences, where the 2D points from only one of the two available stereo frames are typically used. Due to dead-reckoning, the independently estimated path of the left and right camera may differ. A balanced reprojection error score involving both stereo frames and the known relative orientation was used in (Nistér et al., 2006). This generalized method for 3D-2D motion estimation based on 2D image points from extrinsically calibrated stereo cameras (non-concurrent rays) was proposed in (Nistér et al., 2004).

2.7 Pose Refinement Techniques

In time window-based or local SBA the last *n* camera poses or keyframes are refined based on 3D-2D point correspondences. Since the intrinsic camera parameters, poses and the 3D points are optimized using a least squares minimization, this method is not robust against outliers. Another approach is the estimation of relative poses not only based on consecutive frames but also between all possible pairs of the last *n* frames. Choosing the best combination among them, for example based on the mean

reprojection error, actually represents a type of keyframe selection. In (Olson et al., 2006) a posterior iterative but fast global alignment of the pose graph is proposed. Compared to other approaches, like for example extended Kalman filtering, better trajectories in less computational time could be estimated. Nevertheless, instead of doing posterior optimization, a window-based PGO during the flight would be more helpful for navigation tasks. A general C++ framework for graph optimization was proposed in (Kümmerle et al., 2011), since the optimization of graph-based non-linear error functions has shown to be applicable to many problems (for example SBA, PGO and SLAM). A performance comparable to implementations of state-of-the-art approaches for the specific problem was observed.

2.8 Outlier Removal

The accuracy of relative pose estimation as well as optimization depends on the quality of features and their matches. Removing outliers in a stereo frame based on the epipolar constraint is a simple and effective approach. A popular method for additional outlier removal in consecutive, i.e. temporal, stereo frames is the RANSAC scheme in the relative pose estimation step. A problem of RANSAC is that it tends to favor degenerated configurations. In (Frahm and Pollefeys, 2006) a framework that estimates the correct relation even for (quasi-)degenerate data (QDEGSAC) is proposed. The trifocal geometry can also be exploited for feature matching, since it defines a point constraint instead of the epipolar geometry of image pairs that only defines an ambiguous line constraint. The computational costs for three images in (Heinrichs et al., 2008) are not significantly larger than for two images. But the third image helps to identify and eliminate wrong image matches and disambiguates path estimation in critical configurations. In further VO optimization (step 5) only inliers should be used. Triangulated points with a high standard deviation may additionally be excluded.

3. STEREO VO AND ENHANCEMENT STRATEGIES

The final goal of this study is to derive an efficient and accurate VO workflow for real-time navigation tasks (Figure 1).

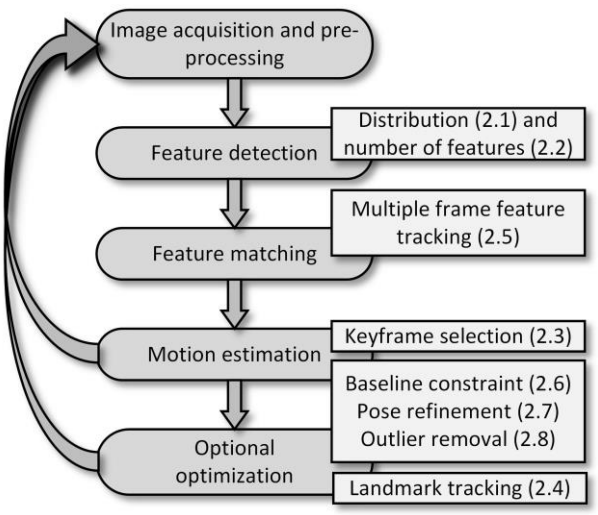

Figure 1. General VO workflow and the relation to enhancement strategies as discussed in section 2

In the following, our proposed frame-to-frame stereo VO chain is presented. Furthermore, in section 3.2 a selection of the most promising enhancement strategies in terms of real-time on-board UAS stereo VO are identified. The VO chain is implemented in C++ using the OpenCV library.

3.1 Frame-to-Frame VO

In monocular VO solely based on 2D information, the relative orientation can only be estimated up to an unknown scale factor. Therefore, we use a calibrated stereo camera system enabling a calibrated relative pose estimation. The stereo scheme is known to exhibit less drift than the monocular one in case of small motions (Scaramuzza and Fraundorfer, 2011).

Several different feature detectors based on corners, regions or blobs offering rotation-, scale- and affine invariants are available. For a choice the expected surrounding conditions have to be considered carefully. Typical light-weight UAS camera frame rates range around 100 Hz inducing fairly small relative viewpoint and scale changes. Furthermore, multi-rotor platforms tend to align horizontally during slow indoor manoeuvres. Amongst many, the Harris or Förstner feature detector (Harris and Stephens, 1988; Förstner, 1994) provide interest points that are relatively stable under small to moderate image distortions (Schmid et al., 2000). Since they offer a high robustness, repeatability and – particularly important for VO – sub-pixel localization accuracy while being moderately computational efficient they are used here.

SIFT (Lowe, 2004) is a very powerful but also computationally expensive feature descriptor, e.g. for structure-from-motion and 3D reconstruction applications, i.e., where large pose changes are expected. In case of small motions the KLT optical flow tracker (Lucas and Kanade, 1981; Tomasi and Shi, 1994) is a common choice to track sparse feature sets in image sequences. The features to track only have to be detected once in the first image circumventing further feature detection steps as well as descriptor computation and matching. Since standard KLT assumes small pixel displacements, we use the pyramidal version (Bouguet, 2000), which also allows larger displacements and is robust to the presence of image blur caused by motion. As stated above, feature points don't have to be detected in each frame individually and the expensive descriptor computation and matching is not necessary. On the other hand efficient adding of new features is required. We found empirically that adding new interest points leads to good results when the absolute number of tracked feature points is less than 75 % of the initially found number of points.

Relative pose changes can in general be estimated using AO or PnP approaches. As pointed out in the early days of VO (Nistér et al., 2004) and (Alismail et al., 2010), PnP approaches tend to be more accurate than AO, since the image reprojection error instead of the 3D feature position error is minimized (Fraundorfer and Scaramuzza, 2012). Using image-based quantities (2D coordinates), the usually introduced uncertainty effects in depth cancel to a large amount (Nistiér, 2006). For better performance, we use the efficient PnP (EPnP) solution proposed by (Lepetit et al., 2009) in conjunction with the well-known RANSAC scheme for outlier detection.

Given a calibrated stereo camera rig, our VO method sequentially processes every incoming stereo image pair in the following manner. The initial absolute pose of the stereo system is defined as the global coordinate origin represented as

$$\mathbf{P}_0 = \begin{bmatrix} \mathbf{R}_0 & \mathbf{T}_0 \\ \mathbf{0} & 1 \end{bmatrix}, \quad (1)$$

where $\mathbf{R}_0$ is a 3×3 identity matrix representing the absolute orientation and $\mathbf{T}_0$ is a three element zero vector representing the stereo camera position (left camera is defined as reference). In each iteration the four images of two consecutive frames (*t* and *t+1*) are transformed to line-preserving normal images based on the known camera matrices, radial distortion coefficients and relative orientation. Interest points and the pyramidal KLT tracker are used for finding and recovering corresponding points in all four images. Stereo normal images enable outlier removal based on thresholding *y*-coordinate differences of matching candidates. Then, metric 3D coordinates of all points in frame *t* are triangulated. Using the 3D-2D correspondences in the left image of frame *t+1*, the EPnP solver estimates the relative pose changes **R** and **T**. The pose update can be obtained using the relative rotation $\Delta\mathbf{R} = \mathbf{R}^T$ and translation $\Delta\mathbf{T} = -\mathbf{R}^T\mathbf{T}$ change:

$$\mathbf{P}_{t+1} = \mathbf{P}_t \cdot \Delta\mathbf{P}_{t+1}, \quad \Delta\mathbf{P}_{t+1} = \begin{bmatrix} \Delta\mathbf{R}_{t+1} & \Delta\mathbf{T}_{t+1} \\ \mathbf{0} & 1 \end{bmatrix} \quad (2)$$

3.2 Selection of Enhancement Strategies

For our experiments the following enhancement strategies of the above described basic workflow were identified. In order to achieve a uniform-like distribution of feature points a simple bucketing approach is used, in which interest points are independently detected in each rectangular region. This may be valuable for example in case of images containing vegetation or moving objects. A rejection of moving points by RANSAC is more likely when we ensure a uniform-like point distribution which covers more rigid parts of the scene. Further outliers may be removed by thresholding the reprojection error obtained by projecting triangulated 3D points of frame *t* into frame *t+1* using the projection matrix estimated by EPnP. This is of special interest when optimization, i.e. SBA, is used.

To reduce uncertainties, for example in case of small motions, we use a simple keyframe selection method similar to (Alcantarilla et al., 2012). The current frame *t+1* is a new keyframe, when the accumulated translation or rotation with respect to the last keyframe exceeds a defined threshold.

Optimization can be achieved using bundle adjustment. In order to keep computational efforts low, usually a time window-based adjustment using the latest *n* frames is applied. We use the fast and efficient simple SBA (SSBA) implementation described in (Zach, 2014) for our experiments.

PGO (Olsen et al., 2006; Kümmerle et al., 2011) is a general concept that can be applied for solving several problems including SBA and SLAM. One drawback of SBA is the sensitivity to outliers due to the least squares minimization. PGO offers an easy way to optimize graphs without any 2D and 3D point observations. On the other hand, trying to optimize a pose chain without any further constraints will not affect the graph at all. One approach to introduce more constraints could be achieved by a window-based optimization using all possible transitions (edges) between the current *n* poses (vertices) instead of just using transitions between consecutive poses. As this requires the estimation of all corresponding relative orientations, we decided to introduce selected keyframes as further constraints. Furthermore, the stereo baseline constraint can be introduced, optimizing the path of the left and right camera simultaneously. PGO is often used for offline posterior optimization. Similar to SSBA, PGO is applied in a time window-based manner in this study.

4. EXPERIMENTAL RESULTS

4.1 Datasets and Evaluation Criteria

Experiments were conducted using the stereo benchmark datasets New Tsukuba with daylight illumination (Martull et al., 2012) and the KITTI stereo sequence 00 (Geiger et al., 2012). Additionally, an own dataset acquired with a manually carried stereo system consisting of two BlueFox-MLC200wC cameras with a resolution of 752×480 pixels, a field of view of 100° and a maximum framerate of 90 Hz was tested. The baseline of the system is around 18.5 cm. Examples of each dataset are shown in Figure 2.

Figure 2. Example images from New Tsukuba (upper left), own stereo data (upper right) and KITTI (bottom)

For the evaluation we used mean values of the error metrics proposed in (Geiger et al., 2012) expressing the errors in rotation and translation as a function of different defined trajectory lengths. A maximum number of 750 additional features per frame was added to the currently existing set in order to ensure an average of around 1000 features. In all configurations features were removed when the *y*-coordinate difference of homologous points in a stereo frame was larger than 0.5 pixels (epipolar constraint).

4.2 New Tsukuba Dataset

The simulated sequence consists of 1800 stereo frames (30 Hz) moving along a complex known trajectory within an indoor office environment. The last 80 frames contain few textures and a large moving object. If not explicitly mentioned, these frames are omitted in the experiments, since the proposed workflow assumes rigid scenes. The following configurations were tested: (a) basic frame-to-frame VO (BVO), (b) feature bucketing (BT) using 3×2, 12×8 and 24×16 non-overlapping rectangles, (c) re-projection-based outlier removal (OR) with a reprojection threshold of 0.5 pixels, (d) OR + BT, (e) keyframe selection (KS), (f) KS + BT, (g) KS + window-based SSBA, (h) KS + window-based SSBA + BT, (i) KS + PGO and (j) KS + PGO + BT. The window size for SSBA and PGO was set to *n=20*. Based on own empirical observations, a new keyframe was

initialized when the accumulated translation and rotation exceeded 150 mm or 5 degrees, respectively. Additionally, a new keyframe was defined when the number of tracked features was lower than 50 % of the points in the last keyframe in order to ensure enough point correspondences. Following the approach in (Badino et al., 2013), the evaluation path lengths were defined as (1, 2, …, 8) meters. The resulting translation and rotation errors are summarized in Table 1. The selected trajectories are shown in Figure 3.

Configuration	New Tsukuba: Mean translation and rotation errors *r* and *t*	
	r [deg/m]	*t* [%]
BVO	0.0207	4.0276
BVO all frames	0.0288	6.9974
BT 3×2	**0.0206**	3.8811
BT 12×8	0.0208	4.4118
BT 24×16	0.0208	4.2010
OR	0.0207	3.9961
OR + BT 3×2	**0.0206**	3.5935
KS	0.0218	3.5452
KS + BT 3×2	0.0224	**3.3382**
KS + SSBA	0.0243	5.2437
KS + SSBA + BT 3×2	0.0242	4.9192
KS + PGO	0.0213	3.7622
KS + PGO + BT 3×2	0.0211	3.8166

Table 1. Mean translation and rotation error rates for Tsukuba: the lowest error rates are highlighted

Our basic VO method yields accurate trajectories of good quality. A mean rotation error of 0.021 deg/m is comparable good to the value of ~0.02 deg/m in (Badino et al., 2013) obtained using their baseline stereo VO method in which also the KLT tracker is utilized. Our translation error is even ~2 % lower. The best result in terms of mean rotation and translation errors could be obtained by applying outlier removal and bucketing (OR + BT 3×2) as well as keyframe selection and bucketing (KS + BT 3×2), respectively.

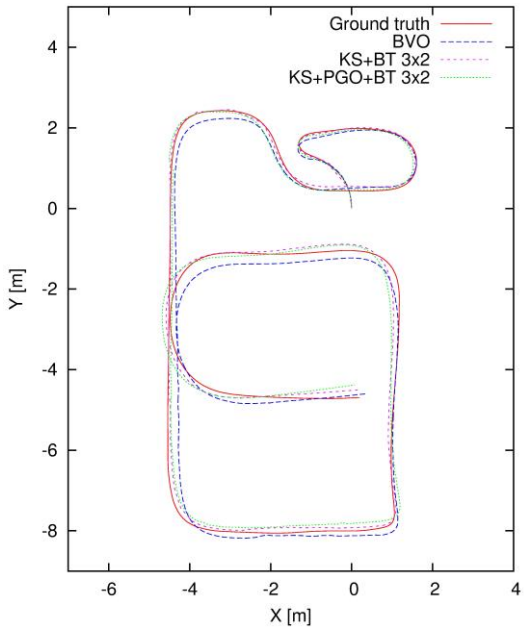

Figure 3. Selected result trajectories for New Tsukuba

4.3 KITTI Dataset

The first KITTI stereo sequence (00) containing 4541 stereo image pairs with a size of 1241×376 pixels was used for the evaluation. The 10 Hz sequence contains real world data acquired from a driving car. Infrastructure but also vegetation and moving objects, like pedestrians and vehicles, are contained. Compared to the New Tsukuba dataset, these moving objects are much smaller and the rigid parts of the scenes are well textured, so that we used the complete sequence. The following configurations were tested: (a) BVO, (b) BT using 6×2, 24×8 and 48×16 non-overlapping rectangles, (c) OR with a threshold of 0.5 pixels, (d) OR + BT (6×2 and 24×8), (e) KS, (f) KS + BT, (g) window-based SSBA, (h) window-based SSBA + BT with *n=20*, (i) PGO and (j) PGO + BT with *n=5*. A new keyframe is initialized, when the accumulated translation and rotation exceed 5 m or 5 degrees, respectively. Since the motion of the camera system is much faster and the frame rate is lower compared to Tsukuba, SSBA was tested without KS here.

Configuration	KITTI: Mean translation and rotation errors *r* and *t*	
	r [deg/m]	*t* [%]
BVO	0.0096	1.9490
BT 6×2	0.0086	1.9035
BT 24×8	0.0088	1.9333
BT 48×16	0.0088	1.9923
OR	0.0099	1.9756
OR + BT 6×2	0.0089	1.9119
OR + BT 24×8	0.0082	1.8693
KS	0.0086	2.4589
KS + BT 6×2	0.0073	2.0124
KS + BT 24×8	0.0088	2.5675
SSBA	0.0124	2.7092
SSBA + BT 6×2	0.0126	2.8027
PGO	0.0083	1.8682
PGO + BT 6×2	**0.0067**	**1.5961**

Table 2. Mean translation and rotation error rates for KITTI sequence (00): the lowest error rates are highlighted

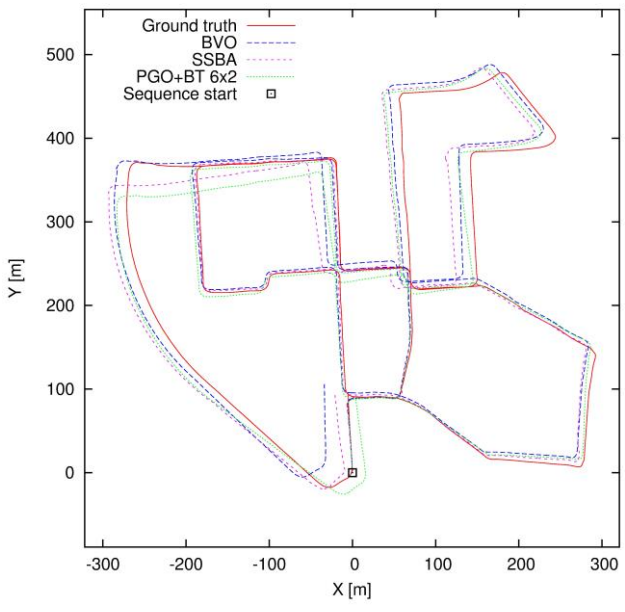

Figure 4. Selected result trajectories for KITTI

Since keyframe selection has shown to decrease the performance in terms of translation, PGO was applied in order to optimize the left and right camera path simultaneously using the stereo baseline constraint (section 3.2). Following the suggestions of the dataset providers, the evaluation path lengths are defined as (100, 200, …, 800) meters.

According to Table 2 and Figure 4, our basic VO method again provides quite accurate results. The achieved mean translation and rotation errors of 1.95 % and 0.0096 deg/m are comparable good with respect to those reported for the KLT-based benchmark method in (Badino et al., 2013) ($t \approx 1.7$ % and $r \approx 0.011$ deg/m). Please note that the results of Badino et al. were obtained using all 11 KITTI sequences. OR and BT (OR + BT 24×8) decreased the translation error to 1.87 %. KS and BT reduced the rotation error to 0.073 deg/m. However, the significantly best result in terms of both errors could be obtained by PGO + BT 6×2.

4.4 Own Dataset

Our outdoor dataset consists of 1243 grayscale stereo image pairs. The trajectory describes a closed loop around the Digital Bauhaus Lab in Weimar.

Configuration	Loop closure errors: coordinate differences (end - start) and absolute distance *S*			
	dX [m]	*dY* [m]	*dZ* [m]	*S* [m]
BVO	-0.791	0.139	2.294	2.444
BT 3×2	0.069	1.922	1.197	2.266
BT 6×4	-0.109	1.674	1.639	2.592
OR	-0.070	0.595	2.517	2.588
OR+BT 3×2	0.225	3.327	1.470	3.644
KS	-1.124	0.502	3.970	4.156
KS+BT 3×2	-0.041	1.429	1.418	**2.014**
PGO	0.070	1.463	1.492	2.091
PGO+BT 3x2	-1.030	-0.005	1.766	2.045

Table 3. Loop closure values in *X*, *Y*, and *Z* for our own dataset: the lowest value *S* is highlighted

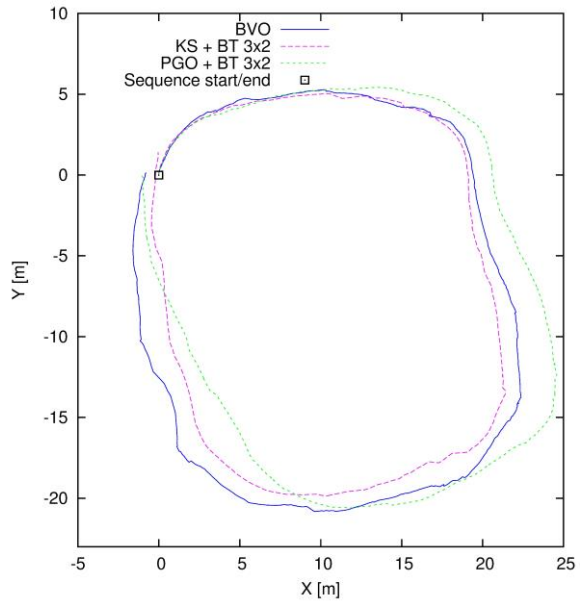

Figure 5. Selected result trajectories for own data

Since no ground truth data is available here, we use the loop closure expressed in coordinate differences of the start and end position as evaluation criterion. The following configurations were tested: (a) BVO, (b) BT using 3×2 and 6×4 buckets, (c) OR with a threshold of 1.5 pixels, (d) OR + BT 3×2, (e) KS, (f) KS + BT, (g) window-based PGO (baseline constraint) with $n=5$. A new keyframe is initialized, when the accumulated translation and rotation exceed 50 cm or 5 degrees, respectively (found empirically). The corresponding loop closure results are listed in Table 3.

The coordinate differences significantly differ and it was not possible to obtain absolute distances *S* below 2 m. Furthermore, the trajectories tend to drift in height, since only positive values for *dZ* were observed. KS+BT 3×2 yields the lowest absolute difference. However, it remains unclear which of the method is the best. The trajectories in Figure 5 exhibit significant differences and therefore also demonstrate that additional criterions should be used for the evaluation of data without ground truth information.

5. DISCUSSION

The experimental results presented in section 4 show that the quality of VO results in this study mainly depend on the components chosen for frame-to-frame VO and partially can be enhanced incorporating additional strategies and optimization steps. A further important factor is the characteristic of each dataset or application.

Processing the New Tsukuba dataset turned out to be challenging due to unconstrained camera manoeuvres, illumination changes, fast rotations combined with low textured image content as well as an opening door at the end of the sequence. After excluding this moving object, our basic VO chain is able to obtain quite accurate results. Processing all frames increased the mean errors *r* and *t* significantly (Table 1). A rough bucketing using 3×2 rectangles turned out to be useful in combination with OR as well as KS. The simple KS method leads to a significant gain in translation (0.48 %) and reduced the set of used poses to around 800. A critical point in KS is the tracking of features. In order to ensure a successful relative pose estimation, we had to increase the number of feature points by a factor of two in order to ensure that enough features are available. Nonetheless, pose change estimation using EPnP in a RANSAC scheme sometimes failed in our tests. The reason for this is that too many features were identified as outliers due to the epipolar constraint. This indicates that KLT provides a good tracking in consecutive frames with small pose changes, but tend to accumulate small location inaccuracies in case of tracking points over multiple frames. Similar to this, also OR pruned too much features in our tests with KS. Optimization of the keyframe poses using SBA downgraded the results – even with filtering of 3D points, which exceed a standard deviation in *X*, *Y* or *Z* of 15 mm. The main reason is that in our current basic implementation the optimized 3D coordinates obtained by SBA were not used in the following estimations. This will be improved in future works. Another reason may be the above mentioned accumulated inaccuracies from KLT. Since the constraints introduced in PGO are solely related to the selected and estimated keyframe poses, the results can obviously only be improved within the range of the quality of KS. For the New Tsukuba dataset, our approach outperforms the method proposed in (Bellavia et al., 2015). Furthermore, a better rotation accuracy than reported in (Badino, et al. 2013) could be achieved. Using an Intel i7 CPU with 8 cores and 32 GB RAM,

minimum and maximum average computational times per frame of 0.16 s (BVO) and 0.32 s (KS+SSBA) were measured.

The KITTI dataset 00 was less challenging compared to New Tsukuba, due to the constrained and forward directed motion on a plane and the well textured images. Nonetheless, decreasing both, translation and rotation error, was only possible using BT or PGO. BT (6×2) reduced *r* and *t* by 0.001 deg/m and 0.05 %, respectively. The reason for this might be twofold. First, in BT stable, locally thresholded features have a better chance to survive in case of natural image content, e.g., vegetation, which is known to yield larger corner responses but might not be static. Second, with a more uniform-like distribution features detected on moving objects have a better chance to be filtered out by RANSAC. On the other hand the experimental results show that choosing a good bucket size is not trivial. Applying KS and BT together, the mean rotation error could be reduced to 0.0073 deg/m, but in turn the translation error increased slightly compared to BVO. However, due to the fast motion of the camera and the relatively low frame rate, KS is not required. The significantly best results could be obtained by introducing the stereo baseline constraint with PGO. Minimum and maximum average computational times per frame were measured in a range of 0.13 s (BVO) and 0.38 s (PGO+BT).

Our dataset was challenging because of the discontinuous trajectory as a result of manual image acquisition while walking. Other challenging properties of the sequence are repeated patterns (e.g. grids), reflective and translucent surfaces as well as large untextured image areas (pavement). However, this represents a realistic scenario and is therefore important to investigate. The obtained results (Table 3) are not satisfactory and the trajectory plots (Figure 5) indicate that a loop closure itself is not enough for the evaluation of data without ground truth information. Hence, for accurate indoor navigation tasks the method has to be further investigated.

6. CONCLUSION AND OUTLOOK

In this paper, a frame-to-frame VO chain is proposed. Furthermore, several different strategies of VO enhancement and optimization are reviewed. Five different strategies (feature bucketing, outlier removal, keyframe selection, window-based bundle adjustment and window-based pose graph optimization) were tested in different combinations. The results demonstrate, that the proposed VO method is able to yield accurate results within short computational times, even using our basic implementation. However, especially the results obtained using real own data demonstrate the complexity and difficulty of VO.

The proposed baseline VO chain provides better results than initially expected. The additionally incorporated and evaluated approaches for VO enhancement further improved these results. The most significant improvements could be achieved using BT, KS and PGO introducing the stereo baseline constraint. We expect comparable good improvements with SBA when optimized 3D coordinates from preceding frames are used. Of course each additionally added component in VO also leads to more free parameters to be defined in advance. This may affect the transferability to other datasets.

The discussion in section 5 indicates several different threads for further progress. Multi-frame feature integration (Badino et al., 2013) could help to improve the location accuracy of tracked features. In order to avoid (quasi-)degenerated configurations in relative pose estimation, QDEGSAC (Frahm and Pollefeys, 2006) should be utilized. The very simple keyframe selection used in this study should be replaced by an adaptive version. Here, the feature temporal flow (Bellavia et al., 2015) is a promising approach. The detection of moving objects may be another task for future works, for example based on the dense scene flow (Alcantarilla et al., 2012). Finally, VO results should directly be coupled with IMU and other measurements, for example using the well-known Kalman filter.

REFERENCES

Achtelik, M., Bachrach, A., He, R., Prentice, S. and Nicholas, R., 2009. Stereo vision and laser odometry for autonomous helicopters in GPS-denied indoor environments. In: Grant, G. R., Gage, D. W. and Shoemaker, C. M. (Eds.), Unmanned Systems Technology, *SPIE-The International Society for Optical Engineering*, Orlando, FL, USA.

Alcantarilla, P. F., Yebes, J. J., Almazán, J. and Bergasa, L. M., 2012. On combining visual SLAM and dense scene flow to increase the robustness of localization and mapping in dynamic environments. *IEEE International Conference on Robotics and Automation (ICRA)*, Saint Paul, MN, pp. 1290-1297.

Alismail, H., Browning, B. and Dias, M.B., 2010. Evaluating pose estimation methods for stereo visual odometry on robots. In: *Proceedings of the International Conference on Intelligent Autonomous Systems (IAS-11)*, Ottawa, Canada.

Bailey, T. and Durrant-Whyte, H., 2006. Simultaneous localization and mapping: part II. *IEEE Robotics & Automation Magazine*, 13(3), pp. 108-117.

Bellavia, F., Fanfari, M. and Colombo, C., 2015. Selective visual odometry for accurate AUV localization. *Autonomous Robots*, pp. 1-11.

Bouguet, J.-Y., 2000. Pyramidal implementation of the Lucas Kanade feature tracker: Description of the algorithm. Technical report, *Intel Corporation Microprocessor Research Labs*.

Brockers, R., Hummenberger, M., Weiss, S. and Matthies, L., 2014. Towards autonomous navigation of miniature UAV. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, pp. 645-651.

Carrillo, L. R. G., López, A. E. D., Lozano, R. and Pégard, C., 2012. Combining stereo vision and inertial navigation system for a quad-rotor UAV. In: *J. Intell. Robotics Syst.*, 65(1-4), pp. 373-387.

Chen, L.-H. and Chiang, K.-W., 2015. The performance analysis of stereo visual odometry assisted low-cost INS/GPS integration system. *Smart Science*, 3(3), pp. 148-156.

Durrant-Whyte, H. and Bailey, T., 2006. Simultaneous localization and mapping: part I. *IEEE Robotics & Automation Magazine*, 13(2), pp. 99-110.

Frahm, J. M. and Pollefeys, M., 2006. RANSAC for (quasi-) degenerate data (QDEGSAC). In: *Proc. CVPR*, pp. 453-460.

Förstner, W., 1994. A framework for low level feature extraction, In: *Proc. ECCV*, Springer, pp. 383-394.

Fraundorfer, F. and Scaramuzza, D., 2012. Visual odometry: part II - Matching, robustness, and applications. *IEEE Robotics and Automation Magazine*, 19(2), pp. 78-90.

Geiger, A., Lenz, P. and Urtasun, R., 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: *Proc. CVPR*, pp. 3354-3361.

Glover, A., Maddern, W., Warren, M., Reid, S., Milford, M. and Wyeth, G., 2012. OpenFABMAP: An open source toolbox for appearance-based loop closure detection. In: *Proc. ICRA*, pp. 4730-4735.

Han. S.-J. and Choi, J., 2014. Real-time precision vehicle localization using numerical maps. *ETRI Journal*, 6(6), pp. 968-978.

Heinrichs, M., Hellwich, O. and Rodehorst, V., 2008. Robust spatio-temporal feature tracking. *IAPRS*, 37(B3a), pp. 51-56.

Harris, C. and Stephens, M., 1988. A combined corner and edge detector. In: *Proc. Alvey Vision Conf*., pp. 147-151.

Haykin, S. S., 2001. Kalman filtering and neural networks. *John Wiley & Sons*, Inc., New York, NY, USA.

Huh, S., Shim, D.H. and Kim, J., 2013. Integrated navigation system using camera and gimbaled laser scanner for indoor and outdoor autonomous flight of UAVs. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3158-3163.

Klein, G. and Murray, D., 2007. Parallel tracking and mapping for small ar workspaces. In: *Proc. Int. Symp. Mixed and Augmented Reality*, pp. 225-234.

Kuemmerle, R., Grisetti, G., Strasdat, H., Konolige, K. and Burgard, W., 2011. g2o: A general framework for graph optimization. In: *Proc. ICRA*, Shanghai, pp. 3607-3613.

Lepetit, V., Moreno-Noguer, F. and Fua. P., 2009. EPnP: An accurate O(n) solution to the PnP problem. *Int. J. Comput. Vision*, 81(2), pp. 155-166.

Lourakis, M. I. A., 2010. Sparse non-linear least squares optimization for geometric vision. In: Daniilidis, K., Maragos, P. and Paragios, N. (Eds.), *Proc. European Conference on Computer Vision: Part II* (ECCV), Springer, Berlin, Heidelberg, pp. 43-56.

Lowe, D., 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2(60), pp. 91-110.

Lucas, B. and Kanade, T., 1981. An iterative image registration technique with an application to stereo vision. In: *Proc. International Joint Conference on Artificial Intelligence*, pp. 674-679.

Lynen, S., Sattler, T., Bosse, M., Hesch, J., Pollefeys, M., Siegwart, R., 2015. Get out of my lab: Large-scale, real-time visual-inertial localization. *Robotics Science and Systems* (RSS).

Martull, S., Martorell, M. P. and Fukui, K., 2012. Realistic CG stereo image dataset with ground truth disparity maps. *ICPR workshop TrakMark2012*, pp. 40-42.

Nannen, V. and Oliver, G., 2012. Optimal number of image keypoints for real time visual odometry. In: *IFAC Worksh. Navig. Guid. Control Underw. Veh. (NGCUV)*, Porto, Portugal, pp. 331-336.

Nistér, D., Naroditsky, O. and Bergen, J., 2004. Visual odometry. In: *Proc. CVPR*, pp. 652-659.

Olson, E., Leonard, J. and Teller, S., 2006. Fast iterative optimization of pose graphs with poor initial estimates. In: *Proc. ICRA*, pp. 2262-2269.

Pollefeys, M., Gool, L.V., Vergauwen, M., Cornelis, K., Verbiest, F. and Tops, J., 2002. Video-to-3d. In: *Proc. IAPRS*, 34, pp. 252-258.

Ristic, B., Arulampalam, S. and Gordon, N. J., 2004. Beyond the Kalman filter: Particle filters for tracking applications. *Artech house*.

Scaramuzza, D. and Fraundorfer, F., 2011. Visual odometry: part I - The first 30 years and fundamentals. *IEEE Robotics and Automation Magazine*, 18(4), pp. 80-92.

Schmid, C., Mohr, R. and Bauckhage, C., 2000. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2), pp. 151-172.

Schmid, K., Tomic, T., Ruess, F., Hirschmüller, H. and Suppa, M., 2013. Stereo vision based indoor/outdoor navigation for flying robots. In: *Proc. IROS*, pp. 3955-3962.

Strasdat, H., Montiel, J. and Davison, A., 2010. Real time monocular SLAM: Why filter? In: *Proc. IRCA*, pp. 2657-2664.

Tardif, J. P., Pavlidis, Y. and Daniilidis, K., 2008. Monocular visual odometry in urban environments using an omnidirectional camera. In: *Proc. IROS*, pp. 2531-2538.

Thormählen, T., Broszio, H. and Weissenfeld, A., 2004. Keyframe selection for camera motion and structure estimation from multiple views. In: *Proc. ECCV*, Springer, pp. 523-535.

Tomasi, C. and Shi, J., 1994. Good features to track. In: *Proc. CVPR*, pp. 593-600.

Tuytelaars, T. and Mikolajczyk, K., 2008. Local invariant feature detectors: a survey. *Found. Trends. Comput. Graph. Vis.*, 3(3), pp. 177-280.

Zach, C., 2014. Robust bundle adjustment revisited. In: *Proc. ECCV*, Springer, pp. 772-787.