Bundle Adjustment with Feature Scale Constraints for Enhanced Estimation Accuracy

Vladimir Ovechkin

Bundle Adjustment with Feature Scale Constraints for Enhanced Estimation Accuracy

Research Thesis

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Technion Autonomous
Systems Program (TASP)

Vladimir Ovechkin

Submitted to the Senate
of the Technion — Israel Institute of Technology
Sh'vat 5778 Haifa February 2018

This research was carried out under the supervision of Assistant Prof. Vadim Indelman from the Faculty of Aerospace Engineering as part of the Technion Autonomous Systems Program at the Technion - Israel Institute of Technology.

Publications:

- V. Ovechkin and V. Indelman. BAFS: Bundle Adjustment with Feature Scale Constraints for Enhanced Estimation Accuracy. IEEE Robotics and Automation Letters (RA-L), 3(2):804-810, 2018.
- V. Ovechkin and V. Indelman. BAFS: Bundle Adjustment with Feature Scale Constraints for Enhanced Estimation Accuracy. In IEEE International Conference on Robotics and Automation (ICRA); submission via IEEE Robotics and Automation Letters (RA-L), May 2018.

ACKNOWLEDGEMENTS

I would first like to thank my thesis advisor Vadim Indelman. He consistently allowed this thesis to be my own work, but steered me in the right direction whenever he thought I needed it.

I would also like to acknowledge my friend Dmitry Kopitkov as the second reader of this thesis, and I am gratefully indebted to his very valuable comments on this thesis.

Finally, I must express my very profound gratitude to my parents and to my wife for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

The generous financial help of the Technion is gratefully acknowledged.

Contents

List of Figures

List of Tables

Abstract

Accurate pose estimation and 3D reconstruction are important in a variety of applications such as autonomous navigation or mapping in uncertain or unknown environments. Bundle adjustment (BA) and simultaneous localization and mapping (SLAM) are commonly used approaches to address these and other related problems. Given a sequence of images, BA is the problem of simultaneously inferring the camera poses and the observed 3D landmarks. BA is typically solved by minimizing the re-projection error between image observations (image features) and their prediction obtained by projecting the corresponding landmarks onto the camera frame. This optimization is typically realized using non-linear least-squares (NLS) approaches. Different techniques exist for detecting image features, including the well-known SIFT features. Yet, state of the art BA and visual SLAM approaches formulate the constraints in the NLS optimization utilizing only image feature coordinates.

In this work we propose to incorporate within BA a new type of constraints that use feature scale information that is readily available from a typical image feature detector (e.g. SIFT, SURF). While feature scales (and orientation) play an important role in image matching, they have not been utilized thus far for estimation purposes in BA framework. Our approach exploits feature scale information and uses it to enhance the accuracy of bundle adjustment, especially along the optical axis of the camera in a monocular setup. Specifically, we formulate constraints between the measured and predicted feature scale, where the latter depends on the distance from the camera and the corresponding 3D point, and optimize the system variables to minimize the residual error in these constraints in addition to minimizing the standard re-projection error. We study our approach both in synthetic environments and real-image ground and aerial datasets (KITTI and Kagaru), demonstrating significant improvement in positioning error.

Abbreviations

BA	:	Bundle Adjustment
BAFS	:	Bundle Adjustment with Feature Scale constraints
DTAM	:	Dense Tracking and Mapping
GPS	:	Global Positioning System
GTSAM	:	Georgia Tech-Smoothing and Mapping
HOG	:	Histogram of Oriented Gradients
IMU	:	Inertial Measurement Unit
iSAM	:	Incremental Smoothing and Mapping
MAP	:	Maximum a Posteriori
MSER	:	Maximally Stable Extremal Regions
PDF	:	Probability Distribution Function
RANSAC	:	Random sample consensus
SfM	:	Structure from Motion
SIFT	:	Scale Invariant Feature Transform
SLAM	:	Simultaneous Localisation and Mapping
SUFR	:	Speeded Up Robust Features
SVO	:	Semi-direct Visual Odometry
VAN	:	Vision Aided Navigation
VO	:	Visual Odometry

Chapter 1

Introduction

1.1 Introduction

Accurate pose estimation and structure reconstruction are important in a variety of applications, including vision aided navigation (VAN) [23], simultaneous localization and mapping (SLAM) [20, 32], visual odometry (VO) [13], augmented reality, structure from motion (SfM), tracking and robotic surgery. Building a spatially consistent model is a key functionality to endow a mobile robot with autonomy. Without an initial map or an absolute localization means, it requires to concurrently solve the localization and mapping problems. For this purpose, vision is a powerful sensor, because it provides data from which stable features can be extracted and matched as the robot moves. But it does not directly provide 3D information, which is a difficulty for estimating the geometry of the environment. Bundle adjustment (BA) is a commonly used approach to address these and other related problems, and as such, has been extensively investigated over the years; see [49] for an extensive review of different aspects in BA.

Standard BA approaches typically assume a pinhole camera model [22] and minimize re-projection errors between measured and predicted image coordinates. This minimization is typically obtained using iterative nonlinear optimization techniques that, provided proper initial guess, converge to the maximum a posteriori (MAP) solution over camera poses and landmarks that represent the observed environment.

Two different approaches address the SLAM problem using vision: one with stereovision, and the other with monocular images. Both approaches rely on a robust interest point matching algorithm that works in very diverse environments. The stereovision based approach is a classic SLAM implementation, whereas the monocular approach is more challenging. A stereo camera setup provides higher accuracy (in proper conditions only) with less computations: as baseline between cameras is known, it allows to reconstruct the environment absolute scale. The monocular case requires alternative approaches for landmark initialization and the solution is available up to scale only. Current work deals with a single camera case.

In lack of sources of absolute information such as GPS or a priori available map, maintaining high-accuracy estimation over time is a challenging task. Other types of prior information

include known motion model, knowledge about the observed environment, such as certain object sizes or awareness of the observed environment structure. All these assumptions might help to create additional constraints for reconstruction and pose estimation and enhance accuracy as a result. However, for a common algorithm it is preferable to solve the SLAM problem without the mentioned assumptions, although their impact on accuracy improvement is significant if this prior information is indeed available.

Current work is dedicated to development of an algorithm independent of prior knowledge about the environment. In lack of observations and constraints that are able to correct the global scale it is necessary to prevent scale drift during an exploration scenario. This is particularly the case for a monocular camera setup due to *scale drift*: without assuming any additional or prior information, camera motion and 3D map can be only estimated up to scale, which drifts over time. SLAM approaches resort to loop closure observations to reset estimation error to prior values. However, this constrains the camera motion to occasionally revisit previously seen areas and to reliable detect these loop closure observations. Moreover, one cannot rely on loop closures in exploration scenarios, where the camera/robot continuously operates in unseen environments. This research proposes an alternative approach to reduce scale drift in a monocular camera setting, without relying on loop closure observations.

1.2 Related Work

Visual SLAM

Early definitions of environment mapping were introduced by Smith in [45], where using constraints formulated from different observations, localization uncertainty was reduced. Further mainstream works [2] on probabilistic SLAM included an extended Kalman filter (EKF). Later it was shown that marginalizing past variables leads to increased fill-in (i.e. less sparse matrices) and thus results in heavier calculations.

In [8], Dellaert investigated a smoothing approache as a viable alternative to extended Kalman filter-based solutions to the problem. In particular, he looks at approaches that factorize either the associated information matrix or the measurement Jacobian into a square root form. Such techniques have several significant advantages over the EKF: they are faster yet exact, they can be used in either batch or incremental mode, are better equipped to deal with non-linear process and measurement models, and yield the entire robot trajectory, at lower cost for a large class of SLAM problems.

A smoothing approach to SLAM involves not just the most current robot location, but the entire robot trajectory up to the current time. A number of authors considered the problem of smoothing the robot trajectory only (Pose-SLAM) [18], [35], [40] particularly suited to sensors such as laser-range finders that easily yield pairwise constraints between nearby robot poses. Visual Pose-SLAM approaches are also investigated, e.g. [24], [31] are some of the latest works dealing with visual Pose-SLAM.

More generally, one can consider the full SLAM problem [48], i.e., the problem of optimal

estimation the entire set of sensor poses along with the parameters of all features in the environment. In fact, this problem has a long history in photogrammetry [4], [17], [5], where it is known as "bundle adjustment" (BA). The bundle adjustment optimization is tightly related to visual full SLAM, and is typically performed in batch mode, which means that when a new portion of observation is added, optimization over all variables is performed from scratch. Pose-SLAM differs from BA and full SLAM since in the former only the camera past and current poses are estimated, without explicit structure reconstruction (e.g. landmarks are not estimated).

The SLAM problem can be represented as a *factor graph*. Whereas each node of the graph represents a state variable to optimize, each edge between two variables represents a pairwise observation of the two nodes it connects. In the literature, many approaches have been proposed to address this class of problems. A naive implementation using standard methods like Gauss-Newton, Levenberg-Marquardt (LM) or variants of gradient descent, cannot be used for online applications. To achieve efficiency it is necessary to exploit the sparse connectivity of the graph and use advanced methods to solve sparse linear systems while re-using calculations as much as possible, as described next.

Online Inference

To be useful for a mobile robot, SLAM calculations have to be performed online, preferably in real-time. To that end, it has been shown in [8, 32, 33] that calculations can be performed incrementally as new data comes in, in contrast to operating in batch mode where the algorithm recalculates the entire solution from scratch. The core insight in these works was that new measurements typically only affect a small part of the state variables and most of the variables are unaffected. The latest of these approaches is known as iSAM2 [32] and is considered by many as the state of the art approach for computationally efficient SLAM.

Another idea to speed up the computations is to perform them in parallel. Parallelization is the key to providing real-time state updates with the ability to incorporate arbitrary loop closures. Williams et al. [51] suggested such an architecture, which allows the low-latency inference and high-latency inference to be viewed as sub-operations of a single optimization performed within a single graphical model (a factor graph). A specific factorization of the full joint density is employed that allows the different inference operations to be performed asynchronously while still recovering the optimal solution produced by a full batch optimization.

Alternative formulations have been also developed in recent years. These include, for example, structureless BA approaches, such as Light Bundle Adjustment (LBA) [25–29] that algebraically eliminate the 3D points and minimize the residual error in multiple view geometry constraints. In contrast, dense BA approaches, such as DTAM [42] and SVO [13], minimize the photogrammetric errors for each overlapping image. Dense approaches demonstrate higher accuracy but are unable to perform large scale optimizations due to high complexity.

Many state of the art approaches [16,44,46] incorporate prior information into BA techniques, such as assuming fixed camera height, and achieve higher estimation accuracy for specific conditions, e.g. for ground vehicle datasets. On the other hand such approaches are limited to

problems where these assumptions are met, for example to known motion model for planar environments.

Direct methods

The so called direct methods use either all image pixels (dense) [42], or all pixels with sufficiently large intensity gradient (semi-dense) [10], or sparsely selected pixels (sparse) [9](DSO) and minimize a photometric error obtained by direct image alignment on the used pixels. Camera poses and pixel depths are estimated by minimizing the photometric error using non-linear optimization algorithms. Since much image information can be used, direct methods are very robust in low-texture scenes and can deliver relatively dense 3D reconstructions. Consequently, due to the direct image alignment formulation, direct methods are very sensitive to unmodeled artifacts such as rolling shutter effect, camera auto exposure and gain control. More crucially, the brightness constancy assumption does not always hold in practice, which drastically reduces the performance of direct methods in environments with rapid lighting change.

Unlike other direct VO algorithms which use dense formulations, DSO performs a novel sparse point sampling across image areas with sufficient intensity gradient. Reducing the amount of data enables real-time windowed bundle adjustment (BA) which jointly optimizes for all model parameters, including camera poses, depths, camera intrinsics and affine brightness transformation factors. The optimal parameters are obtained by minimizing the photometric error using the Gauss-Newton method, which achieves a good trade-off between speed and accuracy. Obsolete and redundant information is marginalized with the Schur complement [37], and the First Estimate Jacobians technique is involved in the non-linear optimization process to avoid the inconsistency and keep the observability of the system. As a direct method, DSO is fundamentally based on the brightness constancy assumption, thus the authors proposed a photometric camera calibration pipeline to recover the irradiance images [9], [11]. Performing direct image alignment on the irradiance images instead of the original images removes artifacts polluting the brightness constancy assumption, hence drastically increases the tracking accuracy [9].

Scale drift

Existing approaches that explicitly address scale drift typically require loop closure observations. For example, Strasdat et al. [20] incorporated within bundle adjustment optimization explicit scale drift correction. Other related approaches exploit nonholonomic motion constraints (e.g. [44]), or fuse information from additional sensors (such as IMU). For example [43], [1], [41] and [30] incorporate inertial measurements into a visual framework. Adding IMU data makes estimation more robust and partially addresses the scale drift problem of monocular SLAM, but does not eliminate this source of error entirely, especially while operating over long time durations. However, the complexity of the system grows vastly with each extra sensor; one has to take into account synchronization issues, inter-sensor calibration and intelligent management of all the additional data that becomes available.

Frost et al. [14] develop an object-aware bundle adjustment approach, and use prior knowledge regarding the size of the observed objects (e.g. cars) to correct scale drift. While their approach does not require loop closure events for scale correction, it has a limitation - the mentioned prior knowledge must be available and accurate. In contrast to the above-mentioned approaches, we do not require any prior knowledge about the environment or loop-closures.

Deep learning

In the last 2-3 years, (deep) learning approaches revolutionized many problems in computer vision and other communities. Deep learning based approaches for camera pose estimation and vision based navigation are actively being developed. In particular, Kendall et al. [34] present a robust and real-time monocular six degree of freedom relocalization system. The system trains a convolutional neural network to regress the 6-DOF camera pose from a single RGB image in an end-to-end manner with no need of additional engineering or graph optimisation. The algorithm can operate indoors and outdoors in real time. This is achieved using a 23 layer pre-trained network. However, operating in exploration scenarios in environments significantly different from the training environments can lead to poor performance while using pre-trained networks. Overall, deep learning approaches to address SLAM are currently actively developed and are not yet sufficiently mature or stable.

1.3 Contribution

In this work we formulate novel image feature scale constraints and incorporate these within BA to improve estimation accuracy, especially along the optical axis of the camera in a monocular setup. This concept leverages the scale invariance property of SIFT [39] (and similar) detectors, and is based on the *key observation* that the detected feature scale changes consistently across a sequence of images. In particular, we show the detected feature scale can be predicted as a function of camera pose, landmark 3D coordinates and the corresponding 3D environment patch (see Figures 3.1 and 3.2), with the latter, according to the scale invariance property, remaining the same for different images observing the same landmark. Incorporating the mentioned feature scale constraints within BA allows to drastically reduce scale drift without requiring loop closures or any other information, given that the detected feature scales are sufficiently accurate. We show the latter can be attained simply by increasing the resolution of Gaussian kernels within the SIFT detector.

It is important to note that feature scale is already typically calculated by common feature detectors (e.g. SIFT) but is only used for image matching - see a diagram of a typical BA and SfM pipeline in Figure 2.1. Here, we propose to exploit this available information for improving the performance of BA. Note we do not interfere with image matching process, but rather propose to make better use of its products.

The idea of using feature scale has been proposed in the past, but in different contexts. For example, Ta et al. [47] use feature scale to determine if a landmark is sufficiently far away to

consider it for rotation updates and to steer the robot to these far-away landmarks (termed Vistas) to avoid collisions in an indoor environment. Guzel et al. [19] recently suggested to use SIFT feature scale for distance estimation.

However, to the best of our knowledge, incorporating image feature scale constraints within BA is novel. In addition to improving accuracy, our method, termed, bundle adjustment with feature scale (BAFS), has also the capability to estimate the actual landmark (object) sizes, up to an overall scale.

Chapter 2

Problem formulation

We consider a sequence of N images captured from different and unknown camera poses. As schematically shown in Figure 2.1, given this image sequence, existing approaches typically first extract track image features (e.g. SIFT [38] or SURF [3] detectors) from each of the images. These features are then matched, typically using the Random Sample Consensus algorithm (RANSAC) [12] to filter out outliers. The result of this phase are corresponding image features across different images which are then fed into the bundle adjustment process as described below. An example for matched features in two images is shown in Figure 2.2.

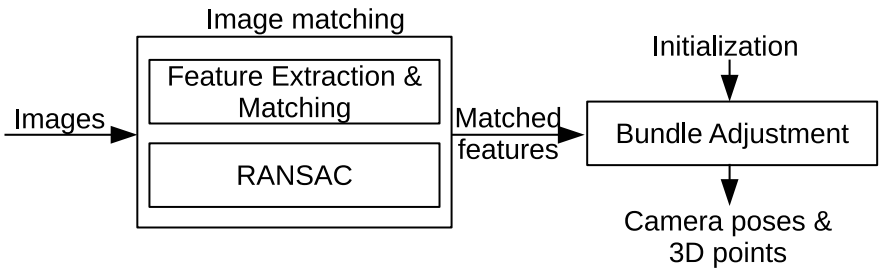


Figure 2.1: Overall framework of monocular BA

Denote the camera pose that captured the i-th image by $x_i = \{R_i, t_i\}$, with rotation matrix R_i and translation vector t_i, and let Z_i represent all the landmark observations of that image, with a single image observation of some landmark l_j denoted by $z_i^j \in Z_i$. Let X represent all the camera poses and L represent all the observed landmarks,

$$X \doteq \{x_1, \ldots, x_i, \ldots, x_N\} \quad , \quad L \doteq \left\{l_1, \ldots, l_j, \ldots, l_M\right\}, \tag{2.1}$$

where M is the number of observed landmarks. These landmarks represent 3D scene points that generate the detected 2D visual features.

We denote by $\pi(x, l)$ the standard projection operator [22], and write the measurement likelihood for an image observation z given camera pose x and landmark l as

$$\mathbb{P}(z|x, l) = \frac{1}{\sqrt{det(2\pi\Sigma)}} \exp\left(-\frac{1}{2} \|z - \pi(x, l)\|^2_{\Sigma_v}\right), \tag{2.2}$$

Figure 2.2: Matched SIFT features.

where we conventionally assumed image noise is sampled from a zero-mean Gaussian distribution $N(0, \Sigma_v)$, and $\|a\|^2_{\Sigma_v} \doteq a^T \Sigma_v^{-1} a$ is the squared Mahalanobis distance.

The joint probability distribution function (pdf) for N camera frames can now be written as

$$\mathbb{P}(X, L|\mathcal{Z}) \propto priors \cdot \prod_{i=1}^{N} \prod_{j \in \mathcal{M}_i} \mathbb{P}(z_i^j | x_i, l_j) \tag{2.3}$$

where $\mathcal{Z} \doteq \{Z_i\}_{i=1}^N$ is the set of all image observations from all images and $\mathcal{M}_i$ is a set of indexes of the landmarks observed from camera pose i. The $priors$ term includes all the prior available information; this term will be omitted from now on for conciseness.

The MAP estimation of X and L is given by

$$X^\star, L^\star = \arg\max_{X,L} \mathbb{P}(X, L|\mathcal{Z}), \tag{2.4}$$

and can be calculated using state of the art computationally efficient solvers [7, 36] that solve the following non-linear least-squares problem:

$$J_{BA}(X, L) \doteq \sum_{i}^{N} \sum_{j \in \mathcal{M}_i} \left\| z_i^j - \pi\left(x_i, l_j\right) \right\|^2_{\Sigma_v}. \tag{2.5}$$

A key problem in the described monocular camera setup is scale drift as information provided by a single camera, without considering any additional information, can only be used to recover the camera motion and the 3D environment up to a common scale, which drifts over time.

Figure 2.3 demonstrates bundle adjustment results in a real-imagery monocular scenario

with a forward-facing camera from KITTI dataset [15]. Estimated path is marked with dashed blue line and black solid line corresponds to ground truth. One can notice that estimated track length is stretched relatively to ground truth, a well-known issue called *scale drift*. For a forward-facing camera, scale drift corresponds to stretching of the estimated trajectory along optical axis. Referring again to Figure 2.3, right after the start and before the first right turn (denoted as area 1) position estimate drifts about 8% with respect to ground truth, while towards the end of trajectory (area 2), position drift grows to 94%.

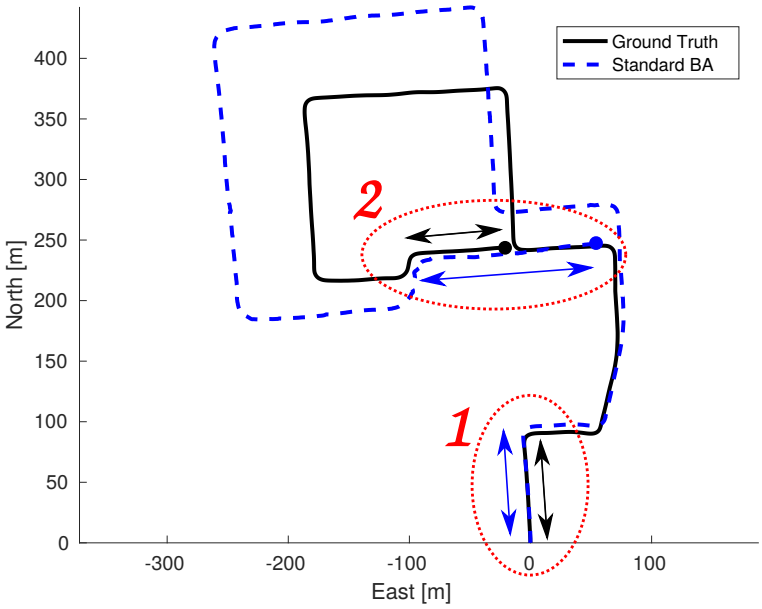


Figure 2.3: Scale drift along optical axis. Top view. Forward-facing camera moves along the trajectory. Black is GT, blue - estimated path. For the 1st 100 frames estimated path is 8% longer than ground truth (zone 1). For the last 100 of frames estimated path is 94% longer than ground truth (zone 2).

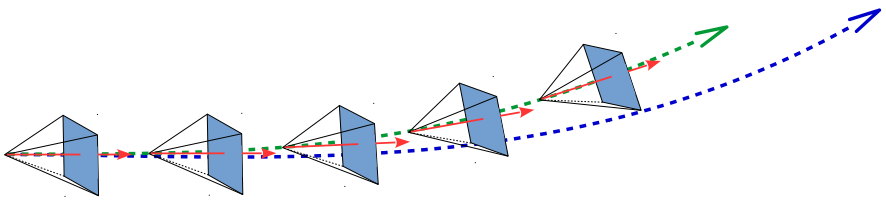


Figure 2.4: Zoomed-in sequence of poses for a singe forward-facing camera. Red arrows show camera optical axis direction. Green dashed line is estimated track. Blue line is a ground truth track. Drift along optical axis is cumulative position error component along movement direction.

Drift along the optical axis is indeed a well known problem, which is often addressed only upon identifying a loop closure event or considering availability of additional sensors or prior knowledge. In contrast, in the next chapter we formulate a new type of constraints that allow to enhance estimation accuracy, in particular along the camera optical axis, without requiring loop

closures or additional prior knowledge.

Chapter 3

Approach

3.1 Feature Scale Constraint Formulation

Standard bundle adjustment formulation exploits only partial information extracted from images by typical image matching approaches: only image coordinates from corresponding views are used, while an image feature (e.g. SIFT feature) is typically accompanied also with two additional parameters - scale and orientation. We propose to incorporate scale information into bundle adjustment optimization by formulating appropriate constraints that describe how feature scale changes for different views according to camera motion and observed landmarks. The corresponding idea, that we call Bundle Adjustment with Feature Scale (BAFS), is schematically illustrated in Figure 3.1.

Our *key observation* is that the detected scales of matched features from different frames capture the *same* portion (patch) of the 3D environment, as illustrated in Figure 3.1. This observation leverages the scale invariance property that typical feature detectors (e.g. SIFT [39]) satisfy. As an example, we consider the image sequence shown in Figure 3.2, where a single feature is tracked and its detected scale across different images is explicitly shown. One can note that, indeed, in all of the frames, the detected scale represents an identical portion of the environment, i.e. the contents inside of the circle with radius equals to detected scale is identical in all frames.

We shall consider the mentioned 3D environment patch extent as virtual landmark size and denote it for the jth landmark by S_j. Based on the above key observation, we argue the detected feature scales in different images change consistently and can be predicted. Specifically, letting s_i^j denote the detected feature scale of the jth landmark in the ith image frame, and considering a perspective camera, we propose the following observation model for s_i^j

$$s_i^j = f\frac{S_j}{d_i^j} + v_i, \tag{3.1}$$

where f is the focal length and v_i is the measurement noise which is modelled to be sampled from a zero-mean Gaussian distribution with covariance Σ_{fs}, i.e. $v_i \sim N\left(0, \Sigma_{fs}\right)$. In Eq. (3.1) we use d_i^j to denote the distance along optical axis from the camera pose x_i to landmark l_j. In

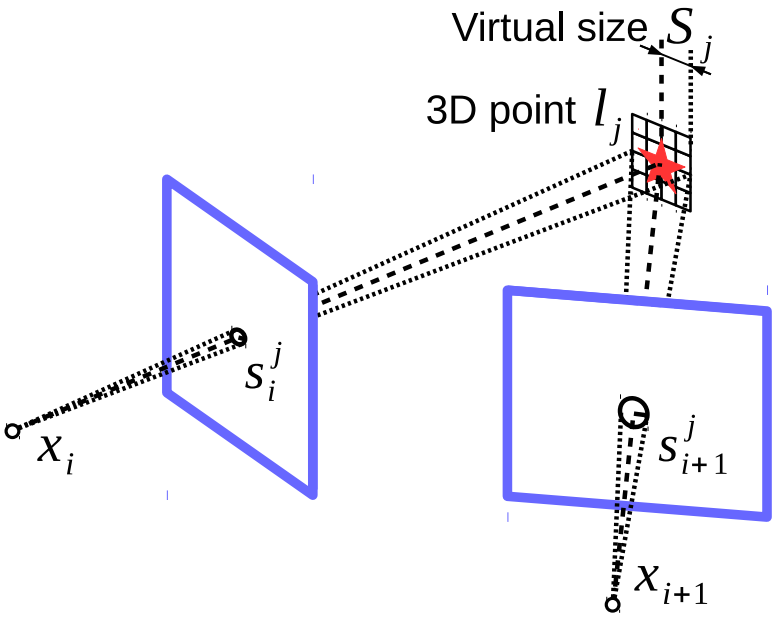


Figure 3.1: Feature scale is modeled as a projection of a virtual landmark size in 3D environment onto the image plane. We leverage the scale invariance property of typical feature detectors, according to which, detected scales of matched features from different images correspond to the same virtual landmark size in the 3D environment, and incorporate novel feature scale constraints within BA.

other words, assuming the optical axis is the z axis in the camera frame,

$$d_i^j(x_i, l_j) \doteq z_c \ , \ \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = R_i l_j + t_i, \tag{3.2}$$

where R_i and t_i are the ith camera rotation matrix and translation vector, i.e. $x_i = \{R_i, t_i\}$.

We note one might be tempted to consider d_i^j to be simply the range between the camera optical center and the landmark 3D position. However, this model is incorrect as we discuss now. To see that, consider again the sequence of images shown in Figure 3.2, where the same landmark is tracked. The landmark is relatively distant and the camera (car) is performing an almost pure rotation motion, such that the range to the landmark is approximately constant. As the camera rotates, the landmark is projected closer and closer to the center of the image while the corresponding detected feature scales are shown in the zoom-in figures. One can observe that these decrease as the features move closer to the center of the image. Figure 3.3 illustrates this scenario schematically. It is shown geometrically that the same landmark (means $S_1 = S_2 = S_3$) observed at the same range from the camera optical center produces different feature scales, so $s_i^1 < s_i^2 < s_i^3$. Now, modeling d_i as range and given some value for S_j in Eq. (3.1) would yield identical, up-to-noise, feature scale predictions, contradicting the detected feature scales $s_i^1 < s_i^2 < s_i^3$. In contrast, modeling d_i as distance along optical axis would and noting $d_1 > d_2 > d_3$, correctly predicts the observed feature scales.

In Eq. 3.1 we assumed the scale measurement noise is sampled from a zero mean Gaussian

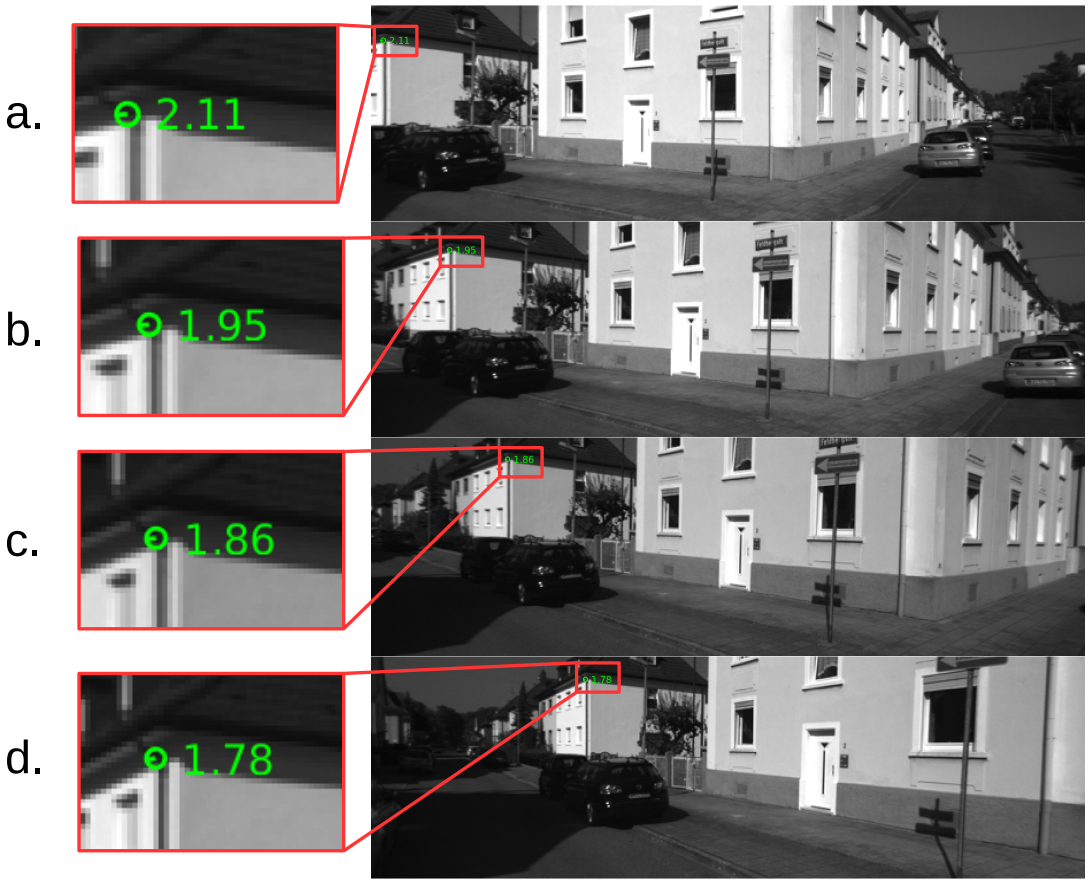


Figure 3.2: A landmark is observed while the camera performs a left turn, from (a) to (d). The detected feature scale in each frame is shown in the zoom-in figures.

distribution. While convenient, it is not obvious to what extend this is a valid assumption for the considered scale measurement model. Using statistics for scale measurements from a dataset of real images we show that, indeed, a Gaussian distribution is a good approximation of the error between scale measurements and our model. This is shown in Figure 3.4, which presents the scale error distribution for 500 frames (approximately 20000 measurements) from KITTI [15] dataset. Scale error measure is estimated for each landmark observation as difference between measured value and predicted according to model (Eq.3.1) value.

One can note that the distribution is zero-mean. Empirically we fitted a sigma parameter to get an upper bound approximation for the scale error. Modeling sigma with a larger value, the measurement impact will decrease, while modeling sigma with too small value might introduce additional error into optimization.

Based on the observation model (3.1) we can now define the corresponding feature scale measurement likelihood as

$$\mathbb{P}(s_i^j | S_j, x_i, l_j) \doteq \frac{1}{\sqrt{|2\pi\Sigma_{fs}|}} \exp\left[-\frac{1}{2}\left\| s_i^j - f\frac{S_j}{d_i^j}\right\|^2_{\Sigma_{fs}}\right] \tag{3.3}$$

As seen, the above likelihood is conditioned on the virtual landmark size S_j. Since the latter is

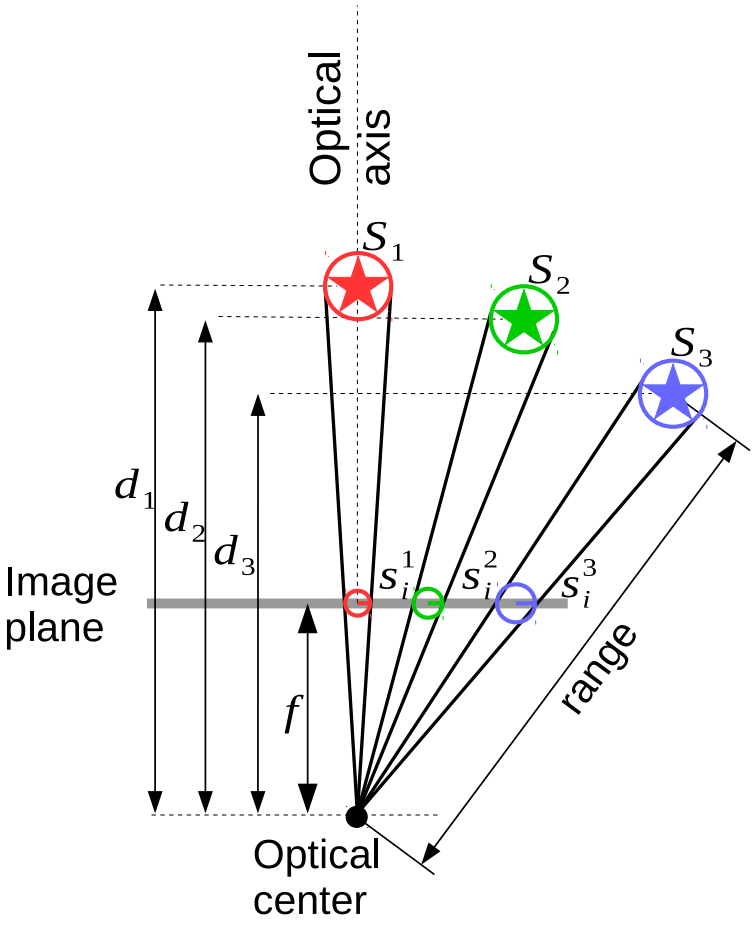


Figure 3.3: Landmark of the same virtual size S_j is observed at a constant range from the camera's optical center, producing different scale projections depending on the distance along optical axis.

actually unknown, we treat it as random variable and infer it, along other variables.

We can now formulate the feature scale constraint and the corresponding likelihood for each landmark observation. Letting $S \doteq \{S_j\}$ denote the virtual landmark sizes for all observed landmarks, and incorporating all the measurement likelihood terms (3.3) yields the following joint pdf (omitting the priors terms)

$$\mathbb{P}(X, L, S|\mathcal{Z}) \propto \prod_{i}^{N} \prod_{j \in \mathcal{M}_i} \mathbb{P}(z_i^j|x_i, l_j)\mathbb{P}(s_i^j|S_j, x_i, l_j). \tag{3.4}$$

As seen, for each landmark observation we now have two types of constraints: projection and scale constraints.

Taking $-log\left[p\left(X, L, S|\mathcal{Z}\right)\right]$ we get the following corresponding non-linear least-squares problem

$$J_{BAFS}\left(X, L, S\right) \doteq \sum_{i}^{N} \sum_{j \in \mathcal{M}_i} \left\| z_i^j - \pi\left(x_i, l_j\right) \right\|_{\Sigma_v}^2 + \left\| s_i^j - f\frac{S_j}{d_i^j} \right\|_{\Sigma_{fs}}^2,$$

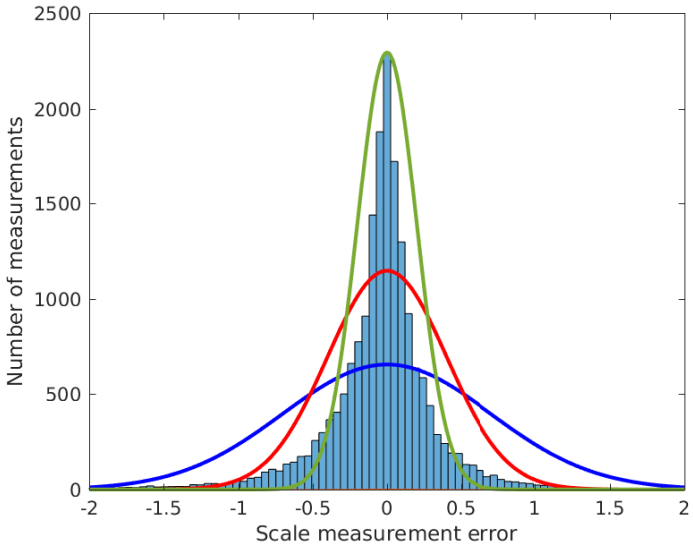


Figure 3.4: Scale error distribution for real images dataset. Red curve is Gaussian curve chosen to approximate data distribution. Green curve over estimates measurements accuracy, blue curve - underestimates.

and we can use state of the art efficient solvers to find the MAP solution $X^\star, L^\star, S^\star$.

3.2 Computational Complexity and Factor Graph Reduction

The obtained joint pdf can be conventionally represented with a factor graph model [8]. A single landmark observation is now used to formulate a projection and feature scale factors. Adding a feature scale factor for each landmark observation corresponds to the factor graph shown in Figure 3.5b. However, for a scenario of N camera frames and M landmarks, this naïve approach increases the number of variables in the optimization from $6M + 3N$ to $6M + 4N$, and doubles the number of factors, which can severely impact optimization time.

Instead, we propose the following simple heuristic. We add feature scale factors and new virtual landmark size variables only for long-term landmarks that are observed for long period of time (number of images above a threshold). Moreover, empirically we notice that these long-term landmarks correspond to ”strong” features which are usually measured more accurately. This property allows to model Σ_{fs} with a lower value than usual, giving more weight to scale constraints in the optimization. Figure 3.5c illustrates a factor graph that corresponds to this heuristic.

3.3 Variable initialization

As the MAP solution is obtained via iterative optimization, each of the optimized variables needs to be initialized. While initialization of camera poses and landmarks can be done using conventional approaches [21], the following method can be used to initialize the virtual landmark

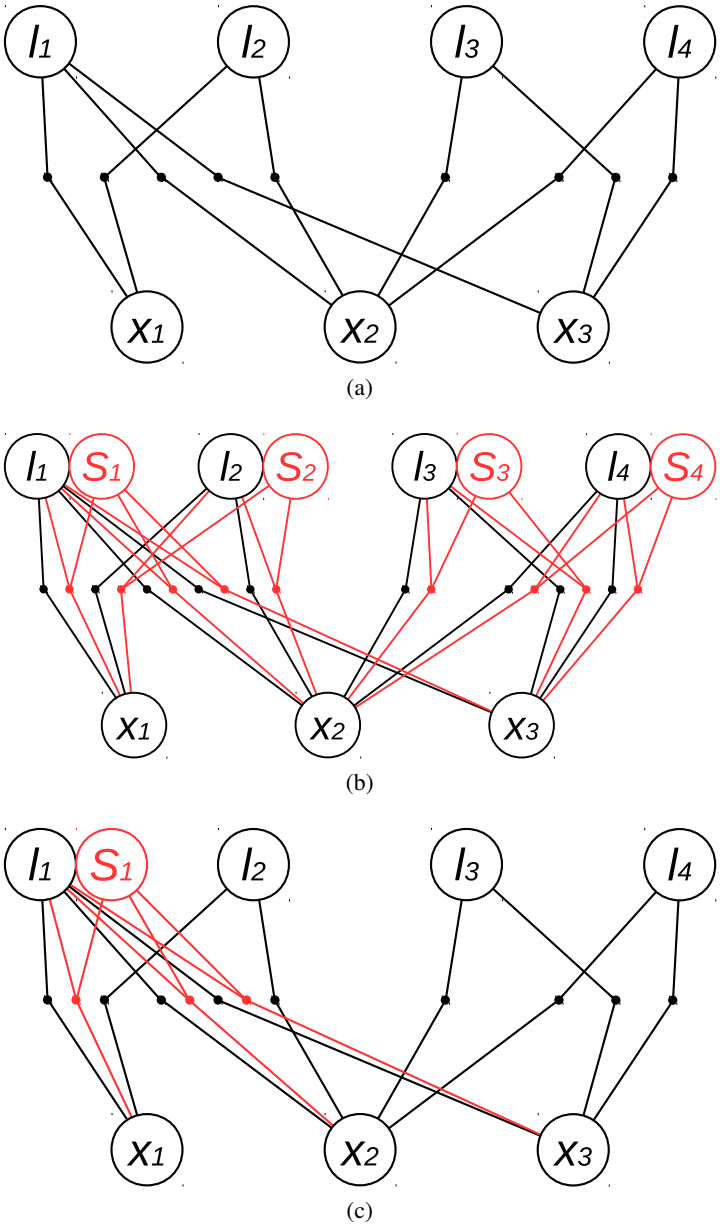


Figure 3.5: Factor graph representations: (a) standard BA with projection factors only; (b) BAFS with naïvely added all feature scale factors; (c) BAFS with feature scale factors added only for long-term landmarks (l_1, in this case).

size. After a new landmark l^j is observed and initialized (e.g. via triangulation which requires two landmark observations), the distance along optical axis d_i^J from camera pose to the landmark can be estimated. We then initialize the corresponding virtual landmark size variable, S_j, using the equation

$$S_j = s_i^j \frac{d_i^j}{f}, \tag{3.5}$$

which is obtained from Eq. (3.1) while neglecting the noise.

In our implementation, we initialize each new landmark via triangulation given two land-

mark observations, and initialize S_j by taking an average value of Eq. (3.5) considering the corresponding two detected feature scales.

3.4 Enhancement of Feature Scale Measurement Accuracy

Thus far, we incorporated our novel feature scale constraints (3.5) within bundle adjustment, but did not discuss when these constraints will actually have impact on estimation accuracy. This aspect naturally depends on how accurate the feature scale observations are in the first place, as modelling this determines Σ_{fs}, the measurement noise covariance from Eq. (3.3).

Intuitively, more accurate feature scale observations and the corresponding lower values of Σ_{fs} will yield better estimation accuracy. In fact, there is a scenario-dependent upper threshold for Σ_{fs} at which the feature scale constraints will have no contribution at all. We study this statement using a synthetic dataset of a downward-facing camera flying at constant height and observing landmarks scattered in 3D-space, occasionally performing loop closures (see Figure 3.6). In this scenario we assigned each landmark l_j a corresponding (ground truth) size S_j, and simulated image and feature scale observations while corrupting the latter with sampled Gaussian noise considering different values of Σ_{fs}.

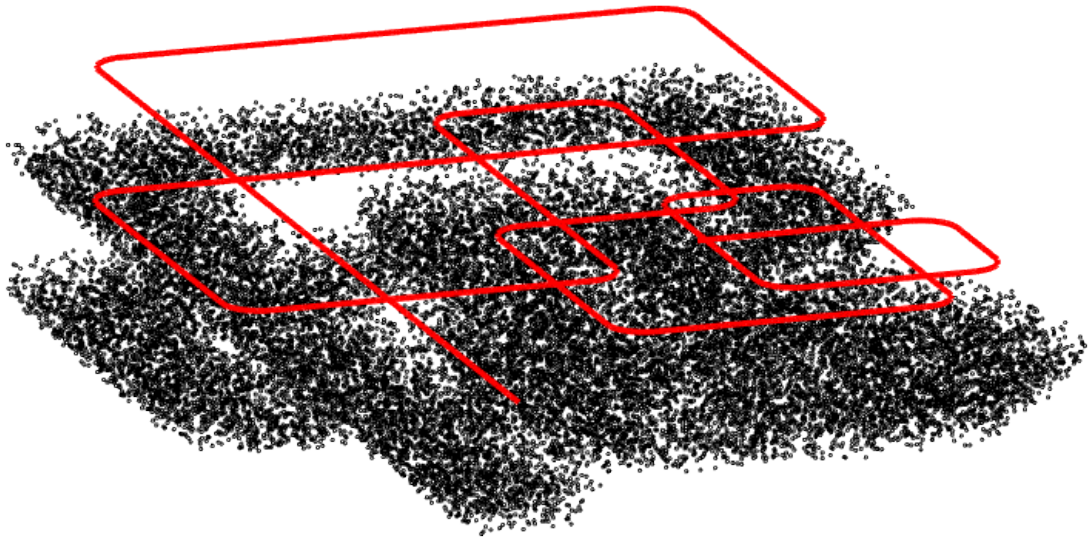


Figure 3.6: Simulated scenario of an aerial downward-facing camera observing randomly-scatted landmarks. Camera's trajectory is shown in red.

As expected, running on this synthetic data showed that accuracy of feature scales is extremely important to improve standard BA precision. Figure 3.7 shows results, in terms of position estimation error, for different simulated values of Σ_{fs} such that $\Sigma_{fs1} > \Sigma_{fs2} > \Sigma_{fs3} > \Sigma_{fs4} > \Sigma_{fs5}$. One can observe the dashed curves, that correspond to two highest values of Σ_{fs} (i.e. Σ_{fs1} and Σ_{fs2}) are very close to the standard BA curve, and do not improve estimation accuracy. As we consider smaller values of Σ_{fs}, estimation accuracy gets dramatically improved.

While the above discussion referred to simulated feature scale observations, in reality these are produced by feature detectors such as SIFT. Unfortunately, we empirically observed that

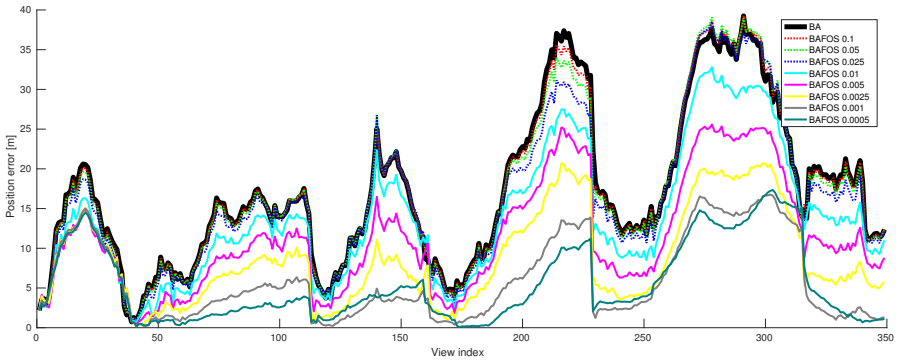


Figure 3.7: Position estimation error. Each curve corresponds to BA with feature scale constraints with noise in simulated feature scale measurements sampled from a Gaussian with different Σ_{fs}. Black solid curve corresponds to standard BA.

incorporating scale constraints into the optimization does not yield any significant improvement, thereby indicating the actual feature scale measurements are not of sufficient quality (i.e. too noisy).

We propose a simple method to address this difficulty. Recall that a SIFT detector first blurs the image with different Gaussian kernels, calculates difference between blurred images with successive kernels, and searches for maxima both spatially and across different kernels. The former determines the feature coordinates, while the latter determines the scale (see Figure 3.8). Therefore, feature scale can be determined only up to resolution of the Gaussian kernels used in this process. To increase accuracy of the detected feature scales, we propose to use a finer resolution of the Gaussian kernels. This simple idea is illustrated in Figure 3.8, where additional kernels and corresponding blurred images are shown in red. Furthermore, while in this work Σ_{fs} is specified manually given detected feature scales, we envision the utilized Gaussian kernels resolution could be used to determine Σ_{fs}. However, exploring this aspect is left for future research. As we show in the sequel, using feature scales with enhanced resolution yields a significant improvement in position estimation accuracy.

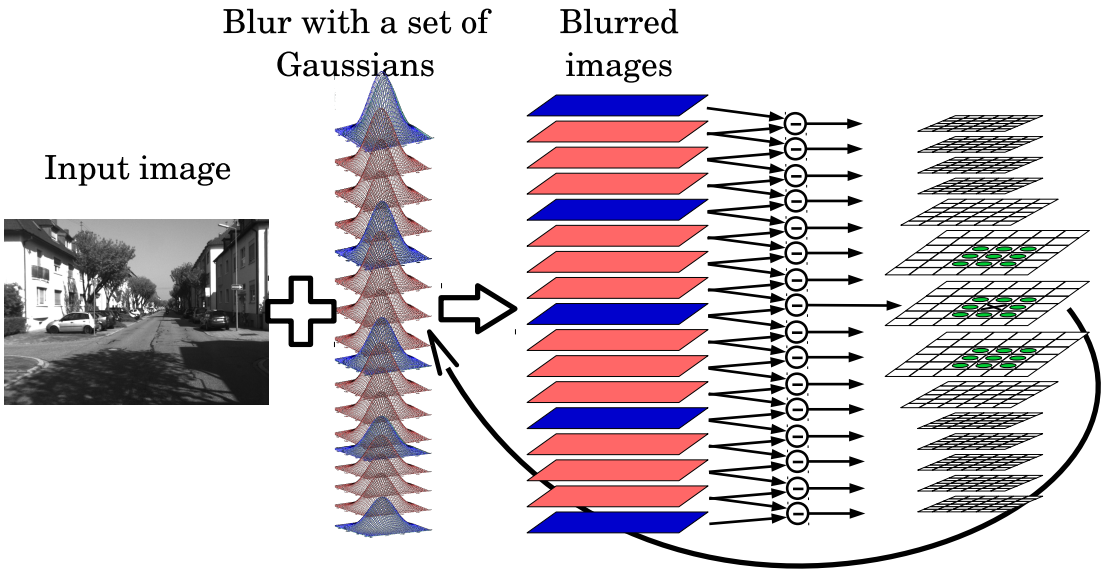


Figure 3.8: SIFT scale estimation process. (i) Blur each input image with a set of Gaussian kernels. (ii) Calculate Difference of Gaussians (DoG). (iii) Feature scale is set as the average of the two Gaussian kernels that correspond to the local-maxima DoG layer.

3.5 Flat Environments

In this section we study empirically the performance of the proposed method in environments that are either flat, or appear as flat, e.g. when flying at a relatively high altitude. In such scenarios, feature scales tend to have constant values with negligible variations from frame to frame, while the the distance along optical axis up to observed landmarks is also roughly fixed. We simulated a flat scenario (see Figure 3.9) to examine difference of requirements to measurements accuracy.

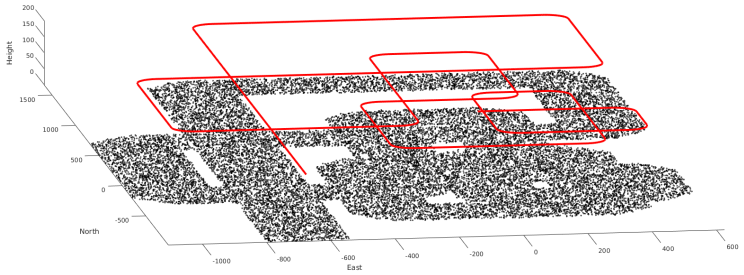
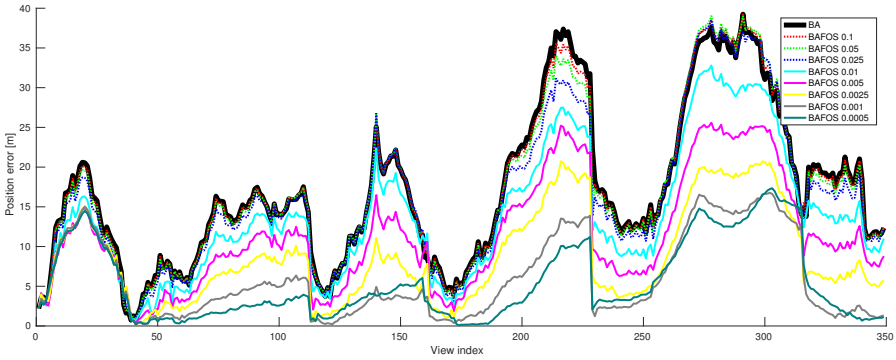
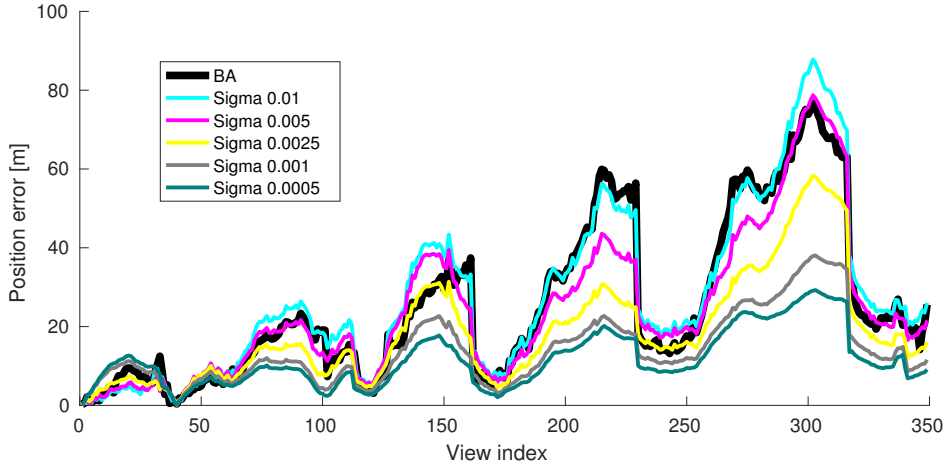


Figure 3.9: Flat simulation scenario. Downward facing camera observes flat environment.

Similar to simulation in section 3.4 we incorporated our novel feature scale constraints (3.5) within bundle adjustment to identify conditions when these constraints will actually have impact on estimation accuracy. Modeling the feature scale observations accuracy with different measurement noise covariance from Eq. (3.3) we got an interesting result for flat scenario defining higher sufficient measurement accuracy threshold to benefit the estimation. In Figure 3.10 we show two different simulation worlds with corresponding error results with a set of simulated measurement noise covariances. Figure 3.10a corresponds to non-flat world shown previously in Figure 3.6 and contains results for the following simulation runs: black solid curve corresponding to standard BA accuracy, dotted curves corresponding to BAFS approach with feature scale measurements of insufficient accuracy and solid coloured curves corresponding to improved accuracy, where one can notice than low Σ_{fs} is modelled than lower error is obtained. Comparing to Figure 3.10b where runs with insufficient previously accuracy are dropped it might be noticed that cyan and magenta curves are also of insufficient accuracy giving error of the same rate as standard BA. Only starting with the yellow curve, accuracy stably improves which means that for flat environments much more accurate scale measurements are necessary in order to get any benefit using our proposed feature scale constraints.



(a)



(b)

Figure 3.10: Position estimation error. Each curve corresponds to BA with feature scale constraints with noise in simulated feature scale measurements sampled from a Gaussian with different Σ_{fs}. Black solid curve corresponds to standard BA. Subfigure (a) corresponds to landmarks scattered in space shown in Figure 3.6, Subfigure (b) correstonds to scenario in Figure 3.9

3.6 Application to Object-based Bundle Adjustment

The proposed concept of feature scale constraints is applicable also using alternative scale invariant quantities detected in the images. Here, we briefly describe one such application, considering object-level BA while using detected object bounding boxes in the images (see Figure 3.11).

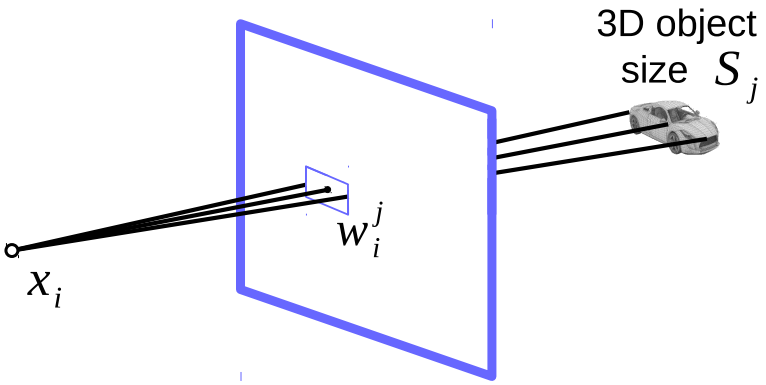


Figure 3.11: Object detected bounding box in the image plane.

Specifically, considering the detected bounding boxes of far away stationary objects as scale invariant, we formulate object scale constraints in a similar manner to feature scale constraints (3.1). Close objects are not taken into account as the corresponding detected bounding boxes might be obtained from significantly different viewpoints and provide inconsistent scale measurements. In our implementation, we use HOG object detector [6] to identify bounding boxes, and formulate the scale constraint considering the detected width and height instead of feature scale. For example, for the *j*th object observed at the *i*th frame, the scale constraint is

$$w_i^j = f\frac{S_j^{obj}}{d_i^j} + v_i^{obj}, \tag{3.6}$$

where w_i^j is the detected bounding box width (see Figure 3.11), and v_i^{obj} is a Gaussian noise that corresponds to the accuracy of bounding box picked by the object detector. Interestingly, the virtual landmark size variable S_j^{obj} now corresponds to object size, which is inferred as part of the optimization process, up to an overall scale.

In addition to the above scale constraints, we formulate projection constraints considering projection of virtual landmarks onto the image. A virtual landmark is defined as the center of the corresponding object, while its projection is found as the center of the detected bounding box. Similar to standard 3D landmarks, object position and width variables can be initialized and introduced into optimization after at least 2 object observations.

Overall, the cost function in our implementation for BA with object scale constraints has the

following form

$$J_{BA+OS} = \sum_i \sum_{j \in \mathcal{M}_i} \left\| z_{i,j} - \pi(x_i, l_j) \right\|^2_{\Sigma_v} + \sum_i \sum_{k \in \mathcal{M}_i^{obj}} \left(\gamma \left\| z_{i,k}^{obj} - \pi(x_i, l_k^{obj}) \right\|^2 + \beta \left\| w_{i,k} - f \frac{S_k^{obj}}{d_i} \right\|^2 . \right) \tag{3.7}$$

Here, $z_{i,k}^{obj}$ is a virtual landmark projection on the image plane, $\pi(x_i, l_j^{obj})$ is the predicted object position in the image, while S_k^{obj} is the object size in 3D, and $w_{i,j}$ is the measured object width. Similar to $\mathcal{M}_i$, we represent by $\mathcal{M}_i^{obj}$ the object indices observed in the i-th frame. The coefficients β and γ are weights defining impact of object scale and object projection position, respectively, on the estimation. It is interesting to note that using scale constraints at object level allows to estimate object size S_k^{obj} for free as it is a part of optimization process.

Each object detection is accompanied with a score, representing accuracy of the detected bounding box. This enables flexible weights tuning, individually for each observation, where higher scores correspond to larger β and γ values. Object size (S^{obj}) estimation accuracy also depends on the object detection scores.

Remark: As a limitation of using object width or height for formulating scale constraints, we note that the measured object width (bounding box width) is not always scale invariant; as an example, consider car observations from significantly different viewpoints (e.g. front and side observation). It is for this reason that in this work we used only width of far-away objects, when their projection change is insignificant and can be neglected.

Chapter 4

Results

We implemented a classical sparse feature based BA framework using the GTSAM [7] solver and the provided Matlab wrapper. As GTSAM supports only projection factors out of box, we implemented a scale factor, which corresponds to the feature scale measurement likelihood (3.3). As described in Section 3.4, we enhance standard SIFT feature scale resolution by increasing the number of layers per octave from default value 3 up to 15. In the reported results we used $\Sigma_v = 0.5$ and manually set Σ_{fs} to 0.2 while adding feature scale constraints for all landmarks. We were able to drop Σ_{fs} up to 0.1 when adding these constraints only for long-term landmarks, as empirically we observed the corresponding detected feature scales are typically of higher quality.

To test the performance of our approach we used two outdoor sequences from the KITTI dataset [15]; typical images are given in the Figure 4.1. Contrary to many other methods tested on this dataset, we do not involve any prior knowledge like camera height or typical object sizes about the environment and solve pure standard bundle adjustment problem with our novel feature scale constraints. Moreover, in this work we do *not* use any loop closures, thereby examining the contribution of the developed scale constraints on estimation accuracy over time.

Figure 4.1: Typical images from KITTI dataset.

4.1 Results with enhanced SIFT scale resolution

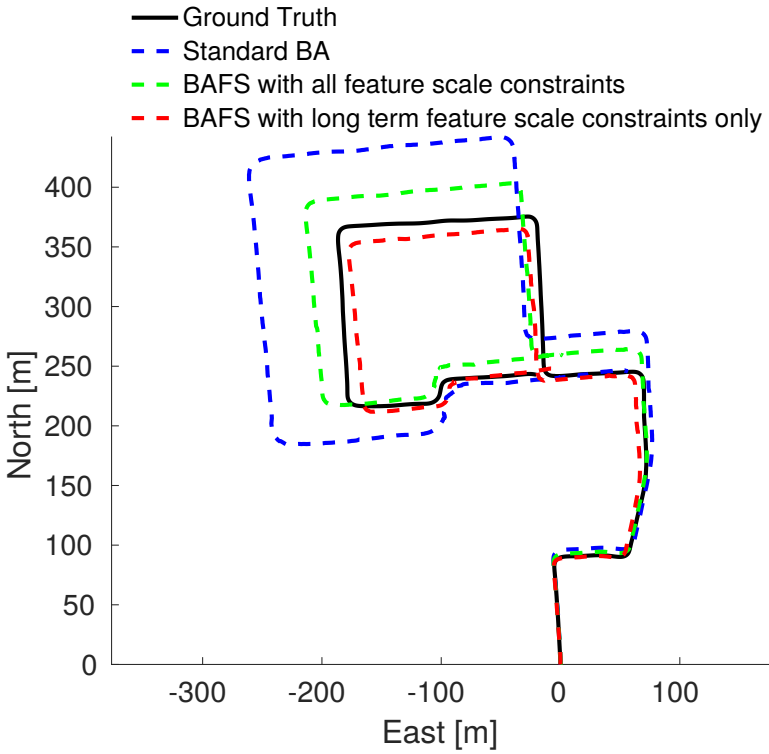
The results for both of the considered sequences are shown in Figures 4.2, 4.2b and compared to ground truth, and standard BA. Additionally, we show our approach with feature scale constraints added for all landmarks, or only for long-term landmarks (see Section 3.2). Specifically, Figure 4.2 shows the estimated trajectories (top view), Figure 4.3 position estimation errors, and Figure 4.4 optimization time. The shown results are obtained in an incremental fashion that is suitable for online applications, i.e. the k camera pose is estimated given available data only up to that time. The reported optimization times for all methods correspond to batch Levenberg-Marquardt optimization with identical settings; we expect running time to drastically drop upon switching to iSAM2 [32] but leave this endeavor to future research.

As seen in Figures 4.2a and 4.2b, standard BA suffers from significant drift along optical axis which is manifested in continuous stretching of the estimated trajectory compared to ground truth. One can notice that position estimation perpendicular to motion heading is more accurate than along the optical axis.

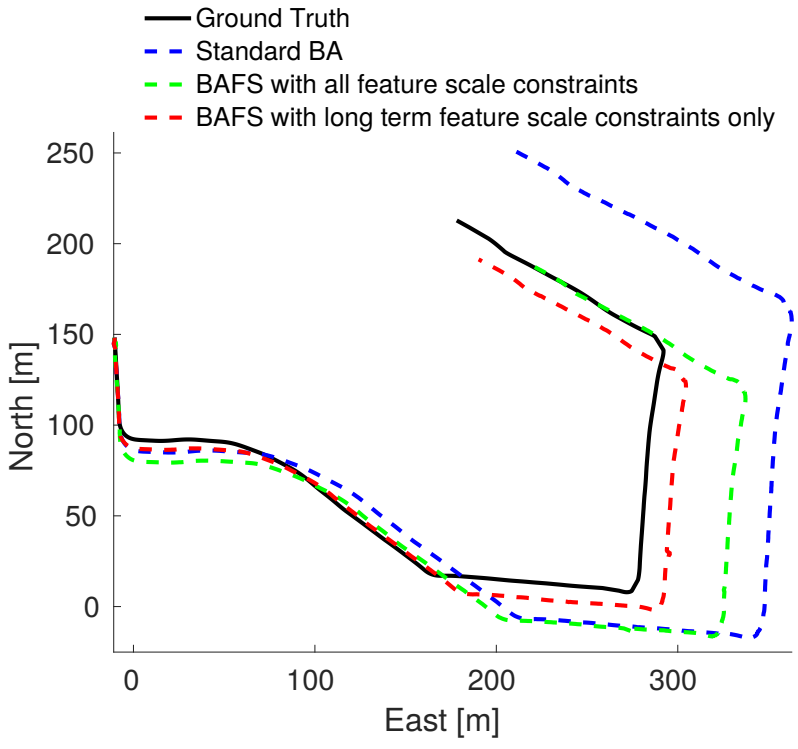
The green curve, which corresponds to BAFS with feature scale constraints for all landmarks, is obviously closer to ground truth and the main improvement is caused by discarding the stretching along optical axis, i.e. reducing scale drift. This result corresponds to our approach using both projection and scale constraints. The corresponding absolute position error is significantly improved (green curve in Figures 4.3a and 4.3b) compared to standard BA approach which only exploits feature projection factors. In particular, position estimation error is often reduced by a factor of about 2.5, e.g. from around 90 meters to 40 meters around frame number 950 in Figure 4.3a.

Estimation performance is even further improved by BAFS with feature scale constraints added only for long-term landmarks, as shown by the red curves in the figures. For example, the above-mentioned 40 meters position error is reduced to 30 meters at the same time instant (see Figure 4.3a). This is perhaps somewhat surprising result, as we use less constraints but obtain higher accuracy. We hypothesize this happens since long term feature scales tend to be more robust and accurate.

Figures 4.4a and 4.4b provide the optimization time for both sequences. One can observe that naïvely using all feature scale constraints considerably increases optimization time compared to standard BA, while adding feature scale constraint only for long-term landmarks does not increase optimization time significantly.

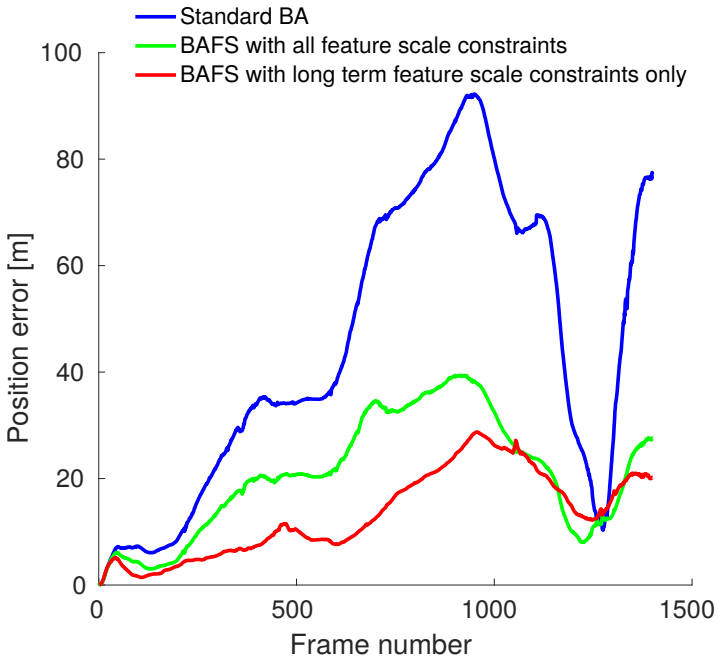


(a)

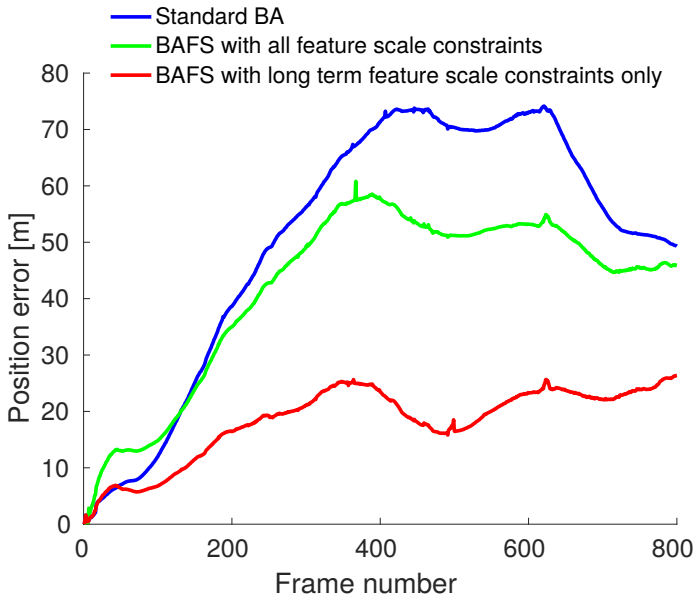


(b)

Figure 4.2: Each image provides top view of estimated trajectory for a different KITTI dataset sequence using SIFT features with enhanced scale resolution. Estimation with standard BA is shown with blue, BA + all Feature Scale constraints - green, BA + long term feature scale constraints - red, black solid line is ground truth

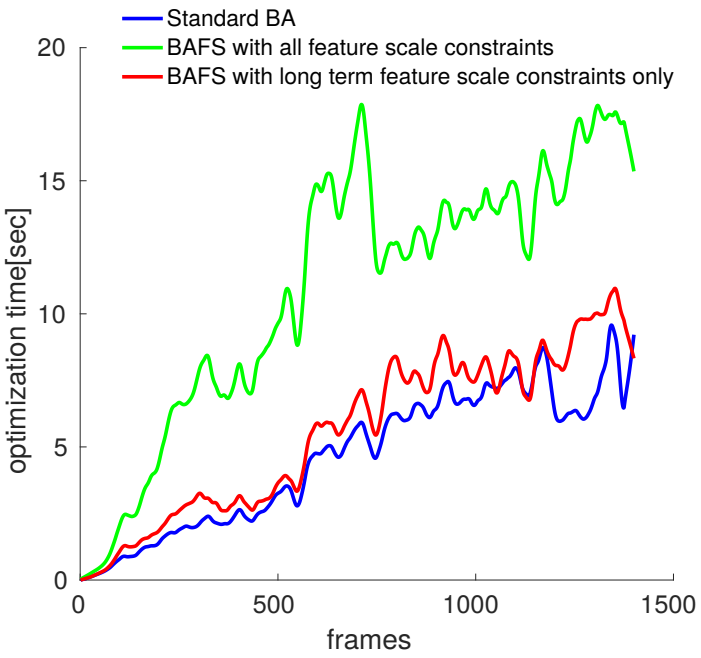


(a)

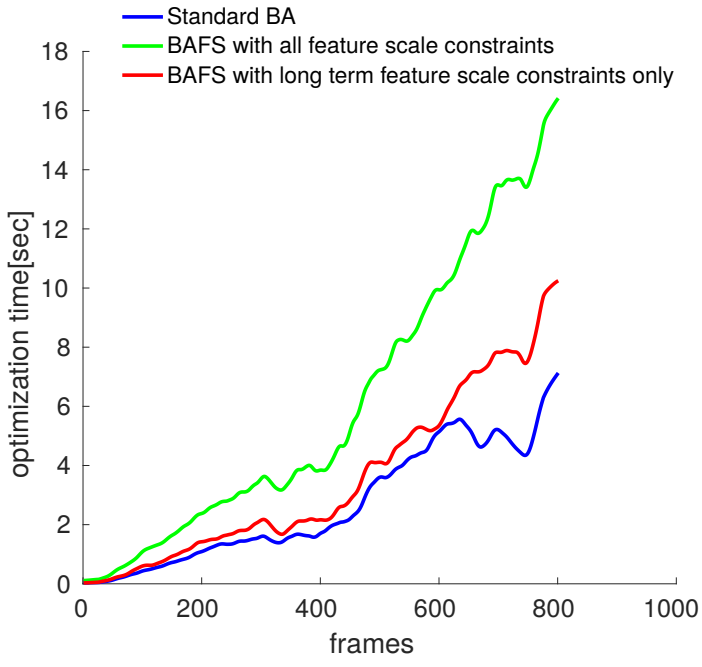


(b)

Figure 4.3: Each image provides norm of position estimation error as a function of time for the 2 different KITTI dataset sequences using SIFT features with enhanced scale resolution. Estimation with standard BA is shown with blue, BA + all Feature Scale constraints - green, BA + long term feature scale constraints - red.



(a)



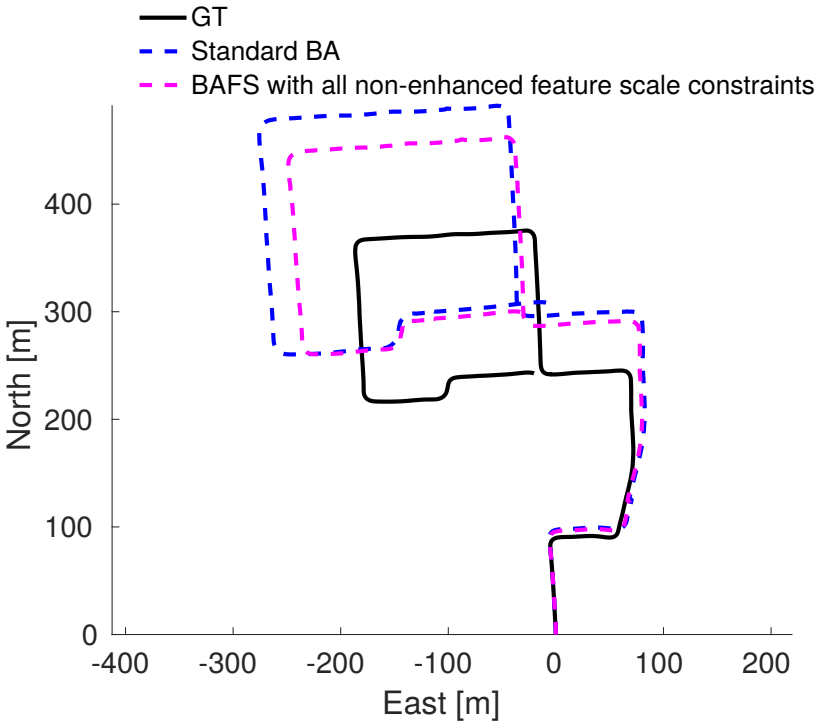
(b)

Figure 4.4: Each image provides optimization time for each frame for the 2 different KITTI dataset sequences using SIFT features with enhanced scale resolution. Estimation with standard BA is shown with blue, BA + all Feature Scale constraints - green, BA + long term feature scale constraints - red.

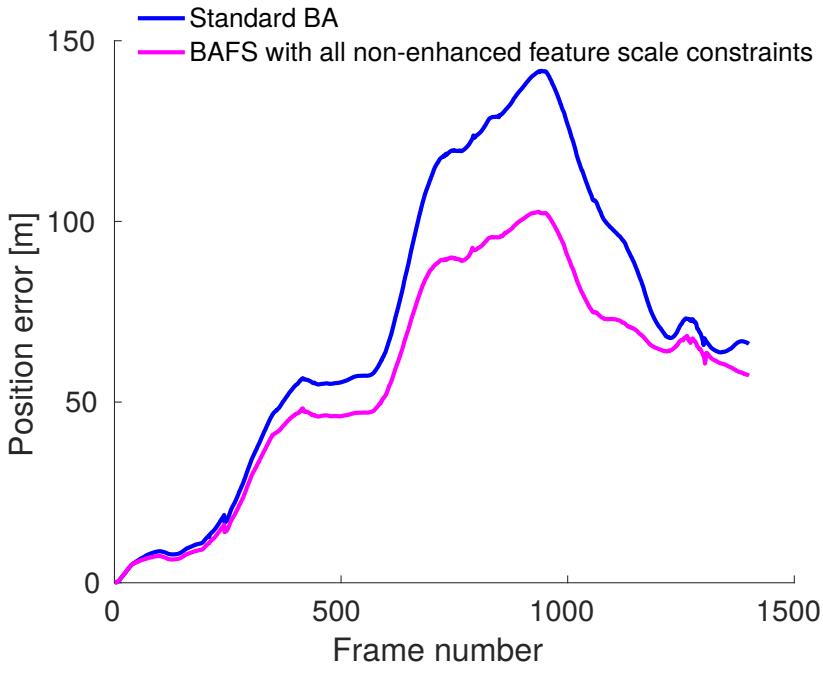
4.2 Results with non-enhanced SIFT scale resolution

The results above were obtained using enhanced-resolution feature scales (see Section 3.4). To demonstrate the importance of improving the accuracy of detected feature scales, we show in Figures 4.5a and 4.5b results of our approach without such enhancement, i.e. using default SIFT settings. It is evident that, while there is still improvement in position estimation compared to standard BA, the obtained results are by far inferior to those reported in Figure 4.2a and Figure 4.3a.

Moreover, using only long term feature scales is no longer possible because of scale outliers, i.e. correctly matched features in terms of image coordinates might have inconsistent scales. In particular, while using feature scales to generate additional constraints we faced the phenomenon of significant deviation of some feature scales from our model given in Eq. 3.1. Enhanced SIFT-scale resolution decreases scale outliers number significantly.



(a)



(b)

Figure 4.5: (a) image provides top view of estimated trajectory for KITTI dataset sequence using SIFT features with **non-enhanced** scale resolution. Estimation with standard BA is shown with blue, BA + all Feature Scale constraints - magenta, black solid line is ground truth. (b) image provides norm of position estimation error as a function of time using SIFT features with **non-enhanced** scale resolution.

4.3 Results with object scale constraints

We modified our framework and added an object detector at the image preprocessing step. It was necessary to extract cars bounding boxes and track objects across sequence of images. Within scale constraints we treated bounding box measurements as scale measurements and object size as virtual landmark size. For consistency we had to introduce car center position as a landmark. Scale and size initialization were performed in the same way as for landmark scale and size, and object center of mass was initialized as a triangulation result of corresponding bounding boxes centers observed from different poses.

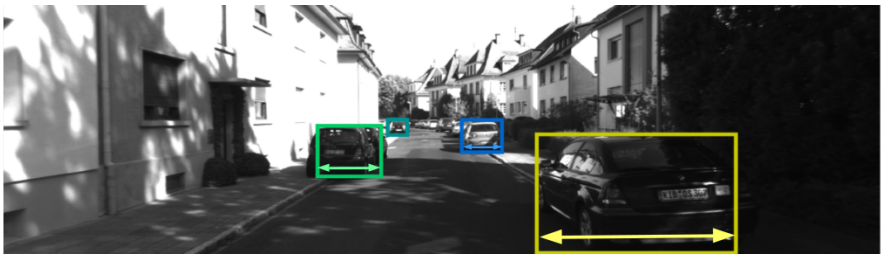


Figure 4.6: Cars tracked across sequence of images in KITTI dataset.

Finally, Figure 4.7 provides position estimation error for BA using object scale constraints, as discussed in Section 3.6, compared to BA with feature scale constraints and to standard BA. As seen, while estimation accuracy is slightly improved compared to standard BA, using feature scale constraints provides significantly better accuracy. On the other hand improving estimation using objects does not increase the optimization time significantly as number of objects is negligible comparing to number of features.

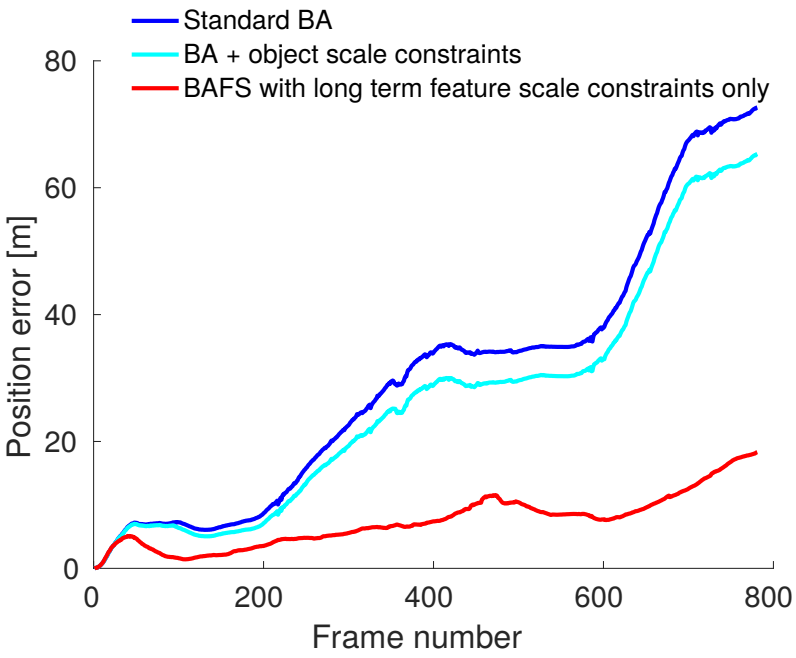


Figure 4.7: Norm of position estimation error as a function of time. Blue - standard BA, cyan - BA with object scale constraints, red - BAFS in modification with long term scale constraints only.

4.4 Flat scenario results

To test the performance of our approach in flat environments we used an outdoor Kagaru dataset [50], which provides a sequence of images from a downward-facing camera and ground truth camera position from GPS and IMU censors. Images are taken from a plane flying at constant hight and observing fields and rare trees and houses. Figure 4.8 shows typical images from the dataset.

Figure 4.8: Typical images from aerial dataset Kagaru.

The result for an aerial image sequence is shown in Figures 4.9 and 4.10, and compared to ground truth, and standard BA. As seen in Figure 4.9, standard BA suffers much less from scale drift along optical axis comparing to drift of trajectory given for KITTI dataset (e.g. Figure 4.2).

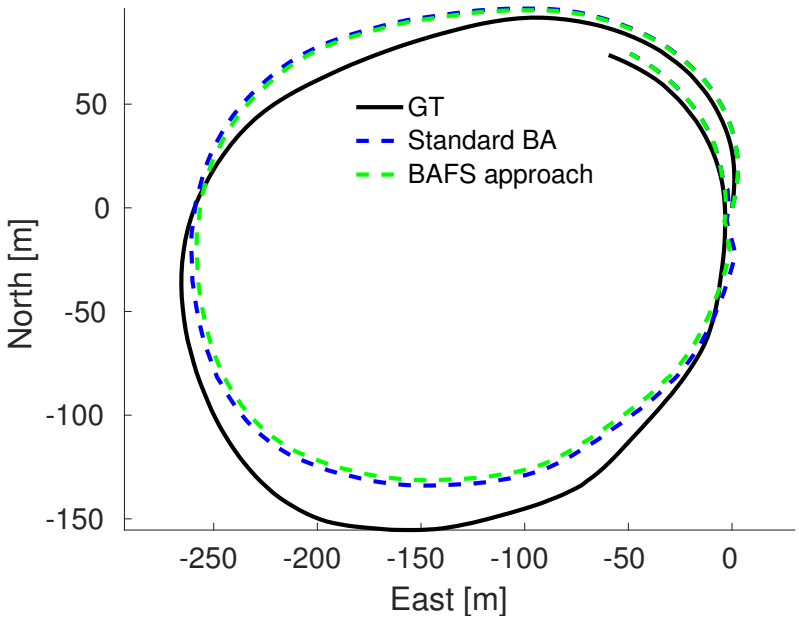


Figure 4.9: Top view of estimated trajectory for Kagaru dataset sequence. Estimation with standard BA is shown with blue, BA + Feature Scale constraints - green, black solid line is ground truth.

It is clear from Figure 4.9 that estimation is not improved at all in spite of enhanced SIFT scale resolution applied at feature extraction step. Figure 4.10 confirms this observation - as

mentioned in Section 3.5, more accurate measurements are required for flat scenarios. A naïve attempt to model scale measurements with too low Σ_{fs} value (which corresponds to the very high desired accuracy) leads to failure.

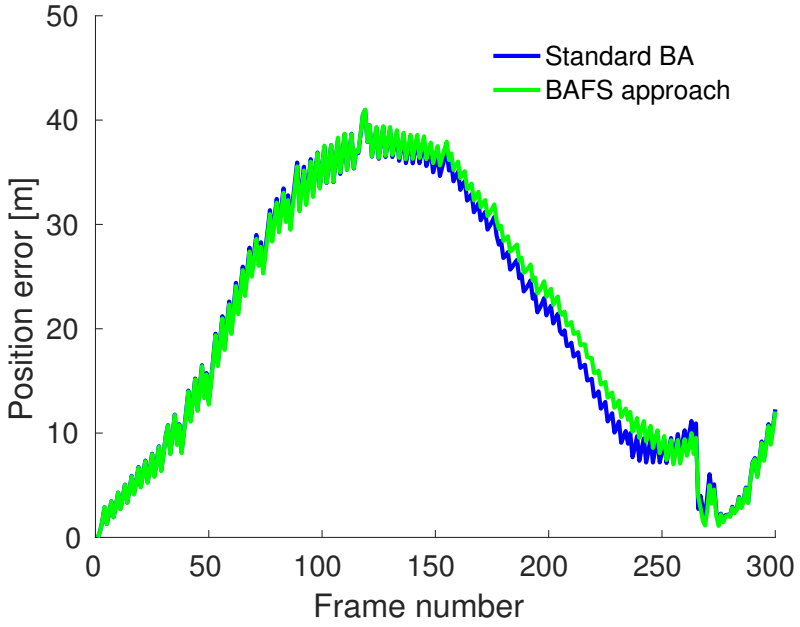


Figure 4.10: Position error for Kagaru dataset as function of time. Estimation with standard BA is shown with blue, BA + Feature Scale constraints - green.

Chapter 5

Conclusions and Future Work

We developed novel feature scale constrains and incorporated them within bundle adjustment, leveraging the scale invariance property typical feature detectors (e.g. SIFT) satisfy. Our approach does not require any additional or prior information, as it exploits already available feature scale information, which was used thus far only for image matching, and was not utilized for estimation purposes. We also proposed a method to improve feature scale accuracy by simple resolution enhancement at detection step. Using these feature scales as measurements, our approach significantly improves position estimation, especially along the optical axis in a monocular setup without requiring loop closures. Specifically, we demonstrated on KITTI datasets position estimation error can be reduced by a factor of 3, compared to standard bundle adjustment, e.g. from 90 meters to 30 meters after 950 frames. The suggested concept of exploiting scale information for improving estimation accuracy is applicable also to other scale-invariant measurements, and we demonstrated one such application, considering object-level bundle adjustment.

Future work

While in this work we focused on feature scale information, typical detectors also calculate feature orientation (local image gradient directions). Future research will investigate how the latter can be used to improve estimation accuracy even further. The idea of scale constraints might be developed further: one can utilize any scale invariant substance (objects, features) stackable across a sequence of frames.

One of the main problems is to define the measurement scale accuracy and to define if the measurement is indeed scale invariant for the case. It is important both for objects as they are not scale invariant in all cases and for features due to scale outliers. In section 3.6 we discussed the idea of using object scale constraints. This idea might be developed further. As we do not need any prior knowledge about object size, objects might be detected ”on the fly”. This means no specific pre-trained detector is needed. MSER-feature is a simplified variation of such kind of objects.

Another possible direction is related to scale outliers aspects, as mentioned in Section 4.1. We hypothesize that accounting for consistent changes in detected feature scales can lead to

enhanced robust data association.

Bibliography

[1] M. W. Achtelik, S. Lynen, S. Weiss, L. Kneip, M. Chli, and R. Siegwart. Visual-inertial slam for a small helicopter in large outdoor environments. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2651–2652, 2012.

[2] Tim Bailey, Juan Nieto, Jose Guivant, Michael Stevens, and Eduardo Nebot. Consistency of the ekf-slam algorithm. In *J. of Intelligent and Robotic Systems*, pages 3562–3568. IEEE, 2006.

[3] H. Bay, T. Tuytelaars, and L. V. Gool. Surf: speeded up robust features. In *European Conf. on Computer Vision (ECCV)*, 2006.

[4] Duane C. Brown. The bundle adjustment - progress and prospects. *Int. Archives Photogrammetry*, 21(3), 1976.

[5] M.A.R. Cooper and S. Robson. Theory of close range photogrammetry. In K.B. Atkinson, editor, *Close range photogrammetry and machine vision*, chapter 1, pages 9–51. Whittles Publishing, 1996.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 886–893, 2005.

[7] F. Dellaert. Factor graphs and GTSAM: A hands-on introduction. Technical Report GT-RIM-CP&R-2012-002, Georgia Institute of Technology, September 2012.

[8] F. Dellaert and M. Kaess. Square Root SAM: Simultaneous localization and mapping via square root information smoothing. *Intl. J. of Robotics Research*, 25(12):1181–1203, Dec 2006.

[9] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[10] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *European Conf. on Computer Vision (ECCV)*, pages 834–849. 2014.

[11] Jakob Engel, Vladyslav C. Usenko, and Daniel Cremers. A photometrically calibrated benchmark for monocular visual odometry. *CoRR*, abs/1607.02555, 2016.

[12] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. *Commun. ACM*, 24:381–395, 1981.

[13] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. Svo: Fast semi-direct monocular visual odometry. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 15–22, 2014.

[14] Duncan P Frost, Olaf Kähler, and David W Murray. Object-aware bundle adjustment for correcting monocular scale drift. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 4770–4776. IEEE, 2016.

[15] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

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

[17] S. Granshaw. Bundle adjustment methods in engineering photogrammetry. *Photogrammetric Record*, 10(56):181–207, 1980.

[18] J.-S. Gutmann and K. Konolige. Incremental mapping of large cyclic environments. In *IEEE Intl. Symp. on Computational Intelligence in Robotics and Automation (CIRA)*, pages 318–325, November 2000.

[19] Mehmet Serdar Guzel and Panus Nattharith. New technique for distance estimation using sift for mobile robots. In *Electrical Engineering Congress (iEECON), 2014 International*, 2014.

[20] Strasdat H., Montiel J. M. M., and Davison A. J. Scale drift-aware large scale monocular SLAM. In *Robotics: Science and Systems (RSS)*, Zaragoza, Spain, June 2010.

[21] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[22] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

[23] J. A. Hesch, D. G. Kottas, S. L. Bowman, and S. I. Roumeliotis. Camera-imu-based localization: Observability analysis and consistency improvement. *Intl. J. of Robotics Research*, 2013.

[24] V. Ila, J. M. Porta, and J. Andrade-Cetto. Information-based compact Pose SLAM. *IEEE Trans. Robotics*, 26(1):78–93, 2010. In press.

[25] V. Indelman. Bundle adjustment without iterative structure estimation and its application to navigation. In *IEEE/ION Position Location and Navigation System (PLANS) Conference*, April 2012.

[26] V. Indelman and F. Dellaert. Incremental light bundle adjustment: Probabilistic analysis and application to robotic navigation. In *New Development in Robot Vision*, volume 23 of *Cognitive Systems Monographs*, pages 111–136. Springer Berlin Heidelberg, 2015.

[27] V. Indelman, A. Melim, and F. Dellaert. Incremental light bundle adjustment for robotics navigation. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, November 2013.

[28] V. Indelman, R. Roberts, C. Beall, and F. Dellaert. Incremental light bundle adjustment. In *British Machine Vision Conf. (BMVC)*, September 2012.

[29] V. Indelman, R. Roberts, and F. Dellaert. Incremental light bundle adjustment for structure from motion and robotics. *Robotics and Autonomous Systems*, 70:63–82, 2015.

[30] V. Indelman, S. Wiliams, M. Kaess, and F. Dellaert. Information fusion in navigation systems via factor graph based incremental smoothing. *Robotics and Autonomous Systems*, 61(8):721–738, August 2013.

[31] H. Johannsson, M. Kaess, M. Fallon, and J. J. Leonard. Temporally scalable visual slam using a reduced pose graph. In *2013 IEEE International Conference on Robotics and Automation*, 2013.

[32] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. Leonard, and F. Dellaert. iSAM2: Incremental smoothing and mapping using the Bayes tree. *Intl. J. of Robotics Research*, 31:217–236, Feb 2012.

[33] M. Kaess, A. Ranganathan, and F. Dellaert. iSAM: Incremental smoothing and mapping. *IEEE Trans. Robotics*, 24(6):1365–1378, Dec 2008.

[34] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: Convolutional networks for real-time 6-dof camera relocalization. In *Intl. Conf. on Computer Vision (ICCV)*, 2015.

[35] K. Konolige. Large-scale map-making. In *Proc.* 21^{th} *AAAI National Conference on AI*, San Jose, CA, 2004.

[36] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard. g2o: A general framework for graph optimization. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, Shanghai, China, May 2011.

[37] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale. Keyframe-based visual-inertial odometry using nonlinear optimization. *Intl. J. of Robotics Research*, 2014.

[38] D.G. Lowe. Object recognition from local scale-invariant features. In *Intl. Conf. on Computer Vision (ICCV)*, pages 1150–1157, 1999.

[39] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *Intl. J. of Computer Vision*, 60(2):91–110, 2004.

[40] F. Lu and E. Milios. Globally consistent range scan alignment for environment mapping. *Autonomous Robots*, pages 333–349, Apr 1997.

[41] T. Lupton and S. Sukkarieh. Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions. *IEEE Trans. Robotics*, 28(1):61–76, Feb 2012.

[42] R.A. Newcombe, S.J. Lovegrove, and A.J. Davison. DTAM: Dense tracking and mapping in real-time. In *Intl. Conf. on Computer Vision (ICCV)*, pages 2320–2327, Barcelona, Spain, Nov 2011.

[43] J. Nikolic, J. Rehder, M. Burri, P. Gohl, S. Leutenegger, P. T. Furgale, and R. Siegwart. A synchronized visual-inertial sensor system with fpga pre-processing for accurate real-time slam. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 431–437, 2014.

[44] Davide Scaramuzza, Friedrich Fraundorfer, Marc Pollefeys, and Roland Siegwart. Absolute scale in structure from motion from a single vehicle mounted camera by exploiting nonholonomic constraints. In *Intl. Conf. on Computer Vision (ICCV)*, pages 1413–1419, 2009.

[45] R. Smith, M. Self, and P. Cheeseman. A stochastic map for uncertain spatial relationships. In *Proc. of the Intl. Symp. of Robotics Research (ISRR)*, pages 467–474, 1988.

[46] S. Song, M. Chandraker, and C. C. Guest. Parallel, real-time monocular visual odometry. In *2013 IEEE International Conference on Robotics and Automation*, pages 4698–4705, May 2013.

[47] D.-N. Ta, K. Ok, and F. Dellaert. Vistas and parallel tracking and mapping with wall–floor features: Enabling autonomous flight in man-made environments. *Robotics and Autonomous Systems*, 62(11):1657–1667, 2014.

[48] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. The MIT press, Cambridge, MA, 2005.

[49] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment – a modern synthesis. In W. Triggs, A. Zisserman, and R. Szeliski, editors, *Vision Algorithms: Theory and Practice*, LNCS, pages 298–375. Springer Verlag, Sep 1999.

[50] Michael Warren, David McKinnon, Hu He, Arren Glover, Michael Shiel, and Ben Upcroft. Large Scale Monocular Vision-only Mapping from a Fixed-Wing sUAS. In *International Conference on Field and Service Robotics*, Matsushima, Japan, 2012.

[51] S. Williams, V. Indelman, M. Kaess, R. Roberts, J. Leonard, and F. Dellaert. Concurrent filtering and smoothing: A parallel architecture for real-time navigation and full smoothing. *Intl. J. of Robotics Research*, 33:1544–1568, 2014.

התסריט המישורי הנצפה מגובה קבוע לערך.

יתר על כן, על מנת להמחיש את גמישות אילוצי קנה־המידה החדישים שלנו אנו מריצים את מסד הנתונים KITTI תוך שימוש באילוצי קנה המידה חלופיים של אובייקט ומשפרים את דיוק החישובים.

בקנה מידה גדול בשל סיבוכיות גבוהה. חלק מהשיטות המודרניות משלבות מידע מוקדם, כגון גובה מצלמה קבוע, ומשיגות תוצאות מדויקות. יחד עם זאת, פתרון שכזה מוגבל למערכות קרקעיות בגובה קבוע מעל פני השטח. עבור המקרה הכללי)ללא מידע מוקדם(, קיימת גישה מודרנית שמבצעת אופטימיזציה על קנה המידה לתיקון סחיפת קנה מידה. אולם גישה זו משפרת ביצועים רק במקרה של סגירות מעגל (LOOP-CLOSURE) , כלומר כאשר המצלמה חוזרת להסתכל על אזור שכבר נצפה בעבר.

הרעיון להשתמש בקנה המידה של נקודות עניין הוצע בעבר, אך בהקשרים שונים. לדוגמה, אחד החוקרים משתמש בקנה מידה של נקודות עניין בכדי לקבוע אם נקודת ציון הינה רחוקה מספיק לטובת עדכוני סיבוב בניווט מקורה, בעוד בשיטה אחרת הוצע לאחרונה להשתמש בקנה מידה SIFT של נקודות עניין לצורך הערכת המרחק. עם זאת, למיטב ידיעתנו, שילוב אילוצים על קנה מידה בתוך BA הוא חדש. בנוסף לשיפור הדיוק, לשיטה שלנו, הנקראת, BUNDLE ADJUSTMENT WITH FEATURE SCALE CONSTRAINT (BAFS), יש גם את היכולת להעריך את הגודל של נקודת ציון או אובייקט בפועל, עד כדי קנה המידה הכולל של הבעיה.

בעבודה זו אנו מציעים לשלב בתוך BA סוג חדש של אילוצים המשתמשים במידע על קנה המידה של נקודות העניין, מידע זה זמין מכל גלאי תכונות ויזואליות אופייני (למשל SIFT, SURF). למרות שלקנה המידה של נקודות עניין תפקיד חשוב בהתאמת תמונות, למיטב ידיעתנו תכונה זו לא נוצלה עד כה למטרות שערוך במסגרת BA. תפיסה זו מנצלת את אחת התכונות הבסיסיות והחשובות של גלאי SIFT (ודומיו) - אינווריאנטיות לקנה המידה (SCALE INVARIANCE). התפיסה מבוססת על האבחנה המרכזית שקנה המידה של נקודת עניין משתנה באופן עקבי על פני רצף של תמונות. פרט, אנו מראים שקנה המידה של נקודת עניין ניתן לחיזוי כפונקציה של מיקום מצלמה, קואורדינטות של נקודת הציון ואזור)טלאי(תלת ממדי התואם לקנה המידה שמסביב לנקודת הציון. על פי עקרון האינווריאנטיות של קנה המידה, הטלאי שמסביב לנקודת הציון נשאר זהה עבור תמונות שונות של אותה נקודת ציון.

בתצורת מצלמה בודדת, השימוש בקנה המידה של נקודות העניין כאמצעי לשיפור הדיוק של BA התגלה כמניב את מירב השיפור לאורך הציר האופטי. השימוש הנזכר לעיל נעשה באמצעות יצירת אילוצים בין קנה המידה הצפוי, שתלוי במרחק בין המצלמה לנקודת העניין, לקנה המידה הנמדד. לאחר מכן מתבצעת אופטימיזציה על משתני המערכת במטרה למזער את השגיאה השיורית של האילוצים הללו בנוסף למזעור שגיאת ההטלה הסטנדרטית. בהנחה שקנה המידה של נקודות העניין ניתן בדיוק מספק, שילוב אילוצי קנה המידה בתוך BA מאפשר לצמצם משמעותית את סחיפת קנה המידה ללא צורך בסגירת לולאה או כל מידע אחר. אנו מראים שניתן להשיג קנה מידה בדיוק מספק על ידי הגדלת הרזולוציה של הגרעינים הגאוסיינים בתוך גלאי ה- SIFT. גישתנו נבחנת הן בסביבות מלאכותיות והן באמצעות מסדי נתונים קרקעיים ואוויריים מהעולם האמיתי (KITTIו-KAGARU), ומדגימים שיפור משמעותי בהפחתת הסחיפה לאורך הציר האופטי.

בעזרת סימולציה אנו מדגימים כיצד הדיוק במדידות קנה המידה משפיע על החישוב, ובאמצעות מגוון תסריטים אנו מגדירים את מגבלות גישתנו. אנו מציגים, בפרט, כי עבור תסריט הסביבה המישורית, גישתנו דורשת מדידה מדויקת הרבה יותר של קנה המידה. ע"י הרצה של גישתנו על מסדי נתונים של תמונות אמיתיות אנו מאמתים את תוצאות ההדמיה על ידי שיפור משמעותי של מסד הנתונים הקרקעיים - KITTI, אך לא מניבים תוצאות עבור מסד נתונים אוויריים KAGARU אשר תואם את

תקציר

בעבודה זו אנו מציגים גישה לשיפור דיוק (BA) Bundle Adjustment. רצף של תמונות שצולמו ממצלמה נעה המתקבל כקלט, ו־ BA משחזר את מסלול המצלמה ואת הסביבה התלת ממדית. שערוך של מסלול ושחזור תלת־ממדי של סביבה חשובים במגוון יישומים כגון ניווט או מיפוי אוטונומיים בסביבות לא ידועות. שיטות של BA ושיטות של לוקליזציה ומיפוי סימולטניים (SLAM) בדרך כלל משמשים לטיפול בבעיות אלה ובעיות דומות.

הפתרון של BA בדרך כלל כרוך במזעור שגיאת ההטלה בין תצפיות תמונה (נקודות העניין ־ features, הנמצאות במישור התמונה) לבין החיזוי המתקבל על ידי הטלת נקודות הציון (landmarks, נמצאות בעולם התלת ממדי) חזרה למערכת המצלמה. מזעור זה מתקבל בדרך כלל באמצעות טכניקות אופטימיזציה איטרטיביות ולא ליניארית. בהינתן תנאי התחלה מספקים, שיטות אלו מתכנסות לפתרון מקסימום אפוסטריורי (MAP) של מיקומי המצלמה ונקודות הציון המייצגות את הסביבה הנצפית. חשוב לציין שכל נקודת עניין בתמונה מבוטאת באמצעות מיקום, קנה מידה, כיוון וקידוד ייחודי. בעוד מידע זה משמש להתאמה בין נקודות עניין על פני תמונות שונות, גישות קיימות של BA ו Visual SLAM , משתמשות רק במידע לגבי מיקום נקודת העניין בתמונה לצורך ניסוח אילוצי הטלה.

שיטות Visual SLAM טיפוסיות מניחות קלט ממצלמה בודדת (monocular camera) או ממצלמת סטראו (stereo camera) , כאשר אחת הסיבות העיקריות לשימוש וחקר SLAM monocular טהור היא הפשטות של המכשור הנדרש ליישומו. אף על פי כן, החיסרון הוא בכך שהאלגוריתמים הנדרשים עבור SLAM monocular מורכבים בהרבה יותר. וזאת מפני שהעומק אינו יכול להיות מוסק ישירות מתמונה בודדת וממצלמה בודדת.

בעיה נוספת מוכרת בגישות SLAM של מצלמה בודדת היא סחיפת קנה המידה. ללא ידיעת הטרנספורמציה בין המצלמות של מערכת סטריאו המשמש כעוגן, קנה המידה של רצף התצלומים המקומי והחישובים של התנועה התואמת עשויים להסחף עם הזמן. הסחיפה נאגרת לאורך הזמן ובדרך כלל היא המקור העיקרי של טעויות במערכות SLAM עם מצלמה בודדת. התופעה קיימת גם עבור מצלמות סטראו, אם כי במידה מועטה יותר.

בשנים האחרונות פותחו ניסוחים חלופיים הכוללים, בין היתר, גישות BA ללא שערוך סביבה, כגון Light Bundle Adjustment (LBA). גישות אלו מעלימות בצורה אלגברית את נקודות הציון וממזערות את השגיאה השיורית של אילוצי גיאומטריה רב־תמונתית (Multiple View Geometry). לעומת זאת, גישות BA המשלבות שערוך של נקודות ציון, כגון DTAM ו SVO , ממזערות את השגיאות הפוטוגרמטריות עבור תמונות חופפות. גישות אלו מספקות דיוק גבוה יותר, אך אינן מסוגלות לבצע אופטימיזציה

המחקר נעשה בהנחיית פרופסור ואדים אינדלמן בפקולטה להנדסת אוירונוטיקה וחלל

תודות

אני מודה לטכניון על התמיכה הכספית הנדיבה בהשתלמותי.

אופטימיזצייה מבוססת תמונות עם סקלה של נקודות עניין לצורך שיפור דיוק שערוך

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר
מגיסטר למדעים בתוכנית מערכות אוטונומיות

ולדימיר אבצ'קין

הוגש לסנט הטכניון — מכון טכנולוגי לישראל
שבט תשע"ח חיפה פברואר 2018

אופטימיזצייה מבוססת תמונות עם סקלה של נקודות עניין לצורך שיפור דיוק שערוך

ולדימיר אבצ'קין