BAFS: Bundle Adjustment With Feature Scale Constraints for Enhanced Estimation Accuracy

Vladimir Ovechkin and Vadim Indelman

***Abstract*—We propose to incorporate within bundle adjustment (BA) a new type of constraint that uses feature scale information, leveraging the scale invariance property of typical image feature detectors (e.g., SIFT). While feature scales play an important role in image matching, they have not been utilized thus far for estimation purposes in a BA framework. Our approach exploits the already-available feature scale information and uses it to enhance the accuracy of BA, especially along the optical axis of the camera in a monocular setup. Importantly, the mentioned feature scale constraints can be formulated on a frame to frame basis and do not require loop closures. We study our approach in synthetic environments and the real-imagery KITTI dataset, demonstrating significant improvement in positioning error.**

***Index Terms*—Localization, mapping, SLAM.**

I. Introduction

ACCURATE pose estimation and structure reconstruction are important in a variety of applications, including vision aided navigation (VAN) [10], simultaneous localization and mapping (SLAM) [8], [16], visual odometry (VO) [4], augmented reality, structure from motion (SfM), tracking and robotic surgery. 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 [22] for an extensive review of different aspects in BA.

Standard BA approaches typically assume a pinhole camera model [9] and minimize re-projection errors between measured and predicted image coordinates. This minimization is typically obtained using iterative nonlinear optimization techniques that, provided a proper initial guess, converge to the maximum a posteriori (MAP) solution over camera poses and landmarks that represent the observed environment. Alternative formulations have been also developed in recent years. These include, for example, structureless BA approaches, such as Light Bundle Adjustment (LBA) [11]–[15] that algebraically eliminate the 3D points and minimize the residual error in multiple view geometry constraints. In contrast, dense BA approaches, such as DTAM [19] and SVO [4], minimize the photogrammetric errors for each overlapping image.

Manuscript received September 7, 2017; accepted December 5, 2017. Date of publication January 11, 2018; date of current version January 25, 2018. This letter was recommended for publication by Associate Editor U. Frese and Editor C. Stachniss upon evaluation of the reviewers' comments. *(Corresponding author: Vadim Indelman.)*

V. Ovechkin is with the Technion Autonomous Systems Program, Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: vladimir.o@campus.technion.ac.il).

V. Indelman is with the Department of Aerospace Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: vadim.indelman@technion.ac.il).

Digital Object Identifier 10.1109/LRA.2018.2792141

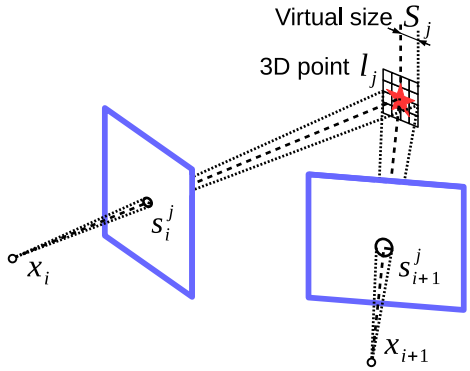

Fig. 1. Feature scale is modeled as a projection of a virtual landmark size in 3D environment onto the image plane. We leverage the scale invariance property of typical feature detectors, according to which, detected scales of matched features from different images correspond to the same virtual landmark size in the 3D environment, and incorporate novel feature scale constraints within BA.

In cases where sources of absolute information such as GPS or an a priori map are unavailable, maintaining high-accuracy estimation over time is a challenging task. 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. Existing approaches address this issue by explicitly correcting scale drift at loop closures (e.g., [8]), exploiting nonholonomic motion constraints (e.g., [20]), or fusing information from additional sensors (such as IMU). Frost *et al.* [5] 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 this letter 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 [18] (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 Figs. 1 and 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

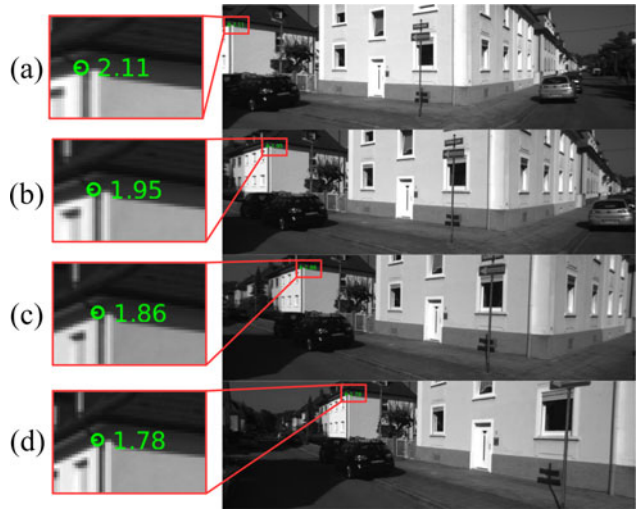

Fig. 2. A landmark is observed while the camera performs a left turn, from (a) to (d). The detected feature scale in each frame is shown in the zoom-in figures.

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. Here, we propose to exploit this available information for improving the performance of BA. Note we do not interfere with the 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.* [21] use feature scale to determine if a landmark is sufficiently far away to consider it for rotation updates in indoor navigation, while Guzel *et al.* [7] recently suggested to use SIFT's 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.

II. Notations and Problem Formulation

We consider a sequence of N images captured from different and unknown camera poses. 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,

$$X \doteq \{x_1, \ldots, x_i, \ldots, x_N\} \quad , \quad L \doteq \{l_1, \ldots, l_j, \ldots, l_M\}, \tag{1}$$

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 [9], and write the measurement likelihood for an image observation z given camera pose x and landmark l as

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

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

$$\mathbb{P}(X, L|\mathcal{Z}) \propto \text{priors} \cdot \prod_{i=1}^{N} \prod_{j \in \mathcal{M}_i} \mathbb{P}(z_i^j | x_i, l_j) \tag{3}$$

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

$$X^\star, L^\star = \arg\max_{X,L} \mathbb{P}(X, L|\mathcal{Z}), \tag{4}$$

and can be calculated using state of the art computationally efficient solvers [2], [17] that solve the following non-linear least-squares problem:

$$J_{BA}(X, L) \doteq \sum_{i}^{N} \sum_{j \in \mathcal{M}_i} \left\| z_i^j - \pi(x_i, l_j) \right\|^2_{\Sigma_v}. \tag{5}$$

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. 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 section we formulate a new type of a constraint that allows enhanced estimation accuracy, particularly along the camera optical axis, without requiring loop closures or additional prior knowledge.

III. Approach

A. Feature Scale Constraint Formulation

The standard bundle adjustment formulation exploits only a subset of the 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 also accompanied by two additional parameters - scale and orientation. We propose to incorporate this 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 Fig. 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 Fig. 1. This observation leverages the scale invariance property that typical feature

detectors (e.g., SIFT [18]) satisfy. As an example, we consider the image sequence shown in Fig. 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

$$s_i^j = f\frac{S_j}{d_i^j} + v_i, \tag{6}$$

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., $v_i \sim N(0, \Sigma_{fs})$. In (6) we use d_i^j to denote the distance along the optical axis from the camera pose x_i to landmark l_j. In other words, assuming the optical axis is the z axis in the camera frame,

$$d_i^j(x_i, l_j) \doteq z_c, \quad \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = R_i l_j + t_i, \tag{7}$$

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 Fig. 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. Fig. 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 (6) 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.

Based on the observation model (6) we can now define the corresponding feature scale measurement likelihood as

$$\mathbb{P}(s_i^j | S_j, x_i, l_j) \doteq \frac{1}{\sqrt{|2\pi\Sigma_{fs}|}} \exp\left[-\frac{1}{2}\left\| s_i^j - f\frac{S_j}{d_i^j} \right\|_{\Sigma_{fs}}^2\right]. \tag{8}$$

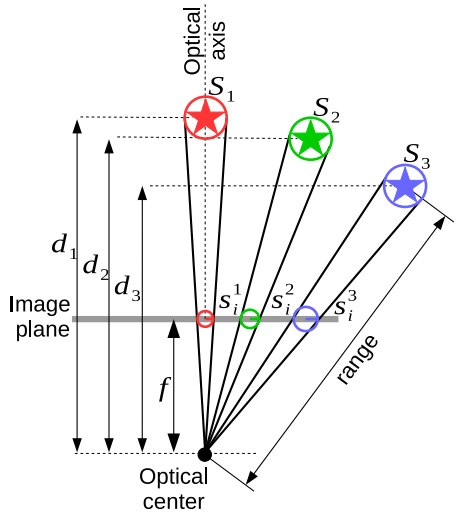

Fig. 3. Landmark of the same virtual size S_j is observed at a constant range from the camera's optical center, producing different scale projections depending on the distance along optical axis.

As seen, the above likelihood is conditioned on the virtual landmark size S_j. Since the latter is 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 (8) yields the following joint pdf (omitting the priors terms)

$$\mathbb{P}(X, L, S|\mathcal{Z}) \propto \prod_i^N \prod_{j \in \mathcal{M}_i} \mathbb{P}\left(z_i^j | x_i, l_j\right) \mathbb{P}\left(s_i^j | S_j, x_i, l_j\right). \tag{9}$$

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

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

$$J_{BAFS}(X, L, S) \doteq + \sum_i^N \sum_{j \in \mathcal{M}_i} \left\| z_i^j - \pi(x_i, l_j) \right\|_{\Sigma_v}^2 + \left\| s_i^j - f\frac{S_j}{d_i^j} \right\|_{\Sigma_{fs}}^2, \tag{10}$$

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

B. Computational Complexity and Factor Graph Reduction

The obtained joint pdf can be conventionally represented with a factor graph model [3]. 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 Fig. 4(b). 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.

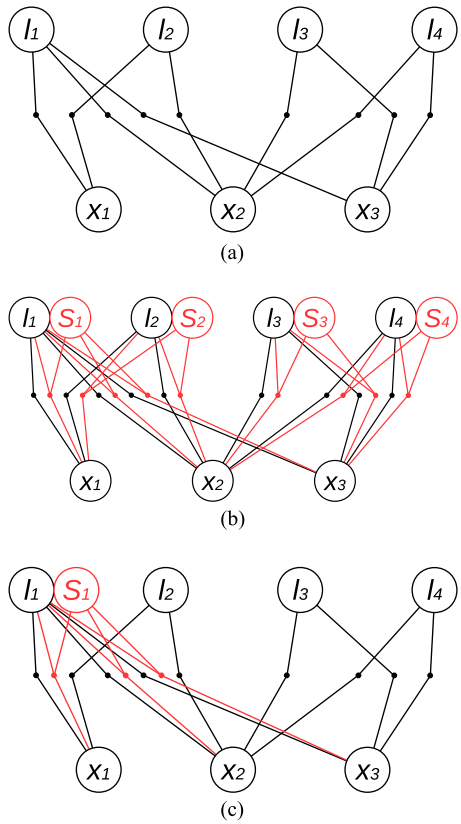

Fig. 4. Factor graph representations: (a) standard BA with projection factors only; (b) BAFS with naïvely added all feature scale factors; (c) BAFS with feature scale factors added only for long-term landmarks (l_1, in this case).

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. Fig. 4(c) illustrates a factor graph that corresponds to this heuristic.

C. 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 [9], the following method can be used to initialize the virtual landmark 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

$$S_j = s_i^j \frac{d_i^j}{f}, \tag{11}$$

which is obtained from (6) while neglecting the noise.

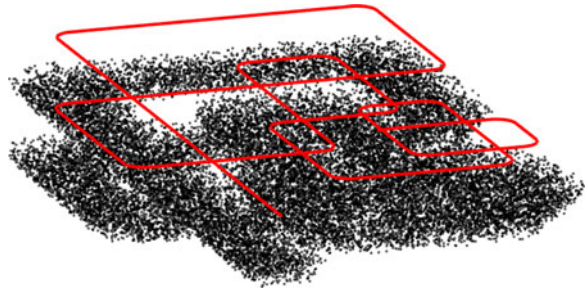

Fig. 5. Simulated scenario of an aerial downward-facing camera observing randomly-scatted landmarks. Camera's trajectory is shown in red.

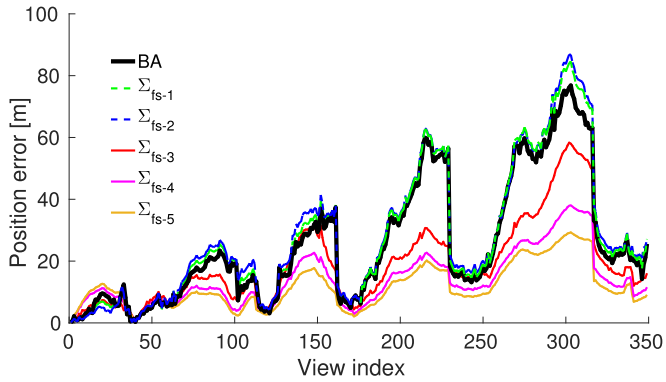

Fig. 6. Position estimation error. Each curve corresponds to BA with feature scale constraints with noise in simulated feature scale measurements sampled from a Gaussian with different Σ_{fs}. Black solid curve corresponds to standard BA.

In our implementation, we initialize each new landmark via triangulation given two landmark observations, and initialize S_j by taking an average value of (11) considering the corresponding two detected feature scales.

D. Enhancement of Feature Scale Measurement Accuracy

Thus far, we incorporated our novel feature scale constraints (10) 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 (8).

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 Fig. 5). 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}.

As expected, running on this synthetic data showed that accuracy of feature scales is extremely important to improve standard BA precision. Fig. 6 shows results, in terms of position estimation error, for different simulated values of Σ_{fs}

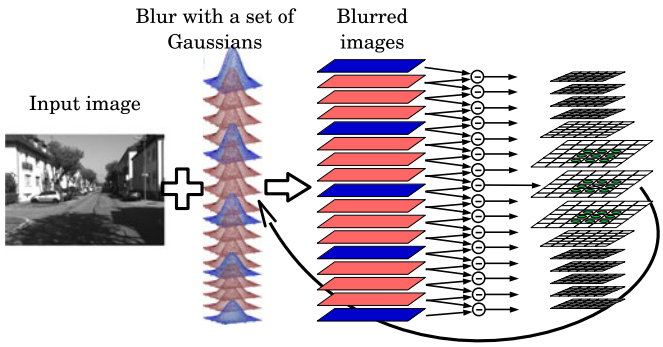

Fig. 7. SIFT scale estimation process. (a) Blur each input image with a set of Gaussian kernels. (b) Calculate Difference of Gaussians (DoG). (c) Feature scale is set via interpolation or as the average of the two Gaussian kernels that correspond to the local-maxima DoG layer.

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 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 Fig. 7). 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 Fig. 7, 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.

E. 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 Fig. 8).

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 (6). Close objects are not taken into account as the corresponding detected bounding boxes might be obtained from significantly

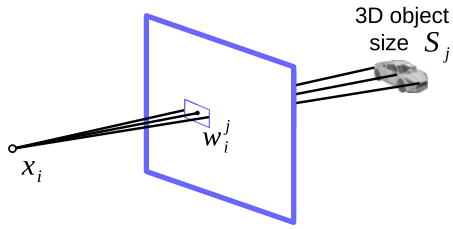

Fig. 8. Object detected bounding box in the image plane.

different viewpoints and provide inconsistent scale measurements. In our implementation, we use HoG object detector [1] to identify bounding boxes, and formulate the scale constraint considering the detected width and height instead of feature scale. For example, for the jth object observed at the ith frame, the scale constraint is

$$w_i^j = f\frac{S_j^{\text{obj}}}{d_i^j} + v_i^{\text{obj}}, \tag{12}$$

where w_i^j is the detected bounding box width (see Fig. 8), and v_i^{obj} is a Gaussian noise that corresponds to the accuracy in 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.

IV. Experimental Results

We implemented a classical sparse feature based BA framework using the GTSAM [2] 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 (8). As described in Section III-D, we enhance standard SIFT feature scale resolution by increasing the number of layers per octave from default value 3 up to 15, which, however, consumed about 2.5 times more runtime of SIFT feature extraction. 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} down 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 [6]. 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.

In our experiments we evaluate the performance of the developed method to reduce scale drift across a sequence of frames. We initialize the *global* scale with ground truth range between the first two camera frames, although odometry information could also be used. It makes it easy to evaluate estimation performance versus time compared to ground truth, though actual solution remains up to scale, as in any other monocular SLAM approach.

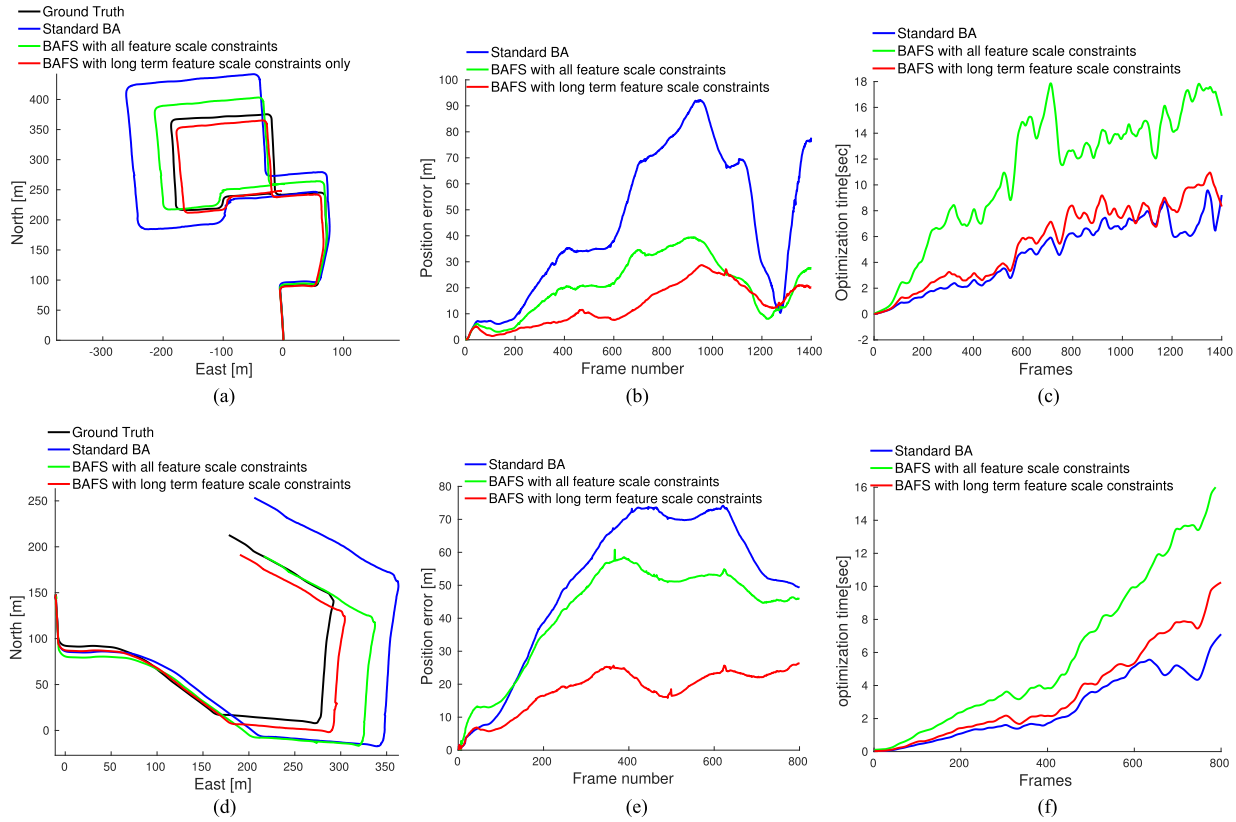

Fig. 9. Each row describes results for a different KITTI dataset sequence using SIFT features with enhanced scale resolution. (a) and (d) top view of estimated trajectory; (b) and (e) norm of position estimation error as a function of time; (c) and (f) optimization time for each frame.

The results for both of the considered sequences are shown in Fig. 9, 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 III-B). Specifically, Fig. 9 shows the estimated trajectories (top view), position estimation errors, and 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 [16] but leave this endeavor to future research.

As seen in Fig. 9(a) and (d), 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 Fig. 9(b) and (e)] 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 to 40 meters around frame 950.

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 Fig. 9(e)]. This is perhaps a 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.

Fig. 9(c) and (f) 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.

The above results were obtained using enhanced-resolution feature scales (see Section III-D). To demonstrate the importance of improving the accuracy of detected feature scales, we show in Fig. 10(a) and (b) 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 Fig. 9.

Finally, Fig. 10(c) provides position estimation error for BA using object scale constraints, as discussed in Section III-E, 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.

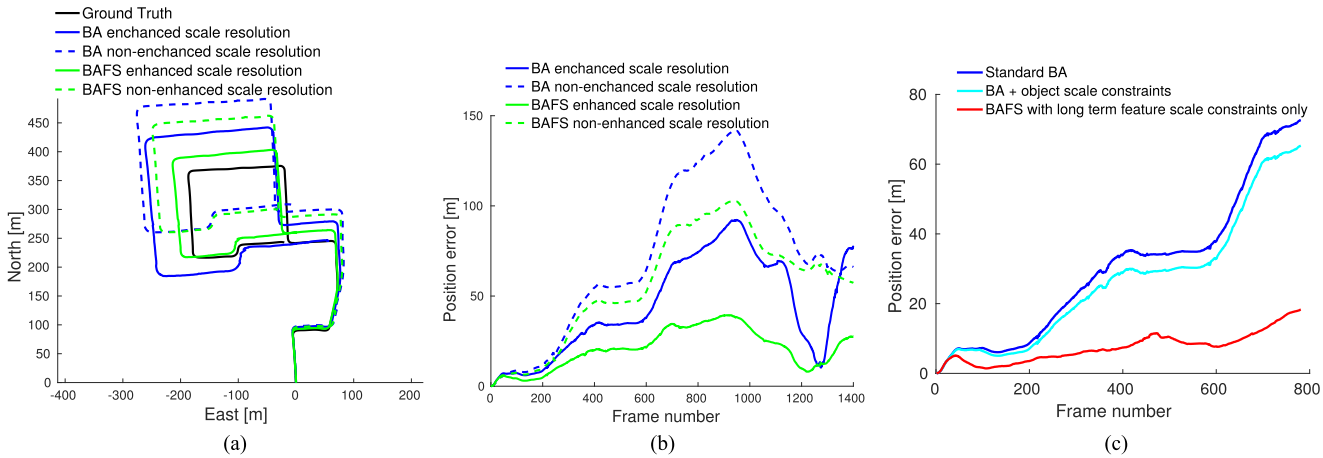

Fig. 10. (a), (b) Results with non-enhanced and enhanced scale resolution: (a) top view of estimated trajectory; (b) norm of position estimation error. (c) norm of position estimation error for BA with object scale constraints, compared with standard BA, and BAFS with long term features using enhanced feature scale resolution.

V. Conclusion

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. 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.

References

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[2] F. Dellaert, "Factor graphs and GTSAM: A hands-on introduction," Georgia Inst. Technol., Atlanta, GA, USA, Tech. Report GT-RIM-CP&R-2012–002, Sep. 2012.

[3] F. Dellaert and M. Kaess, "Square Root SAM: Simultaneous localization and mapping via square root information smoothing," *Int. J. Robot. Res.*, vol. 25, no. 12, pp. 1181–1203, Dec. 2006.

[4] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2014, pp. 15–22.

[5] D. P. Frost, O. Kähler, and D. W. Murray, "Object-aware bundle adjustment for correcting monocular scale drift," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2016, pp. 4770–4776.

[6] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, pp. 1231–1237, 2013.

[7] M. S. Guzel and P. Nattharith, "New technique for distance estimation using sift for mobile robots," in *Proc. Int. Elect. Eng. Congr.*, 2014, pp. 1–4.

[8] H. Strasdat, J. M. M. Montiel, and A. J. Davison, "Scale drift-aware large scale monocular SLAM," in *Proc. Robot.: Sci. Syst.*, Zaragoza, Spain, Jun. 2010.

[9] R. I. Hartley and A. Zisserman," *Multiple View Geometry in Computer Vision*. 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[10] J. A. Hesch, D. G. Kottas, S. L. Bowman, and S. I. Roumeliotis, "Camera-IMU-based localization: Observability analysis and consistency improvement," *Int. J. Robot. Res.*, vol. 33, pp. 182–201, 2014.

[11] V. Indelman, "Bundle adjustment without iterative structure estimation and its application to navigation," in *Proc. IEEE/ION Position Locat. Navig. Syst. Conf.*, Apr. 2012, pp. 748–756.

[12] V. Indelman and F. Dellaert, "Incremental light bundle adjustment: Probabilistic analysis and application to robotic navigation," in *New Development in Robot Vision*, vol. 23, Cognitive Systems Monographs, Berlin, Germany: Springer, 2015, pp. 111–136.

[13] V. Indelman, A. Melim, and F. Dellaert, "Incremental light bundle adjustment for robotics navigation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 748–756.

[14] V. Indelman, R. Roberts, C. Beall, and F. Dellaert, "Incremental light bundle adjustment," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2012.

[15] V. Indelman, R. Roberts, and F. Dellaert, "Incremental light bundle adjustment for structure from motion and robotics," *Robot. Auton. Syst.*, vol. 70, pp. 63–82, 2015.

[16] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. Leonard, and F. Dellaert, "iSAM2: Incremental smoothing and mapping using the Bayes tree," *Int. J. Robot. Res.*, vol. 31, pp. 217–236, Feb. 2012.

[17] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "g2o: A general framework for graph optimization," in *Proc. IEEE Int. Conf. Robot. Autom.*, Shanghai, China, May 2011, pp. 3607–3613.

[18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[19] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," in *Proc. Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 2320–2327.

[20] D. Scaramuzza, F. Fraundorfer, M. Pollefeys, and R. Siegwart, "Absolute scale in structure from motion from a single vehicle mounted camera by exploiting nonholonomic constraints," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 1413–1419.

[21] 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," *Robot. Auton. Syst.*, vol. 62, no. 11, pp. 1657–1667, 2014.

[22] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon, "Bundle adjustment—A modern synthesis," in *Vision Algorithms: Theory and Practice*, LNCS, W. Triggs, A. Zisserman, and R. Szeliski, Eds. New York, NY, USA: Springer, Sep. 1999, pp. 298–375.