Photometric Bundle Adjustment for Vision-Based SLAM

Hatem Alismail(✉), Brett Browning, and Simon Lucey

The Robotics Institute, Carnegie Mellon University, Pittsburgh, USA
{halismai,brettb,slucey}@cs.cmu.edu

Abstract. We propose a novel algorithm for the joint refinement of structure and motion parameters from image data directly without relying on fixed and known correspondences. In contrast to traditional bundle adjustment (BA) where the optimal parameters are determined by minimizing the reprojection error using tracked features, the proposed algorithm relies on maximizing the photometric consistency and estimates the correspondences implicitly. Since the proposed algorithm does not require correspondences, its application is not limited to corner-like structure; any pixel with nonvanishing gradient could be used in the estimation process. Furthermore, we demonstrate the feasibility of refining the motion and structure parameters simultaneously using the photometric error in unconstrained scenes and without requiring restrictive assumptions such as planarity. The proposed algorithm is evaluated on range of challenging outdoor datasets, and it is shown to improve upon the accuracy of the state-of-the-art VSLAM methods obtained using the minimization of the reprojection error using traditional BA as well as loop closure.

1 Introduction

Photometric, or image-based, minimization is a fundamental tool in a myriad of applications such as: optical flow [1], scene flow [2], and stereo [3,4]. Its use in vision-based 6DOF motion estimation has recently been explored demonstrating good results [5–8]. Minimizing the photometric error, however, has been limited to frame–frame estimation (visual odometry), or as a tool for depth refinement independent of the parameters of motion [9]. Consequently, in unstructured scenes, frame–frame minimization of the photometric error cannot reduce the accumulated drift. When loop closure and prior knowledge about the motion and structure are not available, one must resort to the Gold Standard: minimizing the reprojection error using bundle adjustment.

Bundle adjustment (BA) is the problem of jointly refining the parameters of motion and structure to improve a visual reconstruction [10]. Although BA is a versatile framework, it has become a synonym to minimizing the reprojection error across multiple views [11,12]. The advantages of minimizing the reprojection error are abundant and have been discussed at length in the literature [11,12]. In practice, however, there are sources of systematic errors

S.-H. Lai et al. (Eds.): ACCV 2016, Part IV, LNCS 10114, pp. 324–341, 2017.
DOI: 10.1007/978-3-319-54190-7_20

in feature localization that are hard to detect and the value of modeling their uncertainty remains unclear [13,14]. For example, slight inaccuracies in calibration exaggerate errors [15], sensor noise and degraded frequency content of the image affect feature localization accuracy [16]. Even interpolation artifacts play a non-negligible role [17]. Although minimizing the reprojection is backed by sound theoretical properties [11], its use in practice must also take into account the challenges and nuances of precisely localizing keypoints [10].

In this work, we propose a novel method that further improves upon the accuracy of minimizing the reprojection error and even state-of-the-art loop closure [18]. The proposed algorithm brings back the image in the loop, and jointly refines the motion and structure parameters to maximize the photometric consistency across multiple views. In addition to improved accuracy, the algorithm does not require correspondences. In fact, correspondences are estimated automatically as a byproduct of the proposed formulation. The ability to perform BA without the need for precise correspondences is attractive because it can enable VSLAM applications where corner extraction is unreliable [19], as well as additional modeling capabilities that extend beyond geometric primitives [20,21].

1.1 Preliminaries and Notation

The Reprojection Error. Given an initial estimate of the scene structure $\{\boldsymbol{\xi}_j\}_{j=1}^{N}$, the viewing parameters per camera $\{\boldsymbol{\theta}_i\}_{i=1}^{M}$, and $\mathbf{x}_{ij}$ the projection of the j^{th} point onto the i^{th} camera, the reprojection error is given by

$$\epsilon_{ij}(\mathbf{x}_{ij}; \boldsymbol{\theta}_i, \boldsymbol{\xi}_j) = \|\mathbf{x}_{ij} - \pi\left(\mathbf{T}(\boldsymbol{\theta}_i), \mathbf{X}(\boldsymbol{\xi}_j)\right)\|, \tag{1}$$

where $\pi(\cdot, \cdot)$ is the image projection function. The function $\mathbf{T}(\cdot)$ maps the vectorial representation of motion to a rigid body transformation matrix. Similarly, $\mathbf{X}(\cdot)$ maps the parameterization of the point to coordinates in the scene.

In this work, we assume known camera calibration parameters as is often the case in VSLAM and parameterize the scene structure using the usual 3D Euclidean coordinates, where $\mathbf{X}(\boldsymbol{\xi}) := \boldsymbol{\xi}$, and

$$\boldsymbol{\xi}_j^\top = \begin{pmatrix} x_j & y_j & z_j \end{pmatrix} \in \mathbb{R}^3. \tag{2}$$

The pose parameters are represented using twists [22], where the rigid body pose is obtained using the exponential map [23], *i.e.*:

$$\boldsymbol{\theta}_i^\top \in \mathbb{R}^6 \quad \text{and} \quad \mathbf{T}(\boldsymbol{\theta}) := \exp(\widehat{\boldsymbol{\theta}}) \in SE(3). \tag{3}$$

Our algorithm, similar to minimizing the reprojection error using BA, does not depend on the parameterization. Other representations for motion and structure have been studied in the literature and could be used as well [24–26].

Geometric Bundle Adjustment. Given an initialization of the scene points and motion parameters, we may obtain a refined estimate by minimizing the squared reprojection error in Eq. (1) across tracked features, *i.e.*:

$$\{\Delta\boldsymbol{\theta}_i^*, \Delta\boldsymbol{\xi}_j^*\} = \underset{\boldsymbol{\theta}_i, \boldsymbol{\xi}_j}{\operatorname{argmin}} \sum_{i=1}^{M}\sum_{j=1}^{N} \frac{1}{2}\delta_{ij}\epsilon_{ij}^2(\mathbf{x}_{ij}, \Delta\boldsymbol{\theta}_i, \Delta\boldsymbol{\xi}_j), \tag{4}$$

where $\delta_{ij} = 1$ if the j^{th} point is visible, or tracked, in the i^{th} camera. We call this formulation *geometric* BA.

Minimizing the reprojection error in Eq. (4) is a large nonlinear optimization problem. Particular to BA is the sparsity pattern of its linearized form, which can be exploited beneficially for both large– and medium–scale problems [11].

1.2 The Photometric Error

The use of photometric information in Computer Vision has a long and rich history dating back to the seminal works of Lucas and Kanade [27] and Horn and Schunk [28]. The problem is usually formulated as a pairwise alignment of two images. One is the reference $\mathbf{I}_0$, while the other is the input $\mathbf{I}_1$. The goal is to estimate the parameters of motion $\boldsymbol{p}$ such that the sum of the squared intensity error is minimized

$$\boldsymbol{p}^* = \underset{\boldsymbol{p}}{\operatorname{argmin}} \sum_{\mathbf{u}\in\Omega_0} \frac{1}{2}\|\mathbf{I}_0(\mathbf{u}) - \mathbf{I}_1(\mathbf{w}(\mathbf{u};\boldsymbol{p}))\|^2, \tag{5}$$

where $\mathbf{u} \in \Omega_0$ denotes a subset of pixel coordinates in the reference image frame, and $\mathbf{w}(\cdot,\cdot)$ denotes the warping function [29]. Minimizing the photometric error has recently resurfaced as a robust solution to visual odometry (VO) [6,7,30]. Notwithstanding, minimizing the photometric error has not yet been explored for the *joint* optimization of the motion and structure parameters for VSLAM in unstructured scenes. The proposed approach fills in the gap by providing a photometric formulation for BA, which we call BA *without* correspondences.

2 Bundle Adjustment Without Correspondences

BA is not limited to minimizing the reprojection error [10]. We reformulate the problem as follows. First, we assume an initial estimate of the camera poses $\boldsymbol{\theta}_i$ as required by geometric BA. However, we do not require tracking information for the 3D points. Instead, for every scene point $\boldsymbol{\xi}_j$, we assign a *reference* frame denoted by $r(j)$. The reference frame is used to extract a fixed square patch denoted by $\boldsymbol{\phi}_j \in \mathbb{R}^D$ over a neighborhood denoted by $\mathcal{N}$. In addition, we compute an initial *visibility* list indicating the frames where the point may be in view. The visibility list for the j^{th} point excludes the reference frame and is denoted by:

$$\mathbf{V}_j = \{k \;:\; k \neq r(j) \text{ and } \boldsymbol{\xi}_j \text{ is visible in frame } k\}, \text{ for } k \in [1,\ldots,M]. \tag{6}$$

Given this information and the input images $\{\mathbf{I}_i\}_{i=1}^{M}$, we seek to estimate an optimal update to the motion $\Delta\boldsymbol{\theta}_i{}^*$ and structure parameters $\Delta\boldsymbol{\xi}_j{}^*$ that satisfy

$$\{\Delta\boldsymbol{\theta}_i^*, \Delta\boldsymbol{\xi}_j^*\} = \underset{\Delta\boldsymbol{\theta}_i, \Delta\boldsymbol{\xi}_j}{\operatorname{argmin}} \sum_{j=1}^{N} \sum_{k \in V(j)} \mathcal{E}(\boldsymbol{\phi}_j, \mathbf{I}_k; \Delta\boldsymbol{\theta}_k, \Delta\boldsymbol{\xi}_j), \text{ where} \tag{7}$$

$$\mathcal{E}(\boldsymbol{\phi}, \mathbf{I}'; \boldsymbol{\theta}, \boldsymbol{\xi}) = \sum_{\mathbf{u} \in \mathcal{N}} \frac{1}{2} \|\boldsymbol{\phi}(\mathbf{u}) - \mathbf{I}'(\pi(\boldsymbol{\theta}, \boldsymbol{\xi}) + \mathbf{u})\|^2. \tag{8}$$

The notation $\mathbf{I}'(\pi(\cdot,\cdot)+\mathbf{u})$ indicates sampling the image intensities in a neighborhood about the current projection of the point using an appropriate interpolation scheme (bilinear in this work). The objective is illustrated in Fig. 1.

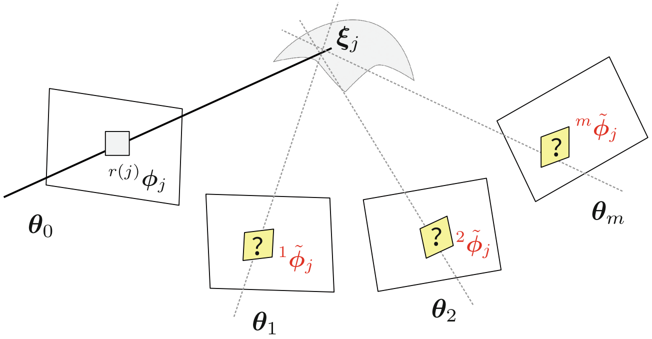
$\boldsymbol{\xi}_j$
$^{r(j)}\boldsymbol{\phi}_j$
$\boldsymbol{\theta}_0$
?
$^{1}\tilde{\boldsymbol{\phi}}_j$
$\boldsymbol{\theta}_1$
?
$^{2}\tilde{\boldsymbol{\phi}}_j$
$\boldsymbol{\theta}_2$
?
$^{m}\tilde{\boldsymbol{\phi}}_j$
$\boldsymbol{\theta}_m$

Fig. 1. Schematic of the proposed approach. We seek to optimize the parameters of motion $\boldsymbol{\theta}_i$ and structure $\boldsymbol{\xi}_j$ such that the photometric error with respect to a fixed patch at the reference frame is minimized; correspondences are estimated implicitly

Linearization and Sparsity. The optimization problem in Eq. (7) is nonlinear and its solution proceeds with standard techniques. Let $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ denote the current estimate of the camera and the scene point, and let the current projected pixel coordinate in the image plane be given by

$$\mathbf{u}' = \pi(\mathbf{T}(\boldsymbol{\theta}), \mathbf{X}(\boldsymbol{\xi})), \tag{9}$$

then taking the partial derivatives of the 1st-order expansion of the photometric error in Eq. (8) with respect to the motion and structure parameters we obtain:

$$\frac{\partial \mathcal{E}}{\partial \boldsymbol{\theta}} = \sum_{\mathbf{u} \in \mathcal{N}} \mathbf{J}^\top(\boldsymbol{\theta}) \left| \boldsymbol{\phi}(\mathbf{u}) - \mathbf{I}'(\mathbf{u}' + \mathbf{u}) - \mathbf{J}(\boldsymbol{\theta}) \Delta\boldsymbol{\theta} \right| \tag{10}$$

$$\frac{\partial \mathcal{E}}{\partial \boldsymbol{\xi}} = \sum_{\mathbf{u} \in \mathcal{N}} \mathbf{J}^\top(\boldsymbol{\xi}) \left| \boldsymbol{\phi}(\mathbf{u}) - \mathbf{I}'(\mathbf{u}' + \mathbf{u}) - \mathbf{J}(\boldsymbol{\xi}) \Delta\boldsymbol{\xi} \right|, \tag{11}$$

where $\mathbf{J}(\boldsymbol{\theta}) = \nabla\mathbf{I}(\mathbf{u}' + \mathbf{u})\frac{\partial\mathbf{u}'}{\partial\boldsymbol{\theta}}$, and $\mathbf{J}(\boldsymbol{\xi}) = \nabla\mathbf{I}(\mathbf{u}' + \mathbf{u})\frac{\partial\mathbf{u}'}{\partial\boldsymbol{\xi}}$. The partial derivatives of the projected pixel location with respect to the parameters are identical to those obtained when minimizing the reprojection error in Eq. (1), and $\nabla\mathbf{I} \in \mathbb{R}^{1\times 2}$ denotes the image gradient. By equating the partial derivatives in Eqs. (10) and (11) to zero we arrive at the normal equations which can be solved efficiently using standard methods [31].

We note that the Jacobian involved in solving the photometric error has a higher dimensionality than its counterpart in geometric BA. This is because the dimensionality of intensity patches ($D \geq 3 \times 3$) is usually higher than the dimensionality of feature projections (typically 2 for a monocular reconstruction problem). Nonetheless, the Hessian remains *identical* to minimizing the reprojection error and the linear system remains sparse and is efficient to decompose. The sparsity pattern of the photometric BA problem is illustrated in Fig. 2.

Since the parameters of motion and structure are refined jointly, the location of the patch at the reference frame $\boldsymbol{\phi}(\mathbf{u})$ in Eq. (8) will additionally depend on the pose parameters of the reference frame. Allowing the reference patch to "move" during the optimization adds inter-pose dependencies in the linear system and might cause the location of the reference patch to drift. For instance, the solution may be biased towards image regions with brighter absolute intensity values in an attempt to obtain the minimum energy in low-texture areas.

To address this problem, we fix the patch appearance at the reference frame by storing the patch values as soon as the reference frame is selected. This is equivalent to assuming a known patch appearance from an independent source. Under this assumption, the optimization problem now becomes: given a known and fixed patch appearance of a 3D point in the world, refine the parameters of the structure and motion such that photometric error between the fixed patch and its projection onto the other frames is minimized. This assumption has two advantages: (1) the Hessian sparsity pattern remains identical to the familiar form when minimizing the reprojection error using traditional BA, and (2) we can refine the three coordinates (or the full four projective coordinates [10]) of the scene points as opposed to only refining depth along a fixed ray in space.

In addition to improving the accuracy of VSLAM, the algorithm does not require extensive parameter tuning. This is possible by allowing the algorithm to determine the correct correspondences, hence eliminating the many steps required to ensure outlier-free correspondences with traditional BA. The current implementation of the proposed algorithms is controlled by the three parameters summarized in Table 1 and explained next.

Table 1. Configuration parameters for the proposed algorithm shown in Algorithm 1.

Parameter	Value
Patch radius	1 or 2
Non maxima suppression radius	1
Max distance to update V_j	2

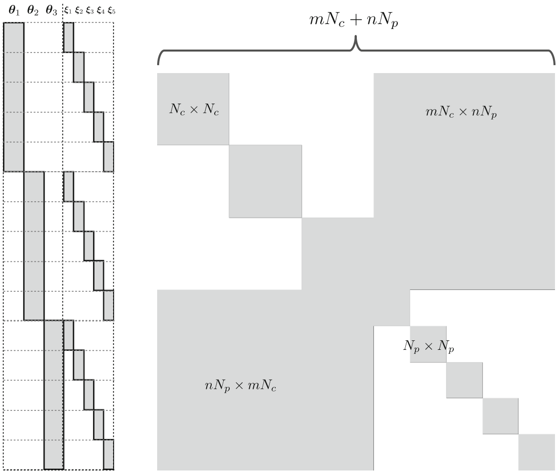

Fig. 2. Shown on the left is the form of the Jacobian for a photometric bundle adjustment problem consisting of 3 cameras, 4 points, and using a 9-dimensional descriptor, with $N_c = 6$ parameters per camera, and $N_p = 3$ parameters per point. The form of the normal equations is shown on the right. The illustration is not up to scale across the two figures.

Selecting Pixels. While it is possible to select pixel locations at every frame using a standard feature detector, such as Harris [32] or FAST [33], we opt to use a simpler and more efficient strategy based on the gradient magnitude of the image. This is performed by selecting pixels with a local maxima in a 3×3 neighborhood of the absolute gradient magnitude of the image. The rationale is that pixels with vanishing intensity gradients do not contribute to the linear system in Eqs. (10) and (11). Other strategies for pixel selection could used [34,35], but we found that the current scheme works well as it ensures an even distribution of coordinates across the field-of-view of the camera [36]. The proposed pixel selection strategy is also beneficial as it is not restricted to corner-like structure and allows us to use pixels from low-texture areas. We note that this pixel selection step selects pixels at integer locations; there is no need to compute accurate subpixel positions of the selected points at this stage.

In image-based (photometric) optimization there is always a distinguished reference frame providing fixed measurements [9,37,38]. Selecting a single reference in photometric VSLAM is unnecessary and may be inadvisable. It is unnecessary as the density of reconstruction is not our main goal. It is inadvisable because we need the scene points to serve as tie points [39] and to form a strong network of constraints [10]. Given the nature of camera motion in VSLAM, selecting points from every frame ensures the strong network of connections between the tie points. For instance, typical hand-held and ground robots motions are mostly forward with points leaving the field-of-view rapidly.

Nonetheless, selecting new scene points at every frame using the aforementioned non maxima suppression procedure has one caveat. If we always select pixels with strong gradients between consecutive frames, then we are likely to track previous scene points rather than finding new ones. This is because pixels with locally maximum gradient magnitude at the consecutive frame are most likely images of previously selected points. Treating projections of previously initialized scene points as new observations is problematic because it introduces unwanted dependencies in the normal equations and superficially increases the number of independent measurements in the linearized system of equations.

To address this issue, we assume that the structure and motion initial estimates are accurate enough to predict the location of the current scene points in the new frame. Prior to initializing new scene points, we use the provided pose initialization to warp all previously detected scene points that are active in the optimization sliding window onto the new frame. After that, we mark a 3×3 square area at the projection location of the previous scene points as an invalid location for initializing new points. Finally, The number of selected points per frame varies depending on the image resolution and image content. In our experiments, this number ranges between ≈4000–10000 points per image.

Determining Visibility. Ideally, we would like to assume that newly initialized scene points are visible in all frames and to rely on the algorithm to reliably determine if this is the case. However, automatically determining the visibility information along with structure and motion parameters is challenging, as many scene points quickly go out of view, or become occluded.

An efficient and reliable measure to detect occlusions and points that cannot be matched reliably is the normalized correlation. For all scene points that are close to the current frame i, we use the pose initialization $\mathbf{T}_i$ to extract a 5×5 intensity patch. The patch is obtained by projecting the scene points to the new frame and its visibility list is updated if the zero-mean normalized correlation score (ZNCC) is greater than 0.6. We allow ±2 frames for a point to be considered close, $i.e.|i - r(j)| \leq 2$. This procedure is similar to determining visibility in multi-view stereo algorithms [4] and is best summarized in Algorithm 1.

Optimization Details. We use the Ceres optimization library [40] to optimize the objective in Eq. (7). We use the Levenberg-Marquardt algorithm [41,42] to minimize a Huber loss function instead of squared loss to improve robustness. Termination tolerances are set to 1×10^{-6}, and automatic differentiation facilities are used. The image gradients used in the linearized system in Eqs. (10) and (11) are computed using central-differences. Finally, we also make use of the Schur complement for a more efficient solution.

Since scene points do not remain in view for an extended period in most VSLAM datasets, the photometric refinement step is performed using a sliding window of five frames [43]. The motion parameters of the first frame in the sliding window is held constant to fixate the Gauge freedom [10]. The 3D parameters of the scene points in the first frame, however, are included in the optimization.

Algorithm 1. Summary of image processing in our algorithm

```
procedure ProcessFrame($\mathbf{I}_i, \mathbf{T}_i$)
    Step 1: establish connections to the new frame
    mask = all_valid(rows($\mathbf{I}$), cols($\mathbf{I}$))
    for all scene points $\mathbf{X}_j$ in sliding window do
        if reference frame $r(j)$ is too far from $i$ then
            continue
        $\mathbf{x} := $ projection of $\mathbf{X}_j$ onto image $\mathbf{I}_i$ using pose $\mathbf{T}_i$
        $\boldsymbol{\phi}' :=$ patch at $\mathbf{x}$ and $\boldsymbol{\phi} :=$ reference patch for $\mathbf{X}_j$
        if zncc($\boldsymbol{\phi}, \boldsymbol{\phi}'$) > threshold then
            add frame $i$ to visibility list $V_j$
            mask($\mathbf{u}$) = invalid

    Step 2: add new scene points
    $\mathbf{G} :=$ gradient magnitude of $\mathbf{I}_i$
    for all pixels $\mathbf{u}$ in $\mathbf{I}_i$ do
        if $\mathbf{u}$ is a local maxima in $\mathbf{G}$ then
            if location $\mathbf{u}$ is valid in mask then
                initialize a new point $\mathbf{X}$ with reference patch at $\mathbf{I}(\mathbf{u})$
```

3 Experiments

In this section, we evaluate the performance of the proposed algorithm on two commonly used VSLAM benchmarks to facilitate comparisons with the state-of-the-art. The first is the KITTI benchmark [44], which contains imagery from an outdoor stereo camera mounted on a vehicle. The second is the Malaga dataset [45], which is particularly challenging for VSLAM because the baseline of the camera (12 cm) is small relative to the scene structure.

3.1 The KITTI Benchmark

Initializing with Geometric BA. Torr and Zisserman [12] convincingly argue that the estimation of structure and motion should proceed by feature extraction and matching to provide a good initialization for BA-based refinement techniques. Here, we use the output of ORB-SLAM [18], a recently proposed state-of-the-art VSLAM algorithm, to initialize our method. ORB-SLAM not only performs geometric BA, but also implements loop closure to reduce drift.

We only use the pose initialization from ORB-SLAM. We do not make use of the refined 3D points as they are available at selected keyframes only. This is because images in the KITTI benchmark are collected at 10 Hz, while the vehicle speed exceeds 80 km/h in some sections. Subsequently, the views are separated by a large baseline, which violates the small displacement assumption required for the validity of linearization in Eqs. (10) and (11).

Hence, to initialize 3D points we use the standard block matching stereo algorithm implemented in OpenCV. This is a winner-takes-all brute force search strategy based on the sum of absolute intensity differences (SAD). The algorithm is configured to search for 128 disparities using a 7×7 aggregation window.

The choice of initializing the algorithm with ORB-SLAM is intentional to assess the accuracy of the algorithm in comparison to the Gold Standard solution from traditional BA. We note, however, that a correspondence-free system is possible by initializing the pose parameters with a direct method [5], or possibly a low-quality GPS.

Performance of the algorithm is shown in Fig. 3 and not only does it outperform the accuracy of (bundle adjusted and loop-closed) ORB-SLAM, but it also outperforms other top performing algorithms, especially in the accuracy of estimating rotations. Compared algorithms include: ORB-SLAM [18], LSD-SLAM [5,30], VoBA [46], and MFI [47].

We note that sources of error in our algorithm are correlated with faster vehicle speeds. This is to be expected as the linearization of the photometric error holds only in a small neighborhood. This could be mitigated by implementing the algorithm in scale-space [48], or improving the initialization quality of the scene structure (either by better stereo, or better scene points obtained from a geometric BA refinement step). Interestingly, however, the rotation error is reduced at high speeds which can be explained by lack of large rotations. The same behavior can be observed with LSD-SLAM's performance as both methods rely on the photometric error, but our rate of error reduction is higher due to the joint refinement of pose and structure parameters.

Initializing with Frame–Frame VO. Surprisingly, and contrary to other image-based optimization schemes [15,50], our algorithm does not require an accurate initialization. Figure 5 demonstrates a significant improvement in accuracy when the algorithm is initialized using frame–frame VO estimates with unbounded drift. Here, we used a direct method to initialize the camera pose without using any feature correspondences [49].

Interestingly, however, when starting from a poor initialization our algorithm does not attain the same accuracy as when initialized using a better quality starting point as shown in Fig. 3. This leads us to conclude the algorithm is sensitive to the initialization conditions more so than traditional BA. Importantly, however, the algorithm is able to significantly improve upon a poor initialization.

Convergence Characteristics and Runtime. As shown in Fig. 6, most of the photometric error is eliminated during the first five iterations of the optimization. While this is by no means a metric of quality, it is reassuring as it indicates a well-behaved optimization procedure. The number of iterations and the cumulative runtime per sliding window of 5 frames is shown in Fig. 7. The median number of iterations is 34 with a standard deviation of ≈6. Statistics are computed on the KITTI dataset frames. The runtime is ≈2 s per sliding window (400 ms per frame) using a laptop with a dual core processor clocked at 2.8 GHz and 8 GB of RAM, which limits parallelism. We note that it is possible to improve the runtime of the proposed method significantly using the CPU, or the GPU. The bottleneck

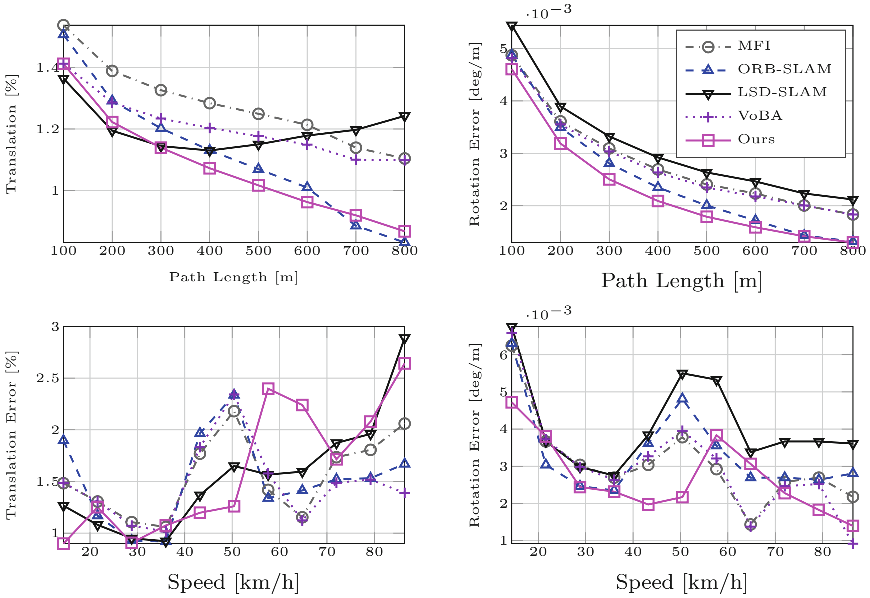

Fig. 3. Comparison to state-of-the-art algorithms on the KITTI benchmark. Our approach performs the best. Error in our approach correspond to segments of the data when the vehicle is driving at a high speed, which increases the magnitude of motion between frames and affects the linearization assumptions. No loop closure, or keyframing is performed using our algorithm. Improvement is shown qualitatively in Fig. 4

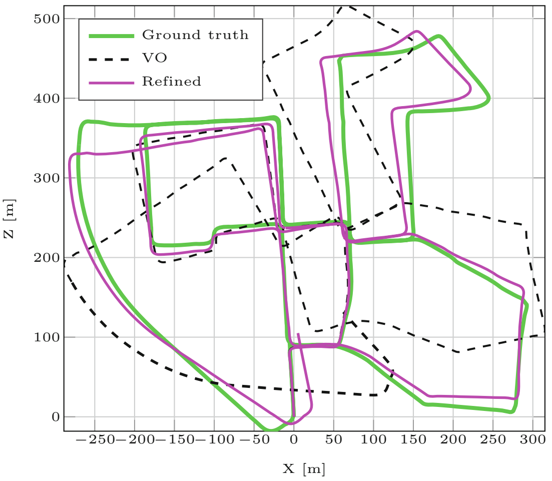

Fig. 4. Improvement starting from a poor initialization shown on the first sequence of the KITTI benchmark. Quantitative evaluation is shown in Fig. 3. We used a direct (correspondence-free) frame–frame VO method to initialize the pose parameters [49].

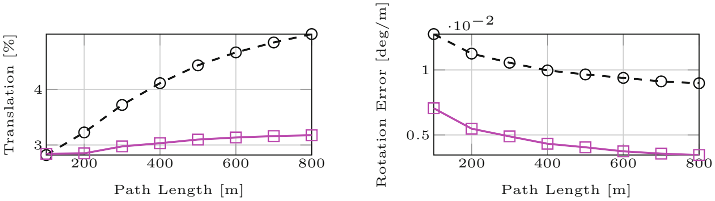

Fig. 5. Improvement in accuracy starting from a poor initialization using a frame–frame direct VO method with unbounded drift.

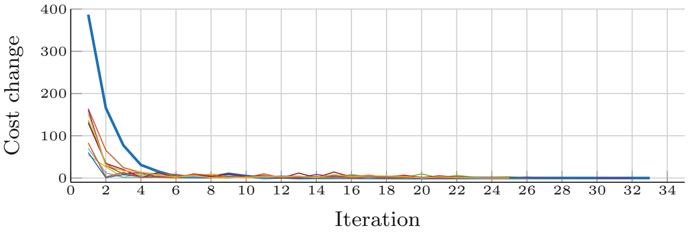

Fig. 6. Rate of error reduction at every iteration shown for the first 10 sliding windows, each with 5 frames. The thicker line shows the first bundle, which has the highest error. Most of the error is eliminated with the first 5 iterations.

of the proposed algorithm is image interpolation (which can be done efficiently with SIMD instructions) and the reliance on automatic differentiation (which limits any code optimization as the code must remain simple for automatic differentiation to work).

3.2 The Málaga Stereo Dataset

The Málaga dataset [47] is a particularly challenging dataset for VSLAM. The dataset features driving in urban areas using a small baseline stereo camera at resolution 800 × 600. The stereo baseline is 12 cm which provides little parallax for resolving distal observations. We use extracts 1, 3, and 6 in our evaluation.

Our experimental setup is similar to the KITTI dataset. However, we estimate the stereo using the SGM algorithm [51], as implemented in the OpenCV library. The stereo is used to estimate 16 disparities with a SAD block size of 5 × 5. We did not observe a significant difference in performance when using block matching instead of SGM.

The Malaga dataset provides GPS measurements, but they are not accurate enough for quantitative evaluation. The GPS path, however, is sufficient to qualitatively demonstrate precision. Results are shown in Fig. 8 in comparison with ORB-SLAM [18], which we used its pose output to initialize our algorithm. We note that in extract 3 of the Malaga dataset (shown on the left in Fig. 8), ORB-SLAM loses tracking during the turn and our algorithm continues *without* initialization.

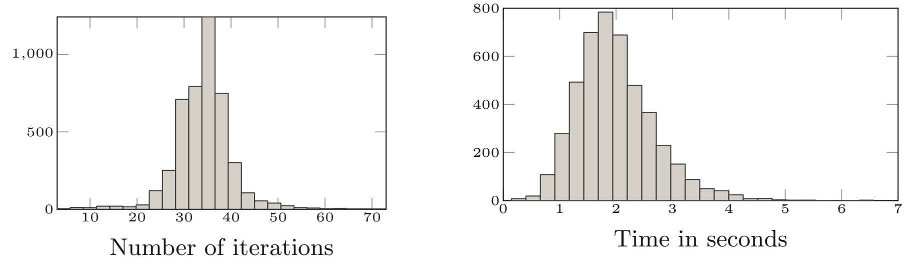

Fig. 7. Histogram of the number of iterations (on the left) and runtime (on the right). The median number of iterations is 34, with a standard deviation of 6.02. The median run time is 1.89, mean 1.98 and standard deviation of 0.69. The runtime is reported for sliding window of 5 frames on the KITTI benchmark.

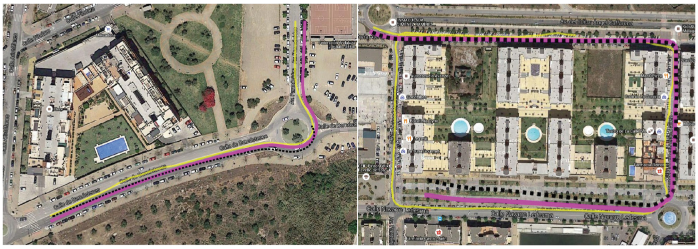

Fig. 8. Our algorithm (magenta) compared with ORB-SLAM (dashed) against GPS (yellow) on extracts 3 and 6 of the Malaga dataset. For extract 3 ORB-SLAM loses tracking during the roundabout, where our algorithm continues without an initialization. Results for extract 6 are shown up to frame 3000 as ORB-SLAM loses tracking. The figure is best viewed in color. (Maps courtesy of Google Maps.) (Color figure online)

To assess the quality of pose estimates, we demonstrate results on a dense reconstruction procedure shown in Fig. 9. Using the estimated camera trajectory, we chain the first 6 m of the disparity estimates to generate a dense map. As shown in Fig. 9, the quality of pose estimates appears to be good.

4 Related Work

Geometric BA. BA has a long and rich history in computer vision, photogrammetry and robotics [10]. BA is a large geometric minimization problem with the important property that variable interactions result in a *sparse* system of linear equations. This sparsity is key to enabling large–scale applications [52,53]. Exploiting this sparsity is also key to obtaining precise results efficiently [54,55]. The efficiency of BA has been an important research topic especially when handling large datasets [56,57] and in robotics applications [58–60]. Optimality and convergence properties of BA have been studied at length [11,61,62] and remain

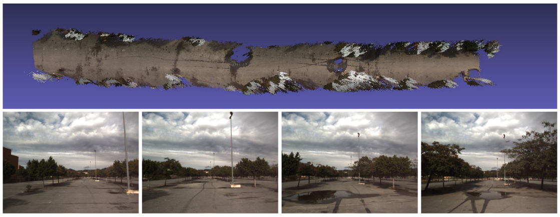

Fig. 9. Dense map from Malaga dataset extract 1. The map is computed by stitching together SGM disparity with the refined camera pose.

of interest to date [63]. All the aforementioned research in geometric BA could be integrated into the proposed photometric BA framework.

Direct Multi-frame Alignment. By direct alignment we mean algorithms that estimate the parameters of interest from the image data directly and without relying on sparse features as an intermediate representation of the image [64]. The fundamental differences between direct methods (like the one proposed herein) and the commonly used feature-based pipeline is how the correspondence problem is tackled and is not related to the density of the reconstruction. In the feature-based pipeline [12], structure and motion parameters are estimated from known and fixed correspondence. In contrast, the direct pipeline to motion estimation does not use fixed correspondences. Instead, the correspondences are estimated as a byproduct of directly estimating the parameters of interest.

The use of direct algorithms for SFM applications was studied for small–scale problems [38,65–67], but feature-based alignment has proven more successful in handling wide baseline matching problems [12] as small pixel displacements is an integral assumption for direct methods. Nonetheless, with the increasing availability of high frame-rate cameras and the increasing computational power, direct methods are demonstrating great promise [5,6,9].

To date, however, the use of direct methods in VSLAM has been limited to frame–frame motion estimation. Approaches that make use of multiple frames are designed for dense depth estimation only and multi-view stereo [4,9], which assume a correct camera pose and only refine the scene structure. Other algorithms can include measurements from multiple frames, but rely on the presence of structures with strong planarity in the environment [37,68] (or equivalently restricting the motion of the camera to rotation only [69]).

In this work, in contrast to previous research in direct image-based alignment [9,38], we show that provided good initialization, it is possible to jointly refine the structure and motion parameters by minimizing the photometric error and without restricting the camera motion or the scene structure.[1]

The LSD-SLAM algorithm [5] is a recently proposed direct VSLAM algorithm. The fundamental difference in comparison to our work is that we refine

[1] While this work was under review, Engel *et al.* proposed a similar photometric (direct) formulation for VSLAM [70].

the parameters of motion and structure jointly in one large optimization problem. The joint optimization of motion and structure proposed herein is important in future work concerning the optimality and convergence properties of photometric structure-from-motion (SFM) and photometric, or direct, VSLAM.

Dense Multi-view Stereo (MVS). MVS algorithms aim at recovering a dense depth estimate of objects or scenes using many images with known pose [4]. To date, however, research on simultaneous refinement of motion and depth from multiple frames remains sparse. Furukawa and Ponce [15] were among the first to demonstrate that relying on minimizing the reprojection error is not always accurate enough. Recently, Delaunoy and Pollefeys [50] proposed a photometric BA approach for dense MVS. Starting from a precise initial reconstruction and a mesh model of the object, the algorithm is demonstrated to enhance MVS accuracy. The imaging conditions, however, are ideal and brightness constancy is assumed [50]. In our work, we do not require a very precise initialization and can address challenging illumination conditions. More importantly, the formulation proposed by Delaunoy and Pollefeys requires the availability of an accurate dense mesh, which is not possible to obtain in VSLAM scenarios.

5 Conclusions

In this work, we show how to improve on the accuracy of the state-of-art VSLAM methods by minimizing the photometric error across multiple views. In particular, we show that it is possible to improve results obtained by minimizing the reprojection error in a bundle adjustment (BA) framework. We also show, contrary to previous image-based minimization work [5,7,9,30,38], that the *joint* refinement of motion and structure is possible in unconstrained scenes without the need for alternation or disjoint optimization.

The accuracy of minimizing the reprojection using traditional BA is limited by the precision and accuracy of feature localization and matching. In contrast, our approach — BA without correspondences — determines the correspondences implicitly such that the photometric consistency is maximized as a function of the scene structure and camera motion parameters.

Finally, we show that accurate solutions to geometric problems in vision are not restricted to geometric primitives such as corners and edges, or even planes. We look forward to more sophisticated modeling of the geometry and photometry of the scene beyond the intensity patches used in our work.

References

1. Sun, D., Roth, S., Black, M.: Secrets of optical flow estimation and their principles. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2432–2439 (2010)
2. Vedula, S., Baker, S., Rander, P., Collins, R., Kanade, T.: Three-dimensional scene flow. IEEE Trans. Pattern Anal. Mach. Intell. **27**, 475–480 (2005)

3. Seitz, S., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 519–528 (2006)
4. Furukawa, Y., Hernndez, C.: Multi-view stereo: a tutorial. Found. Trends Comput. Graph. Vis. **9**, 1–148 (2015)
5. Engel, J., Schöps, T., Cremers, D.: LSD-SLAM: large-scale direct monocular SLAM. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8690, pp. 834–849. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10605-2_54
6. Kerl, C., Sturm, J., Cremers, D.: Robust odometry estimation for RGB-D cameras. In: International Conference on Robotics and Automation (ICRA) (2013)
7. Steinbrucker, F., Sturm, J., Cremers, D.: Real-time visual odometry from dense RGB-D images. In: IEEE International Conference on Computer Vision, ICCV Workshops (2011)
8. Meilland, M., Comport, A.: On unifying key-frame and voxel-based dense visual SLAM at large scales. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3677–3683 (2013)
9. Newcombe, R., Lovegrove, S., Davison, A.: DTAM: dense tracking and mapping in real-time. In: IEEE International Conference on Computer Vision (ICCV), pp. 2320–2327 (2011)
10. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle adjustment — a modern synthesis. In: Triggs, B., Zisserman, A., Szeliski, R. (eds.) IWVA 1999. LNCS, vol. 1883, pp. 298–372. Springer, Heidelberg (2000). doi:10.1007/3-540-44480-7_21
11. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge (2004)
12. Torr, P.H.S., Zisserman, A.: Feature based methods for structure and motion estimation. In: Triggs, B., Zisserman, A., Szeliski, R. (eds.) IWVA 1999. LNCS, vol. 1883, pp. 278–294. Springer, Heidelberg (2000). doi:10.1007/3-540-44480-7_19
13. Kanazawa, Y., Kanatani, K.: Do we really have to consider covariance matrices for image features? In: Proceedings of the Eighth IEEE International Conference on Computer Vision, 2001. ICCV 2001, vol. 2, pp. 301–306. IEEE (2001)
14. Brooks, M.J., Chojnacki, W., Gawley, D., Van Den Hengel, A.: What value covariance information in estimating vision parameters? In: Proceedings of the Eighth IEEE International Conference on Computer Vision, 2001. ICCV 2001, vol. 1, pp. 302–308. IEEE (2001)
15. Furukawa, Y., Ponce, J.: Accurate camera calibration from multi-view stereo and bundle adjustment. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008, pp. 1–8. IEEE (2008)
16. Deriche, R., Giraudon, G.: Accurate corner detection: an analytical study. In: Proceedings of the Third International Conference on Computer Vision, pp. 66–70. IEEE (1990)
17. Shimizu, M., Okutomi, M.: Precise sub-pixel estimation on area-based matching. In: ICCV, pp. 90–97 (2001)
18. Mur-Artal, R., Montiel, J.M.M., Tardós, J.D.: ORB-SLAM: a versatile and accurate monocular SLAM system. CoRR abs/1502.00956 (2015)
19. Milford, M., Wyeth, G.: SeqSLAM: visual route-based navigation for sunny summer days and stormy winter nights. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 1643–1649 (2012)
20. Reid, I.: Towards semantic visual SLAM. In: 13th International Conference on Control Automation Robotics Vision (ICARCV), p. 1 (2014)

21. Salas-Moreno, R., Newcombe, R., Strasdat, H., Kelly, P., Davison, A.: SLAM++: simultaneous localisation and mapping at the level of objects. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1352–1359 (2013)
22. Murray, R.M., Li, Z., Sastry, S.S., Sastry, S.S.: A Mathematical Introduction to Robotic Manipulation. CRC Press, Boca Raton (1994)
23. Ma, Y., Soatto, S., Kosecka, J., Sastry, S.S.: An Invitation to 3-D Vision: From Images to Geometric Models. Springer, New York (2003)
24. Hartley, R., Trumpf, J., Dai, Y., Li, H.: Rotation averaging. Int. J. Comput. Vis. **103**, 267–305 (2013)
25. Civera, J., Davison, A.J., Montiel, J.M.: Inverse depth parametrization for monocular SLAM. IEEE Trans. Robot. **24**, 932–945 (2008)
26. Zhao, L., Huang, S., Sun, Y., Yan, L., Dissanayake, G.: ParallaxBA: bundle adjustment using parallax angle feature parametrization. Int. J. Robot. Res. **34**, 493–516 (2015)
27. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision (DARPA). In: Proceedings of the 1981 DARPA Image Understanding Workshop, pp. 121–130 (1981)
28. Horn, B.K., Schunck, B.G.: Determining optical flow. Artif. Intell. **17**, 185–203 (1981)
29. Baker, S., Matthews, I.: Lucas-Kanade 20 years on: a unifying framework. Int. J. Comput. Vis. **56**, 221–255 (2004)
30. Engel, J., Stueckler, J., Cremers, D.: Large-scale direct SLAM with stereo cameras. In: International Conference on Intelligent Robots and Systems (IROS) (2015)
31. Nocedal, J., Wright, S.J.: Numerical Optimization, 2nd edn. Springer, New York (2006)
32. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey Vision Conference, Manchester, vol. 15, p. 50 (1988)
33. Rosten, E., Drummond, T.: Machine learning for high-speed corner detection. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 430–443. Springer, Heidelberg (2006). doi:10.1007/11744023_34
34. Dellaert, F., Seitz, S.M., Thorpe, C.E., Thrun, S.: Structure from motion without correspondence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 557–564. IEEE (2000)
35. Meilland, M., Comport, A., Rives, P.: A spherical robot-centered representation for urban navigation. In: IROS (2010)
36. Nister, D., Naroditsky, O., Bergen, J.: Visual odometry. In: Computer Vision and Pattern Recognition (CVPR) (2004)
37. Irani, M., Anandan, P., Cohen, M.: Direct recovery of planar-parallax from multiple frames. In: Triggs, B., Zisserman, A., Szeliski, R. (eds.) IWVA 1999. LNCS, vol. 1883, pp. 85–99. Springer, Heidelberg (2000). doi:10.1007/3-540-44480-7_6
38. Stein, G., Shashua, A.: Model-based brightness constraints: on direct estimation of structure and motion. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 992–1015 (2000)
39. Agouris, P., Schenk, T.: Automated aerotriangulation using multiple image multipoint matching. Photogramm. Eng. Remote Sens. **62**, 703–710 (1996)
40. Agarwal, S., Mierle, K., et al.: Ceres solver (2016). http://ceres-solver.org
41. Levenberg, K.: A method for the solution of certain non-linear problems in least squares. Q. J. Appl. Maths. **2**, 164–168 (1944)
42. Marquardt, D.W.: An algorithm for least-squares estimation of nonlinear parameters. J. Soc. Ind. Appl. Math. **11**, 431–441 (1963)

43. Snderhauf, N., Konolige, K., Lacroix, S., Protzel, P.: Visual odometry using sparse bundle adjustment on an autonomous outdoor vehicle. In: Levi, P., Schanz, M., Lafrenz, R., Avrutin, V. (eds.) Autonome Mobile Systems 2005. Informatik aktuell, pp. 157–163. Springer, Heidelberg (2006)
44. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
45. Blanco, J.L., Moreno, F.A., González-Jiménez, J.: The málaga urban dataset: high-rate stereo and lidars in a realistic urban scenario. Int. J. Robot. Res. **33**, 207–214 (2014)
46. Tardif, J.P., George, M., Laverne, M., Kelly, A., Stentz, A.: A new approach to vision-aided inertial navigation. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4161–4168. IEEE (2010)
47. Badino, H., Yamamoto, A., Kanade, T.: Visual odometry by multi-frame feature integration. In: IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 222–229 (2013)
48. Lindeberg, T.: Scale-space Theory in Computer Vision. Springer, New York (1994)
49. Alismail, H., Browning, B., Lucey, S.: Direct visual odometry using bit-planes. CoRR abs/1604.00990 (2016)
50. Delaunoy, A., Pollefeys, M.: Photometric bundle adjustment for dense multi-view 3D modeling. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1486–1493. IEEE (2014)
51. Hirschmuller, H.: Accurate and efficient stereo processing by semi-global matching and mutual information. In: Computer Vision and Pattern Recognition (2005)
52. Agarwal, S., Snavely, N., Seitz, S.M., Szeliski, R.: Bundle adjustment in the large. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6312, pp. 29–42. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15552-9_3
53. Konolige, K., Garage, W.: Sparse sparse bundle adjustment. In: BMVC, pp. 1–11 (2010)
54. Jeong, Y., Nister, D., Steedly, D., Szeliski, R., Kweon, I.S.: Pushing the envelope of modern methods for bundle adjustment. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 1605–1617 (2012)
55. Engels, C., Stewnius, H., Nister, D.: Bundle adjustment rules. In: Photogrammetric Computer Vision (2006)
56. Wu, C., Agarwal, S., Curless, B., Seitz, S.M.: Multicore bundle adjustment. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3057–3064. IEEE (2011)
57. Ni, K., Steedly, D., Dellaert, F.: Out-of-core bundle adjustment for large-scale 3D reconstruction. In: IEEE 11th International Conference on Computer Vision, pp. 1–8 (2007)
58. Konolige, K., Agrawal, M.: FrameSLAM: from bundle adjustment to real-time visual mapping. IEEE Trans. Robot. **24**, 1066–1077 (2008)
59. Kaess, M., Ila, V., Roberts, R., Dellaert, F.: The Bayes tree: an algorithmic foundation for probabilistic robot mapping. In: Hsu, D., Isler, V., Latombe, J.C., Lin, M. (eds.) Algorithmic Foundations of Robotics IX. Springer Tracts in Advanced Robotics, vol. 68, pp. 157–173. Springer, Heidelberg (2011)
60. Kaess, M., Ranganathan, A., Dellaert, F.: iSAM: incremental smoothing and mapping. IEEE Trans. Robot. (TRO) **24**, 1365–1378 (2008)
61. Kahl, F., Agarwal, S., Chandraker, M.K., Kriegman, D., Belongie, S.: Practical global optimization for multiview geometry. Int. J. Comput. Vis. **79**, 271–284 (2008)

62. Hartley, R., Kahl, F., Olsson, C., Seo, Y.: Verifying global minima for L_2 minimization problems in multiple view geometry. Int. J. Comput. Vis. **101**, 288–304 (2013)
63. Aftab, K., Hartley, R.: LQ-bundle adjustment. In: IEEE International Conference on Image Processing (ICIP), pp. 1275–1279 (2015)
64. Irani, M., Anandan, P.: About direct methods. In: Triggs, B., Zisserman, A., Szeliski, R. (eds.) IWVA 1999. LNCS, vol. 1883, pp. 267–277. Springer, Heidelberg (2000). doi:10.1007/3-540-44480-7_18
65. Horn, B.K.P., Weldon, E.J.: Direct methods for recovering motion (1988)
66. Oliensis, J.: Direct multi-frame structure from motion for hand-held cameras. In: Proceedings of the 15th International Conference on Pattern Recognition, vol. 1, pp. 889–895 (2000)
67. Mandelbaum, R., Salgian, G., Sawhney, H.: Correlation-based estimation of egomotion and structure from motion and stereo. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 1, pp. 544–550 (1999)
68. Silveira, G., Malis, E., Rives, P.: An efficient direct approach to visual SLAM. IEEE Trans. Robot. **24**(5), 969–979 (2008). doi:10.1109/TRO.2008.2004829
69. Lovegrove, S., Davison, A.J.: Real-time spherical mosaicing using whole image alignment. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6313, pp. 73–86. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15558-1_6
70. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. ArXiv e-prints (2016)