Efficient Large-Scale Structure From Motion by Fusing Auxiliary Imaging Information

Hainan Cui, Shuhan Shen, Wei Gao, and Zhanyi Hu

Abstract—One of the potentially effective means for large-scale 3D scene reconstruction is to reconstruct the scene in a global manner, rather than incrementally, by fully exploiting available auxiliary information on the imaging condition, such as camera location by Global Positioning System (GPS), orientation by inertial measurement unit (or compass), focal length from EXIF, and so on. However, such auxiliary information, though informative and valuable, is usually too noisy to be directly usable. In this paper, we present an approach by taking advantage of such noisy auxiliary information to improve structure from motion solving. More specifically, we introduce two effective iterative global optimization algorithms initiated with such noisy auxiliary information. One is a robust rotation averaging algorithm to deal with contaminated epipolar graph, the other is a robust scene reconstruction algorithm to deal with noisy auxiliary data for camera centers initialization. We found that by exclusively focusing on the estimated inliers at the current iteration, the optimization process initialized by such noisy auxiliary information could converge well and efficiently. Our proposed method is evaluated on real images captured by unmanned aerial vehicle, StreetView car, and conventional digital cameras. Extensive experimental results show that our method performs similarly or better than many of the state-of-art methods, in terms of reconstruction accuracy and completeness, but is more efficient and scalable for large-scale image data sets.

Index Terms—Structure from motion, auxiliary imaging information, potential inlier, 3D reconstruction.

I. INTRODUCTION

STRUCTURE from motion (SfM) approaches have been widely used to build 3D scene models from images in the past few years. The state-of-art IBA (incremental bundle adjustment) approaches [1]–[3] start by selecting a few seed images for initial reconstruction, then repeatedly add new images to incrementally reconstruct the scene and refine the result by bundle adjustment. Although such an incremental mode finds its success in a variety of applications, it may suffer from large error accumulation, scene drift and heavy computational load. Contrary to IBA, many global algorithms [4]–[9] which simultaneously operate on all images are reported recently, in which the bundle adjustment, a time consuming module, is activated only once rather than repeatedly. However, sometimes such global methods do not work well because the estimated parameters are not accurate enough for the final bundle adjustment. Fortunately, with the progress of sensors technology, many imaging devices come with built-in sensors, such as Global Positioning System (GPS), compass and IMU (inertial measurement unit), and from which approximate camera poses can be easily obtained. These sensor data, though too noisy to be directly used for 3D reconstruction [4], [5], may in fact provide many informative and valuable priors for the SfM solving. Thus in this paper, we present a novel global strategy to solve SfM problem by fully exploiting available noisy auxiliary imaging information, such as GPS, IMU info, and compass angle. One key advantage of our method is its versatility, applicable to both unordered (Fig. 1a) and ordered images (Fig. 1b and Fig. 1c), and another is its remarkable computational efficiency, and the scalability for large-scale scene reconstruction.
reconstruction. For example, as shown in our experiments, we successfully reconstruct the image dataset SV2 which contains 16600 images. Our proposed method consists of three main steps. The first one is to build an EG (epipolar graph), and the second one is a robust iterative rotation estimation, where global camera rotations are estimated by iteratively discarding gross epipolar edges because pairwise geometry estimates always contain errors even under RANSAC paradigm. The last step is to iteratively perform triangulation and bundle adjustment. In order to tackle the problem of gross errors in pairwise geometry estimates as well as the inaccuracy of initializing camera centers with noisy GPS data, we introduce a concept called “potential inlier” for the iterative optimization process, which constitutes one of our major novelties.

Our proposed method is validated on various image datasets, including images captured by UAVs, StreetView cars and conventional digital cameras, and compared with many of the state–of–art methods, such as Bundler [1], MRF-based [4], VSFM [7], OpenMVG [8] and Linear Method [9]. The preliminary version of this article appeared in [11].

II. RELATED WORK

A. Incremental SFM

With an incremental mode, many approaches [1]–[3], [12], [13] repeatedly invoke the bundle adjustment. The state-of-art representative of incremental SFM is Bundler [1], which may suffer from drift due to the accumulation of errors in addition to its heavy computational load when handling large image datasets. The worst-case time complexity of image matching part and bundle adjustment part is respectively \(O(n^2)\) and \(O(n^4)\) in the number of images, which becomes prohibitive when the number of images is large. Many remedies are proposed to tackle this problem recently. For the image matching module, graph-based algorithms [14], [15] are proposed to improve the efficiency by pruning the original image set. Snavely et al. [14] compute a small skeletal graph on a subset of images at first, then reconstruct the skeletal set and register the remaining images subsequently. However, the graph construction is always time-consuming, and sometimes the completeness of the reconstructed scene cannot be guaranteed. The other typical solution is to employ image retrieval technology [16]–[20]. Nister and Stewenius [16] build a hierarchical vocabulary tree to explore candidate matching image pairs, and Havlena and Schindler [18] also cast feature matching as image indexing when sufficiently large visual vocabularies are available. These image retrieval methods render the time complexity of image matching part linearly with the number of images. Moreover, based on the rank of Hamming distance, Cheng et al. [21] propose a Cascade Hashing strategy of three layers to accelerate the pairwise image matching. For the bundle adjustment part, parallelism techniques [22], [23] are usually employed to accelerate the matrix computations. For example, Wu et al. [23] present a new inexact Newton type bundle adjustment algorithm to fully exploit hardware parallelism for efficiently solving large-scale 3D scene reconstruction problems.

Additionally, the reconstruction result of incremental approaches largely depends on the seed images selection rule which in turn usually depends on the estimated fundamental and homography models. Since the quality of such estimated entities directly affects the quality of 3D reconstruction, Moulon et al. [3] propose to improve their estimation through a contrario methodology. Haner and Heyden [12] present a new selection and addition rule based on covariance propagation, and point out that a well-determined camera should have both small estimated covariance and low reprojection error for the next view planning.

B. Global SFM

Global reconstruction methods [7]–[9], [24]–[27], which only optimize the reconstruction result once for all images, have risen to prominence in the last few years. These approaches usually take three steps to solve the SFM problem. The first one is to compute camera rotations by rotation consistency, the second is to calculate camera translations, and the third one is to refine camera poses and 3D points by performing a final bundle adjustment.

Many linear methods [9], [24]–[26] are proposed for solving global camera orientations or translations in the literatures. For example, Jiang et al. [9] propose a linear method for global camera pose registration, in which a lot of accurate pairwise geometries are required to perform the SVD decomposition for accurate camera positions estimation. Similarly, a linear algorithm based on a novel decomposition of the Essential Matrix to recover the global translation vectors is reported in Arie-Nachimson et al. [24]. While efficient, such linear methods are rather sensitive to outliers.

Given known rotations, some \(l_\infty\) methods are presented for solving for camera (and possibly point) positions [28]–[30]. The joint estimation of translations and 3D points is formulated under the \(l_\infty\) norm and solved by using second-order cone programming, as proposed by Hartley and Shaffalitzky [29], and later generalized by [30]. However, we find such \(l_\infty\) methods are also sensitive to outliers, and they could fail on noisy image collections.

In order to increase the reconstruction accuracy, many approaches [8], [31], [32] resort to accurate triple geometries. For example, Moulon et al. [8] propose a contrario trifocal tensor estimation method to extract stable translations. While the accuracy could be increased, such triplet-based methods are more likely to discard many useful images since the corresponding epipolar graph may not be dense and accurate enough. For example, for StreetView images, pairwise geometry estimates are usually sparse and noisy because the StreetView car is always driving fast. As a result, many images may be discarded by triplet-based reconstruction method due to their weak visual connections with others.

C. SFM With Auxiliary Imaging Information

Auxiliary imaging information, such as GPS, IMU or the compass angle, are always useful for the SFM solving. Carceroni et al. [33] compute camera rotations using GPS directly. Pollefeys et al. [34] report a real-time SFM in
urban scene reconstruction with the support of GPS/IMU sensors. However, these two methods rely on high-precision GPS sensors which are not available in common consumer-level devices. Several methods [4], [35], [36] are proposed to reconstruct a 3D scene by exploiting noisy auxiliary imaging information. Crandall et al. [4] propose a discrete-continuous optimization method, in which noisy auxiliary info (GPS and vertical vanishing point) is incorporated into the SfM process. In their method, vertical vanishing points are used to estimate the tilt angle, and BP (belief propagation) is used on a discretized space of camera orientations and 2D camera positions to find a good parameter initialization, followed by a nonlinear least squares and bundle adjustment to refine the initial estimates. Similarly, Sinha et al. [35] also propose a linear method in which vanishing points are exploited to solve SfM problem. However, these two methods are not applicable to UAV images because the vertical vanishing points cannot be estimated when the UAV faces a large tract of land where evident lines are not always available. Besides, the extent of the scene should be predetermined in [4], and discrete position labels require a huge amount of memory when the scene covers a large area.

In this paper, we present an efficient and versatile global approach to improve the SfM solving, which fully exploits noisy auxiliary imaging information obtained from consumer-level sensors. Here by versatility, we mean our approach is applicable to various kinds of images, including common digital images, UAV images and StreetView images.

III. A Global Approach by Iteratively Optimizing Potential Inliers

The input of our reconstruction problem consists of (a) a set of images \( I = \{I_1, I_2, \ldots, I_N \} \), (b) auxiliary imaging information for each image, such as GPS, compass angle, and focal length. Our goal is to estimate an absolute camera pose for each image, and a 3D position for each scene point, consistent with all the input measurements as much as possible in a geo-referenced coordinate system. Our SfM method, shown in Fig. 2, consists of three main steps. Step1 is a pre-processing step, aiming at getting noisy point correspondences between images and then building an EG (epipolar graph).

In Step2, global camera rotations are iteratively estimated through rotation consistency. In order to increase the percentage of real epipolar edge inliers in the optimization, gross edge outliers are filtered out in each iteration. Then, camera poses and 3D scene points are iteratively estimated in Step3 by initializing camera centers with noisy GPS data. Next these three steps will be elaborated.

A. Step 1: Pre-Processing

Firstly, the raw GPS data, which is always in the form of longitude, latitude and altitude defined in the WGS84 coordinate system, is converted into the ECEF (Earth Centered, Earth Fixed) coordinate system. Here the ECEF, which is usually called the local east-north-up, is used as our global geo-referenced coordinate system. Then, we compute SIFT matches using Cascade Hashing strategy [21], for which a vocabulary tree [16] is built to detect candidate matching image pairs. Note that too distant image pair are directly discarded based on the noisy geotags. After matching relevant images, geometric verification is performed for each candidate pair, where RANSAC is exploited to improve the robustness and the residual error threshold is set to 2.5 pixels. Two images are considered as a matched pair if the number of matched feature points is more than a threshold (in our work, it is set to 20), and then their pairwise geometry is computed.

Given the feature correspondences of every matched pair, 3D scene points are identified by finding their corresponding track–interest points across multiple images which have similar SIFT descriptors. However, sometimes a feature point may be contained by different tracks due to features mismatching, thus such tracks are considered unstable and discarded for subsequent bundle adjustment due to their ambiguous status. The final matching result is represented by an epipolar graph called EG, whose vertices \( V = \{I_1, I_2 \cdots I_N \} \) correspond to images and edges \( E = \{e_{ij} | i, j \in V \} \) link matched pairs. For each edge, the pairwise geometry \( \{R_{ij}, t_{ij}\} \) is further computed by decomposing essential matrix.

Given the pairwise geometry of every epipolar edge in EG, a triplet-based voting scheme is further performed to eliminate gross inconsistent epipolar edges. For each triplet, for example \((i, j, k)\), its three edges \(\{(i, j), (j, k), (k, i)\}\) are...
given a vote of +1 if the geodesic distance [37] between the loop rotation matrix [38] and the identity matrix is smaller than a threshold \( \delta_1 \), or else given a vote of −1. Here, the loop rotation matrix in our case is simply \( R_{ij} \ast R_{jk} \ast R_{ki} \). Then for each epipolar edge, a score is assigned as the ratio of the number of the −1 votes to the total received votes. The edges that received at least 3 votes and with the score greater than a threshold \( \delta_2 \) are detected and discarded. Typically, \( \delta_1 \) and \( \delta_2 \) is set to 60° and 90° respectively. Finally, the LCC (largest connected component) of EG is extracted and used in our subsequent reconstruction steps.

B. Step 2: Robust Iterative Rotation Estimation

Coarse initial camera rotations defined under the ECEF coordinate system can be easily obtained from camera orientations. In our work, the vertical vanishing point is used to estimate the tilt angle as in [4]. If a camera is equipped with a compass sensor, then its orientation is initialized by compass and tilt angle; otherwise, the method proposed by Crandall et al. [4] is used to get a rough orientation.

Given a pairwise geometry estimate \( (R_{pq}, t_{pq}) \) between cameras \( p \) and \( q \), the problem of rotation estimation can be formulated as a search for the absolute orthonormal rotations \( R_p, R_q \), such that the following constraint is satisfied:

\[
R_{pq} = R_p R_q^T
\]

Every edge forms such a constraint. Thus, an overdetermined constraint system is obtained since EG always consists of redundant edges. As proposed by Martinec and Pajdla [39], the solution of this system can be initially computed without considering the orthonormality constraint and then enforced subsequently by projecting it to the closest rotation under Frobenius norm using SVD decomposition. While efficient, this method is sensitive to edge outliers. However, the epipolar edge outliers, whose geometry estimates are either incorrect or actually non-existent, are always present due to features mismatching.

In order to tackle edge outliers and increase the percentage of real edge inliers in the optimization, we propose a robust iterative rotation estimation algorithm by exclusively optimizing the so-called potential edge inliers at each iteration. An edge in EG is regarded as a potential edge inlier in the \( i^{th} \) iteration if its corresponding residual \( \|R_{pq} - R_p R_q^T\|_F \) is less than a threshold \( T^{(i)} \). Given a threshold \( \alpha \), \( T^{(i)} \) in the \( i^{th} \) iteration is initially computed as follows:

\[
T^{(i)} = \min \{ T : \frac{\sum_{j=1}^{M} \eta_j^{(i)}}{M} \geq \alpha \}
\]

subject to that each matrix in \( \mathbf{R} \) is orthonormal. \( e_{pq}^{(i)} \) is set to 1 if the residual of the edge between image \( p \) and image \( q \) is smaller than \( T^{(i)} \), otherwise set to 0. With the iterations going on, the camera rotations would become more and more accurate and then more real edge inliers would be involved in the optimization. For the sake of efficiency, the iteration is stopped when the overlap rate \( O_{\text{edge}} \) of the potential edge inliers (\( PEI \)) between the two consecutive iterations is more than a threshold \( \tau_1 \) (in our work, \( \tau_1 \) is set to 99.0%).

\[
O_{\text{edge}}^{(i)} = \frac{PEI^{(i)} \cap PEI^{(i-1)}}{PEI^{(i)} \cup PEI^{(i-1)}}
\]

C. Step 3: Robust Iterative Scene Reconstruction

Given the camera rotations \( \mathbf{R} \), the initial camera projection matrix set \( \mathbf{P} = \{ P_i; i = 1 \ldots N \} \) is set as \( P_i = K_i R_i (I - C_i) \), where \( K_i \) is the intrinsic matrix of camera \( i \) (focal length obtained from EXIF), \( R_i \) the rotation of camera \( i \), and \( C_i \) the converted geotags of image \( i \). Then, 3D scene points are initially reconstructed by triangulating their corresponding tracks. For each track, we estimate the angles of all pairs of rays that could be used to triangulate the 3D point, and find the image pair having maximal angle of separation \( \theta_{\text{max}} \). If \( \theta_{\text{max}} \) is larger than a threshold \( \gamma_1 \), then the point is triangulated and kept as a candidate for further processing. For the robustness concern, a 3D point is ignored when its current average
reprojection error across all visible images is greater than \(\gamma_2\) pixels. (in our work, \(\gamma_1 = 2^\circ\) and \(\gamma_2 = 100\)).

Given the camera projection matrix set \(P\) and the set of currently reliable reconstructed 3D points \(X\), the discrepancy between the measured 2D image point locations and projection of the predicted 3D scene points is minimized subsequently. For \(N\) images and \(K\) tracks, the cost function \(G\) is formulated as the weighted geometric projection errors:

\[
G(P, X) = \sum_{i=1}^{N} \sum_{j=1}^{K} v_{ij}\|x_{ij} - \gamma(P_i, X_j)\|^2 \tag{7}
\]

where 2D image point locations \(x_{ij}\) are the observation of the 3D point \(X_j\) in the \(i^{th}\) image; \(v_{ij}\) is set to 1 if \(X_j\) is visible in the \(i^{th}\) image, otherwise set to 0. \(\gamma(P_i, X_j)\) denotes the projection of \(X_j\) in the \(i^{th}\) image. Note that in our work only the first two camera radial distortion parameters are used.

The non-linear least square problem defined in Eq. (7) needs a good parameter initialization. However, converted GPS locations are not accurate enough to be used as the camera positions initialization, and track outliers caused by features mismatching are inevitable. In most case, we found one-time bundle adjustment is not sufficient to produce satisfactory reconstruction result. Thus to tackle this problem, we propose an iterative approach by exclusively performing optimization on potential track inliers at the current iteration, where additional triangulations and bundle adjustments are carried out.

A track is regarded as a potential track inlier in the \(i^{th}\) iteration if its average reprojection error across visible images is less than \(H^{(i)}\). Given a threshold \(\beta\), \(H^{(i)}\) is initially calculated as:

\[
H^{(i)} = \min\{H : \frac{\sum_{j=1}^{K}\delta_j^{(i)}}{K} \geq \beta\} \tag{8}
\]

s.t. \(\delta_j^{(i)} = \begin{cases} 0, & \text{if } r_j^{(i)} > H; \\ 1, & \text{if } r_j^{(i)} \leq H; \end{cases} \tag{9}\)

where \(r_j^{(i)}\) denotes the averaged reprojection error of \(j^{th}\) track in the \(i^{th}\) iteration; \(j = 1 \ldots K\); \(K\) denotes the number of tracks. Similar as Section III-B, a covering criteria should be further satisfied here: the current potential track inliers should cover all the edges in the LCC of EG. If this criteria is not satisfied, we use a modified K-cover algorithm [40], called 1-cover here, to obtain a subset satisfying the criteria. Firstly, tracks are sorted in an ascending order according to their average reprojection errors. Then, we greedily select a minimal subset of tracks that covers each epipolar edge in the LCC of EG at least 1 time. Note that the difference between our selection strategy and that proposed in [4] mainly includes two aspects: one is that gross tracks have been filtered by our triangulation part, the other is that our search algorithm is along the ordered tracks rather than along the original tracks in [4]. Finally, the maximal average reprojection error of the tracks in this subset is used to update the current threshold. Thus, the final threshold \(H^{(i)}\) that used to classify a track is calculated as:

\[
H^{(i)} = \max\{1\text{-cover}^{(i)}, H^{(i)}\} \tag{10}\]

Since there are still outliers in the potential tracks inliers, we use a robust Huber function by setting its parameter as 25 pixels on the reprojection error. Moreover, although the focal lengths obtained from image EXIF tags and the geotags are not accurate enough, they should not be too erroneous either, hence two penalty terms are added to the cost function (Eq. (7)). Thus, at the \(i^{th}\) iteration, our cost function is formulated as:

\[
F(P^{(i)}, X^{(i)}) = \sum_{i=1}^{N} \sum_{j=1}^{K} v_{ij}(x_{ij} - \gamma(P_i, X_j)) + \sum_{i=1}^{N} \lambda(\sum_{j=1}^{K} \delta_j^{(i)} - \gamma(P_i, X_j))^{2} \tag{11}\]

where \(\delta_j^{(i)}\) is set to 1 if the \(j^{th}\) track is considered as a potential track inlier in the \(i^{th}\) iteration, otherwise set to 0; \(f_j^{(i)}\) and \(f_{ei}\) are the focal length of the \(i^{th}\) image in the \(i^{th}\) iteration and the corresponding EXIF tag; \(\lambda\) is a scaling constant; \(C_i^{(i)}\) denotes the position of the \(i^{th}\) camera in the \(i^{th}\) iteration, which should be not too far from its geotags \(GPS_i\). In our work, \(\beta\) is set to 90%, which means the 10% of tracks with the largest average reprojection errors, are considered as potential track outliers, and are discarded in the current optimization. The determination of this threshold is discussed in the experimental section. For the sake of efficiency, the iteration is stopped when the overlap rate \(O_{\text{track}}\) of the potential track inliers (PTI) between the two consecutive iterations is more than a threshold \(\tau_2\) (in our work, \(\tau_2\) is set to 99.0%).

\[
O_{\text{track}}^{(i)} = \frac{\sum_{j=1}^{K} \text{PTI}^{(i)} \cap \text{PTI}^{(i-1)}}{\sum_{j=1}^{K} \text{PTI}^{(i)} \cup \text{PTI}^{(i-1)}} \tag{12}\]

Conventionally, repeated bundle adjustment is regarded as the most time-consuming part in 3D reconstruction. However, as our following experimental part shows, the time-cost of repeated bundle adjustment in this step is acceptable. This is because on the one hand, only a part of tracks are optimized in each iteration, and we find the number of the iterations is always less than 5; on the other hand, the sparse structure of SFM problem is taken into account. In our work, the weighting factor \(\lambda\) in Eq. (11) is set to 10^−4, and ceres-solver [41] is adopted to perform the bundle adjustment.

IV. EXPERIMENTAL RESULTS

A. Datasets

The extensive experiments are carried out on a PC with an Intel Xeon E5-2603 2.50GHz CPU(4 cores) and 32G RAM. Since the reconstruction results may get influenced by many unpredictable factors, such as the accuracy of auxiliary imaging information, the cleanliness of image tracks and the distance from the cameras to the real scenes, thus it is nearly impossible to conduct synthetic experiments to simulate realistically. Instead, we validate our reconstruction method on extensive real image experiments. In this section, three kinds of real images, including (1) UAV images; (2) images captured by free shooting; (3) StreetView images, are used, and their specifications are listed in Table I. For UAV image, compass
TABLE I
SPECIFICATIONS OF IMAGE DATASETS

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th># of images in LCC</th>
<th>Capturing device</th>
<th>GPS accuracy</th>
<th>compass accuracy</th>
<th>same initial focal length?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Land1</td>
<td>122</td>
<td>Canon SD Mark III</td>
<td>3-5 m</td>
<td>5-10 m</td>
<td>no</td>
</tr>
<tr>
<td>Land2</td>
<td>144</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Land3</td>
<td>942</td>
<td></td>
<td>5-10 m</td>
<td>5-10 m</td>
<td>yes</td>
</tr>
<tr>
<td>UAV1</td>
<td>145</td>
<td>UAV</td>
<td>5-10 m</td>
<td>5-10 m</td>
<td>yes</td>
</tr>
<tr>
<td>UAV2</td>
<td>453</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UAV3</td>
<td>501</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UAV4</td>
<td>1124</td>
<td></td>
<td>5-10 m</td>
<td>5-10 m</td>
<td></td>
</tr>
<tr>
<td>SV1</td>
<td>2468</td>
<td>StreetView Car</td>
<td>3-5 m</td>
<td></td>
<td>yes</td>
</tr>
<tr>
<td>SV2</td>
<td>16600</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SV3</td>
<td>3270</td>
<td></td>
<td>3-5 m</td>
<td>5-10 m</td>
<td>yes</td>
</tr>
</tbody>
</table>

angle is a necessity and tilt angle is set to 0°. Yet for the other two kinds of images, compass angle is just an alternative. If it is not available, the method proposed by Crandall et al. [4] is used to get a rough orientation.

We also use three datasets (Land1, Land3 and the Arts Quad) where the camera location ground truth is available. For Land1, 122 images have geotags recorded by a consumer-level GPS receiver, and among them 120 images have very precise GPS coordinates obtained by a survey-quality differential GPS sensor (error under 10cm), which can be used for ground truth. For Land3 of 942 images, 300 images have ground truth camera positions. For the Arts Quad dataset, the images and 348 truth GPS locations are publicly available in [42].

B. Comparison Methods and Criteria

We compare our method with both the state-of-art incremental SfM approaches, including Bundler [1] and VSFM [7], and some recent representative global SfM approaches, including MRF-based method [4], OpenMVG [8] and the Linear Method [9]. Note that since OpenMVG in [43] requires images to have the same initial focal length, it cannot be run on Land1, Land2 and Land3 where the images are with different focal lengths. In addition, the tilt plays an important role in MRF-based approach [4], thus it is not performed on UAV images due to the lack of vertical vanishing points for such images.

For the comparison criteria, both qualitative and quantitative comparisons are carried out. In the qualitative comparisons, both the scene structures and the camera trajectories are assessed. Gross calibration errors or evident artifacts are the direct indicators of the algorithm’s inadequacy. In the quantitative comparisons, we evaluate the accuracy of the reconstructed cameras by comparing their positions to the ground truth locations. Moreover, the running-times spent by the reconstruction approaches, setting aside image feature detection and matching module, are recorded to compare the computational load. Note that in all our subsequent experiments, we use red and green points to alternately show the calibrated camera centers, which is the same way as Bundler [1].

C. Parameter Settings

Our proposed method mainly has two important parameters: $\alpha$ in Section III-B and $\beta$ in Section III-C, which should be influenced by many unpredictable factors, such as the accuracy of auxiliary imaging information and the cleanliness of tracks. Since it is difficult to theoretically set these two values, we determine them experimentally. Note that considering the parameters generality, both the images captured on the land and the images captured in the air should be explored in this probe experiment. Thus, four image datasets (Land1, Land2, UAV2 and UAV4), which can be successfully reconstructed by Bundler [1], are selected for the determination. Since the true camera poses are usually unavailable for large-scale SfM problems, here the result of Bundler is regarded as the ground truth, as it is a state-of-the-art SfM system and its obtained results are usually faithful. In our experiments, these two parameters are explored step by step from $(70\% + k \times 5\%, k = 0, 1, \ldots, 6)$.

For the camera rotation accuracy showed in Fig. 3–Rotation Accuracy, a decrease appears on Land1 when $\alpha$ comes to about 85%. For the computational efficiency showed in Fig. 3–Running-time, less constraints, e.g. less epipolar edges, almost means less time-demand spent by the non-linear optimization. Thus considering both the criteria, a tradeoff initial $\alpha$ is set to 90%. Fig. 4 shows the camera center accuracy, rotation accuracy and running-time change with respect to different settings of $\beta$. For Land2, the center accuracy decreases when $\beta$ drops to about 85%. Yet when it is set to 90% or 95%, both the center and rotation accuracy on four image datasets are comparable. Thus, further considering
the running-time, we set \( \beta \) to 90% also. Note that these two thresholds are fixed in all our subsequent experiments.

D. Comparison Results

For better evaluating the two vital steps (rotation estimation step and scene reconstruction step), two datasets (Land2 and UA V1) are selected to show the iterative processing details.

1) Results of Step 2: In this step, our goal is to obtain accurate camera rotations. Here we use the epipolar edge residual histogram, where the residual is the Frobenius norm of consistency measurement, to illustrate the rotation accuracy. For example, for the edge linked camera \( p \) and camera \( q \), the residual is computed by \( \| R_{pq} - R_{q}R_{q}^{T} \|_{F} \). The iterative results of Land2 and UA V1 are showed in Fig. 6(a) and Fig. 6(b). For example for UA V1, we can infer from the leftmost three histograms that the estimated rotations become more and more accurate with the iterations going on, which further verifies the necessity of iterations. Additionally comparing with the ground-truth rotations (the fourth histogram from left), our result is much better than the result produced by optimizing all of the epipolar edges (\( \alpha = 100\% \)), which is showed at the rightmost histogram of Fig. 6(b). For the convergence, the Fig. 5a shows this step converges with less than five iterations.

In addition, Fig. 7 shows the epipolar graphs before and after our potential inlier selection process. We found that for images captured on the ground, most of the epipolar edge outliers are produced by featured mismatching. For example for Land2 in Fig. 7(a), two images showed in the example of potential outlier do not have any common visible area. However, these two images have quite similar textures and lots of features are mismatched, thus the pairwise geometry is still obtained. In addition, for UAV images, most of the epipolar edge outliers are caused by small common visible area. For example for UAV1 in Fig. 7(b), two images showed in the example of potential outlier have very small common visible area. In this situation, the obtained pairwise geometry is always gross because the common area does not reflect the full geometry relationship between two images.

2) Results of Step 3: Some reconstruction results with respect to the iteration time are shown in Fig. 8, from which it can be seen that one-off bundle adjustment is obviously not enough when camera centers are directly initialized with noisy GPS. Since initial parameters are not good enough, only a subset of tracks are regarded as potential track inliers at the first iteration. With iterations going on, more potential track inliers appear in the subsequent iterations, which means the camera poses become more and more accurate. In particular, both camera positions and scene structure are unreasonable after the first iteration for Land2 (Iter_1). However, with the iterations going on, the scene structure becomes more and more reasonable. For the convergence, Fig. 5b shows this step converges in less than five iterations.

Tracks are always clean in UAV images as no occlusions exist in the view, but are contaminated in images captured by free shooting or the StreetView car because of the large changes of view angles or the existence of numerous self-symmetric features. In our experiment, the percentage of final potential track inliers to the whole tracks on Land2 and UAV1 is 54.07% and 83.71% respectively, which shows our proposed method works satisfactorily on both cluttered and clean scenes.
3) Qualitative Comparison: Due to the space limit, four datasets (Land2, UAV1, UAV3, and SV1) are chosen to give a qualitative comparison among different reconstruction methods, as shown in Fig. 9. For OpenMVG and Linear Method, the camera poses estimation requires the existence of many accurate triplets. However, there may be not enough triplets in some UAV and StreetView datasets because the speed of UAV or car is usually fast. Especially for StreetView images, many pairwise geometry estimates may not be accurate enough due to the existence of numerous self-symmetric features. As shown in Fig. 9, both OpenMVG and Linear Method produce erroneous reconstructions on UAV1, UAV3 and SV1. But for Land2, since buildings are captured many times by the moving cameras, the reconstruction results are comparable among the five methods (Bundler, MRF-based approach, VSFM, Linear Method and our method), indicating that most existing SfM methods are suited for this kind of scenarios.

For the scene structure of UAV1, the result produced by the openMVG and that by the Linear Method are obviously incomplete. For the camera trajectory of UAV1, our result is similar to the Bundler’s result. However, compared with
Fig. 8. Sparse reconstruction results with respect to the iteration number (from Iter_1 to Iter_5). Red and green points denote the camera positions.

Fig. 9. Sparse reconstruction results on 4 image datasets (Land2, UAV1, UAV3 and SV1). Red and green points denote the camera positions. Blue ellipses mark the sampled unreasonable areas in the results. For the images in Land2 have different focal length, they cannot be reconstructed by openMVG. For the tilt angles in the datasets UAV1 and UAV3 are not available, these two datasets are not reconstructed by MRF-based method.

VSFM’s result on UAV1, the camera trajectory of our result is more reasonable (unreasonable jitters appear in VSFM’s result highlighted by a blue circle). Similarly for UAV3, the scene structure reconstructed by openMVG is apparently wrong. The reason is likely that the estimation of image distortion in OpenMVG is not accurate, and more elaborate reasons are reported by Wu [44]. Results produced by the Linear Method on UAV3 are obviously wrong, and there are also
some obvious calibration errors (sudden leap on camera centers) in the results produced by VSFM and Bundler. Thus, considering the performance of our method on both UAV1 and UAV3, our SfM system produces more reasonable results than others.

For the reconstructed scene on SV1, the results produced by Bundler, openMVG and Linear Method are obviously incomplete or wrong. In order to make the comparison more evident, the scene structure and camera trajectory of other results are respectively shown in Fig. 10. For the scene structure, some obvious errors, which are highlighted by red circles, appear in the results produced by the MRF-based approach and VSFM. For the camera trajectory, our result is more reasonable as no obvious jitters appear on the route of car. Two blue ellipses mark the unreasonable parts on the results of VSFM and MRF-based method. The reason of why our results are better than those produced by the MRF-based method is mainly due to the following two factors. Firstly, the 2D camera positions in the MRF-based method may be not dense enough (in our experiments, a label corresponds a 4m*4m square). As a result, parameter initializations may be not accurate enough for the final bundle adjustment. Secondly, the MRF-based method is more sensitive to track outliers since the accuracy of initial translations largely depends on the selected tracks.

In sum, in term of qualitative comparison, our method outperforms the other five ones, and some other reconstruction results produced by our method are additionally showed in Fig. 11. Note that these results can be used as a stepping-stone for further processing, notably for dense reconstruction, for example, some dense reconstructions produced by PMVS2 are shown in Fig. 1.
4) Quantitative Evaluation: Ideally, high-quality 3D ground truth points (e.g., from a laser scanner) are required to compare with the reconstruction results. However, it is expensive to collect such data for large-scale scenes. Instead, we evaluate our method by comparing our calibrated camera positions with the ground-truth locations collected by highly accurate differential GPS sensors, and by comparing it with Bundler [1]. It is noted that the IBA solution has errors and is not ground truth, but since Bundler is a state-of-the-art SfM system, a comparison with it is nevertheless informative.

Given the ground-truth camera positions and corresponding calibrated camera centers, RANSAC is used to estimate a 3D similarity transformation between them. Table II summarizes the comparison results, including the specifications of each dataset and median distances between corresponding camera positions. Note that the dataset Arts Quad is publicly available in [4], where the reconstruction accuracy is reported as 1.16m, similar to ours. From the median center error in Table II, we can further conclude that the accuracy of our proposed method is comparable with the state-of-art Bundler.

5) Time Efficiency and Scalability: As shown in Fig. 9, none but Land2 is successfully reconstructed by the most of the methods, and Linear Meathod [9] is fastest one for Land2. But for the other datasets, OpenMVG [8] and Linear Meathod [9] always leave some images uncalibrated due to their dependency on accurate triplets. Thus, in order to fairly compare the efficiency of reconstruction, the time-cost of our method is compared with the other methods, including Bundler [1], MRF-based method [4] and VSFM [7]. Note that these three methods are parallel performed on all the datasets in Table I. The running-times and the number of calibrated images of the corresponding method are all recorded in Table III, from which we can see that our method performs better than other approaches on all the datasets except Land2 and UAV1. In particular for VSFM, although the proposed preemptive matching method speeds up the process of SfM solving, it makes the visual connection between images weak, which in turn may make more images uncalibrated and increase the risk of scene incompleteness.

For UAV datasets, the computational efficiency of our method is more salient. For example for UAV4, our method is about 40 times faster than the conventional incremental SfM method Bundler [1] but with similar accuracy, indicating that GPS is a piece of quite useful and effective prior information for this application, and its underlying reasons are likely two-fold: one is the tracks of UAV are more clean as no occlusions exist in the view; the other is that the impact caused by noisy GPS in term of the reprojection error is lessened when the flying height of UAV is relatively high, and for such cases, initializing camera centers as GPS is a fairly good choice for the bundle adjustment. Table IV shows the detailed comparison between our method and the MRF-based method on SV1. These two methods are all global reconstruction methods, and the MRF-based approach spends a lot of time in estimating translations while the bulk of the running time of our approach is spent in the iterative bundle adjustment.
From this comparison, we can conclude that our method has a better scalability than the MRF-based approach.

E. Discussion

Before ending this section, here is a short discussion on the global convergence of our proposed iterative optimization method, as well as on the possible underlying reasons of why our method performs well on various kinds of datasets.

Two related questions could be arose: Can the global convergence of our method be guaranteed for all kinds of datasets? If yes, why does so it? For the first question, we cannot theoretically prove that it has reached a global-optimal solution. However, to verify the effectiveness of our proposed method, we collect three kinds of images, including images captured by free shooting, UAV images and street view images. According to the experimental results, we can see that our system performs well on many real image datasets. For the second question, the answer is rather tricky. We think although the auxiliary imaging information is not accurate enough, it should nonetheless contain sufficient truthfulness on the imaging condition to act as a reasonable initializer for our potential inliers selection. Note that in our work, a constraint is considered as a potential inlier if its residual at the current iteration is less than an adaptive threshold. By such a setting, a potential inlier is not necessarily meant a real inlier. It is merely meant that it should not be a gross outlier, and its probability being a real inlier is much larger than a potential outlier. In addition, potential inlier is meaningful only at the current iteration because a potential inlier at the current iteration could change to a potential outlier at the next iteration.

By iteratively filtering out the potential outliers, the iterative convergence is accelerated. To some degree, our proposed iterative method possesses some analogy with the well-known Boosting scheme. In Boosting, by iteratively combining weak classifiers, a strong classifier is obtained. In our method, by iteratively filtering potential outliers, potential inliers converge to real inliers with high probability, and the parameters, such as camera poses and 3D scene points, become more and more accurate. Unlike Boosting where the convergence is slow due to the less impact of later weak classifiers, our method is quite computationally efficient, which mainly due to the following two interlinked factors: firstly, only potential inliers are used, which is a subset of the total constraints; secondly, with the iteration going on, the set of the selected potential inliers contains less and less real outliers, and the estimated parameters become closer and closer to the correct ones.

V. Conclusion

In this paper, we propose an efficient large-scale SfM method by fully exploiting auxiliary imaging information. The main novelty of our work is the exclusive use of the so-called potential inliers at each iterative optimization step to effectively deal with the inevitable constraint outliers, which is made possible in turn by employing auxiliary imaging information. Experimental results show that our approach outperforms the state-of-art reconstruction approaches, especially for UAV and StreetView images. In future work, the iterative convergence will be further investigated and we will do more research on how to tackle occasional gross geotags, for example, exploring the consistency between the camera centers and the estimated pairwise translations to refine them, as proposed in [4] and [27].

References


W. Gao received the B.S. degree in computational mathematics and the M.S. degree in pattern recognition and intelligent systems from Shanxi University, in 2002 and 2005, respectively, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, in 2008. He has been with the Robot Vision Group, National Laboratory of Pattern Recognition, Institute of Automation, since 2008, where he is currently an Associate Professor. His current research interests include computer vision, 3D reconstruction, and image processing.

Haihan Cui received the B.S. degree from the South China University of Technology, in 2011. He is currently pursuing the Ph.D. degree with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interest is computer vision, with a particular interest in 3D reconstruction.

Zhanyi Hu received the B.S. degree in automation from the North China University of Technology, Beijing, China, in 1985, and the Ph.D. degree in computer vision from the University of Liege, Belgium, in 1993. Since 1993, he has been with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, where he is currently a Professor. His research interests are in robot vision, which include camera calibration and 3D reconstruction, and vision guided robot navigation. He was the Local Chair of ICCV2005, an Area Chair of ACCV2009, and the PC Chair of ACCV2012.

Shuhan Shen received the B.S. and M.S. degrees from Southwest Jiao Tong University, in 2003 and 2006, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, in 2010. Since 2010, he has been with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, where he is currently an Associate Professor. His research interests are image based 3D modeling.

Wei Gao received the B.S. degree in computational mathematics and the M.S. degree in pattern recognition and intelligent systems from Shanxi University, in 2002 and 2005, respectively, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, in 2008. He has been with the Robot Vision Group, National Laboratory of Pattern Recognition, Institute of Automation, since 2008, where he is currently an Associate Professor. His current research interests include computer vision, 3D reconstruction, and image processing.

Haihan Cui received the B.S. degree from the South China University of Technology, in 2011. He is currently pursuing the Ph.D. degree with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interest is computer vision, with a particular interest in 3D reconstruction.

Zhanyi Hu received the B.S. degree in automation from the North China University of Technology, Beijing, China, in 1985, and the Ph.D. degree in computer vision from the University of Liege, Belgium, in 1993. Since 1993, he has been with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, where he is currently a Professor. His research interests are in robot vision, which include camera calibration and 3D reconstruction, and vision guided robot navigation. He was the Local Chair of ICCV2005, an Area Chair of ACCV2009, and the PC Chair of ACCV2012.