Efficient structure from motion with weak position and orientation priors

Conference Paper · July 2011

DOI: 10.1109/CVPRW.2011.5981775 · Source: IEEE Xplore

Christof Hoppe
Graz University of Technology
21 PUBLICATIONS **424** CITATIONS
SEE PROFILE

Horst Bischof
Graz University of Technology
794 PUBLICATIONS **33,023** CITATIONS
SEE PROFILE

Stefan Kluckner
Siemens Medical Solutions, Inc.
38 PUBLICATIONS **591** CITATIONS
SEE PROFILE

Efficient Structure from Motion with Weak Position and Orientation Priors

Arnold Irschara, Christof Hoppe, Horst Bischof
Graz University of Technology
{irschara, hoppe, bischof }@icg.tugraz.at

Stefan Kluckner
Siemens Corporate Technology, Graz, Austria
stefan.kluckner@siemens.com

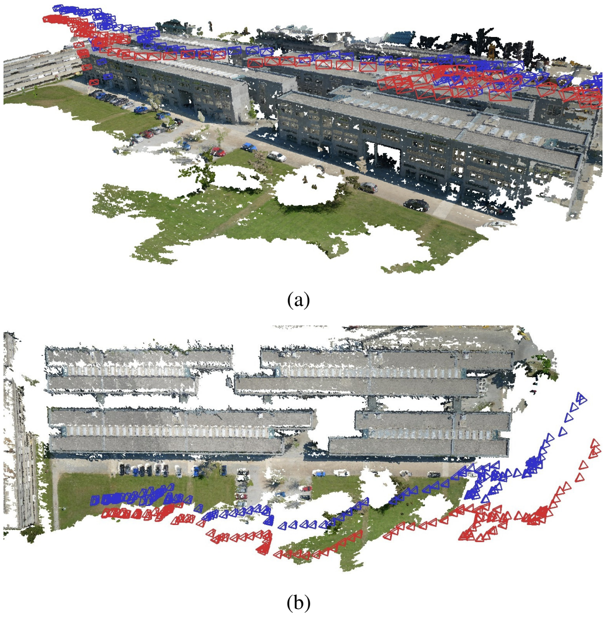
Abstract

In this paper we present an approach that leverages prior information from global positioning systems and inertial measurement units to speedup structure from motion computation. We propose a view selection strategy that advances vocabulary tree based coarse matching by also considering the geometric configuration between weakly oriented images. Furthermore, we introduce a fast and scalable reconstruction approach that relies on global rotation registration and robust bundle adjustment. Real world experiments are performed using data acquired by a micro aerial vehicle attached with GPS/INS sensors. Our proposed algorithm achieves orientation results that are subpixel accurate and the precision is on a par with results from incremental structure from motion approaches. Moreover, the method is scalable and computationally more efficient than previous approaches.

1. Introduction

We observe an ever increasing amount of Global Positioning System (GPS) and Inertial Measuring Units (IMU) attached to camera systems that deliver instant geo-referenced images in a 3D World Geodetic System (WGS84). For instance, reliance on direct geo-referencing is nowadays a standard in aerial photogrammetry [3] and often allows the avoidance of aerial triangulation (AT) and ground control measurements. Recently, Pollefeys et al. [17] demonstrated real-time structure from motion (SfM) and dense matching in urban scenes based on a GPS/IMU supported reconstruction system. These systems rely on highly accurate geo-referening devices that are calibrated and the delivered pose and orientation estimates are often superior to the one obtained by image based methods (i.e. subpixel accurate image registration). Furthermore, relying on absolute orientation information, these systems are not prone to drift and loop closure can be easily handled.

On the contrary, there exists a variety of direct geo-referencing platforms that provide instant, but often inaccurate and imprecise position and orientation information. Examples of such platforms are mobile devices including Smartphones and Micro Aerial Vehicles (MAVs). Unfortunately, GPS/IMU data from such devices does not reach the required level of precision for pixel accurate image alignment that can be used for practical computer vision task such as 3D object reconstruction and navigation. However, it delivers a rough estimate of the camera poses and orientations.


(a)

(b)

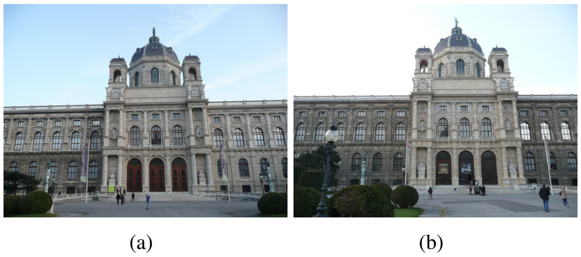
Figure 1. (a) Side view and (b) top view of GPS/IMU camera orientations (bule) and adjusted camera orientations (red) obtained by our global structure from motion approach. Dense reconstruction computed by PMVS [5].

In this paper we present an approach that leverages such imprecise prior information to speedup structure from motion computation in terms of feature matching and geometric estimation. The problem of SfM computation, i.e. recovering the (sparse) structure of a 3D scene and camera orientations has reached maturity over the past years. Current state of the art is represented among others by Pollefeys [18] and Nister [15]. It is nowadays even possible to automatically reconstruct a scene from unorganized and very inhomogeneous image datasets such as community photo collections gathered from the web [20]. The success of these methods is mainly based on substantial progress in wide baseline matching, scale invariant feature detectors [11] and advances in multiple view geometry [22, 14]. However, large scale SfM is computationally demanding and often prone to drift and error accumulation. For non sequential image acquisition (i.e. unordered image collections) and camera networks with loops, exhaustive pairwise feature matching is normally the most time demanding operation of current structure from motion pipelines. If no prior knowledge about camera positions and orientations is available, it requires matching of $O(n^2)$ image pairs. To overcome these limitations, current large scale SfM systems [1] take advantage of a two stage matching approach. First, a coarse view matching based on image retrieval and bag of feature concepts [19] is performed. Second, only image pairs that achieve a high similarity score are used for detailed matching.

Recently, Strecha et al. [21] have demonstrate that city scale 3D reconstruction is feasible by incorporating GPS and geo-tags for image matching. In their approach GPS information is used to partition images based on their locations into manageable and overlapping datasets that can be efficiently reconstructed. This is in contrast to the approach of Carceroni et al. [2] that study the geometric problem of estimating camera orientations from multiple views, given the positions of known camera locations. In this paper we provide a framework that addresses both problems. In particular, we (i) provide a criterion for effective view selection based on GPS/IMU information and (ii) present an efficient method for structure from motion computation.

2. GPS/INS supported matching

Feature matching of unordered images is often the most time consuming part of a SfM algorithm. A typical solution to restrict the number of images for detailed feature matching is to use a coarse matching strategy based on vocabulary tree concepts [16] in order to efficiently determine a subset of potentially matching image pairs. However, due to repetitive structure, the result of vocabulary tree based image retrieval can often be arbitrarily wrong. Images that achieve a high similarity score do not necessarily show the same part of an environment (e.g. the images shown in Figure 2 are visually similar but are taken from two different buildings). Using global pose information as provided by a GPS/IMU unit, such ambiguous images can be filtered which further reduces the matching effort. Given the set of images $\mathcal{I} = I_1, \dots, I_n$, associated with approximate knowledge of external pose measurements $\mathcal{G} = G_1, \dots, G_n$, we select a subset V_i of potentially matching view pairs.



(a) (b)

Figure 2. Visual similar facade images from two different buildings, (a) Museum of Art History and (b) Museum of Natural History, both located in Vienna.

2.1. External Pose Information

The external pose $G_i = [\ R \mid \mathbf{t}\ ]$ is achieved by GPS and IMU, where R is a 3×3 rotation matrix and $\mathbf{t}$ a 3-space vector representing camera orientation and translation, respectively. Global position information is delivered by a standard GPS receiver in the WGS84 coordinate system which describes a position on earth as longitude, latitude and height. For further processing, the GPS datum is transformed in the Earth Centered, Earth Fixed (ECEF) coordinate system, which is a Cartesian system capable of representing reconstructions in global scale. The camera orientation is described by three angles yaw, pitch, and roll, where the yaw angle is aligned to magnetic north. Thus, the external pose G_i is composed of a GPS datum transformed to the ECEF coordinate system and three rotation angles. In conjunction with the known intrinsic parameter K of the camera, we obtain the full projection matrix $\hat{P}_i$ for each image I_i,

$$\hat{P}_i = KG_i = K[\, R \mid \mathbf{t}\,]. \tag{1}$$

The retrieved projection matrices give a rough estimate of the camera position and orientation that is used for further processing.

2.2. View Selection

To identify images that potentially share corresponding points, we select for each image I_i a set $T_i = T_1 \dots T_k$ images that achieve a sufficient high probabilistic similarity score [7]. Next, images in T_i are filtered according to their GPS/IMU information using a coarse overlap criterion. If a detailed 3D model of the environment is available, the image overlap can be easily obtained by projecting and back-projecting the view frustum of view I_i into view I_j. In case that no detailed model of the environment is present, we can only make weak assumptions on the maximum scene depth S_i that restrict the area observed by an image I_i. For instance, given a rough Digital Surface Model (DSM) the height above ground can be estimated which limits the maximal depth range for cameras looking towards the earth-surface (e.g. aerial image surveys). For terrestrial data (i.e. horizontal looking cameras) the S_i can be fixed to a user de-

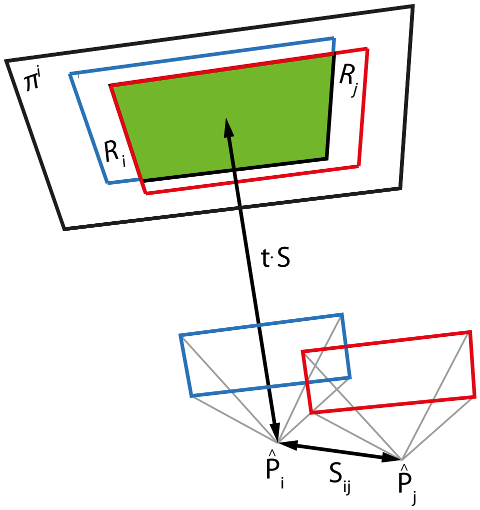


Figure 3. View overlap estimation. In front of camera $\hat{P}_i$, a plane π^i is defined with distance $t \cdot S$. R_i and R_j are the visible areas of π^i in $\hat{P}_i$ and $\hat{P}_j$ respectively. The image overlap is determined according to Equation (4).

fined threshold that is based on the maximal expected scene size. Furthermore, the maximum scene depth S_{ij} that can be recovered by an image pair $< I_i, I_j >$ depends on their baseline. We define,

$$S_{ij} = t \cdot d(G_i, G_j), \tag{2}$$

where $d(\cdot,\cdot)$ denotes the Euclidean distance and t is factor that determines the required reconstruction accuracy. Given these constraints, we estimate the maximum scene depth S that can be reconstructed by an image pair $< I_i, I_j >$,

$$S = min(S_{ij}, S_i, S_j). \tag{3}$$

Furthermore, the images must have an overlap. To calculate a coarse overlap criterion $\mathcal{O}^i_j$, we define a plane π^i that is parallel to image I_i and whose distance to the camera center G_i is S. R_i and R_j denote the image extend of view I_i and I_j on the plane π^i. The image overlap $\mathcal{O}^i_j$ is computed by,

$$\mathcal{O}^i_j = \frac{a(R_i \cap R_j)}{a(R_i \cup R_j)}, \tag{4}$$

where $a(\cdot)$ returns the area of the projected rectangle.

Since feature descriptors like SIFT may tolerate only a maximum view angle change of approx. 30°, we require that the view vector of $\hat{P}_j$ and the normal of π^i enclose a maximum angle α. Otherwise $\mathcal{O}^i_j$ is set to zero. An illustration of the overlap calculation is shown in Figure 3. For every image pair $< I_i, I_j >$ with $I_j \in T_i$ we compute the overlap $\mathcal{O}^i_j$. If the overlap is above a fixed threshold, I_j is inserted into the set V_i which are later used for detailed feature matching. The first eight images of the set T_i and the set V_i for a sample image are depicted in Figure 4. The corresponding GPS positions are shown in Figure 5.

3. Reconstruction Method

Unlike incremental reconstruction methods [20] that start from a small reconstruction which is subsequently expanded by adding more images, we follow a global reconstruction approach that considers all images, simultaneously. In order to determine feature correspondences over multiple views, an epipolar graph is build which encodes two view relations. We use the view selection routine described in the previous section that largely reduces the matching effort for the graph reconstruction. Based on the epipolar graph, connected components are extracted and point tracks over multiple views are generated. The tracks are later used for structure initialization. Next, rotational components of two-view epipolar geometries are consistently registered and aligned with the known GPS positions. Finally, robust bundle adjustment is used to refine camera orientations and 3D structure. A detailed description of the individual processing steps is described in the following sections.

3.1. Building the Epipolar Graph

First of all, scale invariant feature points are extracted from every image. Our method utilizes the very effective SIFT keypoint detector and descriptor [11] which achieves excellent repeatability performance for wide baseline image matching [13]. In particular we rely on the publicly available SiftGPU[1] software. Keypoint correspondences between image pairs are determined by employing a GPU accelerated feature matching approach based on the CUBLAS library with subsequent instructions that apply the distance ratio test.

After matching relevant images V_i to each query view I_i, geometric verification based on the Five-Point algorithm [14] is performed. Since matches that arise from descriptor comparisons are often highly contaminated by outliers, we employ a RANSAC [4] algorithm for robust estimation. The matching output is a graph structure denoted as epipolar graph $\mathcal{EG}$, that consists of the set of vertices $\mathcal{V} = \{I_1 \ldots I_N\}$ corresponding to the images and a set of edges $\mathcal{E} = \{e_{ij} | i,j \in \mathcal{V}\}$ that are pairwise reconstructions, i.e. relative orientations between view i and j, $e_{ij} =< P_i, P_j >$,

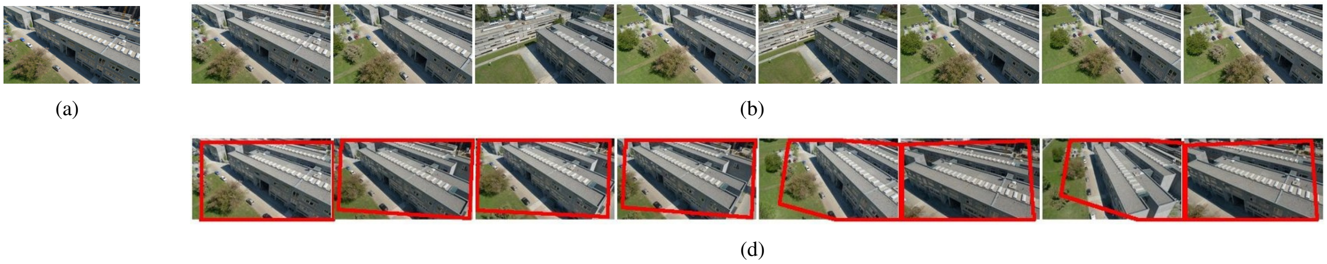
$$P_i = K_i[\,I \mid \mathbf{0}\,] \text{ and } P_j = K_j[\,R \mid \mathbf{t}\,] \tag{5}$$

and a set of triangulated points with respective image measurements.

3.2. Consistent Rotations

Given the epipolar graph $\mathcal{EG}$, the initial camera positions and orientations remains to be determined. First, relative rotations $\{R_{ij}\}$ between view pairs i and j are upgraded into

[1] http://cs.unc.edu/∼ccwu/siftgpu


(a)

(b)

(d)

Figure 4. Vocabulary tree vs. overlap criterion. The first row shows the first eight images returned by the vocabulary tree given the query image (a). The second row shows the first eight images of the filtered vocabulary tree result sorted by their overlap value. The red box depicts which part of the rectangle R_i is visible in the current image.

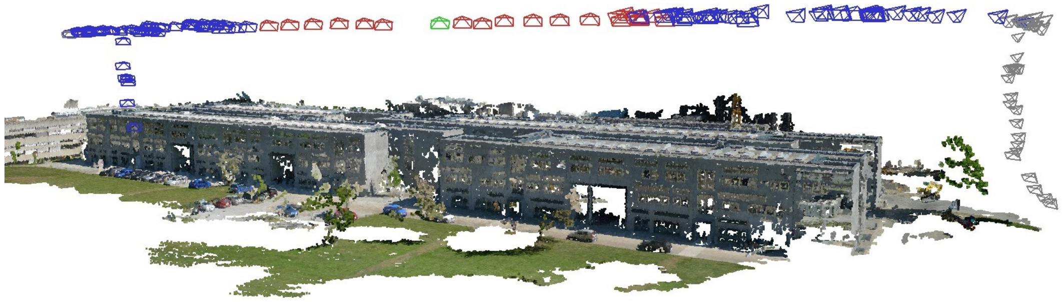


Figure 5. View selection of a trajectory containing 196 images. Absolute pose information G_i is provided by a GPS and IMU (colored pyramids indicate camera positions and orientations). Given a query image I_i (green) results in a set T (blue + red) of images with similar appearance delivered by the vocabulary tree. Images with an overlap value $\mathcal{O}_j^i$ of at least 50% are depicted in red.

a consistent set of rotations $\{R_i\}$ by solving the (overdetermined) system of equations,

$$R_{ij}R_i = R_j \tag{6}$$

subject to the constraint that R_i are orthonormal. As described in [12], the solution can be obtained by solving the system initially for approximate rotation matrices $\hat{R}_i$ (without satisfying the orthonormality constraint) and subsequently projecting the approximate rotation $\hat{R}_i$ to the closest rotation in the Frobenius norm. This is done by using the singular value decomposition (SVD). Equation (6) is normally overdetermined since the epipolar graph consists of a redundant set of relative orientations that contribute to the global structure. Figure 6 shows a typical epipolar graph for aerial data.

Not all epipolar geometries are equally important. In general, relative rotations that are determined by many correspondences can be rated as more confident than orientations that are only supported by a small number of measurements. As suggested in [12], we consider the number of inliers and reweight each row of (6) according to a quality criterion that determines the accuracy of an epipolar geometry e_{ij}. Rather than using the raw number of inliers $|\mathcal{F}_{ij}|$ as suggested in [12], we compute the weights ω_{ij} as follows,

$$\omega_{ij} = \sqrt{N}\min(c_i, c_j) \tag{7}$$

where $N = |\mathcal{F}_{ij}|$ is the number of inliers between view i and j and c_i, c_j is a measure of feature coverage,

$$c_*(\mathcal{F}_{ij}) = \begin{cases} 0 & |\mathcal{F}_{ij}| < \alpha \\ \frac{A(\mathcal{F}_{ij}, r)}{A_\boxdot} & \text{otherwise} \end{cases} \tag{8}$$

where α is a minimal number of required inliers (e.g. $\alpha = 10$ in our experiments), $A_\boxdot$ denotes the total image area and $A(\mathcal{F}_{ij}, r)$ is the resulting area that the feature points $\mathcal{F}_{ij}$ cover after applying a dilation operation with a circular structuring element of radius $r = \sqrt{\frac{A_\boxdot}{|F_{ij}|}}$. In addition to the raw number of inliers that determines the confidence of the relative orientation result, the coverage criterion [7] further takes the spatial distribution of correspondences into account. As a consequence, convergent views that have well distributed correspondences produce a higher score than epipolar pairs with the same number of correspondences but with random point distribution. Hence, system (6) is extended to a reweighted form,

$$\omega_{ij}(R_{ij}\mathbf{r}_i^k - \mathbf{r}_j^k) = \mathbf{0}_{3\times 1} \tag{9}$$

for $k = 1, 2, 3$, where $\mathbf{r}_i^k$ are columns of R_i. The system can be efficiently solved by a sparse least squares solver (e.g. using the ARPACK library). Figure 7 depicts raw GPS/IMU camera orientations $\hat{P}$ and the final result after consistent rotation alignment.

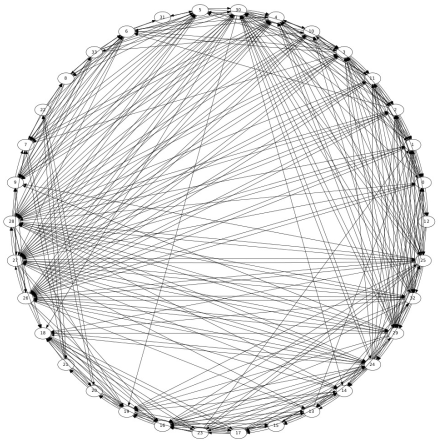


Figure 6. Epipolar graph showing valid relative orientations between view pairs.

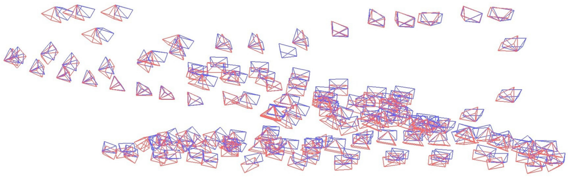


Figure 7. Raw camera positions and orientations from GPS/IMU (blue) and adjusted camera rotations by global rotation registration (red).

3.3. Camera Center Initialization

Given the registered rotation matrices R_i, camera centers are initialized with the GPS datum G_i in the (ECEF) coordinate system as described in Section 2.1. One transformation still remains to be determined, the rotations R_i have to be aligned to fit the GPS path. This is accomplished by aligning relative translations $\mathbf{v}_{ij}$ to corresponding GPS orientations $\hat{\mathbf{v}}_{ij}$,

$$\mathbf{v}_{ij} = R\hat{\mathbf{v}}_{ij} \tag{10}$$

an instance of the well known orthogonal Procrustes problem which can be solved using the singular value decomposition.

3.4. Track Generation

The epipolar graph $\mathcal{EG}$ stores a set of relative orientations and feature correspondences between view pairs $< I_i, I_j >$. Every image I_i is matched to a number of neighboring images and the matching information is stored locally in every node. Note, $\mathcal{EG}$ is a directed graph, a match $I_i \rightarrow I_j$ does not necessarily imply $I_j \leftarrow I_i$. Next, for each image node I_i of the graph, point measurements are aggregated to tracks $m = (< x_1^i, y_1^i >, < x_2^j, y_2^j > \dots, < x_n^k, y_n^k >)$, where $f =< x^k, y^k >$ represent feature locations of image I_k. Since point tracks are generated for each image and stored locally, at first instance, the set of point tracks $m \in \tilde{\mathcal{M}}$ is redundant, i.e. a feature point f from image I_k can take part in different tracks. The point tracks are later used for global optimization in bundle adjustment. From a practical viewpoint, redundant measurements are not desired since it involves more parameters in the optimization framework, hence we are interested in a minimal representation. To this end we determine a subset of tracks $\mathcal{M} \subseteq \tilde{\mathcal{M}}$ that covers every matched feature correspondence of the epipolar graph only once. This is an instance of the set cover problem [9], one of the earliest problems known to be NP-complete. We use a greedy approach [8] to efficiently determine a minimal set of tracks that are subsequently used to initialize the sparse 3D structure.

3.5. 3D Structure Initialization

From the previous processing steps a set of camera orientations $P_{i=1:N} \in \mathcal{P}$ (i.e. calibration and poses) and point tracks $m \in \mathcal{M}$ is obtained. It remains to be determined the coordinates of 3D points $\mathbf{X}_{j=1:M} \in \mathcal{X}$ according to every track. Given the fact that camera orientations $\mathcal{P}$ in general do not offer a pixel accurate alignment and outliers are still present in $\mathcal{M}$, a linear triangulation method based on $\mathcal{P}$ will result in an arbitrary large reconstruction error (i.e. the 3D structure is weakly initialized). In practice, we observed that a direct triangulation approach does not provide a sufficiently high accuracy for structure initialization, often even the cheirality constraint [6] is not satisfied. However, (sub)-pixel accurate camera orientation between view pairs is provided from the epipolar graph $\mathcal{EG}$ that allows accurate two view triangulation. Hence, from every point track we determine the image pair that has a maximal baseline (a large baseline ensures a low relative error of GPS coordinates) and perform a two view triangulation based on these relative orientations. Next, the 3D point is transformed according to the global registered camera positions. This leads to an initial 3D structure that is subsequently optimized by a robust bundle adjustment algorithm taking all measurements of the point tracks into account.

3.6. Robust Bundle Adjustment

Given a set of measurements, bundle adjustment [22, 6, 10] optimizes camera orientations and structure by minimizing the reprojection error,

$$\mathcal{C}(P_i, \mathbf{f}_j) = \sum_i \sum_j v_{ij} d(P_i \mathbf{X}_j, \mathbf{x}_{ij})^2 \tag{11}$$

where the 2D point measurements $\mathbf{x}_{ij}$ are the observations of unknown 3D points $\mathbf{X}_j$ in the unknown cameras P_i and v_{ij} is a binary variable that is 1 if the point $\mathbf{X}_j$ is visible in image P_i and 0 otherwise. In practice, bundle adjustment involves adjusting the bundle of rays between each 3D point and the set of camera centers by minimizing the sum of squares of a large number of nonlinear, real-valued functions. Bundle adjustment is a large, but sparse geometric parameter estimation problem which is tolerant to missing data (i.e. not every 3D point must be visible in each camera). In the case of Gaussian measurement noise, a nonlinear least-squares minimization achieves the Maximum Likelihood estimate (subject to the constraint that the initialization is sufficiently close to the global minimum). However the assumption of Gaussian Noise in image measurements is a strong assumption and is not true for real world structure from motion problems that rely on natural feature matching techniques. In our system feature tracks are geometrically verified by using the epipolar constraint, however mismatches may still occur. Thus least squares is not an appropriate norm and we require a robust cost function to handle outliers. In its basic implementation bundle adjustment minimizes a squared vector norm $\sum_i ||\epsilon||^2$ with $\epsilon = ||\mathbf{x}_i - \hat{\mathbf{x}}_i||$. Hence, the robust cost function can be implemented by re-weighting the error vector $\epsilon_i' = w_i \epsilon_i$ such that,

$$||\epsilon_i'||^2 = w_i^2 ||\epsilon_i|| = C(||\epsilon_i||). \tag{12}$$

Therefore it follows $\sum_i C(||\epsilon_i||) = \sum_i ||\epsilon_i||^2$ as desired where,

$$w_i = \frac{\sqrt{C(||\epsilon_i||)}}{||\epsilon_i||}. \tag{13}$$

The weighting w_i is often called attenuation factor since it seeks to attenuate the cost of the outliers. In our experiments we evaluate the performance of different cost function for the task of bundle adjustment. In particular we use the following cost functions,

- **Squared error cost function**

$$C(\epsilon) = \epsilon^2 \tag{14}$$

- **Huber cost function**

$$C(\epsilon) = \begin{cases} \epsilon^2 \text{ for } |\epsilon| < b \\ 2b|\epsilon| - b^2 \text{ otherwise} \end{cases} \tag{15}$$

- **Blake-Zisserman cost function**

$$C(\epsilon) = \begin{cases} \epsilon^2 \text{ for } |\epsilon| < b \\ b^2 \text{ otherwise} \end{cases} \tag{16}$$

- **Sigma cost function**

$$C(\epsilon) = \begin{cases} \epsilon^2 \text{ for } |\epsilon| < b \\ 2b|\epsilon| - b^2 \text{ for } b < |\epsilon| < \sigma b \\ b^2(2\sigma - 1) \text{ otherwise} \end{cases} \tag{17}$$

- **Cauchy cost function**

$$C(\epsilon) = b^2 \log(1 + \epsilon^2/b^2) \tag{18}$$

Graphs corresponing to the individual cost functions are depicted in Figure 8.

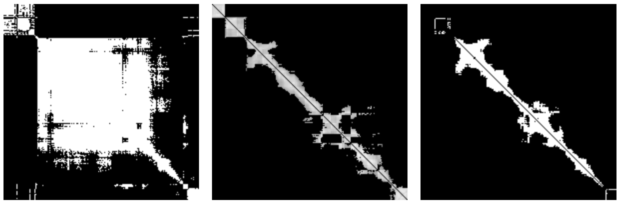


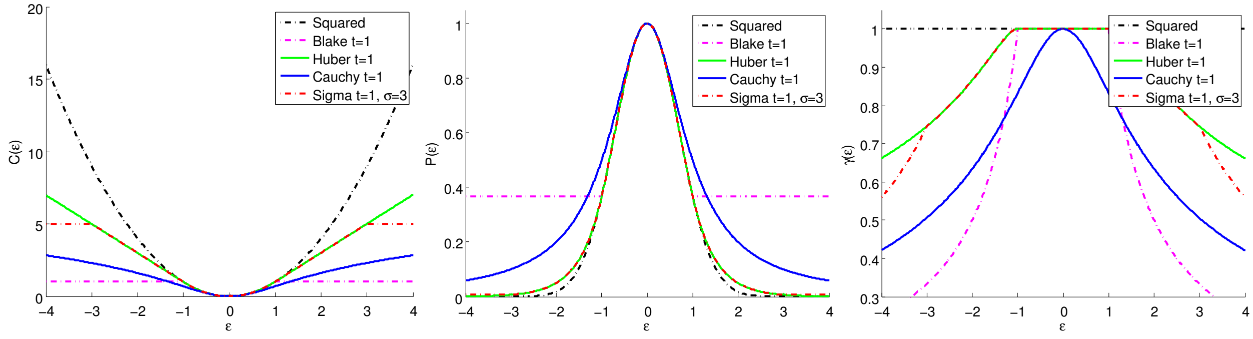
Figure 9. (a) Adjacency matrix showing potential matching candidates (bright pixel) between view i,j from a vocabulary tree scoring and (b) potentially matching image pairs as computed by our geometric view selection strategy. (c) Epipolar geometries determined by exhaustive image matching representing the ground truth.

4. Experiments

We evaluate our reconstruction approach on real world data acquired by an unmanned aerial vehicle[2] with integrated GPS/IMU sensors. The attached camera system delivers images of resolution 3968×2232, the accuracy of the GPS receiver is about $2m$ and the relative position precision approximately $0.5m$. We assume that the camera origin and GPS/IMU sensor shares the same 3D position in space. This is a valid approximation if the distance to the scene is much larger than the offsets between the individual sensors. Images are acquired at an interval of 3s and a GPS/IMU tag is stored for each image. During one flight mission of 10-15 minutes, about 200 images are acquired. For each input image we use a maximal number of d neighbors from the GPS view selection strategy (as described in Section 2) and compute the epipolar graph (we use $d = 20$). Our view selection approach reduces the matching effort from $O(n^2)$ to $O(nd)$. Compared to exhaustive image matching that considers $\sim$ 40000 view pairs, only 4000 matching operations are performed which gives a ten-fold speedup. Figure 4 shows potential matching candidates from a vocabulary tree approach and matching pairs computed by our proposed geometric view selection criterion.

We use the method described in Section 3.2 to initialize camera orientations and 3D points. Next, bundle adjustment is executed to minimize the reprojection error. Different cost functions are evaluate with respect to the final reprojection error. The results are summarized in Table 2. While robust cost functions (e.g. Huber, Blake-Zisserman, Sigma,Cauchy) achieve comparable performance in terms of average and median reprojection error, the Squared Error cost function leads to a wrong reconstruction, i.e. bundle adjustment does not converge to a reasonable geometric solution. This can be seen from Figure 10(b) and Figure 10(d), respectively. Note, consistent rotation registration (see Section 3.2) is an important processing step. Through our experiments, robust bundle adjustment was not able to con-

[2] AscTec Falcon 8 http://www.asctec.de



(a)

(b)

(c)

Figure 8. Comparison of different cost functions $C(\epsilon)$ dependent on the measurement error ϵ. (a) Cost functions $C(\epsilon)$, (b) corresponding PDFs and (c) attenuation factors.

verge to a true solution from raw IMU initialized projection matrices $\hat{P}_i$. While the Root Mean Squared Error (RMSE) between global registered rotation matrices and the final optimized bundle adjustment result is about $0.1°$, the RMSE of the initial IMU orientation is on average more than $10°$.

We compare the reconstruction results of our proposed approach to the one obtained by an incremental structure from motion pipeline. After aligning both results with a robust similarity transform in a metric coordinate system, the mean as well as the median residual error between camera centers is about $0.023m$. On average, the deviation between the view vectors of the reconstructed camera positions is $0.03°$, only. Both values show that our approach does not differ to an incremental SfM approach according to accuracy and the qualitative results are also equivalent as depicted in Figure 11. Moreover, our approach is computationally more efficient. Table 1 gives detailed timings of our system.

While an incremental structure from motion pipeline (e.g. the bundler software[3]) requires a repeated call of a bundle adjustment optimizer (with time complexity O(n^3), where n is the number of frames), our proposed algorithm only requires rotation an track initialization and one single bundle adjustment call.

Operation	time [s]
SIFT (3968 × 2232 pixel)	0.4
Coarse Matching	0.05
Matching (5000 × 5000)	$d \times 0.044$
RANSAC (5-pt, N=2000)	$d \times 0.12$
Incremental SfM (200 views)	1800
Our approach (200 views)	190

Table 1. Timings for individual processing steps (per image) and comparison to an incremental structure from motion approach. Note, d reflects the number of considered images for detailed feature matching and geometric verification.

Cost function	$\epsilon_{avg.}$	ϵ_{median}	inliers [%]
Before bundle	403.56	312.67	0.13
Squared error	11.46	4.72	71.89
Huber	4.015	0.724	98.24
Blake-Zisserman	4.56	0.66	98.20
Sigma	4.51	0.677	98.52
Cauchy	4.592	0.662	98.46

Table 2. Evaluation of bundle adjustment with respect to different cost functions for a fixed number of 150 iterations. While robust cost functions (Huber, Blake-Zisserman, Sigma, Cauchy) achieve comparable results in terms of average ($\epsilon_{avg.}$) and median (ϵ_{median}) reprojection errors and detect a comparable fraction of inliers (we denote a measurement as inlier if the reprojection error is below 3 pixel), the squared error cost function does not properly converge and the minimization fails (this is shown in Figure 10(d)).

5. Conclusion

In this paper we introduced a reconstruction algorithm that effectively takes advantage of GPS/IMU information as a weak prior to speedup structure from motion computation. The main contributions of the proposed method are (i) a view selection strategy based on global position/orientation information that limits the matching effort and (ii) a fast and scalable reconstruction approach that relies on global rotation registration and robust bundle adjustment. We tested our approach on real world scenarios using data from an unmanned aerial vehicle. From our experiments we conclude that our proposed approach is robust and scalable. Furthermore our approach is computationally more efficient than previous methods and achieves reconstruction results that are on a par with state of the art SfM approaches.

Acknowledgements

This work was supported by the Austrian Research Promotion Agency (FFG) FIT-IT projects Construct (830035), Holistic (830044) and Pegasus (825841).

[3] http://phototour.cs.washington.edu/bundler/

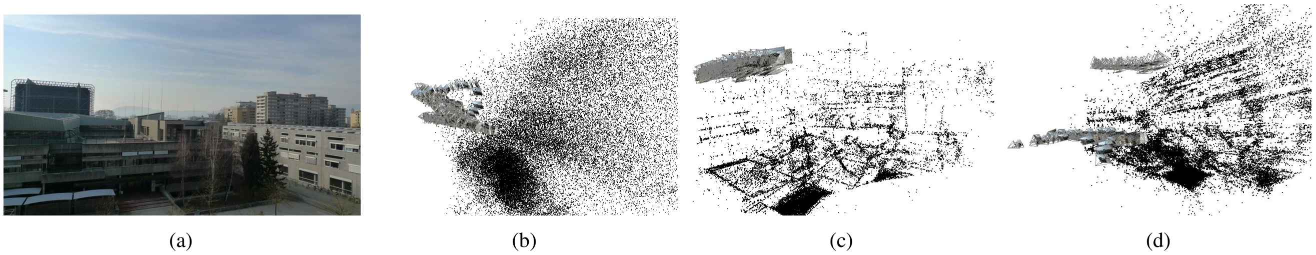

(a) (b) (c) (d)

Figure 10. (a) Sample image of the scene and (b) initial camera orientation and 3D structure by our proposed approach before bundle adjustment. Reconstruction result after bundle adjustment using the robust Cauchy (c) and non-robust Squared Error (d) cost function. Note, while in (c) a true geometric configuration is found, the Squared Error cost function leads to a wrong geometric configuration (d).

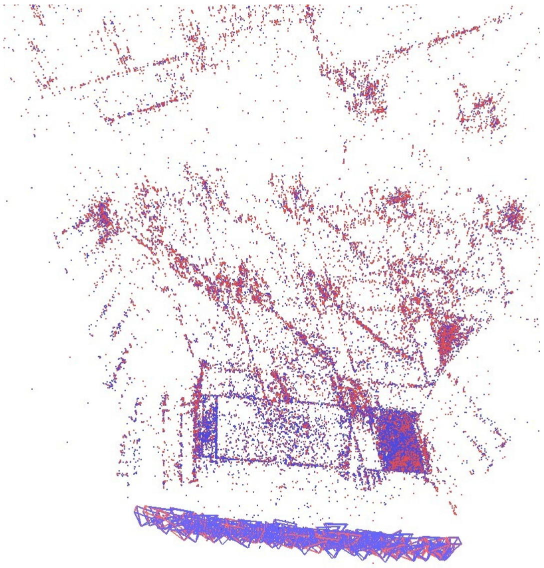


Figure 11. Top view of the reconstruction result obtained by an incremental SfM approach (red) registered to the Cauchy optimized bundle adjustment result (blue) using our proposed approach.

References

[1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building Rome in a day. In *Proc. ICCV*, 2009.

[2] R. L. Carceroni, A. Kumar, and K. Daniilidis. Structure from motion with known camera positions. In *Proc. CVPR*, pages I: 477–484, 2006.

[3] M. Cramer. Experiences on operational gps/inertial system calibration in aiborne photogrammetry. *Journal GIS - Geoinformationssysteme*, pages 37–42, 2002.

[4] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. *Communication Association and Computing Machine*, 24(6):381–395, 1981.

[5] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *PAMI*, 2009.

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

[7] A. Irschara, C. Zach, J. M. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. In *Proc. CVPR*, pages 2599–2606, 2009.

[8] D. S. Johnson. Approximation algorithms for combinatorial problems. *J. of Comput. System Sci.*, 9:256–278, 1974.

[9] R. M. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, pages 85–103. Plenum Press, NY, 1972.

[10] M. A. Lourakis and A. Argyros. SBA: A Software Package for Generic Sparse Bundle Adjustment. *ACM Trans. Math. Software*, 36(1):1–30, 2009.

[11] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[12] D. Martinec and T. Pajdla. Robust rotation and translation estimation in multiview reconstruction. In *Proc. CVPR*, 2007.

[13] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *IJCV*, 65:43–72, 2005.

[14] D. Nistér. An efficient solution to the five-point relative pose problem. *PAMI*, 26(6):756–770, 2004.

[15] D. Nistér, O. Naroditsky, and J. Bergen. Visual odometry. In *Proc. CVPR*, pages 652–659, 2004.

[16] D. Nistér and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proc. CVPR*, pages 2161–2168, 2006.

[17] M. Pollefeys, D. Nister, J. M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S. J. Kim, P. Merrell, C. Salmi, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewenius, R. Yang, G. Welch, and H. Towles. Detailed real-time urban 3D reconstruction from video. *IJCV*, 78(2-3):143–167, July 2008.

[18] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch. Visual modeling with a hand-held camera. *IJCV*, 59(3):207–232, 2004.

[19] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, pages 1470–1477, 2003.

[20] N. Snavely, S. M. Seitz, and R. S. Szeliski. Modeling the world from internet photo collections. *IJCV*, 80(2):189–210, Nov. 2008.

[21] C. Strecha, T. Pylvänäinen, and P. Fua. Dynamic and scalable large scale image reconstruction. In *Proc. CVPR*, pages 406–413. IEEE, 2010.

[22] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment – A modern synthesis. In *Vision Algorithms: Theory and Practice*, pages 298–375. 2000.