Fast Structure from Motion for Sequential and Wide Area Motion Imagery

Hadi AliAkbarpour ,* hd.akbarpour@gmail.com
Kannappan Palaniappan,* palaniappank@missouri.edu
Guna Seetharaman † guna@ieee.org

Abstract

We present a fast and efficient Structure-from-Motion (SfM) pipeline for refinement of camera parameters and 3D scene reconstruction given initial noisy camera metadata measurements. Examples include aerial Wide Area Motion Imagery (WAMI) which is typically acquired in a circular trajectory and other sequentially ordered multiview stereo imagery like Middlebury [46], Fountain [50] or body-worn videos [27]. Image frames are assumed (partially) ordered with approximate camera position and orientation information available from (imprecise) IMU and GPS sensors. In the proposed BA4S pipeline the noisy camera parameters or poses are directly used in a fast Bundle Adjustment (BA) optimization. Since the sequential ordering of the cameras is known, consecutive frame-to-frame matching is used to find a set of feature correspondences for the triangulation step of SfM. These putative correspondences are directly used in the BA optimization without any early-stage filtering (i.e. no RANSAC) using a statistical robust error function based on co-visibility, to deal with outliers (mismatches), which significantly speeds up our SfM pipeline by more than 100 times compared to VisualSfM.

1. Introduction

The use of Global Positioning System (GPS) and Inertial Measurement Unit (IMU) sensors to track the 3D path of platforms and cameras is becoming more widely available and is routinely used in aerial navigation and imaging [10]. The camera path and pose information is used to support robust, real-time 3D scene reconstruction using Structure-from-Motion algorithms (SfM) [24, 27, 43, 45, 6, 5, 7] and direct geo-referencing. Irschara *et al*. in [24] observe that, "These systems rely on highly accurate geo-referencing 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)". However, many (inexpensive) aerial platforms produce IMU and GPS values of limited accuracy due to measurement and timing errors which then need to be refined for accurate SfM [24]. Extracting and incorporating 3D information in WAMI processing [39, 11] will be very useful for mitigating parallax effects in video summarization [52], better stabilization and appearance models for tracking [38], depth map filtering of motion detections [53], and improving video analytics like object tracking [41, 40].

For robust computer vision tasks and accurate aerial photogrammetry, it is essential to refine the camera poses[1]. Bundle Adjustment (BA) is the most popular first stage and a gold standard [49, 32] for obtaining precise camera poses. BA uses initial estimates of camera poses to improve the metadata by minimizing reprojection errors [14, 33, 51]. But is computationally expensive requiring $O((N_c+N_{3D})^3)$ operations for N_c cameras or views and N_{3D} structure, 3D scene feature or tie points (cubic in time and quadratic in memory). Conventional BA shows satisfactory convergence if sufficiently accurate initial estimates are provided either from image feature matching-based essential matrix estimation [23, 1] or combined with on-board sensor measurements [29, 15, 45].

In this paper we propose a new pipeline which leverages weak pose and path information available from imprecise IMU and GPS measurements to both speedup the camera pose refinement process and make it more robust. Figure 1 shows both the conventional SfM and the proposed fast Bundle Adjustment for Sequential imagery (BA4S) pipelines. Contributions of this paper include:

1. We show that weak camera parameters provided by inaccurate sensors on airborne platforms can be directly used for BA as initial values (and not as extra constraints [29, 15, 45]), provided that a proper robust function is used. It will be shown that there is no need to apply a camera estimation method (e.g. Five-Point algorithm [36]). Or to apply filtering methods such as Extended Kalman Filter (EKF) [29, 15] before using noisy sensor measurements in the optimization step.

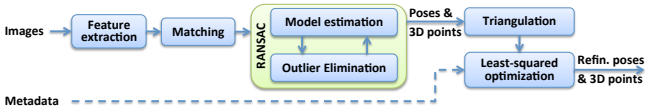
2. We demonstrate that the putative feature correspondences obtained from a sequential matching paradigm can be directly used as observations and initial 3D points for BA optimization. There is no need to filter outliers from the set of putative matches prior to optimization. Specifically, we bypass RANSAC and other combinatorial outlier filtering methods.

*Dept. of Computer Science, Univ. of Missouri, Columbia, MO, USA
†Adv. Computing Concepts, Naval Research Lab., Washing. DC, USA

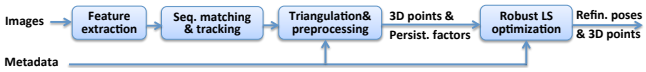
[1] In photogrammetry camera pose is also known as *exterior orientation*.

3. We show that BA4S is robust in: (1) dealing with imprecise and noisy camera parameters due to inaccurate GPS and IMU sensors, and (2) using the set of all feature correspondences without explicit outlier filtering.

4. BA4S uses a new adaptive robust error function based on weighted feature track-length to mitigate the influence of outliers for fast BA optimization. Each residual weight is based on its feature track to population statistics using a novel 3D feature *co-visibility* measure.



(a) Conventional SfM pipeline where camera poses and outliers are simultaneously estimated using RANSAC. Metadata maybe used as extra constraints in optimization.



(b) Proposed SfM pipeline (BA4S) where camera metadata is used directly. There is no model estimation, explicit outlier elimination or RANSAC filtering of mismatches.

Figure 1: Conventional versus our proposed BA4S SfM pipelines.

Background

In the computer vision community the camera parameters are known as *intrinsic* and *extrinsic*, while in photogrammetry, these same metadata are referred to as *interior* and *exterior* parameters and the estimation process is referred to as *resectioning*. Precise estimates of these parameters are critical for practical computer vision applications particularly dense 3D reconstruction. BA, considered as the gold standard for refinement of camera metadata [49, 32, 21], is a classical and well studied problem in computer vision and photogrammetry dating back more than three decades [30, 51, 21]. A comprehensive introduction to BA in [51] covers a wide spectrum of methods and issues.

Due to the recent surge in developing large scale 3D reconstructions using internet photos, smartphones as well as aerial imagery, there has been renewed interest in making BA more robust, scalable and accurate [3, 25, 28, 23, 47]. Recent methods include Sparse BA [31, 26, 13], Incremental BA [29] and Parallel BA [56, 55]. Several optimization methods for BA are compared in [25] and the conjugate gradient approach is shown to produce better results in terms of speed and convergence. In [45] a SfM system using low-resolution images taken by micro-aerial vehicles (MAVs) is described. In [24], the authors use a view selection strategy to speedup SfM but had limited success using robust BA, "Through our experiments, robust bundle adjustment was not able to converge to a true solution from raw IMU initialized projection matrices". Although a robust BA was tested their approach was not sufficient to improve the camera parameters when raw metadata was used. In our work, we successfully used robust BA to refine the cameras due to a different SfM pipeline and camera motion model that enabled inaccurate sensor measurements to be directly used as initial values for BA and without any early-stage feature mismatch filtering (i.e. no RANSAC or its variants).

Recently multi-view stereo techniques have been successfully used in large scale 3D scene reconstruction. For example, 3D reconstruction from large collections of consumer cameras has enabled visualization of city-scale models and photo tourism [1, 48]. In aerial imagery, similar photo tourism techniques has been adapted [37] and true volumetric approaches [42] have shown promising 3D reconstructions. The topic of dense matching in oblique multi-camera systems is discussed in [44]. Semi-global matching in airborne video sequences is discussed in [22, 19].

Many approaches to use GPS and IMU measurements for refining camera parameters have been proposed, especially in the robotics community. However, GPS and IMU measurements have been used mostly as ancillary information along with other pose estimation methods through the essential matrix (e.g. Five-Point algorithm) in computer vision [29, 15] or resectioning in photogrammetry. For example, in [15, 29, 43, 12], platform and sensor GPS and IMU measurements are fused with an SfM approach using an Extended Kalman Filter or as extra constraints in BA in order to produce 3D reconstructions and not directly as in our proposed robust BA4S approach.

2. Building Feature Tracks

In sequential image capture, we know which frames are adjacent to each other, as in persistent aerial WAMI [39] or hyper-lapse first person videos [27]. By leveraging this powerful temporal consistency constraint between images as prior information, we reduce the time complexity of matching, N_c cameras, from $O(N_c^2)$ to $O(N_c)$, without compromising the quality of BA results [45]. In our proposed approach, interest points are extracted from each image using a robust local feature detector. Starting from the first frame, for each two successive image frames, the descriptors associated with interest points are compared with successive matches building up a set of *feature tracks* without using RANSAC. A feature track provides evidence that a potentially unique 3D point in the scene has been observed in a consecutive set of image frames.

In our approach, along with sequential feature tracking, we compute a persistency factor, γ_j, that measures the temporal *co-visibility* of the j-th 3D point (length of a trajectory) in the image sequence. Temporal co-visibility was used in the literature for other purposes such as object recognition [20]. Here we exploit it as a robustness parameter reflecting the reliability in identifying a 3D scene point. Each track (i.e. estimated 3D feature point trajectory) has an associated persistency factor. After building all tracks, the populations statistics of track persistency factor for, N_{3D}, 3D points are estimated including the mean, $\mu_F = \frac{1}{N_{3D}} \sum_{j=1}^{N_{3D}} \gamma_j$ and standard deviation, σ_F. These first and second order track

persistency statistics are used to appropriately weight each track in a novel manner in the BA optimization formulation.

3. Bundle Adjustment for Sequential Imagery

3.1. BA Formulation

Bundle Adjustment (BA) refers to the problem of jointly refining camera parameters and 3D structure in an optimal manner often using reprojection error as the quality metric. Given a set of N_c cameras, with arbitray poses (translations, orientations) and, N_{3D}, points BA optimization is defined as minimizing the sum-of-squared reprojection errors:

$$E = \min_{\mathbf{R}_i,\mathbf{t}_i,\mathbf{X}_j} \sum_{i=1}^{N_c} \sum_{j=1}^{N_{3D}} \|\mathbf{x}_{ji} - g(\mathbf{X}_j, \mathbf{R}_i, \mathbf{t}_i, \mathbf{K}_i)\|^2 \qquad (1)$$

where $\mathbf{R}_i$, $\mathbf{t}_i$, $\mathbf{K}_i$ are respectively the rotation matrix, translation vector and (intrinsic) calibration matrix of the ith camera, $\mathbf{X}_j$ is the j-th 3D point in the scene and observation $\mathbf{x}_{ji}$ is the 2D image coordinates of feature $\mathbf{X}_j$ in camera i and the L_2 norm is used. The mapping $g(\mathbf{X}_j, \mathbf{R}_i, \mathbf{t}_i, \mathbf{K}_i)$ is a transformation model which projects a 3D point $\mathbf{X}_j$ onto the image plane of camera i using its extrinsic, $\mathbf{R}_i$ and $\mathbf{t}_i$, and intrinsic parameters, $\mathbf{K}_i$, defined as:

$$g(\mathbf{X}_j, \mathbf{R}_i, \mathbf{t}_i, \mathbf{K}_i) \sim \mathbf{P}_i \mathbf{X}_{ji} \qquad (2)$$

where $\mathbf{P}_i$ is the projection matrix of camera i, defined as:

$$\mathbf{P}_i = \mathbf{K}_i \left[\mathbf{R}_i | \mathbf{t}_i\right]. \qquad (3)$$

Due to errors in the metadata and feature matching outliers, $g(\mathbf{X}_j, \mathbf{R}_i, \mathbf{t}_i, \mathbf{K}_i) \neq \mathbf{x}_{ji}$, and the minimization problem seeks a statistically optimal estimate for the camera matrices $\mathbf{P}_i$ and 3D feature points $\mathbf{X}_j$. The L_2 minimization of the reprojection error involves adjusting the bundle of rays between each camera center and the set of 3D points which is a non-linear constrained optimization problem. Note that the above minimization is equivalent to finding a maximum likelihood solution assuming that the measurement noise is Gaussian; see [51, 21] for more details. There exist various methods to solve the non-linear least squares problem including implicit trust region and Levenberg-Marquardt methods that are well established in the BA literature [31, 25].

3.2. Adaptive Robust Error Function

The selection of 2D feature point correspondences is one of the most critical steps in image-based multi-view reconstruction [4]. Feature correspondences are usually contaminated by outliers, that is matching errors or incorrect data associations. The outliers or mismatches may be caused by occlusions, repetitive patterns, illumination changes, shadows, image noise and blur for which the assumptions of the feature detector and descriptors are not satisfied [16]. On the other hand, BA which is usually solved using the Levenberg-Marquardt numerical method [51] is highly sensitive to the presence of feature correspondence outliers [4]. Mismatches can cause problems for the standard least squares approach; as stressed in [9] even a single mismatch can globally affect the result. This leads to sub-optimal parameter estimation, and in the worst case a feasible solution is not found [4, 35]. This is even more problematic in high resolution images that have a large number of features and potential correspondences which increases the probability of association or matching errors. Furthermore, aerial images have a high degree of parallax making matching and feature tracking a much more difficult problem.

Generally the mismatches are explicitly excluded from the set of potential feature correspondence in the early stages of the conventional SfM pipeline (Figure 1a) well before the BA optimization stage. In this approach the initial camera parameters are simultaneously estimated while explicitly detecting and eliminating outliers usually by applying different variations of RANSAC. In our proposed SfM approach (Figure 1b), we show that we can bypass the explicit RANSAC-based outlier elimination step by using an appropriate robust error measure. Robust error functions also known as M-estimators are popular in robust statistics and reduce the influence of outliers in estimation problems. We have observed that not every choice of a robust function works well [8] and a proper robust function is critical to achieve a robust minimization of the reprojection error when the initial parameters are too noisy and outliers are not explicitly eliminated beforehand.

The following novel robust function is proposed which uses the *weighted persistency factor* of each feature track, number of consecutive observations compared to the set of all tracks, to reduce the effects of outliers in the optimization error metric:

$$\rho_{ji}(s_{ji}, \gamma_j, \mu_F, \sigma_F) = \left(\frac{\gamma_j}{\mu_F + \sigma_F}\right)^2 \log\left(1 + \left(\frac{\mu_F + \sigma_F}{\gamma_j}\right)^2 s_{ji}^2\right) \qquad (4)$$

where $s_{ji} = \|\mathbf{x}_{ji} - g(\mathbf{X}_j, \mathbf{R}_i, \mathbf{t}_i, \mathbf{K}_i)\|^2$ denotes the residual of the j-th 3D point in the i-th camera (i.e. feature track), γ_j stands for the persistency factor related to j-th 3D point, and μ_F and σ_F are the mean and standard deviation of the persistency factor, respectively, for the population of feature tracks. Substituting (4) into (1) we obtain a new robust error function which leads to the global minimization:

$$E_{BA4S} = \min_{\mathbf{R}_i,\mathbf{t}_i,\mathbf{X}_j} \sum_{i=1}^{N_c} \sum_{j=1}^{N_{3D}} \Big\{ \left(\frac{\gamma_j}{\mu_F + \sigma_F}\right)^2 \cdot \log\left(1 + \left(\frac{\mu_F + \sigma_F}{\gamma_j}\right)^2 \|\mathbf{x}_{ji} - g(\mathbf{X}_j, \mathbf{R}_i, \mathbf{t}_i, \mathbf{K}_i)\|^2\right) \Big\}. \qquad (5)$$

The proposed robust function is inspired by the Cauchy or Lorentzian robust function [24, 51] which has an influence function very similar to the Geman-McClure robust function [34] that decreases rapidly reducing the effect of large outlier values. The residuals, s_{ij} associated with each feature track

Dataset specification			BA4S		Time: BA4S						Time: VisualSfM	
					Per stage				Total			
Dataset	**Image size**	N_c	N_o	N_{3D}	**Iter.**	**Tracking**	**Triangulation**	**Optim.**	**Whole sequence**	**Per image**	**Whole sequence**	**Per image**
Four hills	6048×4032	100	262,828	80,661	36	42 s	8 s	16 s	66 s (∼3 m)	0.66 s	36 m	21.6 s
Columbia (subset)	6600×4400	202	655,593	115,897	10	235 s	13 s	15 s	263 s (∼4 m)	1.3 s	*NA*	*NA*
Albuquerque (subset)	6600×4400	215	668,000	141,559	30	223 s	15 s	35 s	273 s (<5 m)	1.27 s	265 m(> 4h)	73.95 s
Berkeley	6600×4400	220	683,123	138,743	24	185 s	16 s	43 s	244 s (∼4 m)	1.11 s	280 m (> 4.5 h)	76.37 s
LA	6600×4400	351	1,115,603	207,391	10	230 s	23 s	39 s	292 s (<5 m)	0.83 s	485 m (∼ 7 h)	78.29 s
Coit Tower (SanFrans.)	6600×4400	629	2,059,711	344,923	10	370 s	40 s	93 s	503 s (<9 m)	0.79 s	*NA*	*NA*
Albuquerque (full)	6600×4400	1,071	3,473,122	603,119	30	467 s	63 s	222 s	752 s (<13 m)	0.7 s	1,596 m (> 26h)	85.37 s
Columbia-II	6600×4400	5,322	17,437,897	2,509,670	10	2329 s	270 s	521 s	3140 s (∼52 m)	0.59 s	*NA*	*NA*

Table 1: Datasets specifications and timings for individual processing steps (per image) and overall with comparison to an incremental structure from motion approach. N_c, N_o and N_{3D} stand for 'number of cameras/images', 'number of observations (2D points)' and 'number of feature tracks (3D points)', respectively. The speed performance of each dataset is presented per stage for BA4S. The total taken time and per image speeds are presented for both BA4S and for VisualSFM as an incremental SfM algorithm.

are weighted adaptively, with longer lived feature tracks being favored (larger γ_j) over residuals with shorter length feature tracks (smaller γ_j). So a larger persistency weight favors longer co-visibility features that are more likely to represent the same real 3D structure point in the scene. The proposed weighted persistency factor using a modified Cauchy robust function in (4) performed the best compared to the standard Cauchy or Huber robust functions *without persistency* [8].

4. Experiments on Real Datasets

In this section we evaluate the proposed BA4S SfM pipeline. The sample aerial WAMI data were collected (by Transparent Sky) using an aircraft with on-board pose sensors flying over five different urban areas including downtown Albuquerque, NM, Four Hills, NM, Columbia, MO, Los Angeles, CA and Berkeley, CA. In addition to sequential aerial imagery datasets, the BA4S pipeline has been tested on several publicly available vision benchmark datasets with multiview imagery acquired in a sequential and circular trajectory, including *Dino* from the Middlebury MVS evaluation project [46]. As discussed in [17, 4], *Dino* is a very interesting dataset since the object lacks salient features and is a challenging example to test the BA4S pipeline. Fountain-P11 from EPFL [50] is another dataset with eleven images taken from side views of a wall-attached fountain.

4.1. Evaluation Methods

In aerial WAMI, it is not always practical to provide a quantitative evaluation of the results due to a lack of available 3D ground truth which is both expensive and difficult to collect [17]. Generally, reprojection error is commonly used for evaluating SfM results. However, in our pipeline the standard L_2 reprojection error is not an appropriate measure for evaluation since outliers have not been eliminated. That is all spurious scene points as well as valid 3D points across all cameras will contribute to the reprojection error while our primary objective is to recover accurate camera pose.

Instead, we introduce a new pixel-based error measure to evaluate the SfM results referred to as the Euclidean Epipolar Error (EEE) which uses the classical epipolar constraint to evaluate the quality of the refined camera parameters on the image plane. We also include a few additional error measures to compare the camera poses in 3D. We also consider a qualitative assessment of the refined camera metadata by recovering dense multiview 3D scene using PMVS ([18]).

4.1.1 Euclidean Epipolar Error (EEE) Measure

We generate image-based manual groundtruth, N_g feature tracks, for each WAMI dataset (typically N_g = 11 to 50). Given reference camera, l, then for each possible camera pair (l, m), the fundamental (transformation) matrix is directly computed using extrinsic parameters (not estimated) as:

$$\mathbf{F}_{lm} = \mathbf{K}_m^{-T}\,\mathbf{R}_l\,\mathbf{R}_m^T\;\mathbf{K}_l'\;skew(\mathbf{K}_m\,\mathbf{R}_m\,\mathbf{R}_l^T\,(\mathbf{t}_m - \mathbf{R}_l\,\mathbf{R}_m^T\,\mathbf{t}_l)) \tag{6}$$

For the k-th 3D groundtruth points, $\mathbf{g}_{km}$, with $k = 1 \ldots N_g$, each projected into camera, m, its corresponding epipolar line is computed and plotted in the reference image l. The sum of the perpendicular Euclidean distances between the each epipolar line and its associated groundtruth point, averaged over all points, is used as the error measure between the camera pair:

$$\epsilon_{lm} = \frac{1}{N_g}\sum_{k=1}^{N_g} d(\mathbf{g}_{km}\,,\,\mathbf{F}_{lm}\,\mathbf{g}_{kl}) \tag{7}$$

This error is computed over all possible pairs of cameras in the sequence, $\{(l,m)|\; l,m = 1..N_c\}$. Ideally the ϵ_{lm} should be zero due to the fundamental geometric constraint. However, the triple product $\mathbf{g}_{km}^T\,\mathbf{F}_{lm}\,\mathbf{g}_{kl} \neq 0$, in real scenarios due to errors in either point correspondences or camera parameters. The errors ϵ_{lm} can be treated as a matrix and visualized using colored picture elements (pels). In addition to computing, ϵ_{lm}, between cameras, l and m, the mean μ_ϵ and standard deviation σ_ϵ of the error over all cameras is:

$$\mu_\epsilon = \frac{1}{N_c^2}\sum_{l=1}^{N_c}\sum_{m=1}^{N_c}\epsilon_{lm} \tag{8}$$

$$\sigma_\epsilon = \left(\frac{1}{N_c^2}\sum_{l=1}^{N_c}\sum_{m=1}^{N_c}(\epsilon_{lm} - \mu_\epsilon)^2\right)^{1/2}. \tag{9}$$

4.2. Implementation

The BA4S pipeline was implemented in C++. The computer used for experiments was a server with CPU Intel Xeon

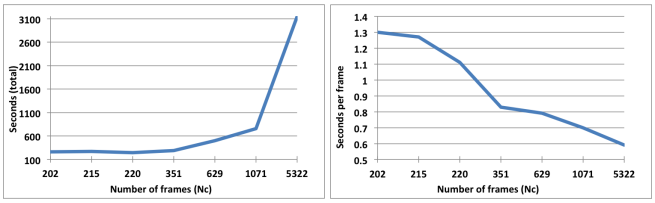
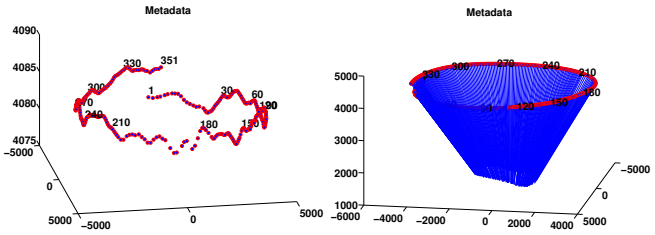
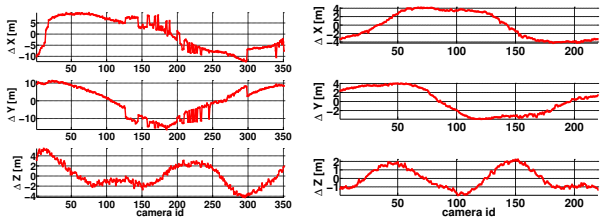


Figure 2: Timing performance for BA4S corresponding to Table 1 showing total time (left) and per frame (right); note non-linear horizontal scale. The time complexity is linear in the number of frames. The per-frame timing decreases as the number of frames increases which is very promising for large scale aerial imagery applications.



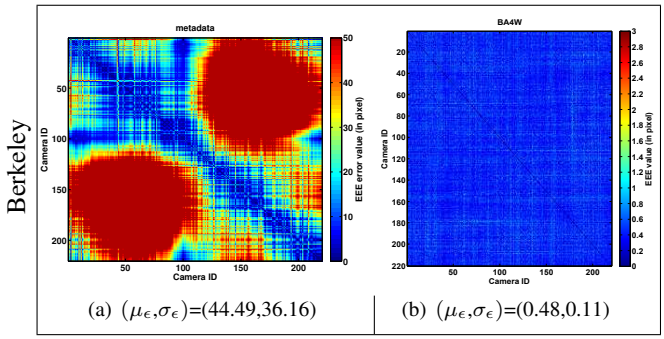
(a) Positions (b) Position and look directions

Figure 3: Camera trajectories corresponding to the LA WAMI dataset.



(a) LA (b) Berkeley

Figure 4: Difference between camera positions of metadata and BA4S output. They basically indicate how much the camera positions have been corrected after BA.



(a) $(\mu_\epsilon,\sigma_\epsilon)$=(44.49,36.16) (b) $(\mu_\epsilon,\sigma_\epsilon)$=(0.48,0.11)

Figure 5: Evaluation of camera parameters using EEE measure before (a) and after BA (b) for the Berkeley WAMI dataset. The pel in each matrix, ϵ_{lm}, indicates the error between the l-th camera (matrix rows) and m-th camera (columns), computed using (7). The ϵ_{lm} pel values were truncated beyond the maximum value (ie 50 and 5). The mean and standard deviation of the errors (8) and (9) over all cameras are shown below each plot. Notice that each plot uses a different scale for better visualization of errors. Color bars shown to the right.

5650, 2.66 GHz, 12 cores (24 threads), 24 GB RAM and nVidia GTX480/1.5GB GPU. SIFT-GPU [54] was used for fast feature extraction. The Ceres Solver library [2] was used for non-linear least-squares estimation; Schur's complement, Cholesky factorization and Levenberg-Marquardt algorithms were used for trust region step computation.

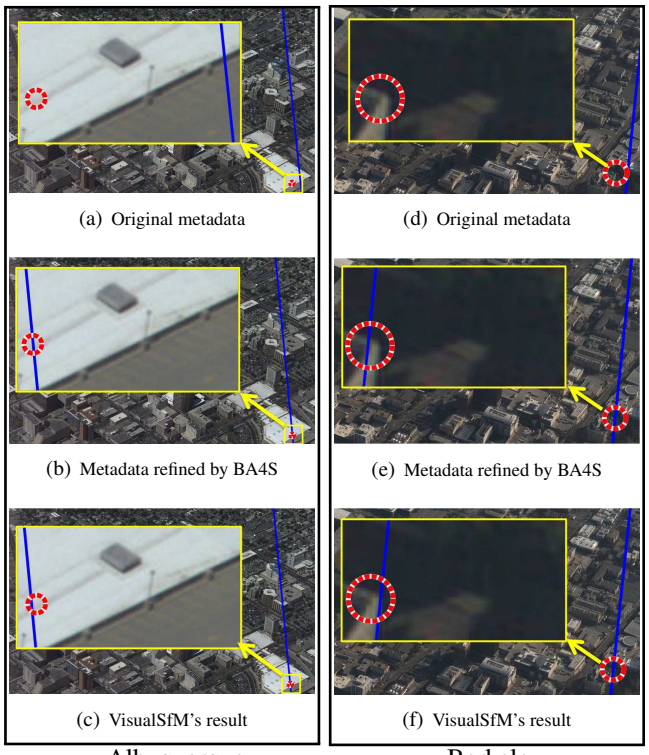
4.3. Results

The characteristics of the test dataset along with timing results are given in Table 1. Each dataset includes platform-based camera orientation matrices and translation vectors provided by imprecise IMU and GPS measurements along with intrinsic camera parameters that we refer to as metadata. The BA4S pipeline was run on each dataset. A non-linear triangulation algorithm [21] was used to estimate and initialize 3D points. The persistency factors of the tracks and their related statistics are used as weights in the adaptive robust function for the BA optimization. We compare BA4S with VisualSfM [56] a state-of-the-art SfM implementation for some datasets; For VisualSfM we provided imagery without including metadata which is not a fully supported feature. VisualSFM uses two strategies for matching including pre-emptive and exhaustive [55]. With preemptive matching VisualSFM generated several fragments of cameras and only a fraction of cameras could be recovered while for other cameras it failed. This is consistent with similar observations about VisualSFM's limited performance on sequential aerial images [45]. We ran VisualSFM in its exhaustive/expensive matching mode in order to recover all cameras.

BA4S performance in terms of overall and per frame execution times are plotted in Figure 2. The time per frame is approximately constant (see column 11 in Table 1) and is independent of the number of cameras (or views) which is a surprising result compared to other methods in the literature. In fact for the longer sequence the per-frame time is decreasing which is very promising for large scale aerial imagery mosaicing and reconstruction applications. For the largest dataset (Columbia-II) with 5,322 frames BA4S uses only 0.59 sec per frame.

The camera positions and their viewing directions (optical axes) for the LA dataset are plotted in Figure 3. Figure 4 shows the degree of camera pose correction recovered by BA4S. The EEE evaluation metric perviously explained was applied to the Berkeley dataset and the error matrix shown as colored pels in Figure 5. The left plot shows the EEE measure of the camera parameters using metadata (uncorrected platform camera parameters); the range of errors is truncated to 50 pixels. The initial raw metadata is very noisy but after refinement using the proposed BA4S pipeline there is significant improvement in quality (see Figure 5b). The EEE μ_ϵ and σ_ϵ statistics using all the cameras (see (8) and (9)) are also given. BA4S was quite successful in refining the metadata while using significantly less time (Table 1).

Figure 6 shows the EEE graphically to assess camera parameters for Albuquerque (left column) and Berkeley (right column) datasets. Point correspondence #2 in the ground truth between 50th and 150th cameras within the sequence were used. The epipolar lines corresponding to image #50 in each dataset is computed using the camera parameters and plotted on image #150. The first row shows original raw metadata (unrefined). The middle row shows the epipolar lines after the metadata were refined using the BA4S



(a) Original metadata (d) Original metadata

(b) Metadata refined by BA4S (e) Metadata refined by BA4S

(c) VisualSfM's result (f) VisualSfM's result

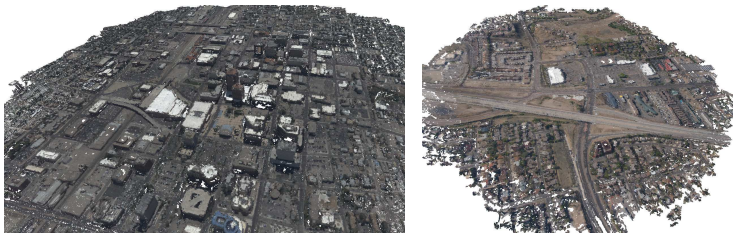
Albuquerque Berkeley

Figure 6: Visual assessment of camera parameters using epipolar lines for the corresponding ground truth points between a pair of cameras (#50,#150). The first and second columns correspond to Albuquerque and Berekeley datasets. The images correspond to camera #150 within each dataset. On each image the corresponding ground truth point is indicated by red circle. Epipolar lines corresponding to the ground truth point from camera #50 (the other camera in the pair) is calculated using camera parameters and plotted on the image of camera #150, for each dataset. The camera parameters from three different sources are used in each row; top: metadata, middle: BA4S (refined metadata) and bottom: VisualSFM.

pipeline. The epipolar lines should ideally pass through the ground truth points (center of the marked circles in each plot). As can be seen, the noise in metadata is significantly reduced after applying BA4S. The errors values in these plots are consistent with the EEE values in Figure 5; look at the pair (#50,#150) in the matrix. The epipolar lines in the last row were plotted using the camera parameters estimated by VisualSFM for comparison.

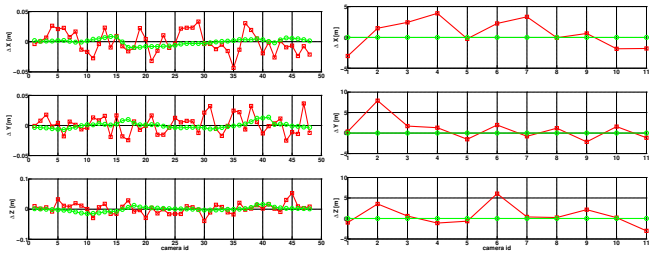
Usually a dense 3D reconstruction algorithm such as PMVS[18] is applied after BA in order to obtain a dense and colored point cloud. We also applied PMVS for some of the datasets to visually assess reconstructed point clouds. The optimized metadata from BA4S is used as input to PMVS (or CMVS). Figures 7-a and -b show the PMVS dense point clouds for Albuquerque and Four Hills respectively.

In addition to testing BA4S on aerial WAMI datasets, we have applied it to the Middlebury benchmark datasets for mulitview 3D reconstruction which are not WAMI but the images are acquired sequentially. *Dino* [46] is one of the challenging datasets for a classical BA due to the lack of salient features for tracking across views [17, 4]. The proposed BA4S pipeline was tested on this dataset to evaluate its applicability and performance for non-WAMI trajectory images. The camera parameters in the Middlebury ground truth were synthetically perturbed for both rotation and translation. The perturbed camera parameters along with the images were input to the BA4S pipeline. The *Dino* dataset has 48 cameras with image resolution of 640×480 pixels. The position errors for the metadata (perturbed camera parameters before optimization) and the optimized ones by BA4S are plotted in Figure 8a. Figure 9 shows the visual assessment of two point correspondences. The errors for the corresponding epipolar lines are significantly reduced after BA4S optimized camera parameters. The epipolar lines have very large errors (third and fourth columns of each row) when the noisy camera parameters are used. A dense version of the point cloud using PMVS with BA4S optimized camera parameters is shown in Figure 11a.



(a) Albuquerque downtown (b) Four Hills - Albuquerque

Figure 7: Dense 3D point clouds obtained by applying PMVS using BA4S outputs



(a) Dino (b) FountainP11

Figure 8: Position errors before (red curve) and after (green curve) optimization using BA4S to refine the metadata for Dino (left) and Fountain-P11 (right) datasets.

FountainP11 [50] is another non-WAMI dataset. As with the Dino dataset the ground truth camera parameters were perturbed prior to running the BA4S algorithm. There are 11 cameras each image has a resolution of 3072×2048 pixels. The position errors for the metadata (perturbed camera parameters) and the refined results using BA4S are plotted in Figure 8b. Figure 10 shows two point correspondences. The initial epipolar lines have very large errors in the views (third and fourth columns of each row) when the perturbed camera parameters were used. The errors for the corresponding epipolar lines become significantly smaller once BA4S refined camera parameters are used. A dense version of the point cloud for FountainP11 using PMVS and optimized camera parameters is shown in Figure 11b.

5. Conclusions

We describe BA4S a fast, robust and efficient SfM pipeline that we developed for 3D reconstruction from im-

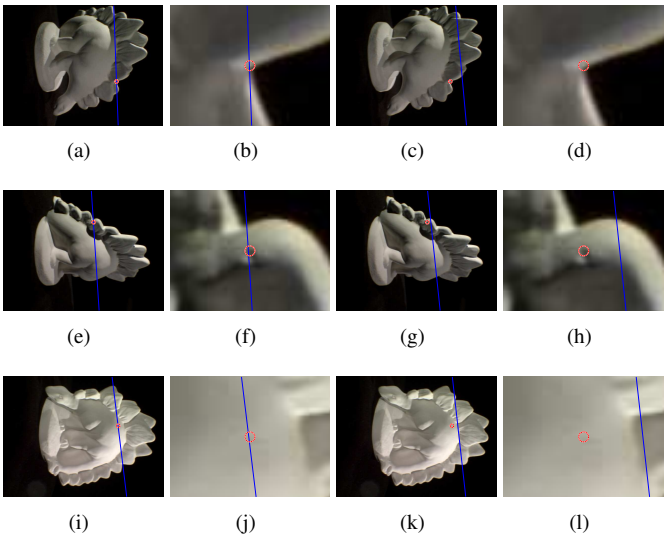


Figure 9: Visual assessment of camera parameters using epipolar lines in Dino dataset. Each row shows the results for one correspondence. First and second columns of each row show the full and zoomed views using camera parameters after BA4S. Third and fourth columns are using the original camera metadata parameters.

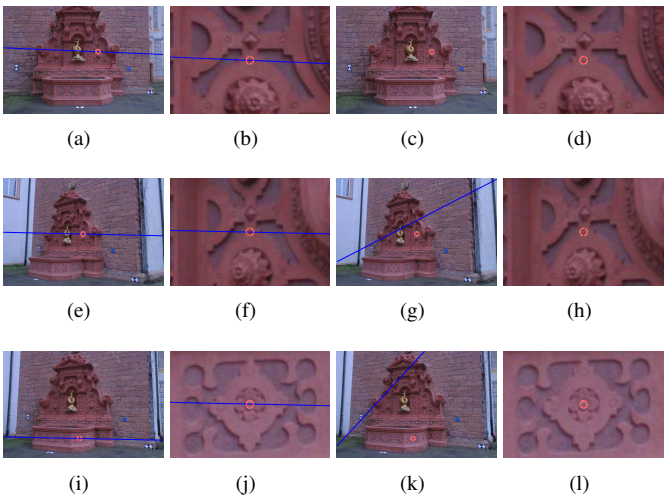


Figure 10: Visual assessment of camera parameters using epipolar lines for the Fountain-P11 dataset. Each row shows the results for one correspondence. First and second columns of each row show the visual assessment when camera parameters were obtained from BA4S. Third and fourth columns show the significantly larger errors when the original camera metadata are used.

ages acquired using sequentially ordered camera motion with approximate camera metadata. Significant performance gains for 3D reconstruction are possible since we do not require RANSAC-like combinatorial feature correspondences and outlier rejection nor estimating the initial camera parameters (i.e. no essential matrix estimation). A new robust reprojection error function was introduced that is adaptive to feature track or 3D scene point quality and measures the *co-visibility* persistency factor of each track relative to the population statistics. Using the BA4S pipeline it is possible to efficiently refine noisy camera parameters more 100 times faster than VisualSfM, taking on average just 0.59 sec per

(a) Dino (b) Fountain-P11

Figure 11: Dense reconstruction using PMVS after optimized camera parameters are estimated using BA4S.

frame. The proposed SfM pipeline is highly suitable and scalable for 3D urban reconstruction for wide area motion imagery in which high resolution geo-tagged aerial imagery are sequentially acquired.

Acknowledgments

This research was partially supported by the U.S. Air Force Research Laboratory under agreement AFRL FA875014-2-0072. The aerial WAMI was collected by Transparent Sky, LLC in Edgewood, NM and provided by Steve Suddarth. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of AFRL, NRL, or the U.S. Government.

References

[1] S. Agarwal, Y. Furukawa, and N. Snavely. Building rome in a day. *Communications of the ACM*, 54:105–112, 2011.

[2] S. Agarwal and K. Mierle. Ceres solver. `http://ceres-solver.org`.

[3] S. Agarwal, N. Snavely, S. M. Seitz, and R. Szeliski. Bundle adjustment in the large. In *European Conference Computer Vision*, pages 29–42, 2013.

[4] A. Albarelli, E. Rodola, and A. Torsello. Imposing semi-local geometric constraints for accurate correspondences selection in structure from motion: A game-theoretic perspective. *Int. Journal of Computer Vision*, 97:36–53, 2012.

[5] H. Aliakbarpour, L. Almeida, P. Menezes, and J. Dias. Multi-sensor 3D volumetric reconstruction using CUDA. *3D Research, Springer*, 2(4):1–14, 2011.

[6] H. Aliakbarpour and J. Dias. Three-dimensional reconstruction based on multiple virtual planes by using fusion-based camera network. *Jour. of IET Computer Vision*, 6(4):355, 2012.

[7] H. Aliakbarpour, K. Palaniappan, and J. Dias. Geometric exploration of virtual planes in a fusion-based 3D data registration framework. In *Proc. SPIE Conf. Geospatial InfoFusion III (Defense, Security and Sensing: Sensor Data and Information Exploitation)*, volume 8747, page 87470C, 2013.

[8] H. Aliakbarpour, K. Palaniappan, and G. Seetharaman. Robust camera pose refinement and rapid SfM for multi-view aerial imagery without RANSAC. *IEEE Journal of Geoscience and Remote Sensing Letters*, Accepted, 2015.

[9] A. Aravkin, M. Styer, Z. Moratto, A. Nefian, and M. Broxton. Student's t robust bundle adjustment algorithm. In *Int. Conf. on Image Processing*, pages 1757–1760, 2012.

[10] E. Blasch, P. Deignan, S. Dockstader, M. Pellechia, K. Palaniappan, and G. Seetharaman. Contemporary concerns in geographical/geospatial information systems (GIS) processing. In *Proc. IEEE National Aerospace and Electronics Conference (NAECON)*, pages 183–190, 2011.

[11] E. Blasch, G. Seetharaman, K. Palaniappan, H. Ling, and G. Chen. Wide-area motion imagery (WAMI) exploitation tools for enhanced situational awareness. In *Proc. IEEE Applied Imagery Pattern Recognition Workshop*, 2012.

[12] M. Bryson, A. Reid, F. Ramos, and S. Sukkarieh. Airborne vision-based mapping and classification of large farmland environments. *Journal of Field Robotics*, 27(5):632–655, may 2010.

[13] A. Delaunoy and M. Pollefeys. Photometric Bundle Adjustment for Dense Multi-View 3D Modeling. apr 2014.

[14] T. Dickscheid, T. Läbe, and W. Förstner. Benchmarking automatic bundle adjustment results. In *21st Congress of the International Society for Photogrammetry and Remote Sensing (ISPRS). Beijing, China*, pages 7–12, 2008.

[15] J.-M. Frahm, M. Pollefeys, S. Lazebnik, D. Gallup, B. Clipp, R. Raguram, C. Wu, C. Zach, and T. Johnson. Fast robust large-scale mapping from video and internet photo collections. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(6):538–549, 2010.

[16] F. Fraundorfer and D. Scaramuzza. Visual odometry: Part ii: Matching, robustness, optimization, and applications. *IEEE Robotics & Automation Magazine*, 19(2):78–90, 2012.

[17] Y. Furukawa and J. Ponce. Accurate camera calibration from multi-view stereo and bundle adjustment. *Int. Journal of Computer Vision*, 84(3):257–268, 2009.

[18] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(8):1362–76, 2010.

[19] M. Gerke. Dense image matching in airborne video sequences. *ISPRS*, pages 639–644, 2008.

[20] M. Grabner. Object recognition based on local feature trajectories. *Proceedings of the International Cognitive Vision*, 2005.

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

[22] H. Hirschmüller, M. Buder, and I. Ernst. Memory efficient semi-global matching. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume I, pages 371–376, Melbourne, Australia, 2012. XXII ISPRS Congress.

[23] V. Indelman, R. Roberts, C. Beall, and F. Dellaert. Incremental light bundle adjustment. In *British Machine Vision Conference*, pages 134.1–134.11, 2012.

[24] A. Irschara, C. Hoppe, H. Bischof, and S. Kluckner. Efficient structure from motion with weak position and orientation priors. *IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2011.

[25] Y. Jeong, S. Member, D. Niste, and I.-s. Kweon. Pushing the envelope of modern methods for bundle adjustment. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 34(8):1605–1617, 2012.

[26] K. Konolige. Sparse Bundle Adjustment. *Proc. of the British Machine Vision Conference*, pages 102.1–102.11, 2010.

[27] J. Kopf, M. Cohen, and R. Szeliski. First-person hyper-lapse videos. *ACM Trans. Graphics (SIGGRAPH)*, 33(4), 2014.

[28] R. Lakemond, C. Fookes, and S. Sridharan. Resection-intersection bundle adjustment revisited. *ISRN Machine Vision*, 2013:1–8, 2013.

[29] M. Lhuillier. Incremental fusion of structure-from-motion and GPS using constrained bundle adjustments. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 34(12):2489–2495, 2012.

[30] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections, 1981.

[31] M. Lourakis and A. Argyros. SBA: A software package for sparse bundle adjustment. *ACM Transactions on Mathematical Software*, 36(1):30, 2009.

[32] J. McGlone, E. Mikhail, J. Bethel, and R. Mullen, editors. *Manual of Photogrammetry, Fifth Ed.* American Society of Photogrammetry and Remote Sensing, 2004.

[33] J. Mundy. The relationship between photogrammetry and computer vision J.L. Mundy GE Corporate Research and Development Schenectady, NY 12309. *In Integrating Photogrammetric Techniques With Scene Analysis and Machine Vision, SPIE*, 1944, 1993.

[34] S. Nath and K. Palaniappan. Adaptive robust structure tensors for orientation estimation and image segmentation. *Lecture Notes in Computer Science (ISVC)*, 3804:445–453, 2005.

[35] S. Niko and P. Protzel. Towards using sparse bundle adjustment for robust stereo odometry in outdoor terrain. In *TARO*, volume 2, pages 206–213, 2006.

[36] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(6):756–770, 2004.

[37] E. Ontiveros, C. Salvaggio, D. Nilosek, N. Raqueno, and J. Faulring. Evaluation of image collection requirements for 3D reconstruction using phototourism techniques on sparse overhead data. In *Proc. SPIE 8390. Algorithms Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XVIII*, 2012.

[38] K. Palaniappan, F. Bunyak, P. Kumar, I. Ersoy, S. Jaeger, K. Ganguli, A. Haridas, J. Fraser, R. Rao, and G. Seetharaman. Efficient feature extraction and likelihood fusion for vehicle tracking in low frame rate airborne video. In *13th Int. Conf. Information Fusion*, 2010.

[39] K. Palaniappan, R. Rao, and G. Seetharaman. Wide-area persistent airborne video: Architecture and challenges. In B. Banhu et al., editors, *Distributed Video Sensor Networks: Research Challenges and Future Directions*, chapter 24, pages 349–371. Springer, 2011.

[40] R. Pelapur, S. Candemir, F. Bunyak, M. Poostchi, G. Seetharaman, and K. Palaniappan. Persistent target tracking using likelihood fusion in wide-area and full motion video sequences. In *15th Int. Conf. Information Fusion*, pages 2420–2427, 2012.

[41] R. Pelapur, K. Palaniappan, and G. Seetharaman. Robust orientation and appearance adaptation for wide-area large format video object tracking. In *9th IEEE Int. Conf. Advanced Video and Signal-Based Surveillance*, 2012.

[42] T. B. Pollard. *Comprehensive 3D change detection using volumetric appearance modeling*. PhD thesis, Brown University, USA, 2009.

[43] M. Pollefeys, D. Nistér, J.-M. Frahm, A. Akbarzadeh, et al. Detailed real-time urban 3D reconstruction from video. *Int. Journal of Computer Vision*, 78(2-3):143–167, oct 2007.

[44] E. Rupnik, F. Nex, and F. Remondino. Oblique multi-camera systems - orientation and dense matching issues. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-3/W1:107–114, 2014.

[45] J. Schönberger, F. Fraundorfer, and J.-M. Frahm. Structure-from-Motion for MAV image sequence analysis with photogrammetric applications. In *International Archives of the Photogrammetry, Remote Sensing, and Spatial Information Sciences*, volume XL-3, pages 305–312, 2014.

[46] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 519–528, 2006.

[47] K. I. J. Shawn Recker, Christiaan Gribble, Mikhail Shashkov, Mario Yepez, Mauricio Hess-Flores. Depth Data Assisted Structure-from-Motion Parameter Optimization and Feature Track Correction. In *Applied Imagery Pattern Recognition Workshop (AIPR)*, 2014.

[48] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from Internet photo collections. *International Journal of Computer Vision*, 80:189–210, 2008.

[49] H. Stewenius, C. Engels, and D. Nistér. Recent developments on direct relative orientation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 60(4):284–294, 2006.

[50] C. Strecha, W. von Hansen, L. V. Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. *IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.

[51] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment – a modern synthesis. In *Vision Algorithms: Theory and Practice (W. Triggs, A. Zisserman, and R. Szeliski Eds.)*, pages 298–372, 2000.

[52] R. Viguier, C.-C. Lin, K. Swaminathan, S. Pankanti, et al. Resilient mobile cognition: Algorithms, innovations, and architectures. In *IEEE Int. Conf. on Computer Design*, 2015.

[53] R. Wang, F. Bunyak, G. Seetharaman, and K. Palaniappan. Static and moving object detection using flux tensor with split gaussian models. In *IEEE CVPR Change Detection Workshop*, pages 420–424, 2014.

[54] C. Wu. SiftGPU: A GPU implementation of Scale Invaraint Feature Transform (SIFT), 2007.

[55] C. Wu. Towards linear-time incremental structure from motion. *Int. Conf. 3D Vision*, pages 127–134, 2013.

[56] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz. Multicore bundle adjustment. In *IEEE Conf. Computer Vision Pattern Recognition*, pages 3057–3064, 2011.