STRUCTURE-FROM-MOTION FOR MAV IMAGE SEQUENCE ANALYSIS WITH PHOTOGRAMMETRIC APPLICATIONS

Johannes L. Schönberger[a,b,*], Friedrich Fraundorfer[a], Jan-Michael Frahm[b]

[a] Technische Universität München, Remote Sensing Technology, 80333 München, Germany
(johannes.schoenberger, friedrich.fraundorfer)@tum.de
[b] University of North Carolina at Chapel Hill, Department of Computer Science, Chapel Hill, NC 27599, USA
(jsch, jmf)@cs.unc.edu

KEY WORDS: structure-from-motion, 3D reconstruction, MAV, image sequences, low-resolution, close-range

ABSTRACT:

MAV systems have found increased attention in the photogrammetric community as an (autonomous) image acquisition platform for accurate 3D reconstruction. For an accurate reconstruction in feasible time, the acquired imagery requires specialized SfM software. Current systems typically use high-resolution sensors in pre-planned flight missions from far distance. We describe and evaluate a new SfM pipeline specifically designed for sequential, close-distance, and low-resolution imagery from mobile cameras with relatively high frame-rate and high overlap. Experiments demonstrate reduced computational complexity by leveraging the temporal consistency, comparable accuracy and point density with respect to state-of-the-art systems.

1. INTRODUCTION

Micro-aerial-vehicles (MAV) provide an autonomous and cost-efficient platform for mapping inaccessible or dangerous areas. Their ability to acquire data from close-distance enables high-resolution mapping with low-resolution sensors. However, the physical constraints of MAV systems present a set of challenges; size, weight, and power represent the most important design factors. Typical unmanned systems are equipped with propellers, a processing unit, a battery, communication units, an inertial-measurement-unit (IMU), and a set of sensors. MAV image sequences present SfM systems with several challenges. Due to their limited form-factor, MAVs typically have unstable trajectories, their SLAM cameras produce low-resolution imagery, and have unstable aberration parameters. Reconstructions suffer from accumulated drift effects, difficult geometry due to small baselines, and data gaps in image sequences, e.g., caused by loss of connection to ground control stations or spontaneous adaption of exposure settings. Hence, accurate and robust 3D mapping typically relies on additional higher-resolution cameras as additional payload to the platform, and specialized 3D reconstruction software is necessary for the processing. Over the last couple of years, incremental SfM systems have emerged as a comparatively robust technology. Yet, they typically produce poor results in very challenging environments, and it is still a rather computationally expensive procedure if the goal is to yield highly accurate results for large scenes.

This paper describes and evaluates an offline SfM system specifically designed for sequential, low-resolution imagery from mobile cameras with relatively high frame-rate and high overlap (w.r.t. traditional photogrammetric systems). Our proposed SfM pipeline fills a gap in current 3D reconstruction software, which is primarily targeted at high-resolution imagery from flight missions planned in advance. However, lower resolution sensors enable MAV missions, where additional payload of high-resolution cameras is not feasible, extended flight duration and high maneuverability is a must, and prior path planning is not possible due to lack of localization sensors (e.g., GPS). By leveraging the temporal consistency of the imagery, the system achieves significantly reduced computational complexity while producing accurate results comparable to current state-of-the-art aerial platforms. It performs well in challenging environments, where other systems fail to produce reasonable reconstructions. The key features of the proposed system are: (1) Usage of an image retrieval system to reduce matching complexity, for automatic loop-closure, and for automatic merging of separate sub-sequences with overlapping trajectories. (2) Reduced computational complexity by leveraging the sequence's temporal consistency, redundant viewpoint detection, and adequate algorithms. (3) Reconstruction initialization without prior information (e.g., GPS and IMU measurements). (4) Flexible models for different cameras, including wide-angle cameras, with self-calibration capability of camera model parameters. (5) Definition of control-points for georegistration and measurement of unknown points. (6) Inclusion of prior rotation constraints from IMUs.

Optionally, the reconstruction results of the pipeline can be used as input for commercial photogrammetry software, e.g., for refined bundle adjustment (BA), to derive surface models and orthophotos, or for distance and volume measurements. The results of this work and a sample data-set are available as part of the open-source software `MAVMAP` at `https://github.com/mavmap`.

2. RELATED WORK

Recently, MAV systems have found increased attention in the photogrammetric community as an image-acquisition platform for accurate 3D reconstruction, e.g., mapping of inaccessible or dangerous areas (Eisenbeiss, 2009), for DEM generation (Greiwe et al., 2013), surveying of archeological sites (Fallavollita et al., 2013), or glacier mapping (Solbø and Storvold, 2013). MAV mapping has been shown to yield comparable results to traditional aerial systems (Barry and Coakley, 2013, Küng et al., 2011). However, current MAV systems rely on carrying higher-resolution cameras with low frame-rate as additional payload, or need GPS and IMU measurements for the initialization of the reconstruction process. Moreover, these systems require prior path and image acquisition planning. Due to the camera's high-resolution and its additional payload, these systems typically fly at far-distance

[*] Corresponding author.

from the object of interest, they capture images at a low framerate, and the MAV platform should stand still during image acquisition. Mobile cameras with low-resolution sensors and high frame-rate do not impose those constraints, i.e. MAV systems can (autonomously) explore and map a region in a continuous manner from close-distance, without careful prior planning or intermediate stops for image acquisition. These properties enable us to use MAVs for new mapping scenarios, e.g., for indoor environments (no GPS), time-critical missions (no prior planning and continuous, fast overfly), or environments divided into small, occluded objects (requires close-distance acquisition).

Nowadays, numerous MAV systems use cameras as primary sensors for real-time navigation and exploration. However, accurate, image-based 3D SLAM as an online application is still impractical and does not scale to large environments due to high computational complexity of SfM. To reduce payload, power consumption, and computational complexity, most online SLAM systems use low-resolution cameras and simplified, approximate mapping solutions (Davison, 2003, Choi and Lee, 2011, Dryanovski et al., 2013, Meier et al., 2012), or, keyframe-based incremental BA (Weiss et al., 2013). Many systems leverage multiple cameras with fixed baselines for direct stereo reconstruction (Meier et al., 2012), some use wide-angle cameras to cover a larger field-of-view (Shen et al., 2013, Weiss et al., 2013). Yet, mapping systems for this kind of low-resolution imagery are primarily designed for real-time navigation and exploration as an on-board (Dryanovski et al., 2013, Shen et al., 2013, Weiss et al., 2013) or (partially) off-board (Meier et al., 2012) solution. However, it seems natural to use this data for accurate 3D reconstruction.

Traditional photogrammetric software entails a relatively high level of human interaction, and is not suitable for processing large MAV image data-sets (Greiwe et al., 2013). Barazzetti et al. (Barazzetti et al., 2010) propose an automated tie-point extraction scheme, and also commercial solutions with a (semi-)automatic work-flow have appeared, e.g., Erdas Imagine Photogrammetry (Hexagon Geospatial, 2014), PhotoScan (Agisoft LLC, 2014), or Pix4Dmapper (Pix4D SA, 2014). These systems typically use GPS and IMU information for reconstruction initialization and apply an exhaustive matching approach for tie-point extraction, which is needless for sequential imagery and computationally prohibitive for large image sequences due to its quadratic computational complexity. Abdel-Wahab et al. (Abdel-Wahab et al., 2012) propose a system with lower computational complexity for unordered collections of high-resolution imagery.

In parallel to the aforementioned efforts, the computer vision community has developed SfM systems to reconstruct large-scale environments from a huge number of unordered photos in feasible times with no human interaction (Snavely et al., 2006, Frahm et al., 2010). Frahm et al. (Frahm et al., 2010) leverage iconic image selection through clustering and Agarwal et al. (Agarwal et al., 2010) image retrieval systems to reduce the computational burden of exhaustive matching. VisualSFM (Wu, 2013) emerged as the probably most advanced, publicly accessible system for automated and efficient 3D reconstruction from unordered photo collections. However, it provides limited flexibility for photogrammetric applications (limited camera models and self-calibration capability, no estimation of unknown points, no integration of IMU priors, etc.), and is also based on an exhaustive matching approach (no exploitation of temporal consistency of data).

3. PIPELINE

In this section we describe our SfM pipeline for sequential image data. Figure 1 shows a flowchart of the reconstruction process. We use m, n, and o to denote the number of 3D points, images, and image measurements during the reconstruction, respectively. Additionally, we assume that the monocular imagery is sequentially ordered according to its acquisition time.

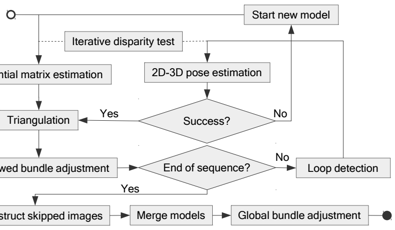


Figure 1: SfM pipeline overview.

3.1 Image description and matching

Feature type SIFT has been proven to be a very robust feature w.r.t. rotation and scale change. However, SIFT extraction is comparatively expensive; consequently, we employ the slightly less discriminative SURF features as a compromise between robustness and speed (Heinly et al., 2012). Both feature types can be used for building efficient image retrieval systems. For loop closure (see Section 3.5), we use such a system to index images on-the-fly during the incremental reconstruction.

Feature detection Image streams from mobile cameras tend to change contrast and brightness abruptly over time due to adaptive exposure settings of the sensor. Conventional corner detection, such as peak finding in Harris or Hessian corner measure responses (Mikolajczyk et al., 2005), naturally detects more features in high contrast regions of an image. Hence, if an image region is captured once in low and once in high contrast between two consecutive images, a typical feature detector will not detect features at the same locations. As a consequence, a system would fail to continue tracks of those features; resulting in weak or, in the worst case, to a total loss of connectivity between two images. The KLT tracker (Tomasi and Kanade, 1991) is able to overcome this problem, but generally only works for very high-frequency image sequences with very small feature displacement. Alternatively, we propose to partition an image into fixed-size regions and enforce a minimum and maximum amount of features in each region by adaptive adjustment of the SURF corner thresholds for each region. Depending on the image capture frequency, we mostly face small changes in content between consecutive images. Accordingly, there is seldom need to adjust the thresholds and the adaptive detection has only minimal impact on the overall detection performance. However, the adaptive detection significantly increases connectivity between images.

Feature matching Feature extraction and exhaustive matching is typically still the most computationally expensive part in SfM systems for arbitrary photo collections (Wu, 2013). Image matching includes the brute-force distance computation between all possible feature pair combinations $O(o^2)$ for all possible image pair combinations $O(n^2)$. The conventional exhaustive approach is necessary in order to establish connections between all parts of a scene; this is especially important for sequential image data, which typically has very loose connectivity. However, the time consistent structure of sequential image data gives us prior information about which images are likely to match; namely those images with small time offsets between their acquisitions. By leveraging this fact, we can reduce the computational complexity of matching from $O(n^2)$ to $O(n)$. In addition, feature displacement is small between consecutive images, which enables us to significantly reduce the number of feature pairs for which to compute the descriptor similarity measure by restricting the search space to feature pairs with small displacement. To obtain robust putative matches we employ the ratio test (Lowe, 2004), and perform a cross-check, i.e. we only accept correspondences that match to each other from one image to another and vice-versa.

3.2 Initial reconstruction

Essential matrix estimation Assuming that we are given a sufficient guess for the intrinsic camera parameters, we can reconstruct the initial camera poses up to an unknown scale by estimating the essential matrix $\mathbf{E}$ (Longuet Higgins, 1981) from 5 normalized corresponding observations. We robustly estimate $\mathbf{E}$ using the five-point algorithm by Nister (Nister, 2003); this algorithm yields very good performance and stability for general-purpose application. To obtain relative rotation $\mathbf{R}$ and translation $\mathbf{t}$, we decompose $\mathbf{E}$ and use the cheirality constraint (Hartley, 1993) to recover the correct solution, i.e. 3D points must have positive depth. The RANdom SAmple Consensus (RANSAC) algorithm (Fischler and Bolles, 1981) serves as a framework for robust estimation. While more sophisticated RANSAC techniques (Raguram et al., 2013) have been developed over the last years, conventional RANSAC still performs best for data with high inlier ratio, as we mostly face for sequential imagery.

Initial image pair Choosing a good initial image pair, by making a trade-off between geometric quality (camera registration and structure uncertainty) and model size o (number of 3D points), is a critical step in incremental SfM systems and different strategies exist (Beder and Steffen, 2006, Snavely et al., 2006). We try to find such a pair at the beginning of the sequence.

Geometric quality (uncertainty of reconstruction) is mainly determined by triangulation angles between intersecting viewing rays (Hartley and Zisserman, 2004, ch. 12.6); a larger baseline between two images increases triangulation angles. The amount of spatial displacement of corresponding image measurements and the number of inliers in estimating a homography between two images are indicators for the amount of viewpoint change in terms of parallax and disparity (Nister, 2000). While feature displacement correlates with viewpoint change, it typically yields less meaningful results for scenes with large depth-variation. In the case where the transformation between two images can be described by a homography $\mathbf{H}$, all 3D points lie in a plane (Hartley and Zisserman, 2004, ch. 18.5). The homography between calibrated cameras a and b is defined as

$$\mathbf{H}_{ab} = \mathbf{R} - \mathbf{t}\mathbf{n}^T/d \tag{1}$$

where $\{\mathbf{R}, \mathbf{t}\}$ describe the relative motion from a to b, and $\{\mathbf{n}, d\}$ denote the normal vector and distance from camera a to the 3D plane $\mathbf{n}^T\mathbf{x} + d = 0$, respectively. Knowing the homography between images is equivalent to knowing the 3×3 part $\mathbf{M}$ of the 3D transformation from camera system a to b

$$\mathbf{P}_{ab} = [\mathbf{M}|\mathbf{t}] \tag{2}$$

So, a homography fully describes the transformation $\mathbf{P}_{ab}$, if the camera motion $\mathbf{t}$ is infinitesimal. Though, the homography criteria (upper bound for number of inliers) fails to produce meaningful results, if the scene is planar; in this case it will always have a high inlier ratio. We therefore enforce a combination of sufficient feature displacement $\Delta\mathbf{X}$ and a small number of homography inliers N_H (Snavely et al., 2006) for the selection of the initial pair, i.e. we reject image pairs that satisfy the following condition

$$N_H < \hat{N}_H \vee \Delta\mathbf{X} < \Delta\hat{\mathbf{X}} \tag{3}$$

where $\hat{N}_H$ is set to a high fraction (e.g., 0.9) of the number of feature matches, and $\Delta\hat{\mathbf{X}}$ to a pixel distance. Additionally, we avoid strong forward motion (by evaluating the z-component of $\mathbf{t}$), since it degrades stability of 3D points observed in the image center due to small triangulation angles. Another factor for the stability of the initial model is the uncertainty of the image measurements, which propagates to the accuracy of their corresponding 3D points. A smaller reprojection threshold in RANSAC improves uncertainty of the measurements.

Improving geometric quality inevitably reduces the size of the initial model due to a typically reduced number of correspondences. Larger model size is desirable, since, on the one hand, it leads to higher redundancy in the estimation of the geometric configuration; on the other hand, a sufficient model size is necessary as a starting point for the following sequential reconstruction described in Section 3.3.

3.3 Sequential reconstruction

Based on the existing 3D model, sequential reconstruction incrementally registers new cameras from 2D-3D correspondences. This is commonly known as the Perspective-n-Point (PnP) problem or 2D-3D pose estimation (Fischler and Bolles, 1981). Typical SfM systems with an exhaustive matching approach combine image matching with a geometric verification using robust fundamental or essential matrix estimation. The procedure then incrementally extends the model by registering new camera poses from geometrically verified image pairs by solving the PnP problem (Agarwal et al., 2010, Frahm et al., 2010). Robustness is enforced by RANSAC, which has exponential computational complexity in the number of model parameters. So, in order to sample at least one outlier-free set of corresponding image measurements, one must run a minimum of iterations

$$d = \log(1 - p) / \log(1 - e^s) \tag{4}$$

where p denotes the confidence of sampling at least one outlier-free set of measurements, s the number of model parameters and e the inlier ratio of our measurements.

2D-3D pose estimation We reduce computational complexity by including the geometric verification (typically through essential matrix estimation) in the sequential 2D-3D pose estimation. The well-determined 2D-3D camera pose estimation (P3P) only requires four correspondences between image observations and known 3D points (Gao et al., 2003), which reduces the necessary number of RANSAC iterations d. We eliminate the computational burden of essential matrix estimation once we reconstructed the initial pair. Suppose the measurements in our estimation are corrupted with 30% outliers and we want to sample at least one outlier-free set of samples with 99% confidence; reducing the model parameters from 5 (essential matrix) to 4 (P3P) leads to a decrease in minimum RANSAC iterations from 26 to 17, which is equivalent to a speedup of more than 30%. Furthermore, we use the very efficient and stable closed-form solution to the P3P problem by Gao et al. (Gao et al., 2003), which is much less computationally expensive than the most-efficient solvers for essential matrix estimation.

Pose refinement and triangulation After we estimated a rough camera pose in the previous step, next we employ a non-linear pose refinement by using all inliers from the 2D-3D RANSAC procedure. The pose refinement is critical for the accurate and reliable triangulation of new 3D points from image measurement correspondences, which do not have a corresponding 3D point in the existing model. We use a linear approximation for the triangulation (Hartley and Zisserman, 2004) and refine the 3D points by using BA (see Section 3.4). To accept new 3D points as being valid, they must satisfy the following conditions: sufficiently large triangulation angle, small reprojection error, and they must pass the cheirality constraint w.r.t. to the refined camera pose. We also continue existing 3D point tracks to increase redundancy, but we only enforce the latter two constraints.

Redundant viewpoints To ensure stable registration, we skip images with small baselines using the same constraints as described in Section 3.2. Note, that skipped frames are registered before final BA. In addition, by skipping redundant viewpoints (with a high degree of overlap), we gain in terms of speed, since rejecting similar images is computationally cheap in comparison to the 2D-3D RANSAC procedure, with the non-linear pose refinement, the triangulation of new 3D points and the windowed BA (see Section 3.4). Given the smooth temporal viewpoint change, we follow its temporal order; another reason for this is to ensure smooth trajectories during BA (see Section 3.4).

3.4 Bundle adjustment

The previous sections described the process of camera registration and structure computation, which are subsequently refined in a non-linear optimization, known as bundle adjustment (BA) (Luhmann et al., 2013, Hartley and Zisserman, 2004, Triggs et al., 2000). As described in the following, our system integrates optional methods as a complement to traditional BA to improve reconstruction results: definition of flexible camera models, camera parameter self-calibration, inclusion of rotation constraints from IMU sensors, and automatic geo-registration and measurement of unknown 3D points through control-point definition.

Estimation framework BA is a sparse geometric estimation problem with 3D features $\mathbf{P}_i = \{\mathbf{P}_1, ..., \mathbf{P}_m\}$ and camera parameters $\mathbf{C}_j = \{\mathbf{C}_1, ..., \mathbf{C}_n\}$ as unknowns, and 2D image features $\mathbf{X}_k = \{\mathbf{X}_1, ..., \mathbf{X}_o\}$ as uncertain measurements, where k denotes the index of the true measurement of 3D feature i in image j. A projection function $\mathbf{Q}(\mathbf{P}_i, \mathbf{C}_j)$ maps 3D features to their corresponding 2D image features and serves as a mathematical model. BA minimizes the cost function S in terms of the overall reprojection error

$$S = \frac{1}{2} \sum_{k=1}^{o} \rho_k \left(\|\mathbf{X}_k - \mathbf{Q}(\mathbf{P}_i, \mathbf{C}_j)\|_2^2 \right) \tag{5}$$

by simultaneously refining 3D features and camera parameters according to the loss function $\rho_k(\cdot)$. BA is a high-dimensional, non-linear problem, where the state vector $\mathbf{U} = \{\mathbf{P}_i, \mathbf{C}_j\}$ lies on a manifold and is iteratively reparameterized using local, linear approximations. On the one hand, it is essential to use parameterizations that can be locally approximated by linear and quadratic functions (e.g., we use the axis-angle representation as a parameterization for rotations). On the other hand, the parameter space has multiple local minima and hence one needs a good initial guess for the parameter vector (camera parameters and 3D features) in order to converge to the global minimum. Attainment of these initial estimates is covered in Sections 3.1, 3.2 and 3.3.

We use the `Ceres-Solver` library (Agarwal et al., 2014) as a framework for non-linear least-squares estimation. Despite careful filtering of image feature matches, there is a considerable amount of outlier observations in BA. Consequently, we assume that the observations follow a heavy-tailed Cauchy distribution, modeled as

$$\rho_k(x) = \log(1 + x) \tag{6}$$

Note, that maximum likelihood (ML) estimation is naturally robust as long as the correct error model is employed. For numerically stable and efficient optimization it is essential to make use of the special sparsity of BA, since cameras in image sequences typically only see a small excerpt of the complete scene. We use the Schur complement trick (Brown, 1958), maximize sparsity structure by column and row re-ordering (Li and Saad, 2006), use sparse Cholesky factorization (Chen et al., 2008), and use the Levenberg-Marquardt algorithm for trust region step computation (Levenberg, 1944, Marquardt, 1963).

Without constraining the geometry of the reconstruction, BA can perturb the geometric structure arbitrarily. In general, the structure can be perturbed by an arbitrary 3D similarity transformation (7 degrees of freedom: scale, rotation and translation) without changing the scene's image projections, which results in a rank-deficient Jacobian and is known as a datum defect. There are two methods to overcome this defect: fixed- and free-network adjustments (Triggs et al., 2000). We use the former by fixing at least 7 DoF during the optimization; we set the parameters of certain camera poses (rotation and translation) as fixed. In addition, we must avoid configuration defects, which occur in case a parameter is not well-determined by its measurements. Examples include 3D points observed only from one camera, which can be moved arbitrarily along the viewing ray, or cameras with an insufficient number of image projections (e.g., in case of windowed BA).

Camera models Highly accurate results in BA rely on a precise mathematical modeling of the underlying geometry of the image formation process, modeled as $\mathbf{Q}(\mathbf{P}_i, \mathbf{C}_j)$. The standard pinhole camera model is often a too limited description of these processes and more sophisticated models that take internal camera aberrations into account typically provide better abstraction thereof; however, over-parameterized models tend to yield biased results. The intrinsic parameters of a camera model can be estimated in photogrammetric calibrations. Low-cost cameras, as often used on mobile platforms, are especially subject to temporal change in their calibration parameters, and need complex models for accurate reconstruction. The rigidity of the scene provides constraints which one can exploit for self-calibration of camera parameters during the BA. Since we employ automatic feature detection (see Section 3.1) and thus have a comparatively large number of image observations distributed over the entire image plane, we can achieve very good calibration estimates. Our library provides an easy means to define arbitrary mathematical camera models and to specify prior calibration parameters as fixed, or variable for self-calibration. Separate camera models and parameters can be specified for each image separately, or one can use the same parameters for specific sets of images. For an evaluation of the benefits of using different camera models, see Section 4.

Rotation constraints Nowadays, a large number of camera platforms are equipped with inertial measurement units (IMU), which yield orientation information through gyroscopes, accelerometers, and magnetometers. Depending on the accuracy of the IMU, this information can help to reduce accumulated drift effects by constraining the camera rotation during BA (see Section 4.). The rotation constraint is modeled for each camera in the sequence as

$$\Delta r_j = \left\| \hat{\mathbf{R}}_j - \mathbf{R}_j \right\|_F \tag{7}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. $\mathbf{R}_k$ and $\hat{\mathbf{R}}_k$ respectively denote the extrinsic 3×3 rotation matrix of the estimate and the IMU measurement of camera k. The constraint is added to the overall cost function of the BA by weighting it with the regularization parameter λ_r and transitioning from a ML to a maximum a posteriori (MAP) estimate:

$$S_r = S + \frac{\lambda_r}{2} \sum_{j=1}^{n} \rho_k (\Delta r_j) \tag{8}$$

Control points and geo-registration The incremental SfM procedure is based upon automatic detection and matching of image observations. Our method provides an optional and easy way to manually define additional image observations of specific 3D points, denoted as control points. Control points with unknown 3D coordinates are estimated during BA (e.g., for measurement of specific, new points); control points with known 3D coordinates are used for geo-registration and hence introduced as fixed parameters during BA. At least three fixed control points are necessary for geo-registration to set the datum; in this case all camera poses are introduced as free parameters during BA.

Initial bundle adjustment Following the approximate reconstruction of the initial camera pair using essential matrix estimation (see Section 3.2), we perform BA to refine the 3D points and the camera pose configuration. We set the datum by fixing six parameters (rotation and translation) of the first and one parameter of the second camera translation. These seven DoF equal the required minimum number of fixed parameters to set the datum and to allow a maximum refinement of the camera configuration.

Windowed bundle adjustment After each sequentially registered image (see Section 3.3) we perform windowed BA, i.e. the optimization w.r.t. the most recently registered images. For temporally windowed BA, it is crucial that the pipeline reconstructs images in sequential order and that we fix at least two poses (12 degrees of freedom) at the beginning of the window, i.e. the poses of the oldest images. Suppose we allowed to optimize the poses of the oldest image. Then the camera pose and the 3D points could be changed arbitrarily without any relation to the remaining part of the known reconstruction; resulting in non-continuous transitions in trajectory and scene solely caused by windowed BA. For windowed BA, we employ a comparatively relaxed convergence criteria accounting for a strictly limited number of iterations and reduced computational complexity. This does not degrade the final reconstruction quality, since cameras and 3D points normally stabilize quickly and BA naturally optimizes unstable parts of the problem the most.

Global bundle adjustment Before applying a final, global BA over the complete reconstruction, we reconstruct the remaining poses of skipped images, as described in sections 3.2 and 3.3, but without intermediate windowed BA. In contrast to windowed BA, we enforce a stricter convergence criteria for the final BA in order to achieve higher accuracy.

3.5 Loop closure

Even when leveraging BA, SfM inevitably suffers from accumulated drift. Hence, loop closure is essential to compensate for this effect by strengthening weakly connected geometry of the structure. Our loop closure method is based on an incrementally built-up image retrieval system, which is used for detection of similar images. Similar image pairs are geometrically verified and BA minimizes the drift w.r.t. to the new geometric constraints.

Image retrieval The image retrieval system is built upon a visual word based vocabulary tree (Nister and Stewenius, 2006). We incrementally index newly registered images in the retrieval system using the same local SURF features as for image matching to avoid duplicate computational burden. The local features (visual words) of an image are quantized using the vocabulary tree and represent a sparse document vector in the image database, which is implemented as an inverse file. The sparseness of the vectors results in a low memory footprint and allows for efficient search in the database. For image retrieval, a query document is compared against all documents in the database according to an L_2 based similarity measure. The weight of the different visual words is based on the length of the inverted file (inverse document frequency). Ordering the images by the score value yields a ranking of the images in terms of visual similarity. The most similar images can then be used as hypotheses for loop closure.

Geometric verification and optimization We invoke image retrieval every l frames during the reconstruction. Naturally, image retrieval detects images with small time offset in the image sequence, as those images are likely to have large overlap. While this stabilizes local connectivity of the trajectory, it is especially important to also close loops between images with large temporal offsets (i.e. when the camera revisits the same location multiple times). Consequently, we retrieve a large number of similar images from the database, restrict geometric verification to a maximum number of temporally local images and allow an unlimited amount of loop closure attempts to images with large temporal offsets. The newly established geometric constraints are only considered in the final, global BA (see Section 3.4). Note, that the chosen cost-function in BA only down-weights and not neglects large errors. Hence, BA optimizes accumulated drift effects, when loops are closed successfully.

3.6 Merging of separated, overlapping sequences

Sequential reconstruction occasionally fails to connect consecutive images, e.g., due to errors in the image capture, texture-less regions, or over-exposed images. In this case, the pipeline starts a new, separate model by invoking the initial reconstruction on the remaining set of images. However, camera trajectories often overlap when the platform revisits the same spatial location multiple times. Detection of these overlaps enables us to connect and merge the separated reconstructions into one combined model. We leverage loop detection as described in Section 3.5 to find the connections between the different models. If we successfully verify a connection of at least three common images between two models, we can use their common images to estimate a similarity transformation from one model into the other. We find such connections over the entire sequence in order to estimate a reliable transformation and to establish strong connections for BA.

4. EXPERIMENTS

To evaluate our method we use experiments both on low- and high-resolution imagery. These show that our system robustly produces reconstructions for challenging data-sets, captured by low-resolution, wide-angle, and high frame-rate cameras. An experiment on a data-set consisting of multiple, separated subsequences illustrates the utility of automatic reconstruction merging. In another experiment we demonstrate the theoretical benefits of using prior IMU measurements as orientation constraints. Additionally, our system is able to estimate the camera pose with cm-accuracy w.r.t. a geodetically measured ground-truth trajectory. Finally, we assess the accuracy of unknown point estimation based on a manually evaluated high-resolution data-set with a net of geodetically measured ground-truth control-points, which were not used in the registration. All experiments were performed on a computer system with an Intel Xeon E5645 6-Core processor, 24GB RAM, and a nVidia GTX 560 graphics card.

4.1 Low-resolution imagery

In this experiment we apply our SfM pipeline on an image sequence acquired by the PIXHAWK system (Meier et al., 2012). The system has a downward-looking camera, captures images with a resolution of 0.3MP (752×480) at a frame rate of 15Hz from an average distance of $\approx 10m$ above ground, and yields orientation through an IMU.

	# residuals	# cameras	# points	Total time
VisualSFM (preemptive)	10,756,887	2,395 (67%)	488,350	30.2h
VisualSFM (exhaustive)	–	–	–	$\approx$ 19d
MAVMAP	7,039,224	3,574 (100%)	348,482	1.5h

Table 1: Reconstruction results for Rubble 1 data-set.

Non-contiguous sequence (Rubble 1) This data-set comprises 3574 images (undistorted with parameters from a lab-calibration) of a rubble field, has multiple loops and was acquired in three independent flight missions The data-set cannot be reconstructed with simple, sequential reconstruction techniques (see figure 2). Reconstruction results can be found in Table 1 and Figure 2. Our pipeline merges the three sub-models into a single model, closes loops between multiple flyovers ($l = 30$) and reconstructs all camera poses in a reasonable amount of time. We attain high redundancy through automatic feature detection and matching in terms of a mean track-length of 12 observations per 3D point. More than 10% of the 348,482 3D points have a track-length of > 20, and the mean reprojection error is $0.36px$. We can see that the estimated camera orientation suffers from comparatively little accumulated drift. Additionally, depending on the accuracy of the IMU, constraining the camera rotation with the IMU orientation measurements can reduce noise. Using preemptive matching, VisualSFM can only connect two of the three sub-sequences, and it only registers a fraction of 67% of all cameras. Exhaustive matching is not practicable due to excessive computational complexity. We were not able to produce a reasonable reconstruction with traditional photogrammetric software for this data-set; despite the significant amount of human interaction required, which is impractical for sequences of this size.

Contiguous sequence with ground-truth camera trajectory (Rubble 2) The reconstruction results for another sequence consisting of 4234 images can be found in Figure 3. We use the original, distorted images and totally rely on self-calibration (with initial guesses for focal length and principal point from the camera specifications) using a 9 DoF model suitable for wide-angle cameras. For the first 1100 frames, we synchronously tracked the camera position using a tachymeter (Leica TotalStation 15i) with sub-cm accuracy. As a post-processing step we aligned the reconstructed and ground-truth trajectories using an over-determined similarity transformation estimation. Again, our system registers all cameras and we achieve a mean residual of $\Delta\mathbf{P} = 11.8cm$ for the reconstructed w.r.t. the ground-truth trajectory. Additionally, in comparison to a model suitable for wide-angle cameras (9 DoF) (Mei and Rives, 2007), a standard camera model with radial and tangential distortion parameters (8 DoF OpenCV model) yields significantly inferior results with $\Delta\mathbf{P} = 17.8cm$. We can see that the reconstructed trajectory follows a curvature in the altitude direction due to large radial distortion at the image borders, while the actual altitude of the MAV stays approximately at the same level.

Ground sampling distance Since MAVs are able to acquire images from close-range, even low-resolution imagery can be exploited for high-resolution mapping. Suppose the MAV flies by the object of interest at a distance of 10m, and with a field-of-view of 90°, we achieve a ground sampling distance of $GSD = 2.7cm$ (sensor resolution of 752 × 480). This is comparable to current state-of-the-art aerial imaging systems (e.g., UltraCamXp WA yields a $GSD = 8.6cm$ at a flight altitude of 1000m) and MAV platforms, e.g., as used by the senseFly drone system, which achieves a $GSD = 4cm$ while using a camera with a sensor resolution of 16MP (12MP), but flies at an altitude of 130m (115m).

4.2 High-resolution imagery

In addition, we carried out experiments on a set of 31 high-resolution images of a concrete platform (Platform data-set) acquired by a SONY NEX-7 camera with a resolution of 6000 × 4000 pixels mounted on a MAV. The entire scene is covered by a network of 48 geodetically measured control points with sub-mm precision. Manual evaluation of the scene using Erdas Imagine Photogrammetry yields a mean point accuracy of $\sigma_E = 0.2cm$. We ran our pipeline on the original, distorted images by using a subset of 10 fixed control points for geo-registration, and evaluated the residuals of the remaining control points w.r.t. the estimated Erdas Image Photogrammetry results (see Figure 4). The complete BA problem comprises 25,777 3D points with a total of 128,282 observations and a mean reprojection error of $0.29px$. Hence, the inclusion of 48 control points has negligible impact on the overall estimation result. Our system achieves accurate results within σ_E. Moreover, we obtained results for two camera models of differing complexity (4 DoF pinhole model without distortion parameters, and a 8 DoF OpenCV model with two radial and tangential distortion parameters, respectively). In this case, usage of an appropriate camera model improves accuracy by more than 20%.

5. CONCLUSION

In this paper we present a publicly accessible SfM pipeline for the robust 3D reconstruction from low-resolution image sequences. It fills a gap in currently available 3D reconstruction software packages. We show that the system scales well with large image data-sets by leveraging its temporal consistency, while achieving comparable accuracy to manual, photogrammetric evaluation.

6. ACKNOWLEDGMENT

We thank Markus Achtelik (ASL, ETH Zrich) for providing the Rubble 2 dataset and Konrad Eder (FPF, TU Mnchen) for providing the Platform dataset. This material is based upon work supported by the National Science Foundation under Grant No. IIS-1252921 and No. IIS-1349074.

REFERENCES

Abdel-Wahab, M., Wenzel, K. and Fritsch, D., 2012. Efficient reconstruction of large unordered image datasets for high accuracy photogrammetric applications. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences I-3, pp. 1–6.

Agarwal, S., Furukawa, Y., Snavely, N., Curless, B., Seitz, S. M. and Szeliski, R., 2010. Reconstructing rome. Computer 43(6), pp. 40–47.

Agarwal, S., Mierle, K. and Others, 2014. Ceres solver. https://code.google.com/p/ceres-solver/.

Agisoft LLC, 2014. Photoscan. http://www.agisoft.ru/products/photoscan.

Barazzetti, L., Remondino, F., Scaioni, M. and R., B., 2010. Fully automatic uav image-based sensor orientation. Int. Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences.

Barry, P. and Coakley, R., 2013. Field accuracy test of rpas photogrammetry. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XL-1/W2, pp. 27–31.

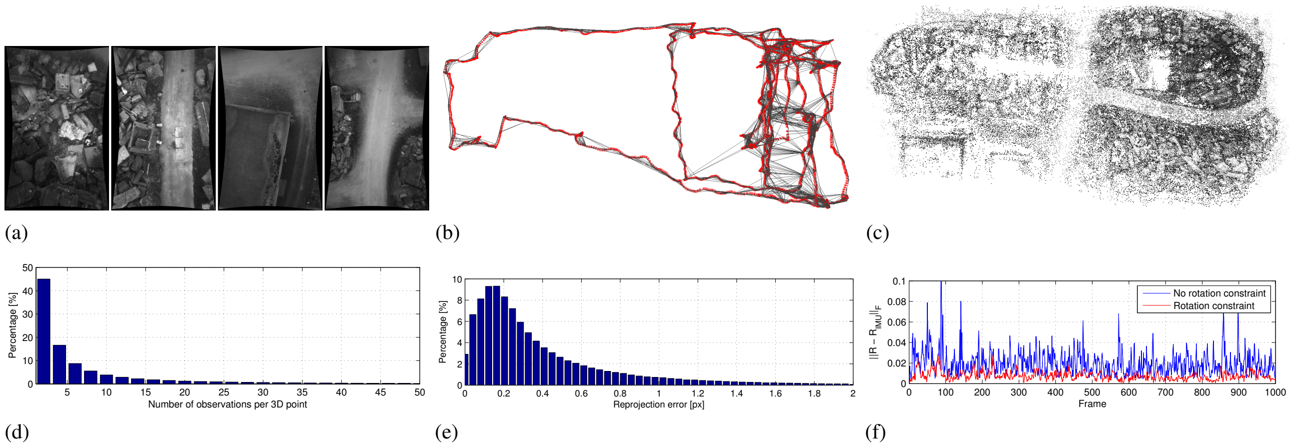


Figure 2: Rubble 1 data-set: Low-resolution, undistorted MAV image sequence (a), reconstructed trajectory with loop-closure connections (b), resulting sparse 3D point cloud (c), clipped track-length histogram with maximum track-lengths up to 567 (d), the reprojection error histogram (e), and the deviation of rotation estimates from IMU orientation ($\lambda_r = 10$) for one of the sub-sequences (f).

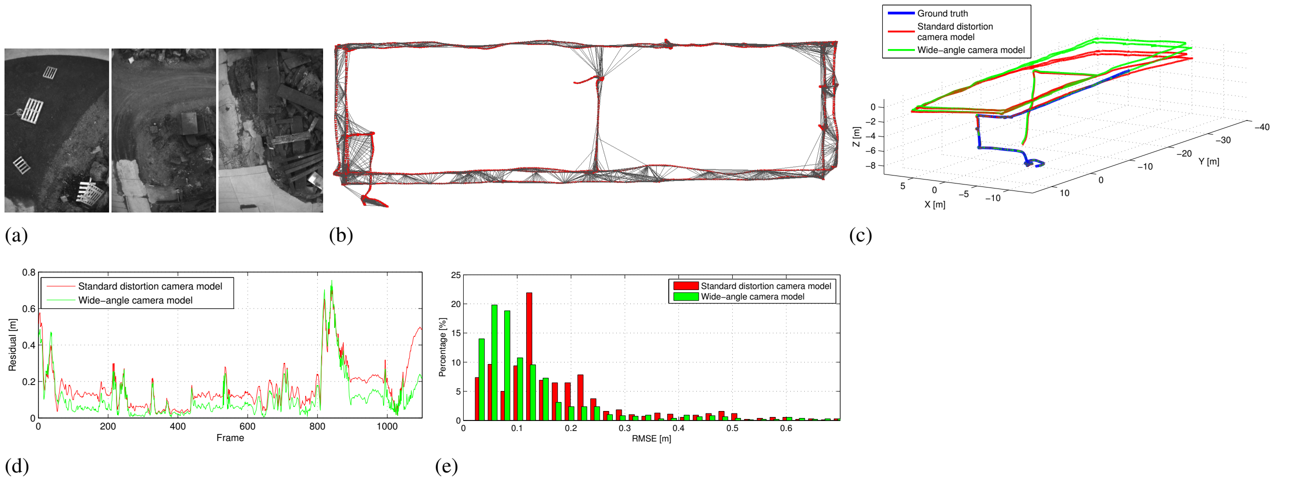


Figure 3: Rubble 2 data-set: Low-resolution, distorted MAV image sequence (a), the reconstructed trajectory with loop-closure connections (b), the ground-truth and reconstructed trajectory (c), and residuals of the reconstruction results w.r.t. the ground-truth (d,e).

Beder, C. and Steffen, R., 2006. Determining an initial image pair for fixing the scale of a 3d reconstruction from an image sequence. In: Proceedings of the 28th Conference on Pattern Recognition, DAGM'06, Springer-Verlag, Berlin, Heidelberg, pp. 657–666.

Brown, D. C., 1958. A solution to the general problem of multiple station analytical stereo triangulation. Technical Report Technical Report No. 43 (or AFMTC TR 58-8), Technical Report RCA-MTP Data Reduction, Patrick Airforce Base, Florida.

Chen, Y., Davis, T. A., Hager, W. W. and Rajamanickam, S., 2008. Algorithm 887: Cholmod, supernodal sparse cholesky factorization and update/downdate. ACM Trans. Math. Softw. 35(3), pp. 22:1–22:14.

Choi, K. and Lee, I., 2011. Real-time georeferencing of image sequence acquired by a uav multi-sensor system. In: Multi-Platform/Multi-Sensor Remote Sensing and Mapping (M2RSM), 2011 International Workshop on, pp. 1–6.

Davison, A., 2003. Real-time simultaneous localisation and mapping with a single camera. In: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, pp. 1403–1410 vol.2.

Dryanovski, I., Valenti, R. and Xiao, J., 2013. An open-source navigation system for micro aerial vehicles. Autonomous Robots 34(3), pp. 177–188.

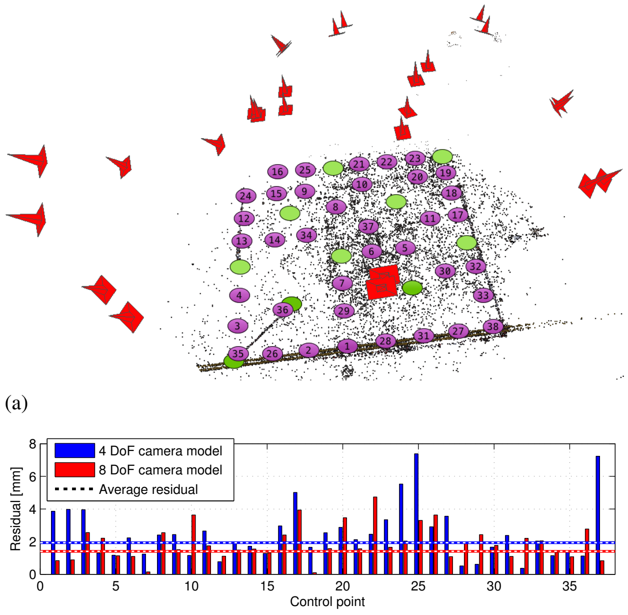
Eisenbeiss, H., 2009. UAV Photogrammetry. phdthesis, Institute of Geodesy and Photogrammetry, ETH Zurich, Zurich, Switzerland.

Fallavollita, P., Balsi, M., Esposito, S., Melis, M. G., Milanese, M. and Zappino, L., 2013. Uas for archaeology & new perspectives on aerial documentation. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XL-1/W2, pp. 131–135.

Fischler, M. A. and Bolles, R. C., 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24(6), pp. 381–395.

Frahm, J.-M., Fite-Georgel, P., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y.-H., Dunn, E., Clipp, B., Lazebnik, S. and Pollefeys, M., 2010. Building rome on a cloudless day. In: K. Daniilidis, P. Maragos and N. Paragios (eds), Computer Vision ECCV 2010, Lecture Notes in Computer Science, Vol. 6314, Springer Berlin Heidelberg, pp. 368–381.

Gao, X.-S., Hou, X.-R., Tang, J. and Cheng, H.-F., 2003. Complete solution classification for the perspective-three-point problem. IEEE Trans. Pattern Anal. Mach. Intell. 25(8), pp. 930–943.



(a)

(b)

Figure 4: Platform data-set: Camera, point cloud and control point configuration (fixed: green, estimated: purple) (a), and accuracy of control points for different camera models w.r.t. ground-truth data (b).

Greiwe, A., Gehrke, R., Spreckels, V. and Schlienkamp, A., 2013. Aspects of dem generation from uas imagery. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XL-1/W2, pp. 163–167.

Hartley, R. I., 1993. Cheirality invariants. In: IN PROC. DARPA IMAGE UNDERSTANDING WORKSHOP, pp. 745–753.

Hartley, R. I. and Zisserman, A., 2004. Multiple View Geometry in Computer Vision. Second edn, Cambridge University Press, ISBN: 0521540518.

Heinly, J., Dunn, E. and Frahm, J.-M., 2012. Comparative evaluation of binary features. In: A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato and C. Schmid (eds), Computer Vision ECCV 2012, Lecture Notes in Computer Science, Springer Berlin Heidelberg, pp. 759–773.

Hexagon Geospatial, 2014. Erdas imagine photogrammetry. http://www.hexagongeospatial.com/products/imagine-photogrammetry/Details.aspx.

Küng, O., Strecha, C., Beyeler, A., Zufferey, J.-C., Floreano, D., Fua, P. and Gervaix, F., 2011. The accuracy of automatic photogrammetric techniques on ultra-light uav imagery. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XXXVIII-1/C22, pp. 125–130.

Levenberg, K., 1944. A method for the solution of certain non-linear problems in least squares. Quarterly Journal of Applied Mathmatics II(2), pp. 164–168.

Li, N. and Saad, Y., 2006. Miqr: A multilevel incomplete qr preconditioner for large sparse least-squares problems. SIAM J. Matrix Anal. Appl. 28(2), pp. 524–550.

Longuet Higgins, H., 1981. A computer algorithm for reconstructing a scene from two projections. Nature.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2), pp. 91–110.

Luhmann, T., Robson, S., Kyle, S. and Boehm, J., 2013. Close-range Photogrammetry and 3D Imaging. 2nd edn, De Gruyter, Berlin, Germany.

Marquardt, D. W., 1963. An algorithm for least-squares estimation of nonlinear parameters. Journal of the Society for Industrial and Applied Mathematics 11(2), pp. pp. 431–441.

Mei, C. and Rives, P., 2007. Single view point omnidirectional camera calibration from planar grids. In: Robotics and Automation, 2007 IEEE International Conference on, pp. 3945–3950.

Meier, L., Tanskanen, P., Heng, L., Lee, G., Fraundorfer, F. and Pollefeys, M., 2012. Pixhawk: A micro aerial vehicle design for autonomous flight using onboard computer vision. Autonomous Robots 33(1-2), pp. 21–39.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T. and Gool, L., 2005. A comparison of affine region detectors. International Journal of Computer Vision 65(1-2), pp. 43–72.

Nister, D., 2000. Reconstruction from uncalibrated sequences with a hierarchy of trifocal tensors. In: In ECCV, pp. 649–663.

Nister, D., 2003. An efficient solution to the five-point relative pose problem. In: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, Vol. 2, pp. II–195–202 vol.2.

Nister, D. and Stewenius, H., 2006. Scalable recognition with a vocabulary tree. In: IN CVPR, pp. 2161–2168.

Pix4D SA, 2014. Pix4dmapper. http://pix4d.com/.

Raguram, R., Chum, O., Pollefeys, M., Matas, J. and Frahm, J., 2013. Usac: A universal framework for random sample consensus. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35(8), pp. 2022–2038.

Shen, S., Mulgaonkar, Y., Michael, N. and Kumar, V., 2013. Vision-based state estimation for autonomous rotorcraft mavs in complex environments. In: Robotics and Automation (ICRA), 2013 IEEE International Conference on, pp. 1758–1764.

Snavely, N., Seitz, S. M. and Szeliski, R., 2006. Photo tourism: Exploring photo collections in 3d. In: ACM SIGGRAPH 2006 Papers, SIGGRAPH '06, ACM, New York, NY, USA, pp. 835–846.

Solbø, S. and Storvold, R., 2013. Mapping svalbard glaciers with the cryowing uas. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XL-1/W2, pp. 373–377.

Tomasi, C. and Kanade, T., 1991. Detection and tracking of point features. Technical report, International Journal of Computer Vision.

Triggs, B., McLauchlan, P., Hartley, R. and Fitzgibbon, A., 2000. Bundle adjustment a modern synthesis. In: B. Triggs, A. Zisserman and R. Szeliski (eds), Vision Algorithms: Theory and Practice, Lecture Notes in Computer Science, Vol. 1883, Springer Berlin Heidelberg, pp. 298–372.

Weiss, S., Achtelik, M. W., Lynen, S., Achtelik, M. C., Kneip, L., Chli, M. and Siegwart, R., 2013. Monocular vision for long-term micro aerial vehicle state estimation: A compendium. Journal of Field Robotics 30(5), pp. 803–831.

Wu, C., 2013. Towards linear-time incremental structure from motion. In: 3D Vision, 2013 International Conference on, pp. 127–134.