Computer Vision and Image Understanding 000 (2016) 1–19

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance

Markus Rumpler[a,*], Alexander Tscharf[b], Christian Mostegel[a], Shreyansh Daftry[c], Christof Hoppe[a], Rudolf Prettenthaler[a], Friedrich Fraundorfer[a], Gerhard Mayer[b], Horst Bischof[a]

[a] Institute for Computer Graphics and Vision, Graz University of Technology, Austria
[b] Chair of Mining Engineering and Mineral Economics, Montanuniversität Leoben, Austria
[c] Robotics Institute, Carnegie Mellon University, USA

ARTICLE INFO

Article history:
Received 4 December 2015
Revised 30 March 2016
Accepted 18 April 2016
Available online xxx

Keywords:
Photogrammetric computer vision
Unmanned aerial vehicles
Image-based 3D reconstruction
Mapping
Camera calibration
Image acquisition
Online feedback
Structure-from-motion
Georeferencing
Fiducial markers
Accuracy evaluation

ABSTRACT

During the last decades photogrammetric computer vision systems have been well established in scientific and commercial applications. Recent developments in image-based 3D reconstruction systems have resulted in an easy way of creating realistic, visually appealing and accurate 3D models. We present a fully automated processing pipeline for metric and geo-accurate 3D reconstructions of complex geometries supported by an online feedback method for user guidance during image acquisition. Our approach is suited for seamlessly matching and integrating images with different scales, from different view points (aerial and terrestrial), and with different cameras into one single reconstruction. We evaluate our approach based on different datasets for applications in mining, archaeology and urban environments and thus demonstrate the flexibility and high accuracy of our approach. Our evaluation includes accuracy related analyses investigating camera self-calibration, georegistration and camera network configuration.

1. Introduction

Creating and visualizing realistic and accurate 3D models is becoming a central ambition of research in the field of spatial data acquisition. In this domain passive cameras have become very popular as measurement devices due to their inherent flexibility and low cost compared to traditional surveying equipment. As passive cameras are also very light-weight they can be mounted on UAVs (unmanned aerial vehicles), which have emerged in recent years as a promising platform to perform close range aerial data acquisition and surveying tasks (Rehak et al., 2013).

UAVs help to overcome geometric constraints and combine the advantages of both aerial and terrestrial photogrammetry and also serve as a low-cost alternative to the classical manned surveying. Typical applications reach from agriculture and environmental monitoring, surveying tasks for mining, archaeology or architecture as well as inspection of objects that are difficult and dangerous to reach for human operators. Multi-copter UAVs in particular, are able to capture highly overlapping images from almost terrestrial camera view points to oblique and nadir aerial images, due to the ability to navigate at very low airspeed and hover at nearly any position.

Together with an automated multi-view processing pipeline, dense 3D point clouds from images are generated in a flexible, fast and cheap way and can compete with point clouds from laser scans in terms of accuracy (Leberl et al., 2010). Such multi-view

* Corresponding author.
E-mail addresses: rumpler@icg.tugraz.at (M. Rumpler), alexander.tscharf@unileoben.ac.at (A. Tscharf), mostegel@icg.tugraz.at (C. Mostegel), daftry@cmu.edu (S. Daftry), hoppe@icg.tugraz.at (C. Hoppe), rudolf.prettenthaler@tugraz.at (R. Prettenthaler), fraundorfer@icg.tugraz.at (F. Fraundorfer), gerhard.mayer@unileoben.ac.at (G. Mayer), bischof@icg.tugraz.at (H. Bischof).

http://dx.doi.org/10.1016/j.cviu.2016.04.008
1077-3142/

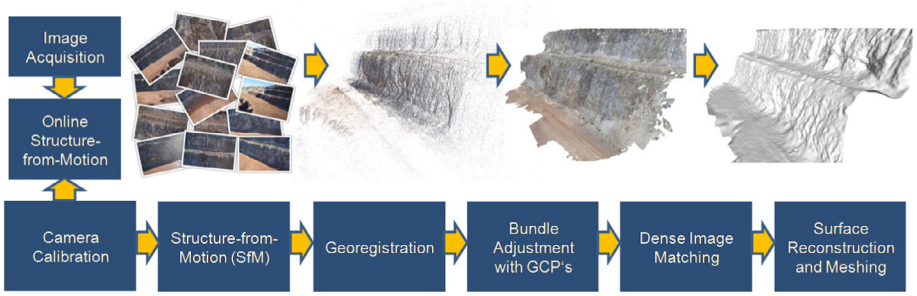


Fig. 1. Automated processing workflow for geo-accurate reconstructions. Top row: Image set, sparse reconstruction, dense point cloud and triangle-based surface mesh of a quarry wall.

processing pipelines have been integrated in many software packages (e.g. VisualSfM, Acute3D, Pix4D, Agisoft PhotoScan, PhotoModeler, etc.). These pipelines are able to calculate the intrinsic and extrinsic camera parameters as well as scene structure represented as a (sparse) 3D point cloud from an unordered set of images. In subsequent steps, the model gets refined to generate a denser point cloud (Furukawa and Ponce, 2009; Hirschmueller, 2005).

Many of the afore mentioned software packages show increasing robustness and result in high quality and visually appealing 3D models (Tscharf et al., 2015). However, there are three main drawbacks.

The first drawback is that the reconstruction errors of the models are not always clear and so they are often not directly suited for photogrammetric applications.

The second drawback is that nearly all of the software packages include manual steps e.g. for indirect georeferencing to establish tie point correspondences and aligning the reconstructions in a world coordinate frame.

The third drawback is that none of these software packages provides the operator with sufficient information to judge the completeness of the reconstruction at acquisition time. Especially complex object geometries require high overlap and a very dense image network to guarantee completeness, which cannot be ensured by using terrestrial or aerial nadir images exclusively. Only a combination of terrestrial and aerial viewpoints is able to guarantee completeness of the model.

In this context, we see the need for a user-friendly, fully automated processing pipeline including user guidance during image acquisition, an easy to use camera calibration procedure and accurate georeferencing of reconstructed models in a world coordinate system having absolute geographic position and orientation with predictable reconstruction accuracy (Fig. 1).

In the first part of our paper (Section 2) we present our fully automated multi-scale end-to-end workflow (Fig. 1) to create precise and geo-accurate reconstructions in complex environments by the combined use of different camera platforms (aerial and terrestrial). Our contribution is three-fold. Firstly, we present and advocate the use of planar fiducial markers as a target pattern to obtain accurate and reliable camera calibration. Secondly, with known ground control point (GCP) positions or GPS measurements of the camera positions, we are able to automatically set the generated 3D model into its geographic reference coordinate frame. By additionally integrating GCPs and camera self-calibration in the bundle adjustment optimization, we are able to create precise and geo-accurate 3D reconstructions without any manual interaction. Lastly, we integrate user guidance (Hoppe et al., 2012) into the data acquisition step to ensure that acquired images are suited for automated photogrammetric processing (Section 3). Photogrammetric methods need to cope robustly with unordered sets of images, where scale, depth and ground sampling distance (GSD) changes are immanent. Our online feedback approach thus ensures that the final Structure-from-Motion (SfM) reconstruction meets predefined accuracy requirements and results in a complete reconstruction of the object. To our knowledge, this is the first approach of this kind.

The largest part of this paper (Section 4) gives insights how we evaluated and describes what is important during image acquisition to obtain highly accurate reconstructions. We present accuracy related analyses investigating camera self-calibration, georegistration, camera network configuration and ground sampling distance and show how to obtain geo-accurate reconstructions for complex object geometries with high precision using aerial UAV imagery in combination with terrestrial images.

We evaluate our approach based on five different scenarios for applications in mining and archaeology, as well as urban environments and demonstrate the flexibility and high accuracy of our approach.

2. Reconstruction pipeline

In this section, we describe our fully automated multi-view processing pipeline to reconstruct geo-accurate 3D models and camera positions with input images captured with different cameras at different scales and view points. The reconstruction pipeline is roughly divided into four parts: camera calibration; determination of the exterior parameters of camera positions and orientations together with the reconstruction of sparse 3D object points; geo-registration by transforming the 3D model into a geographic reference coordinate frame; densification of the object points and finally generation of a polygonal surface model and texturing. The reconstruction step takes pre-calibrated images from different sources, groups them according to their intrinsic parameters and processes them jointly to finally generate a textured polygonal surface model.

2.1. Calibration

Accurate intrinsic camera calibration is critical to computer vision methods that involve image based measurements. Traditional SfM pipelines such as Bundler, Agisoft, etc. employ a direct use of uncalibrated views for 3D reconstruction, and can inherently deal with a dataset having images taken at varying focal length, scale and resolution. However, in our experience, we have found that accuracy of Structure-from-Motion computation is expected to be higher with an accurately calibrated setup (Irschara et al., 2007; Strecha et al., 2008).

In most of the calibration literature (Zhang, 2000), a strict requirement on the target geometry and a constraint to acquire the entire calibration pattern has been enforced. This is often a source of inaccuracy when calibration is performed by a typical end-user. Additionally, these methods tend to fail when images are taken at

Please cite this article as: M. Rumpler et al., Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.04.008

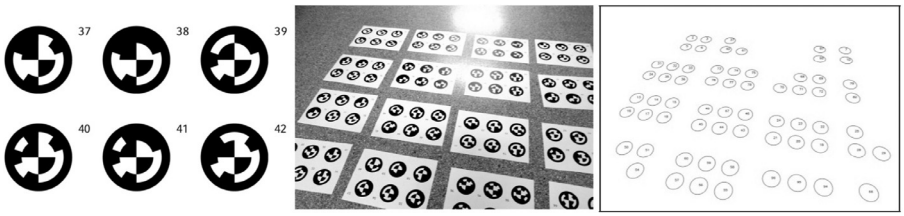


Fig. 2. Fiducial markers, typical calibration image with printed marker sheets arranged on the planar surface of a floor and reliably detected markers with center position and ID.

Fig. 3. Histogram for an unrolled circular marker and rotation invariant binning of the code stripe. The numbers from top to bottom indicate the probabilities for center, binary code and outer circle. The marker with ID 20 has been successfully decoded.

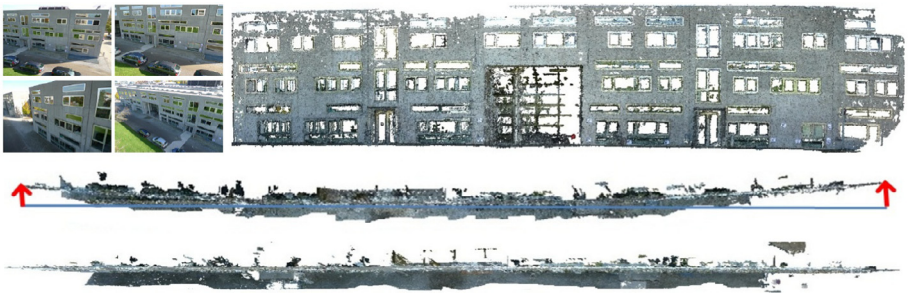


Fig. 4. Significant bending is prevalent in the facade reconstruction using OpenCV calibration (middle). In contrast, accurate camera parameters delivered by our method (Daftry et al., 2013) results in a straight facade reconstruction (bottom).

considerably different distances to the object. Hence, aiming at the accuracy of target calibration techniques while factoring out image space variations due to occlusion, reflection, etc., we advocate the use of a recently proposed fiducial marker based camera calibration method (Daftry et al., 2013). The calibration routine follows the basic principles of planar target based calibration and thus requires simple printed markers to be imaged in several views (Fig. 2).

Each marker includes a unique identification number as a machine-readable black and white circular binary code, arranged rotationally invariant around the marker center (Fig. 3).

A novel technique for robustly estimating the focal length and determining the calibration matrix K is employed, where an error function is exhaustively evaluated to obtain globally optimal values of focal length f, principal point and radial distortion. For details on the calibration routine please refer to the original paper (Daftry et al., 2013).

There are significant qualitative and quantitative benefits of the presented calibration method towards a multi-scale robust image sequence. Fig. 4 shows a reconstruction of a facade that, although visually correct in appearance, suffers from geometric inconsistencies (significant bending) that is prevalent along the fringes when using standard calibration and undistortion results from OpenCV (Bradski, 2000). In contrast, our method results in an almost straight wall.

In our findings, this method works very robustly and performs much better for a multi-scale image sequence acquired at varying depths to the object, as compared to traditional methods that employ a non-linear minimization technique for intrinsic parameter estimation. In addition, we facilitated the method with an easy to use GUI. Our calibration software is available online.[1]

2.2. Structure-from-motion

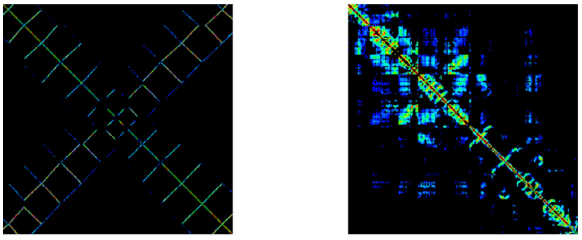
Calculation of the exterior camera orientations called Structure-from-Motion (SfM), or Aerial Triangulation (AT), includes feature extraction and feature matching, estimation of relative camera poses from known point correspondences and incrementally adding new cameras and computation of 3D object coordinates of the extracted feature points (Hartley and Zisserman, 2004). Camera orientations and 3D coordinates of the object points are then optimized using bundle adjustment.

For our method we assume pre-calibrated images, i.e. images that have already been undistorted together with an initial guess of the focal length (see Section 2.1). We group all input images into subsets sharing the same camera and focal length in a preprocessing step. The grouping and assignment to an initial calibration and focal length is performed according to meta information from specific tags provided with the image file (e.g. Exif information in JPEG or TIFF images), or given by the user.

The first processing step in our pipeline is feature extraction on every image in all subsets. A variety of methods exist for automated detection of feature points. The scale-invariant feature transform (SIFT) (Lowe, 2004) proved to be very robust against rotation, illumination changes and view point variations and scaling. It is therefore ideally suited to match images automatically from

[1] https://aerial.icg.tugraz.at/

Please cite this article as: M. Rumpler et al., Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.04.008



(a) (b)

Fig. 5. Rows and columns of the epipolar graph represent individual cameras and their connections to each other based on shared feature matches and image overlap. (a) shows a traditional aerial survey with regular flight pattern, (b) the connections between cameras for an unordered oblique image data set.

different view points, i.e. aerial images from a UAV and terrestrial images as well as inside views of an object taken with different cameras into one single reconstruction. The only prerequisite is that there is overlap between the images showing sufficient texture and salient features that can be matched across the views. The extracted features for all images are then stored and further processed.

Matching of the extracted features is performed between all images and all subsets. Since we assume that no further information about the images is known, feature matching would require an exhaustive comparison of all the extracted features in an unordered image set between all possible pairs of images to preserve as many image measurements as possible for an object point. Exhaustive comparison in an unordered set of images, however, requires a lot of computation time and is the most time consuming step in every Structure-from-Motion pipeline. The expense related to correspondence search and matching is thus quadratic in terms of the number of extracted feature points in the scene, which can lead to a critical amount of time in data sets with several thousands of images. To speed up the correspondence analysis in large data sets, methods based on vocabulary trees are applied to achieve a rough pre-selection of similar image pairs (Nistér and Stewenius, 2006; Sivic and Zisserman, 2003). The computation time for feature extraction and matching is additionally reduced through the extensive use of graphics processing hardware (GPUs).

Established feature correspondences between images are then used to estimate the relative camera orientations between pairs of images. Geometric verification of the relative camera orientations is performed using the five-point algorithm (Nistér, 2003) within a RANSAC loop (Fischler and Bolles, 1981). Starting from an initial image pair, new images are incrementally added to the reconstruction using the three-point algorithm (Haralick et al., 1991). The relative orientations between cameras are represented in a graph structure, the so-called epipolar connectivity graph (Fig. 5). Images in the graph are represented by the nodes and the relationships between them (based on common feature points and overlap) are represented by the edges of the graph that correspond to the relative orientations between cameras.

Camera orientations and triangulated 3D feature points are then simultaneously refined by minimizing the reprojection error between the projected 3D point and its corresponding 2D feature measurement in the image in a bundle adjustment step (Triggs et al., 2000). Optimization in the bundle adjustment is carried out based on Google's Ceres Solver for non-linear least squares problems Agarwal et al.

2.3. Automatic georeferencing

Reconstructions created by purely image-based approaches like the method described here are initially not metric due to the lack of scale information in the images. A metric scale of the reconstruction can be accomplished easily for example by one known distance in the scene. This might be a distance measure between two distinct points that is also easily recognizable in the digitally reconstructed 3D model, or a known distance between two camera positions.

However, in most cases and in surveying applications in particular, the absolute position of object points is important. In addition, we want the created 3D model stored and displayed in position and orientation in its specific geographic context.

Georegistration is achieved by a rigid similarity transformation (also called 3D Helmert transformation (Watson, 2006) or 7-parameter transform) of the model into a desired metric target coordinate system using at least 3 known non-collinear point correspondences between model points and points in the reference coordinate system (control points). A more robust transformation result is obtained by a larger number of points and a robust estimation of the transformation parameters for rotation, translation and scaling. The method of least squares within a RANSAC loop (Fischler and Bolles, 1981) improves clearly the registration quality of the model in the presence of noise and outliers.

2.3.1. Georegistration and GPS alignment

Flying platforms for aerial data acquisition are often equipped with a GPS receiver, that allows positioning of the aircraft in flight, stabilization and, depending on the application autonomous navigation between waypoints. Recording of GPS data during the flight enables to track and monitor positions and travelled distances of the UAV. It is then necessary to link the recorded images to the corresponding position data in order to use the GPS information for georeferencing (Nilosek and Salvaggio, 2012). This is achieved by synchronized timestamps of the images with the GPS signal. Several professional products instead offer a direct interface between on-board GPS receiver and camera to instantly assign a GPS location to a captured image and store the information in the meta data of the image file. Recorded information from inertial sensors may also be available in the meta data, providing information about the platform orientation of the aircraft at the time of capturing the image, given by the rotation angles for roll, pitch and yaw. During SfM, this additional information of approximate camera positions is used for guidance to speed up the search for neighboring images and match image features through guided matching.

Position data stored for each image is now used to metrically scale the previously calculated reconstruction and to transform the model into a desired reference system. However, the quality and accuracy of location data is not sufficient in most cases to allow an accurate three-dimensional reconstruction and reliable measurements in the scene solely based on GPS positions and IMU (inertial measurement unit) data. Due to weight restrictions of UAVs and a maximum payload depending on the used aircraft, usually very small GPS receivers are used that allow only limited accuracy in the range of 1-2 meters in position and 1-2° orientation accuracy (Pfeifer et al., 2012). But, the accuracy is sufficiently high for a rough positioning and metric scaling of the image-based reconstruction because transformation parameters can be estimated robustly when using a large number of images. The more images and GPS positions, the more robust the transformation gets. The accuracy of the absolute positioning of the reconstruction might be low, but the precision of the metric scaling is high enough, because relative position errors between GPS positions are better distributed and compensated, the larger the number of position measurements, i.e. the number of images.

Please cite this article as: M. Rumpler et al., Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.04.008

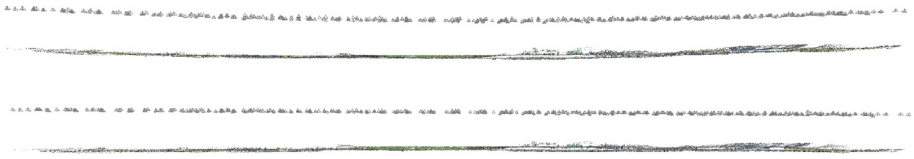


Fig. 6. Illustration of the bending-effect for the railway dataset. Camera positions and 3D points drift away from fixed control points due to systematic errors (top). The surveying area has an extent of about 2.1 × 0.6 kilometers. 403 images were taken with a senseFly eBee fixed-wing drone at a constant flying height of 85 meters above ground level. We used 19 ground control points (GCPs) to geo-register the scene. Errors caused by the bending in this dataset without external information resulted in positional shifts of 3D points and camera positions of up to 8 meters from their measured GPS position (bottom).

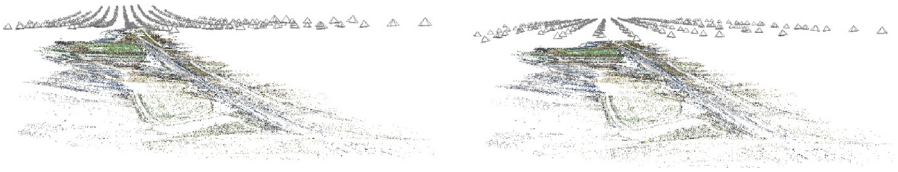


Fig. 7. Results of the photogrammetric reconstruction without (left) and with additional GPS positions and optimization of camera intrinsics in the bundle adjustment (right). The direct comparison shows the reduction of the initially clearly visible distortion of image block and object points.

Table 1
Quantitative accuracy improvements for the railway dataset using GPS information and ground control points (GCPs). With external constraints in the bundle adjustment the reconstruction error decreases from almost 8 meters to less than 25 cm.

	Error [m]		
Method	Mean	Std.dev.	Median
BA + rigid GPS alignment	7.272	2.170	7.514
BA + rigid GCP alignment	4.824	6.499	2.309
Constrained BA with GPS	2.708	0.789	2.622
Constrained BA with GCPs	0.247	0.550	0.232

2.3.2. Constrained bundle adjustment with GPS and ground control points

Pure image-based approaches suffer from systematic errors. We experienced that especially for a few datasets showing long elongated, large-scale scenes our pipeline resulted in large errors up to a few meters due to a deformation of the whole image block introduced in the bundle adjustment. Depending on the fixed reference point locations from the georegistration step, the errors drift away from these fixed points and cause a bending-effect like in the railway dataset shown in Figs. 6–7 and Table 1.

Observed camera block deformations are very often caused by incorrectly estimated radial distortion parameters of the camera. As a consequence the reprojections of 3D points onto the image plane are not correct and thus cause wrong error measures in the bundle adjustment process. Furthermore, the reprojection error as the sole evaluated error measure has impact on many independent parameters (3D positions of the object points as well as intrinsic and extrinsic camera parameters). Errors can be passed back and forth during the optimization and camera positions may undergo large changes.

These systematic errors can be avoided by either a more accurate initial camera calibration or by adding external constraints in the bundle adjustment. For photogrammetric applications, we therefore use (roughly) known GPS positions of the cameras determined by an on-board GPS receiver and fixed control points to allow for camera self-calibration within the optimization.

Georegistration of the reconstruction as described in the previous subsection alone does not solve this issue. The model deformations are still present due to the shape-preserving character of the transformation. Instead, after rough georegistration and GPS-alignment, we use known GPS locations of the images in an additional bundle adjustment step to constrain the positions of the cameras and to reduce an initial distortion of the image block. We do that by calculating the deviations of the calculated camera positions from the Structure-from-Motion result and penalize the deviation to their measured GPS positions in the optimization step. The influence of the deviation between the measured position is weighted by a Huber loss function (Huber, 1964). The camera positions can move only within a certain range around their measured positions and thus, are softly linked to their measured GPS positions. This leads to smaller residuals on the one hand, and on the other, a direct transition from the model coordinate system into a desired geographic reference system is accomplished simultaneously. The bundle adjustment step is again carried out based on Google's Ceres Solver for non-linear least squares problems Agarwal et al.

In addition, ground control points (GCPs) are also used to correct distortions or a small geographic misalignment of the model and to tie the reconstruction to a certain geographic position. Besides camera positions and 3D points we therefore use the GCPs also for self-calibration in the bundle adjustment step and optimize common intrinsic camera and distortion parameters for each camera group.

GCPs signal points that are usually easily recognizable natural or artificial landmarks in the scene. Their position is known and for example determined by means of conventional survey methods or DGPS (Differential Global Positioning System) with high accuracy. For this purpose, the bundle adjustment is further extended to the use of control points and their corresponding image measurements. The additional information is seamlessly integrated into the reconstruction process. The reprojection error between the image measurements and projected control points is additionally weighted and penalized in the bundle adjustment in a similar way to the mass of natural features obtained by the SIFT keypoint detector. Important in this case is an appropriate weighting that balances the residual reprojection errors of the GCPs compared to the SIFT-generated points. Usually, low number of GCPs (around a couple of dozen) is confronted with a large number of natural feature points (hundreds of thousands or millions of points).

For GPS positions and ground control points together, the optimization problem is defined as

$$f^* = \min_{\{R,t,K\},\{P\},\{R\},\{S\}} \sum_P E(P) + \lambda \sum_R E^{gcp}(R) + \omega \sum_S E^{gps}(S), \qquad (1)$$

with R, t as the rotation and translation of the cameras and K the intrinsic camera matrix. $\{P\}$, $\{R\}$ are the sets of natural 3D points and reference points represented by the GCPs, including their 2D image measurements. $\{S\}$ denotes the set of GPS measurements

Please cite this article as: M. Rumpler et al., Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.04.008

and the corresponding camera positions. E, E^{gcp} and E^{gps} are the error functions for 3D points, GCPs and GPS camera positions. λ and ω denote scalar weighting factors.

The error function for 3D points is defined as

$$E(X) = \sum_{x^i \in X_p} \rho(C^P(\Gamma_i(X), x^i)). \tag{2}$$

ρ denotes the robust Huber loss function, Γ_i projects a 3D point X into image I_i. For the point case C^P, a common choice is the 2D reprojection error defined as the Euclidean distance between the projected 3D point (Γ_i(X)) and the observed 2D measurement (x_i). GCPs are handeled in the similar way.

In case of GPS measurements for camera positions, Eq. 2 simplifies to

$$E(C) = \sum_{M \in S} \rho(C^{gps}(M, C)). \tag{3}$$

C^{gps} is the Euclidean distance in 3D between the measured GPS position (M) of the camera and the reconstructed camera center (C).

The weighting terms λ and ω in Eq. 1 are dynamically selected depending on the number of 3D points $|\{P\}|$, GCPs $|\{R\}|$ and GPS measurements $|\{S\}|$ to balance the weights between them such that all parts contribute equally.

Integrating both mechanisms (using ground control points and GPS positions of the cameras) distributes the residual errors equally over all cameras and object points and allows for 3D reconstructions with very low geometric distortions. Furthermore, in the case of regular camera networks we experience that an additional cross flight and images at different scales taken at different distances to the object help to stabilize the intrinsic camera parameters. 2D image measurements, feature matches across overlapping images and triangulated 3D points are then better constrained. This leads to a more robust self-calibration result and furthermore to a more stable image block and increased point position accuracy even for very large, elongated surveying areas (Figs. 6 and 7).

2.4. Surface reconstruction and texturing

The results of the previous steps so far are the external orientations of the cameras, optimized intrinsic camera parameters and a 3D point cloud from triangulated object feature points. Due to the comparably low number of triangulated feature points (approximately 5000 features per image, depending on the texture) and their non-uniform distribution on the surface compared to the dense number of pixels in one image (millions of pixels), the modeling of the surface is only an approximation of the real surface. To increase the number of 3D points, stereo (Hirschmueller, 2005) or multi-view methods (Furukawa and Ponce, 2009; Irschara et al., 2012) are used for pixel-wise image matching.

For better visualization and for further use as a digital surface model (DSM) for surveying tasks, we extract a closed surface from the point cloud using a method based on 3D Delaunay triangulation and graph cuts (Labatut et al., 2007). The method produces watertight triangle meshes from unstructured point clouds very robustly even in the presence of noise and gross outliers in the raw 3D sample points. The meshes can then be textured (Waechter et al., 2014) from the input images to generate a photorealistic representation of the scene. Fig. 1 shows a comparison between a sparse reconstruction, a densified point cloud and a reconstructed triangle surface mesh of a quarry wall consisting of about 10 million 3D points.

3. Data acquisition with user guidance

To evaluate the presented workflow and the achieved accuracy, several image flights were carried out to record datasets typical for urban, mining and archaeological applications.

We used different platforms and cameras to acquire each of the datasets. One is a Falcon 8 octocopter by AscTec, equipped with a Sony Nex-5N digital system camera or alternatively with a GoPro Hero 3+ action camera. The second flying platform is a senseFly eBee, a small fixed-wing UAV with a Canon IXUS 127HS compact camera. The main advantages of multi-copters are their flexibility and the ability to fly at very low airspeed to record datasets with high overlap, hover and observe objects from any possible position, even very close to an object to capture images at a very high level of detail. The fixed-wing UAV, however, is able to fly and survey large areas in short time with certain details not been detected due to the in general larger flying altitude and higher airspeed. In addition we use a Canon EOS 5D full-frame digital SLR and a consumer-grade Panasonic compact camera for terrestrial images in areas, where highly detailed views from the inside of an object are required and an airborne mission cannot be performed. A compiled summary of cameras and sensors used is given in Table 2.

To guarantee a certain accuracy, a desired image overlap and minimum ground sampling distance has to be defined beforehand. Based on Eqs. 4 and 5 for nadir image acquisition in aerial photogrammetry,

$$PixelSize = \frac{SensorWidth\,[mm]}{ImageWidth\,[px]}, \tag{4}$$

$$GSD = \frac{PixelSize\,[\frac{mm}{px}] * ElevationAboveGround\,[m]}{FocalLength\,[mm]}, \tag{5}$$

we estimate a maximum flying height above ground and imaging distance to the object, respectively.

The field of view (FOV) calculates from Eq. 6,

$$\alpha = 2 \cdot \arctan \frac{SensorWidth\,[mm]}{2 \cdot FocalLength\,[mm]}. \tag{6}$$

The scene coverage for one image captured from height h above ground is calculated from Eq. 7,

$$c = 2 \cdot h \cdot \tan \frac{\alpha}{2} \approx ImageWidth\,[px] \cdot GSD. \tag{7}$$

The baseline b between views then calculates from the overlap ratio $o_r = \frac{o}{c}$ with o being the overlap $o = 2 \cdot h \cdot \tan \frac{\alpha}{2} - b$ to $b = (1 - o_r) \cdot c$.

To enable analysis of which parameters influence the reconstruction accuracy we oversample the scenes and record images at the minimum of 70% overlap in previously defined distances and heights from the object.

Apart from the imaging distance, the baseline between particular cameras has a strong influence on the triangulation geometry and ray intersection. Especially for the canonical stereo configuration with parallel optical axes, the distance to baseline ratio is a good parameter to quantify the quality of a camera network. Small baselines lead to small triangulation angles and to high depth uncertainty. But to enable feature matching, high image overlap and intersection angles below 30° are optimal (Zeisl et al., 2009).

3.1. Online-feedback for image acquisition

We support image acquisition by an online feedback system to assess the recorded images with respect to the parameters of image overlap, ground sampling distance and scene coverage defined in the previous section to ensure completeness and redundancy of the image block.

Please cite this article as: M. Rumpler et al., Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.04.008

Table 2
Camera and sensor specifications (focal length given in 35mm equivalent).

Camera	Sensor size	Resolution	Focal length	Pixel size
Sony Nex-5N	$23.4 \times 15.6\,mm$	4912×3264	$24\,mm$	$4.76\,\mu m$
Canon IXUS 127HS	$6.16 \times 4.62\,mm$	4608×3456	$24\,mm$	$1.35\,\mu m$
Canon EOS 5D	$36.0 \times 24.0\,mm$	4368×2912	$24\,mm$	$8.24\,\mu m$
Panasonic DMC-TZ22	$6.2 \times 4.6\,mm$	4320×3240	$24\,mm$	$1.44\,\mu m$
GoPro Hero 3+	$6.25 \times 4.68\,mm$	4000×3000	$17\,mm$	$1.55\,\mu m$

The automated offline pipeline described in Section 2.2 yields high accuracy, as we will show in Section 4. However, processing takes several hours, thus, results are only available hours later. If the reconstruction result does not meet the desired expectations, e.g. due to a lack of images in areas that would have been relevant to the survey, a repetition of the flight is necessary which cause additional costs and time delays. In order to be able to already judge the results on site whether the captured images are suited for fully automated processing, we apply a recently developed method for online Structure-from-Motion (Hoppe et al., 2012). The method calculates the exterior orientation of the cameras and a sparse point cloud reconstruction already during or right after the flight on site.

We expect that a user does not acquire images in a totally random order. Further, we assume that a new input image has an overlap to an already reconstructed scene part. We can then split the SfM problem into two tasks that are easier to solve: localization and structure expansion (Irschara et al., 2009). More formally, given a freshly acquired input image I and a reconstructed scene M, we find the position of I within M and finally, we expand the map M. The presented method is similar to visual SLAM, but it matches wide-baseline features instead of tracking interest points. Since some of the features are already used for the triangulation of 3D points, we can directly establish 2D-3D image correspondences between I and M. Given a set of 2D-3D correspondences and a calibrated camera, we solve the absolute pose problem robustly in a RANSAC loop (Fischler and Bolles, 1981).

The images may be streamed down from the UAV via Wi-Fi to a laptop computer on the ground. High-resolution images can be processed in real time on a standard notebook computer. The method requires a stream of consecutively captured images and needs about two seconds to determine the outer orientation of a 10 megapixel image and calculating new object points. The online reconstruction does not claim high accuracy, because we restrict image matching and bundle adjustment to immediate neighboring cameras. However, the method allows online feedback to estimate the achievable reconstruction quality and is very beneficial to determine completeness of the final reconstruction during image recording.

For the user, the quality of the reconstruction can be judged only poorly from the triangulated sparse feature points. Two main relevant parameters determine the accuracy: redundancy, which states how often a surface point is seen, and the ground resolution. To determine both parameters from the actual reconstruction, a surface is incrementally extracted from the sparse points (Daftry et al., 2015; Hoppe et al., 2013). The surface extraction method is based on Labatut et al. (2007), which uses a Delaunay triangulation of 3D points and robustly labels the tetrahedra into free and occupied space using a random field formulation of the visibility information. Having defined all terms for our random field formulation, we are then able to derive a globally optimal labeling solution for our surface extraction problem using dynamic graph cuts. The surface is extracted as the interface between free- and occupied space.

We visualize the current ground sampling distance and image redundancy as quality indicators on the surface model to guide the user throughout the acquisition (see Fig. 8). For the user it is then obvious, how often parts of the scene are observed and at which ground resolution they can be reconstructed. This assists the pilot in planning the next steps of the flight so that a uniform coverage of the scene with constant ground resolution can be achieved.

To demonstrate the benefits of our feedback system, we asked a user experienced in image-based 3D reconstruction to acquire images with sufficient overlap (from the inside and outside) of a complex (non-convex shaped) building entrance which we then processed with our online SfM system. We performed the experiment twice: One time with user feedback and one time without. For the first experiment 100 images were acquired without feedback. 74 of them were successfully integrated in the reconstruction. 26% of the images could not be integrated, mainly because of missing features on the object through underexposure, and overexposure in the background which the user did not recognize during acquisition time.

With user feedback, missing correspondences between images and the reconstruction are instantly reported to the user (see Fig. 9). After three images that failed to be aligned, the acquisition strategy and camera settings where adapted successfully. In the end, 100 images out of 118 were successfully integrated into the reconstruction, which is a rate of 15% missed images compared to 26% in the experiment without feedback (Hoppe et al., 2012).

Fig. 10 shows a sample image of a reconstructed building overlaid by the redundancy and resolution information. The images were acquired by a manually controlled UAV. The redundancy map shows that images where mostly captured at the center part of the building whereas the resolution is distributed equally within the atrium of the building. The overall shape of the inner part of the building is extracted correctly, but the outer parts are missing. The visualization of the 3D geometry and color coding helps to select new camera positions and obtain a complete 3D model of the building.

Quantitatively, our method achieves the same accuracy as state-of-the-art methods but reduces the computational effort significantly. The difference in computational effort is mainly caused by the definition of the energy function. Other methods such as Labatut et al. (2007) have to perform a full raycast for each ray, the used methods from Daftry et al. (2015); Hoppe et al. (2013) only have to identify the tetrahedra in front and behind the vertex and the first triangle that is intersected by the ray. Hence, the combination of the dynamic graph cut with the energy formulation of Daftry et al. (2015) allows to extract the surface from an increasingly growing point cloud independent of the overall scene size in real-time.

4. Experiments and results

In this section we analyze the performance of the presented workflow based on different datasets. For our investigations we chose five different test sites: A facade dataset of a building in an urban environment, two datasets are showing mining applications. One of them is located at the "Styrian Erzberg", another one is a small gravel pit situated in Upper Austria. Furthermore, we

Please cite this article as: M. Rumpler et al., Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.04.008

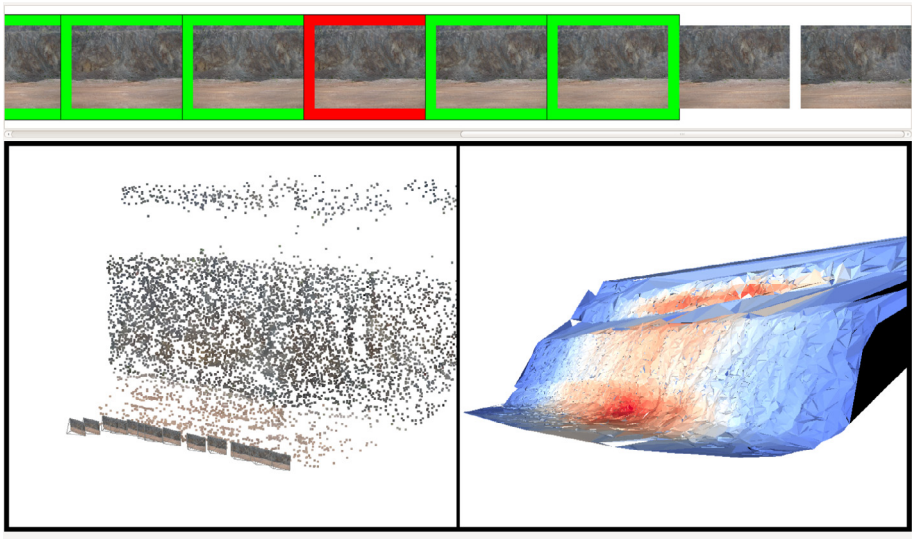


Fig. 8. Visualization of ground resolution using an online Structure-from-Motion system to assess reconstruction accuracy and scene coverage during image acquisition (color coding: red = high, blue = low resolution).

Fig. 9. Images that could not be aligned to the reconstruction are indicated by a red frame around the image (left). With feedback the user can instantly recognize problems and adapt the image acquisition strategy to recover image alignment (right).

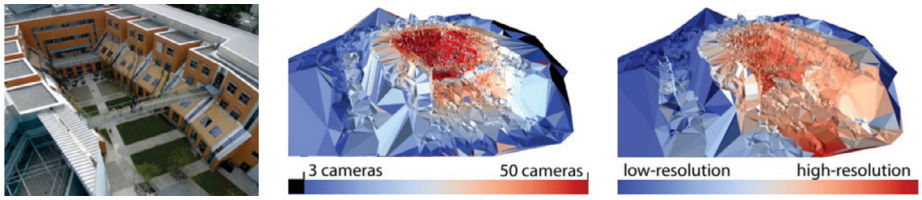


Fig. 10. Sample image from a building acquired by a manually controlled UAV (left). Visualization of the redundancy map (middle) and resolution map (right).

recorded two archaeological sites, one in Valcamonica, Northern Italy and one in Side, Turkey.

For photogrammetric applications the accuracy of reconstructed object points is of prime interest. Thus we perform a point-wise comparison of reconstructed object points to corresponding, clearly identifiable 3D reference point coordinates. The two mining sites are therefore equipped with a dense network of ground truth points to assess the quality of the reconstruction. For both the facade dataset and the archaeological site in Italy, we have dense surface scans for a ground truth comparison.

4.1. Urban facade dataset

We acquired a dataset (Graz500 Multi-scale Facade) with a multi-scale image sequence of an outdoor facade scene consisting of 500 images (Daftry et al., 2015). For image acquisition we used the Falcon octocopter from AscTec as a flying platform. The overview of the multi-scale camera network design is depicted in Fig. 11. Images were acquired at different depths, heights and viewing angles to the facade using the online feedback method described above. The dataset thus also offers an opportunity for detailed analysis and evaluation of various factors in image-based facade reconstruction. The dataset is publicly available online.[2]

We have acquired accurate terrestrial laser scanning (LIDAR) data, having a GSD of 1.5 cm and point measurement uncertainty of 2 mm using a Leica Total Station that will serve as geometrical ground truth to evaluate the quality of the image based reconstructions. The data was acquired in a single scan and hence does not involve any irregularities due to scan registrations. To assess the achieved absolute accuracy in 3D, the facade (which is about 30 m high and 50 m long) is equipped with a reference network (17 fiducial targets) of ground control points. The ground truth data for each GCP is measured using a theodolite and has an uncertainty of less than 1 mm.

Multi-scale camera network. In 3D reconstruction literature, the distance to the reconstruction object has always been considered an important and contributing factor but seldom has been studied in an empirical way. The closer we go to the object, more fine details are captured and more information is gained, and thus, accuracy is improved. However, our experience with Structure-from-Motion has shown that also drift increases when we get closer to the object. A bending of the reconstruction is introduced (see Section 2.1 and 2.3.2). We performed a systematic study on the ground control point accuracy with respect to distance of image acquisition from the facade and ground sampling distance. Our facade dataset was further quantified into 3 row-subsets based on the distance of acquisition from close to distant (4 m, 6 m and 10 m), and reconstruction was performed on each subset independently using the proposed pipeline.

[2] https://aerial.icg.tugraz.at/

Please cite this article as: M. Rumpler et al., Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.04.008

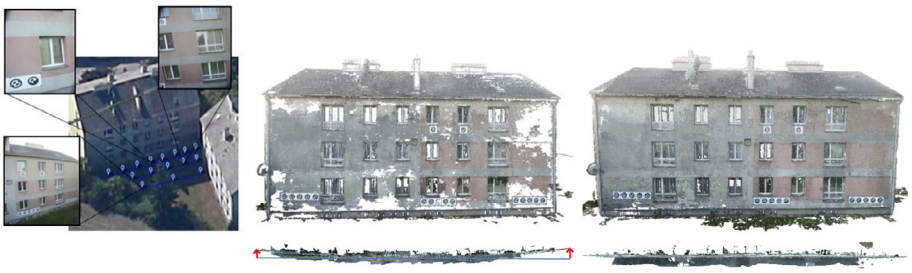


Fig. 11. Left: Graz500 multi-scale facade dataset. Middle: The reconstruction of a facade computed with a well-known state-of-the-art method (Snavely et al., 2008) shows, though visually correct in appearance, holes and geometric inconsistencies (significant bends along the fringes). Right: In comparison, using our multi-scale approach and online feedback support results in a straight wall and complete scene reconstruction.

Fig. 12. Resolution (i.e. point density) for SfM results from individual row subsets at distance: 4 m (left), distance: 6 m (middle) and distance: 10 m (right). A larger number of finely textured feature points are only visible in the close-by images due to a higher GSD.

Table 3
Accuracy results on the surveyed ground control points for individual row subsets and for the complete multi-scale camera network.

	Mean Error [mm]		
Images	Ours	Bundler	AgiSoft
Near Row (4 m)	**45.2**	51.3	57.1
Middle Row (6 m)	**23.1**	27.2	32.3
Far Row (10 m)	**5.7**	11.2	16.1
Multi-scale BA	**9.1**	15.5	21.6

4.1.1. Resolution versus high accuracy

It can be observed from the results in Table 3 that the mean absolute error on the GCPs decreases significantly as we go further away from the facade. Images taken further away from an object reduce camera drift and bending. This is contrary to the belief that the closer one gets to the object i.e. the higher the resolution the greater will be the accuracy. Thus after an exhaustive evaluation and study of various parameters we can state that the influence of the geometric configuration of the multi-view camera network on the resulting accuracy is very high and there is a significant accuracy gain as we go away from the facade. This is because of the strong drift effect caused in the camera pose estimation when the distance between the camera and the object is very small. However, we also observe that as we go closer to the facade the point density of the reconstruction is greatly improved as can be seen in Fig. 12. This is because a larger number of finely textured feature points are only visible in the close-by images due to a higher GSD. It can be thus concluded that there is a trade-off between accuracy and resolution (i.e. point density) as we change the distance between the image acquisition and facade. We generalize this as a systematic behavior as they can also be consistently observed in the standard software packages. Hence, we infer that a model incorporating the knowledge of this trade-off could help in improving the metric accuracy of the final reconstruction.

In order to give a full quantitative evaluation of the influence of our interactive SfM framework on reconstruction accuracy we compare our methodology to state-of-the-art pipelines using ground truth 3D data. The Bundler (open source) (Snavely et al., 2008) and Agisoft[3] (commercial) software packages were used as our primary reference, as they represent the most popular methods for SfM within the computer vision community.

4.1.2. Absolute error

We perform a point-wise comparison and evaluation. First, we calculate the absolute mean over all the points as the 3D Euclidean distance between the corresponding ground control point and the reconstructed and georeferenced point (see Fig. 13).

Multi-scale camera network benefits. We extend our evaluation to quantitatively and qualitatively assess the benefits of a multi-scale camera network based acquisition when applied to incremental 3D reconstruction methods. Experiments on accuracy evaluation are performed with and without a multi-scale network based matching and bundle adjustment. The results of the experiments are shown in the last row of Table 3. We observe that the proposed multi-scale camera network using the constrained bundle block formulation helps to overcome drift. It facilitates accurate reconstructions without compromising on scene completeness. The qualitative benefits on the geometric fidelity of the reconstruction has been shown in Fig. 11. As a ground truth, we know that the reconstructed wall of the facade should be straight. However on a detailed inspection, we clearly see that the reconstructed wall suffers from significant bending using a uni-scale acquisition approach, owing mainly to the drift due to map building in an incremental SfM framework. In contrast, the use of a multi-scale approach helps to constrain the bundle block from deformation due to error accumulation and consequently results in an accurate and complete reconstruction.

4.1.3. Relative error

Next, we calculate the one way Hausdorff distance (similarity measure) between the reconstructed point cloud (densification was done using PMVS (Furukawa and Ponce, 2009)) and the point cloud

[3] http://www.agisoft.com/

Please cite this article as: M. Rumpler et al., Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.04.008

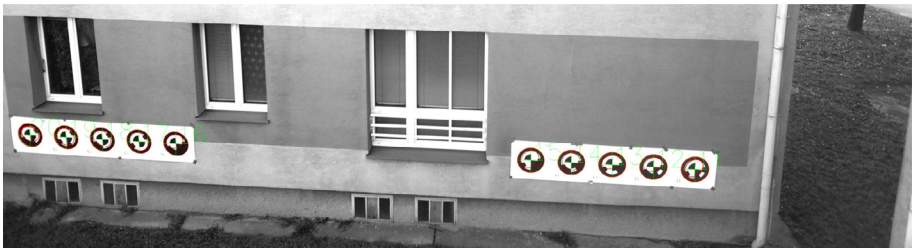


Fig. 13. Automatically detected ground control points with plotted marker centers and corresponding marker IDs.

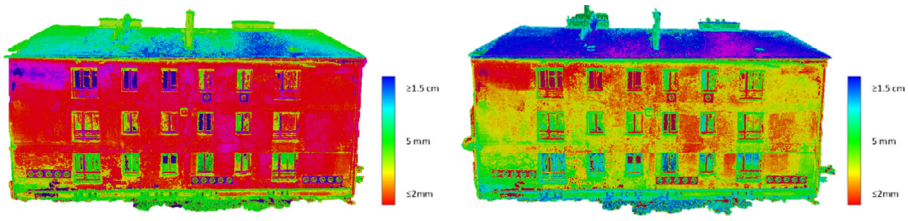


Fig. 14. Color coded dense 3D point clouds based on Hausdorff distance obtained using Ours (left) and Bundler (right).

Table 4
Time performance.

Time Performance	Time Taken [sec]		
	Ours	Bundler	AgiSoft
SfM	**880**	6220	6455

obtained from the laser scanner. The number of points in the reconstructed point cloud was about 9 million with a GSD of 1 mm.

Similar steps were performed for the sparse point cloud obtained from the Bundler software. The reconstructed point clouds were then color coded based on the Hausdorff distance. The results are shown in Fig. 14.

It can be seen that using our method the absolute mean error for surveyed GCPs is in the range of 9 mm and the overall relative accuracy of 90 % of the facade is within 2 mm error range with respect to the laser point cloud, which is within the uncertainty range of the total station. A closer inspection reveals that the high errors are only along the sections of the point cloud missing in the laser scanner such as roof, missing window panes, etc. Thus, we observe that our method considerably outperforms the state-of-the-art methods in both the absolute and relative error analysis to get highly accurate results comparable to the uncertainty of the laser point cloud.

4.1.4. Time performance

To evaluate the performance of the online SfM approach, we compare the runtime of the presented online SfM to a state-of-the-art batch-based SfM approach. For both methods, we use 5000 SIFT features per image with the largest scale. The features are extracted by the SiftGPU[4] implementation. Our approach requires 880 seconds to process all 500 images of the dataset, which is 7.1 times faster than Bundler, see Table 4. On average, our approach requires 1.75 seconds to integrate a new image into the structure and to extend the map. This is within the latency of the time constraints of the UAV to transmit back a new image from the next possible camera network position, and hence we conclude that the online SfM method is approximately in real-time.

[4] http://www.cs.unc.edu/~ccwu/siftgpu/

4.2. Mining datasets

Here we will investigate what are the relevant parameters determining accuracy in general and try to answer the following questions: How does accuracy increase with the use of external information in the reconstruction process given by ground control points and, how many control points are necessary to achieve satisfactory results with respect to absolute position accuracy and how should they be distributed.

As reference points in the two presented mining datasets we use binary coded, individually identifiable fiducial markers (Rumpler et al., 2014) printed on durable weather proof plastic foil, already introduced in Section 2.1. In addition, non coded red circular targets are used to densify the reference network in certain parts of the two mining datasets. Different subsets of the points are used as ground control points (GCPs) for automated georeferencing, and others are used as check points (CPs) to evaluate the achieved accuracy. All reference points were conventionally surveyed using a Trimble S6 total station with an average precision of 10 mm for 3D point surveying without prism.

Styrian Erzberg. The Styrian Erzberg is the biggest iron ore open pit mine in central Europe. Our test site represents one quarry wall, which is about 24 m high and 100 m long with the typical geometry of an open pit hard rock mine. It is equipped with 129 reference points with known ground truth positions. 45 are realized as fiducial markers on the top and bottom of the wall and on the adjacent benches and are used as temporary GCPs. Additionally, the wall is equipped with 84 circular targets, which are used to evaluate the reconstruction accuracy. This dense network (see Fig. 15) enables an extensive evaluation of accuracy and allows us to quantify systematic deformations of the image block and reconstructed 3D geometry.

Due to complex geometry and steep slopes at the test site, we used the AscTec Falcon 8 octocopter for image acquisition. Using the octocopter we were able to approach and hold any possible camera position, enabling the opportunity to acquire images under stable conditions for our further investigations. All together 850 images were recorded in different flying altitudes, viewing angles and distances to the object with a mean GSD of 1.5 cm.

Gravel Pit. Our second test site is a small gravel pit situated in Upper Austria. As shown in Fig. 16 the scene includes the actual pit as well as the surroundings and covers an area of about 0.43 km^2.

Please cite this article as: M. Rumpler et al., Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.04.008

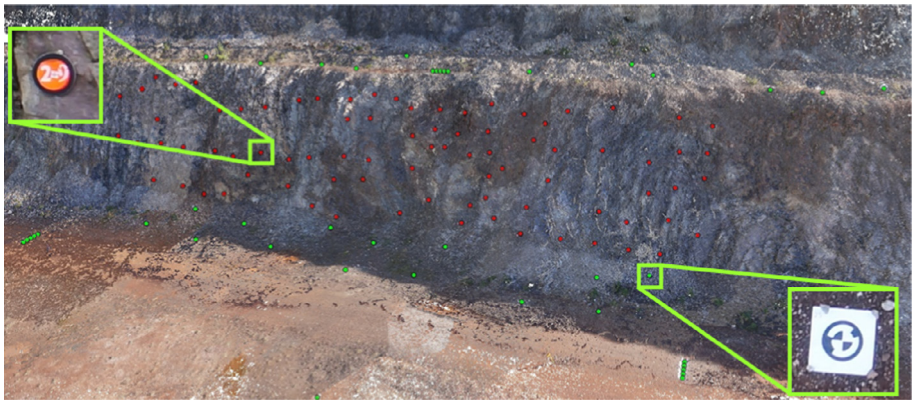


Fig. 15. The reference point network allows an extensive accuracy evaluation. Markers (right) indicating GCP positions are shown in green, circular targets (left) for quantitative evaluation are in red.

Fig. 16. Colored model of a gravel pit with surroundings.

Reference points are temporarily signalled in the same manner as described for the Erzberg dataset and are evenly distributed over the whole site. 27 control points are realized as fiducial markers and 19 as red circular targets. Additionally a small part of the pit was scanned at high level of detail (4 points per m^2) using the autonomous scan function of a Trimble S6 total station.

Images were recorded using a senseFly eBee fixed-wing UAV in different flying altitudes (75, 100 and 140 m). Due to camera specifications and higher elevation above ground the mean GSD is about 3.5 cm in this test scenario. The dataset consists of 533 images in total with an overlap within each altitude held constant at 70%. The resulting 3D model (Fig. 16) includes more than 400 million points and represents the scene at a level of detail not achievable with manual surveying methods.

4.2.1. Absolute position error

Fig. 17 shows the absolute point error for each check point of the Erzberg dataset, where a mean accuracy of less than 2.5 cm is reached using all 850 images and constrained bundle adjustment with GCPs.

Table 5 shows the improvement in accuracy by comparing the mean absolute point error for rigid similarity transform and optimization using GCP constrained bundle adjustment. The mean error which is already very good before GCP bundling then decreases further. The decreasing standard deviation indicates an equalization of the error distribution and a less deformed image block after the optimization.

For the gravel pit dataset an overall accuracy of 14 cm is achieved, primarily due to a higher flying altitude, a different camera with lower resolution (see Table 2) and different camera network.

Table 5
Accuracy improvement by GCP constrained bundle adjustment.

Method	Error [cm]		
	Mean	Std.dev.	Median
BA without GCPs	4.54	1.64	4.40
Constrained BA with GCPs	**2.45**	**1.18**	**2.16**

For a better understanding of block stability and accuracy we investigate in the following relevant parameters influencing the reconstruction quality. For this purpose, a high oversampling of the scene was performed, as already described in Section 3. Parameters with large impact on accuracy are, besides image overlap and triangulation angle, foremost the ground sampling distance determined by image resolution and imaging distance to the object and the distance to baseline ratio given by the camera network. In order to quantify the influence of these parameters and to give guidelines for image acquisition, a systematic parameter analysis is carried out based on different subsets of the previously described datasets.

4.2.2. Georegistration

One of the most important and critical steps with respect to the absolute position accuracy in the presented workflow is georegistration. Because of the fact that results of a Structure-from-Motion pipeline are initially in a local Euclidean coordinate system, georegistration or at least scaling has to be done every time, regardless

Please cite this article as: M. Rumpler et al., Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.04.008

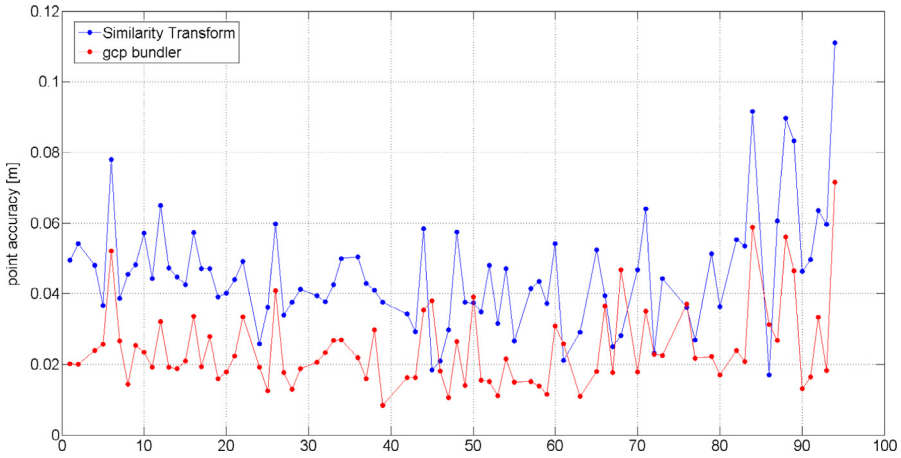


Fig. 17. Using all 850 images and all available GCPs in the constrained bundle adjustment, a mean measurement error dropping from 4.54 cm (bundle adjustment without GCPs) to below 2.45 cm (constrained bundle adjustment with GCPs) is reached (Rumpler et al., 2014).

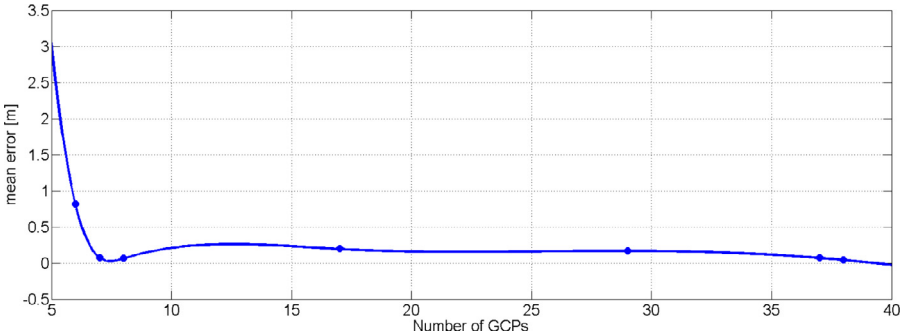


Fig. 18. The overall reconstruction error after pure rigid georegistration decreases very quickly and saturates at a low level after 7 to 8 ground control points. A higher number of GCPs is not necessarily needed for accuracy reasons. The GCPs were sampled randomly each time. The slight performance drop and increasing error after the minimum at 7 GCPs results from a non-optimal selection of the randomly chosen control points within the scene.

of how images are recorded. As already mentioned, accurate georegistration is possible by integrating GCPs in the bundle adjustment. The number of points and their spatial distribution within the scene strongly affects the achievable accuracy. Fig. 18 shows that the error clearly decreases with an increasing number of GCPs, but it is also apparent that even a small number of seven or eight GCPs is sufficient to get good results. In our case studies, adding more GCPs does not necessarily improve the result with respect to the overall accuracy.

Regarding the spatial distribution, the GCPs should be evenly distributed over the whole scene, especially concerning the height-component. Height tie points are at least as important as control points of position. If for example all GCPs are along one row systematic deformations can be observed because the reconstruction can tilt around that axis. Moreover, in contrast to traditional bundle adjustment approaches (Kraus, 1994), control points should be situated not entirely at the boundaries of the scene, because of less image coverage and a weak triangulation network. To guarantee a desired accuracy the used ground control point should be robustly detected in at least 10 images.

Our investigations also show that georeferencing using GPS information of the aircraft exclusively without any additional position constraints is not sufficient for surveying tasks with respect to the absolute pose of the reconstruction. Indeed, integrating a large number of camera positions in the reconstruction process mitigates systematic deformations of the image block and might result in highly precise metric scaling, but it is not possible to achieve absolute position accuracies below the meter range due to the high uncertainty of the small on-board GPS sensors on UAVs.

4.2.3. Number of observations

Rumpler et al. (2011) shows in a synthetic simulation on a traditional regular aerial flight pattern that accuracy increases with a higher number of image measurements and with increasing triangulation angles. Fig. 19 derived from the Erzberg dataset including oblique views shows as well, that the mean object point error decreases with increasing total number of images used for the reconstruction. But it is also obvious, that there is a fast saturation in accuracy improvement within larger datasets.

Thus, a higher number of images in the dataset leads to an accuracy improvement, but considering the number of image measurements per reference point does not necessarily reduce the error, as already shown in Rumpler et al. (2014). In contradiction to synthetic results of Rumpler et al. (2011), it is not possible to exemplify the achievable accuracy alone on the number of used images or observations for unordered and oblique datasets. The changing camera configuration influences feature matching, triangulation angle and ray intersection geometry, and from this we argue, opposing to Fraser (1996), that not every additional image measurement necessarily leads to an improvement in accuracy in practice with real world image data.

4.2.4. Camera network and resolution

We have shown that the influence of geometric configuration of the multi-view camera network on the resulting accuracy is higher than the influence of redundancy in image acquisition. In this section we present further investigations on the influence of camera network and resolution and compare a terrestrial dataset with different aerial networks for the Erzberg scene.

Please cite this article as: M. Rumpler et al., Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.04.008

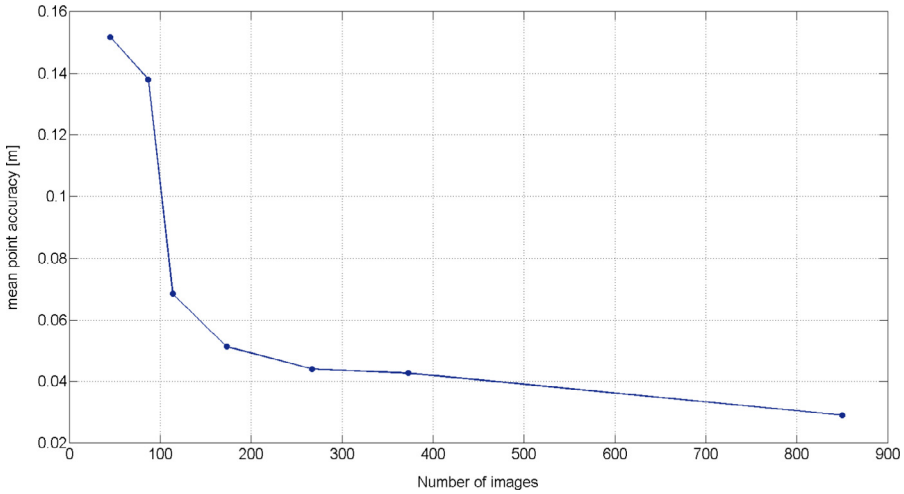


Fig. 19. Error curve for different image subsets. With increasing total number of images used for the reconstruction, the mean point error decreases.

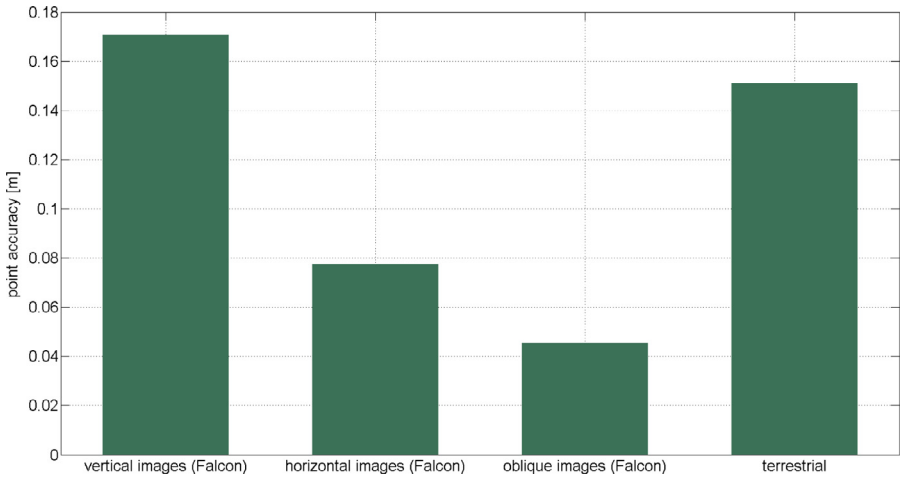


Fig. 20. Adjustable camera angle, low airspeed and high image overlap using a multi-rotor UAV for image acquisition enables best results.

The ground sampling distance, or resolution respectively, has a strong influence on the achievable accuracy. The uncertainty of a point in 3D increases with increasing distance to the camera, thus images that are further away introduce larger errors. First, this is because of a lower ground sampling distance, and thus, lower level of detail in the images. Secondly, the influence of localization errors on the reconstruction uncertainty increases with point depth. Image noise is approximately constant for all images, however, the resulting positional error increases with larger distances due to a larger angular error and smaller triangulation angles. See Eq. 8 with b being the baseline, f the focal length, d the disparity and z the the point depth.

$$\epsilon_z \approx \frac{bf}{d} - \frac{bf}{d+\epsilon_d} \approx \frac{z^2}{bf}. \tag{8}$$

Fig. 20 shows the mean error for all targets of the Erzberg dataset with respect to the different subsets. It clearly shows that the viewing angle has to be carefully adapted to the object geometry. Using exclusively vertical images, the steep wall is shadowed and the mean error increases to 17.1 cm. The smallest error is achieved using a combination of different views (vertical, horizontal and oblique), which is only possible by using a multi-copter UAV. Because of the adjustable camera angle and low airspeed, images can be always optimally adapted with respect to the surface geometry and a high overlap and level of detail can be achieved easily.

It is apparent that pure terrestrial photogrammetric systems are not flexible enough compared to data acquisition with UAVs. Because of imaging positions bound to ground level it is mostly not possible to observe the object completely or from a certain distance or view point due to geometric or safety reasons, especially in hazardous environments. The combination of different distances and image resolutions in a multi-scale camera network also affects the achievable accuracy positively. Images taken further away mitigate the error propagation within the first row, they help connecting the camera network over longer tracks reducing drift and the image block is stabilized. In general, flying at different altitudes is a common approach in airborne photogrammetry, to optimize the intrinsic camera parameters, which furthermore also results in better reconstruction accuracy.

4.3. Turkey

Our next test site is an archaeological excavation in Side, Turkey, where we show a qualitative performance analysis. The site shows complex geometry with arches, partly collapsed walls and chambers. We used an AscTec Falcon 8 equipped with a Sony Nex-5N camera for aerial image acquisition, together with

Please cite this article as: M. Rumpler et al., Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.04.008

Fig. 21. Image acquisition with an AscTec Falcon 8 octocopter for archaeological site documentation and reconstruction.

Fig. 22. Rendered views from an automatically reconstructed and textured 3D model of an archaeological excavation site in Side, Turkey, obtained from 4.722 terrestrial and aerial images captured with 3 different cameras from the air and from the ground.

terrestrial images in areas which could not be observed from the air (Fig. 21).

The terrestrial images were recorded from the inside and outside of the object with a Canon EOS 5D DSLR with a 24 mm wide angle lens for high resolution terrestrial images and a small consumer-grade Panasonic DMC-TZ22 zoom camera.

We took 5.014 images within four days in total with all three cameras, giving 38.4 GB of raw image data. Aerial images were captured in a classic raster flight pattern with cross flights in two different heights (40 and 90 meters above ground with a minimum overlap of about 80%) and in a hemisphere flight around the object with tilted camera to ensure enough overlap with terrestrial images for automated matching. We were able to align 4.722 images fully automatic into one single reconstruction of the site. Seven markers as ground control points were used to georeference the model. An overview image of the reconstruction together with detail views of the object are presented in Fig. 22.

4.4. Italy

We use this experiment to evaluate three factors of the acquisition pipeline. First, we analyze the reconstruction accuracy using multiple combinations of cameras. Then, we benchmark the georeferencing performance of our system using the fiducial markers. Finally, we demonstrate the benefit of the optimization of the intrinsic camera parameters for the reconstruction accuracy as well as the georeferencing performance.

4.4.1. Dataset details

This dataset shows a rock formation (Seradina R.12C) surrounded by vegetation in the region of Valcamonica in Northern Italy, shown in Fig. 23. The rock surface is covered with prehistoric rock carvings and has a size of approximately 17 × 13 meters. A ground truth mesh of the rock was obtained through terrestrial laser scanning (TLS) by Arctron3D[5]. The mesh has a resolution of 8 mm edge length and the accuracy of the laser scanner (Riegl VZ-400) is 5 mm.

For registering the ground truth mesh and the SfM reconstructions, we used a local coordinate system with four surveying points around the rock surface. First, the four points were measured with the laser scanner and then the points were used (a year later) for positioning a Leica total station. This second total station was then used to measure the center of 13 fiducial markers, which were placed circular around the rock (see Fig. 23).

For image acquisition, we used a UAV (Asctec Falcon 8) and three different cameras (Sony Nex-5N, Panasonic DMC-TZ22 and a GoPro Hero 3+). Two cameras were simultaneously mounted on the UAV (GoPro and Nex-5N) and with the third camera (TZ22) we acquired images in a hand-held manner by walking circularly

[5] http://www.arctron.de/en/

Fig. 23. Rock 12C in Seradina, Valcamonica, Italy. 13 fiducial markers were placed around the rock.

around Rock 12C. During the acquisition 601 images were taken with the Nex-5N camera in a regular time-interval of 2 seconds. The GoPro was operated with 24 frames per second at a resolution of 1920x1440. The video was stored with lossless compression and shows a high signal-to-noise ratio. From the 20 minutes video footage from the GoPro camera we also extract frames in the same time interval, which resulted in roughly the same number of images. With the hand-held camera (TZ22) 498 images were acquired.

The whole SfM pipeline was executed independently four times with different sets of images. It is executed once for each of the cameras mounted on the UAV (Nex-5N and GoPro separately), once with the images from Nex-5N and TZ22 combined and finally with all available images (Nex-5N, GoPro and TZ22 combined).

4.4.2. Reconstruction accuracy

After the automatic georeferencing both reconstructions were aligned with the ground truth using the iterative closest point (ICP) implementation of CloudCompare.[6] This eliminates any errors introduced by the georeferencing and allows for a fair comparison of the reconstructions using the ground truth mesh. In Fig. 24 we show the TLS ground truth of Arctron as well as the resulting sparse reconstructions. In Fig. 25 we show the absolute error between the sparse reconstruction and the ground truth mesh. The corresponding error histogram is shown in Fig. 26.

The experiment shows that the SfM Pipeline is quite flexible to the input data and can easily use multiple cameras in a coherent way. The comparison between GoPro and Nex-5N shows the expected effect. On the one hand the GoPro reconstruction shows more of the area surrounding the rock, on the other hand the reconstruction error is approximately twice as high. This factor of two in the uncertainty is very likely due to the fact that the GoPro images are roughly half the size of the Nex-5N images. From the experiments with the TZ22 camera, it can be seen that the combination of aerial images with the hand-held images can significantly boost the reconstruction accuracy. While the vast majority of 3D points for the Nex-5N reconstruction show an accuracy of below 2 cm, the accuracy for the Nex-5N + TZ22 reconstruction lies clearly below 1 cm.

4.4.3. Georeferencing

For all four reconstructions in the previous experiment the automatic georeferencing was accurate enough to work as a sufficiently good initialization for the ICP alignment. In Fig. 27 we show the error distribution of the reconstructions without performing ICP alignment. Even without the ICP alignment the absolute reconstruction error stays clearly below 8 cm in all cases. With the images from the TZ22 cameras the error can even drop below 3 cm. This suggests that registration based on fiducial markers is accurate enough for many applications.

4.4.4. Optimization of camera calibrations

For the cameras Nex-5N and TZ22 the initial camera calibrations were already very accurate (reprojection error approximately 0.1 pixel), but for the GoPro camera with its extreme wide angle lens the reprojection error after calibration was quite significant (larger than 3 pixels). In the case of a good initial calibration the optimization of the higher order distortion parameters (especially the radial distortion) does not result in significantly higher reconstruction accuracy, but in the case of a bad initial calibration the effects are quite drastic. In Fig. 28 we show the benefit of the optimization of the higher order distortion parameters (radial distortion). In this example the reconstruction accuracy is improved by a factor 10.

In Fig. 29 we show the impact of the camera self-calibration routine on the automatic georeferencing. With the routine the absolute error drops by a factor of approximately three.

[6] http://www.cloudcompare.org/

Please cite this article as: M. Rumpler et al., Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.04.008

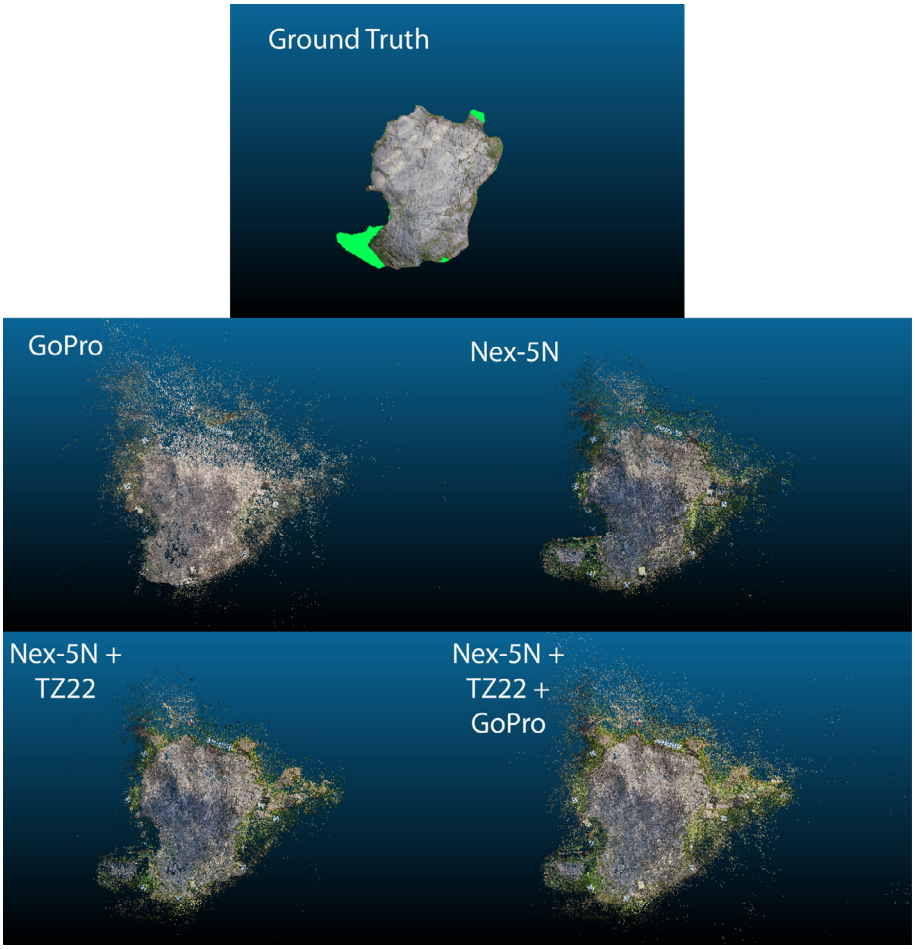


Fig. 24. Seradina Rock 12C. Top: Terrestrial laser scanning (TLS) ground truth mesh. Middle and Bottom: Sparse reconstructions using images from different cameras.

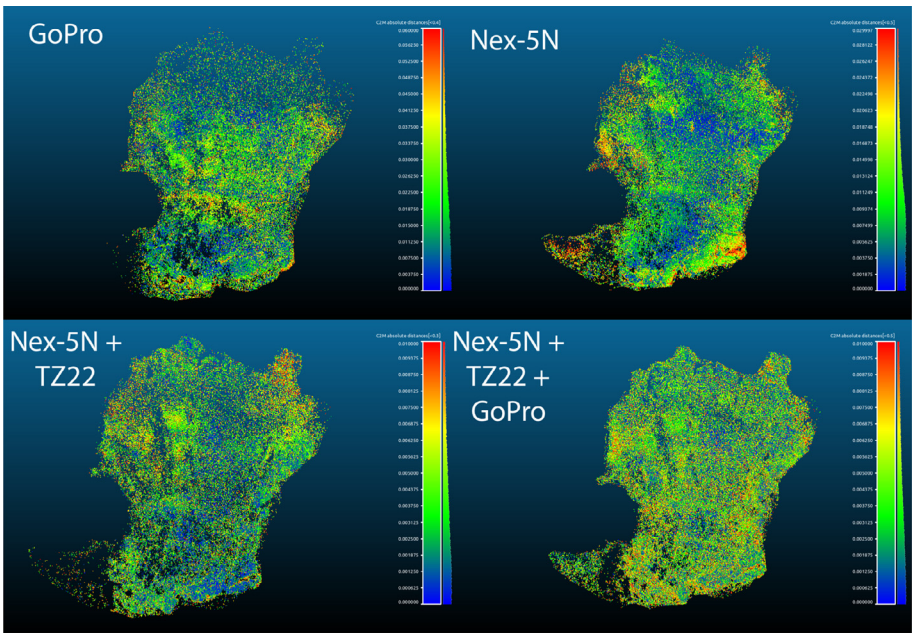


Fig. 25. Absolute error of the sparse reconstructions with images from different cameras. Note that the color is scaled differently for the reconstructions (top left: max = 6 cm, top right: max = 3 cm, bottom: max = 1 cm).

Please cite this article as: M. Rumpler et al., Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.04.008

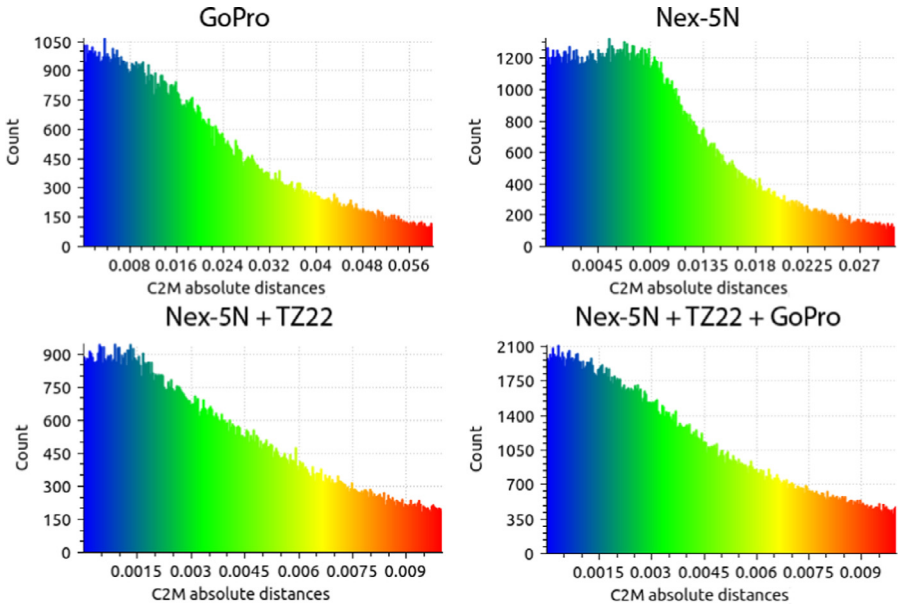


Fig. 26. Error histograms of the sparse reconstructions with images from different cameras. Note that the ranges are scaled differently for the reconstructions (top left: max = 6 cm, top right: max = 3 cm, bottom: max = 1 cm).

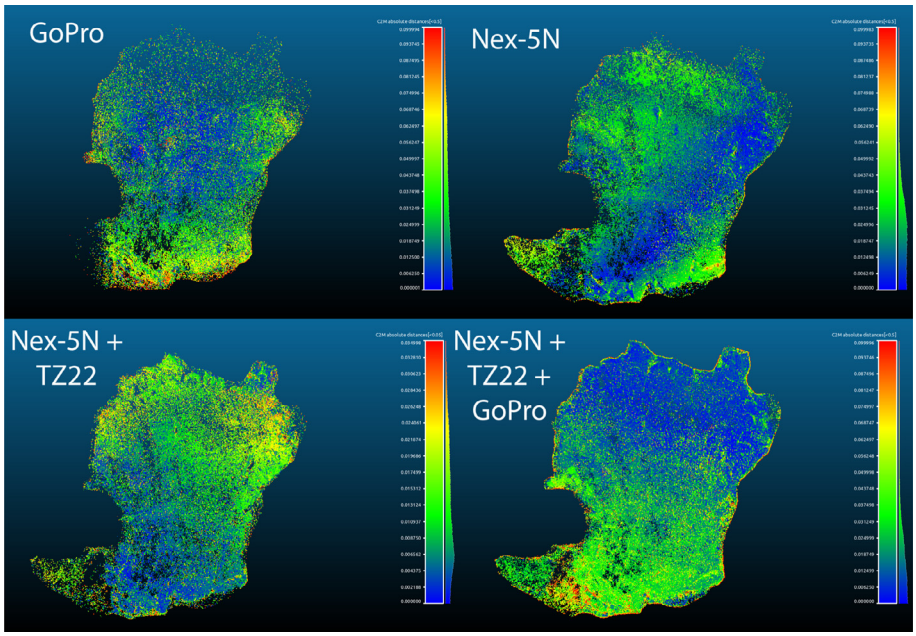


Fig. 27. Comparison of sparse reconstructions with different cameras directly after automatic georeferencing. Note that the bottom left figure is scaled differently (max. 3.5 cm opposed to 1 cm). In all cases the error stays clearly below 8 cm.

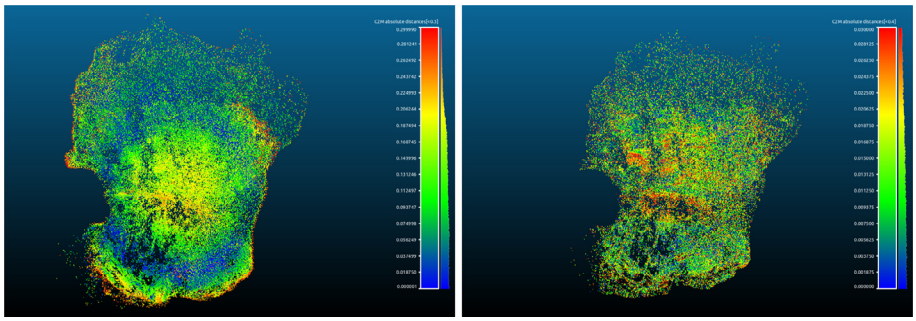


Fig. 28. Benefit of optimizing the radial distortion in the bundle adjustment on the example of the GoPro reconstruction. The absolute error to the ground truth is shown after performing ICP. The left figure shows a reconstruction without, and on the right with radial distortion optimization. The reconstruction accuracy is improved by a factor of approximately 10. The circular error distribution in the left figure indicates that the reconstruction is bent, which can be observed if the estimate of the radial distortion is off.

Please cite this article as: M. Rumpler et al., Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.04.008

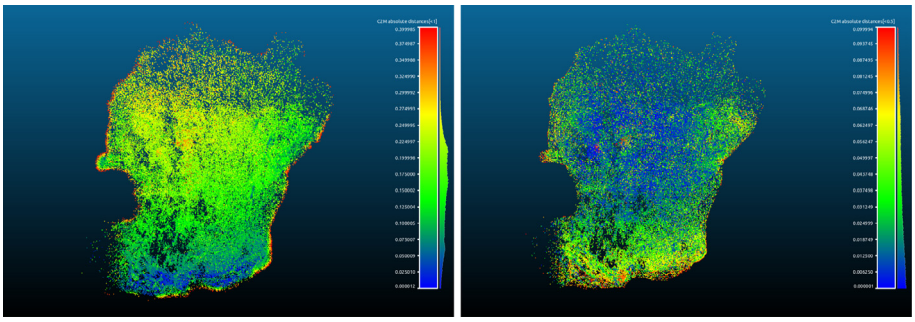


Fig. 29. Benefit of the camera self-calibration routine for georeferencing. The absolute error to the ground truth is shown directly after georeferencing. The left image shows a reconstruction without, and on the right with performing camera self-calibration. With the routine the absolute error drops by a factor of approximately three.

5. Conclusion

In this paper we presented a fully automated processing pipeline for precise, metric and geo-accurate 3D reconstructions of complex geometries using various imaging platforms. Firstly, we advocated the use of planar fiducial markers as a target pattern to obtain accurate and reliable camera calibration. Secondly, we integrated an online feedback to guide the user during data acquisition regarding ground sampling resolution and image overlap to guarantee automated photogrammetric processing, that the final reconstruction meets predefined accuracy requirements and results in a complete reconstruction of the object. Lastly, we utilize known GPS positions and ground control points in the scene and integrate them into our image-based reconstruction pipeline. We use the additional information given by GPS and GCPs for self-calibration in the bundle adjustment step and optimize common intrinsic camera and distortion parameters for each individual camera group.

We show that combining these technologies, adapting the image acquisition strategy and the developments in UAV technology together can return metrically accurate data that has immense applications in architecture, engineering and construction domains.

Low and equally distributed mean point position errors are achieved when integrating additional external constraints in the bundle adjustment to avoid systematic deformations and bending of the reconstruction due to an initially inaccurate camera calibration. We showed that the reconstruction accuracy is not only influenced by ground sampling distance and the image overlap, but is strongly influenced by the structure of the camera network. Images taken further away cause larger errors, but when using only images taken from a very close view point to the object, the reconstruction is more affected by drift and distortions. Combining images taken at different distances, view points and viewing angles stabilizes the image block and mitigates the error propagation.

Although many investigations and concepts discussed in this paper including bundle block adjustment approaches, camera self-calibration or optimal distribution of control points are well known in photogrammetric literature for decades, we presented a best practice example for different use cases, engineered to state-of-the-art performance. Our approach is suited for seamlessly matching and integrating images with different scales from different view points and cameras into one single reconstruction.

Based on five different datasets for applications in mining, archaeology and urban environments, we evaluated our approach and demonstrated its flexibility and high accuracy.

Acknowledgements

This work has been supported by the Austrian Research Promotion Agency (FFG) BRIDGE programme under grant 841298 and EC FP7 project 3D-PITOTI (ICT-2011-600545). We further thank Ute Lohner-Urban, Peter Scherrer and the Institute of Archaeology, University of Graz.

References

Agarwal, S., Mierle, K., Others, Ceres solver. http://ceres-solver.org.
Bradski, G., 2000. The OpenCV Library.
Daftry, S., Hoppe, C., Bischof, H., 2015. Building with drones: accurate 3D facade reconstruction using MAVs. In: IEEE International Conference on Robotics and Automation, ICRA 2015, Seattle, WA, USA, 26-30 May, 2015, pp. 3487–3494.
Daftry, S., Maurer, M., Wendel, A., Bischof, H., 2013. Flexible and user-centric camera calibration using planar fiducial markers. In: British Machine Vision Conference (BMVC).
Fischler, M.A., Bolles, R.C., 1981. Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. Commun. Assoc. Comput. Mach. 24 (6), 381–395.
Fraser, C., 1996. Network Design. In: Atkinson, Close-range Photogrammetry and Machine Vision. Whittles Publishing UK, pp. 256–282.
Furukawa, Y., Ponce, J., 2009. Accurate, Dense, and Robust Multi-View Stereopsis. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI).
Haralick, R.M., Lee, C., Ottenberg, K., Nölle, M., 1991. Analysis and solutions of the three point perspective pose estimation problem. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 592–598.
Hartley, R., Zisserman, A., 2004. Multiple View Geometry in Computer Vision, second edition Cambridge University Press.
Hirschmueller, H., 2005. Accurate and efficient stereo processing by semi-global matching and mutual information. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
Hoppe, C., Klopschitz, M., Donoser, M., Bischof, H., 2013. Incremental surface extraction from sparse structure-from-motion point clouds. In: Proceedings of the British Machine Vision Conference (BMVC). BMVA Press.
Hoppe, C., Klopschitz, M., Rumpler, M., Wendel, A., Kluckner, S., Bischof, H., Reitmayr, G., 2012. Online feedback for structure-from-motion image acquisition. In: British Machine Vision Conference (BMVC).
Huber, P.J., 1964. Robust Estimation of a Location Parameter. Ann. Math. Stat. 35 (1), 73–101.
Irschara, A., Rumpler, M., Meixner, P., Pock, T., Bischof, H., 2012. Efficient and globally optimal multi view dense matching for aerial images. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences.
Irschara, A., Zach, C., Bischof, H., 2007. Towards wiki-based dense city modeling. In: IEEE International Conference on Computer Vision (ICCV).
Irschara, A., Zach, C., Frahm, J.-M., Bischof, H., 2009. From structure-from-motion point clouds to fast location recognition. In: CVPR. IEEE Computer Society, pp. 2599–2606.
Kraus, K., 1994. Photogrammetrie, band 1 grundlagen und standardverfahren, fifth edition Ferd. Duemmlers Verlag, Bonn.
Labatut, P., Pons, J.P., Keriven, R., 2007. Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In: IEEE International Conference on Computer Vision (ICCV).
Leberl, F., Irschara, A., Pock, T., Meixner, P., Gruber, M., Scholz, S., Wiechert, A., 2010. Point clouds: lidar versus 3D vision. Photogramm. Eng. Remote Sens. 76 (10), 1123–1134.
Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. (IJCV) 60, 91–110.
Nilosek, D., Salvaggio, C., 2012. Geo-accurate dense point cloud generation. http://dirsapps.cis.rit.edu/3d-workflow/index.html.
Nistér, D., 2003. An efficient solution to the five-point relative pose problem. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 195–202.
Nistér, D., Stewenius, H., 2006. Scalable recognition with a vocabulary tree. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2161–2168.
Pfeifer, N., Glira, P., Briese, C., 2012. Direct georeferencing with on board navigation components of light weight UAV platforms. Int. Arch. Photogrammetry, Remote Sensing Spat. Inf. Sci. XXXIX-B7, 487–492.

Please cite this article as: M. Rumpler et al., Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.04.008

Rehak, M., Mabillard, R., Skaloud, J., 2013. A micro UAV with the capibility of direct Georeferencing. In: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XL-1/W2, pp. 317–323.
Rumpler, M., Daftry, S., Tscharf, A., Prettenthaler, R., Hoppe, C., Mayer, G., Bischof, H., 2014. Automated end-to-end workflow for precise and geo-accurate reconstructions using Fiducial Markers. In: Photogrammetric Computer Vision (PCV), pp. 135–142.
Rumpler, M., Irschara, A., Bischof, H., 2011. Multi-view stereo: redundancy benefits for 3D reconstruction. 35th Workshop of the Austrian Association for Pattern Recognition.
Sivic, J., Zisserman, A., 2003. Video google: a text retrieval approach to object matching in videos. In: IEEE International Conference on Computer Vision (ICCV), pp. 1470–1477.
Snavely, N., Seitz, S.M., Szeliski, R., 2008. Modeling the world from internet photo collections. Int. J. Comput. Vis. 80 (2), 189–210.
Strecha, C., Von Hansen, W., Van Gool, L., Fua, P., Thoennessen, U., 2008. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8.
Triggs, B., McLauchlan, P., Hartley, R., Fitzgibbon, A., 2000. Bundle adjustment - a modern synthesis. In: Vision Algorithms: Theory and Practice, pp. 298–375.
Tscharf, A., Rumpler, M., Fraundorfer, F., Mayer, G., Bischof, H., 2015. On the use of UAVs in mining and archaeology - geo-accurate 3D reconstructions using various platforms and terrestrial Views. ISPRS Ann. Photogrammetry, Remote Sensing Spat. Inf. Sci. II-1/W1, 15–22.
Waechter, M., Moehrle, N., Goesele, M., 2014. Let there be color! large-scale texturing of 3D reconstructions. In: European Conference on Computer Vision (ECCV), pp. 836–850.
Watson, G.A., 2006. Computing helmert transformations. In: Journal of Computational and Applied Mathematics, 197, pp. 387–395.
Zeisl, B., Georgel, P.F., Schweiger, F., Steinbach, E., Navab, N., 2009. Estimation of location uncertainty for scale invariant feature points. In: British Machine Vision Conference (BMVC).
Zhang, Z., 2000. A flexible new technique for camera calibration. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 22 (11), 1330–1334.

Please cite this article as: M. Rumpler et al., Evaluations on multi-scale camera networks for precise and geo-accurate reconstructions from aerial and terrestrial images with user guidance, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.04.008