Dense 3D Reconstruction from Wide Baseline Image Sets

Helmut Mayer[1], Jan Bartelsen[1], Heiko Hirschmüller[2], and Andreas Kuhn[1,2]

[1] Institute of Applied Computer Science, Bundeswehr University Munich
[2] Institute for Robotics and Mechatronics, German Aerospace Center (DLR)

Abstract. This paper describes an approach for Structure from Motion (SfM) for wide baselines image sets and its combination with the dense Semiglobal Matching (SGM) 3D reconstruction approach. Our approach for SfM relies on given information concerning image overlap, but can deal with large baselines and produces highly precise camera parameters and 3D points. At the core of our contribution is robust least squares adjustment with full exploitation of the covariance information from affine point matching to bundle adjustment. Reweighting for robust adjustment is based on covariance information for each individual residual. We use points detected based on Differences of Gaussians including scale and orientation information as well as a variant of the five point algorithm. A strategy similar to the Expectation Maximization (EM) algorithm is employed to extend partial solutions. The key characteristics of the approach is reliability obtained by aiming at a high precision in every step. The capabilities of our approach are demonstrated by presenting results for sets consisting of images from the ground and from small Unmanned Aircraft Systems (UASs).

1 Introduction

Structure from Motion (SfM) from sets of images in combination with dense 3D reconstruction forms a good basis for photo realistic visualization. For example, Leberl et al. [14] show that high quality models can be generated from aerial images, in particular for Microsoft Bing Maps. Leberl et al. term the resulting model extended by semantic information, for instance concerning windows and cars, 'Virtual habitat'. For generating semantic information, terrestrial images and derived 3D models can be used as well, e.g., for buildings and trees [24,11].

Pollefeys et al. [22] presented one of the first approaches dealing with SfM for a larger number of images in a general configuration, i.e., without known approximate pose. It employed uncalibrated images, i.e., images for which the intrinsic camera parameters such as principal distance (focal length) and principal point are not known. This makes the approach very flexible, yet, on the other hand, reliant on sufficient 3D structure in the scene for the determination of intrinsic parameters.

With the five point algorithm [19], it became feasible to directly compute SfM from calibrated images, i.e., for which the intrinsic parameters are known.

F. Dellaert et al. (Eds.): Real-World Scene Analysis 2011, LNCS 7474, pp. 285–304, 2012.

Pollefeys et al. [20] have used it to build a system that was employed for reconstructing 3D structure from more than one hundred thousand images.

Commonly, image overlap is either known implicitly in the form of the order in a sequence, or explicitly, e.g., from an aerial flight plan. Schaffalitzky and Zisserman [25] presented one of the first approaches which automatically determined the image overlap in image sets.

This has led to methods for very large image collections, the so called 'Community Photo Collections' – CPC [6] on the Internet. These techniques mostly use information from the Exif (Exchangeable image file format) tags of the images to derive approximate intrinsic camera parameters and thus conduct calibrated SfM. Agarwal et al. [1] have approached the challenge of CPC with a large cloud of computers. Yet, 'Building Rome on a Cloudless Day' [5] has dealt with millions of images, for which the only thing known to start with is a tag linking them to a place / city such as Rome. It was shown that the images can be organized in terms of visual similarity. This is used for 3D reconstruction of parts with many images. Everything is computed in one day on one standard computer, albeit with several powerful GPUs – Graphical Processing Units.

While the above work is impressive, one has to note that it is based on certain characteristics of the data and a couple of assumptions which make it tractable:

- Many images at tourist attractions are taken from nearly the same spot and thus look alike, i.e., many similar images can be found even for extremely downscaled versions of the images. Frahm et al. [5] use the GIST operator on 4×4 images, i.e., very little information on texture and color is available.
- The goal is to reconstruct the obvious 3D structure, leading to impressive 3D reconstructions of highlights, such as the Colosseum in Rome. Yet, there might be images, possibly with wider baselines, that could be used to extend the geometrical coverage or even to link the tourist attractions. This is not considered, as it would mean a detailed comparison of many more images.

A preliminary version of our work, comprising also absolute pose estimation, has been published earlier [2]. It focuses on image sets with possibly very large baselines. For the registration of these images, we have to either supply the sequence of images, or sets of overlapping triplets.

The basis of our work (Figure 1) are points with scale and rotation detected based on Difference of Gaussians (Section 2). We start by removing unlikely matches by cross correlation with a very low threshold. Matches are refined by least squares matching [7] using an affine geometric model. This results in subpixel accurate point positions including covariance information.

The points and their covariance information are employed for SfM from pairs and triplets (Section 3). It is based on a variant of the five point algorithm embedded into RANdom SAmple Consensus – RANSAC [4] using the Geometric Robust Information Criterion – GRIC [27]. A strategy similar to the Expectation Maximization (EM) algorithm is used to extend partial solutions. We employ robust bundle adjustment (Section 4), where we reweight based on residuals (distance between reprojected 3D point and measured 2D point) and, particularly, covariance information for each individual residual.

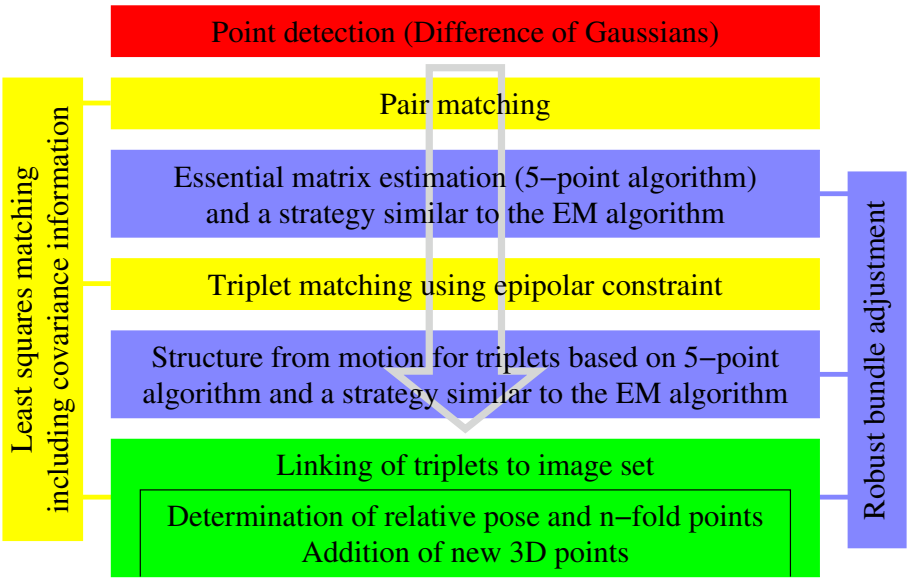

Fig. 1. Structure from Motion based on least squares matching and robust bundle adjustment

Triplets are linked either sequentially or hierarchically to image sets (Section 5). This results in highly precise poses, improved intrinsic parameters, and 3D points including covariance information.

Section 6 presents results for terrestrial images and images acquired from small Unmanned Aircraft Systems (UASs) with a size of less than one meter and a weight of approximately one kilogram. We demonstrate the precision obtained by our approach by means of a loop closing experiment. Wide baseline matching capabilities are shown with results for a combination of terrestrial images and images from a UAS.

The poses and intrinsic parameters are input for Semiglobal Matching – SGM [9] (Section 7) which leads to dense 3D point sets and detailed 2.5D, or 3D surfaces. Finally, results for dense matching with SGM are given. Section 8 concludes the paper with an outlook.

2 Point Detection and Matching

The basis for our approach are points based on Differences of Gaussians (DOG) as proposed by Lowe [15] and implemented in SiftGPU [29]. As we want to deal with situations with very low contrast, such as weak structures on facades, we employ a very low threshold.

We start with image pairs. The point centers as well as their estimated scale and rotation are employed to cut out image patches of size 13×13 pixels from the

images. These patches are correlated by means of (normalized) cross correlation. For all best matches for points in the master image, which exceed a low threshold of 0.5, we compute a histogram of the rotation differences. The histogram is smoothed and its mode determined. As the mode of the histogram was found to rather reliably describe the in-image-plane rotation between image pairs, we use it for normalization: We cut out unrotated patches (though with individual scale) in one image and rotate all patches in the second image according to the difference of rotation as given by the mode of the histogram.

Cross correlation between patches is computed again and the same low threshold of 0.5 is used. Yet, this time the best matches for all points in the master image exceeding the threshold are subject to least squares matching [7]: The sum of the squared intensity differences between patches around the points is minimized by varying the parameters for a geometric and a radiometric transformation between the patches.

We use an affine geometric model with six parameters $(a_0^i, \ldots, b_2^i)$ describing the translation in x- and y-direction as well as two rotations and two scales. Given a square patch in master image 0, this leads to a parallelogram in the matching images (Figure 2). While the general model for a linear mapping between image patches is a homography, we found that the eight parameters of a homography usually cannot be reliably determined for small patches. Small patches are a must, though, because the region around a point in the scene does not have to be planar and the farther one goes from a point, the higher becomes the likelihood for discontinuities and occlusion.

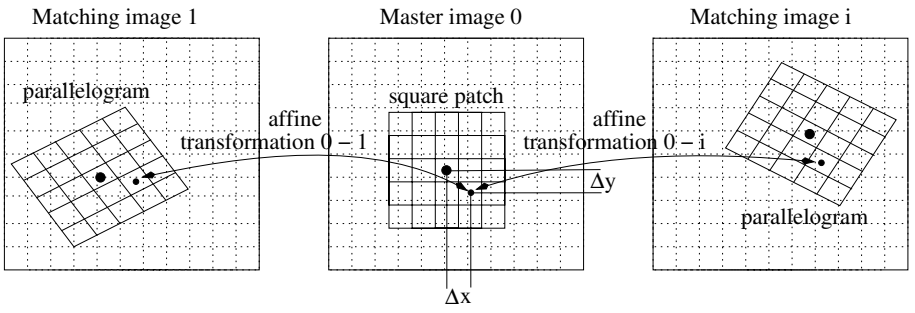

Fig. 2. Least squares matching is based on an affine geometric model. Individual pixels (small dots) of image patches around subpixel precise points (large dots) are transformed based on the affine model. Given a square patch in the master image this leads to parallelograms in the matching images.

The pixel raster of the patch in the master image is defined by Δx and Δy as well as the indices j and k ($-N \leq j \leq N$ and $-N \leq k \leq N$ with $N = 6$). Δx and Δy depend on the scale known from point detection. The coordinates of the pixels in the master image 0 and the matching image i are described by

$$x^0_{jk} = x^0 + j\Delta x$$
$$x^0_{jk} = y^0 + k\Delta y$$
$$x^i_{jk} = x^i + a^i_0 + a^i_1 j\Delta x + a^i_2 k\Delta y \quad (1)$$
$$y^i_{jk} = y^i + b^i_0 + b^i_1 j\Delta x + b^i_2 k\Delta y \ , \quad (2)$$

with x^0, y^0 and x^i, y^i denoting the centers of the patches in master image 0 and matching image i, respectively. We use subpixel coordinates also for x^0 and y^0 to optimally center the patch around the point.

For the subpixel precise point positions, the intensity of the pixels has to be determined by (in our case bilinear) interpolation. Additionally to the six parameters $a^i_0, \ldots, b^i_2$ for the geometry we use bias r^i_0 and contrast r^i_1 for the intensity to radiometrically adapt the patch in matching image i. This leads to the following residuals v^i_{jk} for least squares adjustment ($I^0()$ and $I^i()$ denote the intensity function in master image 0 and matching image i, respectively):

$$v^i_{jk} = I^0(x^0_{jk}, y^0_{jk}) - [r^i_0 + r^i_1 I^i(x^i_{jk}, y^i_{jk})] \quad (3)$$

The goal of least squares matching is to estimate affine parameters $a^i_0, \ldots, b^i_2$ and radiometric parameters r^i_0, r^i_1 minimizing the sum of all squared residuals

$$\sum_{j=-N}^{N} \sum_{k=-N}^{N} [v^i_{jk}]^2 \ . \quad (4)$$

Equation (4) is linear with respect to the radiometric parameters r^i_0 and r^i_1. It is nonlinear in terms of the geometric parameters, because $I^i()$ is nonlinear in general. As there is no closed-form solution, first order Taylor expansion is employed to linearize Equation (4) based on initial values for the parameters. We assume no translation ($a^i_0 = b^i_0 = 0$), a similar intensity ($r^i_0 = 0$ and $r^i_1 = 1$) and take the known scale difference and rotation into account for a^i_1, a^i_2, b^i_1 and b^i_2. Setting the derivative to zero, one obtains a linear system

$$\mathbf{A}\beta = \mathbf{y} \ . \quad (5)$$

Matrix $\mathbf{A}$ consists of the Jacobian of the intensity function in the matching image i with respect to the unknown geometric and radiometric parameters concatenated in vector β. Vector $\mathbf{y}$ comprises the negative measurement errors.

While the linear system (5) can be solved directly, we employ the normal equations

$$\mathbf{N}\beta = (\mathbf{A}^\mathsf{T}\mathbf{A})\beta = \mathbf{A}^\mathsf{T}\mathbf{y} \quad (6)$$

and compute $\beta = \mathbf{N}^{-1}\mathbf{A}^\mathsf{T}\mathbf{y}$. By this means we obtain $\mathbf{C} = \mathbf{N}^{-1}$, i.e., the relative covariance matrix for the unknown parameters. Because the problem is nonlinear, the solution is obtained iteratively. For optimization we use the Levenberg Marquardt algorithm.

The criteria for a valid match obtained by least squares matching are that the cross correlation value is larger than 0.8 as well as that the estimated variance

for the shift is below 0.1 pixels. For the latter, one has to consider that from our experience the estimated variance is always highly optimistic. Cross correlation is known to be not a good descriptor for stronger geometrical distortions. Though, it was found to be very useful if the geometrical distortions are small [18], which is the case after least squares matching.

For more than two images, we link least squares matching for pairs. The image in which the patch is closest to the image center is used as master, as this improves the chance for a frontal view. The patch in the master image is geometrically kept as square and the affine transformations relative to the other images are estimated (Figure 2).

The solutions are linked by substituting $I^0()$ in equation (3) by the average intensity in all images. To account for different average intensities and contrasts of the patches, we take the estimated radiometric parameters r_0^i and r_1^i for each patch into account when computing the average. As the problem is nonlinear and solved iteratively, the average intensity changes due to different geometric transformations as well as different radiometric parameters for each iteration.

Output for the accepted matches are the improved coordinates $x^i + a_0^i$ and $y^i + b_0^i$ as well as their relative covariance information. The latter can improve SfM estimation particularly for stronger in-image-plane rotations [17].

While least squares matching entails more effort than just using the point centers of the SIFT points, we found that the relative coordinates obtained are more precise. This is probably due to the fact, that we look for optimum matches. This reduces the influence of geometrical deformations, partial occlusions, and noise, which influence point centers when they are estimated independently.

3 Two and Three View Geometry

In the remainder of this paper we assume that we have at least an approximate knowledge of the intrinsic parameters. We also implemented an uncalibrated approach in the spirit of Pollefeys et al. [21]. Yet, we found it to be only reliable if sufficient 3D structure is present. Only then, the intrinsic parameters can be reliably determined.

Triplets are the basic geometric building block of our approach due to the following reasons:

- Opposed to pairs where points can only be checked in one dimension by means of their distance from their respective epipolar lines, triplets allow for an unambiguous geometric check of points. This does not only lead to much more reliable points, but also to improved, more reliable information for the cameras.
- Triplets can be directly linked into larger sets by determining their relative pose (translation, rotation, and scale) based on two common images.

Because the combinatorics is worse for triplets than for pairs, we start with pairs and determine essential matrices and thus epipolar lines for them. For known intrinsic parameters, the relative pose of the image pair is determined directly, i.e., with no need for approximate values, by means of the five point algorithm [19].

As usually only a possibly small part of the matched point pairs is actually correct, we employ RANSAC in conjunction with GRIC [27]. The later means, that instead of counting the number of inliers, we attribute a constant penalty to outliers and values proportional to their squared residuals v^2 to inliers. A threshold is used to define where the transition from inliers to outliers occurs. While in RANSAC the number of inliers is maximized, GRIC aims at a minimum corresponding to many points with small residuals. By means of GRIC one can distinguish between solutions with a low precision, but more points, and highly precise solutions, with possibly less points, but smaller residuals, which are more likely correct.

The above combination of RANSAC and GRIC works well for more or less well behaved scenes. Yet, we found that for complex scenes, e.g., involving many very similar points, the above combination is not sufficient to tell good from bad solutions. This happens, e.g., for window corners on facades of buildings, possibly in conjuction with camera movements which conspire with repetitions on the facade. Inspired by Chum et al. [3], we compute a maximum likelihood, i.e., robust bundle adjustment, solution for the best of every couple of hundred RANSAC iterations. Eventually, the bundle solution which leads to the lowest GRIC value is taken as the final result.

But even this gives only a partial solution to the problem. While RANSAC produces a solution from only inliers with a certain probability, it is not guaranteed, that this solution is accurate. Even worse, inaccurate solutions can also be not representative for all, or even the majority of the inliers. E.g., consider a larger image and RANSAC selecting in one sample only inliers from the center of the image. While the geometric solution (of the five point algorithm) will be correct, it will not be precise enough to find also the correct matches closer to the margin of the image. A way to counteract this is to force RANSAC only to use points with a certain minimum distance. Though, this is problematic, because in certain cases there might be just correct matches in the center of the image.

We have devised a strategy similar to the EM algorithm (Figure 3) which employs robust bundle adjustment (cf. next Section) to mitigate the above problem. We robustly bundle adjust the initial direct solution using the inliers determined by RANSAC. The obtained, geometrically improved, solution is employed to compute new inliers based on GRIC. This is iterated until either a predefined number of iterations (here 5) is reached, or no significant improvement in terms of GRIC is obtained.

The above procedure is used for pairs and triplets. For the latter, we employ the result for image pairs to restrict the search space via epipolar lines derived from the essential matrices. This strongly reduces the number of hypotheses for image triplets.

For the geometric computation of triplets, we use one image as master and compute translation and rotation towards the other two images via the five point algorithm for five conjugate points in the three images. This fixes all but one parameter, namely the relative base length between the two pairs. At the moment we assume that we only work with images with a significant base between them.

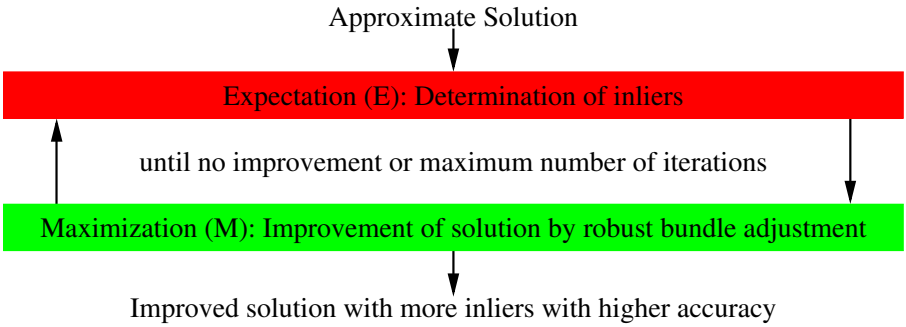

Fig. 3. Strategy similar to the Expectation Maximization (EM) algorithm

While this is a limitation of our approach, we note that there is only a problem with the infinite homography, not with homographies for real planes. Particularly, we triangulate the five points in both pairs and compute the distance from the master image. The ratio of the distances in both pairs is proportional to the ratio of the base lengths. To make the computation robust, we employ the median value of the five ratios computed for the five conjugate points.

4 Robust Bundle Adjustment

While bundle adjustment [28] has not been seen as crucial for early approaches on multi view geometry, since a couple of years it is acknowledged that it is useful and even necessary for large image sets.

This is demonstrated by recent work on generalized preconditioners [13]. They allow for an efficient use of conjugate gradient based solutions for bundle adjustment for very large systems also for the general configurations encountered when collecting data from the ground or in CPC.

Our work goes into another direction, namely robustifying bundle adjustment by means of reweighting. I.e., least squares are generalized in the form of an M-estimator [12]. The particular contribution is, that we compute an estimate for the variance of each individual residual and use this for reweighting when implementing the M-estimator.

The estimation of individual variances for the residuals is costly in terms of computation per iteration. Yet, we found that at least for systems with a limited number of images, i.e., tens of images, it is actually faster in the aggregated run time, because much fewer iterations are needed. What is more, one usually obtains a more precise solution consisting of more points.

Following Jian et al. [13], we define $\mathbf{P} = \{P_i; i = 1, \ldots, M\}$ as the camera parameters, $\mathbf{X} = \{X_j; j = 1, \ldots, N\}$ as the 3D points, and $\mathbf{x} = \{x_k; k = 1, \ldots, K\}$ as the measurement of 3D point X_j in camera P_i). Function $f_k(P_i, X_j)$ projects a 3D point to an image. By

$$v_k = f_k(P_i, X_j) - x_k$$

we define the residual between the projected 3D point and the measured image point. The goal of bundle adjustment is to reduce the sum of the squared residuals

$$\sum_{k=1}^{K} [v_k]^2 \ . \tag{7}$$

Equation (7) is nonlinear. It can be linearized by means of first order Taylor expansion, assuming that appropriate initial estimates for the camera parameters P_i and the 3D points X_j are available:

$$\sum_{k=1}^{K} [f_k(P_i, X_j) + \frac{\partial f_k(P_i, X_j)}{\partial P_i} dP_i + \frac{\partial f_k(P_i, X_j)}{\partial X_j} dX_j - x_k]^2 \ . \tag{8}$$

As above for least squares matching, a linear solution (5) can be obtained by setting the derivatives in (8) to zero. The system consists of a sparse matrix $\mathbf{A}$ made up of the Jacobian of the measurements with respect to cameras and 3D points, the vector β concatenating the parameters of cameras and 3D points, and finally, the vector $\mathbf{y}$ consisting of the negative measurement errors.

While (5) can be solved directly, we solve the normal equations (6). By this means we can introduce the estimated accuracy of the measured image points as derived by least squares matching in Section 2 in the form of a weight matrix. Particularly, we employ as weight the inverse of the relative covariance matrix of the measurements $\mathbf{C}$, leading to

$$\mathbf{N}\beta = (\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A})\beta = \mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{y} \ . \tag{9}$$

For optimizing the solution, we again use the Levenberg Marquardt algorithm. Please note, that $\mathbf{C}$ is a positive definite block diagonal matrix consisting of 2×2 blocks describing the variance of the measured points in x- and y-direction as well as their x-y covariance.

In the M-estimator, we reweight $\mathbf{C}$ by

$$w = \sqrt{2 + \overline{v}^2} \ ,$$

with $\overline{v} = v/\sigma_v$. I.e., the residual is divided by its standard deviation. While usually a common variance is used, we compute an estimate of the covariance of the individual residuals C_v as follows:

$$\mathbf{C}_v = \mathbf{C} - \mathbf{A}(\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T} = \mathbf{C} - \mathbf{A}\mathbf{N}^{-1}\mathbf{A}^\mathsf{T} \tag{10}$$

For an efficient solution, we employ the Schur complement and split up the design matrix in a part for 3D points $\mathbf{A}_X$ and a part for the cameras $\mathbf{A}_C$. This results in the following (symmetric) matrix $\mathbf{N}$ and its inverse $\mathbf{M}$

$$\mathbf{N} = \begin{bmatrix} \mathbf{N}_{XX} & \mathbf{N}_{XC} \\ \mathbf{N}_{XC}^\mathsf{T} & \mathbf{N}_{CC} \end{bmatrix} \quad \text{and} \quad \mathbf{M} = \mathbf{N}^{-1} = \begin{bmatrix} \mathbf{M}_{XX} & \mathbf{M}_{XC} \\ \mathbf{M}_{XC}^\mathsf{T} & \mathbf{M}_{CC} \end{bmatrix} \ .$$

We solve for $\mathbf{M}_{CC} = (N_{CC} - \mathbf{N}_{XC}^{\mathsf{T}}\mathbf{N}_{XX}\mathbf{N}_{XC})^{-1}$, i.e., the inverse for the cameras, at the core of the bundle adjustment. The computation of $\mathbf{M}_{XX} = \mathbf{N}_{XX}^{-1} + \mathbf{N}_{XX}^{-1}\mathbf{N}_{XC}\mathbf{M}_{CC}\mathbf{N}_{XC}^{\mathsf{T}}\mathbf{N}_{XX}^{-1}$ can be done very efficiently, as it only involves the inversion of 3×3 matrices in the block diagonal matrix $\mathbf{N}_{XX}$ and multiplications with 3×6 and 6×6 matrices. The covariance between points and cameras $\mathbf{M}_{XC}$ is for most applications not needed and, thus, not calculated. From $\mathbf{N}\cdot\mathbf{M} = \mathbf{I}$ (with $\mathbf{I}$ the unit matrix) one can derive

$$\mathbf{M}_{XC} = \mathbf{N}_{XX}^{-1}\mathbf{N}_{XC}\mathbf{M}_{CC} \ ,$$

giving the full matrix $\mathbf{M} = \mathbf{N}^{-1}$ needed to solve Equation (10).

As the measurements are 2D image coordinates, the covariance information for residuals corresponds to 2D ellipses. Thus, $\overline{v} = v/\sigma_v$ is computed as ratio of the length of the residual vector and the standard deviation of the residual in the direction of the residual σ_v as shown in Figure 4. $\overline{v}$ is employed to reweight the 2×2 block in matrix $\mathbf{C}$ corresponding to the residual.

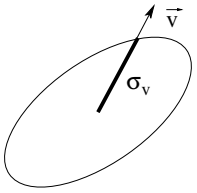

Fig. 4. Error ellipse for residual, direction of the residual and the standard deviation in the direction of the residual σ_v

5 Structure from Motion for Image Sets

We link image sets based on camera information for two common images. We start by linking triplets, but depending on the strategy (cf. below), also sets are linked to sets.

For obtaining approximate values, we first relate the camera information for an image in one set, i.e., the master set, to the camera information for the same image in the other set, i.e., the slave set. As we assume that we know the intrinsic parameters, we can translate and rotate the slave into the master set. The remaining unknown is scale. It is derived from the camera parameters for a second common image, for which in both images the distance to the first common camera is computed. The ratio of the distances gives the ratio in scale of the two coordinate systems. With the obtained approximate values for translation, rotation, and scale, we transform all camera parameters and 3D points from the slave into the master set.

Additionally, we transform also points from the master into the slave set, to obtain more than threefold, i.e., n-fold points[1]. The higher n, the more geometrically stable the solution becomes. For computing n-fold points, we first

[1] The terms twofold, threefold, and n-fold point are used for expressing that the projection of a 3D scene point is detected in two, three, or n images.

note, that it is not useful to compare points in 3D space, because its metric is in general not well defined. Thus, we conduct the comparison in image space. Particularly, we employ trifocal tensors computed from the camera matrices [8] of the slave set and project points from the two common images of the master set into the third, etc., image of the slave set. There, multi-image least squares matching (Section 2) is conducted leading to n-fold points. Finally, we compute a robust bundle adjustment (Section 4) based on the approximate values for translation, rotation, and scale, as well as the n-fold points.

This gives an improved solution for the overlapping part of the combined set. Yet, novel points in the slave set are still missing. Therefore, we compare for the two common images the image coordinates from the slave set with the image coordinates in the master set. Only when there is no nearby point found in image space as implemented by dilation with a radius of two pixels, the corresponding 3D point is introduced. Eventually, again a robust bundle adjustment is computed, this time also including the estimation of improved intrinsic parameters.

We note that the above procedure tracks a point only as long it is visible. While this means that points which are occluded in a frame are lost and possibly re-introduced, we found that this is superior to projecting 3D points into the images. The problem with the latter is, that if one goes around an object, repeating structures, possibly even on the backside, can by chance be at the same location and match very well. As these points are wrong, they can introduce a serious bias in the estimation.

For linking sets, we have implemented a

- sequential strategy and a
- hierarchical strategy.

The sequential strategy is very simple: We just link one triplet after the other to the set with an overlap of always two images. The basic problem with this simple strategy is, that at least for wide baseline sets we found it is necessary that we conduct a robust bundle adjustment each time we add a triplet. This makes the strategy computationally very intensive.

On the other hand, in the hierarchical strategy, sub-sets are grown in parallel and linked one by one (Table 1). As we need two common images, this means that we can extend the set by $2i-2$ images. Starting with 3, we obtain sets with 4, 6, 10, 18, 34, etc. images. This is obviously much more efficient as it entails much fewer robust bundle solutions.

It is less obvious, though, that the hierarchical strategy is also very useful in terms of robust bundle adjustment, particularly for large sets. For robust reweighting (Section 4), it is important, that the variances of the residuals are comparable. If this is not the case, e.g., when linking a large set with multiple overlap and high internal precision with a small set and thus with low precision, there is a strong tendency, that a considerable number of the weaker, but correct points of the smaller set will be thrown out. All this is avoided by hierarchical linking, where sets of approximately the same size and, thus, precision are linked.

Table 1. Hierarchical linking eight image triplets for ten images

1 2 3	2 3 4	3 4 5	4 5 6	5 6 7	6 7 8	7 8 9	8 9 10
1 [**2 3**] 4		3 [**4 5**] 6		5 [**6 7**] 8		7 [**8 9**] 10	
1 2 [**3 4**] 5 6				5 6 [**7 8**] 9 10			
1 2 3 4 [**5 6**] 7 8 9 10							

This was demonstrated by Mayer [16] for a loop of ninety images taken inside the Zwinger, Dresden, Germany. Hierarchical linking has been seven times faster. More important, it produced not only 32,783 compared to 28,582 points for sequential linking, but also many more many-fold points.

6 Results of Structure from Motion

All experiments reported in this section have been conducted using the sequential strategy and the same parameters.

The sequence castle-R20 of Ettlingen castle in Germany consists of twenty images [26]. Some of them are shown at the top of Figure 5. Our SfM approach results in an estimated average back-projection error σ_0 of 0.14 pixels. For demonstrating the high precision of our results, we conducted a loop closing experiment, i.e., we took the last image of the twenty images sequence to be the same as the first image. SfM was conducted without closing the sequence. This means that the differences between the camera parameters for the first and the last image, which should be the same, give an indication of the precision obtained.

Firstly, we note that Figure 5 visually shows, that the differences are small. Table 2 gives a quantitative evaluation. The upper part shows the translation error. It is in the range of 0.1 % of the maximum distance between the camera centers. In terms of an absolute distance this means about 4 centimeters. The absolute angular error after twenty images is only 0.14°. This means that we obtained a relative angular error per image of 0.007°, demonstrating the high precision achieved.

The top of Figure 6 shows three pairs of near infrared images of size 1392×1040 pixel of a sequence of 400 images taken by a mobile mapping system. The pairs have a small overlap due to a diverging imaging configuration and the images a limited quality due to the near infrared. In spite of this and even though the images were not explicitly treated as pairs in SfM, but as sequence, the local geometry of the pairs could be estimated very well. This is mainly due to robustly tracking points over many frames resulting in highly precise many-fold points and camera poses.

The third example is based on images acquired for a village in southern Germany by a small UAS. In one experiment, a building has been captured by terrestrial images which have been linked via ascending images (center of Figure 7 bottom) to a flight line above the village. In spite of the partially strong wide baseline geometry (Figure 7, center row), we could still compute valid and precise camera poses and 3D points.

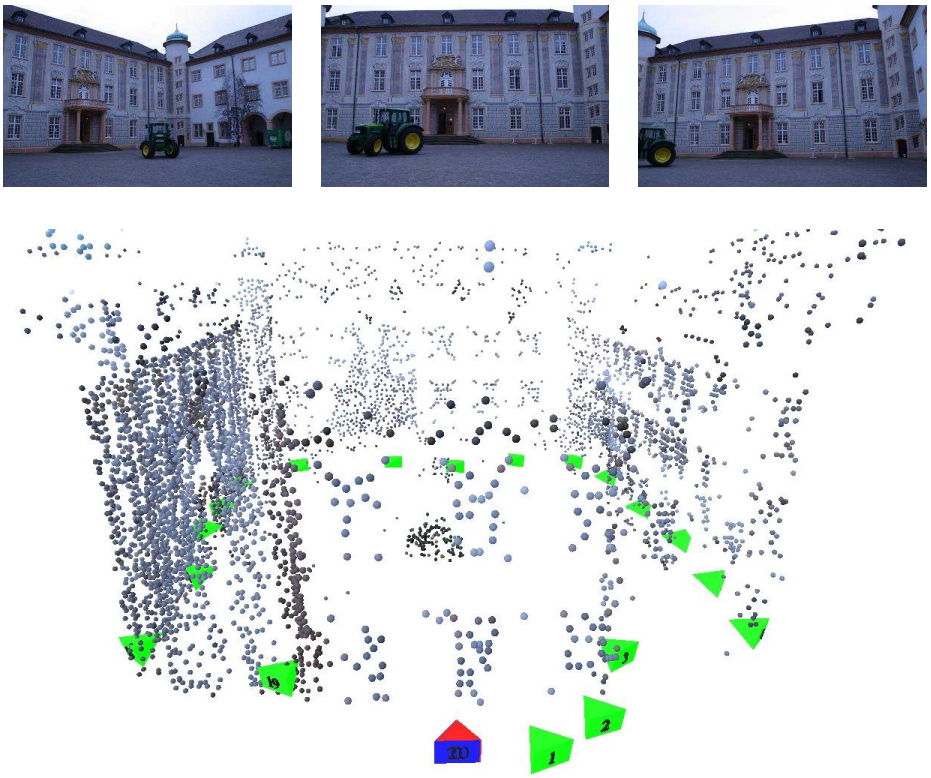

Fig. 5. Top: Images four, seven, and eight of image sequence castle-R20 of Ettlingen castle in Germany, with twenty images [26]. Bottom: Result for SfM ($\sigma_0 = 0.14$ pixels). Cameras are given as pyramids and points are colored from the images. For the loop closing experiment, the first and the last image of the sequence were taken to be the same, depicted in red and blue with numbers 0 and 20. The overlap of the latter demonstrates the high quality of the reconstruction.

7 Dense Reconstruction

For dense reconstruction we employ Semiglobal Matching – SGM [9]. It is based on

- mutual information (MI) or the Census filter for cost computation and
- the substitution of a 2D smoothness term by a combination of 1D constraints (semiglobal).

The mutual information mi_{I_1,I_2} is the sum of the entropies in the two images to be matched $h_{I_1}(i)$ and $h_{I_2}(k)$ minus their joint entropy $h_{I_1,I_2}(i,k)$

$$mi_{I_1,I_2} = h_{I_1}(i) + h_{I_2}(k) - h_{I_1,I_2}(i,k) \quad . \tag{11}$$

Table 2. Evaluation for castle-R20 in terms of loop closing error – Top: Translation in terms of maximum distance of projection centers as well as in absolute distance; Bottom: Absolute angular error (after twenty images) and relative angular error (per image)

Translation	x	y	z
% of maximum distance	0.124	-0.011	-0.053
absolute distance [m]	0.041	-0.004	-0.017

Absolute angular error	0.1398°
Relative angular error per image	0.0070°

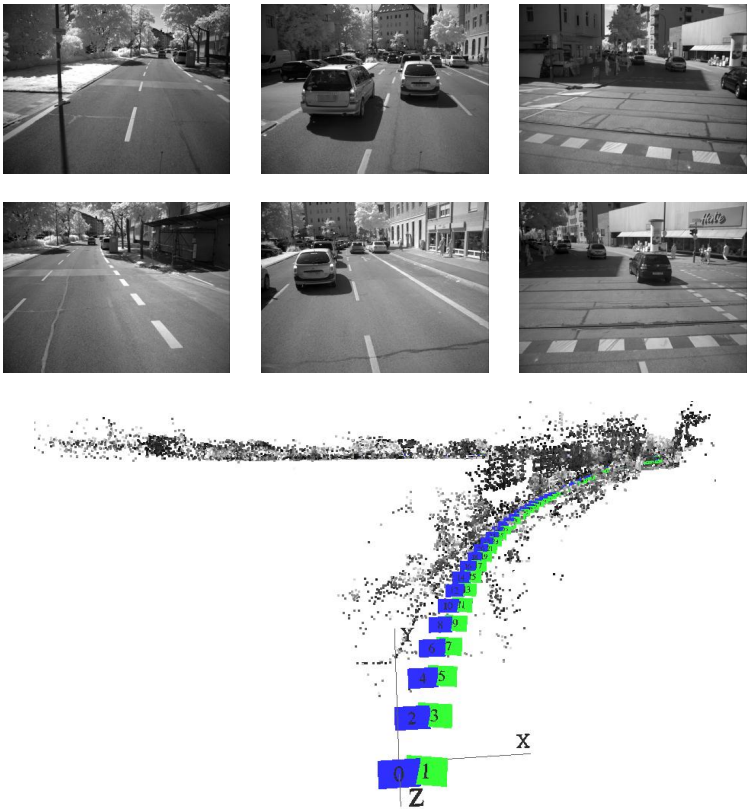

Fig. 6. Top: Image pairs of an infrared sequence of 400 images taken by a mobile mapping system given in the form top / bottom from left to right: Pairs 4 / 5, 118 / 119, and 180 / 181. Bottom: Result of SfM. Points are given with the color taken from the images and camera positions and orientations are marked by colored pyramids.

This leads to the following matching cost (f_D transforms the matching image I_m with an initial disparity image D)

$$C_{MI}(\mathbf{p}, d) = -mi_{I_b, f_D(I_m)}(I_{b\mathbf{p}}, I_{m\mathbf{q}}) \quad , \tag{12}$$

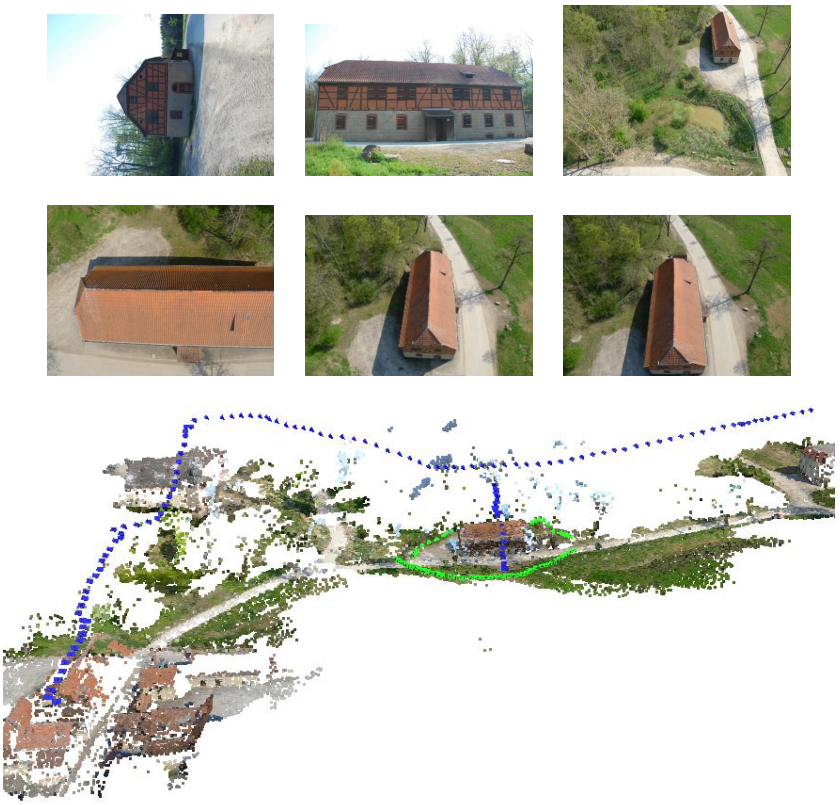

Fig. 7. Top: Images of a German village taken from the ground and from an ascending UAS. Please note the wide baselines between the left and the other two images of the triplet shown on the second row. Bottom: Result for SfM estimation. Cameras are given as pyramids and points are colored from the images. For the building in the center terrestrial images have been linked to the flight line above via ascending images.

where $\mathbf{q}$ is the pixel in the matching image I_m corresponding to the pixel $\mathbf{p}$ in the reference image I_b and the disparity d.

In essence, MI gives the conditional probability distribution for the intensities in the matching image given an intensity in the reference image without resorting to a parametric model. Thus, MI can compensate a large class of global radiometric differences. Though, one has to note that the conditional probability is computed for the whole image which can be a problem for local radiometric changes, e.g., if materials with very different reflection characteristics exist in the scene or lighting conditions change.

The Census filter was found by Hirschmüller and Scharstein [10] to be the most robust variant for matching cost computation. It defines a bit string with each bit corresponding to a pixel in the local neighborhood of a given pixel. A bit is set if the intensity is lower than that of the given pixel. Census thus encodes

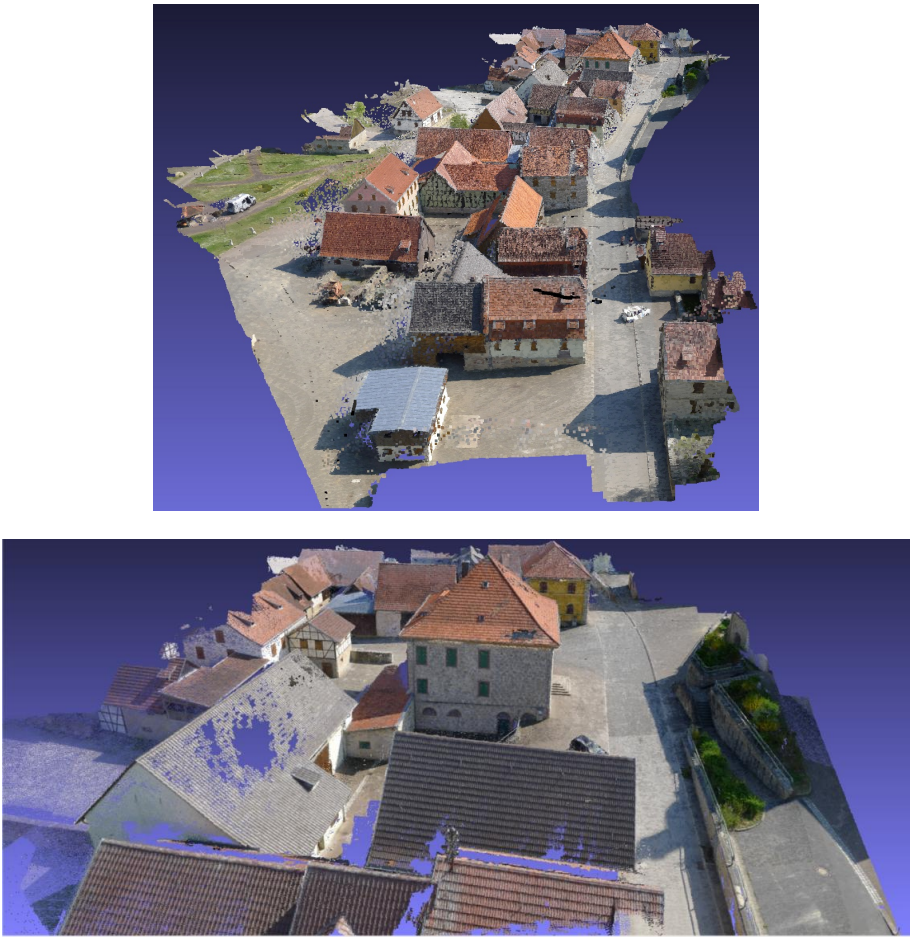

Fig. 8. Top: Dense 3D points generated by SGM. Bottom: Part

the spatial neighborhood structure. A 7×9 neighborhood can be encoded in a 64 bit integer. Matching is conducted via computing the Hamming distance between corresponding bit strings.

The smoothness term punishes changes of neighboring disparities (operator $T[]$ is 1 if its argument is true and 0 otherwise):

$$E(D) = \sum_{\mathbf{p}} \left(C(\mathbf{p}, D_{\mathbf{p}}) + \sum_{\mathbf{q} \in N\mathbf{p}} P_1 T[|D_{\mathbf{p}} - D_{\mathbf{q}}| = 1] + \sum_{\mathbf{q} \in N\mathbf{p}} P_2 T[|D_{\mathbf{p}} - D_{\mathbf{q}}| > 1] \right) \tag{13}$$

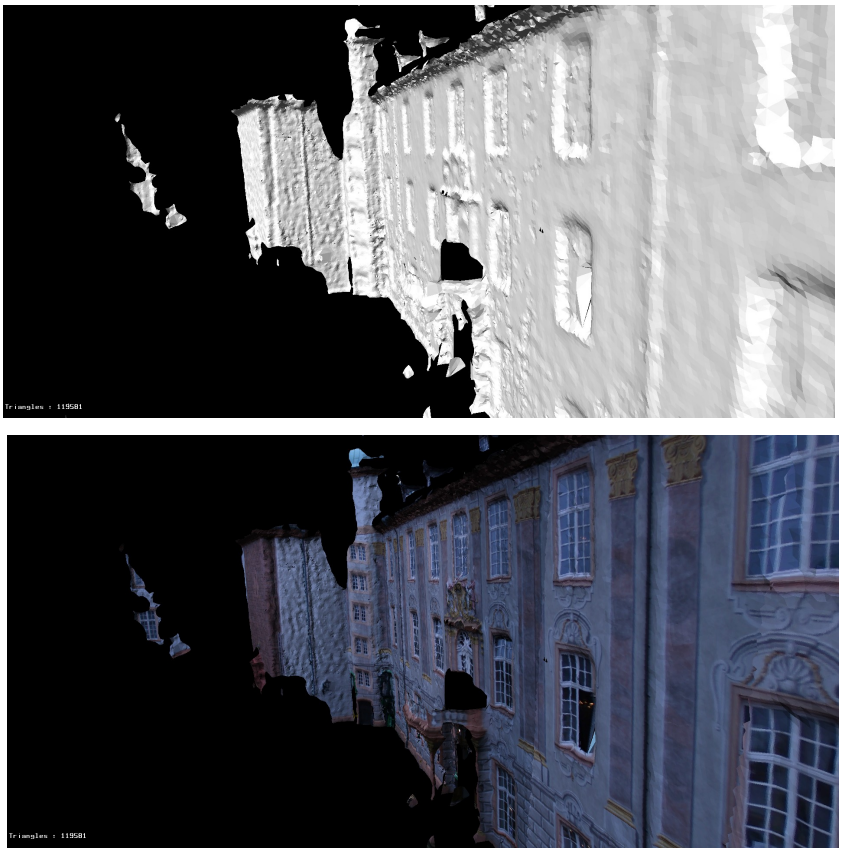

Fig. 9. Result for dense 3D surface mesh reconstruction using SGM of parts of image sequence castle-R20 (Figure 5) – shaded (top) and textured (bottom)

- The first term consists of pixel matching costs for all disparities of D.
- The second term adds a constant penalty P_1 for all pixels $\mathbf{q}$ from the neighborhood $N_\mathbf{p}$ of $\mathbf{p}$, for which the disparity changes only slightly (1 pixel).
- The third term adds a larger constant penalty P_2 for bigger changes of the disparities. Because it is independent of the size of the disparities, it preserves discontinuities.
- As discontinuities in disparity are often visible as intensity changes, P_2 is calculated depending on the intensity gradient in the reference image (with $P_2 \geq P_1$).

In 2D, global minimization is NP hard for many discontinuity preserving energies $E(D)$. In 1D, minimization can be done in polynomial time via dynamical programing, which is usually applied within image lines. Unfortunately, because the solutions for neighboring lines are computed independently, this can lead to streaking.

For the semiglobal solution, 1D matching costs are computed in different, (practically 8) directions which are aggregated without weighting. In the reference image, straight lines are employed, which are deformed in the matching image.

By computing D for exchanged reference and matching image one can infer occlusions or matching errors by means of a consistency check. If more than one pair with the same reference image is matched, the consistency check is conducted for all pairs only once.

With the above methodology, dense disparities can be computed. By using the camera parameters all points can be projected into 3D leading to dense 3D point clouds. While the original work of Hirschmüller [9] has shown how to derive 2.5D surface models, work on the derivation of a 3D surface by means of triangulation of the 3D points dealing also with outliers has been started only recently.

For parts of the village for which camera poses and 3D point clouds have been estimated (Section 6, Figure 7), SGM was used to compute dense 3D points from several pairs. Figure 8 gives an impression of the very high point density and quality obtained.

Finally, Figure 9 shows first results for dense 3D surface mesh reconstruction using SGM. Particularly the shaded visualization shows, that the indentations of the windows could be determined reliably.

8 Conclusions and Outlook

In this paper we have presented an approach for dense reconstruction from wide baseline image sets. As key characteristics it aims at a high precision in every step of the approach from least squares matching to robust bundle adjustment. Particularly for the latter, we take into account the estimated covariance for the residuals, leading to more precise solutions with more points in less time.

Even though we have demonstrated that we can compute SfM for larger scenes consisting of hundreds of images with wide baselines, there are still a couple of shortcomings. The most basic is, that we rely on given information concerning image overlap. While Agarwal et al. [1] and Frahm et al. [5] have shown how the problem can be solved in principle, it is still not clear how to deal with wide baselines. The most obvious way is to compare all possible pairs, but for larger sets this seems to be not feasible even using GPUs.

Yet, also for large scenes with small baselines problems exist. One is in the line of thought of our hierarchical approach for linking image sets (Section 5). Particularly, the question is, which parts of the unordered sets should be linked when, i.e., at which level of the hierarchy.

Then, there are problems with objects of the real world with specific characteristics. E.g., some objects have symmetries, such as that front and back look very similar. This is hard for current approaches for unordered sets, where missing matches are usually attributed to unmodeled occlusions. Thus, the questions arises, how much semantic information is needed for a reliable 3D reconstruction? Should ordering information from the camera, e.g., in terms of known acquisition time be used? If location information, e.g., from GPS is available and reliable, it could be used to circumvent the problem.

Finally, there are also a couple of smaller or larger details in our approach which could be solved in a better way. E.g., at the moment we use one standard value for RANSAC / GRIC for pairs and triplets. While this works in nearly all cases, it can be far from optimal as it does not account for the different precisions possible for images of different sizes, distortions, lighting, contrast and scene characteristics (e.g., facade planes versus trees). Here, estimation by means of RECON [23] could give a more general solution.

Acknowledgment. We thank the reviewers for their comments, which have helped to make important parts much more explicit.

Parts of the presented work were supported by Bundeswehr Geoinformation Office which is gratefully acknowledged.

References

1. Agarwal, S., Snavely, N., Simon, I., Seitz, S., Szeliski, R.: Building Rome in a Day. In: Twelfth International Conference on Computer Vision, pp. 72–79 (2009)
2. Bartelsen, J., Mayer, H.: Orientation of Image Sequences Acquired from UAVs and with GPS Cameras. Surveying and Land Information Science 70(3), 151–159 (2010)
3. Chum, O., Matas, J., Kittler, J.: Locally Optimized RANSAC. In: Michaelis, B., Krell, G. (eds.) DAGM 2003. LNCS, vol. 2781, pp. 236–243. Springer, Heidelberg (2003)
4. Fischler, M., Bolles, R.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Communications of the ACM 24(6), 381–395 (1981)
5. Frahm, J.-M., Fite-Georgel, P., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y.-H., Dunn, E., Clipp, B., Lazebnik, S., Pollefeys, M.: Building Rome on a Cloudless Day. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 368–381. Springer, Heidelberg (2010)
6. Goesele, M., Ackermann, J., Fuhrmann, S., Klowsky, R., Langguth, F., Muecke, P., Ritz, M.: Scene Reconstruction from Community Photo Collections. IEEE Computer 43(6), 48–53 (2010)
7. Grün, A.: Adaptive Least Squares Correlation: A Powerful Image Matching Technique. South African Journal of Photogrammetry, Remote Sensing and Cartography 14(3), 175–187 (1985)
8. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge (2004)
9. Hirschmüller, H.: Stereo Processing by Semiglobal Matching and Mutual Information. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(2), 328–341 (2008)
10. Hirschmüller, H., Scharstein, D.: Evaluation of Stereo Matching Costs on Images with Radiometric Differences. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(9), 1582–1599 (2009)
11. Huang, H., Mayer, H.: Generative Statistical 3D Reconstruction of Unfoliaged Trees from Terrestrial Images. Annals of GIS 15(2), 97–105 (2009)
12. Huber, P.: Robust Statistics. John Wiley & Sons, Inc., New York (1981)

13. Jian, Y.D., Balcan, D., Dellaert, F.: Generalized Subgraph Preconditioners for Large-Scale Bundle Adjustment. In: Thirteenth International Conference on Computer Vision, pp. 295–302 (2011)
14. Leberl, F., Bischof, H., Pock, T., Irschara, A., Kluckner, S.: Aerial Computer Vision for a 3D Virtual Habitat. IEEE Computer 43(6), 24–31 (2010)
15. Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
16. Mayer, H.: Efficiency and Evaluation of Markerless 3D Reconstruction from Weakly Calibrated Long Wide-Baseline Image Loops. In: 8th Conference on Optical 3-D Measurement Techniques, vol. II, pp. 213–219 (2007)
17. Mayer, H.: Issues for Image Matching in Structure from Motion. In: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. (37) B3a, pp. 21–26 (2008)
18. Mikolajczyk, K., Schmid, C.: A Performance Evaluation of Local Descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(10), 1615–1630 (2005)
19. Nistér, D.: An Efficient Solution to the Five-Point Relative Pose Problem. In: Computer Vision and Pattern Recognition, vol. II, pp. 195–202 (2003)
20. Pollefeys, M., Nistér, D., Frahm, J.M., Akbarzadeh, A., Mordohai, P., Clipp, B., Engels, C., Gallup, D., Kim, S.J., Merrell, P., Salmi, C., Sinha, S., Talton, B., Wang, L., Yang, Q., Stewénius, H., Yang, R., Welch, G., Towles, H.: Detailed Real-Time Urban 3D Reconstruction from Video. International Journal of Computer Vision 78(2-3), 143–167 (2008)
21. Pollefeys, M., Van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J.: Visual Modeling with a Hand-Held Camera. International Journal of Computer Vision 59(3), 207–232 (2004)
22. Pollefeys, M., Verbiest, F., Van Gool, L.: Surviving Dominant Planes in Uncalibrated Structure and Motion Recovery. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part II. LNCS, vol. 2351, pp. 837–851. Springer, Heidelberg (2002)
23. Raguram, R., Frahm, J.M.: RECON: Scale-Adaptive Robust Estimation via Residual Consensus. In: Thirteenth International Conference on Computer Vision, pp. 1299–1306 (2011)
24. Reznik, S., Mayer, H.: Implicit Shape Models, Self Diagnosis, and Model Selection for 3D Facade Interpretation. Photogrammetrie – Fernerkundung – Geoinformation 3(08), 187–196 (2008)
25. Schaffalitzky, F., Zisserman, A.: Multi-view Matching for Unordered Image Sets, or How Do I Organize My Holiday Snaps? In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 414–431. Springer, Heidelberg (2002)
26. Strecha, C., von Hansen, W., Van Gool, L., Fua, P., Thoennessen, U.: On Benchmarking Camera Calibration and Multi-View Stereo for High Resolution Imagery. In: Computer Vision and Pattern Recognition, pp. 1–8 (2008)
27. Torr, P.: An Assessment of Information Criteria for Motion Model Selection. In: Computer Vision and Pattern Recognition, pp. 47–53 (1997)
28. Triggs, B., McLauchlan, P., Hartley, R., Fitzgibbon, A.: Bundle Adjustment – A Modern Synthesis. In: Workshop on Vision Algorithms in conjunction with ICCV 1999, pp. 298–372 (1999)
29. Wu, C.: SiftGPU: A GPU Implementation of Scale Invariant Feature Transform (SIFT) (2007), `cs.unc.edu/~ccwu/siftgpu`