AUTOMATED 3D RECONSTRUCTION OF URBAN AREAS FROM NETWORKS OF WIDE-BASELINE IMAGE SEQUENCES

HelmutMayer, JanBartelsen

Institute of Geoinformation and Computer Vision, Bundeswehr University Munich - (Helmut.Mayer, Jan.Bartelsen)@unibw.de, www.unibw.de/ipk

KEY WORDS: Computer Vision, Virtual Landscape, Close Range Photogrammetry, Visualization, Urban Planning

ABSTRACT:

The efficient automated reconstruction of highly detailed 3D models of urban areas for visualization and analysis is an active area of research for diverse applications ranging from surveillance to architecture. A flexible and cheap data source are wide-baseline image sequences generated with hand-held consumer cameras with several to tens of Megapixels. Image sequences are particularly suitable for the reconstruction of 3D structures along linear objects such as roads. This paper presents an approach for 3D reconstruction from image sequences taken with a weakly calibrated camera with no need for approximations for position and attitude, markers on the ground, or even ground control. The generated 3D reconstruction result is relative, i.e., the scale is not known, but Euclidean, that is, right angles are preserved. The paper shows that the approach allows to produce a 3D reconstruction consisting of points, camera positions and orientations, as well as vertically oriented planes from image sequences taken with a Micro Unmanned Aerial Vehicle (UAV) under challenging wind conditions and without navigation information. Finally, the paper discusses how sequences can be linked into networks, or also images into blocks, clarifying which image configurations exist and how they can be adequately treated when prior knowledge about them is available.

1. INTRODUCTION

A recent special issue of the International Journal of Computer Vision on “Modeling and Representations of Large-Scale 3D Scenes” (Zhu and Kanade, 2008) with a special focus on urban areas exemplifies the importance of the field with applications in “mapping, surveillance, transportation planning, archaeology, and architecture” (Zhu and Kanade, 2008). Of particular interest are (Pollefeys et al., 2008, Cornelis et al., 2008) which like us employ images as primary data source, yet with a focus on video data taken from cars and using GPS and INS data.

Contrary to this, our approach for 3D reconstruction is aiming at wide-baseline scenarios with basically no need for approximations for position and attitude or markers in the scene. While our previous work was on uncalibrated cameras (Mayer, 2005), we now assume that the camera is weakly calibrated, meaning that principal distance and point as well as sheer are known up to a couple of percent. Based on this assumption we can use the 5-point algorithm of (Nistér, 2004) which makes the reconstruction much more stable, particularly for (nearly) planar scenes.

While no approximations for position and attitude are needed and also the images are allowed to be rotated against each other, the images of the sequence still have to fulfill certain constraints to obtain a useful result. First of all, all triplets of images in the sequence have to overlap significantly, to allow for the reliable propagation of 3D structure. Additionally for a reliable matching, the appearance of the visible objects should not change too severely from image to image and there should not be large areas with occlusions.

We introduce our approach to 3D reconstruction from wide-baseline image sequences in Section 2. Besides camera orientations we reconstruct 3D points and from them planes which are a good means to describe dense 3D structure in urban areas, e.g., to determine visibility.

This gives way to the 3D reconstruction from image sequences taken from a Micro Unmanned Aerial Vehicle (UAV) presented in Section 3. In spite of the lack of information on strongly varying position and attitude of the camera we could still orient the images and produce a 3D model including textured planes.

The experiences with the UAV led us to an analysis of different imaging configurations, consisting of sequences which can be linked at the ends or also in between, in both cases leading to networks, as well as more random configurations which can give way to image blocks. In Section 4 we show how the different configurations can be adequately treated. We finally end up with conclusions.

2. 3D RECONSTRUCTION FROM WIDE-BASELINE

Our current approach for 3D reconstruction from wide-baseline image sequences extends (Mayer, 2005) to a (weakly) calibrated setup. It starts by extracting points (Förstner and Gülch, 1987). The eigen-vectors of the points are employed to normalize the orientation of the image patches (Mayer, 2008) subsequently used for cross-correlation employing color information. If the correlation score is beyond a low threshold of 0.5, affine least squares matching (LSM) is used. Matches are checked a second time via the correlation score after matching, this time with a more conservative threshold of 0.8.

From corresponding points in two or three images essential matrices or calibrated trifocal tensors (Hartley and Zisserman, 2003) are robustly computed using the five point algorithm by (Nistér, 2004) in conjunction with Random Sample Consensus - RANSAC (Fischler and Bolles, 1981). To obtain a more reliable solution we employ the robust geometric information criterion - GRIC of (Torr, 1997). For three images two times the five point algorithm is employed with the same reference image and the same five points in the reference image. Because of the reference image, the only unknown is the relative scale which is computed as the median of the ratios of the distances to the five 3D points in the two models.

We employ image pyramids to make the procedure more efficient. Because of this we can afford to use as initial search space the whole image, though on the highest pyramid level with a typical resolution of 100×100 pixels. On the second and third highest level the epipolar lines derived from the essential matrices and trifocal constraints are employed, respectively.

After reconstructing triplets they are linked based on the overlapping images. E.g., triplets consisting of images 1 -2 -3 and 2 -3 -4 overlap by images 2 and 3. For those two images projection matrices can be computed from the trifocal tensors (Hartley and Zisserman, 2003) and from them in turn a Euclidean transformation mapping from the first to the second triplet. In (Mayer, 2007b) we have shown how to speed up linking by conducting it hierarchically, at the same time avoiding also a bias in the estimation process due to the combination of sequences of very different lengths (e.g., when one links 3 images to 90 images). During linking we also track points by projecting them into newly linked images and determining the image coordinates via LSM, resulting in highly precise n-fold points.

The linking of the triplets is done on the second or third highest level of the pyramid, depending on the image size. After linking the points are projected into the original resolution images, once again producing highly accurate relative coordinates by means of LSM.

After all steps we employ robust bundle adjustment (McGlone et al., 2004). E.g., also when estimating essential matrices and trifocal tensors we compute a bundle solution every couple of hundred iterations as we found that only the maximum likelihood bundle solution is reliable for difficult sequences (Mayer, 2008)

The outcome of the above process are the relative orientations of cameras as well as 3D points. The coordinate system is fixed to the first camera and the scale is determined by the base from the first to the second camera for which the length is set to one. While this gives basic information about the 3D structure of the scene, it does not allow, e.g., to compute visibility. Ideal for this would be dense depth maps, but there is no standard robust approach for their computation available. Recent approaches such as (Strecha et al., 2004, Lhuillier and Quan, 2005, Hirschmüller, 2008) all have their shortcomings.

(Pollefeys et al., 2008) have shown dense depth maps computed in real-time for extended areas, but the resulting 3D model suffers from occlusions and incorrect shapes as no information about the imaged objects is included. (Cornelis et al., 2008) make use of the knowledge that facades or similar objects are imaged by employing ruled surfaces parallel to the vertical direction. This improves the result, but still some non-vertical objects are not reconstructed with their correct shape. Finally we note, that (Pollefeys et al., 2008) and (Cornelis et al., 2008) both employ dense video data, which considerably restricts the search space, thus allowing for real-time processing on graphical processing units (GPU).

As we focus on urban scenes where planes are abundant and often describe important objects such as walls, we decided to determine planes from the 3D points. Particularly, we follow (Mayer, 2007a). Because the vertical direction is predominant in urban scenes, we determine it first from the image of the vanishing point in the form of the intersection point of the projections of vertical lines in the scene into the images computed by means of RANSAC. Orienting the whole scene vertically helps considerably to determine the boundaries of the partially vertical planes.

The planes themselves are also obtained by RANSAC and additionally least squares adjustment. For the planes two parameters must be given by the user: A threshold determining the distance of points from the plane and the maximum distance between points on the plane, the latter avoiding large planes consisting of a dense cluster of correct points and few randomly distributed points which by chance lie on the plane.

For each plane texture is determined also for partially occluded regions by means of a consensus approach (Mayer, 2007a). The latter allows to reconstruct the correct texture even it is visible in less than 50% of the images which can see the particular region. The results of plane reconstruction have been used for facade interpretation (Mayer and Reznik, 2007, Reznik and Mayer, 2007).

3. 3D RECONSTRUCTION FROM IMAGES FROM A MICRO UAV

A Micro UAV is a very small and light UAV. Thus, it is very appropriate to explore built-up areas. It renders it possible to fly through streets and into courtyards and to take there images of buildings and their facades from an off-ground perspective independently of ground conditions or obstacles on the ground.

In our first experiments we investigated if images from a Micro UAV can be used for 3D reconstruction. We employed a quad-copter, i.e., a UAV with four rotors, with a diameter of 1 meter, and a weight under 1 kg. It carried a ten Megapixel consumer camera. Figure 1 shows the planned image configuration ”Circle”.

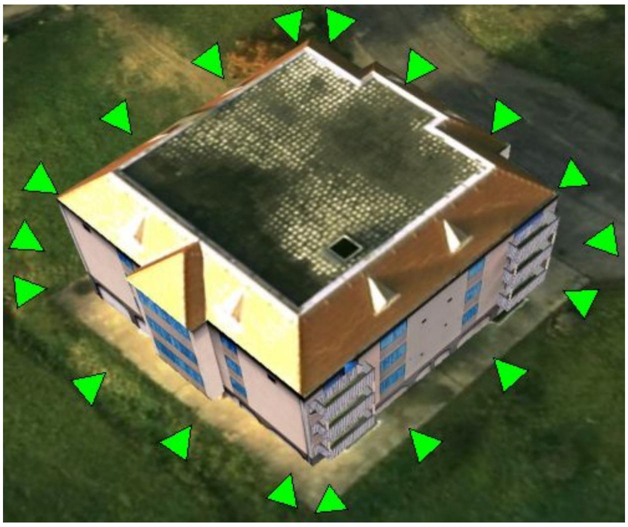

Figure 1. Planned image configuration ”Circle”.

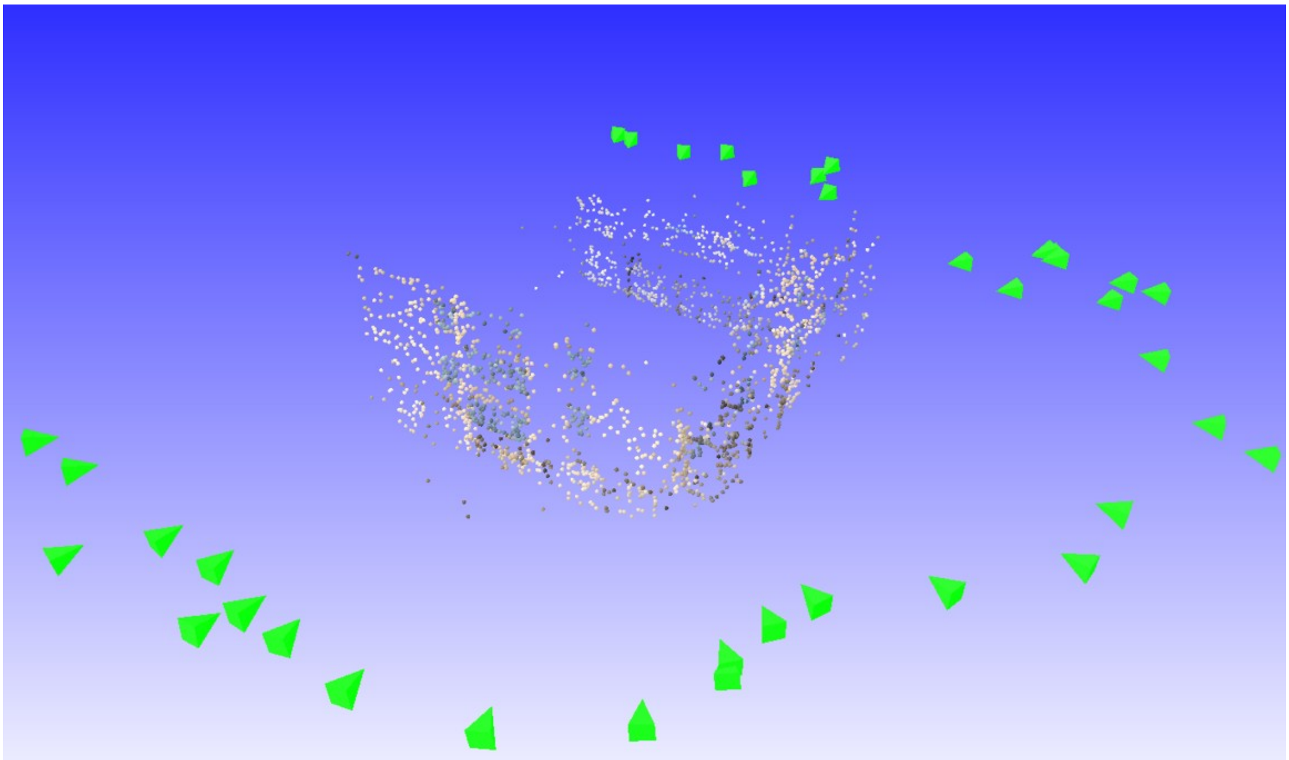

Figure 3. Result for image sequence Three Walls with 36 images, no approximations for position and attitude given, $\sigma0$= 0.67 pixels.

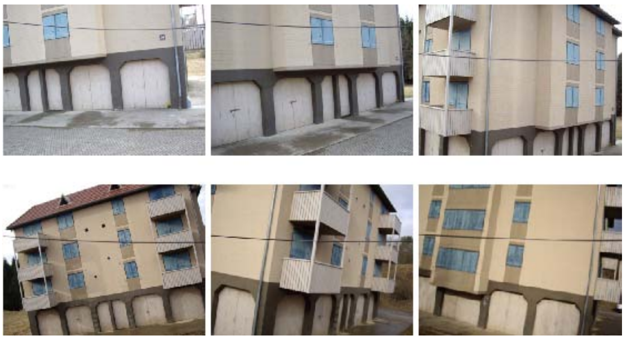

Figure 2. Image sequence Three Walls, images 3, 9, 12, 20, 25, and 28 - Please note the dark approximately horizontal line which stems from the ring around the UAV to secure its rotors against hitting other objects.

The experiments were conducted under challenging wind condi-tions. Due to the very light UAV it was difficult to fly to a position - already slight gusts of wind had a strong effect on the UAV. The images were thus taken only very roughly on a circular path, they had a rather variable orientation (cf. Figure 2), and several of them were not in the correct order. As the order in the sequence is a basic requirement of our approach for 3D reconstruction, we had to organize the images into a continuous sequence with threefold overlap by hand.

A sequence with 36 images could be found for which a 3D model could be reconstructed (cf. Figure 3) showing Three Walls. Figure 4 presents the reconstructed planes also together with the projected texture computed by means of consensus (cf. Section 2).

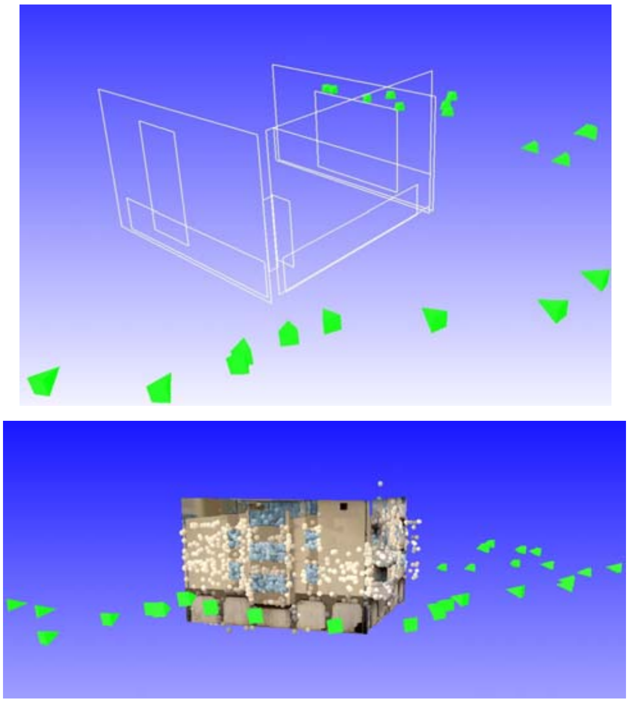

Figure 4. Result for image sequence Three Walls with recon structed planes and projected texture (bottom).

From the fourth, Western Wall no images which could be linked to the Three Walls could be taken due to the strong wind during the experiments. We only realized an additional problem after noticing relatively bad orientation results and a more thorough inspection of the images. It resulted from a ring around the

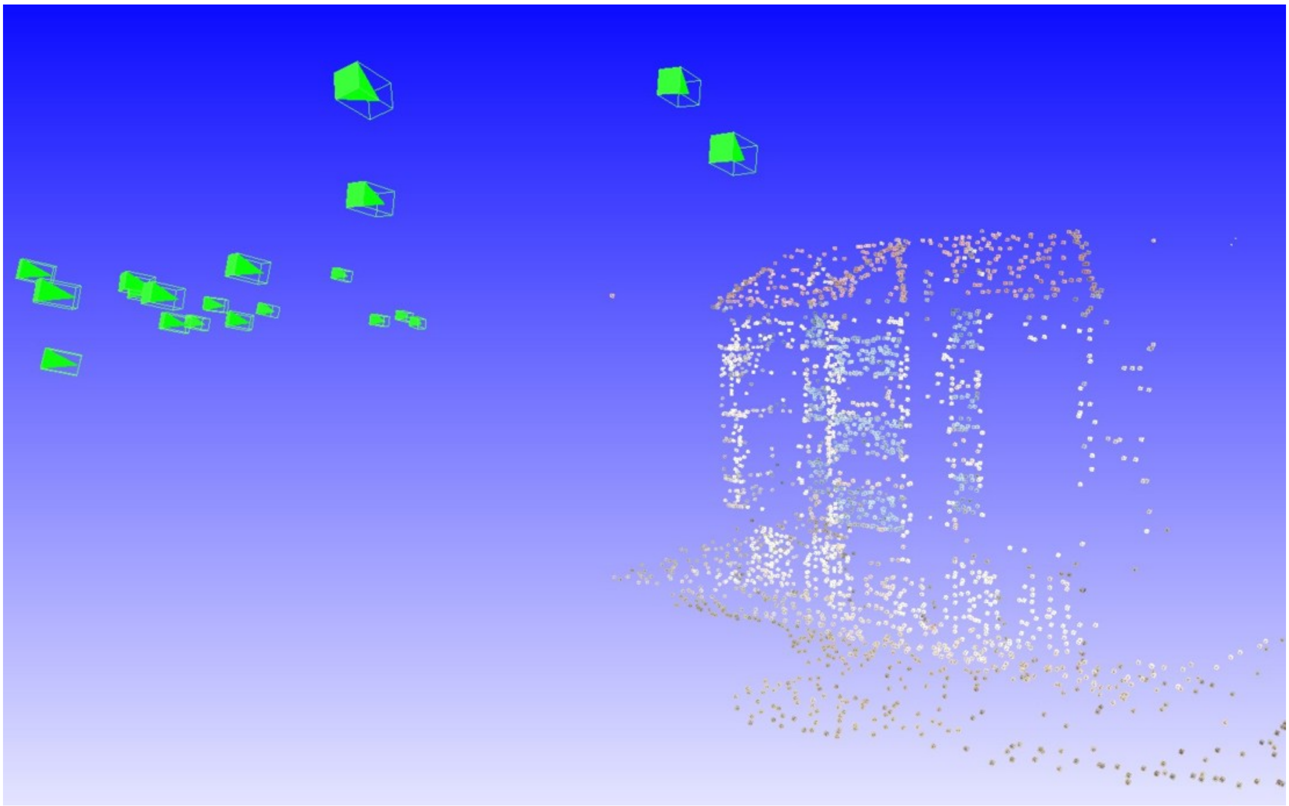

Figure 6. Result for image sequence Southern Wall with 21 images, no approximations for position and attitude, $\sigma 0$= 0.39 pixels.

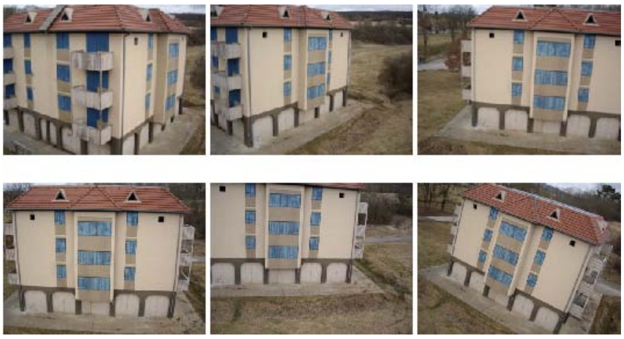

Figure 5. Image sequence Southern Wall, images 3, 6, 13, 15, 17, and 21.

UAV which secures the rotors of the UAV against hitting other objects. This ring was visible for the camera and the camera focused on the ring in several instances leading to blurred images and thus to the very bad standard deviation of 0.67 pixels. (Please note that we usually obtain 0.1 to 0.15 pixel standard deviation in other experiments with a similar camera.)

To get rid of the problem, the camera was tilted downwards for the further experiments. Yet, it was thus necessary to fly higher when taking images (cf. Figure 5). The result in Figure 6 presents camera positions higher above the ground and has a lower standard deviation. In Figure 7 again planes and texture mapped on them is shown.

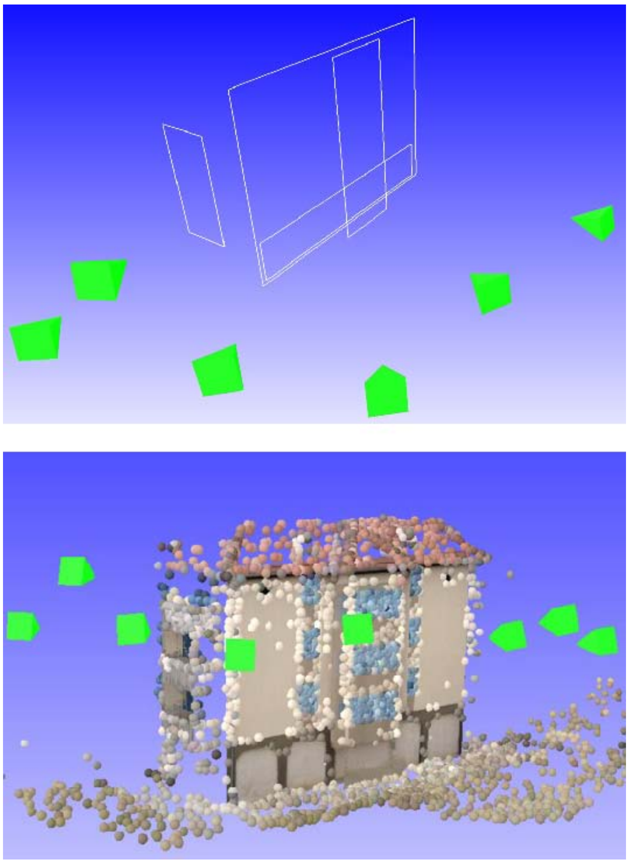

Figure 7. Result for image sequence Southern Wall with reconstructed planes together with the projected texture (bottom).

4. LINKING IMAGES SEQUENCES INTO NETWORKS

Our approach presented in Section 2 assumes that individual sequences of images are given. It has been validated with a large number of wide baseline sequences (Mayer, 2005, Mayer, 2008). Yet, the experiments with the UAV presented above, which led to more or less long independent image sequences, made clear, that it is important to be able to treat extended image configurations, very often in the form of networks of linked sequences, in an adequate way.

Linking all images independently as, e.g., in (Schaffalitzky and Zisserman, 2002), is the generic solution which works for sequences, networks, i.e., sequences linked at certain positions, or general blocks. (Schaffalitzky and Zisserman, 2002) avoid matching all pairs of images by transforming the pairwise matching problem into a correspondence problem in feature space made up of two stages. In the first a table of point features versus views is calculated, giving for each feature point its putative matches over multiple views. The second stage improves the quality of the matches by a number of global ”clean-up” operations outputting a table with considerably more correct matches. The complexity of the method is $O(n + \frac{1}{2}n^2 - \frac{1}{2}n)$.

While (Schaffalitzky and Zisserman, 2002) is a generic method, it is often far from optimal, as still a hard decision problem has to be solved: Are two images related just by chance, or do the few matches actually result from wide baseline perspective distortion, occlusions, and a small overlap? While a full least squares solution can help to improve the reliability (Mayer, 2008), it is still advantageous to make use of the fact that usually images are taken in the form of partial sequences. I.e., images which can be matched are often taken directly after each other. Using this information should improve reconstruction, though we note that it might still be not easy to decide, when a sequence ends. Here, again GRIC (Torr, 1997) might be helpful. We have analyzed the situation for (partial) sequences and have found several configurations, which are shown in Figure 8.

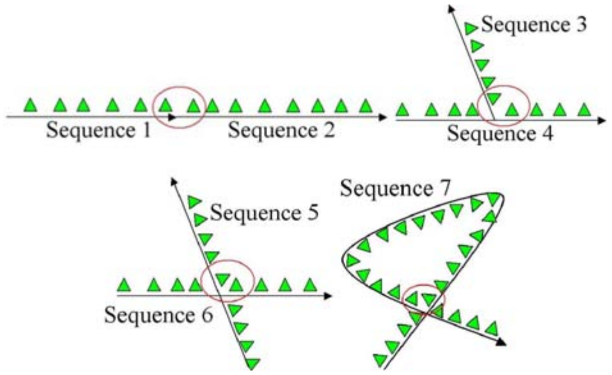

Figure 8. Configurations for the (internal) linking of image sequences.

The most simple configuration in terms of complexity consists of two sequences (1 and 2 in Figure 8) which can be linked at one of their ends. [1] For this configuration only four possibilities have to be checked. More effort is needed if the end of one sequence (Sequence 3 in Figure 8) can be linked to any image in the other sequence. 2*(*n*+*m*) possibilities have to be checked, where *n* and *m* are the numbers of images of the two sequences.

The worst configuration for sequences is if any image of a sequence with *n* images can be linked with any image of the other sequence with *m* images (Sequences 5 and 6 in Figure 8). For this configuration *n***m* possibilities exist. I.e., if *n* is 8 and *m* is 12 for a set of 20 images, there are 96 possibilities. A special configuration is given by Sequence 7 in Figure 8 with an intersection inside the sequence. Here it is possible to use the camera positions and orientations to predict possibly overlapping images and thus to avoid an exhaustive search.

Finally, if nothing about the order of the images is known, i.e., in the extreme case of exhaustive search for threefold overlapping images in a set of *n* images (here we cannot make use of neighboring third images as in sequences)

$$\binom{n}{3} = \frac{n!}{(n-3)! \cdot 3!} = \frac{1}{6}n^3 - \frac{1}{2}n^2 + \frac{1}{3}n$$

checks would be needed. For a set of 20 images there are 1,140 possible triplets. Though the complexity is polynomial, it is too expensive in practice and thus the problem can only be solved by a strategy transforming the problem into the feature space such as (Schaffalitzky and Zisserman, 2002).

The above discussion shows, that it should be advantageous from the point of view of computational complexity, but also reliability, to make use of partial order in a set of images. We particularly plan to let the user decide how the system should proceed. In many cases a strategy should be advantageous where one starts with trying to link adjacent images and by this means constructs partial sequences. Depending on the user knowledge one then ei-ther tries to link the sequences at the ends, ends are tested against all images in the other sequences, or whole sequences are checked image for image against each other all with the goal to construct networks. (Self-intersections are in all cases handled automatically as this is not a combinatorial problem.)

If the user indicates, that there is no order in the images, it is useful to employ an approach such as (Schaffalitzky and Zisserman, 2002). It should also be helpful for linking sequences if not only the ends are to be matched.

5. CONCLUSIONS AND OUTLOOK

We have shown that from images taken with a weakly calibrated camera from a Micro UAV, for which neither a reliable positioning was possible due to challenging weather conditions, nor even any approximations for position and attitude are available, it is still possible to conduct a reliable and precise 3D reconstruction without markers or even ground control.

For the future, we plan to make more experiments with an up-graded Micro UAV more stable against wind. Another possibility for improved results could be a digital camcorder with a resolu-tion in the range of Megapixels, which can be carried by a Micro UAV. Due to the high frame rate the images

[1] We note that also here threefold overlap is necessary. Thus, actually in this and all other configurations a third image neighbored to one of the pair of linked images has to overlap the pair. To make the description more readable, we decided not to discuss this issue in the remainder of the text.

are taken close in space to each other and thus they form natural sequences.

Finally, we are on the way to implement what we have discussed in Section 4. The aim are image configurations consisting of networks of sequences in built-up areas which will be used for highly detailed 3D reconstruction. What is particularly lacking is the possibility to match images independent of image scale. While the Scale Invariant Feature Transform – SIFT (Lowe, 2004) is the obvious solution, we also consider PCA-SIFT (Ke and Sukthankar, 2004) and the findings of (Mikolajczyk et al., 2005, Mikolajczyk and Schmid, 2005).

REFERENCES

Cornelis, N., Leibe, B., Cornelis, K. and Van Gool, L., 2008. 3D Urban Scene Modeling Integrating Recognition and Reconstruction. *International Journal of Computer Vision* 78(2–3), pp. 121-141.

Fischler, M. and Bolles, R., 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM* 24(6), pp. 381–395.

Förstner, W. and Gülch, E., 1987. A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centres of Circular Features. In: *ISPRS Intercommission Conference on Fast Processing of Photogrammetric Data*, Interlaken, Switzerland, pp. 281–305.

Hartley, R. and Zisserman, A., 2003. *Multiple View Geometry in Computer Vision* – Second Edition. Cambridge University Press, Cambridge, UK.

Hirschmüller, H., 2008. Stereo Processing by Semi-Global Matching and Mutual Information. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(2), pp. 328–341.

Ke, Y. and Sukthankar, R., 2004. PCA-SIFT: A More Distinctive Representation for Local Image Descriptions. In: *Computer Vision and Pattern Recognition*, Vol. 2, pp. 516–523.

Lhuillier, M. and Quan, L., 2005. A Qasi-Dense Approach to Surface Reconstruction from Uncalibrated Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(3), pp. 418–433.

Lowe, D., 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60(2), pp. 91–110.

Mayer, H., 2005. Robust Least-Squares Adjustment Based Orientation and Auto-Calibration of Wide-Baseline Image Sequences. In: *ISPRS Workshop in conjunction with ICCV* 2005 “Towards Benchmarking Automated Calibration, Orientation and Surface Reconstruction from Images” (BenCos), Beijing, China, pp. 1–6.

Mayer, H., 2007a. 3D Reconstruction and Visualization of Urban Scenes from Uncalibrated Wide-Baseline Image Sequences. *Photogrammetrie - Fernerkundung - Geoinformation* 3/07, pp. 167– 176.

Mayer, H., 2007b. 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.

Mayer, H., 2008. Issues for Image Matching in Structure from Motion. In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. (37) 3.

Mayer, H. and Reznik, S., 2007. Building Facade Interpretation from Uncalibrated Wide-Baseline Image Sequences. *ISPRS Journal of Photogrammetry and Remote Sensing* 61(6), pp. 371–380.

McGlone, J., Bethel, J. and Mikhail, E. (eds), 2004. Manual of Photogrammetry. American Society of Photogrammetry and Remote Sensing, Bethesda, USA.

Mikolajczyk, K. and Schmid, C., 2005. A Performance Evaluation of Local Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(10), pp. 1615–1630.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T. and van Gool, L., 2005. A Comparison of Affine Region Detectors. *International Journal of Computer Vision* 65(1/2), pp. 43–72.

Nistér, D., 2004. An Efficient Solution to the Five-Point Relative Pose Problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(6), pp. 756–770.

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., Stew´enius, H., Yang, R., Welch, G. and Towles, H., 2008. Detailed Real-Time Urban 3D Reconstruction from Video. *International Journal of Computer Vision* 78(2–3), pp. 143–167.

Reznik, S. and Mayer, H., 2007. Implicit Shape Models, Model Selection, and Plane Sweeping for 3D Facade Interpretation. In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. (36) 3/W49A, pp. 173–178.

Schaffalitzky, F. and Zisserman, A., 2002. Multi-view Matching for Unordered Images Sets, or “How Do I Organize My Holiday Snaps?”. In: *Seventh European Conference on Computer Vision*, Vol. I, pp. 414–431.

Strecha, C., Fransen, R. and Van Gool, L., 2004. Wide-Baseline Stereo from Multiple Views: A Probabilistic Account. In: *Computer Vision and Pattern Recognition*, pp. 552–559.

Torr, P., 1997. An Assessment of Information Criteria for Motion Model Selection. In: *Computer Vision and Pattern Recognition*, pp. 47–53.

Zhu, Z. and Kanade, T., 2008. Modeling and Representations of Large-Scale 3D Scenes. *International Journal of Computer Vision* 78(2–3), pp. 119–120.