DENSE IMAGE MATCHING IN AIRBORNE VIDEO SEQUENCES

M. Gerke

International Institute for Geo-Information Science and Earth Observation – ITC, Department of Earth Observation Science, Hengelosestraat 99, P.O. Box 6, 7500AA Enschede, The Netherlands, gerke@itc.nl

ICWG III/V

KEY WORDS: Video, Surface, Matching, Resolution

ABSTRACT:

The use of airborne video data is gaining increasingly attention in the photogrammetric community. This interest is driven by the availability of low-cost sensor platforms like UAV and low-cost sensors such as digital (video) consumer cameras. Moreover, a wide range of applications are related to this kind of sensor data, e.g. fast mapping in case of disasters, where geometric and semantic information on a particular scene has to be captured within a small timeframe.
The advantage of video data against wide baseline images is that tracking algorithms can be used to derive highly redundant tie point information in a fully automatic manner. One drawback is that due to the reduced resolution and only short exposure time, the image quality is worse compared to the quality provided by mapping cameras. However, the many-fold overlapping enables the use of multiframe super resolution techniques to obtain higher quality textures.
In this paper the focus lies on the dense surface reconstruction using airborne video sequences. The first step in the approach consists of retrieving the structure and motion of the cameras, also incorporating geometric knowledge on the scene. In the subsequent step a dense surface reconstruction is applied. First, appropriate image combinations for the stereo matching are selected. After rectification, the Semi-Global Matching technique is applied, using the Mutual Information approach for retrieving local energy costs. After the matches are linked, super resolution images are computed and 3D point clouds are derived by forward intersection.
The results for two datasets show that the super resolution images have a higher nominal resolution than the original ones. As the accuracy of the forward intersection depends on the actual image acquisition parameters, the unfiltered 3D point cloud could be noisy. Therefore, some further improvements for the 3D point coordinates are identitied.

1 INTRODUCTION

For many applications dense surface reconstruction from images is becoming an interesting alternative to laserscanning. In the context of airborne remote sensing metric digital cameras are available which are able to acquire high resolution images at high overlapping ratio. This availability stimulates the development of sophisticated approaches to dense matching and surface reconstruction (Hirschmüller et al., 2005, Zebedin et al., 2006). The advantage over LIDAR in those cases is that besides the derivation of a DSM, further products like (true) orthoimages of high resultion are computable right away.

The dense surface reconstruction is also interesting in other fields; in close range applications the focus is on the reconstruction of single (man-made) objects or even whole cities. In those cases the high overlapping is often achieved by using video data, see e.g. (Pollefeys et al., 2004). The advantage of video over single wide-baseline shots is the high redundancy of observations through the high overlapping which can be exploited to retrieve correspondences and thus camera pose and calibration information through tracking algorithms (Shape from Motion).

In between those two domains – airborne remote sensing being primarily used for mapping purposes and video based reconstruction of man-made object – one can find the field of airborne remote sensing from low altitude platforms, like helicopters or Unmanned Airborne Vehicles (UAVs) (Eisenbeiss and Zhang, 2006, Förstner and Steffen, 2007). Due to its flexibility and low costs for operation, UAVs are interesting for a lot of applications. Using an UAV equipped with a video camera enables to combine having an overview on a certain area of interest with the advantages of using dense image sequences to retrieve geometric and semantic information. The challenges one is facing when working with this kind of data are manifold, e.g. the motion of the vehicle may not be smooth, and the image scale might be smaller than in the aforementioned cases, influencing the available accuracy and reliability.

The focus of this paper is on the implementation of a strategy for dense image matching in airborne video sequences. The goal is to derive two datasets: one are so-called super resolution images where the multiple observation of the scene of interest is exploited to derive noise reduced images with a higher nominal resolution than the original ones. The second dataset is a dense 3D point cloud as derived from forward intersecting the matched points. The paper is meant as a case study where known approaches and algorithms are used to set-up a practical workflow for the processing of airborne video data. The results will show the potential of the applied techniques, but also reveal some open issues.

The remainder of this paper is organised as follows: The next section describes the established workflow to process the data, including some links to the applied literature. In section 3 some experiments are described: After the outlining of two different datasets, the obtained results are shown and evaluated. Some conclusions from those case studies and an outlook to further work are given in the last section.

2 WORKFLOW AND METHODS

The workflow as currently realized consists of the following steps (cf. Figure 1):

1. Structure and motion recovery: After feature tracking across the sequence the camera matrices are computed through bundle adjustment.

2. Dense stereo matching: From the sequence some stereo pairs are chosen according to a pre-defined strategy, see next point. The pairs are first rectified and afterwards the Semi-Global Matching approach is applied to derive dense disparity information.

3. Linking of matches: The idea behind linking the stereo matches is to increase the effective baseline for forward intersection and to reach redundancy.

4. Computation of super resolution images: With the applied strategy for matching and linking the pixels in the images which participated in the matching process were observed multiple times. This fact is exploited to compute images preserving the same geometry as the original ones, but with an enhanced image quality regarding the noise and the effective resolution.

5. Forward intersection: The correspondence information as retrieved from the matching and the subsequent linking are used for multi-view forward intersection to obtain 3D coordinates for the matched points including colour information.

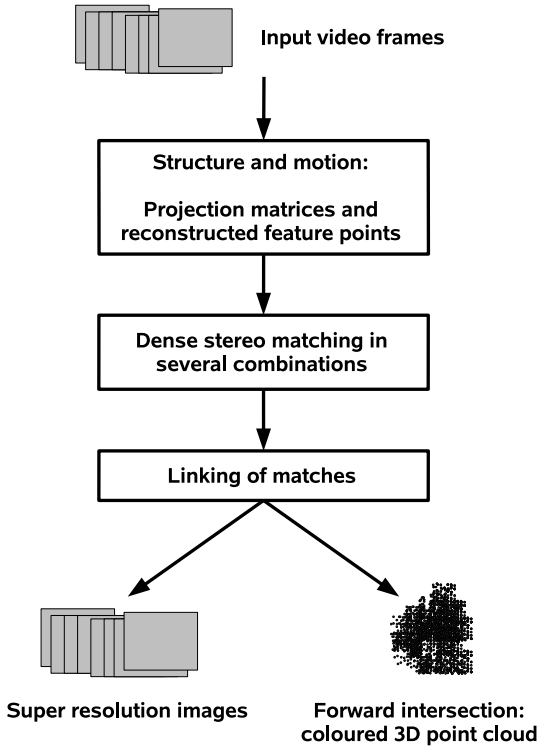

Figure 1: Workflow

2.1 Structure and motion recovery

In most cases where a procedure as described in this paper is applied, uncalibrated, non-metric cameras are used, and in contrast to conventional (airborne) remote sensing, precise navigation information through GPS/IMU is normally not available. Thus, the full information on the individual camera poses throughout the sequence including intrinsic camera parameters need to be recovered from the images. The initial step consists in retrieving image-to-image correspondences by feature tracking. Through a subsequent bundle adjustment including self-calibration, the scene can be reconstructed up to scale if no additional knowledge on the scene geometry is available. Further information on the structure and motion recovery can be found in several sources, e.g. (Hartley and Zisserman, 2004, Pollefeys et al., 2004).

For the implementation of the workflow at hand, the commercial software *Boujou* (2d3, 2008) is currently being used. Besides the fully automatic reconstruction up to scale, it is possible to define constraints on the actual scene geometry, like known distances in object space between feature points. Further, the coordinate frame can be fixed through the definition of plane constraints. As an additional unknown the radial distortion coefficient is estimated and the possibility to compute undistorted images is offered to the user. Refer to (Dobbert, 2005) for detailled information on the approach as implemented in Boujou.

In the subsequent steps the undistorted images, 3D feature coordinates, the corresponding image points and the individual projection matrices are used.

2.2 Matching strategy

The aim of the data processing described in this paper is to derive two final datasets, namely so-called super resolution images and a 3D representation of the scene which can be used e.g. for visualisation tasks. Both products require to establish dense image correspondences. Apart from some special cases (Heinrichs et al., 2007), matching is normally done in stereo image pairs, thus it is required to link stereo correspondences across the sequence. In this paper it is proposed to increase the reliability of matching by applying two kinds of matches: *long baseline matches* and *short baseline matches*, refer also to Figure 2. The basic idea is

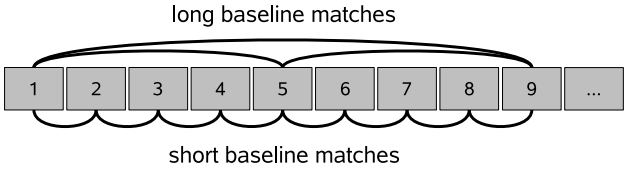

Figure 2: Matching Strategy

that through the short baseline matches correspondences between consecutive matches are established and linked, i.e. a matching pair (m_i, m_{i+1}) in image i is linked with (m_{i+1}, m_{i+2}) and thus establishing the additional match (m_i, m_{i+2}) if the respective pixel m_{i+1} refers to an identical location in image $i+1$. Besides this linking chain, the long baseline matches establish a direct match between the pairs which are already connected through the linked short baseline matches. This procedure results in a higher redundancy of matches and thus helps to increase the reliability: If for instance a correspondence (m_i, m_{i+2}) as derived through short baseline matches does not fit to the direct match (m'_i, m'_{i+2}) from the long baseline match, the correspondences are regarded as wrong and skipped in the subsequent processing.

2.3 Dense stereo matching

The approach to dense stereo matching as applied in the current implementation is the Semi-Global Matching algorithm (Hirschmüller et al., 2005, Hirschmüller, 2008). The basic idea behind this technique is to aggregate local matching costs by a global energy function, which is approximated by an efficient pathwise 1-dimensional optimisation.

The local matching costs can be derived by several methods, like cross-correlation or intensity differences; in the present case they are computed using an hierarchical Mutual Information approach (Viola and Wells, 1997). During cost aggregation not only the local matching cost is considered, but additional penalties are defined by considering disparities in the vicinity of a particular pixel p with the aim to preserve smoothness and height jumps:

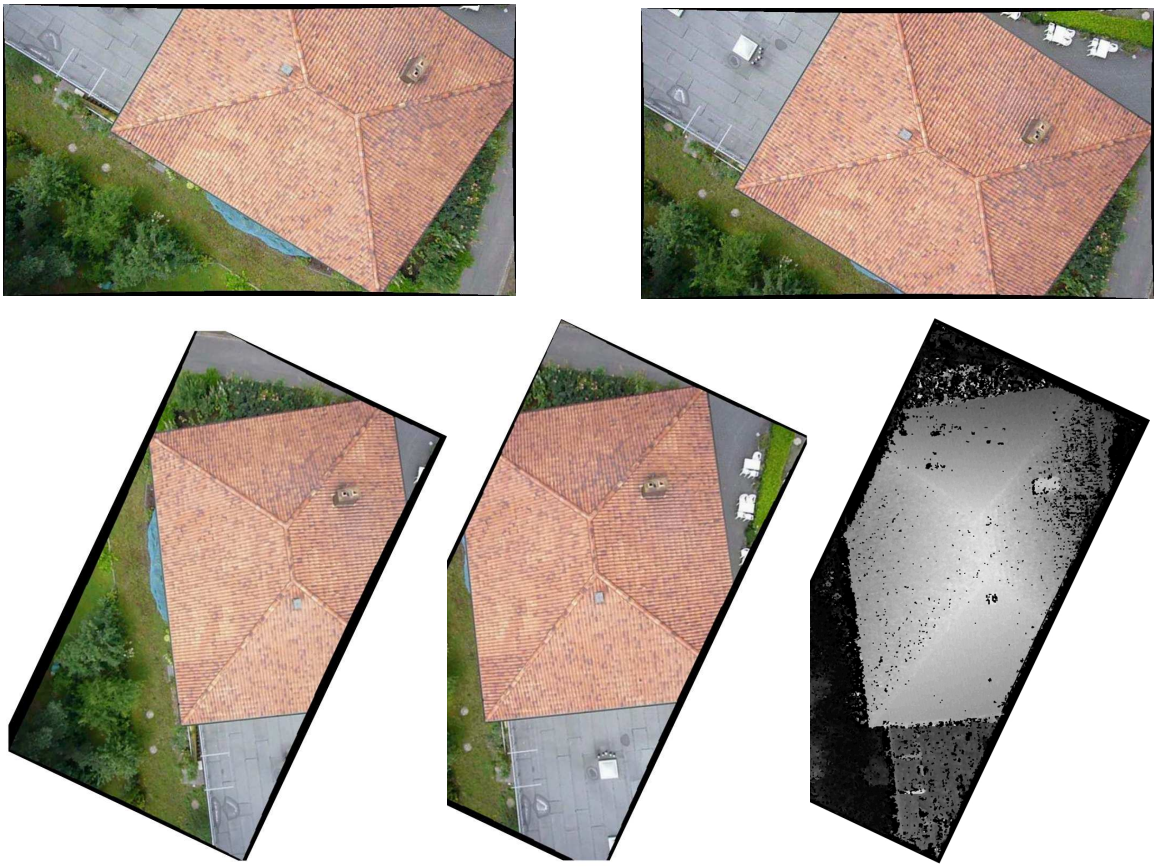

Figure 3: Example for rectification and dense matching. In the upper row the original images are shown, below the rectified pair and the computed disparity map.

one small additional cost P_1 is added if small disparity changes, e.g. 1 pixel, appear in adjacent pixels of p, whereas the larger cost P_2 is applied if larger disparity changes are appearing more far away. The cost P_1 preserves smoothness and due to P_2, height jumps are forced to appear at adjacent pixels, leading to sharp edges in the disparity map. So, for every pixel of interest p of the base image, the aggregated costs need to be computed for every possible disparity, including the penalties from observing the neighbourhood. Finding the final disparity image is equal to the task of minimising the energy for the whole image. Such a procedure would be very inefficient as the complete image must be traversed for every disparity. Instead, the problem is formulated as a 1D-traversing algorithm which sums up the aggregated costs at a particular pixel p and disparity recursively and in different image directions only. In a last step the disparity for a pixel in the base image is selected among all possible disparities as the one causing the least summed-up cost. As subpixel accuracy is desired, a quadratic curve is fitted through the neighboring disparities and the corresponding costs. The minimum of the curve is identified as the optimal disparity value.

To simplify the matching, the images are rectified beforehand. For this purpose the approach proposed in (Oram, 2001) is applied. In contrast to most other techniques for rectification, this approach estimates a non-linear transformation for both images with the aim to minimise perspective distortion effects. Besides the fundamental matrix, the algorithm uses the original matches from the feature tracking to obtain an optimal transformation. In order to further stabilise the transformation, additional features, like available through SIFT (Lowe, 2004) in the case at hand, are incorporated.

In Figure 3 an exemplary dense matching result is shown. The upper row shows the original image from the UAV dataset as described in section 3. The lower row shows the rectified image pair and the disparity map as resulting from the Semi-Global Matching algorithm.

2.4 Computation of super resolution images

In the context of this paper, super resolution images (SRI) refers to the process of computing images preserving the same geometry as the original images from the sequence, but the colour values are computed from the several matches where the image participated. Actually, in the current implementation, two different SRI images are computed: one from the mean value of all corresponding pixel values and one from the median images. As subpixel accuracy is derived from the matching algorithm, the target scale factor for the SRI can be selected larger than 1.

2.5 Forward intersection

A direct solution for the 3D points given observations in multiple images is proposed e.g. in (McGlone et al., 2004, Section 11.1.5). With an unknown 3D point symbolised by X, the corresponding image coordinates in image i by x_i and the respective projection matrix by P_i, the constraint

$$[x_i]_\times P_i X = A_i X = w_i \stackrel{!}{=} 0 \qquad (1)$$

is given ($[x_i]_\times$ defines the skew-symmetric matrix of x_i).

All A_i are assembled in a common matrix A. The error $w^t w$ needs to be minimised, resulting in an optimal point X. This optimal point is the right eigenvector of A belonging to its smallest eigenvalue, computed through a singular value decomposition.

3 EXPERIMENTS

The proposed workflow is demonstrated using two example video datasets. The first dataset *UAV* was obtained from a Microdrone (Microdrones, 2008) platform and captured near the "Drachenfels" close to the city of Bonn (Förstner and Steffen, 2007). The second dataset (*FLI-MAP video*) was captured from a helicopter during a LIDAR flight over Enschede (Fugro, 2008), see Table 1 for some parameters. In that table, the baselength refers to two

Parameter	UAV	FLI-MAP
Flying height H (m)	30	275
Image scale 1 : m_b	1:1,500	1:50,000
Frame size (pix)	848x480	752x582
Pixel size (μm)	12	8.6
Frame rate (Hz)	30	25
Approx. baselength b (m)	0.1	1
Length sequence (img)	280	150

Table 1: Some parameters from the datasets.

consecutive frames. The length of the sequence refers to the number of images which were used for the examples. Noteworthy is the small image scale from the FLI-MAP video; the calibrated focal length of this video device is only 6mm. From this geometric set-up no highly accurate forward intersection can be expected, refer to the section on the resulting point cloud. Some undistorted images from both sequences are shown in Figure 4.

Figure 4: Some frames from both datasets (undistorted images). Upper part: UAV, lower part: FLI-MAP.

3.1 Results

3.1.1 Super resolution images An example for a super resolution image is taken from the UAV dataset. The chosen target scale factor is 1.5. In Figure 5 the gray value profiles (red channel) across the building's roof edge are shown. The left image shows in its upper area a part of the original image, but scaled by factor 1.5 (linear interpolation applied). The line across the edges indicates the location of the grey value profile as shown below the image. The SRI image, computed from the mean value of corresponding points is shown in the right part of Figure 5, including the gray value profile captured at the same image position as in the original image.

In general one can see that the SRI seems to be a bit sharper compared to the original one: The tiles on the roof are less smeared than in the original image. The profile supports the visual impression. Especially in the edge region more details are shown. As an example two points at the profile graph are pointed out by a black arrow. The arrow no. 1 points to the quite salient point in the profile indicating the position of the steep edge where the light grey becomes dark grey. The corresponding area in the profile of the original image is smoother. The arrow no. 2 points to the edge at the eave of the roof where the tiles are showing a lighter colour compared to the red colour on the overall roof area[1]. In the computed SRI image this edge is really existing, but not in the original image.

The SRI as computed from the respective median value of corresponding pixels is not shown here, because no significant difference can be observed compared to the SRI computed from the mean value. This can be explained by the use of solely redundant matches: by this means no gross errors are expected which may influence the SRI from the mean values and thus the robust values from the median computation are close to the mean.

3.2 3D point cloud

In order to evaluate the expected accuracy from forward intersection first a theoretic approximation for the height accuracy is made. In the given examples, especially in the FLI-MAP video, the height component is the critical component.

Generally, given an approximated stereo normal case, the height difference between two points is estimated as

$$H \approx \frac{b \cdot f}{p_x} \tag{2}$$

with p_x: x-parallax.

If only the uncertainty in parallax measures s_{px} is considered, the accuracy for height measurements is derived from the partial derivative with respect to p_x:

$$s_H = \left(\frac{H^2}{b \cdot f}\right) \cdot s_{px} = \frac{H}{b} \cdot m_b \cdot s_{px} \tag{3}$$

In the case at hand more than 2 rays are intersected and thus the expected accuracy and reliability is higher, but nevertheless the approximation reasonably reflects the quality for forward intersection.

To estimate the actual expected accuracy for forward intersection, the parameters from Table 1 need to be inserted into equation 3. The baselength b is chosen in such a way that a sufficient overlapping area is ensured: For UAV it is $b = 2m$ and

[1] The overall intensity of these tiles is larger than on the roof area. However, as the profile shows the red channel, the grey values from the roof area are larger

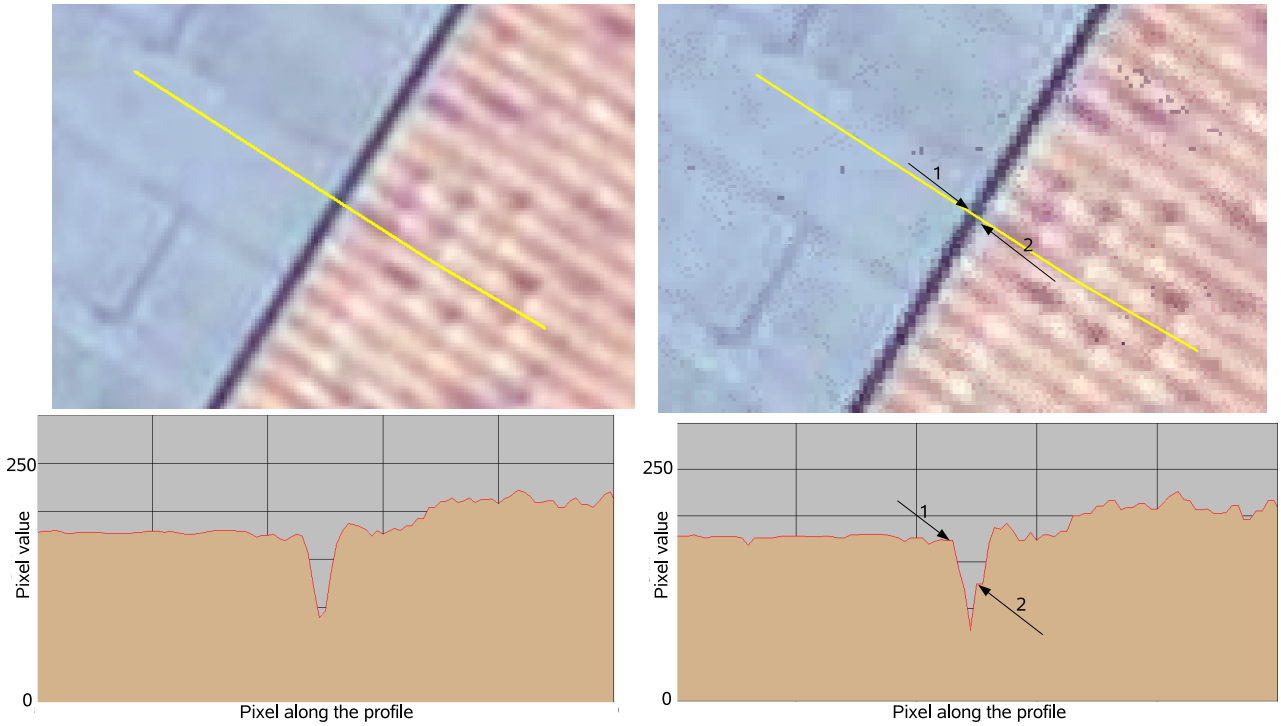

Figure 5: Example for SRI images. The output scale factor is 1.5. Left: original image. Right: SRI computed from the mean values.

for FLI-MAP it is $b = 40m$. The standard deviation for parallax measures is estimated being equal to the pixel size. Setting these numbers into equation 3, the expected accuracy for height measurements is: UAV: $s_H = 0.27m$ and FLI-MAP: $s_H = 2.95m$. The accuracy in X and Y coordinate direction is estimated by $s_X \approx s_Y \approx s_x \cdot m_b$. If s_x is set to the half of the pixelsize, one obtains for the UAV dataset: $s_X \approx s_Y \approx 1cm$ and for the FLI-MAP dataset: $s_X \approx s_Y \approx 20cm$.

In Figures 6 and 7 the resulting point clouds from the UAV and FLI-MAP sequences, respectively are displayed. The upper part shows the colours as captured from the images, i.e. the mean value from all correspondences, and in the lower part the height is coded through the colour, where one full circle in the colour space equals 10m in object space. The FLI-MAP scene is only partly shown.

As no comparison with reference data could be accomplished yet, only a visual inspection of the results is possible. In the point cloud computed from the UAV dataset one could easily identify the roof of the large building. Vertical walls are not visible as this building was only captured from nadir views. In the height coded image also the main structures are clearly visible, but some inaccurate points appear as kind of pepper pattern.

The point cloud as computed from the FLI-MAP dataset is very intersting. If the scene is watched from approximately nadir viewing angle as shown in the figure, the upper point cloud, i.e. with the original colours, looks quite accurate, but the height coded view reveals the problems. There is a lot of noise existing that prevents from identifying the structures. The first observation, namely the good appearance of the true colour model from close nadir view can be explained by the relatively good accuracy of X/Y-coordinate components: a point is projected to a good planar position, but the Z-value is inaccurate. This, however is not observable from a close-to nadir viewing position. In contrast, the error in Z-component is fully reflected in the second view, where the height is coded.

4 CONCLUSIONS AND FURTHER WORK

This paper presents a possible workflow for airborne video data processing up to the computation of so-called super resolution images (SRI) and 3D point clouds. Results are shown for two datasets: the first was captured by a drone at 30m flying height and the second one from a helicopter at 270m flying height above ground. The SRI images shows some more details as the respective original image at a higher nominal resolution. The results for the 3D point cloud shows and confirms that the quality of the computed 3D coordinates largely depends on the flight configuration and camera parameters.

To increase the quality of the obtained 3D coordinates will be the core focus of the further work. One means to filter bad points will be to analyse the covariance matrix of the computed point coordinates. A further way to increase the quality will be to process the points in object space. As shown from the examples, the X/Y-component has a better accuracy than the Z-value. Therefore a strategy could be to correct the Z-component by analysing the colour value of adjacent points in the X/Y-plane.

Next to this, the computation of by-products like 2.5D surface models and orthoimages will be treated in the future, including the semantic interpretation of the scene. According to the results from the SRI computation, it can be expected that the classification and segmentation will benefit from the increased image quality obtained through the large overlapping ratio.

ACKNOWLEDGEMENT

I want to thank Daniel Oram for providing the code for the general rectification on his homepage[2]. Further I like to thank Matthias Heinrichs, TU Berlin, for providing me with his code for Semi-Global Matching. I also want to thank Richard Steffen, University of Bonn, for providing me with the UAV dataset.

[2] http://www.aspa29.dsl.pipex.com (accessed 31 March 2008)

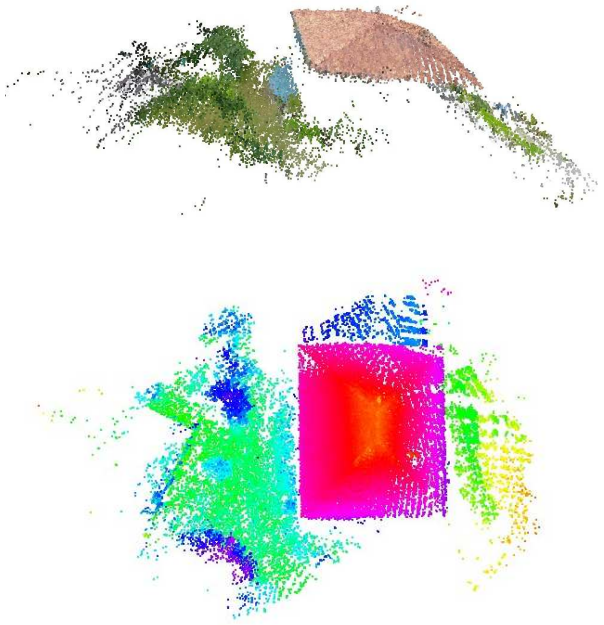

Figure 6: Point cloud derived from the UAV-dataset. Upper image: colour captured from images, lower part: colour used for height coding.

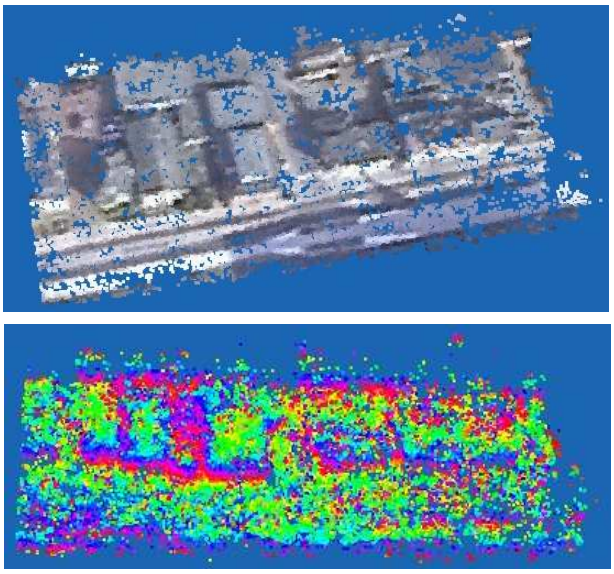

Figure 7: Point cloud derived from the FLI-MAP video-dataset (partly). Upper image: colour captured from images, lower part: colour used for height coding. The background is coloured in blue.

REFERENCES

2d3, 2008. Homepage 2D3: Boujou Software. http://www.2d3.com (accessed 31 March 2008).

Dobbert, T., 2005. Matchmoving: The Invisible Art of Camera Tracking. Sybex.

Eisenbeiss, H. and Zhang, L., 2006. Comparison of DSMs generated from mini UAV imagery and terrestrial laser scanner in a cultural heritage application. In: IAPRS, Vol. 36, pp. 90–96. Part 5 (Comm. 5 Symposium Dresden).

Förstner, W. and Steffen, R., 2007. Online geocoding and evaluation of large scale imagery without GPS. In: Photogrammetric Week, Herbert Wichmann, Heidelberg, pp. 243–253.

Fugro, 2008. Homepage Fugro Aerial Mapping B.V. http://www.flimap.nl (accessed 31 March 2008).

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

Heinrichs, M., Hellwich, O. and Rodehorst, V., 2007. Efficient Semi-Global Matching for trinocular stereo. In: IAPRS, Vol. 36, pp. 185–190. Part 3-W49A (PIA conference, Munich, Germany).

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.

Hirschmüller, H., Scholten, F. and Hirzinger, G., 2005. Stereo vision based reconstruction of huge urban areas from an airborne pushbroom camera (HRSC). In: Lecture Notes in Computer Science: Pattern Recognition, Proceedings of the 27th DAGM Symposium, Vol. 3663, Springer, Berlin, pp. 58–66.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), pp. 91–110.

McGlone, C., Mikhail, E. and Bethel, J. (eds), 2004. Manual of Photogrammetry. 5 edn, ASPRS.

Microdrones, 2008. Homepage Microdrones GmbH. http://www.microdrones.de (accessed 31 March 2008).

Oram, D., 2001. Rectification for any epipolar geometry. In: Proceedings of the 12th British Machine Vision Conference (BMVC).

Pollefeys, M., van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J. and Koch, R., 2004. Visual modeling with a hand-held camera. International Journal of Computer Vision 59(3), pp. 207–232.

Viola, P. and Wells, W. M., 1997. Alignment by maximization of mutual information. International Journal of Computer Vision 24(2), pp. 137–154.

Zebedin, L., Klaus, A., Gruber-Geymayer, B. and Karner, K., 2006. Towards 3D map generation from digital aerial images. ISPRS Journal of Photogrammetry and Remote Sensing 60(6), pp. 413–427.