Evaluation of Digital Surface Models by Semi-Global Matching

HEIKO HIRSCHMÜLLER[1] & TILMAN BUCHER[2]

Summary: This paper considers a processing chain for automatically creating high resolution digital surface models and true ortho-images from aerial and satellite image data. It has been developed at the Institute of Robotics and Mechatronics of the German Aerospace Center (DLR-RM). The processing chain is based on Semi-Global Matching (SGM) that uses a radiometric robust matching cost and an optimization that is based on a global smoothness constraint. SGM is especially suitable for creating models of urban scenes, where sharp depth discontinuities and small details need to be precisely reconstructed. However, the technique also produces very good results in scenes with forest and mountains. In this paper we give an overview of the processing chain and evaluate its results on test data sets from different aerial cameras. It is concluded that SGM permits the creation of high quality surface models that are more accurate and provide much more detail than a surface model from an aerial laser scanner. We also discuss the conditions under which good surface models can be produced by SGM. For very good results, an overlap of 80 % or more along track and 70 % across track should be provided.

1 Introduction

The automatic creation of high quality digital surface models is a topic of active research for applications like large scale city and environment modeling. A processing chain for automatically creating high resolution Digital Surface Models (DSM) from aerial and satellite images has been developed at the Institute of Robotics and Mechatronics of the German Aerospace Center (DLR-RM). The processing chain is based on the Semi-Global Matching (SGM) method (HIRSCHMÜLLER, 2008). It has been implemented on a Linux computer cluster and has processed about 50000 km^2 of data in 5-25 cm/pixel since 2004. The cameras include aerial pushbroom systems like HRSC, which has been developed at the Institute of Planetary Research of the German Aerospace Center, the MFC, that has been developed at the Institute of Robotics and Mechatronics of the German Aerospace Center and the ADS 40 that is manufactured by Leica Geosystems. In the past years, the methods have been extensively applied to many data sets of commercial full frame cameras like UltraCam-D, -X and -Xp from Vexcel or DMC from Intergraph/ZI. Finally, SGM has also been applied to images from commercial satellites, like Quickbird and World View, for creating DSMs with 0.5 m/pixel.

In this paper, we give an overview of the SGM based processing chain in Section 2. Thereafter, we present surface models, computed by different aerial cameras in different ground sampling distances and systematically evaluate them against each other and against a surface model of an aerial laser system in Section 3. Section 4 concludes the paper.

[1] Heiko Hirschmüller, Institut of Robotics und Mechatronics, Department of Perception and Cognition, Oberpfaffenhofen, Germany, heiko.hirschmueller@dlr.de

[2] Tilman Bucher, Institut of Robotics and Mechatronics, Department of Optical Information Systems, Berlin, Germany, tilman.bucher@dlr.de

2 Digital Surface Models by Semi-Global Matching

An important requirement for stereo matching is a precise intrinsic and extrinsic calibration. The remaining error must be less than 1 pixel. However, very good results can be expected if the error is below 0.5 pixel, while higher errors lead to an increased number of wrong matches.
For full frame images, stereo matching is performed between all images which are overlapping by at least 50 %. Typically, we request an overlap of 80% along the flight strip and 70% across. In this setting, each image is automatically matched against six neighbors, i.e. to the two previous and two next images along the flight strip as well as one image of the strips above and below. In this way, most occlusions can be resolved. Additionally, the high redundancy is used for automatically eliminating wrong correspondences during matching. In contrast, images from pushbroom cameras, like HRSC or ADS 40 are processed by matching images captured within the same strip from sensor lines that are arranged in different angles (HIRSCHMÜLLER ET AL., 2005). As for full frame image processing, using more than two sensor lines, arranged in different angles, is beneficial for reducing occlusions and eliminating matching errors.
Regardless of the camera geometry, stereo matching is performed on image pairs by Semi-Global Matching (SGM). The original publication (HIRSCHMÜLLER, 2008) described pixelwise matching using a Mutual Information (VIOLA AND WELLS, 1997) based cost term for compensating radiometric differences. However, a later study that systematically compared matching costs for stereo vision (HIRSCHMÜLLER AND SCHARSTEIN, 2009) indicated that a Census based matching cost (ZABIH AND WOODFILL, 1994) results in almost the same quality, but with increased radiometric robustness.
Census translates both input images individually by encoding the local neighborhood of each pixel into a bit vector. The position of each pixel in the local neighborhood is associated to one bit (Figure 1a). The bit is set, if the pixel has a lower value than the center pixel. Typically, a 9x7 window is used, such that the result is stored into a 64 bit value. The comparison of two pixels is performed by simply computing the Hamming distance[3] of the corresponding bit vectors. It can be seen that Census matching is completely insensitive against a large class of radiometric transformations. In fact, as long as the local order of pixel values does not change, the outcome of matching is exactly the same. Additionally, the property that Census reduces the weight of individual pixels by only storing the information whether or not a value is lower, makes it tolerant against outliers in the window that can be caused by depth discontinuities.
Finding correct correspondences by comparing individual pixels of the Census transformed images is futile, as individual pixels do not contain enough information for unambiguous matching. Local methods like correlation consider a window around each pixel for making it more distinct. This is known to cause errors where the implicit assumption about constant depth within the window is violated. This is best visible at object boundaries that are typically blurred by local methods. In contrast, global methods use a smoothness constraint, which penalizes neighboring pixels that are associated to different depths. This is expressed in a cost function,

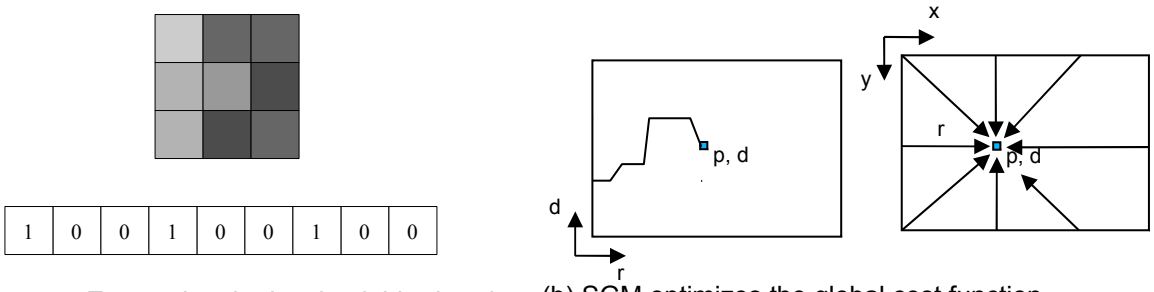
$$E(D)=\sum_p (C_{HAM}(p, D_p)+\sum_{q\in N_p}(P_1 T[|D_p-D_q|=1]+P_2 T[|D_p-D_q|>1])).$$

[3] The Hamming distance is the number of bits that are different.

The cost function takes a disparity image, that encodes the correspondences for all pixels. It sums the pixelwise matching costs (e.g. C_{HAM} for the Hamming distance) over all pixels and adds a small penalty P_1 for neighboring pixels that have slightly different disparities (i.e. depths) and a large penalty P_2 for neighboring pixels with higher disparity differences. The returned value E assesses how well the disparity image fits the encoded constraints. This formulation permits sharp object boundaries, because depth can change abruptly at any pixel in contrast to correlation based methods.

Unfortunately, finding the optimal disparity image for a discontinuity preserving cost is known to be a NP problem for two dimensional images. However, there are a lot of publications in the computer vision community about methods based on Graph Cuts and Believe Propagation that compute approximations. The typical drawback is a long computation time. In contrast, the optimization for the cost function can be done quite efficiently along one dimensional paths through the image using dynamic programming. In the literature, this is commonly applied to image rows, which results in nasty streaking artefacts.

The key idea of SGM (HIRSCHMÜLLER, 2008) is to perform these one dimensional optimizations from all directions through the image as indicated in Figure 1b. For each pixel p, the disparity d is chosen where the sum of costs of paths that reach the pixel at the disparity from eight different directions r is lowest. The quality of SGM is comparable to that of other global methods, but with much higher efficiency regarding computation time.



(a) Census: Expressing the local neighborhood of the center pixel by a bit vector that encodes higher pixels with 1. The value of the center pixel is then replaced by this bit vector.

(b) SGM optimizes the global cost function pathwise from all directions through the volume created by the image dimensions x and y as well as the disparity d.

Figure 1: Computation of the Census matching cost and pathwise optimization of SGM.

The pairwise matching results are fused by selecting the median disparity value for each pixel. The DSM is created by reconstructing the pixels of all disparity images and re-projecting them into an equidistant grid, individually for each image. Thereafter, the information is fused by selecting the median height value in each cell (HIRSCHMÜLLER, 2008).

Since the images are perfectly registered to the DSM, they can be pixelwise re-projected into a true ortho-image. For scenes of cities, the texture at vertical structures like walls is essential for the visual impression. These side-textures are created similar to the ortho-image projection with parallel rays. However, the rays are not orthogonal to the projection plane, but tilted by 20°-25°. In this way, four tilted “ortho”-images are created, that view the scene from left, right, top and bottom (HIRSCHMÜLLER, 2008). These textures are all created fully automatic. Their geometry (i.e. ortho or tilted) is simple enough for texturing the three dimensional reconstruction on-the-fly during visualization (shown in Figure 8 below).

3 Evaluation

Since DSMs can be created from different camera geometries it is important to ask what accuracy can be reached by SGM and what are the differences regarding different cameras.

3.1 Description of the DGPF Data Set

In 2008 a project[4] on the performance of digital airborne cameras has been performed by the German Society of Photogrammetry, Remote Sensing and Geoinformation (DGPF). The test field Vaihingen/Enz has been captured by a wide range of full frame and pushbroom cameras (CRAMER, 2010) for different investigations like the creation of DSMs (HAALA ET AL., 2010). Our paper concentrates on the aerial full frame cameras UltraCam-X (Vexcel), DMC (Intergraph/ZI) and Quattro DigiCAM (IGI), which have captured the test site in a ground sampling distance (GSD) of 8 and 20 cm/pixel. Table 1 describes the most properties of the used data sets.

Table 1: Description of data sets.

Camera	GSD [cm/pixel]	Number of Images used	Image Resolution [MPixel]	Aperture Angle [°]	Radiometric Depth for Matching [Bit]	Overlap [%]	Date [2008]
UltraCam-X	20	52	136	54.7	8	61/ 69	11 Sept.
DMC	20	60	106	69.3	12	63 / 67	06. Aug.
Quattro DigiCAM	20	188 / 4	4 x 39	2 x 33	12	62 / 70	06 Aug.
UltraCam-X	8	215	136	54.7	8	81 / 70	11 Sept.
DMC	8	136	106	69.3	12	63 / 66	24 July
Quattro DigiCAM	8	784 / 4	4 x 39	2 x 33	12	80 / 70	06 Aug.

The Quattro DigiCAM captures four images at the same time, using four tilted cameras, such that the fields of view of the individual cameras are slightly overlapping. However, the four images are not assembled to one large image, but handled as independent images.
The DGPF test also includes data of two pushbroom cameras ADS 40 (Leica Geosystems) and JAS-150 (Jenaoptronic). Unfortunately, we have not been able to process both data sets in time, because the ADS 40 data set did not include the panchromatic channels as level 1, which we require for our software. Furthermore, the JAS-150 data was delivered with extrinsic orientations in a proprietary format together with a Windows library for conversion between image and world coordinates, based on the proprietary format. However, our software is entirely implemented on a Linux cluster. Therefore, we had so far to exclude both pushbroom cameras from our test.
For comparison, the test site has also been captured on 21st August 2008 by the aerial laser ALS 50 (Leica) with 5 points/m^2. The laser data has been projected and interpolated into a regular DSM with 25 cm/pixel. It serves for comparison with the stereo based DSMs.

[4] http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html

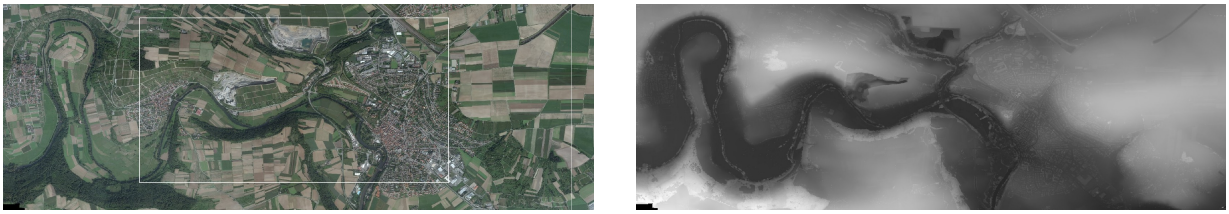


Figure 2: Ortho image (left) and DSM (right) of the UltraCam data set in 20 cm/pixel. The part that is covered in the data sets with 8 cm/pixel is marked by a white square.

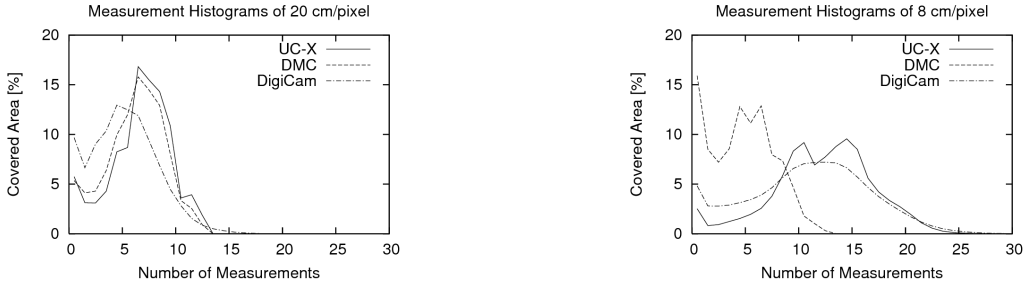


Figure 3: Histograms of the number of pairwise reconstructions that are projected into the DSM cells.

(a) UltraCam-X, 20 cm/pixel

(b) UltraCam-X, 8 cm/pixel

Figure 4: Differences of the stereo and laser DSMs. The range of -10 m (black) to 10 m (white) is shown. The differences of DMC and Quattro DigiCAM are quite similar at this scale and therefore omitted.

3.2 Evaluation of Full Data Sets

All data sets from Table 1 have been processed by SGM using exactly the same parameters that we are always using for processing aerial and satellite data. This means that no parameter tuning has been done. In case of the Quattro DigiCAM data set, the four images that are captured together by the four cameras are taken as independent, individual images. All images are compared to all others. Images that overlap by more than 50 % are matched against each other.

All DSMs have been created in the ground sampling distance of the input images, i.e. with 20 or 8 cm/pixel. We have created the 20 cm DSMs for an area of 7.8 km x 3 km, which is slightly larger than the area captured by laser. The 8 cm DMS's have been created for an area of 4 km x 2.4 km, which is the inner part of the area. Figure 2 shows the ortho image and the DSM of the 20 cm UltraCam-X data set.

As described in Section 2, images are matched pairwise and fused into the DSM. The number of pairwise reconstructions is visualized in Figure 3. Unmatched areas are smoothly interpolated. For the 20 cm data sets, which have a similar overlap, the numbers of reconstructed points (especially UltraCam and DMC) are quite similar (Figure 3). The overall results are comparable. In the 8 cm data sets, the influence of different overlaps is clearly visible. The UltraCam 8 cm

data displays the lowest percentage of unmatched points.
Figure 4 shows the difference of the stereo and the laser DSM. Areas where the stereo models contain larger values are white whereas areas where the laser model contains larger values are black. A comparison to Figure 2 shows that most differences are found in areas with vegetation. Due to the low image overlap of the 20 cm data, large view angles occur, which result in occlusions, especially in the forest and narrow streets. These are interpolated, which tends to result in too large heights. In the 8 cm data with large overlap (e.g. UltraCam), the stereo DSM is more exact than the laser DSM and covers more local maximum and less interpolation. Therefore, the white color mostly means that the ALS data are too low.

3.3 In Detail Evaluation of Scene Parts

Since DSMs of vegetation cannot be properly compared, we have selected a part of the city for detailed comparison. Figure 5 shows the laser DSM of that part. Figure 6 shows the 20 cm stereo

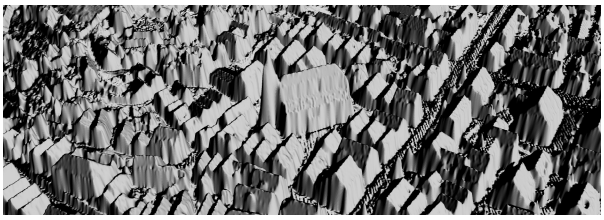
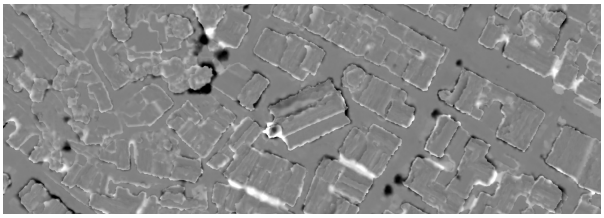


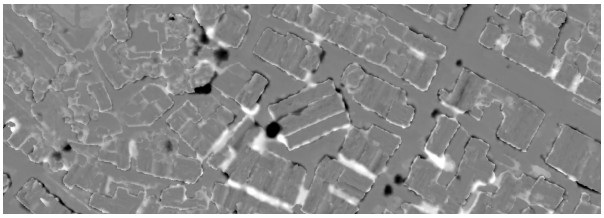
Figure 5: ALS 50, 25 cm/pixel

(a) UltraCam-X, 20 cm/pixel



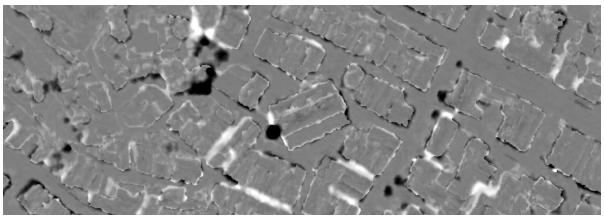
(b) Difference: (a) – Figure 5

(c) DMC, 20 cm/pixel



(d) Difference: (c) - Figure 5

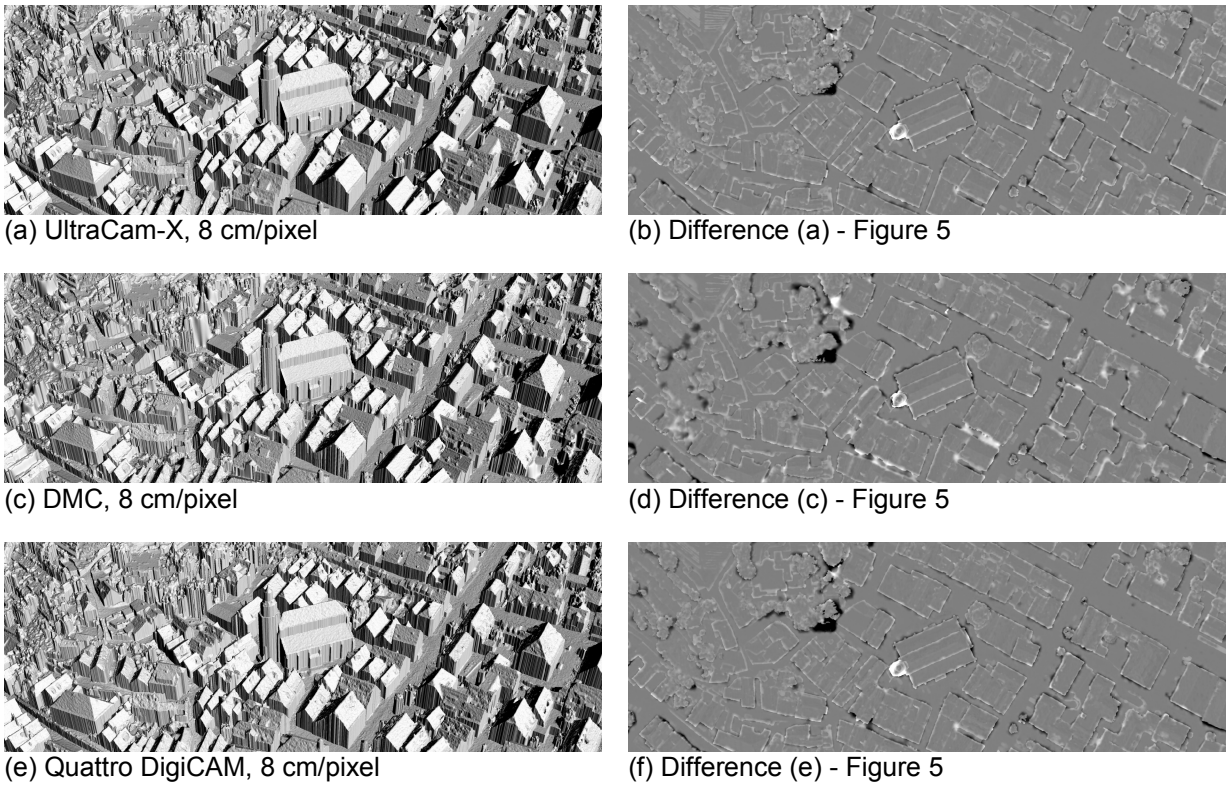
(e) Quattro DigiCAM, 20 cm/pixel



(f) Difference: (e) - Figure 5

Figure 6: Stereo DSMs with 20 cm/pixel and difference to laser DSM.

DSMs as well as the individual differences against the 25 cm laser DSM. The UltraCam DSM appears visually more sharp than the laser DSM. It can be seen in Figure 5b, that almost all differences are at the walls of the houses. A close look reveals that edges of houses are generally lower and the walls are not steep as they should be, but sloped, which seems to be an interpolation effect. Therefore a paired positive/negative effect is observed in the difference images. Even in the data with low image overlap, houses generally appear more accurate than in the laser DSM. The DMC and Quattro DigiCAM data sets are worse than the UltraCam data set in this area (e.g. the tower of the church is missing).
Figure 7 shows the 8 cm stereo DSMs as well as the individual differences against the 25 cm laser DSM. The stereo DSMs appear much more precise than the 20 cm DSMs and the laser. All three DSMs contain much more detail.


(a) UltraCam-X, 8 cm/pixel

(b) Difference (a) - Figure 5

(c) DMC, 8 cm/pixel

(d) Difference (c) - Figure 5

(e) Quattro DigiCAM, 8 cm/pixel

(f) Difference (e) - Figure 5

Figure 7: Stereo DSMs with 8 cm/pixel and difference to laser DSM.

The very high quality that is reached in the 8 cm data sets is also shown in Figure 8 on the example of the UltraCam data set. All buildings appear very precise including many small features on the roofs.

3.4 Evaluation of Surface Profiles

Figure 9 shows a profile through the DSMs. It can be seen that the laser profile contains walls of buildings slightly sloped in contrast to all stereo DSMs. Furthermore, some details are completely missing in the laser DSMs (marked by circle in Figure 9), which is due to the low resolution of the laser system. The detail at the left is actually a power cable that goes over the

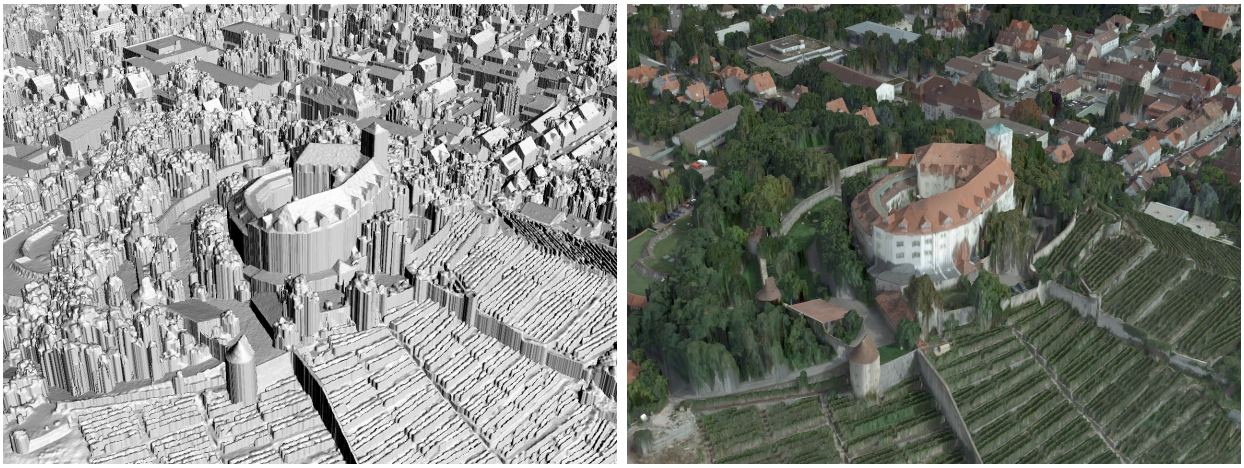


Figure 8: Untextured and automatically textured model from UltraCam-X data set with 8 cm/pixel

roof of the house. The detail on the right is a part of the roof. On the other hand, the 20 cm data of the DMC and Quattro DigiCAM is wrong on the ground near the left part of the church. This is due to the low overlap and large angles that do not allow to properly measure between the houses at least at the resolution of 20 cm/pixel. In contrast, the high resolution of the 8 cm data sets appears very precise in the profile plots.

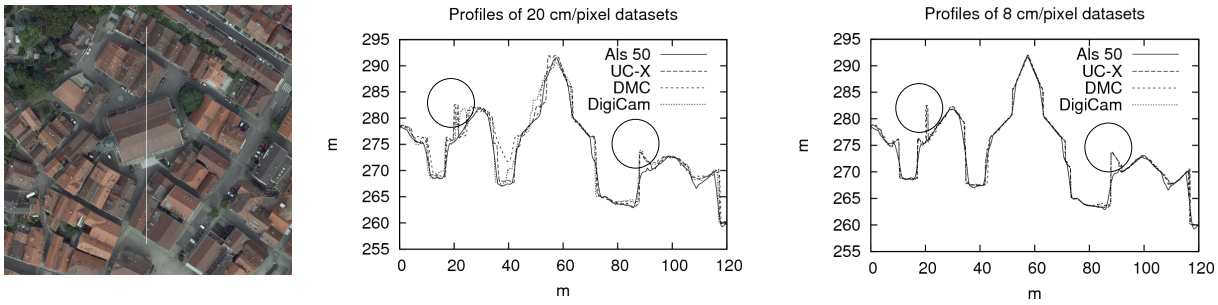


Figure 9: Profiles of the 20 cm and 8 cm data against laser. The circle marks places where all stereo DSMs have correctly picked up real structure that is not detected by the laser.

3.5 Evaluation at Selected Points

We compared the DEM's at 110 selected points that were provided with the test data (Table 2). According to our information, at least a part of these points were used for optimizing the orientation by bundle adjustment. Not all of the points are inside the 8 cm or 20 cm data sets. Additionally, we have ignored one point (3021), because it was interpolated due to missing overlap in the DMC data, which lead to a deviation of 1.34 m.

Since the mean error of most data sets is rather low, we suspect that the large shift of the mean value of the UltraCam 20 cm data set may be due to an inaccuracy of the bundle adjustment. This increases the RMS error of that data set quite much. In contrast, the standard deviation of this data set is just about 8.1 cm, which is quite similar to that of the other 20 cm data sets.

In general, the RMS error is around half of the GSD for all data sets, which is much less than $1/4^{th}$ of a pixel in disparity space (depending on the actual view angles). There is not much more that can be expected from an automatic image processing method.

Table 2: Statistical evaluation on 110 selected points.

Camera	GSD [cm/pixel]	RMS [cm]	Mean [cm]	Min [cm]	Max [cm]
ALS	25	8.9	1.8	-7.1	56.6
UltraCam-X	20	12.5	-9.6	-23.7	14
DMC	20	9.7	1.7	-20.8	28.4
Quattro DigiCAM	20	6.2	0.3	-22.9	13.2
UltraCam-X	8	5.1	-2.6	-10.4	5.9
DMC	8	3.2	-0.1	-7.2	6.4
Quattro DigiCAM	8	2.6	-0.3	-7.3	5.4

3.6 Computation Time

All data sets where computed on a Blade system that features 32 Intel Xeon 5570 Quadcore CPU's using a frequency of 3 GHz. Table 3 shows the computation time for stereo matching, DOM creation and true ortho-image generation.

Table 3: Computation time.

Camera	GSD [cm/pixel]	Number of images	Image Resolution [MPixel]	Matching on 128 CPU cores [hours]	DSM creation on one CPU core [hours]	Ortho image on one CPU core [hours]
UltraCam-X	20	52	136	1.9	0.9	1.2
DMC	20	60	106	1.9	0.7	1.3
Quattro DigiCAM	20	188 / 4	4 x 39	21.0	0.7	1.9
UltraCam-X	8	215	136	15.0	3.3	7.4
DMC	8	136	106	7.7	1.9	3.7
Quattro DigiCAM	8	784 / 4	4 x 39	224.0	3.6	10.8

The computation time for matching is roughly linear to the number of pixels and the relative depth range of the scene, while DSM and ortho-image creation depend mostly on the amount of data. The very large computation time of the Quattro DigiCAM is due to the special geometry of the camera which does not look straight down. In contrast each of the four individual cameras is slightly tilted. This means that even if the camera looks onto flat ground, a part of the scene is closer in one image corner than in the opposite image corner. This special geometry increases the depth range artificially and is responsible for the much larger processing time.

One advantage of SGM is the regularity of its algorithm and the simplicity of the basic operations, which are in fact only comparisons and additions. This allows the implementation on special hardware like GPU (Ernst and Hirschmüller, 2008) or FPGA (Gehrig et al., 2009). Future work will exploit this for speeding up processing of aerial or satellite images in comparison to a purely CPU based implementation.

4 Conclusion

It has been shown that high quality DSMs, which are more accurate and have more detail than a laser DSM, can be created from all tested aerial cameras. The study showed that images should be captured with high overlap, like 80 % along flight strips and 70 % across strips. This was the case for the 8 cm UltraCam data set which has a very high quality. Unfortunately, none of the 20 cm data sets has such a high overlap. Although the quality of the 20 cm data sets is comparable to ALS, from our experience, it is much less than can be reached with 20 cm resolution. Low overlap may cause problems especially in forest or narrow canyons or streets due to large view angles. Furthermore, configurations where the camera does not look straight down leads to increased computation time, at least with our method, due to an increased depth range. Finally, the high radiometric depth that all aerial cameras have, should be used for matching. In this respect, we believe that the UltraCam results could be improved if 12 bit data were used.
Future work includes extending the study to pushbroom cameras like the ADS 40 and JAS-150 as well as satellite images.

5 References

CRAMER, M., 2010: The DGPF-Test on Digital Airborne Camera Evaluation – Overview and Test Design, in the Journal of Photogrammetry, Remote Sensing and Geoinformation Processing (PFG), 02/2010.

ERNST, I. & HIRSCHMÜLLER, H., 2008: Mutual Information based Semi-Global Stereo Matching on the GPU, in Proceedings of the International Symposium on Visual Computing (ISVC08), 1-3 December 2008, Las Vegas, Nevada, USA.

GEHRIG, S., EBERLI, F. & MAYER, T., 2009: A Real-Time Low-Power Stereo Vision Engine Using Semi-Global Matching, in Proceedings of the International Conference on Computer Vision Systems (ICVS), Liege, Belgium, LNCV Volume 5815, pp. 134-143.

HALLA, N., HASTEDT, H., WOLF, K., RESSL, C., BALTRUSCH, S., 2010: Digital Photogrammetric Camera Evaluation – Generation of Digital Elevation Models, in the Journal of Photogrammetry, Remote Sensing and Geoinformation Processing (PFG), 02/2010.

HIRSCHMÜLLER, H., SCHOLTEN, F. & HIRZINGER, G., 2005: Stereo Vision Based Reconstruction of Huge Urban Areas from an Airborne Pushbroom Camera (HRSC), in Proceedings of the 27th DAGM Symposium, 30 August - 2 September 2005, Vienna, Austria, LNCS Volume 3663, pp. 58-66.

HIRSCHMÜLLER, H., 2008: Stereo Processing by Semi-Global Matching and Mutual Information, in IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 30(2), February 2008, pp. 328-341.

HIRSCHMÜLLER, H. & SCHARSTEIN, D., 2009: Evaluation of Stereo Matching Costs on Images with Radiometric Differences, in IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 31(9), September 2009, pp. 1582-1599.

VIOLA, P. & WELLS, W. M., 1997: Alignment by Maximization of Mutual Information, in International Journal of Computer Vision, Volume 24(2), pp. 137-154.

ZABIH, R. & WOODFILL, J., 1994: Non-Parametric Local Transforms for Computing Visual Correspondence, in Proceedings of the European Conference of Computer Vision, Stockholm, Sweden, pp. 151-158.