ENHANCMENT OF DENSE URBAN DIGITAL SURFACE MODELS FROM VHR OPTICAL SATELLITE STEREO DATA BY PRE-SEGMENTATION AND OBJECT DETECTION

Thomas Krauß, Peter Reinartz

German Aerospace Center (DLR),
Remote Sensing Technology Institute
PO Box 1116, 82230 Wessling, Germany
thomas.krauss@dlr.de

Commission I, WG I/4

KEY WORDS: digital surface models, urban, segmentation, object detection, VHR, optical stereo imagery

ABSTRACT:

The generation of digital surface models (DSM) of urban areas from very high resolution (VHR) stereo satellite imagery requires advanced methods. In the classical approach of DSM generation from stereo satellite imagery, interest points are extracted and correlated between the stereo mates using an area based matching followed by a least-squares sub-pixel refinement step. After a region growing the 3D point list is triangulated to the resulting DSM. In urban areas this approach fails due to the size of the correlation window, which smoothes out the usual steep edges of buildings. Also missing correlations as for partly – in one or both of the images – occluded areas will simply be interpolated in the triangulation step. So an urban DSM generated with the classical approach results in a very smooth DSM with missing steep walls, narrow streets and courtyards. To overcome these problems algorithms from computer vision are introduced and adopted to satellite imagery. These algorithms do not work using local optimisation like the area-based matching but try to optimize a (semi-)global cost function. Analysis shows that dynamic programming approaches based on epipolar images like dynamic line warping or semiglobal matching yield the best results according to accuracy and processing time. These algorithms can also detect occlusions – areas not visible in one or both of the stereo images. Beside these also the time and memory consuming step of handling and triangulating large point lists can be omitted due to the direct operation on epipolar images and direct generation of a so called disparity image fitting exactly on the first of the stereo images. This disparity image – representing already a sort of a dense DSM – contains the distances measured in pixels in the epipolar direction (or a no-data value for a detected occlusion) for each pixel in the image. Despite the global optimization of the cost function many outliers, mismatches and erroneously detected occlusions remain, especially if only one stereo pair is available. To enhance these dense DSM – the disparity image – a pre-segmentation approach is presented in this paper. Since the disparity image is fitting exactly on the first of the two stereo partners (beforehand transformed to epipolar geometry) a direct correlation between image pixels and derived heights (the disparities) exist. This feature of the disparity image is exploited to integrate additional knowledge from the image into the DSM. This is done by segmenting the stereo image, transferring the segmentation information to the DSM and performing a statistical analysis on each of the created DSM segments. Based on this analysis and spectral information a coarse object detection and classification can be performed and in turn the DSM can be enhanced. After the description of the proposed method some results are shown and discussed.

1 INTRODUCTION

With the launches of very high resolution (VHR) satellites like GeoEye or WorldView I and II with ground sampling distances (GSD) of about 0.5 m for civil usage the availability of VHR images will increase in the near future. Until now already optical imagery from in-line-stereo scanners like Cartosat-1 or ALOS-Prism with GSD of about 2.5 m and from VHR satellites of the previous generation like Ikonos or QuickBird can already be used for the generation of high resolution digital surface models (DSM). In case of in-line-stereo scanners long stripes of stereo imagery can be acquired in a short time but in case of the satellites mentioned above only in the rather coarse resolution of about 2.5 m and only one spectral channel. Such optical stereo imagery allows the extraction of DSMs with a resolution of about 5 m. In such DSMs larger buildings like in industrial areas are already detectable. But for the generation of city models in a level of detail (LOD) 2 – this level corresponds to the representation of single buildings – these results are not sufficient.

Stereo imagery from VHR satellites like Ikonos, QuickBird or the new GeoEye and WorldView series are acquired using the high agility of the satellites. So in the same orbit two or more images of the same region of interest can be acquired by rotating the satellite during the acquisitions. In this case stereo images can only be acquired of relatively small areas of about 10 km × 10 km and since the satellite has to undergo special manoeuvres for a stereo pair in most cases is charged more than for two single images. But due to the very high ground sampling distance of 1 to 0.5 m DSMs in the ranges of 2 m up to a minimum of only 0.5 m can be derived using special advanced DSM generation and data fusion algorithms.

The generation of digital surface models from very high resolution optical stereo satellite imagery delivers either rather good DSM of relatively coarse resolution of about 1/3 to 1/10 of the original ground sampling distance (GSD) (Lehner et. al., 2008) or high resolution DSM with many mismatches and blunders (Xu et. al., 2008). To generate such high resolution DSMs in the order of the resolution of the GSD of the satellite new approaches have to be analysed. In the last years the classical DSM generation algorithms (Lehner and Gill, 1992) get more and more expanded by using methods first developed in computer vision based on epipolar imagery and dense stereo matching (Scharstein and Szeliski, 2002, Hirschmüller, 2005, Krauß et.al., 2005; d'Angelo et. al., 2008). The generated DSMs still suffer

from many blunders, so after the stereographic DSM extraction step approaches for blunder detection and DSM refinement are highly required and topic of actual research.

Most of these methods use only the DSM data and do outlier detection based for example on statistical approaches (Vincent, 1993). Other methods work on DSM segmentation and extraction of rectangular buildings (Arefi, 2009). Haala et al. (Haala et al., 1998) proposed a method reconstructing building rooftops using surface normals extracted from DSM data. They assumed that building boundaries are detected previously. But most of these methods work only with high resolution and reliable LIDAR DSMs which in general do not suffer from noise, outliers and smoothing effects like stereographic generated DSMs (Schickler et. al., 2001). Also the reconstruction of 3D building structures by hierarchical fitting of minimum boundary rectangles (MBR) and RANSAC based algorithms for line or surface reconstruction are quite common (Arefi et al., 2008). Up to now most approaches for DSM enhancements work by optimizing or modelling directly the DSM or depend on multi-photo approaches.

To overcome this dilemma we propose in the presented paper a method introducing some knowledge from the original image by fusing this information with the calculated DSM. This can be done exploiting some properties of the dense stereo methods used.

2 OVERVIEW OF THE METHOD

The dense stereo DSM generation methods originating from computer vision rely on epipolar imagery. Based on the two epipolar images of a stereo pair a disparity map fitting exactly on one of the epipolar images is generated. This disparity map contains the distances of the feature at this point in the image to the correlating feature in the stereo mate measured in pixels. So per definitionem the disparity map fits exactly on one of the epipolar stereo images – in our case without loss of generality assumed as the "left" image.

This disparity map has to be reprojected using the orbit and ephemeris data to an orthorectified DSM. This DSM is subject to all optimisations described above. In our approach shown in this paper we start the optimisation already one step before – on the disparity map. With our approach already the disparity image can be corrected – not the resulting DSM which will require a calculated ortho image for fusing DSM and image data. In this case the calculation of the ortho image will in turn need a DSM, so the ortho image is already distorted due to the DSM errors.

The disparity map is generated using a hybrid approach fusing the digital line warping after (Krauß et. al., 2005) and the semi global matching introduced by (Hirschmüller, 2005). The fused algorithm work as its precursors only on one stereo image pair (two images, no multiple image stereo) in epipolar geometry. This dense stereo method allows the automatic detection of occlusions directly in the matching approach as described in (Krauß et. al., 2009). The result is a rather perforated disparity map since any collision or mismatch detected in the disparity map is assumed as occlusion and filtered out.

To enhance the disparity map we propose a segmentation of the left stereo image using the algorithm proposed by (Maire, 2010). For each detected segment a plane is interpolated from all non-occlusions in the original disparity map *D* covered by the segment. This generates the filled disparity image *S*. In a second track the disparity map is filled using iterative median filling which preserves the disparities and fills the occlusions with the median of the detected neighbours to disparity map *F*.

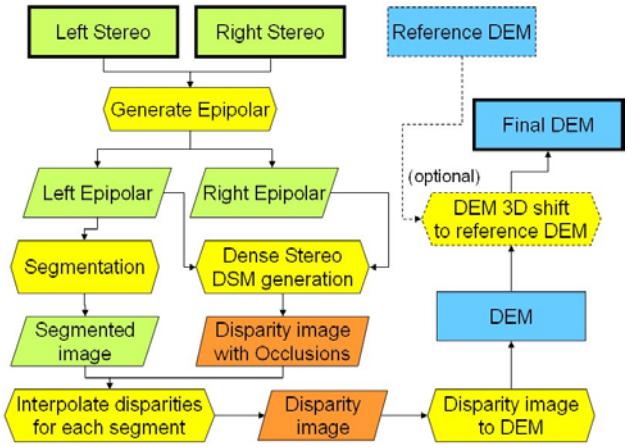

Figure 1.Overview of the process of DEM generation and optimization using image segmention

Applying a DSM-to-DTM algorithm to the segment based filled disparity map *S* results in the ground disparity map *G*. Subtracting *G* from *S* and applying a threshold in disparity pixels corresponding to a height of about 5 meters gives a high objects mask in disparity units. Applying this mask to the interpolated disparity map *F* introduces the sharp building edges derived from the segmentation image into the disparity map.

After this process some more optimisations are possible exploiting the exact correlation of left stereo image and disparity map before ortho projection of the DEM:

- in addition to the pansharpened stereo images satellites like Ikonos, QuickBird or WorldView II provide near infrared channel and the extraction of the normalized digital vegetation index NDVI and in turn a vegetation and water mask is possible
- since the generation of correct disparities in general do not work very well on water bodies these areas can be filled already in the disparity map using the lowest boundary height
- fusing all results together and reprojecting gives the result as an orthographic DSM.

3 DETAILLED PROCESS AND EXPERIMENTAL RESULTS

3.1 Data and preprocessing

For the evaluation of the method an Ikonos stereo pair acquired 2005-07-15 at 10:28 GMT over the city of Munich is used. It provides a ground resolution of 83 cm for the pan channel with viewing angles +9.25° and –4.45° as a level 1A image only corrected for sensor orientation and radiometry. Figure 2 shows the selected sections from the original images covering the city centre of Munich. The Ikonos stereo pair was acquired in forward (left image) and reverse (right image) mode due to the ordered small stereo angles – the standard stereo acquisition geometry for Ikonos uses ±30°. Therefore the first scan line of the left image (top line) is the northernmost line since the satellite travels from

north to south. In the reverse imaging mode the first scan line is southernmost and scanning goes "reverse" of the flying path from south to north. So the topmost line in the right stereo image is the southernmost line.

Figure 2. Section 2000 m × 2000 m from the Munich scene showing the center of the city, left and right stereo image

Fusing the pan channel of the left stereo image with the also supplied multispectral channels a (left) pan sharpened image (Figure 3) together with a NDVI channel (normalized digital vegetation index) is generated. In the NDVI channel values above zero (lighter as the middle gray value) show vivid vegetation whereas darker values represent water bodies (right bottom area of the image with the river Isar).

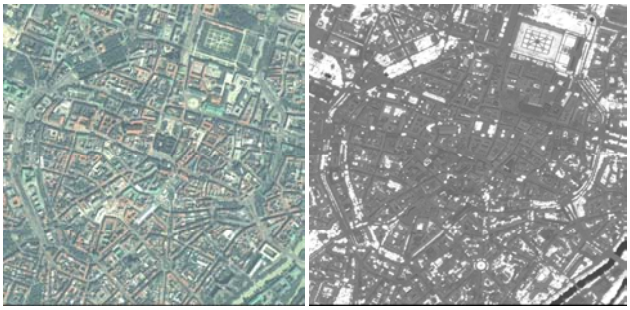

Figure 3. Section 2000 m × 2000 m from the Munich scene showing the center of the city, pan sharpened image (left) and NDVI (right)

After this as a first preprocessing step an epipolar reprojection following (Morgan, 2004) is performed for the left and right pan and the left pansharpened and NDVI images producing Figure 4.

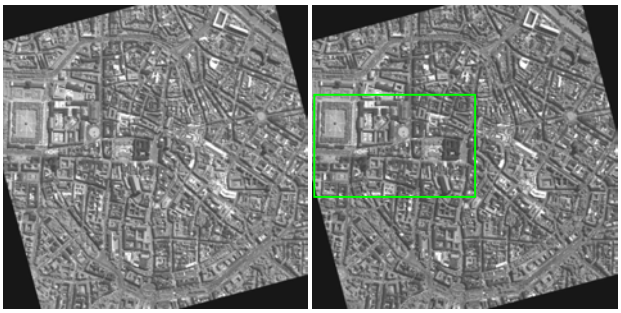

Figure 4. Section of the Munich scene reprojected to epipolar geometry – all subsequent detail sections are the area marked covering the residence gardens, the residences down to the city center with town hall and St. Marys church in the center of the images above (1030 m × 630 m)

The two images of the stereo pair have to be co-registered to an absolute reference or – if not available – relative to each other. This is done by matching all images (stereo mate of the pan image and all multispectral images) to one reference (without loss of generality this is in our case the "left pan" stereo image). The co-registration of the images delivers an affine correction to the absolute reference or relative between the images. In our case the correction is only done relative and shows up that the delivered rational polynomial coefficients (RPCs, see Jacobsen et al., 2005, Grodecki et al., 2004) with both stereo pairs fit already perfectly with less than 10 cm difference relative to each other. Absolute correction (for quality check in the case of the Munich scene) shows nevertheless a simple shift of the scene of about 4.9 m to the east and 8.4 m to the south with respect to an available reference Laser DEM.

3.2 DSM generation

The DSM generation approach combining two computer vision approaches as described in (Krauß et. al., 2009) is used for calculating the disparity map. Since this approach already performs an occlusion detection the resulting disparity maps are quite porous.

Figure 5. Disparity map with occlusions (black), section 1030 m × 630 m from the previously introduced Munich scene

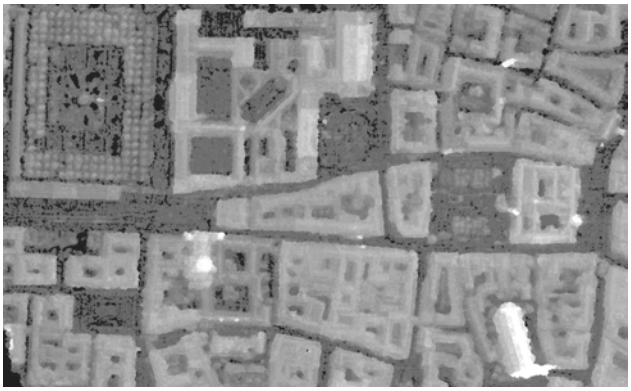

Figure 6. Filled disparity map, section 1030 m × 630 m

For the generation of this disparity map first one map is calculated using a cost function defined on the Birchfield-Tomasi distance (Birchfield and Tomasi, 1998) including the minimum of the absolute difference of the first pixel to the second pixel and within half the distance to it's neighbours. A second map is calculated using a cost function based on the sum of absolute values of gray value distances on a 3 × 3 pixels window. These two disparity

maps look quite the same but show different behaviour in instable areas. So the joined disparity map shown in Figure 5 shows additionally occlusions for deviations in disparity of one pixel or more between the two calculated disparity maps. The disparity map in Figure 6 is generated by filling the occlusions with an iterative median filling using median radiuses of size one.

3.3 Segmentation

As a further pre-processing step towoards the segmentation the pan sharpened image is smoothed using a bilateral filtering after (Tomasi and Manduci, 1998) with the standard parameters as given in (Sirmacek and Unsalan, 2009): a filter size of 5, a distance parameter of 3 and a gray value parameter of 0.1. The bilateral filter is used for filtering out small features but in parallel preserving sharp edges of the objects. The results are shown in Figure 7.

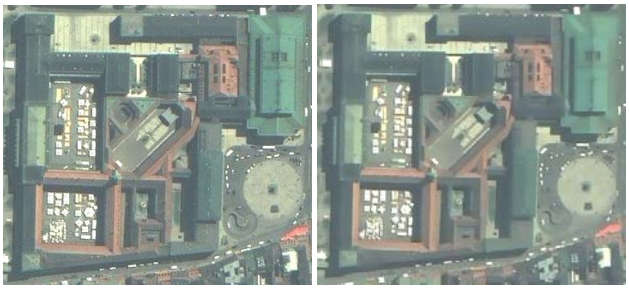

Figure 7. Section of Munich scene (residence area) reprojected to epipolar geometry, left pan sharpened image, right after bilateral filtering

Figure 8 shows the segmented image following the algorithm described in (Maire, 2010). The segmentation is calculated using a fixed parameter λ of 100 000 instead of providing a maximum number of segments. Subject to the segmentation process is a combined image consisting of the left pan sharpened image consisting of a blue, green, red and near infrared channel together with the NDVI map scaled to the same gray value range as the other channels.

Figure 8. Segmented image

3.4 Applying segmentation to disparity map

For the enhancement of the disparity map the segments from the segmented left stereo image (Figure 8) are applied to the disparity map (with occlusions) shown in Figure 5. For each segment an average height of all disparities – ignoring occlusions – is calculated and filled in as the segments height as shown in Figure 9.

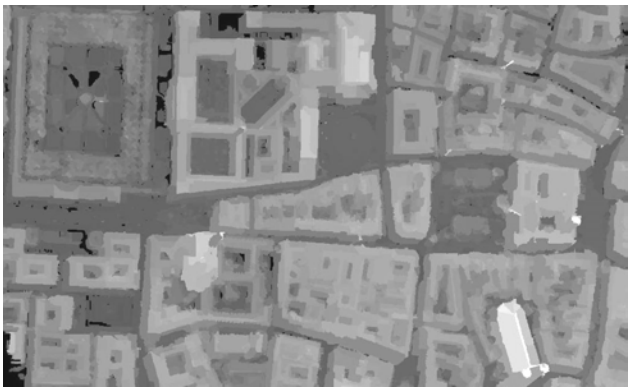

Figure 9. Segmentated disparity map

The segmented disparity map shows now the straight edges derived from the pan sharpened and segmented image together with the heights derived from the calculated disparity map. So the straightened borders of the buildings can be used for detecting the outlines of the buildings and for the 3D reconstruction process.

3.5 Object detection

To enhance the disparity map further an object detection step is introduced. This is done by first deriving a type of digital terrain model (DTM) – referenced here as "ground disparity map" – from the disparity map. This is done by filtering the disparity map with different size median filters and detecting "low" regions by subtracting the medians as shown in diagram Figure 10.

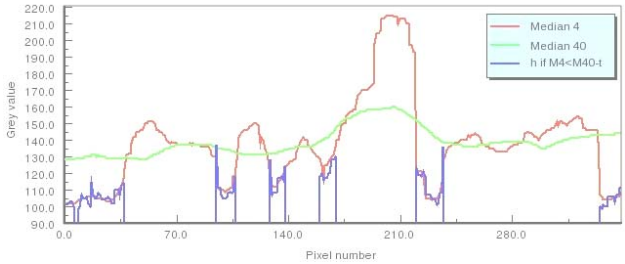

Figure 10. Typical profile showing the calculated medians and the detected street level areas (blue)

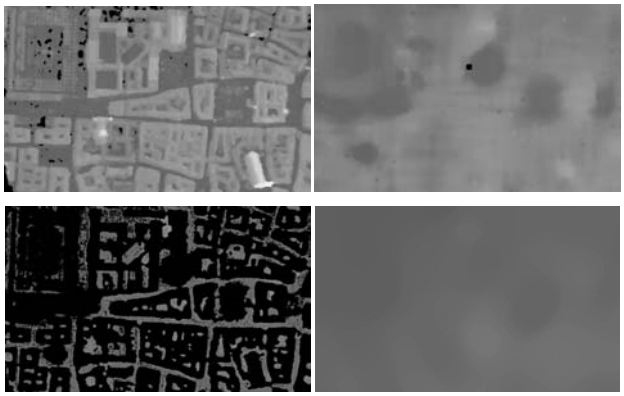

Figure 11.Small (radius 4) median, large (radius 40) median, detected "low" areas (typically streets or courtyards) and finally filled ground disparity map

The remaining "low" regions get eroded by a small amount to eliminate border effects and subsequently filled to the ground disparity map as shown in the bottom right image of Figure 11.

Subtracting the ground disparity map from segmented disparity map and applying a threshold gives a mask of elevated objects. The threshold in disparity pixels can be calculated directly from a height threshold using the provided RPCs. In the given case a threshold of one pixel disparity corresponds to about 5 meters in height.

Figure 12. Derived building masks. Top row based on segmented disparity map, bottom row: filled original disparity map, left top/bottom: calculated map, right top/bottom: after closing and opening operations

As a second object detection step a water mask can be extracted based on the NDVI as shown in Figure 13.

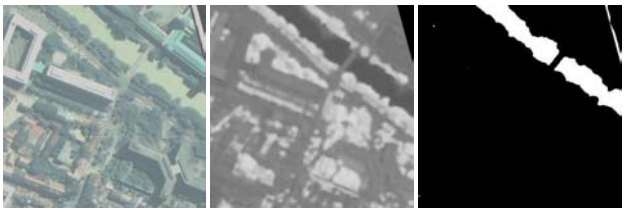

Figure 13. Image section 400 m × 400 m showing the river Isar, the "Deutsches Museum" (above the river) and the European patent office (below), left: image, center: NDVI, right: water mask

Since water areas can not be matched properly and deliver only noisy disparities the filled disparity map (Figure 14, left) gets masked by the water mask and filled with the lowest value of a small border area (Figure 14, right).

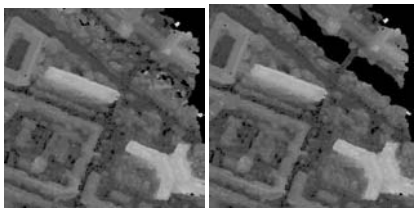

Figure 14. Disparity map before (left) and after (right) the filling of water bodies

As can be seen in Figure 12, left, in the area of the residence gardens (top left area) trees get completely messed up by the segmentation like every dome shaped area since there will be a darker and a brighter half of the dome. So the whole vegetation areas have to be filled to the resulting DSM from the original disparity map using the NDVI vegetation mask. The resulting disparity map consists so the following parts derived from the object detection steps:

- water: filled with lowest neighbours based on the filled disparity map
- vegetation: from the original disparity map
- building areas: from the original disparity map
- non-building, non-water and non-vegetation areas: filled from the ground disparity map

The resulting disparity map gets transformed to an orthographic DSM and filled again using the previous introduced iterative masked median fill. The resulting DSM in UTM zone 32, 1 m resolution of the section shown in Figure 5 through Figure 12 is shown in Figure 15.

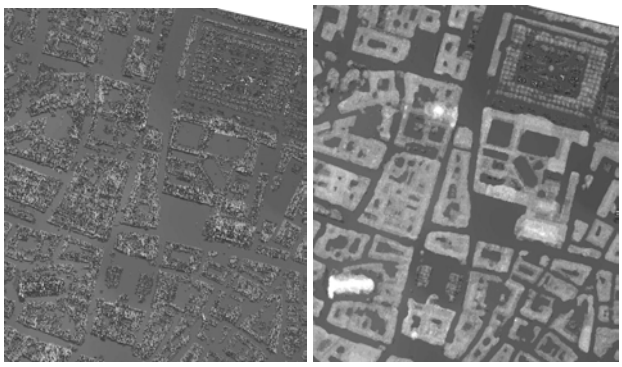

Figure 15. Resulting DSM generated from fusioned disparity maps (left) and filled (right)

4 CONCLUSION AND OUTLOOK

In this paper an approach for DSM enhancement from VHR satellite imagery is presented. The approach is based on fusing an intermediate disparity map generated by the dense stereo matching which fits geometrically onto one of the input images with the image content of this image. For this the image gets classified (vegetation and water) and segmented. Using these segments buildings are detected and a fused disparity map is generated handling objects like buildings, vegetation, water or streets differently. The segmentation shows some problems since too small segments simply rebuild the original input while too large segments fuse together different objects like roofs, walls and shadow areas on the street. Also dome shaped objects – as well as trees – get messed up since the dark side of the dome joins together with shadow areas. So the approach shows some interesting potential but the segmentation has to be improved in future researches for extraction of main directions and objects to extract 3D city models.

REFERENCES

d'Angelo, Pablo, Lehner, Manfred, Krauß, Thomas, Hoja, Danielle and Reinartz, Peter (2008) Towards Automated DEM Generation from High Resolution Stereo Satellite Images. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, International Society for Photogrammetry and Remote Sensing , XXXVII (B4) , pp. 1137-1342. ISPRS Conference 2008 , 2008-07-03 - 2008-07-11 , Peking (China). ISSN 1682-1750

Arefi, H., Engels, J., Hahn, M. and Mayer, H., 2008. Levels of detail in 3d building reconstruction from lidar data. In Proceedings of the International Archieves if the Photogrammetry, Remote Sensing, and Spatial Information Sciences 37, pp. 485–490.

Arefi, H., d'Angelo, P., Mayer, H. and Reinartz, P., 2009. Automatic generation of digital terrain models from Cartosat-1 stereo images. In Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.

Birchfield, S. and C. Tomasi, C., 1998. A Pixel Dissimilarity Measure That is Insensitive To Image Sampling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(4):401-406, April 1998.

Grodecki, J., Dial, G. and Lutes, J., 2004. Mathematical model for 3D feature extraction from multiple satellite images described by RPCs. In: ASPRS Annual Conference Proceedings, Denver, Colorado.

Haala, N., Brenner, C. and Anders, K., 1998. 3D urban GIS from laser altimeter and 2D map data. In Proceedings of International Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences 32, pp. 339–346.

Hirschmüller, H., 2005. Accurate and efficient stereo processing by semi-global matching and mutual information. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Jacobsen, K., Büyüksalih, G. and Topan, H., 2005. Geometric models for the orientation of high resolution optical satellite sensors. In: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. 36 (1/W3). ISPRS Workshop, Hannover.

Krauß, T., Reinartz, P., Lehner, M., Schroeder, M. and Stilla, U., 2005. DEM generation from very high resolution stereo satellite data in urban areas using dynamic programming. In: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. 36 (1/W3). ISPRS Workshop, Hannover.

Krauß, Thomas and Reinartz, Peter, 2009. Refinement of urban digital elevation models from very high resolution stereo satellite images. ISPRS . IPI-Workshop , 02-05 Jun 2009 , Hannover. ISSN 1682-1750

Lehner, M. and Gill, R., 1992. Semi-automatic derivation of digital elevation models from stereoscopic 3-line scanner data. ISPRS, 29 (B4), pp. 68–75.

Lehner, M. and d'Angelo, P. and Müller, R. and Reinartz, P, 2008. Stereo Evaluation of CARTOSAT-1 Data Summary of DLR Results During CARTOSAT-1 Scientific Assessment Program., ISPRS J., 29 (3), p.1295 ff, 21. ISPRS Congress, July 2008, Beijing

Maire, C., 2010. Image Information Extraction and Modeling for the Enhancement of Digital Elevation Models. Phd Thesis, Karlsruhe Institute of Technology (KIT), 09.02.2010.

Morgan, M. F., 2004. Epipolar resampling of linear array scanner scenes. Phd Thesis, Geomatics Engineering, University of Calgary, http://hdl.handle.net/1880/41819

Scharstein, D. and Szeliski, R. 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision, 47(1/2/3):7-42, April-June 2002. Microsoft Research Technical Report MSR-TR-2001-81, November 2001. Middlebury stereo vision page. http://vision.middlebury.edu/stereo. (accessed 04/2010).

Schickler, W. and Thorpe, A., 2001. Surface estimation based on LIDAR. ASPRS Conference, St. Louis, Missouri, USA, April 2001

Sirmacek, B. and Unsalan, C., 2009. Urban-area and building detection using sift keypoints and graph theory. IEEE Transactions on Geoscience and Remote Sensing 47 (4), pp. 1156–1167.

Tomasi, C. and Manduci, R., 1998. Bilateral filtering for gray and color images. In Proceedings of International Conference on Computer Vision 1, pp. 839–846.

Vincent, L., 1993. Morphological grayscale reconstruction in image analysis: applications and efficient algorithms. IEEE Trans Image Process. 1993;2(2):176-201.

Xu, F. and Woodhouse, N. and Xu, Z. and Marr, D. and Yang, X. and Wang, Y., 2008. Blunder elimination techniques in adaptive automatic terrain extraction. ISPRS J., 29 (3), 21. p.1139 ff, ISPRS Congress, July 2008, Beijing