URBAN OBJECT DETECTION USING A FUSION APPROACH OF DENSE URBAN DIGITAL SURFACE MODELS AND VHR OPTICAL SATELLITE STEREO DATA

Thomas Krauß, Peter Reinartz

German Aerospace Center (DLR), Remote Sensing Technology Institute, D-82234 Weßling, Germany
E-Mail: Thomas.Krauss@dlr.de

Commission I/4

KEY WORDS: object detection, fusion, digital surface models, digital terrain models, urban areas, VHR, optical stereo imagery

ABSTRACT:

In this paper we describe a new approach for the extraction of urban objects from very high resolution (VHR) optical stereo satellite imagery. Such data is delivered from sensors like Ikonos, QuickBird, GeoEye or WorldView-II. These sensors provide ground sampling distances (GSD) of 0.5 to 1 m for the pan chromatic channel and 2 to 4 m for the multispectral channels. Normally good digital surface models (DSM) can only be expected at 1/3 to 1/5 of the original GSD. But we present a new approach which uses the generation of dense disparity maps based on computer vision approches and fuse these disparity maps with additional information gained from the original imagery to allow the extraction and afterwards modeling of urban objects. This can be achieved due to the fact that the generated disparity maps are constructed on one of the original images. So a direct pixel to pixel correlation of the height (represented by the disparity) and the spectral information (represented by the pan sharpened original image) can be done. Applying methods for the generation of a digital terrain model (DTM) which represents the ground without elevated objects and spectral classification allows the separation of typical urban classes like buildings, trees, roads, low vegetation, water and so on. These classes will be treated individually in the modeling step to generate a simplified 3D model of the observed urban area. The results are presented, compared to the original imagery and discussed.

1 INTRODUCTION

Introducing dense stereo methods to the generation of digital surface models (DSM) from very high resolution (VHR) optical stereo satellite data gives many new chances for a fully automatic generation of urban 3D city models. One characteristic of the dense stereo methods is the generation of a so called disparity map. Such a disparity map is constructed fitting exactly on one of the original images and gives the distance of each pixel of this image to it's companion in the other stereo image measured in pixels (px) in epipolar direction.

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. Stereo imagery from these satellites 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 the providers charge 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.

In this paper we present a method for extracting typical urban objects based on the generation of dense stereo DSM – more correct the disparity map – and fusion with multispectral information from the pan sharpened original imagery.

The generation of digital surface models from very high resolution optical stereo satellite imagery delivers either a rather good DSM of relative coarse resolution of about 1/3 to 1/10 of the original ground sampling distance (Lehner et al., 2007) or a 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 replaced 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 et al., 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 and Thorpe, 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 using the true ortho imagery or depend on multi-photo approaches, which are mostly not available in the case of satellite imagery.

But for the generation of a correct true ortho image which is needed for these methods an absolutely correct DSM with the same GSD as the imagery is needed. Since such a DSM suffer from many outliers and mismatches also the true ortho image shows these errors and inhibit the correct work of the algorithms mentioned above. So all subsequent work starting with a DSM and a (wrong) true ortho image will show the errors from the DSM generation step.

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. So the object detection and DSM correction start already one step before – on the disparity map and the unchanged original imagery.

2 METHOD

Our proposed method uses the following steps:

- Preprocess the images
  - Transform the original images to epipolar geometry
  - Generate the pan sharpened left stereo image from pan and multispectral channels
  - Calculate the normalized difference vegetation index (NDVI) for this pan sharpened image
- Generate a disparity map fitting on the left original stereo image using a dense stereo algorithm
- Extract the digital terrain model in disparities from the disparity map
- Classify to
  - Vegetation ($NDVI > t_v$)
  - Water ($NDVI < t_w$)
  - High ($(DSM - DTM) \geq t_h$)
  - Low ($(DSM - DTM) < t_h$)
- Extract single objects:
  - Water bodies (all of extracted water class)
  - Trees (vegetation and high)
  - Buildings (non-vegetation and high)
  - Ground vegetation (vegetation and not high)
  - Ground (non-vegetation and not high)
- Model extracted objects
  - Water: set disparity maps to lowest ground value in surroundings [blue]
  - Trees: separate using watershed tranformation for finding tree tops and diameters [dark green]
  - Buildings: extract simplified bounding polygon with one height [gray]
  - Roofs: if building contains slanted roof areas separate these to single planes [red] (future work)
  - Ground: replace with DTM [light gray]
  - Ground vegetation: replace with DTM [light green]
- Create 3D model (VRML format: Virtual Reality Modeling Language; future work)

In this paper the last step – the transformation to the 3D objects is not included. Only the detection based on the fusion of disparity map and one of the original satellite images is shown.

3 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 $+9.25^\circ$ and -4.45° as a level 1A image only corrected for sensor orientation and radiometry. Fig. 1 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 $\pm 30^\circ$. 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.

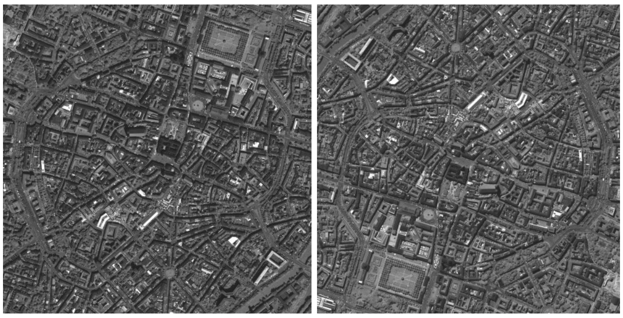

Figure 1: Section 2000 m × 2000 m from the Munich scene showing the center of the city, original left and right stereo image

In the preprocessing step the left and right PAN stereo images are transformed to epipolar images using the provided rational polynomial coefficients (RPCs, (Jacobsen et al., 2005), (Grodecki et al., 2004)) of the input images. The RPCs define functions for samples (x-coordinate in the original image) and lines (y-coordinate in the original image) depending on longitude (λ), latitude (φ) and the ellipsoidal height h over the WGS84 ellipsoid as

$$x = \frac{f_{samp,num}(\lambda, \varphi, h)}{f_{samp,den}(\lambda, \varphi, h)} \qquad \text{and} \qquad (1)$$

$$y = \frac{f_{line,num}(\lambda, \varphi, h)}{f_{line,den}(\lambda, \varphi, h)}. \qquad (2)$$

Each of the functions $f()$ are polynomials of third order in λ, φ and h with 20 coefficients. So over all a RPC is defined by 80 coefficients and 10 scaling and offset parameters for x, y, λ, φ and h.

Calculating geographic coordinates for a pixel in the left image (done by iterative solution of equations (1) and (2)) for different heights h and in turn the x and y coordinates of these in the second image and vice versa give the directions of the epipolar line at this point. Repeating this for all points gives the epipolar lines for the whole image. Satellite geometry is despite of the line scanner very stable so it’s sufficient to do the calculation only for a sparse grid of points (Morgan, 2004).

The same transformation is applied to the multispectral stereo images. Afterwards the PAN and multispectral epipolar images are fusioned to a pan sharpened image. Without loss of generality we construct our disparitiy map on the left image of the stereo pair. So for all following calculations only the left pan sharpened epipolar image P is used.

Using P we can calculate the normalized diffentional vegetation index (NDVI) image N from the NIR and red channel defined as

$$NDVI = \frac{NIR - red}{NIR + red}.$$

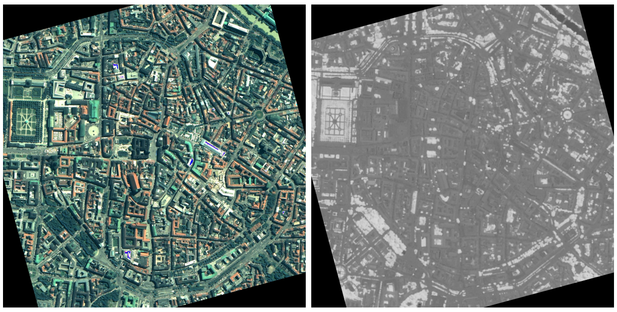

Figure 2: Same section as above – only left stereo image – transformed to epipolar geometry. Left: pan sharpened epipolar image P, RGB channels, right: NDVI image N

In the NDVI image values above zero (lighter than the middle gray value) show vivid vegetation whereas darker values represent water bodies (right top area of the image with the river Isar).

4 GENERATION OF THE DISPARITY MAP

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. In all approaches so far optimizations and object extraction was done using this orthorectified DSM. But since errors originating from the DSM generation process – in our case the dense stereo matching – lead to a wrong orthorectification of the original images we propose in this paper an other way. In our approach we start the optimisation already one step before – on the disparity map. With our approach already the disparity image can be corrected.

The disparity map S 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 and Reinartz, 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.

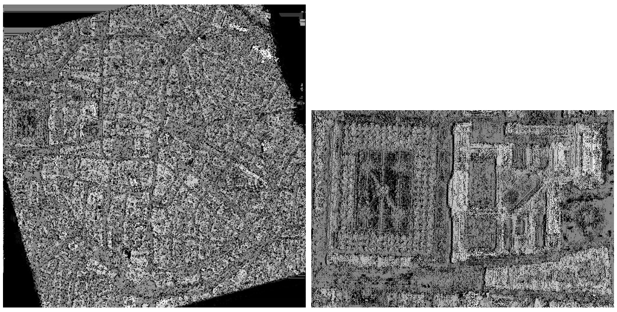

Figure 3: Left: disparity map S fitting on left stereo image, right: detail showing the detected occlusions in black

5 GENERATION OF THE GROUND DISPARITY MAP

In this step a type of digital terrain model (DTM) – referenced here as "ground disparity map" T – is extracted from the disparity map S. This is done by filtering the disparity map with different size median filters to detect "low" regions by subtracting the medians as shown in diagram 4.

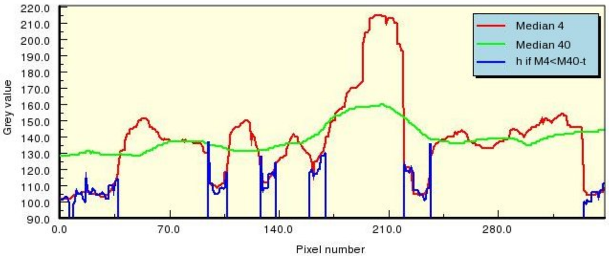

Figure 4: Typical profile showing the calculated medians and the detected street level areas (blue)

Applying two median filters with different sizes show different behaviour at steep edges. The large sized median fills up small holes where the small median filter follows the height structure more strictly. Subtracting the medians and applying a threshold marks areas at the bottom of steep walls.

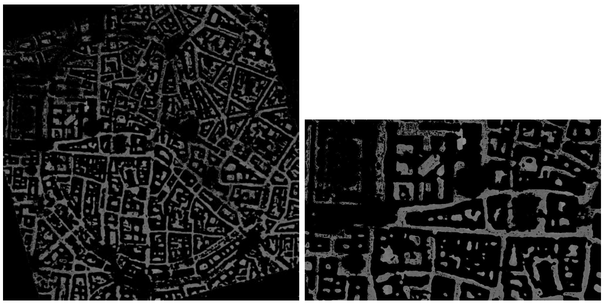

Figure 5: Left: extracted street mask, right: detail

Only taking these "streets" as shown in fig. 5 and filling these deliver the ground disparity map T as shown in fig. 6.

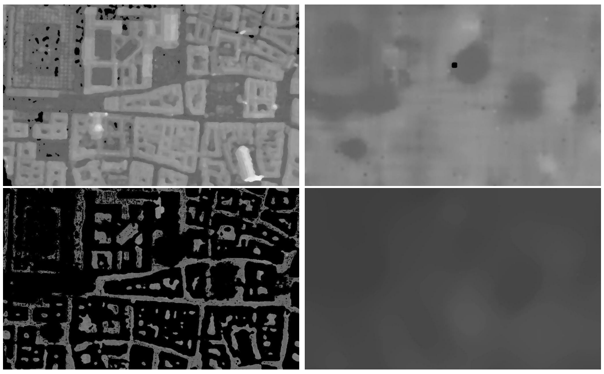

Figure 6: Small (radius 4) median, large (radius 40) median, detected "low" areas (typically streets or courtyards) and filled ground disparity map T

6 CLASSIFICATION

Based on the NDVI image N and the disparity and ground disparity images S and T a coarse classification can be carried out. First from the NDVI image as shown in fig. 7 two masks are computed: the vegetation mask emerging from values $NDVI > 0.3$ and the water mask representing NDVI values $NDVI < -0.15$

(these threshold values where extracted experimentally by iteratively adapting the results to an existing ground truth).

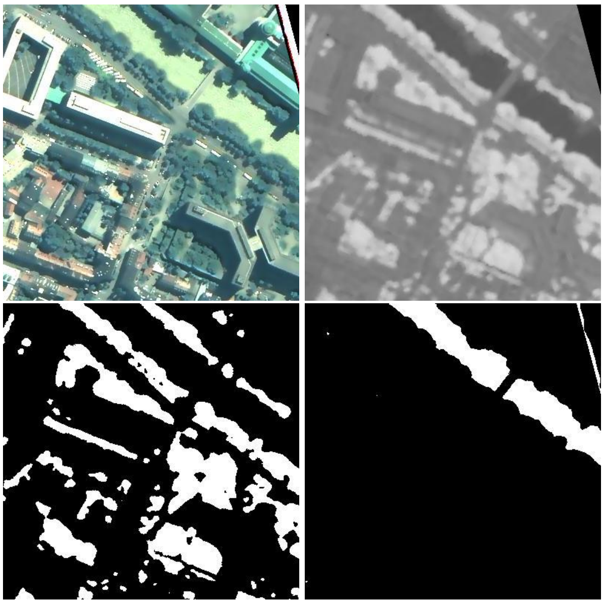

Figure 7: RGB image showing top right section with river Isar, NDVI image N of this part, vegetation and water masks

Subtracting the ground disparity map from the disparity map gives the so called normalized DEM – better called a normalized disparity map (nDEM) in our case. Applying a threshold to this nDEM 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 (but the disparity values in the shown disparity maps are all in units of 0.1 pixels). Additionally for the object extraction as described in the next chapter two different height thresholds are used for the extraction of trees and buildings (see next chapter).

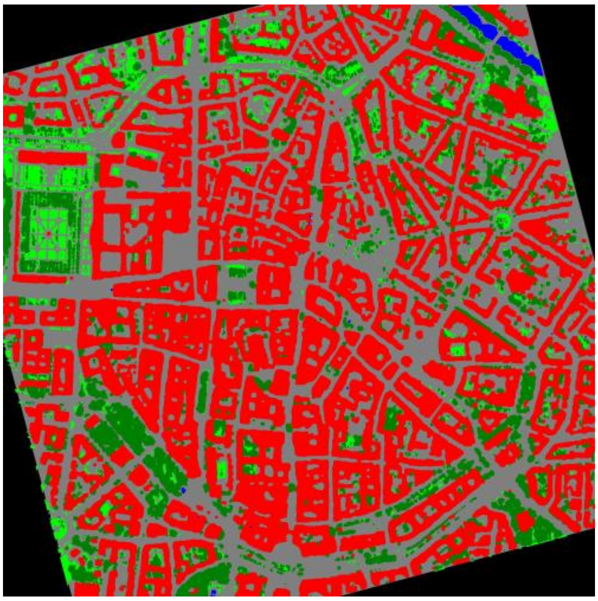

Figure 8: Mask image showing buildings (red), trees (dark green), low vegetation (light green), water (blue) and low non vegetation (gray)

7 OBJECT DETECTION AND MODELING

7.1 Water

The water objects are detected by applying an morphological opening with radius 2 to the extracted water mask to exclude the small erroneous detected "water areas" inside shadows. Since the generation of correct disparities in general do not work very well on water bodies these areas can be filled already now in the disparity map using the lowest boundary height of each water area.

7.2 Trees

The detection of trees is done by extracting only all vegetation areas from the nDEM and applying a morphological dilation with radius 1 to these areas to fill small holes emerging from the occlusion detection. Afterwards a watershed transformation is applied to segment this vegetation nDEM to detached objects with a minimum height of 2.5 m which correspond to 0.5 px disparity in the disparity nDEM. The watershed transformation also delivers a list of all objects with height, center and bounding boxes which can directly be transformed to a vector dataset containing circles at the tree positions with radii corresponding to the area covered by the detected watershed object. See for the resulting tree outlines fig. 9 and fig. 10 (separation of trees in courtyards from buildings).

7.3 Buildings

The buildings get extracted also from the normalized disparity map (nDEM). But in this case a minimum height of a building is assumed as 5 m which in turn compute to 1 px of disparity. To get a better subpixel sampling for the building outlines no simple thresholding of the nDEM is done but the contour line for 5 m respectively 1 px of disparity is extracted from the nDEM. The resulting vector objects are shown also in fig. 9 and fig. 10.

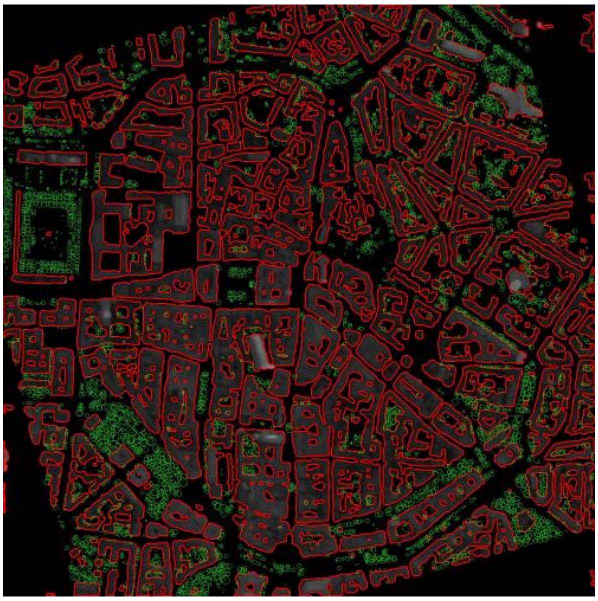

Figure 9: Normalized disparity map with layed over building outlines (red) and detected trees (green)

To model the buildings an approach shown in (Krauß et al., 2007) will be applied. But the outline detection is replaced by the contour line extration described above and added hierarchical splitting of the outlines for included objects like court yards, building in court yards, yards in these buildings and so on. So each building is represented by a vector outline winding around contained yards.

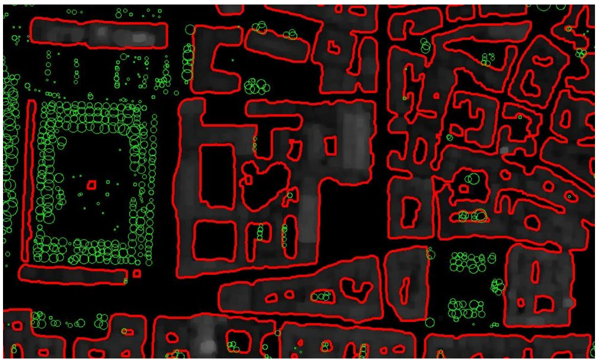

Figure 10: Detail of fig. 9, normalized disparity map with layed over building outlines (red) and detected trees (green)

For the modeling the average height of each building described by such an outline is calculated from the disparity map (not the nDEM!). The 3D object will be a polygon with this height. For additional works which have to be done see chapter Outlook.

7.4 Other objects

The remaining objects contain only low vegetation and non-vegetation areas. These are represented in the object oriented model simply by coloring the ground – the same as for the detected water areas.

8 RESULTS

First preliminary results of the proposed method shown in this paper are the extracted objects "tree" and "building" in vector format for direct conversion to a 3D world model and the ground stuck objects "water", "low vegetation and non-vegetation areas" which get included in the 3D model simply as color coding of the also extracted ground disparity map.

All calculated objects are fitting exactly on one of the two stereo images – in our case without loss of generality the left one. Also the calculated and corrected disparity map together with the derived ground disparity map and normalized disparity map are constructed so that they also fit exactly on the original left stereo image.

After the object extraction step a correction of the disparity map and also the ground disparity map can be conducted by applying existing outlier removal methods – but now these get applied to the disparity maps and not the final DSM. After the correction a better true ortho image can be calculated from the original (pan sharpened) satellite image and the corrected disparity map. The extracted vector objects will be transformed using the same ortho transformation. Applying the vector objects to the corrected DSM will in turn generate the 3D model of the scene.

9 DISCUSSION

The resulting object detection as shown in figs. 8, 9 and 10 was compared manually to the underlying original satellite imagery and a generally good correlation was found. The trees are in general detected very well. The applied water shed transformation shows some problems if the disparities between distinct trees show no saddle points. In this case connected trees are detected as a single tree. This can – if required – be solved in many cases by analyzing the shape of the resulting area and splitting up if much longer than wide.

Due to errors on the boundaries of buildings in the disparity map the extracted building outline polygons have to be rectified using an approach as shown in (Krauß et al., 2007) which was not done in this paper. This will be done in future investigations. But in general the extracted vector outlines show a very close fit of the buildings – much better than also conducted investigations based on a thresholded height mask.

10 CONCLUSION AND OUTLOOK

The approach for fusing the original imagery and derived disparity maps to extract urban objects show many advantages in comparison to an object extraction based on a DSM and a true ortho image generated with this outliers loaded DSM. Also an individual correction of the errors in the disparity map can be done using the detected objects.

Future work needs to be done in the rectification of the building outlines. This contains the correct splitting to rectangular or polygonal individual building parts surrounding the contained yards. Also in this step the detection and separate extraction of high object parts like towers or domes has to be included.

Including these steps and conducting the proposed individual outlier removal per detected urban object will result in much better urban 3D models from very high resolution satellite stereo imagery.

REFERENCES

Arefi, H., d'Angelo, P., Mayer, H. and Reinartz, P., 2009. Automatic generation of digital terrain models from cartosat-1 stereo images.

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

d'Angelo, P., Lehner, M., Krauß, T., Hoja, D. and Reinartz, P., 2008. Towards automated dem generation from high resolution stereo satellite images. Vol. 37, pp. 1137–1142.

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. 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: ISPRS Workshop, Hannover, Vol. 36 (1/W3).

Krauß, T. and Reinartz, P., 2009. Refinement of urban digital elevation models from very high resolution stereo satellite images. Vol. 38.

Krauß, T., Reinartz, P. and Stilla, U., 2007. Extracting orthogonal building objects in urban areas from high resolution stereo satellite image pairs.

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: ISPRS Workshop, Hannover, Vol. 36 (1/W3).

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., Müller, R. and Reinartz, P., 2007. Stereo evaluation of cartosat-1 data for various test sites.

Morgan, M., 2004. Epipolar Resampling of Linear Array Scanner Scenes. PhD thesis. UCGE Reports Nr. 20193.

Scharstein, D. and Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision (IJCV) 47(1/2/3), pp. 7–42.

Schickler, W. and Thorpe, A., 2001. Surface estimation based on LIDAR. Proceedings of the ASPRS.

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

Xu, F., Woodhouse, N., Xu, Z., Marr, D., Yang, X. and Wang, Y., 2008. Blunder elimination techniques in adaptive automatic terrain extraction. Vol. 37, pp. 1139–1144.