IMPROVING IMAGE SEGMENTATION USING MULTIPLE VIEW ANALYSIS

Martin Drauschke, Ribana Roscher, Thomas Läbe, Wolfgang Förstner
Department of Photogrammetry, Institute of Geodesy and Geoinformation, University of Bonn
martin.drauschke@uni-bonn.de, rroscher@uni-bonn.de, laebe@ipb.uni-bonn.de, wf@ipb.uni-bonn.de

KEY WORDS: Image Segmentation, Aerial Image, Urban Scene, Reconstruction, Building Detection

ABSTRACT

In our contribution, we improve image segmentation by integrating depth information from multi-view analysis. We assume the object surface in each region can be represented by a low order polynomial, and estimate the best fitting parameters of a plane using those points of the point cloud, which are mapped to the specific region. We can merge adjacent image regions, which cannot be distinguished geometrically. We demonstrate the approach for finding spatially planar regions on aerial images. Furthermore, we discuss the possibilities of extending of our approach towards segmenting terrestrial facade images.

1 INTRODUCTION

The interpretation of images showing building scenes is a challenging task, due to the complexity of the scenes and the great variety of building structures. As far as human perception is understood today, humans can easily group visible patterns and use their shape to recognize objects, cf. (Hoffman and Richards, 1984) and (Treisman, 1986). Segmentation, understood as image partitioning often is the first step towards finding basic image patterns. Early image segmentation techniques are discussed in (Pal and Pal, 1993). Since then, many other algorithms have been proposed within the image analysis community: The data-driven approaches often define grouping criteria based on the color contrast between the regions or on textural information. Model-driven approaches often work well only on simple scenes e. g. simple building structures with a flat or gabled roof. However, they are limited when analyzing more complex scenes.

Since we are interested in identifying entities of more than two classes as e.g. buildings, roads and vegetation objects, we cannot perform a image division into fore- and background as summarized in (Sahoo et al., 1988). Our segmentation scheme partitions the image into several regions.

It is very difficult to divide an image into regions if some regions are recognizable by a homogenous color, others have a significant texture, and others are separable based on the saturation or the intensity, e. g. (Fischer and Buhmann, 2003) and (Martin et al., 2004). However, often such boundaries are not consistent with geometric boundaries. According to (Binford, 1981), there are seven classes of boundaries depending on illumination, geometry and reflectivity. Therefore, geometric information should be integrated into the segmentation procedure.

Our approach is motivated by the interpretation of building images, aerial and terrestrial, where many surface patches can be represented by low order polynomials. We assume a multi-view setup with one reference image and its intensity based segmentation, which is then improved by exploiting the 3D-information from the depth image derived from all images. Using the determined orientation data, we are able to map each 3D point to an unique region. Assuming, object surfaces are planar in each region, we can estimate a plane through the selected points. The adjacent regions are merged together if they have similar planes. Finally, we obtain an image partition with regions representing dominant object surfaces as building parts or ground. We are convinced that the derived regions are much better for an object-based classification than the regions of the initial segmentation, because many regions have simple, characteristic shapes.

The paper is structured as followed. In sec. 2 we discuss recent approaches of combining images and point cloud information, mostly with the focus on building reconstruction. Then in sec. 3 we briefly sketch our approach for deriving a dense point cloud from three images. So far, our approach is semi-automatic due to the setting of the point cloud's scale, but we discuss the possibility of automatization for all its steps. In sec. 4 we present how we estimate the most dominant plane in the dense point cloud restricted on those points, which are mapped to pixels of the same region. The merging strategy is presented in sec. 5. Here we only study the segmentation of aerial imagery and present our results in sec. 6. Adaptations for segmenting facade images are discussed in each step separately. We summarize our contribution in the final section.

2 COMBINING POINT CLOUDS AND IMAGES

The fusion of imagery with LIDAR data has successfully be done in the field of building reconstruction. In (Rottensteiner and Jansa, 2002) regions of interests for building extraction are detected in the set of laser points, and planar surfaces are estimated in each region. Then the color information of the aerial image is used to merge adjacent coplanar point cloud parts. Contrarily, in (Khoshelham, 2005) regions are extracted from image data, and the spatial arrangement of corresponding points of a LIDAR point cloud is used as a property for merging adjacent regions. In (Sohn, 2004) multispectral imagery is used to classify vegetation in the LIDAR point cloud using a vegetation index. The advantage of using LIDAR data is to work with high-precision positioned points and a very limited portion of outliers. The disadvantage is its expensive acquisition, especially for aerial scenes. Hence, we prefer to derive a point cloud from multiple image views of an object.

Within the last years, the matching of multiple views of an object enabled the reconstruction of 3D object points with high accuracy and high density. Previous approaches as (Kanade and Okutomi, 1994) are based on a low-level preprocessing of the image to extract points of interest. Then, the correspondences of such points are used to estimate the 3D position of the object points. In many applications, Förstner-features (Förstner and Gülch, 1987) or SIFT-features (Lowe, 2004) are used, but the derived point clouds are either sparse or have been extracted from many images or video, e. g. (Mayer and Reznik, 2005) and (Gallup et al., 2007). In (Tuytelaars and Van Gool, 2000), the correspondences are determined over local affinely invariant regions, which were extracted from local extrema in intensity images. This procedure is liable to make matching mistakes if the image noise is relatively high.

Dense point clouds from only a few images are obtained by adjusting the correspondence between pixels by correlation based on (semi-) global methods, e. g. (Hirschmüller, 2005). Assuming the observed objects have a smooth surface, the accuracy of the obtained point clouds gets increased by including information on the relations between the pixels by a Markov random field, e. g. (Yang et al., 2009), or from image segmentation, e. g. (Tao and Sawhney, 2000).

In our approach, we take up the idea of (Khoshelham, 2005) to improve an initial image segmentation using additional 3D information. From multi-view analysis, we derive a point cloud, which is used for deriving additional features for the segmented image regions. We focus on building scenes, whose objects mostly consist of planar surfaces. So, it is reasonable to look for dominant planes in the point cloud, where the search is guided by the image segmentation.

For us, it is important to realize an approach, which has the potential to get automatized since there are many applications with thousands of images. There is a need for a completely automatic procedure if additional features are derived from a reconstructed point cloud to improve the segmentation or interpretation of the images. Our input are only two or more images from the object, which were taken by a calibrated camera. An example is shown in fig. 1.

3 RECONSTRUCTION OF THE 3D SCENE

In this section, we describe the generation of the point cloud $\boldsymbol{C}$ from the given images. For this generation, there are two conditions, which should be fulfilled: (a) the observed objects should be textured sufficiently and (b) the views must overlap, otherwise we have problems to determine the relative orientation between the images. So far, the implemented algorithms need some human interaction for setting the point cloud scale and the disparity range parameters, but under certain conditions, the whole approach could get designed to perform completely automatically.

We describe the procedure with two or three given images $\boldsymbol{I}_1$, $\boldsymbol{I}_2$ and $\boldsymbol{I}_3$. Two views are necessary to reconstruct the

Figure 1: Three aerial views of a building scene consisting of a flat roofed part and a gable roofed part. The initial segmentation of the upper view is shown on its right side. The ground consists of several weirdly shaped regions, and the flat roof is also not well segmented.

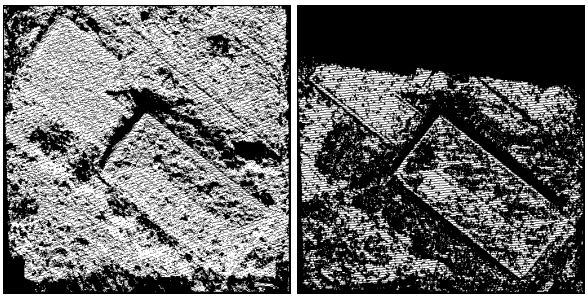

Figure 2: Reconstructed 3D-points are projected back into 2D-image (white). Left: all pairs of matches are shown. The point cloud is very dense with approximately 75% of pixels having a 3D point, but these points are very imprecise. Right: only matches in all three images are shown. The point cloud is still dense with approximately 30% of pixels having a 3D point with higher precision.

observed 3D data, but if the matching is performed over three images, the point cloud is still dense, see fig. 2, and it contains more reliable points, thus less outliers. The reconstruction process can get improved if even more images are considered. If all used images were taken by a calibrated camera, we are able to reconstruct the 3D scene by performing the following steps.

In the first step we determine the relative orientations between the given images. Of course, it can be skipped if the projection matrices have been estimated during image acquisition. Otherwise, due to the calibration of the camera we eliminate automatically the non-linear distortions using the approach of (Abraham and Hau, 1997). The matching of extracted key-points using the approach of (Lowe, 2004) leads to the determination of the relative orientations of all images, i. e. their projection matrices $\boldsymbol{P}_n$, cf. (Läbe and Förstner, 2006). The success of the relative orientation can

be evaluated according to the statistics of the performed bundle adjustment. This step is usually robust enough for typical building scenes, because the facades are often sufficiently textured, and we do not have to deal with total occlusions. Otherwise, problems may occur due to too large mirroring facade parts.

The images are oriented relatively, not absolutely, i. e. the position of the projection centers are not correctly scaled yet. Since we cannot invert a transformation from 3D to 2D, a reasonable assumption about the scale always has to be inserted additionally. The easiest way to set the scale parameter is to measure GPS positions during the image acquisitions. Another strategy would be to measure one or more distances on the object and to identify corresponding points in the images or in the extracted point cloud later. While the first way can easily get automatized, the second one has to be done by human interaction.

From the second step on, we only use three images for a dense trinocular matching and only accept those 3D points, which were matched in all three images. Thus, we reduce many matching errors close to the image borders and avoid points corresponding to occluded surfaces. We use the semi-global matching by (Hirschmüller, 2005) in a realization by (Heinrichs et al., 2007). It is efficient, does not produce too many outliers, and returns a dense point cloud with sufficiently precise points. This approach demands that the images are arranged in a L-shaped configuration with a base image, a further one shifted approximately only horizontally and a third shifted approximately only vertically. Due to the special relation between the three given images, the search space of the matching and 3D estimation of a point is reduced to a horizontal or vertical line, respectively. So far, the two parameters of the one-dimensional search space for the depth have to be set manually before the program is started. Usually, this range lies in a small bound assuming that the flying height or the distance of a facade to the camera are restricted and do not vary much.

The semi-global matching returns a disparity map, which is used to estimate the 3D point cloud by forward intersection. There are a couple of hundred or a thousand gross errors in the determined point cloud, which can be removed under the assumption that all points lie in a certain bounding box. Besides of the remaining outliers the most extracted 3D points form spatial clusters with clearly visible ground and roof planes, cf. fig. 3. Compared with other derived point clouds from stereo aerial imagery, e. g. MatchT[1], the precision of our reconstructed points is significantly lower, but we compensate it by the higher denseness.

4 REGION-WISE PLANE ESTIMATION

In this section, we describe the estimation of the most dominant plane for each detected image region of minimum size. Thereby, any arbitrary image partitioning algorithm

[1] Automated DTM Generation Environment by inpho, cf. www.inpho.de

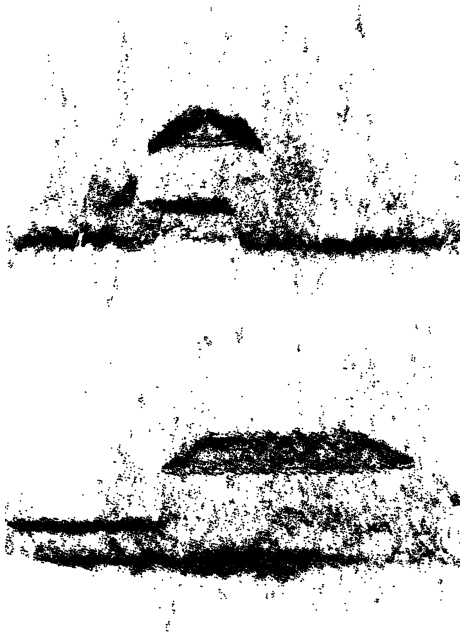

Figure 3: Side- and frontview on a point cloud, derived from scene extracts of the three aerial images from fig. 1. Besides the widely spread points on vegetation objects and some outliers, one can clearly recognize up to four major clusters showing the ground, a flat roof and a gabled roof.

can be chosen. In an earlier experiment, we made good experiences with segmenting aerial images using the watershed algorithm based on the color gradient, cf. (Drauschke et al., 2006). This segmentation approach is also applicable to facade images, cf. (Drauschke, 2009). To overcome oversegmentation at nearly all image parts, we smooth the image with a Gaussian filter with $\sigma = 2$ before determining the watershed regions. Then, oversegmented image parts are highly correlated with vegetation objects, which are not in our focus yet. Such an initial segmentation is shown in fig. 1. For further calculations, we only consider those regions R_k, which have a minimum size of 250 pixels. This parameter should depend on the image size. We have chosen a relatively high value for efficiency reasons.

In the further process, we want to estimate low order polynomial through the 3D points of each region, i. e. its most dominant plane. Therefore, we determine for each region the set of points $\{\boldsymbol{X}_j\}$ from the point cloud, which are projected into the region:

$$\boldsymbol{X}_j \mapsto R_k \;\Leftrightarrow\; \boldsymbol{x}_j = \boldsymbol{P}_n \boldsymbol{X}_j \text{ and } \boldsymbol{x}_j \in R_k. \qquad (1)$$

We assume that most dominant building surfaces and the ground are planar. Hence, we estimate the best fitting plane through the 3D points of a region. A similar procedure can be found in (Tao and Sawhney, 2000). For efficiency reason, we choose a RANSAC-based approach for our plane search, cf. (Fischler and Bolles, 1981). Therefore, we determine the parameters of the plane's normal form from three randomly chosen points $\boldsymbol{X}_{j_1}$, $\boldsymbol{X}_{j_2}$ and $\boldsymbol{X}_{j_3}$:

$$\boldsymbol{n} = (\boldsymbol{X}_{j_2} - \boldsymbol{X}_{j_1}) \times (\boldsymbol{X}_{j_3} - \boldsymbol{X}_{j_1}) \qquad (2)$$

$$d = \left\langle \frac{\boldsymbol{n}}{||\boldsymbol{n}||}, \boldsymbol{X}_{j_1} \right\rangle \qquad (3)$$

and check, how many object points support the determined plane i. e. how many points are near the plane. This depends on the choice of a threshold. Considering aerial images we allowed a maximal distance of 20 cm to the plane. If we want to guarantee with a minimum probability $p_{min} = 0.999$ finding a plane, which is constructed by 3 points and supported by at least half of the points ($\epsilon = 0.5$), we have to perform $m = 52$ trials, because

$$m = \frac{\log(1 - p_{min})}{\log(1 - (1 - \epsilon)^3)} = \frac{\log 0.001}{\log 0.875} \approx 51.7. \quad (4)$$

If no sufficiently high number of supporting points can be found within m trials, the region will no longer be analyzed. In our empirical investigation, segmented regions representing roof parts have always a most dominant plane. Such plane could not get found if e. g. the ground is not planar but forms a small hill or valley, e. g. at and around trees and shrubs. Furthermore, we accepted only those 3D points, which are visible in all three images. Therefore, occluded building parts are also not in further process.

We estimate the best fitting plane using a least-squares adjustment on those points, which support the best proposed plane during the iterations of RANSAC. The statistical reasoning[2] is taken from (Heuel, 2004), p. 145.

5 MERGING OF IMAGE REGIONS

So far, our approach can only handle with merging of regions. If the image is undersegmented in some image parts, i. e. the region covers two or more objects, a splitting criterion must be defined to separate this region parts again. We suggest to search for several dominant planes and to split the regions according to the intersections of these planes. We did not realize the splitting yet, so we only propose our merging strategy.

We determine a region adjacency graph and check for each adjacent pair of regions R_1 and R_2 if a merging of the regions can get accepted. The first test is on equality of the two corresponding estimated planes. We realized that our derived point cloud is too noisy for such statistical reasoning. Therefore, we consider a second test, where we determine the best fitting plane through the set of 3D points from both regions and then we check, if the new plane has a normal vector $\boldsymbol{n}_{12}$ which is similar to the normal vectors $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ of the two previous planes:

$$\angle(\boldsymbol{n}_{12}, \boldsymbol{n}_1) < \theta \ \wedge \ \angle(\boldsymbol{n}_{12}, \boldsymbol{n}_2) < \theta. \quad (5)$$

In our experiments, we used $\theta = 30^o$, which leads to reasonable results with respect to buildings. If one is interested in each individual roof plane, θ should not be more than 10^o. If other applications cannot depend on such a heuristically chosen parameter, we suggest to adapt this condition by a MDL-based approach, cf. (Rissanen, 1989). Then, two regions should be merged, if the encoding of data would decrease when merging.

[2]SUGR: Statistically Uncertain Geometric Reasoning, www.ipb.uni-bonn.de/projects/SUGR

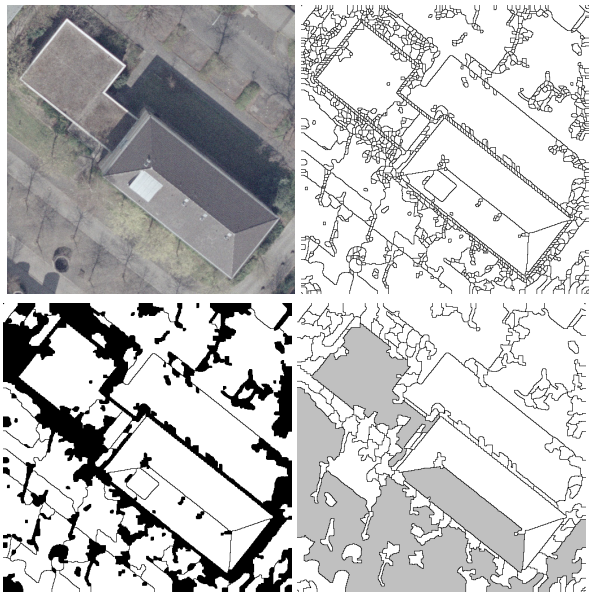

Figure 4: Steps of improving image segmentation. In the upper row, we show the reference image and its initial segmentation. In the bottom row, we show at the left all big regions from the initial partition (in white) and the final segmentation including the MDL-based and the geometry-based grouping of regions. There, the gray-shadowed regions have been merged on the basis on geometric properties.

Until this point, we did not consider small regions whose dominant planes cannot be estimated reliably. Now, we also merge them, too. Small holes can easily merge with their surrounding region, but all others may be merged according to an intensity-based criterion. We implemented a MDL-based strategy according to (Pan, 1994), where we additionally stop the merging as soon as the minimum size of a region has been reached. As alternatives, we could also use strategies for irregular pyramid structures, e. g. (Guigues et al., 2003), which is based on similarity of color intensities or (Drauschke, 2009) which is based on scale-space analysis. Resulting image segmentation is shown in fig. 4.

6 EXPERIMENTS

We have tested our segmentation scheme on 28 extracts of aerial images with known projection matrices showing urban scenes in Germany and Japan. The images from Germany were taken in early spring when many trees are in blossom, but are not covered by leaves yet. The 3D points matched at such vegetation objects are widely spread, cf. fig. 3. In most cases, the corresponding image parts are oversegmented, so that no dominant planes have to get estimated. There is almost no vegetation in the Japanese images, but the ground is often dark from shadows. As mentioned earlier, we have problems with finding precise 3D points in lawn and shadow regions, but with respect to building extraction (i. e. segmenting the major roof parts), our approach achieves satisfying results cf. fig. 5. We are convinced to get better results for matching in dark image parts, if a local enhancement is used to brighten these parts

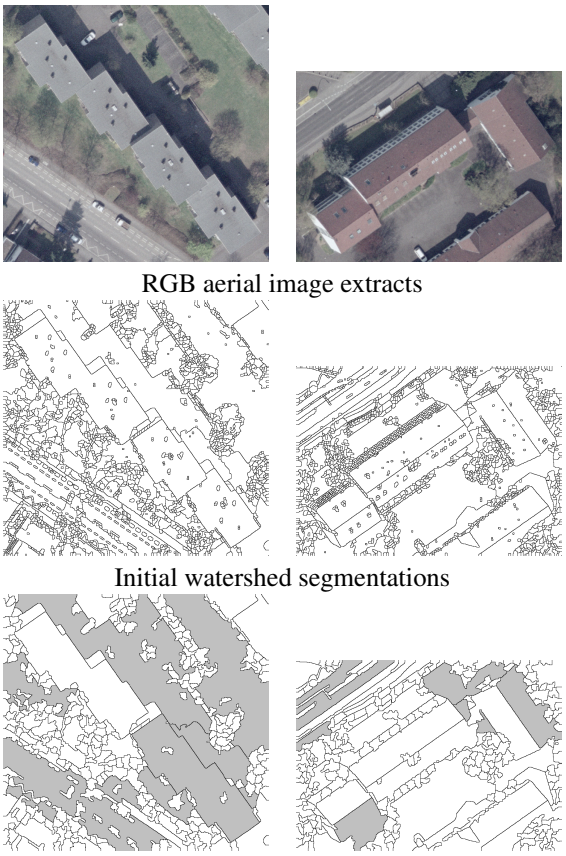

Figure 5: Results of simple building scenes. Again, the gray-shadowed regions have been merged on the basis on geometric properties.

in a preprocessing step, e. g. (Zhao and Lei, 2006). A further improvement should be acchieved, if the whole procedure is repeated, because the MDL-based merged regions are now big enough for determination of their geometric properties.

The noise of the point cloud, which we derive from the semiglobal matching does not disturb the merging of image regions. Considering aerial images, we are faced with large and often planar objects. There, our plane estimation is good enough, because we do not have to many outliers. Otherwise, the plane estimation should be done by a robust estimator. If different object parts have been segmented as one region, then the most dominant plane of the combined region often does not represent one of these object parts. This shows us, that we need to focus in the future on an algorithm for detecting multiple planes (e. g. analysis of the best five planes from RANSAC) and a splitting routine. Furthermore, there are objects as trees or dormers which violate our assumption of having one planar surface. Therefore, we consider to adapt our plane estimation towards extracting general geometric primitives as planes, cylinders, cones and spheres, cf. (Schnabel et al., 2007).

With respect to facade images, we have big trouble with our plane estimation. We ascribe this fact to two major reasons. First, the reconstruction part is challenged by homogenous facades and mirroring or light transmitting windows. Both cases lead to too many outliers. And secondly, the noise of the complete point cloud is too high to differ between planes in the object space, which are parallel, but only a view centimeters apart. Fig. 6 shows a facade image and three views on the dominant planes of given annotated objects. In this case, the supporting points may have a distance of 4 cm to the fitting plane. Dominant planes with distances of more than half of a meter are clearly separable from each other.

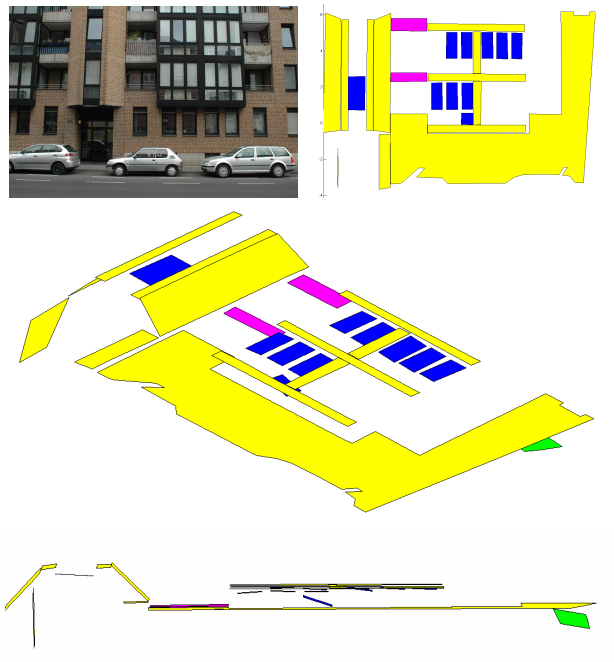

Figure 6: Facade image and different views on fitted planes for hand-labeled object parts. Wall components are drawn in yellow, windows in blue and (if opened) in green, balcony parts in magenta. The planes of overhanging building parts are well distinguishable, but the window planes (if not opened) are very close to its surrounding wall parts. The mirroring and light transmission effects in the window sections lead to geometrically instable plane estimations.

7 CONCLUSION AND OUTLOOK

We presented a novel approach for improving image segmentations for aerial imagery by combining the initial watershed segmentation with information from a 3D point cloud derived from two or three views. For each region, we estimate the most dominant plane, and only the plane parameters are used to trigger the merging process of the regions. With respect to building extraction, our algorithm achieves satisfying results, because the ground and major building structures are better segmented.

In the next steps, we want to search for multiple planes for each region, and we want to implement a splitting routine, so that regions can either get merged or split. If we have such a reliable function, we would start the region merging using the MDL criterion based on the image intensities. So, we can search for geometric descriptions in all, and not only in the big image regions. Furthermore, our approach

can get improved, if we estimate more general geometric primitives for representing the object's surfaces.

Acknowledgements

This work was done within the project *Ontological scales for automated detection, efficient processing and fast visualisation of landscape models*, which is supported by the German Research Foundation (DFG). The authors would also like to thank our student Frank Münster for preparing the data and assisting the evaluation.

REFERENCES

Abraham, S. and Hau, T., 1997. Towards autonomous high-precision calibration of digital cameras. In: SPIE, pp. 82–93.

Binford, T., 1981. Inferring surfaces from images. Artificial Intelligence 17(1-3), pp. 205–244.

Drauschke, M., 2009. An irregular pyramid for multi-scale analysis of objects and their parts. In: GbRPR'09, LNCS 5534, pp. 293–303.

Drauschke, M., Schuster, H.-F. and Förstner, W., 2006. Detectability of buildings in aerial images over scale space. In: PCV'06, IAPRS 36 (3), pp. 7–12.

Fischer, B. and Buhmann, J. M., 2003. Path-based clustering for grouping smooth curves and texture segmentation. PAMI 25(4), pp. 513–518.

Fischler, M. and Bolles, R., 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. ACM 24(6), pp. 381–395.

Förstner, W. and Gülch, E., 1987. A fast operator for detection and precise location of distinct points, corners and centers of circular features. In: ISPRS Conf. on Fast Processing of Photogramm. Data, pp. 281–305.

Gallup, D., Frahm, J.-M., Mordohai, P., Yang, Q. and Pollefeys, M., 2007. Real-time plane-sweeping stereo with multiple sweeping directions. In: CVPR'07.

Guigues, L., Le Men, H. and Cocquerez, J.-P., 2003. The hierarchy of the cocoons of a graph and its application to image segmentation. Pattern Rec. Lett. 24(8), pp. 1059–1066.

Heinrichs, M., Rodehorst, V. and Hellwich, O., 2007. Efficient semi-global matching for trinocular stereo. In: PIA'07, IAPRS 36 (3/W49A), pp. 185–190.

Heuel, S., 2004. Uncertain Projective Geometry. LNCS 3008, Springer.

Hirschmüller, H., 2005. Accurate and efficient stereo processing by semi-global matching and mutual information. In: CVPR, pp. II: 807–814.

Hoffman, D. D. and Richards, W. A., 1984. Parts of recognition. Cognition 18, pp. 65–96.

Kanade, T. and Okutomi, M., 1994. A stereo matching algorithm with an adaptive window: Theory and experiment. PAMI 16(9), pp. 920–932.

Khoshelham, K., 2005. Region refinement and parametric reconstruction of building roofs by integration of image and height data. In: CMRT'05, IAPRS 36 (3/W24), pp. 3–8.

Läbe, T. and Förstner, W., 2006. Automatic relative orientation of images. In: Proc. 5th Turkish-German Joint Geodetic Days.

Lowe, D., 2004. Distinctive image features from scale-invariant keypoints. IJCV 60(2), pp. 91–110.

Martin, D., Fowlkes, C. and Malik, J., 2004. Learning to detect natural image boundaries using local brightness, color, and texture cues. PAMI 26(5), pp. 530–549.

Mayer, H. and Reznik, S., 2005. Building façade interpretation from image sequences. In: CMRT'05, IAPRS 36 (3/W24), pp. 55–60.

Pal, N. R. and Pal, S. K., 1993. A review on image segmentation techniques. Pattern Rec. 26(9), pp. 1277–1294.

Pan, H.-P., 1994. Two-level global optimization for image segmentation. P&RS 49(2), pp. 21–32.

Rissanen, J., 1989. Stochastic Complexity in Statistical Inquiry. World Scientific.

Rottensteiner, F. and Jansa, J., 2002. Automatic extraction of buildings from lidar data and aerial images. In: CIPA, IAPRS 34 (4), pp. 569–574.

Sahoo, P., Soltani, S. and Wong, A., 1988. A survey of thresholding techniques. CVGIP 41(2), pp. 233–260.

Schnabel, R., Wahl, R. and Klein, R., 2007. Efficient ransac for point-cloud shape detection. Computer Graphics Forum 26(2), pp. 214–226.

Sohn, G., 2004. Extraction of buildings from high-resolution satellite data and lidar. In: 20th ISPRS Congress, IAPRS 35 (B3), pp. 1036–1042.

Tao, H. and Sawhney, H. S., 2000. Global matching criterion and color segmentation based stereo. In: Workshop on Applications of Computer Vision, pp. 246–253.

Treisman, A., 1986. Features and objects in visual processing. Scientific American 225, pp. 114–125.

Tuytelaars, T. and Van Gool, L., 2000. Wide baseline stereo matching based on local, affinely invariant regions. In: BMVC, pp. 412–422.

Yang, Q., Wang, L., Yang, R., Stewénius, H. and Nistér, D., 2009. Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling. PAMI 31(3), pp. 492–504.

Zhao, J. and Lei, S., 2006. Automatic digital image enhancement for dark pictures. In: ICASSP, pp. II: 953–956.