Detection of Small Roof Details in Image Sequences

Dimitri Bulatov and Melanie Pohl

Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB),
Gutleuthausstr. 1, 76275 Ettlingen, Germany
{dimitri.bulatov,melanie.pohl}@iosb.fraunhofer.de

Abstract. Detecting smaller elevated objects, like chimneys, in high resolution images has several important applications, such as collision warning. On the other hand, the already existing 3D models (that already include the terrain, buildings and vegetation) can be enriched by new instances. There are not many contributions about extracting fine roof details in the literature. Therefore, we developed a new, modularized algorithm for detecting these details as hot spots in the local elevation maps; such a map is typically obtained by a multi-view dense matching method. We use explicit and implicit assumptions on data in order to tighten the search range for chimneys and reduce the number of false alarms. Finally, filtering hot spots by means of color or intensity images takes place. Thus, good detection rates can be achieved for a data set consisting of several high resolution images taken over a residential area.

Keywords: Chimneys, depth map, hot spot detection, multi-view, modeling, roof details, urban terrain.

1 Introduction and Previous Work

Automatic detection and reconstruction of urban terrain from aerial images has been studied for several decades and has a wide field of applications. Some of them are urban planning, surveillance, but also obstacle and collision warning. For this reason, it is often an advantage to extract not only large elevated objects, such as buildings and vegetation, but also smaller ones, e. g. chimneys or high-voltage transmission lines. The state-of-the-art offers many approaches for detailed reconstruction of buildings [1–3] in aerial images (possibly supported by other sources of information, like airborne laser point clouds), but there are not many contributions for detecting chimneys and other fine roof structures to find, in spite of their good visibility in more than one image. In [4, 5] and other comparable sources, the authors either arrive at the conclusion that too small details can hardly be extracted, or they marginally report about their detection without sufficient quantitative analysis. Indeed, it is not easy to extract these details if the only input is a set of optical images. Even the case of Nadir images, these details can be easily confused with roof-lights, solar cells, shadows of trees or street lamps, and many others.

Beside a pure scientific interest of being able to detect smaller objects – one of the most interesting and challenging topics of image analysis – and thus enrich the 3D models of urban terrain, detection of small objects is useful for various applications.

For example, in [6, 7] chimneys are recognized to be a disturbing entity for change detection respectively building reconstruction. Moreover, automatic detection of chimneys is also an interesting topic in verification of heat pictures with regard to heat isolation of roofs, placing telecommunication lines, and, concerning military and police activities, choice of launching and landing site for helicopters as well as possible hide-outs.

Recently, by combining the generic probabilistic approach for multi-view depth maps computation [8] and a context-based approach of building reconstruction from elevation maps [9], it was possible to create a reconstruction procedure that goes the entire way from image sequences to semantic 3D models of urban terrain. They consist of buildings (modeled by polygons), vegetation (trees) and digital terrain model (DTM). This procedure and its modifications are described in [10, 11]. In this work, we propose an additional module for extracting small elevated objects by a hot spot detector. We use discretized elevation maps as the main input. To avoid the problem of oversmoothing such hot spots using some non-local method, we use the local result of the elevation map in combination with the so called *confidence maps*. These maps are a measure of how reliable the depth value of a given pixel is. The texture image will only be used to verify the hypothesis. The modeling step is the only one for which it becomes relevant what kind of elevated roof details we are dealing with. For the rest of this paper, we identify "small roof segments" with "chimneys" which are modeled by 3D cuboids.

The paper is structured as follows: The necessary background on our procedure for urban terrain reconstruction will be given in Sec. 2 and the step of detecting chimneys – our main contribution – in Sec. 3. Qualitative and quantitative results are given in Sec. 4. Finally, Sec. 5 summarizes main ideas and research directions of the future work.

2   Multi-view Dense Reconstruction and Urban Terrain Modeling

Suppose we have several images with corresponding camera matrices. It has been an important area of research how to obtain precise and reliable 3D positions for a dense set of pixels in these images. Many of the state-of-the-art methods mentioned in the survey [12] presuppose retrieving (discrete) depth or disparity values $j \in \{1, ..., L\}$ for pixels $i \in \{1, ..., N\}$ in some reference image. The collection of depth and disparity values, $\mathcal{D}(i) = j$, is called depth or disparity map $\mathcal{D}$. In order to treat occlusions in a robust way, the method [8] uses *simultaneously* a multi-view configuration and starts by computing a data-cost term $\mathcal{E}^{\text{data}}(j, i)$ for every pixel i and every label of the discretized depth scale j. To do this, dissimilarities (for example, absolute differences of gray values) between pairs of images are measured and $\mathcal{E}^{\text{data}}(j, i)$ is the average value of the lower among them. The term $\mathcal{E}^{\text{data}}(j, i)$ is perturbed by a triangle-based term $\mathcal{E}^{\text{mesh}}(j, i)$ that penalizes the difference between j and j_0, the initialization of the depth label that may stem from triangle interpolation of depths of several available 3D points. Finally, a smoothness term $\mathcal{E}^{\text{smooth}}(j, i)$ that penalizes the difference of depth labels for neighboring pixels, transfers the geometric problem of depth map computation into a multi-labeling problem: Find the mapping $\mathcal{D}(i)$ that minimizes the total energy term over a two-dimensional graph:

$$\mathcal{E} = \sum_{i=1}^{N} \left(\mathcal{E}^{\text{data}} + \mathcal{E}^{\text{mesh}} + \mathcal{E}^{\text{smooth}} \right) (\mathcal{D}(i), i). \tag{1}$$

This problem has L^N solutions in total and cannot be solved in a polynomial time. Therefore, approximating algorithms must be considered. In the current work, the semi-global optimization method of [13] was replaced by the graph-based alpha-beta-swap procedure due to [14, 15] that is available online, because it yields slightly better results for Nadir images where many 3D points lie in fronto-parallel planes. However, because a non-local method often oversmoothes some small hot spots treating them as noise, we will use the *local result* – that is, the minimum value within the i-th column in the $L \times N$ matrix $\mathcal{E}^{\text{data}}(j,i)$ – as the main input for a hot spot detector (hence solving (1) is not needed here). From $\mathcal{E}^{\text{data}}(j,i)$, we can also compute a *confidence map $\mathcal{C}$*, see e. g. [8]. If the confidence value for a pixel lies below a threshold, this means that the minimum of the corresponding column of $\mathcal{E}^{\text{data}}$ is rather shallow. All such pixels are excluded from the detection of hot spots.

Since depth value of a pixel codes a 3D coordinate of the surface point, the local or global result for the depth map can be transformed into an *elevation map* or *digital surface model* (DSM), denoted by $\mathcal{Z}$ or $\mathcal{Z}_{loc}$, respectively. The procedure for detection and reconstruction of buildings remains roughly the same as in [9], though several modules are modified as indicated in [10, 11]. The building detection part consists of extracting DTM, separating buildings from vegetation as well as filtering and labeling elevated regions of the difference DSM – (DTM + Vegetation). This labeled image $\mathcal{B}$ comprises our building hypotheses. The building reconstruction part contains modules for boundary extraction, roof detail analysis and texturing of the resulting 3D polygons by means of the input images. In the roof detail analysis step that we address here, we replaced the normal vector extraction and clustering from local gradient information proposed in [9] by the more global RANSAC procedure [16], modified in an appropriate way. Three points are chosen randomly to constitute a plane and the set of inliers (points that are close enough to the plane) is determined. After evaluation of several such plane hypotheses, the plane with the biggest number of inliers is chosen. With a Least Squares optimization, the plane is adapted successively to the set of inliers after which a new set of inliers is generated. These two subroutines – inlier set (re)computation and non-linear refinement – are repeated iteratively over the whole point cloud. The resulting set of inliers is temporarily removed from the point cloud and the procedure of dominant planes extraction begins again with the remaining points. This technique can be applied either on whole buildings, or, in case the building ground polygon can be cut into pieces in a reasonable way along its diagonals, on each building part.

Because RANSAC often oversmoothes small roof details (dormers, chimneys), we make an effort to identify them in advance, model them separately, and exclude the corresponding 3D-points from further consideration.

3   Detecting and Modeling Chimneys

The road-map of the algorithm with an emphasis on input for chimneys detection module ($\mathcal{Z}_{loc}, \mathcal{B}, \mathcal{C}$ and the reference image I) is visualized in Fig. 1. The three main steps, namely, hypothesis generation (hot spots detection), hypothesis verification and modeling chimneys is described in subsections 3.1, 3.2 and 3.3, respectively.
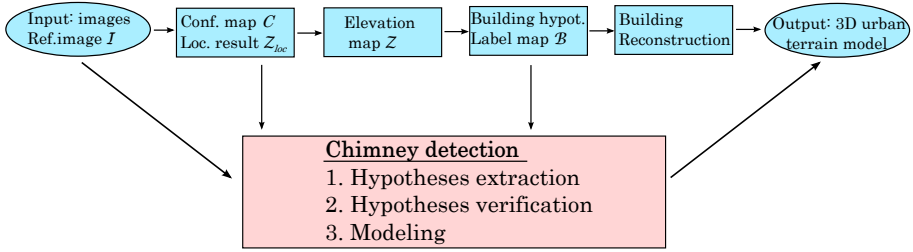
Fig. 1. Road-map for urban terrain reconstruction with a focus on chimney detection

3.1 Hypothesis Generation

Though any hot spot detector can be used for detecting small elevated regions in the difference of Z_{loc} and the DTM, we decided to use the MSER (Maximally Stable Extremal Regions) method [17]. Its performance has shown advantages in comparison to a similar procedure, namely, an enhanced version of [18]. Also, a fast implementation of this algorithm [19] is available. As a main input, MSER requires a discretized (8-bit) intensity image. In the first step, its pixels are sorted by their intensity levels. Then, a temporary image is flooded level by level allowing components to be built, extended and merged. The most stable components are those with very slowly growing areas as a function of the current intensity level g:

$$g = \arg\min_{g'} \frac{\mathcal{A}(g'+\delta) - \mathcal{A}(g'-\delta)}{\mathcal{A}(g')} + \varepsilon|\mathcal{A}(g') - \mathcal{A}_0|, \tag{2}$$

where $\mathcal{A}$ is the area function, $\mathcal{A}_0 = 1\text{m}^2 \cdot$ image resolution^2 is a typical chimney size in pixel, $\delta = 1$ or 2 is an algorithm parameter and ε is a negligibly small scalar needed to resolve ambiguities for the case the numerator in (2) is zero. For a chimney that has a height z_0, the absolute difference in pixels between the intensity values of its seed and the background in the 8-bit image will be:

$$d = 255\frac{z_0}{\Delta h}, \tag{3}$$

where Δh is the height range in the elevation image. In order to increase the resolution of elevations for a better detection of chimneys, our goal must be to reduce Δh. To do this, the minimum and maximum values for elevation ($z_{\min}$ and $z_{\max} = z_{\min} + \Delta h$) are given by the minimum and maximum elevations, respectively, of the building *with chimneys* in $\mathcal{B}$. In case that many buildings in the data set have too different heights, one can subdivide the image in smaller parts or even rescale the data for every single building. In the data set presented in Sec. 4, it was not needed, and an empiric value between 4 and 5 meters was used for $z_{\min}$. Moreover, all pixels for which the confidence map $C < c_0$ and $\underline{\mathcal{B}} = 0$ are excluded from consideration. Here c_0 is a user-specified threshold and $\underline{\mathcal{B}}$ is the morphological closing of $\mathcal{B}$ followed by an erosion. This is done to avoid false

alarms at the building borders that are caused by sudden jumps in the local elevation map (see Fig. 2, top row, middle). The hot spot detector is applied at the image

$$\mathcal{Z}'_{loc} = \frac{255z_0}{\Delta h} \left(\mathcal{Z}_{loc} - DTM - z_{\min}\right) \cdot \left(\mathcal{C} > c_0\right) \cdot \left(\underline{\mathcal{B}} > 0\right), \tag{4}$$

which is only non-zero inside of buildings of sufficient elevation and confidence.
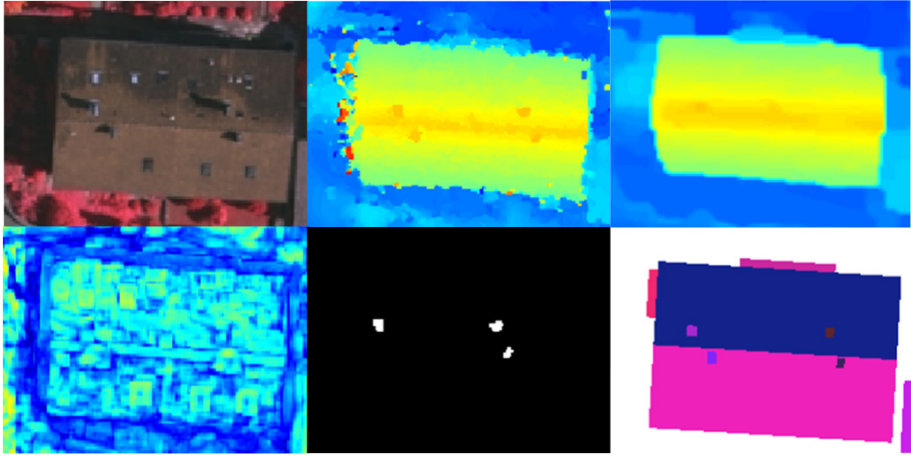
Fig. 2. Top row, left: Part of the reference image. Middle: Part of the local elevation map $\mathcal{Z}_{loc}$. Right: Result of non-local optimization (oversmoothing noise, but also some fine details). Bottom row, left: Confidence map from the local result as a color map. Middle: Binary mask with MSERs colored white. Three of four chimneys of this building were detected. Right: Ground truth of roof plane segmentation.

Finally, we are concerned about the choice of other parameters for the MSER algorithm. The issue of interest is given by fair components on the dark background. The minimum and maximum area of a component is given by $\mathcal{A}_0$ from (2). The difference of intensity values within the component and the contrast are given by the maximal height of a chimney in the data set and equation (3) while a filter on intensity values in the component restricts the position of a chimney. We keep track of these parameters within the computation of MSERs in order to collect only components of suitable properties and to perform necessary filtering afterwards. In Fig. 2, we show for a test data set discussed in the results section: A fragment of the reference image, the corresponding fragments of the local and non-local elevation maps, confidence mapuadra as well as of the result of the MSERs detection with the ground truth of roof plane segmentation.

3.2   Hypothesis Verification

So far we have not used the RGB-image. In this section, we show two possibilities how to reduce the number of false alarms among detected components. In the majority of cases, the false alarms are provoked by noise in the depth map and correspond to rather

homogeneously textured areas in the original image. On the other hand, presence of chimneys exhibits some variations of the color distribution in I. Therefore, the simplest kind of verification consists of considering the standard deviation σ of color or intensity values of pixels in the component: $f = \sigma(I(i \in \overline{C}))$, where C is the connected component previously detected, $\overline{}$ denotes a morphological dilatation and $I(\cdot)$ can denote a scalar or vector function, in the case of intensity or color images, respectively. Only components for which $\|f\|$ exceeds a threshold (≈ 20 for an 8-bit image) are accepted and modeled as chimneys.

An alternative approach consists of computing MSERs in the original image. We compute here both fair and dark hot spots for the gray image I, but only reduced to the regions around the previously detected components. This helps saving computation time and verifying which hypotheses result from the noise and which ones exhibit structure in the textured image and are therefore most likely to be true positives.

3.3 Modeling Chimneys

The height of a chimney is given by the average local elevation value of all pixels – of sufficient confidence – that lie in the corresponding connected component. Since we are only concerned about detecting and not about retrieving the exact structure of the chimney, we consider a quadratic segment whose size is calculated from the area of the component (in pixels) and image resolution. Such an open cuboid, provided a typical texture of a chimney, is added to the model at the position that corresponds to the centroid of the component, as depicted in Fig. 3.
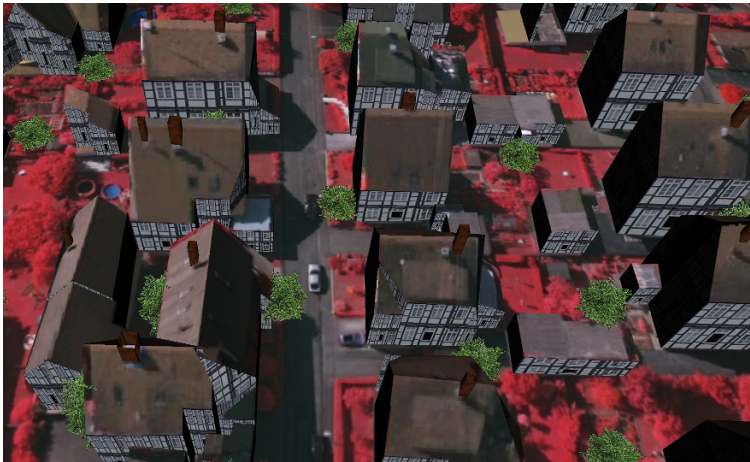
Fig. 3. A detailed view of the 3D model for the data set *Vaihingen*. House roofs and surrounding terrain are textured with the image of Fig. 4, left, and wall textures are standard. Trees are detected and modeled as proposed in [10].

4   Test Data and Computational Results

The test data set is the test area 3 of the ISPRS WG III/4 benchmark on object detection (ISPRS, 2012). Seven 16 bit pan-sharpened color infrared images showing the village *Vaihingen* in southern Germany are given. Moreover, also the corresponding camera matrices, as well as a very detailed ground truth result are available. The area of interest contains 56 buildings and the number of chimneys is 61. The reference image is taken from the almost Nadir perspective and will therefore serve as the coordinate system for the elevation map. The data set is very suitable for the evaluation of results. On the one hand, the detailed ground truth segmentation is available. On the other hand, the data set contains many buildings with almost all kinds of challenges that are characteristic for the discussed task: Shadowy regions (partly shadow of trees on the building roofs), overexposed and underexposed segments, high voltage transmission lines over the house roofs and many others. The data set thus offers a lot of examples all of which should be tackled with the same parameters. The depth map was computed and buildings were extracted as explained in Sec. 2; these input data and intermediate results are visualized in Fig. 4. The subsequent evaluation measures the *correctness* (ratio of true positives among the total number of all detected objects) and the *completeness* (ratio of true positives among all objects of interest in the ground truth) of the extracted candidates. It is clear that in most cases, a choice of parameters that contributes to enhancing correctness (that is, reducing false alarms) will reduce completeness, and vice versa.
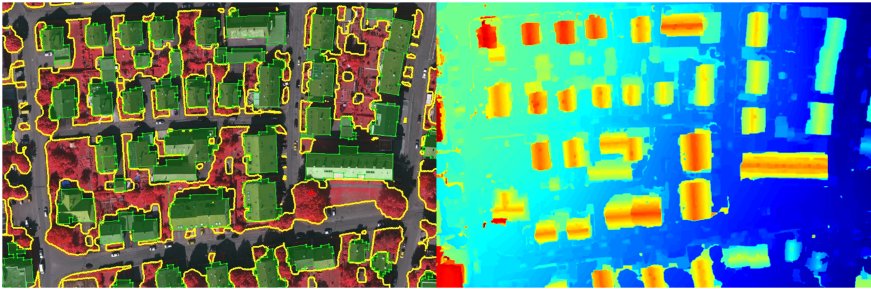
Fig. 4. Left: Part of the reference image from the data set *Vaihingen* and the intermediate results of the buildings (*green polygon interiors*) and vegetation (*yellow*) detection. Right: The corresponding elevation map (without vegetation) computed from seven images with [8].

In case of filtering based on standard deviation, as proposed in the first paragraph of Sec. 3.2, up to 40 out of 61 chimneys could be detected, and there were 12 false alarms. Tightening thresholds yields 36 correct and six false alarms. In other words, the correctness rate of our algorithm takes on values between 0.77 and 0.85 while the completeness is little lower, between 0.59 and 0.66. The result of detecting MSERs in the reference image and considering intersection yields 42 true alarms and 16 false alarms or in values 0.74 and 0.75 for completeness and correctness, respectively. Finally, using the enhanced version of [18] for detection of hot spots instead of MSERs deteriorates the results: The number of false alarms grows (correctness 0.60) while the completeness measure remains at 0.61.

As mentioned before, the main source of false positives lies in image noise: Shadows of the trees or other objects on the building roofs often result in mismatches of our dense matching procedure; neither can they be filtered out by means of the RGB image. In the first strategy, filtering based on standard deviation, roof ridges sometimes provide an additional source of errors. These are actually elevated regions, but the colors of the roof are not the same due to different angles of incidence of the sun. Another reason of obtaining false alarms lies in different kinds of elevated objects on the roof (open garret windows, TV antennas, flower pots on the balconies) that are spuriously accepted as chimneys. The main reason for missed chimneys lies in shortcomings of the hot spot detection. Several chimneys can be filtered out after computation of MSERs in original images because the algorithm in its original implementation can either detect dark regions with a fair background or fair regions with the dark background. Probably, another segmentation algorithm would be more suitable here. Also, because we perform an erosion of $\mathcal{B}$, several chimneys near building outlines have also been lost. This step is, however, necessary since local depth maps have many outliers in shadowy, partly occluded regions around buildings.

5   Summary and Outlook

We presented a fast, straight-forward and fully automatic procedure for chimneys detection in multi-view configurations of high resolution images of urban terrain. To our knowledge, it is the first method that explicitly extracts fine roof details although several possible applications can be outlined. Although we assumed in this work that the camera matrices corresponding to these images and the depth ranges are available, this information can be obtained from images by means of a structure-from-motion approach (such as [20, 21]) in a relative coordinate system; since orientation and scaling is an easy task, one can state that no further information beside images is actually needed. The procedure has a modular structure: Its modules for hot spot detection, verification etc. can easily be replaced or modified if needed.

The procedure uses explicit assumptions – that is, we search for small elevated regions of the DSM, usually with a certain contrast with respect to the background – but also implicit assumptions; for example, the fact that chimneys are usually visible in more than one image of the sequence reflects in a high confidence. Another implicit assumption, namely, that chimneys usually cast shadows, justifies the application of the hypothesis verification module of Sec. 3.2. It was possible to achieve a satisfactory result for completeness and correctness while evaluating a test data set with many buildings and containing a large number of problems the algorithm has to deal with. In Sec. 2, we mentioned that roof planes segmentation is carried out by means of RANSAC, with a consequence that smaller regions tend to be oversmoothed. However, if care is taken and these regions are pre-computed in advance and excluded from further consideration, the qualitative and quantitative results of the reconstruction will become better and an appealing and accurate 3D model (see Fig. 5) can be enriched by new instances.

As a consequence, in our future work, we shall not only trail the results of detection of chimneys but also their reconstruction (height and exact position) as proposed in [22]. In addition, we shall adapt our algorithm for other kinds of small elevated objects (vehicles, transmission lines) and to test its performance on a broader class of data sets.
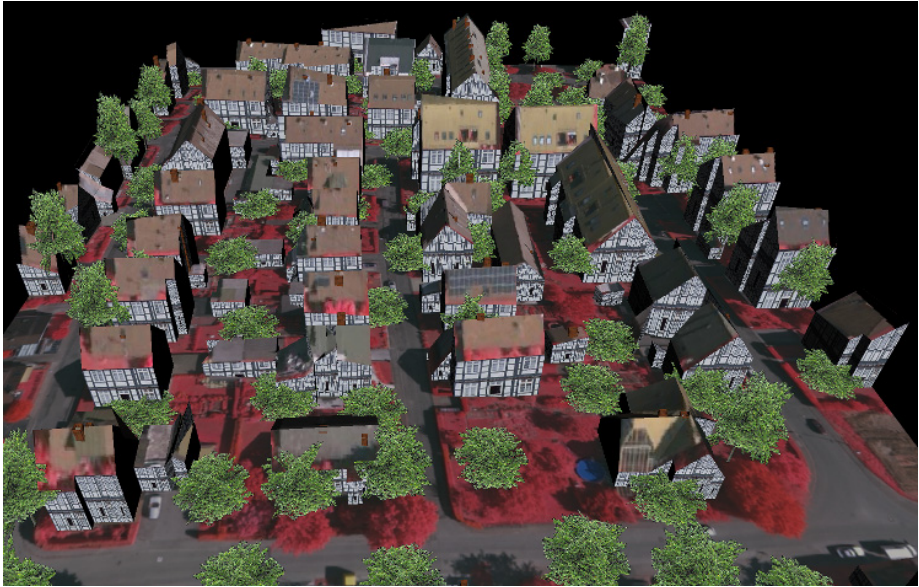
Fig. 5. A view of the 3D model for the data set *Vaihingen*

References

1. Fischer, A., Kolbe, T., Lang, F., Cremers, A., Förstner, W., Plümer, L., Steinhage, V.: Extracting buildings from aerial images using hierarchical aggregation in 2D and 3D. Computer Vision and Image Understanding 72(2), 185–203 (1998)
2. Rottensteiner, F.: Roof plane segmentation by combining multiple images and point clouds. In: Proc. of Photogrammetric Computer Vision and Image Analysis Conference, Int. Arch. of Photogrammetry and Remote Sensing, vol. 38(pt. 3A), pp. 245–250 (2010)
3. Zebedin, L., Bauer, J., Karner, K., Bischof, H.: Fusion of feature- and area-based information for urban buildings modeling from aerial imagery. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 873–886. Springer, Heidelberg (2008)
4. Brenner, C., Dold, C.: Gebäudeextraktion aus luftgestützten und terrestrischen Scandaten. In: Photogrammetrie-Laserscanning-Optische 3D-Messtechnik, Beiträge der Oldenburger 3D-Tage, pp. 312–319. Wichmann Verlag, Heidelberg (2004)
5. Lafarge, F., Descombes, X., Zerubia, J., Pierrot-Deseilligny, M.: Building reconstruction from a single DEM. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 1–8 (2008)

6. Hommel, M.: Detektion und Klassifizierung eingestürzter Gebäude nach Katastrophenereignissen mittels Bildanalyse. PhD thesis, Karlsruhe Intitute of Technologies, KIT (2010)
7. Mosch, M.: 3D-Gebäudeextraktion aus Satellitenbildern suburbaner Regionen. PhD thesis, University of Freiburg (2006)
8. Bulatov, D., Wernerus, P., Heipke, C.: Multi-view dense matching supported by triangular meshes. ISPRS Journal of Photogrammetry and Remote Sensing 66(6), 907–918 (2011)
9. Gross, H., Thönnessen, U., v. Hansen, W.: 3D-Modeling of urban structures. In: Joint Workshop of ISPRS/DAGM Object Extraction for 3D City Models, Road Databases, and Traffic Monitoring, CMRT 2005, Int. Arch. of Photogrammetry and Remote Sensing, vol. 36(pt. 3W24), pp. 137–142 (2005)
10. Bulatov, D., Solbrig, P., Gross, H., Wernerus, P., Repasi, E., Heipke, C.: Context-based urbanterrain reconstruction from m(uav)-videos for geo-information applications. In: Unmanned Aerial Vehicle in Geomatics Conference (ISSN XXXVIII-1/C22), 1682–1777 (2011)
11. Bulatov, D., Wernerus, P., Gross, H.: On applications of sequential multi-view dense reconstruction from aerial images. ICPRAM 2, 275–280 (2012)
12. Scharstein, D., Szeliski, R.: A Taxonomy and Evaluation of Dense Two-frame Stereo Correspondence Algorithms. International Journal of Computer Vision 47(1), 7–42 (2002)
13. Hirschmüller, H.: Stereo processing by semi-global matching and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(2), 328–341 (2008)
14. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Trans. Pattern Anal. Mach. Intell. 26(9), 1124–1137 (2004)
15. Delong, A., Osokin, A., Isack, H.N., Boykov, Y.: Fast approximate energy minimization with label costs. International Journal of Computer Vision 96(1), 1–27 (2012)
16. Fischler, M., Bolles, R.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM, 381–395 (1981)
17. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: BMVC, pp. 761–767 (2002)
18. Anderer, C., Thönnessen, U., Carlsohn, M.F., Klonz, A.: Ein Bildsegmentierer für die echzeitnahe Verarbeitung. In: DAGM-Symposium, pp. 380–384 (1989)
19. Wassenberg, J., Bulatov, D., Middelmann, W., Sanders, P.: Determination of maximally stable extremal regions in large images. In: SPPRA-Conference (2008)
20. Bulatov, D.: Towards Euclidean reconstruction from video sequences. In: Int. Conf. Computer Vision Theory and Applications, vol. (2), pp. 476–483 (2008)
21. Beder, C., Steffen, R.: Incremental Estimation Without Specifying A-priori Covariance Matrices for the Novel Parameters. In: VLMP Workshop on IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, USA (2008)
22. ISPRS: ISPRS gtest project on 3D building reconstruction (2012), http://www.commission3.isprs.org/wg4/
23. Cramer, M.: The DGPF test on digital aerial camera evaluation - overview and test design. Photogrammetrie – Fernerkundung – Geoinformation 2, 73–82 (2010)
24. Spreckels, V., Syrek, L., Schlienkamp, A.: DGPF project: evaluation of digital photogrammetric camera systems – stereoplotting. Photogrammetrie – Fernerkundung – Geoinformation 2, 117–130 (2010)