Towards 3D Map Generation from Digital Aerial Images

Lukas Zebedin, Andreas Klaus, Barbara C. Gruber-Geymayer, Konrad Karner

VRVis Research Center for Virtual Reality and Visualization, Graz, Austria

Abstract

This paper describes the fusion of information extracted from multispectral digital aerial images for fully automatic 3D map generation. The proposed approach integrates spectral classification and 3D reconstruction techniques. The multispectral digital aerial images consist of a high resolution panchromatic channel as well as lower resolution RGB and near infrared (NIR) channels and form the basis for information extraction.

Our land use classification is a 2-step approach that uses RGB and NIR images for an initial classification and the panchromatic images as well as a digital surface model (DSM) for a refined classification. The DSM is generated from the high resolution panchromatic images of a specific photo mission. Based on the aerial triangulation using area and feature-based points of interest we are able to generate a dense DSM by a dense matching procedure. Afterwards an true ortho photo for classification, panchromatic or color input images can be computed.

In a last step specific layers for buildings and vegetation are generated and the classification is updated.

Key words: Classification, Aerial Triangulation, Dense Matching, Information Fusion, True Ortho Photo

1 Introduction

Digital aerial cameras can be used to produce images with a high degree of image overlap in flight direction at almost no additional costs. A terrain point may be visible in 5 to 15 images (depending on strip overlap, velocity and altitude of the airplane). Currently, aerial photogrammetry is undergoing a "paradigm shift" [1] which means the transition from minimizing the number of film photos due to human operator intensive processing to maximizing the

robustness of automation due to high redundant image information using new large format digital aerial cameras.

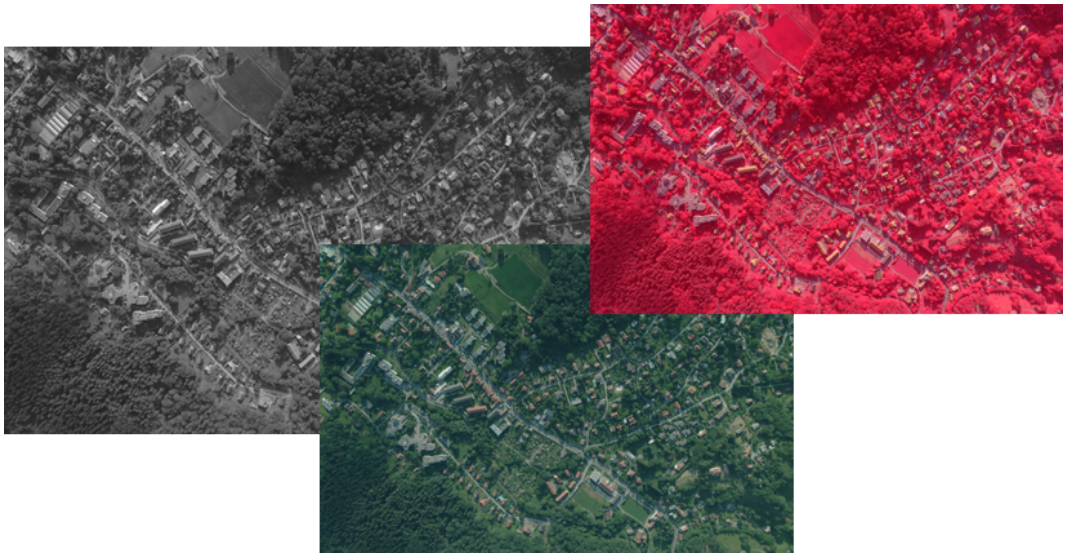


Fig. 1. Image data used: panchromatic high resolution image (left), RGB low resolution image (middle), NIR-R-G low resolution image (right)

This contribution is based on images from the UltraCamD camera from Vexcel Imaging with its multispectral capability. UltraCamD features a multi-head design. It delivers large format panchromatic images composed from nine CCD sensors (11500 pixels cross track and 7500 pixels long track) and simultaneously recorded four additional channels (red, green, blue and NIR) at a frame size of 3680 by 2400 pixels.

Additional ideas about photogrammetric color sensing may be found in [2]. The image data used comprise the panchromatic high resolution images as well as the low resolution multispectral images, see Figure 1.

The workflow is composed of the following steps:

- Initial classification of all images, see Section 2
- Aerial triangulation (AT), see Section 3
- Dense matching to generate a dense DSM, see Section 4
- True ortho photo production, see Section 5
- Object Recognition using initial classification, DSM and ortho photos, see Section 6
- Digital Terrain Model (DTM) and layer generation, Section 7

2 Initial Classification

The initial classification is a supervised classification performed on each of the overlapping color images with 4 color channels (RGB and NIR). The classes rely mainly on color and infrared and will be refined later using the information from the DSM.

The following types of classifiers for supervised classification can be found in literature:

- maximum likelihood classifier
- neural network classifier
- decision tree classifier
- support vector machine (SVM)

One recent paper by Farhad Samadzadegan et al. [3] aims also at classification of aerial images but uses a novel approach based on neuro-fuzzy modelling.

Many - especially newer - papers propose SVM for use in multispectral data. The purpose of [4] is to demonstrate the applicability of SVM to derive land use from operational sensor systems and to evaluate systematically their performances in comparison to other popular classifiers.

The SVM represents a group of theoretically superior machine learning algorithms, see [5]. The SVM employs optimization algorithms to locate the optimal boundaries between classes. Statistically, the optimal boundaries should be generalized to unseen samples with least errors among all possible boundaries separating the classes, therefore minimizing the confusion between classes. An important benefit of the SVM approach is that the complexity of the resulting classifier is characterized by the number of support vectors rather than the dimensionality of the transformed space. As a result, SVMs tend to be less prone to the problem of over-fitting than some other methods. Initial classification uses the SVM library LIBSVM developed at the National Taiwan University, see [6] for software details and [7] for a practical guide to support vector classification.

Initial classification discriminates all classes that are more significantly described by color and NIR values than by texture and spatial relationship. The number and nature of those classes can be adapted to the target area (rural, urban, ...). The classes used for the urban and suburban areas in this paper are:

- Solid: man made structures like streets, buildings with gray or non-colored roofs
- Colored roofs

- Soil, bare earth
- Lake, river, sea
- Vegetation: wood, grassland, fields
- Dark shadows
- Swimming pools

The first step in classification is feature extraction, i.e. the process of generating spectral feature vectors from the 4 input planes. The selection of the features to be extracted is important because it determines the amount of features that have to be computed and processed. In addition to the improved computational speed in lower dimensional feature spaces there might also be an increase in the accuracy of the classification algorithm. The features computed for initial classification include

- Single pixel values of all input planes
- Normalized ratio between image planes. Ratio images may be used to remove the influence of light and shadow on a ridge due to the sun angle. It is also possible to calculate certain indices which can enhance vegetation or geology. NDVI - Normalized Difference Vegetation Index - is a commonly used vegetation index which uses the red and infrared bands of the spectrum.
- Values computed in a circular neighborhood of given radius like minimum, maximum or standard deviation

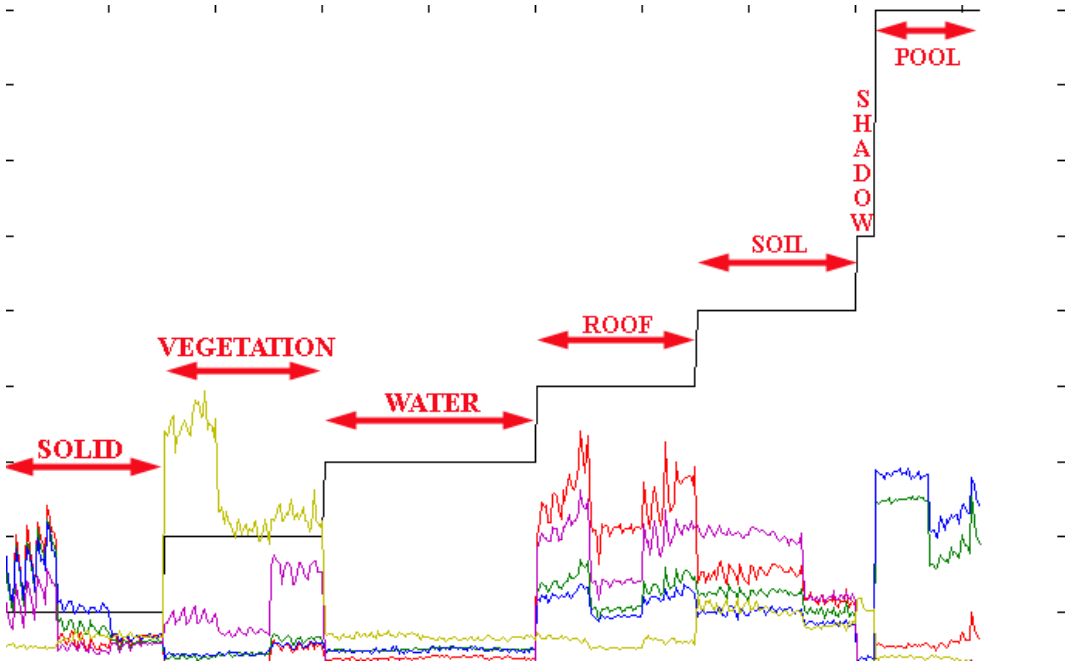


Fig. 2. A subset of feature values used in initial classification

Figure 2 illustrates a subset of feature values for some classified image regions. The colors represent the following features:

- R,G, and B colored red, green and blue
- NIR is colored violet
- NDVI is colored in yellow

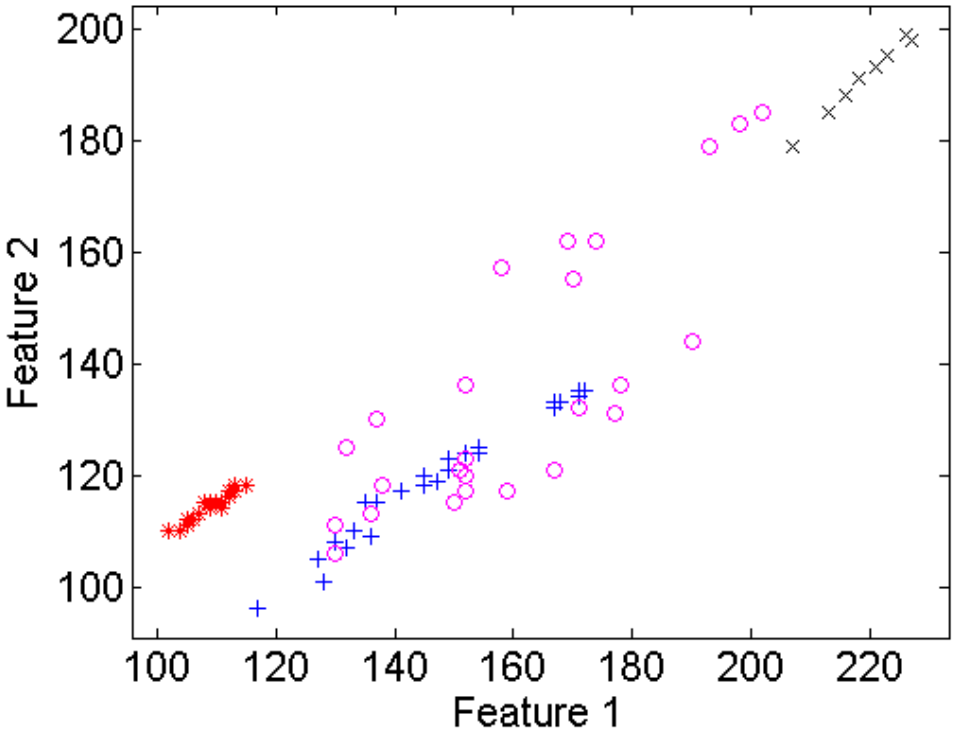


Fig. 3. Distribution of two image features related to the predefined classes

Additional features computed are not included as they would reduce the clearness of Figure 2. The distribution of features can be depicted as shown in Figure 3 for each two features. The SVM is trained to find optimal boundaries between the classes represented by these image features. Initial classification is performed by applying the trained SVM on each pixel.

The result of the initial classification is for each pixel the most probable class including its probability and additionally a second class and its probability if there are two classes with high probabilities. The two classes and their probabilities will be used when the fusion of several initial classification results will be performed, see Section 5.

In a supervised classification the analyst identifies several areas in an image, which represent known features or land use. These known areas are referred to as 'training sites' where groups of pixels are a good representation of the land cover or surface phenomenon. Using the pixel information the classification procedure then looks for other areas which have a similar grouping and pixel value. The analyst decides on the training sites and thus supervises the classification process.

The identification of training sites has to be done

- after radiometric camera calibration
- when the weather and lightening conditions or the land properties significantly change

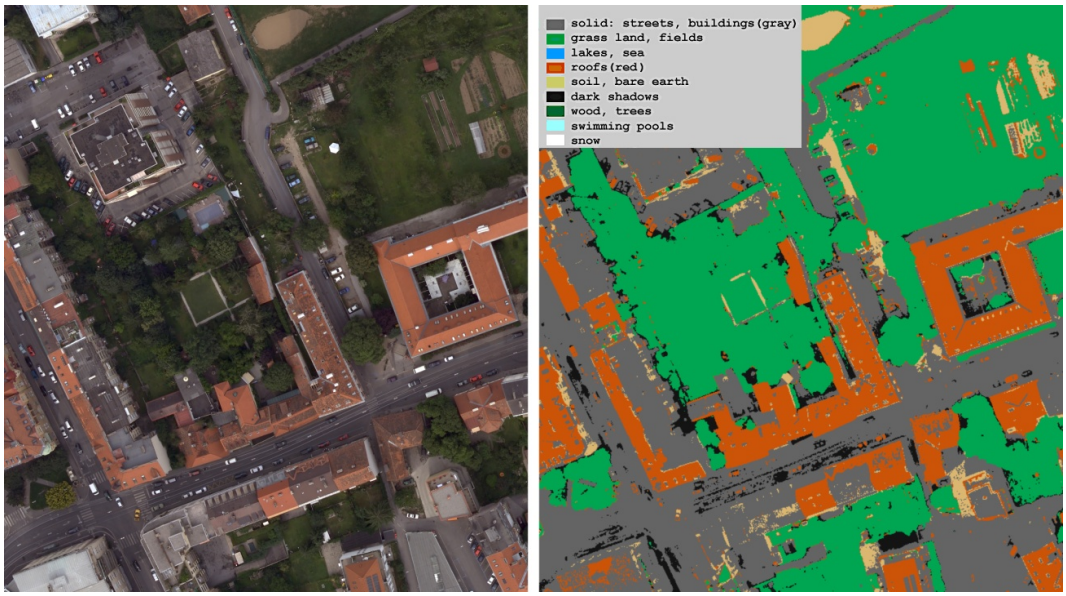


Fig. 4. RGB image detail (left), Initial classification result (right)

Figure 4 depicts the initial classification of the test area in one image. Radiometric classification has problems with gray roofs, as they can not be distinguished from streets. This limitation is overcome in the refined classification which incorporates also the information from the dense match.

3 Automatic Aerial Triangulation

The fundament of aerial triangulation (AT) is to establish correspondences, denoted as tie points, between adjacent images [8]. Digital airborne cameras are able to deliver high redundant images which result in small baselines. Normally, the stripes of images have at least 80% forward overlap and at least 20% side overlap (in urban areas 60% side overlap). Anyhow, the forward overlap is restricted by the trigger rate of the camera and further depends on the aircraft altitude and the flying speed. In the case of low altitude the baseline relative to the viewed scene may become significant as is illustrated in Figure 5. For image pairs with low side overlap the viewpoint change may be even higher. Moreover the epipolar geometry is unknown. Therefore we have to face the uncalibrated wide baseline matching problem in order to establish correct correspondences. Techniques that do not take slanted surfaces into account, such as window based correlation methods, may fail in this case, due

to projective and radiometric distortions. The following assumptions are made to allow a fully automatic processing:

(1) Projective distortion of small surface patches can be approximated by affine transformations. The motivation is that a small planar surface patch undergoes an approximate affine transformation if the viewpoint changes.
(2) Smooth surfaces can be modeled by piecewise planar patches. This approximation is valid for urban areas as well as for rural regions.
(3) Adjacent images of an aerial sensing strip overlap to at least 60%.
(4) Adjacent strips overlap to at least 20%.
(5) The in-plane rotation between adjacent images of a strip is low.
(6) The scale does not differ significantly. This assumption is fulfilled if the altitude is nearly constant.

Fig. 5. Two adjacent images of a strip with significant viewpoint change

Subsequently we will describe our AT workflow, that can be decomposed into 6 consecutive steps:

(1) **Extraction of Points of Interest**
Several thousand Points of Interest (POIs) are extracted in each image. Our POI extraction is based on Harris points and POIs from line intersections [9]. The POIs are sorted by their location and cornerness measure in order to guarantee a good distribution over the image. A sub-pixel refinement of the Harris POIs is calculated by a parabolic fit.
(2) **Calculation of feature vectors**
The feature vectors for POIs from line intersection are calculated from the area enclosed by the two lines which are similar to those proposed by Lowe [10] and are insensitive to affine transformations.
Harris POIs are described by their local neighborhood.
(3) **Establishing matching candidates**
The feature vectors are matched to find a 1 to n mapping between POIs of adjacent images. Each of the n best candidates is evaluated by applying an adaptive area based correlation.

In order to fulfill the non-ambiguous criteria, only matches with a highly distinctive score are retained.

(4) **Outlier elimination**
The robustness of the matching process is enhanced by processing a back-matching and by topological filtering [11]. Another restriction is enforced by the epipolar geometry. Therefore the RANSAC method is applied to the well known five point algorithm [12].

(5) **Relative orientation**
As a result from the last step, we obtain inlier correspondences as well as the essential matrix. By decomposition of the essential matrix the relative orientation of the current image pair can be calculated.

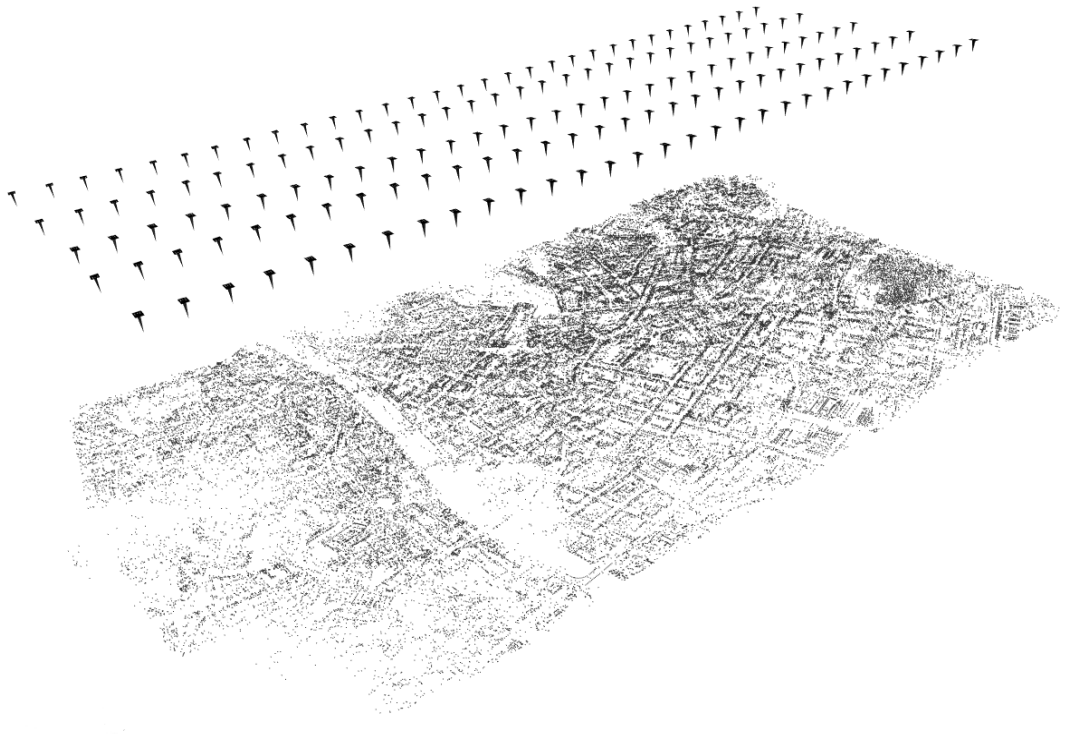


Fig. 6. Oriented block of 5 stripes of 31 images each denoted by small arrows. The reconstructed tie points (about 70.000) are displayed as black dots.

(6) **Final orientation**
The previous steps are accomplished for all images or all adjacent image pairs. In order to get the orientation of the whole set, the scale factor for additional image pairs (except the first) has to be determined. This is done using corresponding POIs available in at least three images. A block bundle adjustment refines the relative orientation of the whole set and integrates other data like GPS or ground control information. Figure 6 shows an oriented block of images. The 5 x 31 aerial images are oriented with respect to each other using about 70.000 tie points on the ground which are shown as black dots in Figure 6. The whole block of images was processed without any human interaction.

4 Dense Matching

Once the AT is finished we perform an area based matching to produce a dense DSM. During the last few years many new dense matching algorithms were introduced. A good comparison of stereo matching algorithms is given in a paper by Scharstein et al. [13].

In our approach we focus on an iterative and hierarchical method based on homographies to find dense corresponding points. First, the assigned area of interest (which should be reconstructed) is tiled into slightly overlapping regions. This procedure has two reasons: (i)Due to insufficient main memory it may not be possible to load all the large format images at the same time. Therefore matching the whole region at once would require a permanent reloading and would decrease the performance. (ii) The tiling enables a distributable calculation and this is important due to dense matching of large regions is very time consuming. Next, the following two-stage approach is applied to each region:

Initialisation using Plane Sweeping
Our matcher has to work with high depth discontinuities (e.g. from high buildings). Therefore we need to initialize the consecutive dense matcher in order to ensure convergence. The correspondences are determined as follows:
First, correlation values for all depth hypotheses are determined by performing a plane sweeping as proposed by [14]. For each potential depth value, all images are projected onto a common plane using projective image warping. A 3D depth space is accumulated by calculating a correlation value for each pixel in the warped image. The range of the plane sweeping (nearest and furthest plane) is derived from the reconstructed POIs determined in the AT.
Next, the best depth values are searched for using semi-global optimization. Unlike in local methods, where the best local hypotheses are taken, we are also using the neighborhood in the depth space by applying an iterative optimization schema.

Dense area based matching
The plane sweeping result is improved and densified by applying an iterative and hierarchical method based on homographies. For each input image an image pyramid is created and the calculation starts at the coarsest level. Corresponding points are determined and upsampled to the next finer level where the calculation proceeds. This procedure continues until the full resolution level is reached. Visibility information is implicitly used for each local patch by excluding images where it might be occluded.
A more detailed description of this algorithm implemented on graphics hardware can be found in [15].

The algorithm works without any human interaction and also reconstructs huge areas by splitting them into small tiles. Each tile is processed independently, therefore directly supporting parallelization on multiple computers. Afterwards the results are fused together into one large height field (Figure 7 illustrates what the output looks like).

Fig. 7. Height field of the inner city of Graz (approximately $3.7km^2$); reconstructed completely automatically without any human interaction. The image is 7141 by 4761 pixels large and has an orientation embedded into its header to reconstruct the 3D model.

5 Ortho photo generation

An ortho photo is obtained by combining an array of aerial images and constructing a realistic map of the terrain below. Current methods rely heavily on a simple stitching procedure considering the DEM, but still introducing irritating errors (like houses with clearly visible facades or mismatches at stitched tile borders). Using the result of the dense match however, it is possible to calculate a true ortho photo.

The true ortho photo has two major advantages over the stitching approach: One benefit is that by fusing various images together, disturbing objects (like moving trams, cars or pedestrians) can be removed in the ortho photo. This feature relies on redundancy in the aerial images, therefore requiring a high overlap (approximately 80% in the test scene, resulting in about 5 views per

strip for one 3D point; twice or three times as much if multiple adjacent strips are available). The second advantage is that because no input image is used directly (as in the stitching approach), depending on the quality of the dense match no facades or other vertical structures are visible.

The output of the dense match is a 2.5D elevation model. This facilitates the generation of an ortho photo, as the model does not need to be orthographically projected into a plane, but already is available as a height field. Together with the orientation of a reference plane for the height field, the 3D points can be recovered. Using a special hierarchical data structure derived from the height field, the visibility test is performed between those 3D points and the available cameras. All views, which pass this visibility test, are used and merged to obtain the texture information. For each view the 3D point is projected into the image and bilinearly interpolated. Now the candidates are transformed into the CIE Lab color space because there the Euclidean distance is proportional to the perceived similarity of colors. Then the entries are divided into clusters and the average of the largest cluster is taken.

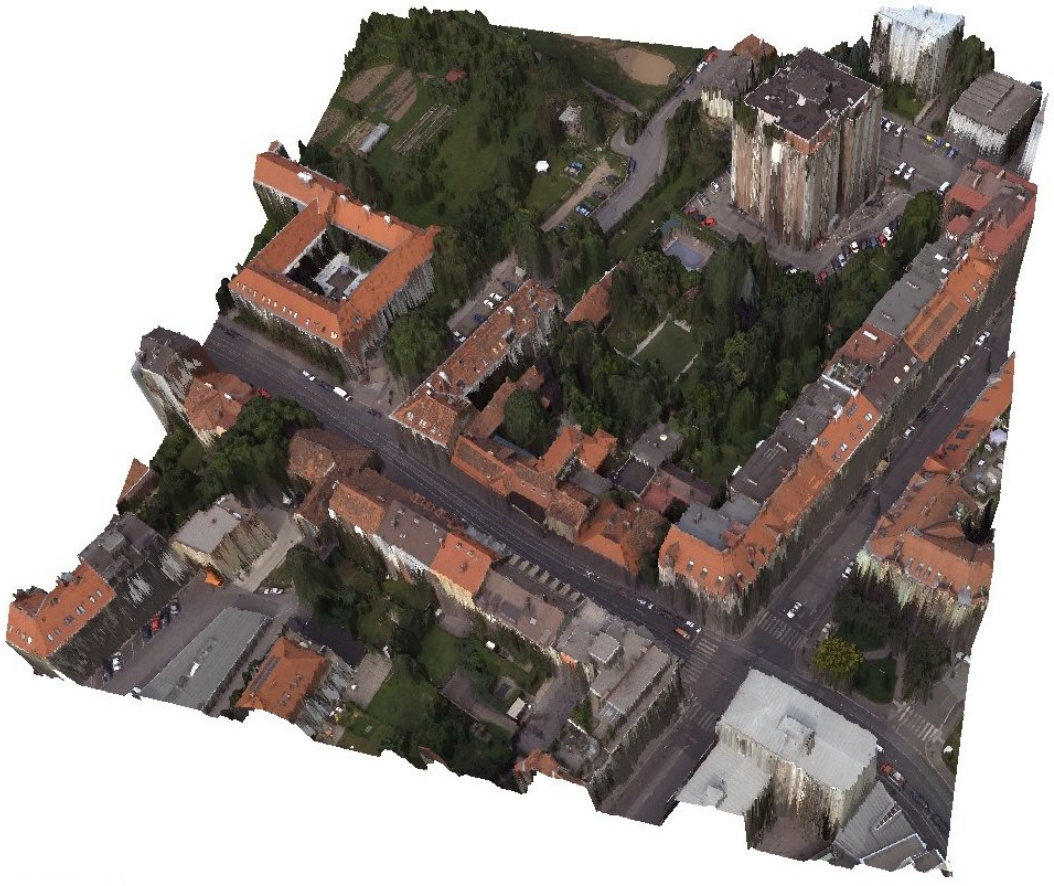


Fig. 8. DSM textured with an ortho photo of the test area.

This ortho photo can be used to texture the DSM, as it can be seen in Figure

8. The resolution of this ortho photo does not need to agree with that of the height field. In general the latter will have a higher ground sampling distance (GSD). Typically the dense match calculates one depth value for a given GSD. Nevertheless additional 3D points can be generated by linear interpolation to exploit the details in the input images.

Figure 9 illustrates the advantage of elimination of moving objects: cars driving though the street are detected and ignored, whereas parking cars are retained as they are stationary objects.

(a) (b)

(c)

Fig. 9. Images (a) and (b) show the same street in different images taken during one flight. In image (c) the resulting ortho photo is depicted illustrating the removal of moving cars.

The feature of object removal is most powerful when all available views are used, as the probability rises that the fusion algorithm chooses the background color and ignores any disturbing object. The visibility test, however, is the performance bottle neck of the algorithm, as the runtime increases linearly with the number of cameras used. Therefore the number of views used can be

capped at an arbitrary threshold (all ortho photos shown in this paper were calculated using nine views at most).

Apart from RGB ortho photo generation the same procedure can be applied to other input image types. The subsequent step of refining the initial classification for example requires an ortho fused image of those initial classification. Only minor changes are required to process them, as no interpolation may be done and the fusing algorithm is simplified as the input data has only one dimension.

6 Object Recognition using initial classification, DSM and ortho photos

(a) (b) (c)

Fig. 10. Input data for the refined classification: (a) ortho initial classification, (b) ortho panchromatic image and (c) height field.

Data fusion and the use of multiple classifiers, see [16], is a topic of special interest. The following data form the basis for data fusion in refined classification (Figure 10 depicts those input images for our test area):

- Ortho initial classification
- Ortho panchromatic image
- Height field image extracted from the DSM

Refined classification updates the results from initial classification arranged into an ortho classification using the spatial properties of the 3D features and the high resolution ortho panchromatic image. Spatial properties allow to differentiate trees from low vegetation, concrete roofs from streets, etc. Data fusion includes the computation of additional information from the ortho input data like height gradients, building blocks and texture measures.

The following refinement of initial classification results is performed:

- Solid gets refined into
  - Streets depicted gray
  - Buildings depicted in yellow
- Vegetation gets refined into
  - Grass land and fields depicted in bright green
  - Wood and trees depicted in dark green

The refined classification of objects of class solid implies the training of a minimal building height to distinguish between objects with low height like cars and small huts. The minimal building height is used to compute building blocks. Building blocks are defined as local height maxima as described in [17] that are restricted to all non vegetation and non water classes. The building blocks are computed in the following way for each pixel classified as solid or roof in initial classification:

- Compute significant minimal height value in a region with specified radius
- Compute maximal height difference, i.e. the difference between height value and significant minimal height value
- If the maximal height difference is higher than the trained minimal building height, then the pixel belongs to a building
- Remove small regions up to a specified size to prevent for example street-lamps or other small but high objects to be classified as buildings

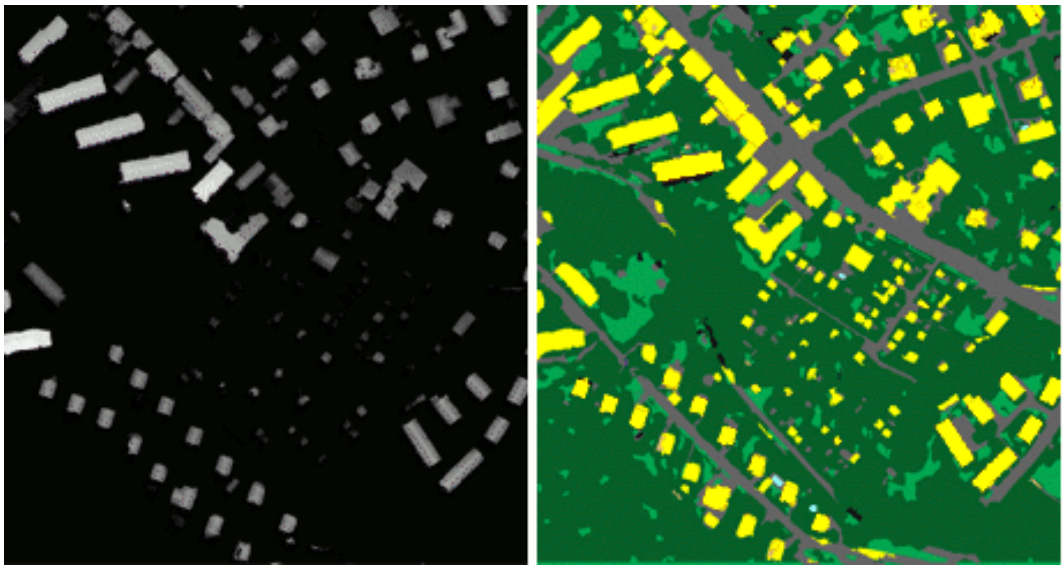


Fig. 11. building blocks (left), refined classification with buildings in yellow as well as trees in dark green and grass in bright green (right)

Refined classification for objects of class solid or roof is based on the computed building blocks. See Figure 11 for an example on building blocks and on refined classification results in which buildings - in yellow - are correctly classified.

Figure 12 shows how small objects like street-lamps or cars are handled in

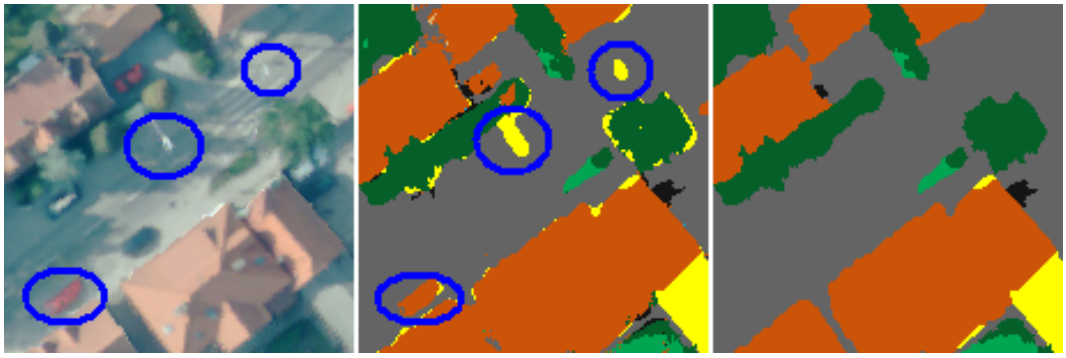


Fig. 12. RGB ortho photo with 2 street lamps and a red car marked (left), refined classification without verification of regions with small size (middle), refined classification result ignoring the street-lamps and the car (right)

refined classification. The two lamps are seen as solid objects with large height values which may lead to a classification as building. Due to the small size of the lamps they are not classified as building but as solid. Additionally small red objects like the red car in Figure 12 that are misclassified as roofs are reclassified as solid in the refined classification due to their small size.

Refined classification detects not only buildings but refines the class vegetation into grass and wood or trees. The refinement is based on a SVM trained using the following features:

- the panchromatic value
- mean and standard deviation of panchromatic values in a specified neighborhood of the pixel
- mean and standard deviation of height gradient values in a specified neighborhood of the pixel

The use of height gradient values improves the detection of wood or trees compared to approaches where only texture measures are used. Figure 13 gives an example of a house as well as trees, grass and solid. The trees are correctly classified.

Refined classification performs data fusion in a way that the classification results are less scattered, see again Figure 13: the initial classification - top middle image - has the roof correctly classified but the chimney and a small roof over a window are classified as solid. The height data - top right image - as well as the height gradients - bottom left image - and the building blocks - bottom middle image - cause a classification of the whole roof as one block, see bottom right image.

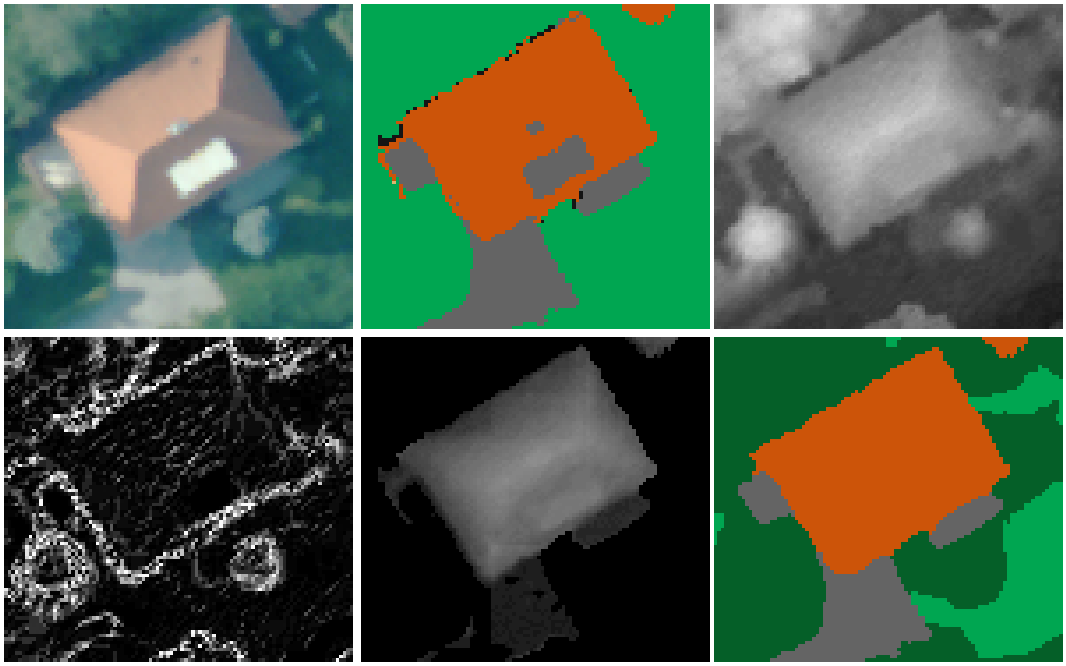


Fig. 13. RGB image of house with red roof (top left), initial classification (top middle), height data (top right), height gradients (bottom left), building blocks (bottom middle), refined classification (bottom right)

7 DTM and layer generation

The generation of a DTM from the DSM can be performed using the refined classification results. Objects of type building and tree have a height that has to be subtracted from the DSM to get the DTM. The height for buildings and trees is computed in the local neighborhood: the difference between height value in the DSM and a significant minimal height value the local neighborhood. The result shown in Figure 14 depicts a valley near Graz. Further work has to be done to smooth the DTM in an optimal way and to handle hills that are mainly covered with wood.

The refined classification result can be used to extract layer information. The most interesting layers are

- Building layer containing building
- Vegetation layer containing trees, wood and grass land
- Street layer containing streets and other places on the ground

Figure 15 illustrates the quality of the obtained information layers and demonstrates that the approach works even for large areas without any human interference. Figure 16 shows two details from the above mentioned figures and demonstrates that even small objects like single trees are classified correctly.

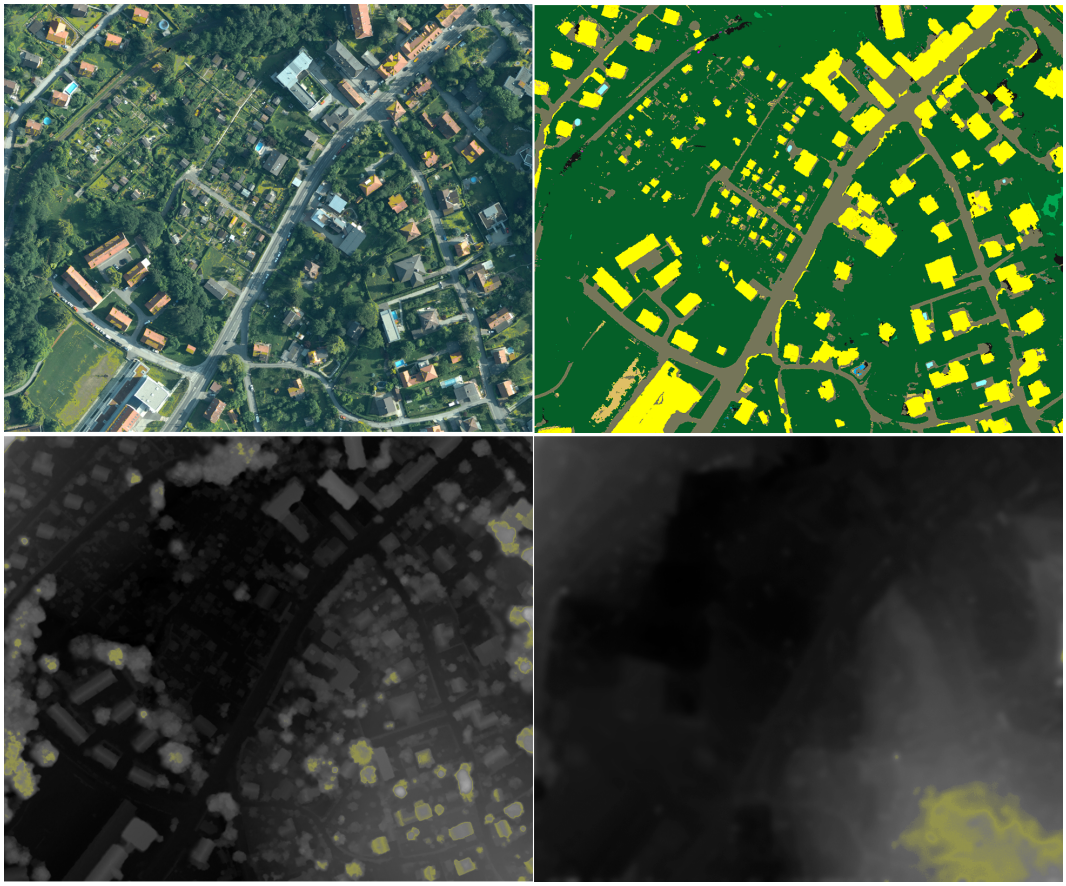
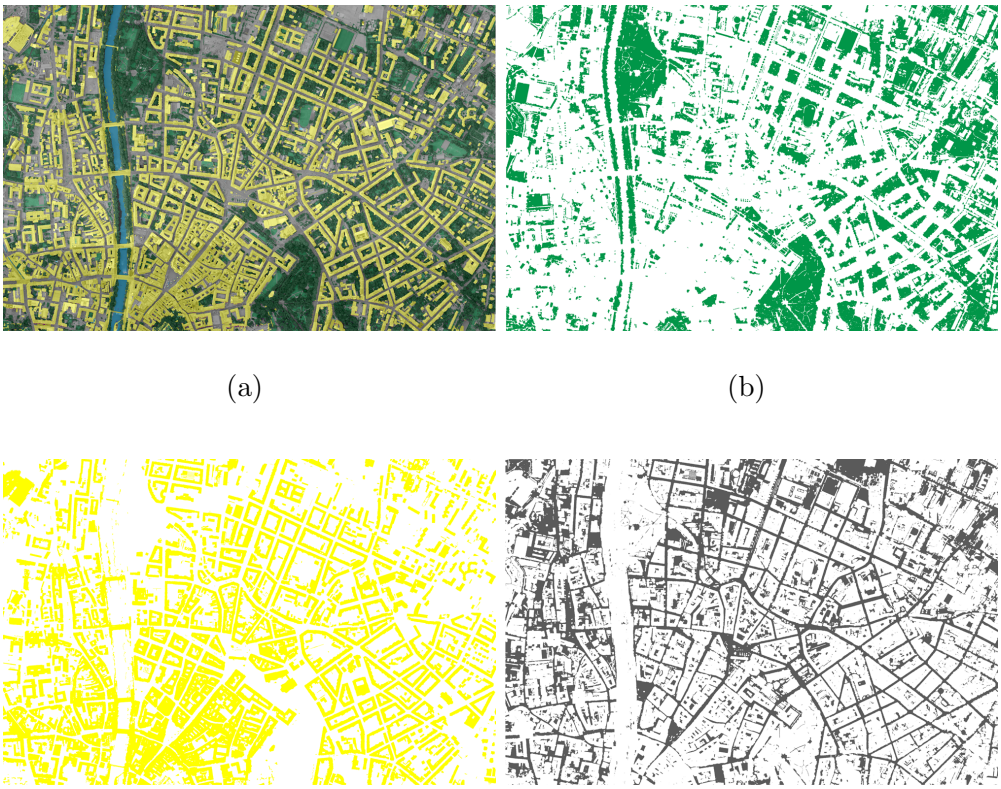


Fig. 14. One example of the generation process of a DTM: in the top left corner the ortho photo of the test area is shown. The classification of the same area is depicted in the top right corner. In the bottom left corner the corresponding height field is placed. The DTM generated by the algorithm outlined in the text is shown in the lower right corner.

8 Conclusions and Future Work

In our approaches we use the high redundancy in the source input images to generate a land use classification, a DSM and true ortho photos. Using these results a digital terrain model from the DSM and the classification can be derived, which represents only bald earth without any objects like trees or buildings. With the exception of the training phase for the initial classification no manual interaction is required. All algorithms run fully automatic and are easily distributable to compute the results in parallel on multiple computers.

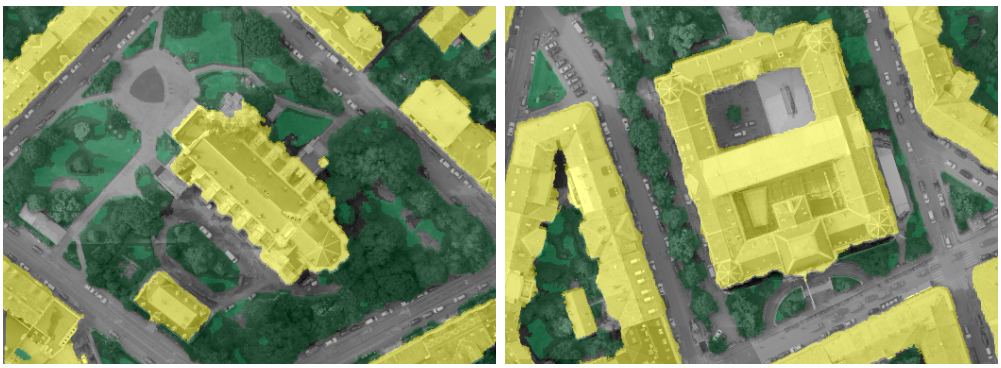
Different layers (building blocks, water regions, vegetation...) like those of a GIS are computed: For visualization and further processing those regions can be replaced by a vector (buildings blocks) or symbolic (vegetation) representation.



(a) (b)

(c) (d)

Fig. 15. Image (a) depicts a gray-scale ortho photo of Graz overlaid with the three color-encoded layers: vegetation is shown in green, buildings in yellow, water in blue and streets are left gray. (b) to (d) show each only one layer: (b) vegetation, (c) buildings and (d) streets.



(a) (b)

Fig. 16. Images (a) and (b) show two magnified areas from Figure 15: single trees are classified correctly, notable by the two different shades of green (grassland and trees).

Currently we work on the integration of edge-based information to improve the quality of the DSM on sharp height discontinuities.

9 Acknowledgements

This work has been done in the VRVis research center, Graz/Austria (http://www.vrvis.at), which is partly funded by the Austrian government research program Kplus. We would also like to thank Vexcel Corporation (http://www.vexcel.com) for supporting this project.

References

[1] F. Leberl and J. Thurgood. The promise of softcopy photogrammetry revisited. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXV, 2004.

[2] F. Leberl, R. Perko, and M. Gruber. Color in photogrammetric remote sensing. In *Proceedings of the ISPRS Commission VII Symposium*, volume 34, pages 59–64, 2002.

[3] Farhad Samadzadegan, Ali Azizi, Michael Hahn, and Curo Lucas. Automatic 3d object recognition and reconstruction based on neuro-fuzzy modelling. *ISPRS Journal of Photogrammetry and Remote Sensing*, 59:255–277, 2005.

[4] C. Huang, L.S. Davis, and J.R.G. Townshend. An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing 2002*, 23(4):725–749, 2002.

[5] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[6] C.-C. Chang and C.-J. Lin. *LIBSVM: a Library for Support Vector Machines*. National Taiwan University, 2005.

[7] C.-.W. Hsu, C.-C. Chang, and C.-J. Lin. *A Practical Guide to Support Vector Classification*. National Taiwan University, Department of Computer Science and Information Engineering, Taipei 106, Taiwan, 2003.

[8] J. Thurgood, M. Gruber, and K. Karner. Multi-ray matching for automated 3d object modeling. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXV, 2004.

[9] J. Bauer, H. Bischof, A. Klaus, and K. Karner. Robust and fully automated image registration using invariant features. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2004.

[10] David G. Lowe. Object recognition from local scale-invariant features. In *Proc. of the International Conference on Computer Vision ICCV, Corfu*, pages 1150–1157, 1999.

[11] Dennis Tell and Stefan Carlsson. Combining appearance and topology for wide baseline matching. In *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part I*, pages 68–81, London, UK, 2002. Springer-Verlag.

[12] D. Nister. An efficient solution to the five-point relative pose problem. *Computer Vision and Pattern Recognition 2003*, pages 195–202, 2003.

[13] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision 2002*, 47:7–42, 2002.

[14] R. T. Collins. A space-sweep approach to true multi-image matching. Technical Report UM-CS-1995-101, 1995.

[15] C. Zach, A. Klaus, and K. Karner. Accurate dense stereo reconstruction using 3d graphics hardware. *Eurographics 2003*, pages 227–234, 2003.

[16] F. Roli and J. Kittler. Fusion of multiple classifiers. *Information Fusion 2002*, 3:243, 2002.

[17] R.Bolter. *Buildings from SAR: Detection and Reconstruction of Buildings from Multiple View High Resolution Interferometric SAR Data*. PhD thesis, TU Graz, 2001.