IMAGE-BASED BUILDING CLASSIFICATION AND 3D MODELING WITH SUPER-PIXELS

Stefan Kluckner and Horst Bischof

Institute for Computer Graphics and Vision
Graz University of Technology, Austria
{kluckner,bischof}@icg.tugraz.at

Commission III/3

KEY WORDS: image classification, detection, aerial, building, modelling, colour, DEM/DTM

ABSTRACT:

Due to an increasing amount of aerial data there is significant demand in automatic large-scale modeling of buildings. This work presents an image-driven method for automatic building extraction and 3D modeling from large-scale aerial imagery. We introduce a fast unsupervised segmentation technique based on super-pixels. Considering the super-pixels as smallest units in the image space, these regions offer important spatial support for an information fusion step and enable a generic modeling of arbitrary building footprints and rooftop shapes. In our three-staged approach we integrate both appearance information and height data to accurately classify building pixels and to model complex rooftops. We apply our approach to datasets, consisting many overlapping aerial images, with challenging characteristics. The classification pipeline is evaluated on ground truth data in terms of correctly labeled pixels. We use the building classification together with color and height for large-scale modeling of buildings.

1 INTRODUCTION

Efficient building classification and 3D modeling from aerial imagery have become very popular in computer vision and photogrammetry due to a rapidly increasing number of applications like urban planning, navigation support, cartography, synthetic or realistic 3D city construction etc. In particular, Internet driven initiatives such as *Google Maps* and *Bing Maps* push the development of efficient, accurate and automatic methods. With the success of the aerial imaging technology, high resolution images can be obtained cost-efficiently. Multiple types of source data such as color or infrared images become available. For instance, the *Microsoft Ultracam* takes multi-spectral images in overlapping strips, providing high redundancy, which adheres every visible spot of urban environments from many different camera viewpoints. The high redundancy within the collected data enables image-based methods for automatic height field generation (Klaus et al., 2006), which offers important support for land-use classification (Zebedin et al., 2006, Kluckner et al., 2009) and 3D modeling of urban environments (Parish and Müller, 2001, Zebedin et al., 2008, Lafarge et al., 2010). Nevertheless, the enormous amount of data, including e.g. color and height information, requires fast methods and sophisticated processing pipelines getting by with a minimum of human interaction.

Considering large scale computation the problem of building extraction in urban environments becomes very difficult for many reasons. Buildings are complex objects with many architectural details, shape variations and a large diversity of appearance. In addition, buildings are located in urban scenes that contain various objects from man-made to natural ones. Therefore, recent approaches heavily differ in the use of data sources, extracted feature types and the applied models. A couple of recently proposed methods exploit 3D information provided by LIDAR data (Matei et al., 2008, Poullis and You, 2009), but already early approaches (Bignone et al., 1996, Cord et al., 1999) used a combination of 2D and 3D information for building extraction and modeling.

An increasing number of methods are based on digital surface models (DSM), directly generated from redundant images. Lafarge et al. (Lafarge et al., 2008) detected rectangular building footprints in DSMs and used symmetry criteria to roughly estimate the geometry of rooftops. In (Lafarge et al., 2010) the authors extended this approach with a library of 3D blocks for improved building generalization from single DSMs. These blocks can be seen as pieces sticked together for building construction and have to be given in advance. In contrast to exploiting a given number of designed models, we consider individual image regions, provided by super-pixel segmentation, as the smallest units representing building parts. While Taillandier (Taillandier, 2005) exploited cadastral maps, aerial images and a DSM, Vosselman and Dijkman (Vosselman and Dijkman, 2001) reconstructed rectangular shaped buildings from points clouds and given ground plans by detecting line intersections and discontinuities between planar faces. More generally, Zebedin et al. (Zebedin et al., 2008) proposed a concept based on fusion of feature and area information for building modeling. The method relies on directly extracting geometric prototypes such as planes and surfaces of revolution, taking into account height data, 3D lines and an individual building mask. A graph cut based optimization procedure refines the final result to form piecewise planar rooftop reconstructions. Other methods additionally involve classification techniques to automatically distinguish between mapped objects. Matikainen et al. (Matikainen et al., 2007) employed a DSM segmentation and a color-driven classification to discriminate buildings from trees. In (Zebedin et al., 2006) the authors fused information from redundant multi-spectral aerial images to generate orthographic images for color, height and land-use classification. Related, in our previous work (Kluckner et al., 2009) we proposed a rapid per-image semantic classification based on statistical description of appearance cues and 3D elevation measurements.

In this work, we focus on efficient, fully image-driven building classification and synthetic 3D modeling in large-scale aerial imagery by using both color and height information as input sources. The main contributions of our work are: We introduce an unsupervised segmentation technique based on super-pixels for generic rooftop construction. Super-pixels are images regions, describing the smallest unit in the image space and are not limited to predefined sizes or shapes. Therefore, a set of super-pixels enables

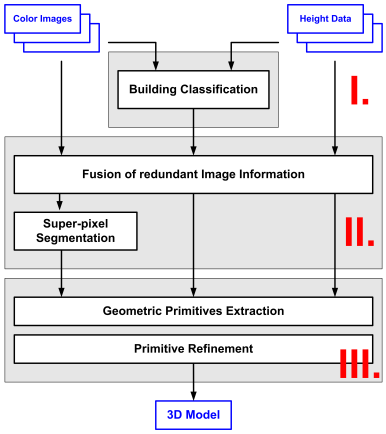


Figure 1: Overview of our classification and modeling approach: We use overlapping color images and height information to detect and construct 3D building models.

a composition of any building footprint. Together with an estimated plane (we exploit the height information), each super-pixel is used to form a part of a rooftop. A final refinement step yields a piece-wise planar approximation taking into account adjacent super-pixels. Our approach exceedingly exploits the redundancy in the data in order to remove outliers in the height data or to improve classification accuracy by image fusion. Apart from some human interaction to label training maps for learning the classifier, the proposed method runs fully automatic with a low number of parameters to adjust.

2 OVERVIEW

Our modeling pipeline is summarized in Figure 1. We consider highly overlapping color aerial images and two types of derived height fields as input sources: A dense matching approach (Klaus et al., 2006) provides corresponding depth information, defining a DSM in a 3D coordinate system, for each pixel in the input images. A digital terrain model (DTM), representing the bald earth, is computed in advance from the DSM by using a similar approach as described in (Champion and Boldo, 2006).

First, we perform building classification in order to obtain an initial interpretation at the pixel level for each image in the dataset (Sec. 3). Due to the high overlap in the aerial imagery, each point on the ground has several (up to ten) class probabilities. Similar as proposed in (Kluckner et al., 2009) we exploit statistical features in combination with random forests (RF) (Breiman, 2001) as classifiers to compactly describe and classify appearance and elevation measurements within local image regions. This classification technique involves a supervised training of the RF in advance using some labeled training maps. Learning a classifier, which discriminates building structures from the background, keeps the approach general and does not require a specific parameter tuning.

Second, a pixel-wise fusion step of multiple views into a common 3D coordinate system generates redundant image tiles for various source modalities like building classification, color and height information (Sec. 4). Following recent trends of integrating unsupervised image segmentation techniques for recognition tasks (Malisiewicz and Efros, 2007, Pantofaru et al., 2008, Fulkerson et al., 2009) we exploit super-pixels (Vedaldi and Soatto, 2008) to improve the fusion of different input sources and to reduce computational complexity for subsequent processing steps.

The third step of our approach involves the generic rooftop construction taking into account the super-pixels, which can be seen as footprint for parts of a building, and fused classification results (Sec. 5). For each building super-pixel, corresponding height data is used to extrude the individual footprints for geometric primitive generation. A spectral clustering step (Frey and Dueck, 2007) detects representative geometric prototypes, which are then used in a refinement step to form spatially consistent piecewise planar rooftops. To show the performance we apply our approach to two different datasets with challenging characteristics (Sec. 6).

3 BUILDING CLASSIFICATION

The first processing step of our approach involves a building classification on each image in the aerial dataset. Due to independent processing of each image this step can be done in a highly parallelized manner. The classification procedure yields class probabilities for each pixel in the processed images by computing statistics over low-level feature cues within small spatial neighborhoods. Due to efficiency we apply RF classifiers (Breiman, 2001) to compute initial building likelihoods at the pixel level.

Random forests are a powerful yet simple method to classify feature vectors by using simple attribute comparisons. In addition, RFs can handle label noise and errors in labeled training data. A forest can be seen as a collection of many random decision trees. The decision nodes of each tree include fast binary splits that give the direction of branching left and right down the tree until a leaf node is reached. By using a greedy optimization strategy the split criteria are learned from a subset of provided input data (which speeds up the training process). After tree construction using the subset of training samples, each tree is refined with the complete set of feature instances in order to generate the final leaf node's class distributions. This technique enables a sophisticated handling of large amount of data and further improves the generalization capability. At runtime, the classifier is evaluated by parsing down a test feature vector in each tree in the forest and accumulating the class likelihoods in the reached leaf nodes.

Each feature instance (P_i, c_i) consists of a computed region descriptor $\mathbf{P}_i$ and a target label $c_i \in \{building, non\text{-}building\}$ directly extracted from training maps. Tuzel et al. (Tuzel et al., 2006) presented a compact descriptor based on local statistics for rapid object detection by exploiting integral structures. A covariance matrix provides a low-dimensional and simple integration of d low-level feature cues. For instance, using a combination of color and height data separates the street regions from gray-valued rooftops or distinguishes between green areas and trees. The diagonal elements of the covariance matrix are the variances of the feature attributes in one channel, whereas the off diagonal elements capture the correlation values between the involved modalities. Thus the statistics up to second order of collected feature vectors can be represented by a mean vector $\mu \in \mathbf{R}^d$ and a covariance matrix $\Sigma \in \mathbf{R}^{d \times d}$. The space of covariance matrices is not a vector space, therefore, simple arithmetic differences between the elements do not measure the real distance between two matrices. Thus covariance descriptors cannot be directly applied to an RF, where simple attribute comparisons are used to construct the classifier. Instead of exploiting manifolds (Tuzel et al., 2006) to obtain a valid covariance similarity measurement, we use Sigma Points (Kluckner et al., 2009), which represent individual covariance matrices directly on Euclidean vector space. The idea relies on extracting specific samples of a given distribution, characterized by μ and Σ, and offers a simple concept for combining first and second order statistics, since the mean vector describes an offset in the Euclidean vector space. We construct

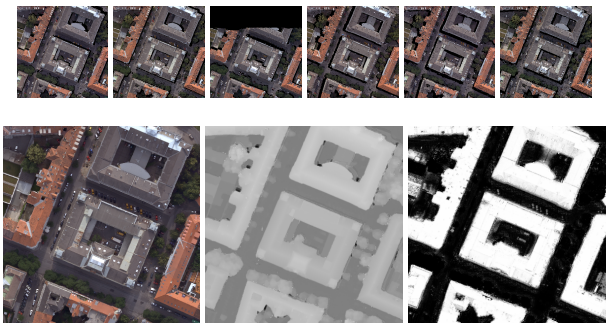


Figure 2: Fusion result: The first row shows six redundant orthographic views of a scene taken from *Graz*. Fused image results are given in the second row for color, height and building classification. Undefined areas are considerably compensated by using the high redundancy.

the set of Sigma Points[1] as follows:

$$\mathbf{p}_0 = \mu \quad \mathbf{p}_i = \mu + \alpha(\sqrt{\Sigma})_i \quad \mathbf{p}_{i+d} = \mu - \alpha(\sqrt{\Sigma})_i, \qquad (1)$$

where $i = 1 \dots d$ and $(\sqrt{\Sigma})_i$ defines the i-th column of the required matrix square root. Due to symmetry of the covariance matrix, we apply the Cholesky factorization to efficiently compute the matrix square root of Σ. The term α defines a weighting for the elements in the covariance matrix and is set to $\alpha = \sqrt{2d}$ as suggested in (Kluckner et al., 2009). Then, a resulting region descriptor $\mathbf{P} = \{\mathbf{p}_0, \dots, \mathbf{p}_{2d}\}$ consists of $2d + 1$ concatenated Sigma Points $\mathbf{p}_i \in \mathbf{R}^d$ and has a dimension of $\mathbf{P} \in \mathbf{R}^{d(2d+1)}$. For details we refer to (Kluckner et al., 2009). The next section describes the fusion of redundant information into a common 3D coordinate system.

4 FUSION OF MULTIPLE IMAGES

Because of the high overlap in the aerial imagery, each point on ground is mapped multiple times from different viewpoints. Since we are interested in large-scale modeling, we generate an orthographic image from many overlapping perspective images by a pixel-wise transformation into a common 3D coordinate system. Taking into account camera data and depth information, provided by a dense matching procedure, corresponding pixels in the perspective images yield multiple observations for color, height and building classification in the orthographic view. Several rectified observations of a scene taken from the imagery *Graz* are shown in Figure 2.

The fusion of redundant information into a common view has the benefit that e.g. reconstruction errors caused by non-stationary objects like moving cars can be compensated. In addition, a projection of many different views produces an orthographic image without undefined image regions caused by perspective occlusions. First, color and height information are fused by computing median values for each pixel from multiple observations. In case of robustly fusing color information per pixel, we use random projections of the color vector onto 1D lines to detect the median of vector-valued data (Tukey, 1974). Though simple mean computation has lower computational complexity, a median will not introduce new colors values as possibly introduced by averaging. In addition, an accurate fused color image is essential for super-pixel segmentation performed at the next step. In order to estimate a final building likelihood for each pixel in the orthographic view, confidences from different views are accumulated and normalized. Figure 2 depicts the final pixel-wise fusion result for color, height and building classification. In the next step we briefly discuss super-pixels and introduce an optimization stage to refine the classification and the prototype labeling on a super-pixel neighborhood.

[1] Code available at http://www.icg.tugraz.at/Members/kluckner

4.1 Super-Pixel Segmentation

A variety of recently proposed methods obtaining state-of-the-art performance on benchmark datasets integrate unsupervised image segmentation methods into classification or object detection. Several approaches utilize multiple segmentations (Malisiewicz and Efros, 2007, Pantofaru et al., 2008) however the generation of many partitions induces enormous computational complexity and is impractical for aerial image segmentation. Recently, Fulkerson et al. (Fulkerson et al., 2009) proposed to use super-pixels, rapidly generated by Quickshift (Vedaldi and Soatto, 2008). These super-pixels accurately preserve object boundaries of natural and man-made objects. Applying Quickshift super-pixel segmentation to our approach offers several benefits: First, computed super-pixels can be seen as the smallest units in the image space. All subsequent processing steps can be performed on a reduced adjacency graph instead of incorporating the full pixel image grid. Furthermore, we consider super-pixels like homogeneous regions providing important spatial support: Due to edge preserving capability, each super-pixel describes a part of only one class, namely *building* or *non-building*. Aggregating data, such as classification and height information, over the pixels defining a super-pixel compensates for outliers and erroneous pixels. For instance, an accumulation of building likelihoods results an improved building classification for each segment. A color averaging within small regions synthesizes the final modeling results and significantly reduces the amount of data. More importantly, we exploit super-pixels, which define parts of the building footprints, for the 3D modeling procedure. Taking into account a derived polygon approximating of the boundary pixels and corresponding height information, classified building footprints can be extruded to form any type of geometric 3D primitives. Therefore, introducing super-pixels for footprint description allows to model any kind of ground plan and in the following the rooftop.

4.2 Refined Labeling using Super-Pixels

Although aggregating the fused building classification or extracting geometric prototypes using super-pixels capture some local information, the regions in the image space are handled independently. In order to incorporate spatial dependencies between nodes defined on the image grid, e.g. *Markov* random field formulations (Boykov et al., 2001) are widely used to enforce an evident final class labeling. In contrast to minimizing the energy on a full image grid (Pantofaru et al., 2008, Kluckner et al., 2009) we apply a conditional *Markov* random field (CRF) stage defined on the super-pixel neighborhoods similar as proposed in (Fulkerson et al., 2009). In our approach we apply the refinement on super-pixels twice: First, we apply the CRF to provide a smooth labeling of the building class taking into account the spatial dependency on an adjacency graph. Second, in a separate processing step the CRF is used for consistent labeling of the geometric prototypes to enforce a piecewise planar rooftop.

Let $G(S, E)$ be an adjacency graph with a super-pixel node $s_i \in S$ and a pair $(s_i, s_j) \in E$ be an edge between the segments s_i and s_j, then an energy can be defined with respect to the class labels $\mathbf{c}$. In this work a label can be a *building/non-building* class or a possible assignment to a specific geometric primitive. Generally,

the energy can be defined as

$$E(\mathbf{c}|G) = \sum_{s_i \in S} D(s_i|c_i) + \omega \sum_{(s_i,s_j)\in E} V(s_i, s_j|c_i, c_j), \quad (2)$$

where $D(s_i|c_i)$ expresses the unary potential of a super-pixel node. In case of the classification refinement, $\mathbf{c}$ represents a binary labeling of the adjacency graph that assigns each graph node s_i a label $c_i \in \{\text{building}, \text{non-building}\}$. The unary potential $D(s_i|c_i) = -\log(H(s_i))$ denotes the class likelihoods $H(s_i)$ of a super-pixel s_i obtained by aggregating pixel-wise confidences. The costs for geometric primitive refinement are described in the next section. The factor ω controls the influence of the regularization and is estimated by using cross validation. In order to consider the region sizes in the minimization process, we compute the pairwise edge term $V(s_i, s_j|c_i, c_j)$ between the super-pixels s_i and s_j with

$$V(s_i, s_j|c_i, c_j) = \frac{b(s_i, s_j)}{1 + g(s_i, s_j)} \delta(c_i \neq c_j). \quad (3)$$

The function $b\,(s_i, s_j)$ computes the number of common boundary pixels of two given segments, $g(s_i, s_j)$ is the L^2 norm of the mean color distance vector and $\delta(\cdot)$ is a simple zero-one indicator function. In this work we minimize the energy defined in Equation 2 by using α-expansion moves (Boykov et al., 2001).

5 BUILDING MODELING

A generation of super-pixels provides footprints for any object in an observed color image. Taking into account the refined building classification and additional height information, 3D geometric primitives describing the smallest unit of a building rooftop can be extracted as the next step. Estimated rooftop hypotheses for each super-pixel in a building (we simply extract connected components on the adjacency graph) are collected and clustered in order to find representative rooftop prototypes. Finally, a CRF optimization assigns consistently the prototypes to each super-pixel in a building considering resulting reconstruction error and neighborhood segments.

5.1 Prototype Extraction

Assuming a set of super-pixels (a super-pixel can be seen as a list of coordinates), classified as parts of an individual building, we initially fit planes to the available corresponding point clouds provided by the fused DSM. In this work we use planes as geometric primitives however the prototype extraction can be extended to any kind of primitives. We apply RANSAC over a fixed number of iterations to find those plane, minimizing the distance to the point cloud, for each building super-pixel. This procedure yields a rooftop hypothesis for each super-pixel defined by a normal vector and single point on the estimated plane (see second row of Figure 3).

5.2 Prototype Clustering and Refinement

As a next step, we introduce a clustering of hypotheses for two reasons: Since the subsequent optimization step can be seen as a prototype labeling problem, similar 3D primitives should provide same labels in order to result a smooth reconstruction of a rooftop. Second, clustering significantly reduces the number of probable labels which benefits the efficiency of the optimization procedure. We apply affinity propagation (Frey and Dueck, 2007) to find representative exemplars of 3D primitives. Affinity propagation takes as input a distance matrix of pairwise similarity measurements and efficiently identify a set of exemplars. Please note that the number of exemplars has not to be given in advance and the similarity matrix can be computed sparsely. We therefore construct the similarity matrix as follows: For each 3D primitive which consists of plane and a 3D point in space, we estimate the reconstruction error for adjacent super-pixels taking into account the current prototype hypothesis and the set of neighboring height data points. Considering only adjacent image regions additionally reduces computational costs for constructing the similarity matrix. The clustering procedure yields a set of representative primitive prototypes which are used to approximate a rooftop shape with respect to the available height information. Next, we reuse the formulation of the energy defined in Eq. 2 to obtain a consistent prototype labeling for building regions. In case of geometric primitive refinement, $\mathbf{c}$ represents a labeling of the adjacency graph that assigns each super-pixel s_i a label $c_i \in T$, where T is the set of geometric prototypes obtained by clustering. Similar as proposed in (Zebedin et al., 2008), the unary potential $D(s_i|c_i)$ denotes the costs, in terms of summed point-to-plane distance measurements, of s_i being assigned the label c_i or prototype, respectively. We compute the pairwise edge term considering appearance and super-pixel sizes in order to obtain a smooth geometric prototype labeling within homogeneous building areas. A refined labeling of prototypes is shown in Figure 3.

5.3 Rooftop Modeling

So far the footprint of each building consists of a set of super-pixels in the image space. In order to obtain a geometric footprint modeling of each super-pixel, we first identify common boundary pixels between adjacent building super-pixels. For each super-pixel, this procedure results a specific set of boundary fragments, which can be individually approximated by straight line segments. A pairwise matching of collected line segments yields a closed yet simplified 2D polygon. Taking account of DTM and the refined geometric primitive assignment, the footprint polygons defined by a number of vertexes are extruded to form small units of a rooftop: distinctive 3D rooftop points are determined by intersecting the plane (given by the geometric primitive) with a line, directed to $(0, 0, 1)^T$, going through the corresponding vertex on ground. For the purpose of visualization, we use a 2D Delaunay triangulation technique to generate the models of the buildings. An individual 3D building model of our approach can be seen as a collection of composed building super-pixels having identical building and rooftop prototype indexes, respectively. A hierarchical grouping of super-pixels could be used to further simplify the resulting building model.

6 EXPERIMENTS

This section evaluates our proposed framework on a large amount of real world data. We first describe the aerial imagery, then the building classification is evaluated on hand-labeled ground truth data. Moreover, we present results of our building generalization and perform quantitative and visual inspection of the constructed models.

Data. We present results for two aerial imageries showing different characteristics. The dataset *Graz* (155 images) shows a colorful appearance with challenging buildings and *San Francisco* (77 images) has suburban occurrence in a hilly terrain. The imageries are taken with the *Microsoft Ultracam* in overlapping strips (80% along-track overlap and 60% across-track overlap), where each image has a resolution of 11500×7500 pixels with a ground sampling distance of approximately 10 cm. We use the color images, the height data computed by dense matching (Klaus

	overall	*building*	*non-buil.*
Graz, pixel level	88.5	90.5	87.3
Graz, with super-pixel	90.6	92.1	90.0
Graz, with CRF	**93.7**	**92.1**	**93.4**
San Fran., pixel level	85.7	86.3	85.3
San Fran., with super-pixel	89.2	89.0	91.8
San Fran., with CRF	**92.1**	**91.8**	**93.4**

Table 1: Building classification accuracy in terms of correctly classified pixels on hand-labeled orthographic test data. It can be clearly seen that use of super-pixels as spatial support improves accuracy. The CRF stage further improves the classification rates using a consistent final labeling of the super-pixels.

et al., 2006) and the derived DTM (Champion and Boldo, 2006). A combination of DTM and DSM yields absolute elevation measurements per pixel from ground which are applied for the building classification and modeling.

Building Classification. For all datasets, we train individual RF classifiers with 8 trees and a maximum depth of 14. The Sigma Points feature vectors are collected within small image patches (11×11 pixels). In this work the Sigma Points describe the statistics of feature cues like color, texture and elevation measurements within small image patches. Texture information is directly obtained by computing first order derivatives on the L channel of CIELab color images. A combination of the color channels, two gradients and the elevation measurements yields a feature vector with 78 attributes, which can be directly trained and evaluated using the RF classifiers.

In our approach we exploit hand-labeled ground truth maps for training of the classifiers. Please note that the labeling of training data involves some human interaction, but since our approach works at the pixel level there is no need to accurately label complete building areas. Hence the labeling of the training data is straightforward and can be efficiently done by applying brush strokes representing either *building* or *non-building* class. For evaluation we additionally label randomly selected orthographic images (we use 9 tiles per dataset). Obtained classification rates are summarized in Table 1. We report both the overall per-pixel classification rate (i.e. the accuracy of all pixels correctly classified) and the average of class specific per-pixel percentages, which gives a more significant measurement due to varying quantity of labeled pixels for each class. On both datasets we obtain overall classification rates of more than 90%. A classification of a single aerial image at full resolution takes approximately 3 minutes on a dual core machine.

The fusion step for color, height and classification, also including the super-pixel generation, of 6 different viewpoints covering an area of 150×150 meters lasts less than 5 minutes. Quickshift is applied to a vector consisting of pixel location and CIELab color. The parameters for Quickshift are set to $\sigma = 2$ and $\tau = 8$. It turned out that these parameters capture nearly all object boundaries in some observed test images. In addition, the parameters generates sufficiently small regions in order to preserve curved boundary shapes. The overall results, adding a CRF stage for classification refinement are given for $\omega = 3.0$.

Figure 3 shows a result for a fused image tile of *Graz*. While the raw pixel-wise fusion of the class probabilities shows higher granularity and blurred object boundaries due to inaccurate 3D information (compare to Figure 2), an integration of super-pixels and CRF improves the final building classification significantly.

Building Modeling. We use the proposed method to model complex rooftops of buildings in 3D. Figure 3 shows a modeling result for a part of *Graz*. In order to obtain a quantitative evaluation

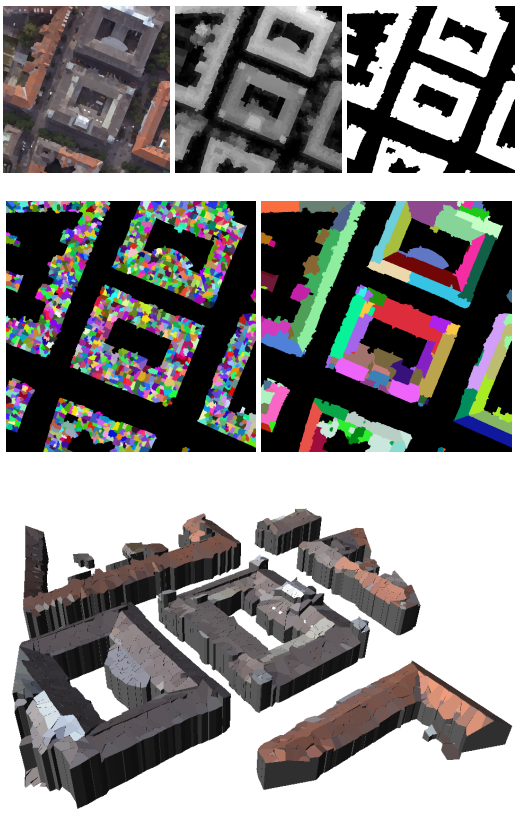


Figure 3: Results for a small part of *Graz*. The first row depicts the input sources like color, elevation measurements and refined classification, aggregated within super-pixels. The second row shows computed super-pixels overlaid with the building mask and the result of the refinement step which groups super-pixels by taking into account the geometric primitives. In the bottom the corresponding constructed 3D building model is given.

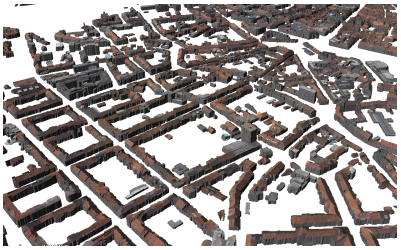
the root mean squared error (RMSE) over all building pixel is computed between fused DSM values and the heights obtained by 3D modeling. For *Graz* we obtain an RMSE of 1.9 meters taking into account all $170.0e6$ building pixels. For *San Francisco* the RMSE is 1.7 meters evaluated on $210.0e6$ pixels. In case of prototype refinement the parameter ω controls the fidelity between the degree of details and geometric simplification. For both datasets the smoothing factor with $\omega = 5.0$ has given reliable results.

In Figure 4 computed 3D models are shown for *San Francisco* and *Graz*. For efficiency and large-scale capability we compute such models in tiles of 1600×1600 pixels. Given the fused color including super-pixel segmentation, height and classification images, the 3D model of *Graz* can be computed within an hour using a subsequent processing.

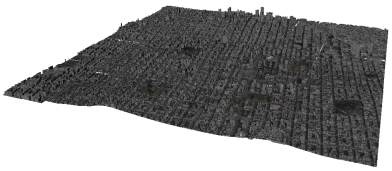
7 CONCLUSION

We have proposed an efficient, purely image-driven approach for constructing synthetic 3D models of buildings by exploiting redundant color and height information. First, an efficient classification at the pixel level has been introduced to separate buildings from the background. A pixel-wise fusion step integrates different modalities from multiple viewpoints into a common orthographic view. In particular, involving a super-pixel segmentation enables a generic modeling of any building rooftop shape and reduces the problem of outliers and computational complexity. We

(a) 3D Model of *Graz*



(b) A detail of *Graz*



(c) 3D model of *San Francisco*

(d) A detail of *San Francisco*

Figure 4: 3D building models of *Graz* and *San Francisco*. The model of *Graz* covers an area of about 4 sqkm, while *San Francisco* has a dimension of 2500×2500 meters. Such models can be constructed within a couple of hours on a standard PC. For *San Francisco* we overlaid the 3D visualization with a triangulated DTM.

applied our approach to two different imageries and demonstrated large-scale capability with low time consumption. Future work will concentrate on handling levels of details and a visualization with e.g. procedural modeling engines like CityEngine (Parish and Müller, 2001). In addition, we will extend our modeling pipeline to other object classes like trees. A direct comparison to GIS data will give an improved indication of accuracy.

ACKNOWLEDGEMENTS

This work was financed by the FFG Project APAFA (813397) and the Austrian Science Fund Project W1209 under the doctoral program Confluence of Vision and Graphics.

REFERENCES

Bignone, F., Henricsson, O., Fua, P. and Stricker, M., 1996. Automatic extraction of generic house roofs from high resolution aerial imagery. In: Proceedings ECCV, pp. 83–96.

Boykov, Y., Veksler, O. and Zabih, R., 2001. Efficient approximate energy minimization via graph cuts. PAMI 20(12), pp. 1222–1239.

Breiman, L., 2001. Random forests. In: Machine Learning, pp. 5–32.

Champion, N. and Boldo, D., 2006. A robust algorithm for estimating digital terrain models from digital surface models in dense urban areas. In: Proceedings ISPRS, on CD-ROM.

Cord, M., Jordan, M., Cocquerez, J.-P. and Paparoditis, N., 1999. Automatic extraction and modelling of urban buildings from high resolution aerial images. In: Proceedings ISPRS Automatic Extraction of GIS Objects from Digital Imagery, pp. 187–192.

Frey, B. J. and Dueck, D., 2007. Clustering by passing messages between data points. Science 315, pp. 972–976.

Fulkerson, B., Vedaldi, A. and Soatto, S., 2009. Class segmentation and object localization with superpixel neighborhoods. In: Proceedings ICCV, on CD-ROM.

Klaus, A., Sormann, M. and Karner, K., 2006. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. Proceedings ICPR, on CD-ROM.

Kluckner, S., Mauthner, T., Roth, P. M. and Bischof, H., 2009. Semantic classification in aerial imagery by integrating appearance and height information. In: Proceedings ACCV, on CD-ROM.

Lafarge, F., Descombes, X., Zerubia, J. and Pierrot Deseilligny, M., 2008. Automatic building extraction from dems using an object approach and application to the 3d-city modeling. IJPRS 63(3), pp. 365–381.

Lafarge, F., Descombes, X., Zerubia, J. and Pierrot-Deseilligny, M., 2010. Structural approach for building reconstruction from a single dsm. PAMI 32(1), pp. 135–147.

Malisiewicz, T. and Efros, A. A., 2007. Improving spatial support for objects via multiple segmentations. In: Proceedings BMVC, on CD-ROM.

Matei, B., Sawhney, H., Samarasekera, S., Kim, J. and Kumar, R., 2008. Building segmentation for densely built urban regions using aerial lidar data. In: Proceedings CVPR, on CD-ROM.

Matikainen, L., Kaartinen, K. and Hyyppä, 2007. Classification tree based building detection from laser scanner and aerial image data. IAPRS 36(3), pp. 280–287.

Pantofaru, C., Schmid, C. and Hebert, M., 2008. Object recognition by integrating multiple image segmentations. In: Proceedings ECCV, on CD-ROM.

Parish, Y. I. H. and Müller, P., 2001. Procedural modeling of cities. In: Proceedings SIGGRAPH, pp. 301–308.

Poullis, C. and You, S., 2009. Automatic reconstruction of cities from remote sensor data. In: Proceedings CVPR, on CD-ROM.

Taillandier, F., 2005. Automatic building reconstruction from cadastral maps and aerial images. IAPRS 36, pp. 105–110.

Tukey, J. W., 1974. Mathematics and the picturing of data. In: Proceedings International Congress of Mathematics, Vol. 2, pp. 523–531.

Tuzel, O., Porikli, F. and Meer, P., 2006. Region covariance: A fast descriptor for detection and classification. In: Proceedings ECCV, on CD-ROM.

Vedaldi, A. and Soatto, S., 2008. Quick shift and kernel methods for mode seeking. In: Proceedings ECCV, on CD-ROM.

Vosselman, G. and Dijkman, S., 2001. 3d building model reconstruction from point clouds and ground plans. IAPRS 34, pp. 37–44.

Zebedin, L., Bauer, J., Karner, K. and Bischof, H., 2008. Fusion of feature- and area-based information for urban buildings modeling from aerial imagery. In: Proceedings ECCV, on CD-ROM.

Zebedin, L., Klaus, A., Gruber-Geymayer, B. and Karner, K., 2006. Towards 3d map generation from digital aerial images. IJPRS 60, pp. 413–427.