LARGE-SCALE AERIAL IMAGE INTERPRETATION USING A REDUNDANT SEMANTIC CLASSIFICATION

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: classification, interpretation, aerial, colour, DEM/DTM, high resolution

ABSTRACT:

This work introduces an efficient classification pipeline, which provides an accurate semantic interpretation of urban environments by using redundant scene observations. The image-based method integrates both appearance and height data to classify single aerial images. Given the initial classification of highly overlapping images, a projection to a common orthographic 3D world coordinate system provides redundant observations from multiple viewpoints and enables a semantic interpretation of large-scale urban environments. In the experimental evaluation we investigate how the use of redundancy influences the accuracy in terms of correctly classified pixels for object classes like building, tree, grass, street and water areas. Moreover, we exploit an efficient yet continuous formulation of the Potts model to obtain a consistent labeling of the pixels in the orthographic view. We present results for the datasets *Dallas* and *Graz*.

1 INTRODUCTION

Semantic image interpretation in large-scale aerial imagery, also referred to as land-use classification, has become very popular due to an increasing number of applications like urban planning, navigation support, cartography, automatic 3D city modeling etc. Recent developments in the aerial imaging technology enable an efficient mapping of urban environments from multiple viewpoints by using different modalities, like color, infrared or panchromatic information. In particular, the *Microsoft Ultracam* takes multispectral high resolution images in highly overlapping strips. The high redundancy within the data enables automatic height field computation (Klaus et al., 2006, Hirschmüller, 2006), but also methods for orthographic image generation (Strecha et al., 2008) by using multiple observations of the same scene. In Figure 1 a scene of *Graz* is shown from multiple viewpoints. In this work we focus on fusing semantic interpretation, obtained by efficiently classifying images taken from different viewpoints, in a common view.

Although appearance driven classification approaches (Shotton et al., 2006, Verbeek and Triggs, 2007, Gould et al., 2008) obtain reliable results on computer vision benchmark datasets such as the MSRC (Shotton et al., 2006) or the PASCAL VOC (Everingham et al., 2007) images, scene interpretation in a real world scale still poses an unsolved and challenging task due to a huge variability in images. Recently proposed methods additionally incorporate 3D information to further improve the classification accuracy. Several approaches, dealing with aerial imagery (Rottensteiner et al., 2004, Zebedin et al., 2006, Kluckner et al., 2009, Nguyen et al., 2010), exploit appearance cues together with elevation measurements (resulting from a combination of a surface and a terrain model). As shown in (Brostow et al., 2008, Xiao and Quan, 2009), an integration of color and 3D information, obtained by *Structure from Motion*, is essential for an accurate semantic interpretation of street-level images. While Brostow et al. (Brostow et al., 2008) handled each image in a sequence separately, the authors in (Xiao and Quan, 2009) combined the obtained classification from multiple views to improve the final semantic interpretation. In spirit similar to (Xiao and Quan, 2009), where the authors proposed to use overlapping street-level images, we introduce an image-based approach for aerial data, which takes redundant image information and height fields as input sources for full scene understanding. In contrast to our previous approach (Kluckner et al., 2009), where classification is performed on individual aerial images, this work focuses on the fusion of the computed interpretation in order to obtain a semantic classification of large-scale urban environments. The fusion of redundant information into a common view holds the benefit, that e.g. missing data, caused by occlusions and non-stationary objects like moving cars can be compensated. More importantly, each point on the ground at the orthographic view is classified multiple times, which enables an improved prediction of the real object class label. In this work we investigate how the redundancy can be exploited to obtain an accurate semantic interpretation in an orthographic view. We show that the classification accuracy can be improved by collecting redundant probabilities for each object class from multiple view points. Moreover, we use a continuous formulation of the Potts model (Pock et al., 2009) to obtain a spatially consistent semantic labeling. Due to the enormous amount of data, we put emphasis on multi-core friendly approaches and efficient data structures, like integral images (Viola and Jones, 2004).

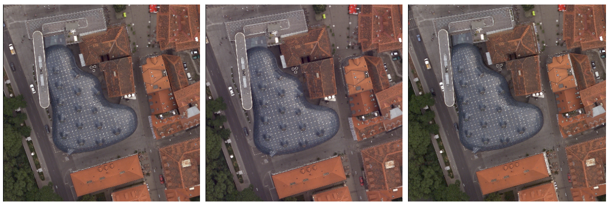

Figure 1: A scene of *Graz*: The *Kunsthaus* is taken from different viewpoints. The redundancy within the data is used to compute discriminative height information. We perform classification on each view in order to obtain an improved semantic interpretation by fusion in a common (orthographic) view.

This paper is organized as follows: In Section 2 we briefly outline required sources modalities, such as the height information. Section 3 highlights our semantic interpretation strategy for aerial images. Section 5 presents the experimental evaluation and Section 6 concludes our work and gives directions to future work.

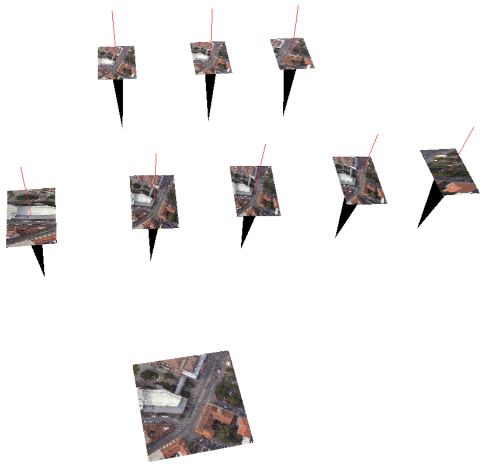

Figure 2: Overlapping camera positions mapping the same scene on ground. The derived range information and camera data are used to compute corresponding pixel locations in the perspective images.

2 SOURCE MODALITIES

This section briefly describes the required image sources. We consider highly overlapping color aerial images and two types of derived height fields as input sources. Figure 2 shows a small part of *Graz* observed from overlapping camera views. The high redundancy within the collected data enables an automatic dense matching method (Klaus et al., 2006), which provides range information of mapped scenes at the pixel level. The corresponding point cloud defines a digital surface model (DSM) in a 3D coordinate system. A digital terrain model (DTM), representing the bald earth, is directly derived from the DSM.

2.1 Digital Surface Model

The 3D data, representing the DSM, is computed from the high resolution aerial images provided by the *Microsoft Ultracam* camera. In order to enable robust and fully automatic processing of the data, the high inter-image redundancy is ensured by capturing images at 80% along-track overlap and 60% across-track overlap. The exterior orientation of the images is achieved by an automatic process similar to the method described in (Zebedin et al., 2006). By using the camera orientation parameters, an area based matching algorithm produces a dense range image for each input image. The range images are computed from three input images (a reference image and its two immediate neighbors) with a plane sweeping approach. We use such triplets to compute the corresponding range information. The plane sweeping is based on the normalized cross correlation as similarity measurement and produces a 3D depth space, which contains the depth hypotheses and their associated correlation values. The final range image is generated by applying a semi-global optimization approach (Klaus et al., 2006).

2.2 Digital Terrain Model

The point cloud, provided by the dense matching process, is an uninterpreted representation of the surface. In order to estimate the bald earth, we separate the points into those that describe the terrain surface and those that represent elevated objects such as buildings and trees. Taking into account the derived surface model from neighboring views (we use an approximated median to efficiently fuse multiple views) and the camera data, a filtering strategy detects local minimums (defining points on ground) over different scales in the fused point cloud. Similar to (Unger et al., 2009), a total variation based in-painting strategy fills areas, formerly elevated by buildings and trees. Subtracting the DTM from the DSM delivers the absolute elevation measurements of the objects, which are used as a discriminative feature modality for our semantic classification procedure.

3 SEMANTIC CLASSIFICATION

The first processing step of our pipeline involves a pixel-wise semantic classification, performed on each image in the aerial dataset. In our approach we use the *Sigma Points* feature representation (Kluckner et al., 2009) together with efficient randomized forest classifiers (Breiman, 2001). A fusion step at the pixel level (simple matrix-vector multiplications) combines the individual results in an orthographic view.

3.1 Sigma Points Feature Representation

Since an aerial imagery usually consists of multiple information cues, such as color, infrared and elevation measurements there is a need to reasonably integrate these low-level feature types. For instance, using a direct combination of color and height data would successfully separate the street regions from gray-valued roof tops or would distinguish between grass-covered areas and trees. We therefore apply a statistical feature representation based on *Sigma Points* (Kluckner et al., 2009) to compactly describe the various cues considering a small local neighborhood defined around each pixel. Due to a nearly fixed object scale a classification at the pixel level is well suited for efficient interpretation of aerial images.

The *Sigma Points*[1] representation can be seen as an efficient approximation of first and second order statistics in Euclidean vector space and is directly derived from covariance matrix descriptors (Tuzel et al., 2006), which can be quickly computed for each pixel using integral image structures (Viola and Jones, 2004). 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. Compared to constructed feature vectors over multispectral data (Zebedin et al., 2006), *Sigma Points* descriptors are low-dimensional and enable a tight integration of appearance and height information in Euclidean vector space. Note that there is no need to quantize the height information into a given number of histogram bins by considering an estimation of the expected height values for each dataset. Moreover, this representation can be simply extended to integrate additional feature cues like more sophisticated filter responses or infrared channels. Then a resulting feature vector consists of $d(2d+1)$ elements, if d modalities are combined. For details we refer to (Kluckner et al., 2009).

3.2 Randomized Forest Classifier

Randomized forests (Breiman, 2001) offer a powerful method to classify collected feature vectors by using simple attribute comparisons (Shotton et al., 2008, Lepetit et al., 2005). Forests are inherently multi-class and can handle errors and noise in the labeled training data. We train the classifier in a supervised manner. A feature instance of the training set consists of a region description, represented by the *Sigma Points* and an assigned target class label.

A randomized forest consists of several binary decision trees, where each tree is composed of split and leaf nodes. The split

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

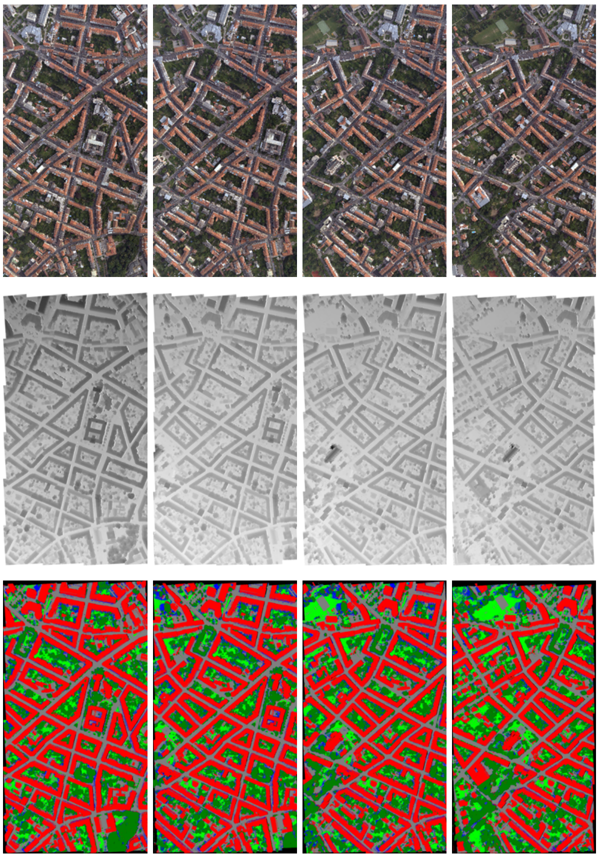

Figure 3: Overlapping aerial images: We exploit color (first row) and dense matching results (second row) to derive a highly redundant semantic classification (third row). Each image has a dimension of 11500×7500 pixels and can be processed within four minutes.

nodes are learned from a random subset of the training data (which speeds up the training process) by using a greedy strategy. Each split criterion then minimizes the weighted information gain (Shotton et al., 2008), considering the class distributions estimated from target labels falling into left and right children nodes. After construction by using the subset of training samples, each tree is refined with the complete set of feature vectors in order to generate the final leaf node class distributions. This technique enables a sophisticated handling of a large amount of data and improves the generalization capability (Shotton et al., 2008).

At runtime, the classifier is evaluated by parsing down a test vector in the forest and accumulating the label histograms in the reached leaf nodes, yielding an averaged class distribution for the corresponding feature representation. Note that the training and evaluation procedure can be accelerated by using multi-core systems as shown in (Sharp, 2008).

3.3 Fusion of Redundant Classification

Since we are interested in large-scale semantic interpretation, we introduce virtual orthographic cameras with a specified pixel resolution (we use a similar sampling distance as provided by the original images) in order to collect the information, provided in the overlapping perspective images in a common 3D coordinate system. Taking into account the available camera data and computed depth information, a simple projection of corresponding pixels (defined by pixel location and depth value) yields multiple observations for the semantic classification, but also for color and height in a common view. In order to estimate a class specific likelihood for each pixel in the orthographic view, class distributions from different views are accumulated and normalized. Figure 4 depicts the pixel-wise accumulation result obtained by our classification pipeline. Due to classification and projection at the pixel level, the result shows a high granularity with respect to the dominant class confidences. We therefore introduce an optimization step to obtain a consistent semantic labeling.

4 REFINED LABELING

Although our feature representation includes some local context information, collected within a small spatial neighborhood, each pixel in the fused image is treated almost independently. The problem of obtaining a smooth and spatially consistent labeling of the whole image scene can also be seen as a task of multi-class image segmentation, where each pixel value is selected from a predefined pool of class labels. In our case the labeling procedure is supported by fused class distributions.

In general, a segmentation problem into multiple object classes can be defined as a minimization of the Potts model (Potts, 1952)

$$\min_{l} \left\{ \lambda \sum_{p,q} V(l_p, l_q) + \sum_{p} D(l_p) \right\}, \tag{1}$$

where $V(l_p, l_q)$ denotes the pairwise interaction potential and $D(l_p)$ is the data term. We aim for finding a final labeling l that assigns each pixel p a class label $l_p \in L$, where the labeling is both piecewise smooth and consistent with the observed data.

Although the problem of multi-class image segmentation is NP-hard (a two class problem can be solved exactly), there exist several algorithms to compute a solution approximately (Boykov et al., 2001, Komodakis and Tziritas, 2007, Pock et al., 2009, Olsson et al., 2009).

Originally, the Potts model was introduced to model phenomena of solid state physics. The continuous formulation of the Potts model for an image domain $\Omega \in R^2$ can be written as

$$\begin{gathered} \min_{s_i} \left\{ \lambda \sum_{i=1}^{N} \mathrm{Per}(s_i; \Omega) + \sum_{i=1}^{N} \int_{s_i} l_i(p) dp \right\}, \\ \text{s.t.} \bigcup_{i=1}^{N} s_i = \Omega, \quad s_i \cap s_j = \emptyset \, \forall i \neq j \,, \end{gathered} \tag{2}$$

where $l_i \in R^2$ are the computed confidence maps for each object class and s_i is the resulting image partition, which corresponds to the object class i. The scalar value λ defines the fidelity between data term and regularization. The first term incorporates the length of the boundary of the segmentation s_i, while the second term considers the data at each point $p \in s_i$ in the segment s_i. In our case the confidence maps l_i with $i = 1, \ldots, N$ are directly obtained by accumulation and normalization of corresponding multiple observations of the semantic intepretation for $N = 5$ object classes.

In this work we use a convex relaxation, based on a primal dual formulation to minimize the energy defined in Equation 2 similar as proposed by Pock et al. (Pock et al., 2009). The minimization scheme exploits an efficient primal dual projected gradient algorithm, which can be accelerated with multi-core systems like

GPUs. According to (Pock et al., 2009), the minimization problem can be rewritten as a total variation functional

$$\min_{u_i} \left\{ \lambda \sum_{i=1}^{N} \int_{\Omega} \sqrt{\nabla u_i(p)^T g(p) \nabla u_i(p)}\; dp + \right. \\ \left. + \sum_{i=1}^{N} \int_{\Omega} u_i(p) l_i(p)\; dp \right\}, \quad (3)$$

where $u_i : \Omega \longrightarrow \{0, 1\}$ is a binary labeling function with $u_i(p) = 1$, if $p \in s_i$ and $u_i(p) = 0$ else.

The first term denotes the total variation of the functions u_i and describes the corresponding anisotropic length of the segment s_i. The term $g \in R^2$ denotes an edge penalty function, which enables a smooth labeling by taking into account strong edges, extracted from e.g. available color or height information.

The second term defines the data term, provided by the class confidence maps. Since the space of binary functions u_i forms a non-convex set, the functional cannot be directly minimized using a convex optimization strategy such as the efficient primal dual algorithm. Pock et al. (Pock et al., 2009) proposed to relax the set of binary functions to a set of functions $u_i : \Omega \longrightarrow [0, 1]$, which can take values between zero and one. They showed that this relaxation scheme yields globally optimal solutions in most practical problems. For details we refer to (Pock et al., 2009). In the experimental evaluation we show that the proposed refinement step significantly improves the classification accuracy.

5 EXPERIMENTS

This section describes experiments, which demonstrate the benefit of exploiting redundancy for semantic classification. We perform experiments on two aerial imageries with different characteristics. The dataset *Graz* (155 images) shows a colorful appearance with challenging building structures and many junctions between trees and rooftops. The second dataset shows part of *Dallas* (79 images) and contains mainly gray valued regions with large buildings. Each image in the datasets has a resolution of 11500×7500 pixels with a ground sampling distance of approximately 10 cm.

In our classification procedure we exploit the color images and elevations measurements, resulting from the combination of the DSM and the DTM. In addition, we apply Sobel filter masks to gray value converted color images in order to compute texture information. The feature vector then consists of 78 attributes, if RGB color, texture and elevation measurements are combined using the *Sigma Points* representation. The size of the spatial neighborhood is set to 7×7 pixels.

For each dataset we separately train a randomized forest with 8 trees, each with a depth of 12. The size of the forest has given a good trade-off between evaluation time and classification accuracy. For the training process we manually label parts of six non-overlapping perspective images, where we consider a partition into five object classes (building, tree, water, grass and street). Figure 4 shows the color coding of the object classes (last row). Note that the feature representation is extracted at the pixel level, therefore the training maps can be generated quickly by annotating objects with simple strokes. Figure 4 depicts an example of a labeled ground truth map. For one dataset feature extraction and training take about 20 minutes on a dual-core computer.

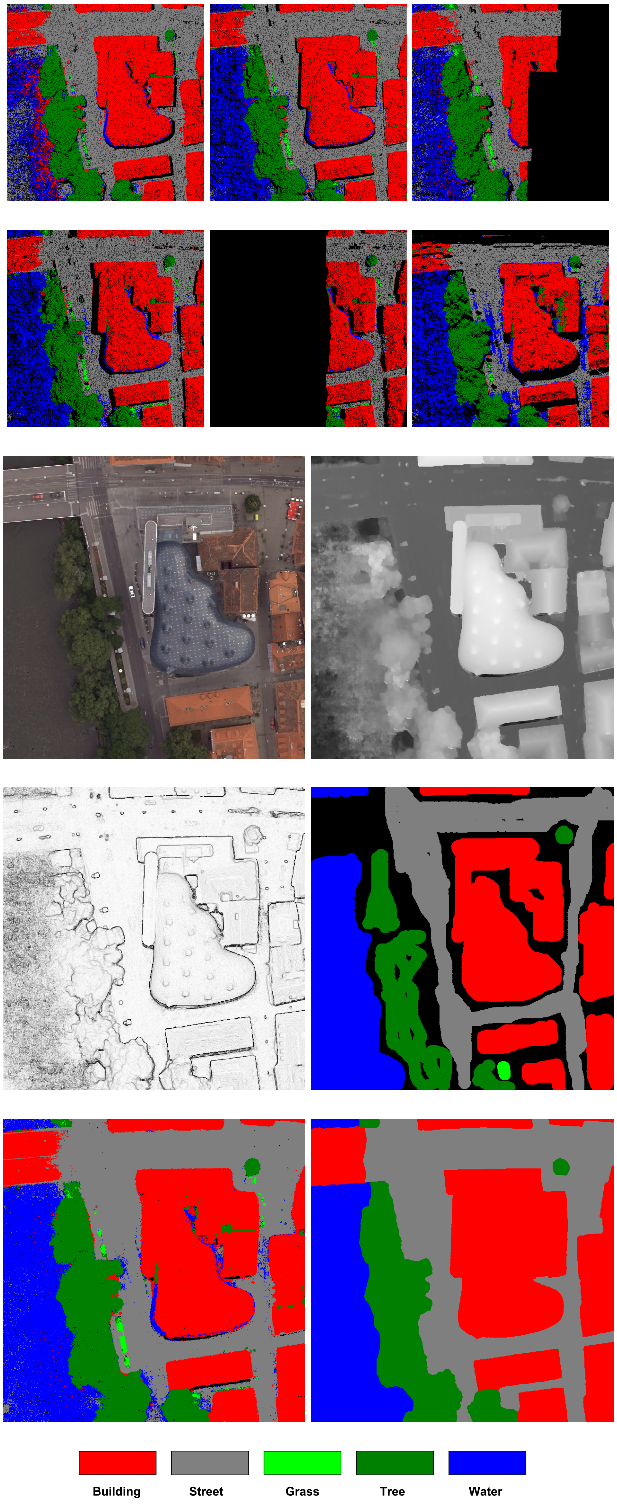

Figure 4: Semantic classification results for the *Kunsthaus* scene: The first two rows show redundant semantic classifications projected to a common orthographic view. Note that these images include many undefined regions caused by occlusions and varying viewpoints. The third row shows a fused color image and height information. The fourth row shows extracted edge information, which is used as penalty function within the Potts model based refinement step and the hand-labeled ground truth map. The raw accumulation over redundant image patches and the final refined classification result are given in the fifth row. The color coding of the object classes is depicted in the last row. Best viewed in color.

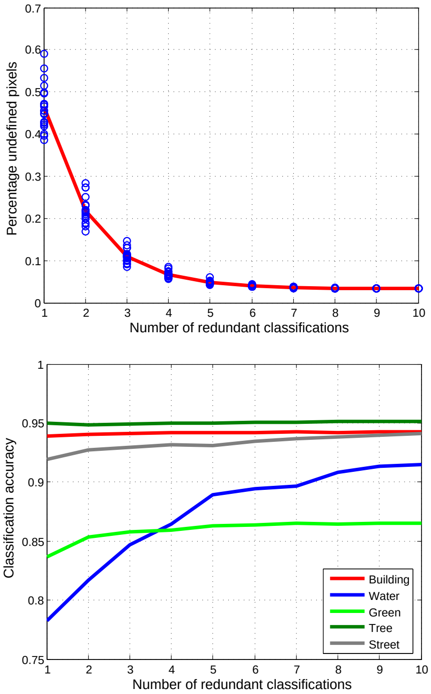

Figure 5: Fusion of redundant semantic classifications: The first plot shows the percentage of undefined pixels as a function of involved observations. The second plot depicts the obtained accuracy in terms of correctly classified pixels for an increasing quantity of involved semantic classifications. Due to varying occlusion, we average the rates over twenty runs.

At runtime we apply the trained classifier to single aerial images, yielding a highly redundant semantic interpretation from different viewpoints. In Figure 3 overlapping classification results are shown. For the purpose of visualization we only depict dominant object classes, evaluated at the pixel level. By taking into account camera data and range information we perform a pixel-wise fusion of the highly redundant images into a common 3D coordinate system. For each pixel on ground, the fusion step provides up to ten observations for color, height and assigned class probabilities. Figure 4 depicts redundant classification results for a scene of *Graz* (first and second row). Black pixels denote undefined areas caused by occlusions.

In our first experiment we investigate how the redundancy influences the classification accuracy in terms of undefined pixels and correctly classified pixels. In order to determine the classification accuracy in the orthographic view, we additionally label ground truth maps (25 image patches for each dataset), covering an area of approximately 500 × 500 meters. Figure 5 shows the obtained results for *Graz*. Due to a varying number of occluded pixels, we repeat the fusion step twenty times (shown as blue dots) with random sets of selected observations. It is obvious that the percentage of undefined pixel significantly decreases with the number of involved observations. This experiment also shows that the challenging task of distinguishing between water and street regions, covered with shadows, benefits from aggregating multiple redundant classifications. In addition, the classification rate of small grass-covered areas can be improved by using multi-view observations.

In the second experiment we apply the continuous formulation of the Potts model, described in Section 4, to obtain a consistent labeling. In order to delineate the boundaries of elevated object classes like buildings and trees accurately, we derive the edge penalty term g by computing the normalized magnitude of the DSM gradient responses according to $g = \exp(-\alpha\sqrt{dx^2 + dy^2})$. Figure 1 depicts a corresponding penalty term image. For both datasets we set $\alpha = 0.2$. The smoothness term is determined empirically with $\lambda = 0.1$. Obtained confusion matrices for *Graz*, before and after applying the refined labeling, are shown in Figure 6. Note that the classification rates can be significantly improved for all objects classes. An evaluation on the dataset *Dallas* shows that the refinement yields an averaged improvement of approximately 6% (from 87.2% to 93.3%). A classification result for a strip of *Dallas* is depicted in Figure 7.

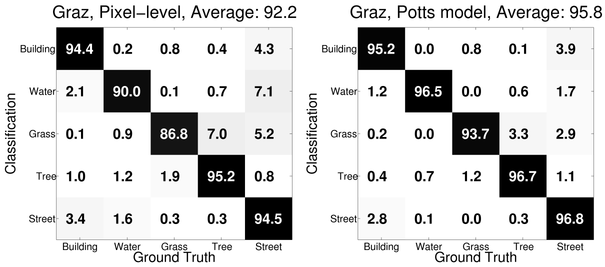

Figure 6: Obtained confusion matrices for *Graz*. The pixel-wise classification rates for all object classes and the accuracy after refined labeling using the continuous formulation of the Potts model. For the evaluation we use ten classified scene observations.

6 CONCLUSION

This work has proposed an efficient approach for semantic interpretation of urban environments. Our approach integrates multiple types of low-level feature cues, such as appearance, texture and elevation measurements and exploits the high redundancy in the data. In this work we have investigated how an overlapping redundant semantic classification into five object classes can be exploited to obtain an accurate large-scale interpretation in an orthographic view. Our approach can be easily extended to additional classes like different types of agriculturally used areas. Future work will concentrate on deriving exact GIS information and on detecting individual trees.

ACKNOWLEDGEMENTS

We would like to thank Dr. Thomas Pock for the fruitful discussions and for providing us with the implementation of the Potts model. This work has been supported by the APAFA Project No. 813397, financed by the Austrian Research Promotion Agency (http://www.ffg.at).

REFERENCES

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.

Brostow, G., Shotton, J., Fauqueur, J. and Cipolla, R., 2008. Segmentation and recognition using structure from motion point clouds. In: Proceedings ECCV, on CD-ROM.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A., 2007. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

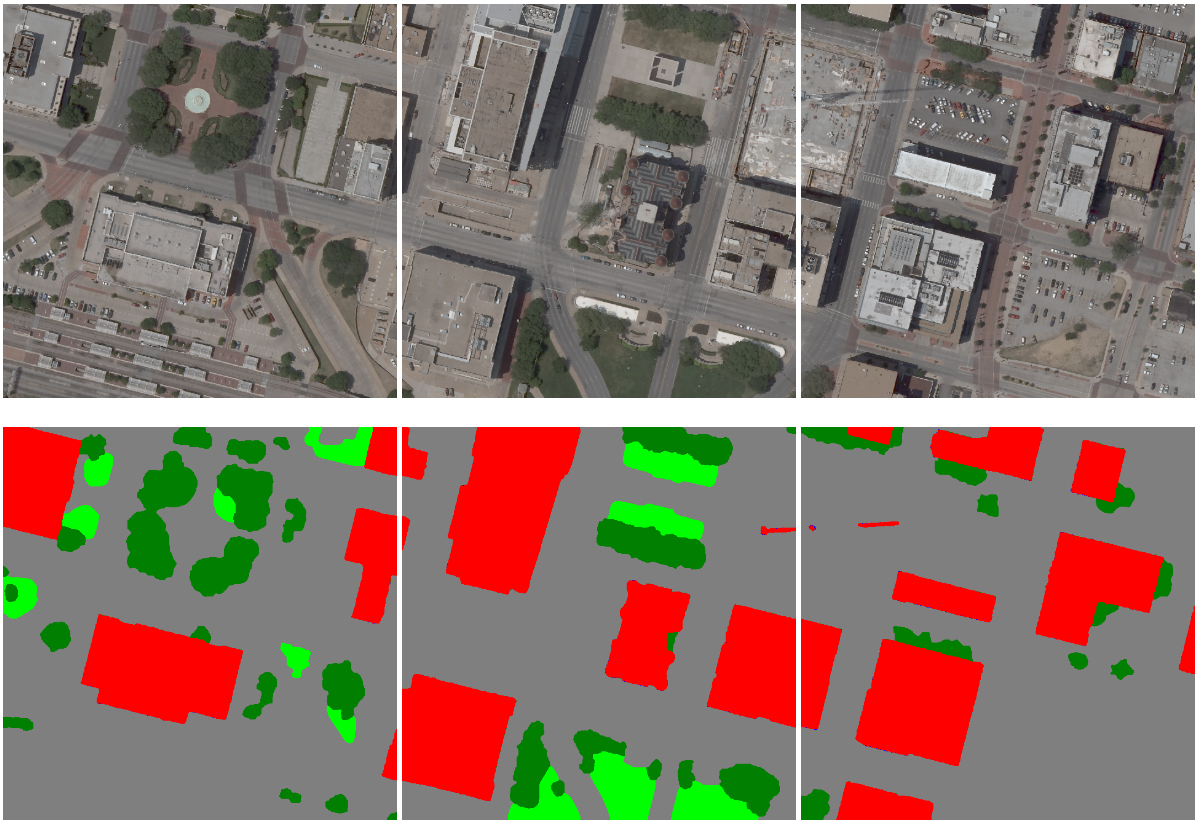

Figure 7: A semantic classification obtained for a strip of *Dallas*: The Potts model accurately preserves objects boundaries like building edges by using a height field driven penalty function. To handle the enormous amount of redundant data we collect the data in 1600 × 1600 pixel image patches.

Gould, S., Rodgers, J., Cohen, D., Elidan, G. and Koller, D., 2008. Multi-class segmentation with relative location prior. IJCV 80(3), pp. 300–316.

Hirschmüller, H., 2006. Stereo vision in structured environments by consistent semi-global matching. In: Proceedings CVPR, 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.

Komodakis, N. and Tziritas, G., 2007. Approximate Labeling via Graph Cuts Based on Linear Programming. PAMI 29(8), pp. 1436–1453.

Lepetit, V., Lagger, P. and Fua, P., 2005. Randomized trees for real-time keypoint recognition. In: Proceedings CVPR, on CD-ROM.

Nguyen, T. T., Kluckner, S., Bischof, H. and Leberl, F., 2010. Aerial Photo Building Classification by Stacking Appearance and Elevation Measurements. In: Proceedings ISPRS, 100 Years ISPRS - Advancing Remote Sensing Science, on CD-ROM.

Olsson, C., Byröd, M., Overgaard, N. C. and Kahl, F., 2009. Extending continuous cuts: Anisotropic metrics and expansion moves. In: Proceedings ICCV, on CD-ROM.

Pock, T., Chambolle, A., Cremers, D. and Bischof, H., 2009. A convex relaxation approach for computing minimal partitions. In: Proceedings CVPR, on CD-ROM.

Potts, R. B., 1952. Some generalized order-disorder transformations. Proceedings of the Cambridge Philosophical Society 48, pp. 106–109.

Rottensteiner, F., Trinder, J., Clode, S., Kubik, K. and Lovell, B., 2004. Building detection by dempster-shafer fusion of lidar data and multispectral aerial imagery. In: Proceedings ICPR, on CD-ROM.

Sharp, T., 2008. Implementing decision trees and forests on a gpu. In: Proceedings ECCV, on CD-ROM.

Shotton, J., Johnson, M. and Cipolla, R., 2008. Semantic texton forests for image categorization and segmentation. In: Proceedings CVPR, on CD-ROM.

Shotton, J., Winn, J., Rother, C. and Criminisi, A., 2006. Textonboost: Joint appearance, shape and context modeling for multiclass object recognition and segmentation. In: Proceedings ECCV, on CD-ROM.

Strecha, C., Gool, L. V. and Fua, P., 2008. A generative model for true orthorectification. IAPRS 37, pp. 303–308.

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

Unger, M., Pock, T., Grabner, M., Klaus, A. and Bischof, H., 2009. A variational approach to semiautomatic generation of digital terrain models. In: Proceedings ISVC, on CD-ROM.

Verbeek, J. and Triggs, B., 2007. Scene segmentation with crfs learned from partially labeled images. In: Proceedings NIPS, on CD-ROM.

Viola, P. and Jones, M., 2004. Robust real-time object detection. IJCV 2(57), pp. 137–154.

Xiao, J. and Quan, L., 2009. Multiple view semantic segmentation for street view images. In: Proceedings ICCV, 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.