LAND COVER CLASSIFICATION OF SATELLITE IMAGES USING CONTEXTUAL INFORMATION

Björn Fröhlich[1,3,*], **Eric Bach**[1,*], **Irene Walde**[2,3,*], **Sören Hese**[2,3], **Christiane Schmullius**[2,3], **and Joachim Denzler**[1,3]

[1] Computer Vision Group, Friedrich Schiller University Jena, Germany
[2] Department of Earth Observation, Friedrich Schiller University Jena, Germany
[3] Graduate School on Image Processing and Image Interpretation, ProExzellenz Thuringia, Germany
[*] Co-first authors
{*bjoern.froehlich,eric.bach,irene.walde,soeren.hese,c.schmullius,joachim.denzler*}*@uni-jena.de*

KEY WORDS: Land Cover, Classification, Segmentation, Learning, Urban, Contextual

ABSTRACT:

This paper presents a method for the classification of satellite images into multiple predefined land cover classes. The proposed approach results in a fully automatic segmentation and classification of each pixel, using a small amount of training data. Therefore, semantic segmentation techniques are used, which are already successful applied to other computer vision tasks like facade recognition. We explain some simple modifications made to the method for the adaption of remote sensing data. Besides local features, the proposed method also includes contextual properties of multiple classes. Our method is flexible and can be extended for any amount of channels and combinations of those. Furthermore, it is possible to adapt the approach to several scenarios, different image scales, or other earth observation applications, using spatially resolved data. However, the focus of the current work is on high resolution satellite images of urban areas. Experiments on a QuickBird-image and LiDAR data of the city of Rostock show the flexibility of the method. A significant better accuracy can be achieved using contextual features.

1 INTRODUCTION

The beginning of land cover classification from aerial images dates back around 70 years (Anderson et al., 1976). Since then aerial and satellite images are used to extract land cover in a broadly manner and without direct contact to the observed area. Land cover is defined as “the observed (bio)physical cover on the earth’s surface” by Di Gregorio (2005). It is an essential information for change detection applications or derivation of relevant planning or modeling parameters. Other fields of applications are the analysis and visualization of complex topics like climate change, biodiversity, resource management, living quality assessment, land use derivation or disaster management (Herold et al., 2008, Hüttich et al., 2011, Walde et al., 2012). Manual digitization of land cover or land surveying methods result in huge effort in time as well as financial and personal resources. Therefore, methods of automated land cover extraction on the basis of area-wide available remote sensing data are utilized and continually improved. High spatial resolution satellite images, such as QuickBird, Ikonos, or WorldView, enable to map the heterogeneous range of urban land cover. By the availability of such high resolution images, OBIA-methods (Object Based Image Analysis) were developed (Benz et al., 2004, Hay and Castilla, 2008, Blaschke, 2010), which are preferred to pixel-based methods in urban context (Myint et al., 2011). Pixel-based methods consider only spectral properties. Object-based classification processes observe, apart from spectral properties, characteristics like shape, texture or adjacency criteria. An overview of automatic labeling methods for land-cover classification can be found in Schindler (2012).

In this work, we present an automatic approach for semantic segmentation and classification, which does not need any human interaction. It extracts the urban land cover from high resolution satellite images using just some training areas. The proposed method is called Iterative Context Forest from Fröhlich et al. (2012). This approach uses besides local features also contextual cues between classes. For instance, the probability of large buildings and impervious surfaces (*e.g.,* parking slots) in industrial areas is much higher than in allotment areas. Using contextual information improves the classification results significantly. The proposed method is flexible in using multiple channels and combinations of those. Therefore, the optimal features for each class are automatically selected out of a big feature pool during a training step. As features we use established methods from computer vision, like integral features from person detection. Iterative Context Forests are originally developed for the problems from image processing like facade recognition and we adapt them for remote sensing data.

The paper is structured as follows. Section 2 describes the study site and the available data set. In Section 3 the method of the semantic segmentation and the modifications made due to remote sensing data are explained. The results are presented and discussed in Section 4. Finally, Section 5 summarizes the work in this paper and mentions further research aspects.

2 STUDY AREA AND DATA SET

In the focus of this study, is the research area of Rostock, a city with more than 200.000 inhabitants on an area of 181 km^2, situated in the north of Germany (Mecklenburg- Vorpommern Statistisches Amt, 2012). A subset of five by five kilometers of a cloud-free Quickbird scene from September 2009 was available for this study to develop and test the method (Figure 1). It represents the south-west part of Rostock, including the Warnow river in the north, parts of the city center, the federal road B103 in the west, and adjacent fields. The Quickbird scene has four multispectral channels (blue, green, red, near infrared), which were pansharpened with the panchromatic channel to a spatial resolution of 60 cm per pixel. The scene was provided in the Ortho-Ready Standard (OR2A) format and was projected to an average elevation (Cheng et al., 2003). The image was corrected for atmospheric effects and orthorectified using ground control points and a digital terrain model. Additionally, a LiDAR normalized digital surface model (nDSM) was available, which was produced

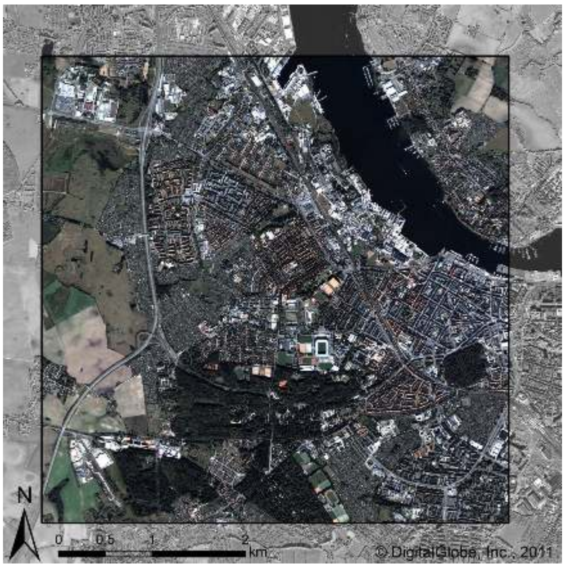


Figure 1: Quickbird satellite image subset of Rostock (©DigitalGlobe, Inc., 2011).

by subtracting the terrain from the surface model (collected in 2006). The relative object heights of the nDSM were provided in a spatial resolution of 2 m per pixel on the ground.

3 SEMANTIC SEGMENTATION

In computer vision, the term *semantic segmentation* covers several methods for pixel-wise annotation of images without a focus on specific tasks. At which, segmentation denotes the process of dividing an images into disjoint group of pixels. Each of those groups is called a region. Furthermore, all pixels in a region are homogeneous with respect to a specific criteria (*e.g.,* color or texture). The target of segmenting an image is to transform the image into a better representation, which is reduced to the essential parts. Furthermore, segmentation can be differed into unsupervised and supervised segmentation.

Unsupervised segmentation denotes that all pixels are grouped into different regions, but there is no meaning annotated to any of them. However, for supervised segmentation or semantic segmentation a semantic meaning is annotated to each region or rather to each pixel. Usually, this is a class name out of a predefined set of class names. The selection of those classes highly depends on the chosen task and the data. For instance, a low resolution satellite image of a whole country can be analyzed, where the classes *city* and *forest* might be interesting. Alternatively, if we classify land cover of very high resolution satellite images of cities, classes like *roof*, *pool*, or *tree* are recognizable in the image.

In this section, we will introduce the Iterative Context Forest (ICF) from Fröhlich et al. (2012). Afterwards, we focus on the differences to the original work. The basic idea of Iterative Context Forest is similar to the Semantic Texton Forests (STF) from Shoton et al. (2008). The basic difference is that the STF context features are computed in advance and can not adapt to the current classification result after each level of a tree.

3.1 Essential foundations

Feature vectors are compositions of multiple features. Each feature vector describes an object or a part of an object. For instance, the mean value of each color channel is such a collection of simple features. To describe more complex structures, we need besides color also texture and shape as important features.

Classification denotes the problem in pattern recognition of assigning a class label to a feature vector. Therefore, a classifier needs an adequate set of already labeled feature vectors. The classifier tries to model the problem out of this training data during a training step. With this model, the classifier can assign to each new feature vector a label during testing.

3.2 Iterative Context Forests

An Iterative Context Forests (ICF) is a classification system which is based on Random Decision Forests (RDF) (Breiman, 2001). Each RDF is an ensemble of decision trees (DT). Therefore, in this section we first introduce DT, subsequently RDF and finally ICF.

3.2.1 Decision trees To solve the classification problem, decision trees (Duda and Hart, 1973, Chap. 8.2) are a fast and simple way. The training data is split by a simple decision (*e.g.,* is the current value in the green channel less than 127 or not). Each subset is split again by another but also simple decisions into more subsets until each subset consists only of feature vectors from one class. Due to these splits, a tree like structure is created, where each subset with only one class in it is called a leaf of the tree. All other subsets are called inner node. The tree is traversed by an unknown feature vector until this vector ends in a leaf. The assigned class to this feature vector is the same as all training feature vectors have in this leaf. To find the best split during training a brute-force search in the training data is done by maximizing the Kullback-Leibler entropy.

3.2.2 Random decision forests It has been exposed that decision trees tend to overfitting to the training data. In the worst case, training a tree on data with high noise let this tree split the data until each leaf only consists of a single feature vector. To prevent this, Breiman (2001) suggest to use RDF, which prevents overfitting by using multiple random selections. First, there is not only one tree learned but many. Second, each tree is trained on a different random subset of the training data. Third, for each split only a random subset from the feature space is used. Furthermore, the data is not split anymore until the feature vectors of one node are from the same class. There are several stop criteria instead: a maximum depth of the tree, a minimum number of training samples in a leaf, and a threshold for the entropy in a leaf is defined. Therefore, an a-posteriori probability can be computed from the distribution of the labels of the feature vectors ended up in the current leaf per tree. A new feature vector traverses all trees and for each tree it ends up in a leaf. The final decision is made by averaging the probabilities of all these leafs (Figure 2).

3.2.3 Millions of features The presented method is based on the extraction of multiple features from the input image. Besides of the single input channels, additional channels can be computed, *e.g.,* gradient image. On each of these channels and on combination of those several features can be computed in a local neighborhood d. For instance, the difference of two random selected pixels relatively to the current pixel position or the mean value of a random selected pixel relatively to the current position (more feature extraction methods are shown in Figure 3).

3.2.4 Auto context features The main difference to a standard RDF is the usage of features changing during traversing the tree. Therfore, the trees have to be created level-wise. After learning a level the probabilities for each pixel and for each

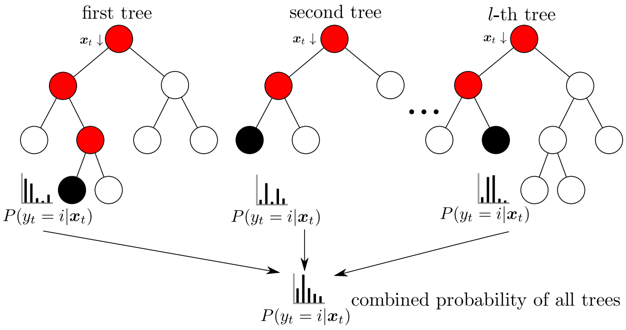

first tree
second tree
l-th tree
$x_t \downarrow$
$P(y_t = i|\boldsymbol{x}_t)$
combined probability of all trees

Figure 2: Random decision forest — l different binary decision tree, traversed node are marked red and the reached leafs are marked black.

class are added to the feature space as additional feature channels. Context knowledge can be extracted from the neighborhood, if the output of the previous level leads to an adequate result. Some of these contextual features are presented in Figure 8.

3.3 Modifications for remote sensing data

The presented method is only used before on datasets presenting facade images (Fröhlich et al., 2012). The challenges in facade images are different to the challenges in remote sensing images. Due to the resolution of the image and the size of the area, the objects are much smaller compared to windows or other objects from facade images. To adapt to this circumstances, the window size d is reduced (*cf.* Section 3.2.3 and Figure 3). Furthermore, some feature channels from the original work are not adaptable to remote sensing data, like the geometric context (Hoiem et al., 2005). Instead, some for the classical computer vision unusual channels can be used. These channels are near infrared and LiDAR nDSM. Due to the flexibility of the proposed method, any kind of channels might be added, like the “Normalized Difference Vegetation Index” (NDVI):

$$\text{NDVI}(x, y) = \frac{\text{NIR}(x, y) - \text{Red}(x, y)}{\text{NIR}(x, y) + \text{Red}(x, y)} \quad . \qquad (1)$$

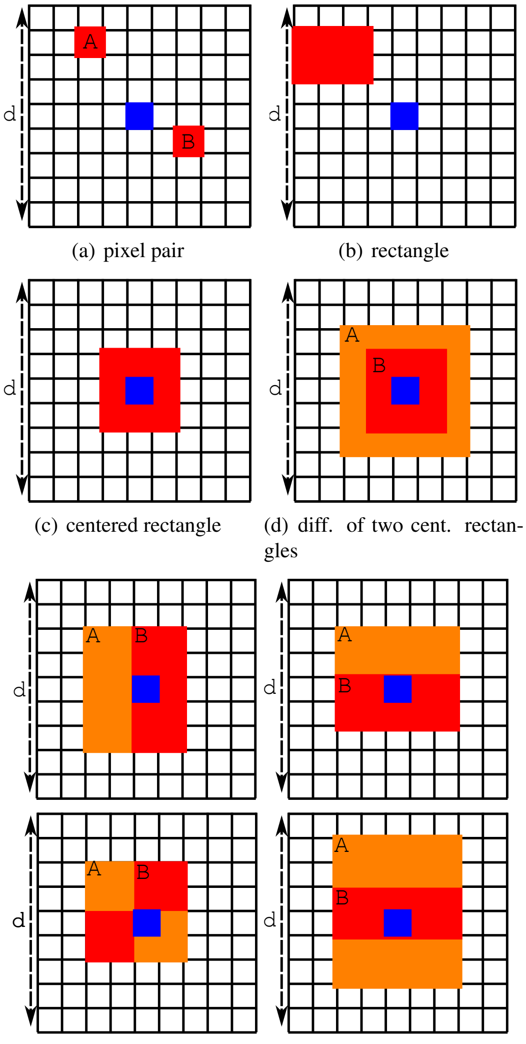
This index is computed from the red and the near infrared channel and allows a differentiation of vegetation and paved areas.

4 RESULTS

For testing, we used some already labeled training areas. On the rest of the dataset 65 points per class are randomly selected for testing (Figure 4). Due to the previous mentioned randomizations, each classification is repeated ten times and the results are averaged. We focused on the classes: *tree, water, bare soil, building, grassland,* and *impervious.*

All tests are made with a fixed window size $d = 30\text{px} = 18\text{m}$ for non-contextual features and $d = 120\text{px} = 72\text{m}$ for all contextual features. Those values are exposed to be optimal in previous tests.

The qualitative results of our proposed method are presented in Figure 5 and the quantitative results in Figure 6. Using only the RGB values, the near infrared (NIR) and the panchromatic channel (PAN) we get an overall accuracy of 82.5% (Figure 6(a)). The main problems are to differ between the classes *impervious* and *building* as well as *grassland* and *bare soil*. The classes *tree* and *water* are already well classified. Adding the nDSM the confusion of *building* and *impervious* rapidly decreases (Figure 6(b)). This accords to our expectations, due to the fact that those both classes look very similar from the bird’s eye view but they differ in the height. Adding the NDVI helps to reduce the confusion between the classes *grassland* and *bare soil* (Figure 6(c)). This is also what we expected, due to the fact that *grassland* has a much brighter appearance in the NDVI image than *bare soil*. But there are still some confusions between *bare soil* and *grassland*. On the other side, adding the NDVI also boosts the confusion between *tree* and *grassland*. This might be a side effect of almost the same appearance of those classes in the NDVI channel and the assignment of shrubs to either of the classes. In Figure 6(d), we added both channels, nDSM and NDVI. The benefits from adding only NDVI or adding only nDSM are still valid.



(a) pixel pair
(b) rectangle
(c) centered rectangle
(d) diff. of two cent. rectangles
(e) Haar-like (Viola and Jones, 2001)

Figure 3: Feature extraction methods from (Fröhlich et al., 2012). The blue pixel denotes the current pixel position and the grid a window around it. The red and orange pixels are used as features. They can be used as simple feature (c and d) or they can be combined, *e.g.*, by $A + B$, $A - B$ or $|A - B|$ (a,b and e).

In Figure 6(e), we used the same settings as in Figure 6(d) besides that we switched off the context features. Almost every value without using context is worse than the values using contextual cues. Especially, *bare soil* and *impervious* benefits from using contextual knowledge. Without contextual knowledge the class *bare soil* is often confused with *grassland* and *impervious*, but using contextual knowledge *impervious* and *bare soil* are well classified. One suitable explanation for this might be that *bare soil* is often found on harvested fields outside the city. Due to this reason, the probability for classes like *grassland* or *impervious* is

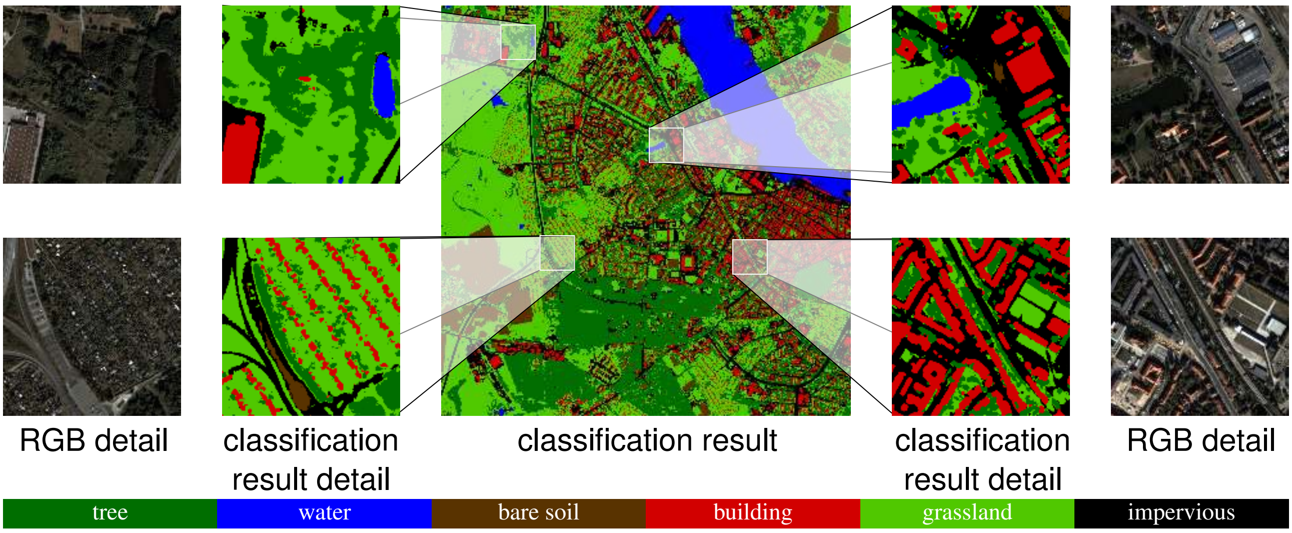
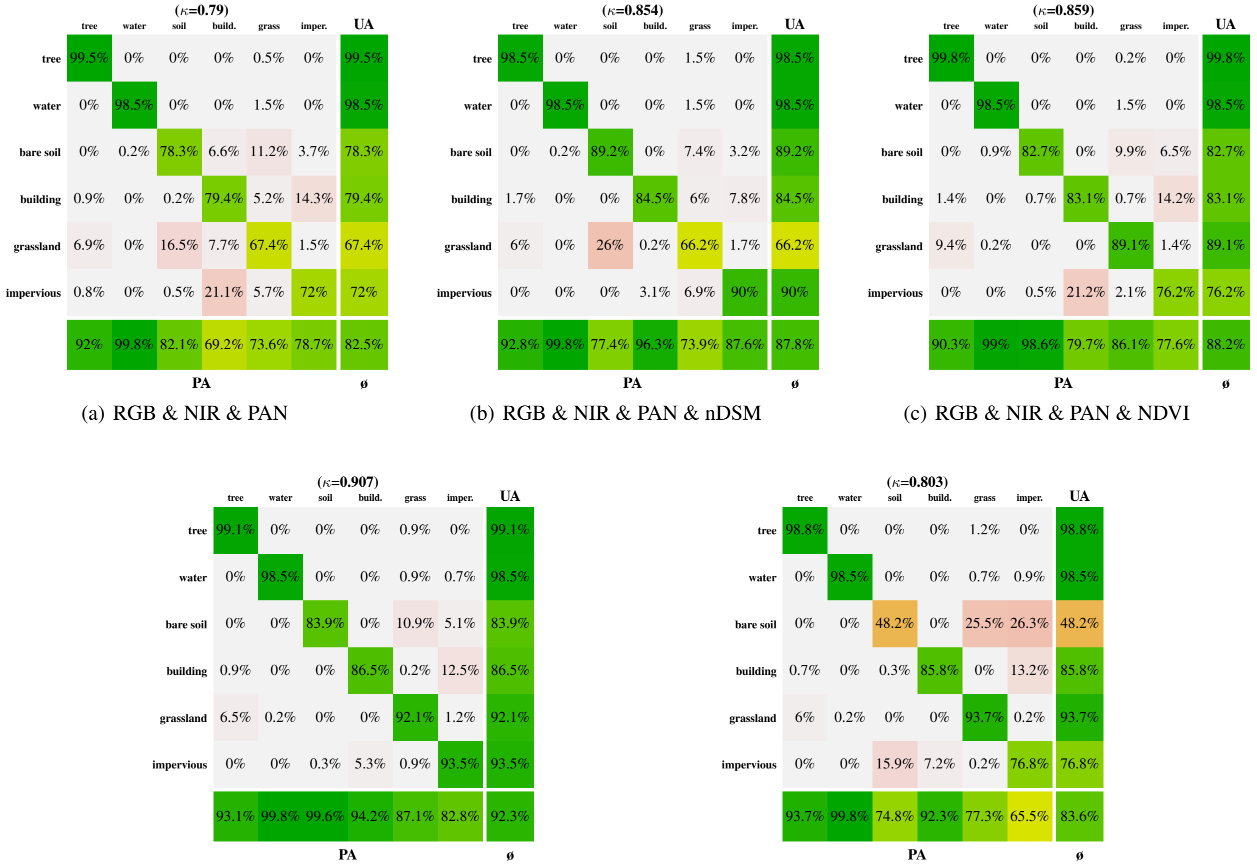


Figure 5: Classification result and four sample areas in full resolution (each 420 × 420m).



(κ=0.79)

	tree	water	soil	build.	grass	imper.	UA
tree	99.5%	0%	0%	0%	0.5%	0%	99.5%
water	0%	98.5%	0%	0%	1.5%	0%	98.5%
bare soil	0%	0.2%	78.3%	6.6%	11.2%	3.7%	78.3%
building	0.9%	0%	0.2%	79.4%	5.2%	14.3%	79.4%
grassland	6.9%	0%	16.5%	7.7%	67.4%	1.5%	67.4%
impervious	0.8%	0%	0.5%	21.1%	5.7%	72%	72%
	92%	99.8%	82.1%	69.2%	73.6%	78.7%	82.5%
			PA				ø

(a) RGB & NIR & PAN

(κ=0.854)

	tree	water	soil	build.	grass	imper.	UA
tree	98.5%	0%	0%	0%	1.5%	0%	98.5%
water	0%	98.5%	0%	0%	1.5%	0%	98.5%
bare soil	0%	0.2%	89.2%	0%	7.4%	3.2%	89.2%
building	1.7%	0%	0%	84.5%	6%	7.8%	84.5%
grassland	6%	0%	26%	0.2%	66.2%	1.7%	66.2%
impervious	0%	0%	0%	3.1%	6.9%	90%	90%
	92.8%	99.8%	77.4%	96.3%	73.9%	87.6%	87.8%
			PA				ø

(b) RGB & NIR & PAN & nDSM

(κ=0.859)

	tree	water	soil	build.	grass	imper.	UA
tree	99.8%	0%	0%	0%	0.2%	0%	99.8%
water	0%	98.5%	0%	0%	1.5%	0%	98.5%
bare soil	0%	0.9%	82.7%	0%	9.9%	6.5%	82.7%
building	1.4%	0%	0.7%	83.1%	0.7%	14.2%	83.1%
grassland	9.4%	0.2%	0%	0%	89.1%	1.4%	89.1%
impervious	0%	0%	0.5%	21.2%	2.1%	76.2%	76.2%
	90.3%	99%	98.6%	79.7%	86.1%	77.6%	88.2%
			PA				ø

(c) RGB & NIR & PAN & NDVI

(κ=0.907)

	tree	water	soil	build.	grass	imper.	UA
tree	99.1%	0%	0%	0%	0.9%	0%	99.1%
water	0%	98.5%	0%	0%	0.9%	0.7%	98.5%
bare soil	0%	0%	83.9%	0%	10.9%	5.1%	83.9%
building	0.9%	0%	0%	86.5%	0.2%	12.5%	86.5%
grassland	6.5%	0.2%	0%	0%	92.1%	1.2%	92.1%
impervious	0%	0%	0.3%	5.3%	0.9%	93.5%	93.5%
	93.1%	99.8%	99.6%	94.2%	87.1%	82.8%	92.3%
			PA				ø

(d) RGB & NIR & PAN & NDVI & nDSM

(κ=0.803)

	tree	water	soil	build.	grass	imper.	UA
tree	98.8%	0%	0%	0%	1.2%	0%	98.8%
water	0%	98.5%	0%	0%	0.7%	0.9%	98.5%
bare soil	0%	0%	48.2%	0%	25.5%	26.3%	48.2%
building	0.7%	0%	0.3%	85.8%	0%	13.2%	85.8%
grassland	6%	0.2%	0%	0%	93.7%	0.2%	93.7%
impervious	0%	0%	15.9%	7.2%	0.2%	76.8%	76.8%
	93.7%	99.8%	74.8%	92.3%	77.3%	65.5%	83.6%
			PA				ø

(e) RGB & NIR & PAN & NDVI & nDSM without context

Figure 6: Results of ICF using different channels. RGB: red, green and blue channel, NIR: near infrared, PAN: panchromatic channel, NDVI: normalized differenced vegetation index, nDSM: normalized elevation model. UA: user accuracy and PA: producer accuracy.

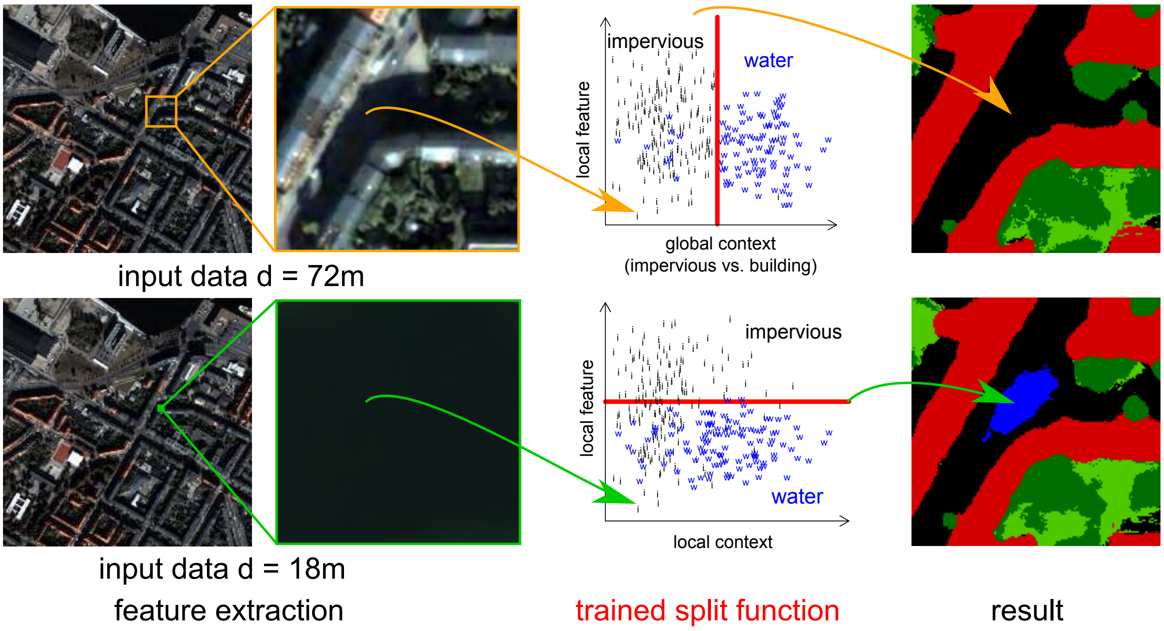


Figure 7: Context vs. no context: first row using contextual features to differ between *impervious* (road) and *water* tends to better results than using no contextual features in the second row.

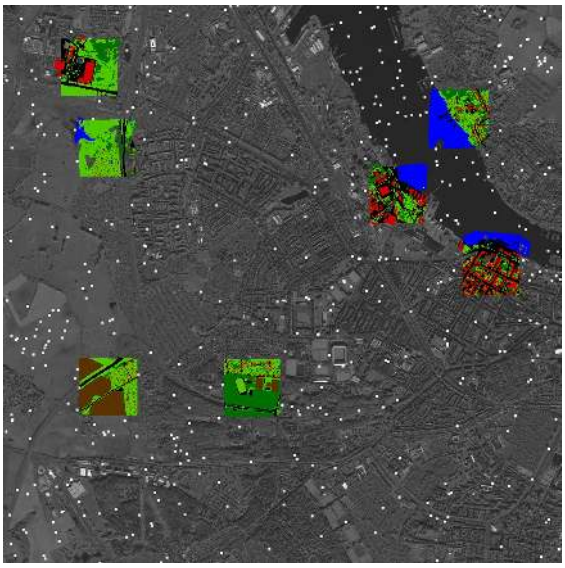


Figure 4: The seven training regions and the 65 evaluation points per class.

much higher in the neighborhood of *buildings*.

The influence of the window size and the usage of contextual features is shown in Figure 7. In this example in the top row, the classes *water* and *impervious* (the road) are well distinguished, but without using contextual knowledge there are some problems in the bottom row, where some pixels in the middle of the street are classified as *water*, due to the fact that in this case the surrounding area is not considered.

Since the time interval from LiDAR, collected in 2006, and the QuickBird satellite image, recorded in 2009, artificial "change" is created, which leads to misclassifications. Some buildings are visible in the satellite image and not in the nDSM and the other way around. There are some problems with the shadow of trees, which are not represented enough in the training data. Furthermore, small objects (like small detached houses) vanish in the classification result with a larger window size. Finally, the object borders are very smooth, this can be fixed by using an unsupervised segmentation.

In Figure 8, we show the final probability maps for all classes and for each pixel of selection of the data. It is not obligatory to use the most probable class per pixel as final decision. It is also possible to use those maps for further processing like filling gaps between streets. However, these specialized methods are not part of this work.

The best classification result (using context on the QuickBird data, nDSM and NDVI) is shown in Figure 5, including some areas in detail.

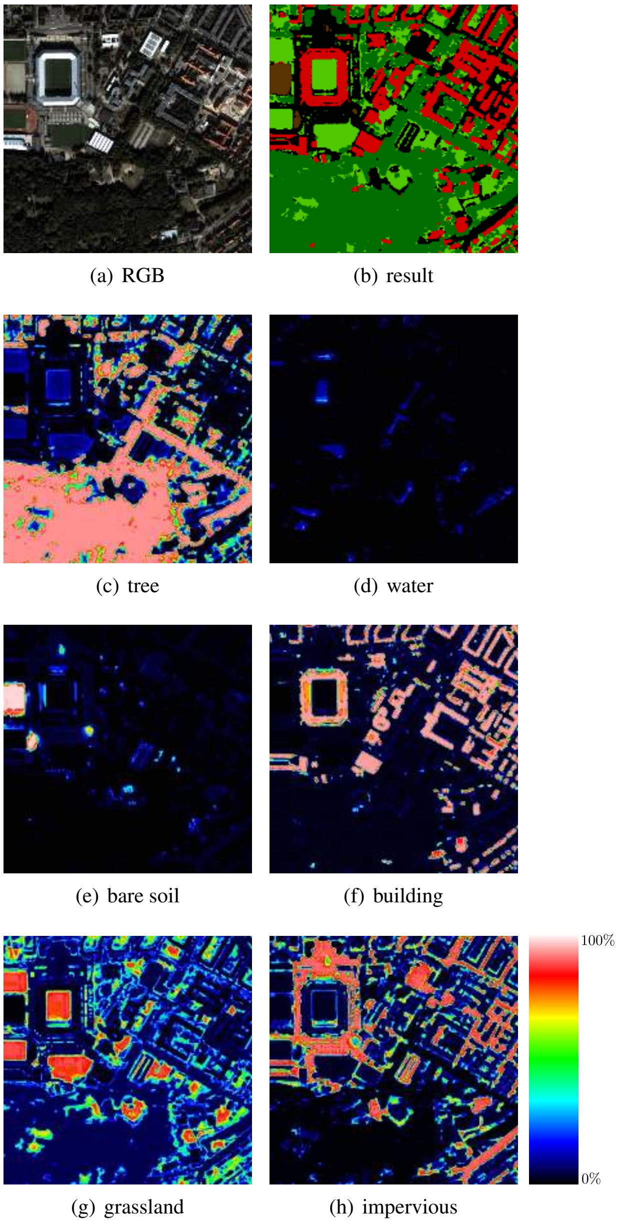
5 CONCLUSIONS AND FURTHER WORK

In this work, we introduced a state of the art approach from computer vision for semantic segmentation. Furthermore, we have presented how to adapt this method for the classification of land cover. In our experiment, we have shown that our method is flexible in using multiple channels and that adding channels increases the quality of the result. The benefits of adding contextual knowledge to the classification has been demonstrated and discussed for some specific problems.

For further work, we are planning to use an unsupervised segmentation to improve the performance especially at the borders of the objects. Furthermore, we are planning to incorporate shape information. Finally, an analysis of the whole QuickBird scene (13.8 × 15.5km) is planned as well as experiments using other scales and classes.

ACKNOWLEDGMENTS

This work was partially founded by the ProExzellenz Initiative of the "Thüringer Ministeriums für Bildung, Wissenschaft und Kultur" (TMBWK, grant no.: PE309-2). We also want to thank the city of Rostock for providing the LiDAR data.



(a) RGB (b) result

(c) tree (d) water

(e) bare soil (f) building

(g) grassland (h) impervious

Figure 8: Probability maps for all classes (each sample area is 840 × 840m).

REFERENCES

Anderson, J. R., Hardy, E. E., Roach, J. T. and Witmer, R. E., 1976. A land use and land cover classification system for use with remote sensor data.

Benz, U. C., Hofmann, P., Willhauck, G., Lingenfelder, I. and Markus, H., 2004. Multi-resolution, object-oriented fuzzy analysis of remote sensing data for gis-ready information. ISPRS Journal of Photogrammetry & Remote Sensing (58), pp. 239–258.

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing 65(1), pp. 2–16.

Breiman, L., 2001. Random forests. Machine Learning 45(1), pp. 5–32.

Cheng, P., Toutin, T., Zhang, Y. and Wood, M., 2003. Quickbird – geometric correction, path and block processing and data fusion. Earth Observation Magazine 12(3), pp. 24–28.

Di Gregorio, A., 2005. Land cover classification system software version 2: Based on the orig. software version 1. Environment and natural resources series Geo-spatial data and information, Vol. 8, rev. edn, Rome.

Duda, R. O. and Hart, P. E., 1973. Pattern Classification and Scene Analysis. Wiley.

Fröhlich, B., Rodner, E. and Denzler, J., 2012. Semantic segmentation with millions of features: Integrating multiple cues in a combined random forest approach. In: Proceedings of the Asian Conference on Computer Vision (ACCV).

Hay, G. J. and Castilla, G., 2008. Geographic object-based image analysis (geobia): A new name for a new discipline. In: T. Blaschke, S. Lang and G. Hay (eds), Object-based image analysis, Springer, pp. 75–89.

Herold, M., Woodcock, C., Loveland, T., Townshend, J., Brady, M., Steenmans, C. and Schmullius, C., 2008. Land-cover observations as part of a global earth observation system of systems (geoss): Progress, activities, and prospects. IEEE Systems Journal 2(3), pp. 414–423.

Hoiem, D., Efros, A. A. and Hebert, M., 2005. Geometric context from a single image. In: Proceedings of the International Conference on Computer Vision (ICCV)), Vol. 1, IEEE, pp. 654–661.

Hüttich, C., Herold, M., Wegmann, M., Cord, A., Strohbach, B., Schmullius, C. and Dech, S., 2011. Assessing effects of temporal compositing and varying observation periods for large-area land-cover mapping in semi-arid ecosystems: Implications for global monitoring. Remote Sensing of Environment 115(10), pp. 2445–2459.

Mecklenburg- Vorpommern Statistisches Amt, 2012. `http://www.statistik-mv.de`.

Myint, S. W., Gober, P., Brazel, A., Grossman-Clarke, S. and Weng, Q., 2011. Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery. Remote Sensing of Environment 115(5), pp. 1145–1161.

Schindler, K., 2012. An overview and comparison of smooth labeling methods for land-cover classification. IEEE Transactions on Geosciences and Remote Sensing.

Shotton, J., Johnson, M. and Cipolla, R., 2008. Semantic texton forests for image categorization and segmentation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8.

Viola, P. and Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, IEEE, pp. 511–518.

Walde, I., Hese, S., Berger, C. and Schmullius, C., 2012. Graph-based urban land use mapping from high resolution satellite images. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. I-4, pp. 119–124.