SEMANTIC SEGMENTATION OF AERIAL IMAGES IN URBAN AREAS WITH CLASS-SPECIFIC HIGHER-ORDER CLIQUES

Javier A. Montoya-Zegarra[a], Jan D. Wegner[a], L'ubor Ladický[b], Konrad Schindler[a]

[a] Photogrammetry and Remote Sensing, ETH Zurich
[b] Computer Vision Group, ETH Zurich

Commission III & VII, WG III/4 & ICWG III/VII

KEY WORDS: semantic aerial segmentation, building detection, road-network extraction, conditional random fields

ABSTRACT:

In this paper we propose an approach to multi-class semantic segmentation of urban areas in high-resolution aerial images with class-specific object priors for buildings and roads. What makes model design challenging are highly heterogeneous object appearances and shapes that call for priors beyond standard smoothness or co-occurrence assumptions. The data term of our energy function consists of a pixel-wise classifier that learns local co-occurrence patterns in urban environments. To specifically model the structure of roads and buildings, we add high-level shape representations for both classes by sampling large sets of putative object candidates. Buildings are represented by sets of compact polygons, while roads are modeled as a collection of long, narrow segments. To obtain the final pixel-wise labeling, we use a CRF with higher-order potentials that balances the data term with the object candidates. We achieve overall labeling accuracies of $> 80\%$.

1. INTRODUCTION

The automatic interpretation of aerial (and satellite) images has been a classic problem of remote sensing and machine vision. Semantically interpreted images, i.e. thematic raster maps, of urban areas are important for many applications, for example mapping and navigation, urban planning and environmental monitoring, to name just a few.

Automatic segmentation into semantically defined classes (also referred to as "image classification" or "semantic labeling") has been an active area of research over the last 40 years. In spite of great progress, the task is far from solved. This is especially true for urban areas, and at high spatial resolutions (on the oder of 0.1 - 1 m). Urban areas exhibit a large variety of reflectance patterns, with large intra-class variations and often also low inter-class variation. The situation gets even more challenging at high spatial resolution. Urban land-cover classes like "road" or "building" are a mixture of many different structures and materials. As small objects such as individual cars, street furniture, roof structures, and even things like traffic signs or road markings become visible, the intra-class variability increases.

In this work, we deal with semantic segmentation of aerial images into broad classes. We put particular emphasis on two classes, roads and buildings. These two object classes are on one hand of particular importance, as they make up a large portion of the urban fabric and account for most of the man-made environment. On the other hand, they also exhibit a significant amount of structure which can be exploited to improve their labeling in the presence of noisy data. *Roads* consist of long, thin segments of slowly varying width, which should form a connected network. *Buildings* are relatively compact blobs with simple (mostly polygonal) boundaries.

We aim to include this a-priori knowledge about object layout in a probabilistic manner, in the form of soft constraints in a Conditional Random Field (CRF) model. Our proposed pipeline starts from a conventional pixelwise prediction of the class likelihoods. However, in order to better account for the rich context and co-occurrence patterns in urban environments we include appearance features sampled from a large spatial neighborhood around each pixel, using ideas from (L'ubor Ladický et al., 2010). On top of the independent class likelihoods of individual pixels, we add a higher-level representation at the level of (putative) objects or pieces of objects, which are derived from object-specific prior assumptions. More specifically, we generate hypotheses for possible road segments and for possible segments of buildings in a data-driven manner: from the raw class likelihoods we sample long, narrow segments that have high cumulative road likelihood, as well as compact blobs with simple boundaries that have high cumulative building likelihood. To guarantee high recall, the set of such object hypotheses is tuned to cover, as far as possible, all roads and buildings, at the cost of being over-complete and redundant. The final labeling step then consists in classifying the image pixels in such a way that the object hypotheses are respected, meaning that (almost) all pixels that belong to a given hypothesis get assigned the same label. Formally, this process can be modeled as a CRF with sparse higher-order cliques, a so-called P^N-Potts model (Kohli et al., 2009). A higher-order prior for roads has already been described in our previous work (Montoya-Zegarra et al., 2014). Here, we first propose a simple a-priori model for buildings. Like earlier work our prior favours simple and smooth (bot not necessarily convex) outlines, but imposes this expectation as a probabilistic soft constraint, rather than as a post-processing heuristic. Second, our framework allows one to handle different, class-specific higher-order correlations in one unified framework. Prior expectations about object layout are taken into account by sampling higher-order cliques in a class-specific fashion; whereas all cliques, independent of how they were generated, are used together in a single CRF inference step to determine the pixel labels. In experiments on the rather challenging *Vaihingen* dataset we obtain $> 80\%$ overall labeling accuracy. We show that powerful context features are vital for good urban classification and outperform standard multi-scale texture filters by a large margin, and that the proposed higher-order priors further reduce the classification error.

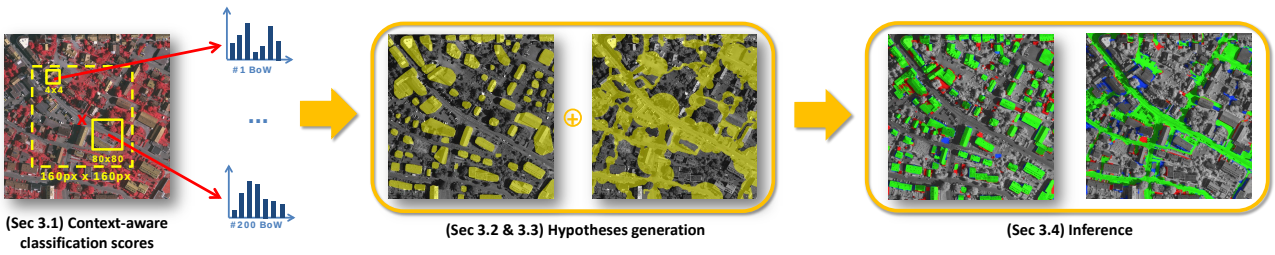


Figure 1: Given an input image, our method first classifies pixels into multiple labels (Sec. 3.1). Next, an over-complete set of building and road candidates is generated (Sec. 3.3 & 3.2). Finally, the candidates are pruned to an optimal subset (Sec. 3.4).

2. RELATED WORK

There is an enormous literature about pixelwise classification of remote sensing imagery, although a large part of it deals with low-resolution images, e.g. from satellites or from hyper-spectral sensors. For an overview, please refer to textbooks such as (Richards, 2013). More recently, there has also been an increasing interest in urban object detection in high-resolution urban scenarios, see for example (Rottensteiner et al., 2014) for an overview.

One possible solution for the interpretation of high-resolution urban data are rule-based approaches (production systems, semantic nets). These methods design custom rules to encode the a-priori knowledge for specific classes. Building detection and extraction has often been approached in 3D, based on multiple views (Herman and Kanade, 1984, Weidner, 1997, Fischer et al., 1998). Building knowledge in 2D has typically been used in a rule-based manner, by assembling edges or image segments to building regions with Gestalt-like grouping rules, e.g. (Fua and Hanson, 1987, Mohan and Nevatia, 1989). In early road detection work, e.g. (Fischler et al., 1981, Stilla, 1995, Steger et al., 1995), roads of specific width, direction, and contrast are extracted by linking responses to gradient or line filters, in the manner of multi-scale line detectors. Such putative road segments have also been combined with locally detected quadrilateral road pieces (Hinz et al., 1999). Some of these works propose to repair gaps in the road network by a minimum-cost path search between high-confidence pieces, somewhat similar to our hypothesis generator (see below). Rule-based methods set hard thresholds at intermediate steps, i.e. evidence that is lost at an early stage can hardly be recovered later on in the process.

To be more robust against noise and missing evidence, probabilistic models aim to avoid hard thresholds. Object knowledge is modeled as a prior distribution over the pixel labels, which is combined with the data likelihood generated by a per-pixel classifier. Probabilistic inference then balances data-driven evidence and the priors. One probabilistic formulation that combines data and topological object knowledge are Marked Point Processes (MPP), used for instance to extract road networks (Stoica et al., 2004, Chai et al., 2013) and building outlines (Ortner et al., 2007). MPPs lead to hard optimization problems (even if object hypotheses are sampled in a data-driven manner (Verdié and Lafarge, 2014)), which can only be solved approximately, and with high computational cost.

Another possibility to model contextual relations are graphical models, especially conditional random fields (CRFs). As opposed to MPPs they are amenable to efficient inference methods such as message passing or graph cuts. (Zhong and Wang, 2007) design a classification framework consisting of multiple CRFs to detect settlement areas in optical satellite images. (Roscher et al., 2010) use Import Vector Machines with CRFs to classify regions of Landsat TM images into multiple land cover classes. (Hoberg et al., 2010) adapt CRFs to multi-temporal multi-class land cover classification by adding temporal interactions to the standard unary and spatial potentials. Only few works exist that apply CRFs to semantic segmentation in urban scenes. (Kluckner et al., 2009) propose an efficient method for multi-label segmentation of aerial images. Covariance descriptors are fed into a Random Forest classifier and contextual information is modeled with a Conditional Random Field. The same labeling method has also been applied at the level of super-pixels (Kluckner and Bischof, 2010).

In our previous work, higher-order P^N-Potts potentials (Kohli et al., 2008) are introduced to represent roads in a CRF energy for road network extraction. Putative roads are either straight line segments and triple junctions (Wegner et al., 2013), or minimum-cost paths (Montoya-Zegarra et al., 2014). In this paper we extend the latter to also include building extraction.

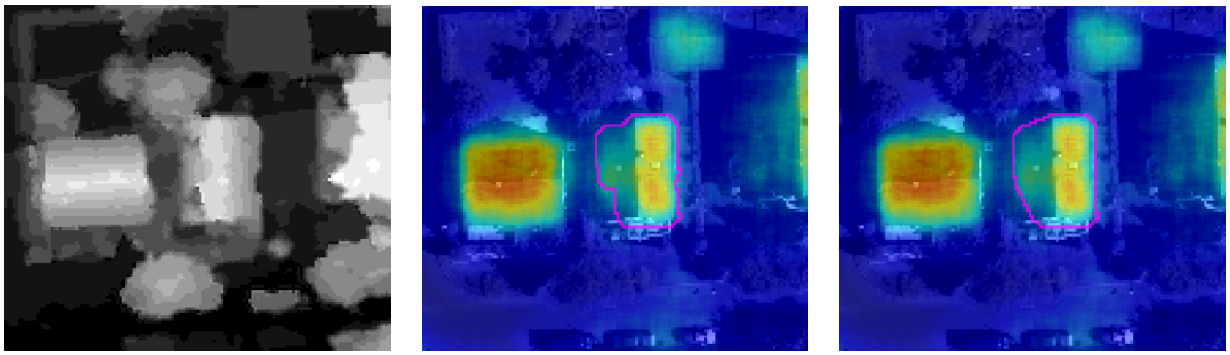
3. MODEL

An overview of our proposed framework is depicted in Figure 1. Given an input aerial image, we run a multi-label classifier that is trained over rich appearance features extracted over large spatial neighborhoods (cf. Subsec. 3.1). The large spatial neighborhoods allow to learn expressive local co-occurrences of feature patterns directly from the data. Consider, for example, the case of roads which are usually surrounded by buildings or trees. Due to the large contextual window of the classifier, building boundaries may be slightly blurred or two closely located buildings can be merged to one. Roads tend to have short gaps that disconnect single road segments. With the next steps we aim at (i) recovering polygonal building shapes and (ii) at fully linking all road pieces to the network while retaining high pixel-wise accuracies.

We follow a recover-and-select strategy. During the recover step we generate an over-complete representation that covers as much as possible of roads and buildings with suitable candidates. Because some candidates may partially cover other objects (e.g., trees, grass), we select the subset that best explains the road and building evidences by energy minimization in a CRF.

We begin with sampling sets of relevant road and building candidates in a data-driven fashion. To extract suitable building candidates, we threshold building likelihoods and seek connected components. The shape of each connected component is approximated with α-shapes (Edelsbrunner et al., 1983) at multiple generalization levels. We obtain multiple overlapping building candidates per connected component at different levels of detail (cf. Subsec. 3.2). Road candidates (*paths* and *blobs*) are generated as in (Montoya-Zegarra et al., 2014) (cf. Subsec. 3.3).

In the *select* step global energy minimization balances unaries and prior term, such that those road and building candidates that have most supporting evidence from the data are boosted (cf. Subsec. 3.4). Each sampled road and building candidate is mapped



(a) (b) (c)

Figure 2: Example of building candidate generation with α-shapes based on (a) the nDSM and (b,c) unary classifier scores (the more red, the higher the building likelihood) with $\alpha = 1$ at (b) and $\alpha = \infty$ at (c).

to a higher-order clique with a robust P^N-Potts potential (Kohli et al., 2008) that encourages clique members to take on the dominant label of the clique. This labeling problem can be solved approximately with α-expansion.

3.1 Local context-aware multi-class scores

We adopt the multi-feature extension (L'ubor Ladický et al., 2009) of the Textonboost classifier (Shotton et al., 2006) to obtain multi-class scores. To do so, a set of dense appearance features are extracted (SIFT (Lowe, 2004), Local Ternary Patterns (Hussain and Triggs, 2012), Textons (Malik et al., 2001), and Self-similarity (Shechtman and Irani, 2007)) and softly quantized into corresponding dictionaries of 512 visual words. To encode local context, a large spatial neighborhood of 160×160 pixels is centered over each pixel and a fixed set of random-generated rectangles is sampled (4×4 to 80×80 pixels). Over each of the 200 sampled rectangles, a histogram of visual words is extracted. The concatenation of the histograms extracted from the random set of rectangles form a feature vector for a single pixel. All extracted histograms are fed to a Boosting Classifier and the raw class scores are mapped to probabilities with a sigmoid function.

3.2 Generation of building candidates

Our goal is to generate a set of building candidates with plausible geometry. Pixel-wise building scores from the unary classifier (Sec. 3.1) well localize buildings, but most pixels on building boundaries have rather uncertain scores. Transition between objects tend to be irregular, due to fluctuations in the pixel-wise likelihoods. Consequently, buildings tend to have jaggy boundaries rather than regular outlines consisting mostly of straight line segments. The building prior shall favour candidates with simple polygonal shapes. First, we seek promising building candidates based on the unary classifier scores. Second, each candidate acts a seed for a set of building shape candidates that are computed exploiting edge information of the normalized digital surface model (nDSM). This models the assumption that abrupt jumps at building boundaries optimally describe sharp building edges.

We start with thresholding the building likelihoods, colored blue (low building likelihood) and red (high building likelihood) in Fig. 2. As a result we obtain a binary mask of connected components that have high building likelihood. Due to the smooth output of the classifier adjacent buildings are sometimes merged. To separate erroneously merged buildings we apply the Watershed algorithm (Meyer and Beucher, 1990) per connected component. For each connected component we extract local maxima and initialize the watershed segmentation at these points. In order to avoid over-segmentation (i.e., multiple closely located local maxima) we smooth the original building likelihoods prior to applying the segmentation. We flood regions based on the gradients of the nDSM, that is watershed lines correspond to distinct height jumps. The generated building segments are then approximated with α-shapes (Edelsbrunner et al., 1983) to generate a set of concave to convex building candidates per segment. By using different values for α we generate multiple generalizations per building segment that will each act as a clique of the CRF energy function of Eq. 1.

3.3 Generation of road candidates

Road candidates are generated as in our previous work (Montoya-Zegarra et al., 2014). First, given a set of road likelihoods obtained from the per-pixel classifier, we compute Laplacian-of-Gaussian (LoG) responses at multiple consecutive scales (Lindeberg, 1994) that well cover the expected range of road widths. We then train a Random Forest classifier (on ground truth) that takes mean, median, and standard deviation of the pixel-wise LoG responses within the LoG filter radius as input. This multi-scale classifier generates a $(x, y, width)$-volume of 3D road likelihoods in which the expected road widths are discretized across scales. Second, minimum cost paths through the volume are computed with the 3D Fast Marching algorithm (Deschamps and Cohen, 2001) to connect road likelihoods to elongated *paths*. Additionally, big *blobs* are sampled from the volume to cover large junctions or squares that cannot be modeled with elongated paths alone. All *paths* and *blobs* together constitute the set of road candidates.

3.4 Maximum a-posteriori labeling

The final step is to infer class labels for all pixels, given their individual class likelihoods from Sec. 3.1 as well as the set of road and building hypotheses. We formulate this as inference in a higher-order CRF which fulfills the robust P^N-Potts model (Kohli et al., 2008), i.e. the higher-order cliques encourage their member pixels to all have the same label, and penalize deviations from the majority label.

Alternatively, this can be thought of as selecting a subset of object candidates that best explain the image evidence, while at the same time correcting the membership of individual pixels in those candidates, but only if the correction is supported by strong enough evidence.

Maximum a-posteriori (MAP) inference in the CRF is equivalent to minimizing the corresponding Gibbs energy, which in our case

is composed of four terms:

$$E = \sum_i E_u(x_i) + \sum_i \sum_{j \in \mathcal{N}(i)} E_p(x_i, x_j) + \\ + \sum_m E_R(Q_m) + \sum_n E_B(Q_n) \ . \quad (1)$$

In this expression, $E_u = -\log P(l_k|x_i)$ denotes the unary term, i.e. the cost of assigning label l_k to pixel x_i. A pairwise term $E_p = [l_i \neq l_j] f(\nabla I_{ij}, \nabla H_{ij})$ encourages local smoothness of the classification through a conventional contrast-sensitive Potts penalty. Here $[\cdot]$ is the Iverson bracket, which returns 1 of the enclosed expression is true and 0 otherwise, and ∇I_{ij} and ∇H_{ij} are the gradients of the image (respectively of the normalized DSM) between pixels x_i and x_j; and $f(\cdot)$ is a linear truncated function that maps the gradient magnitude to a cost.

Furthermore, each of the sampled building and road candidates forms a large clique Q_b, respectively Q_r, over the candidates' member pixels. The clique costs E_R and E_B are robust P^N-Potts potentials. The effect of those cliques is as follows: if within a clique there exists enough evidence for the *road* or *building* class, then also the dissenting member pixels are drawn to that class. Hence, false negatives covered by a clique are corrected. On the contrary, if in a clique there are too few pixels with a preference for *road* or *building*, then also those pixels are drawn to the competing majority label, thereby correcting false positives.

The mapping from the number of deviating pixels to the cost is again linear truncated function, $E(Q) = \min(u, N_k \cdot \frac{u-v}{w} + v)$ with $\{u, v, w\}$ the parameters that define the amount of truncation, and N_k the number of pixels that take on label l_k. In our experiments, we fix the $\{u, v, w\}$-parameters to the same values $\{10, 7, 0.45\}$ for both road cliques E_B and buildings E_R.

The P^N-Potts model (Eq. 1) is amenable to graph cuts. For two labels exact MAP inference is possible efficiently with the min-cut algorithm. For our multi-label problem, a strong local minimum of the energy can be found with α-expansion.

4. EXPERIMENTAL RESULTS

We evaluate our experiments on 1000×1000 pixels tiles of a true orthophoto mosaic (0.25 m pixel size on ground) generated via dense matching from the Vaihingen data set[1]. What makes this scene challenging are (i) many small buildings which are often densely clustered, and that vary strongly in shape and (ii) road networks are irregular, mainly narrow and partially occluded by cast shadows or trees (examples of tiles and nDSM are shown in Fig. 3, ground truth and classification results in Fig. 4). Four tiles are used for training the unary classifier (Sec. 3.1), another four tiles for training the classifier that predicts road-widths (Sec. 3.3), and eight images for testing. For quantitative evaluation we report pixel-wise classification accuracy (Tab. 4.) in terms of precision, recall, and F1 scores for all six classes Asphalt (grey), Background (red), Roads (white), Trees (green), Grass (turquoise), and Buildings (blue).

For building candidate generation we vary the α parameter in an (approximately) exponential sequence (α={1,3,5,7,11,15,25,50, 100,∞}). Note that smaller α values generate concave polygons whereas a higher α leads to more convex candidate shapes. As $\alpha \to \infty$ the shape becomes the convex hull of the object. On

[1] Vaihingen is part of the ISPRS benchmark *http://www.itc.nl/ISPRS_WGIII4/tests_datasets.html* which comprises aerial images covering a semi-urban region in southern Germany

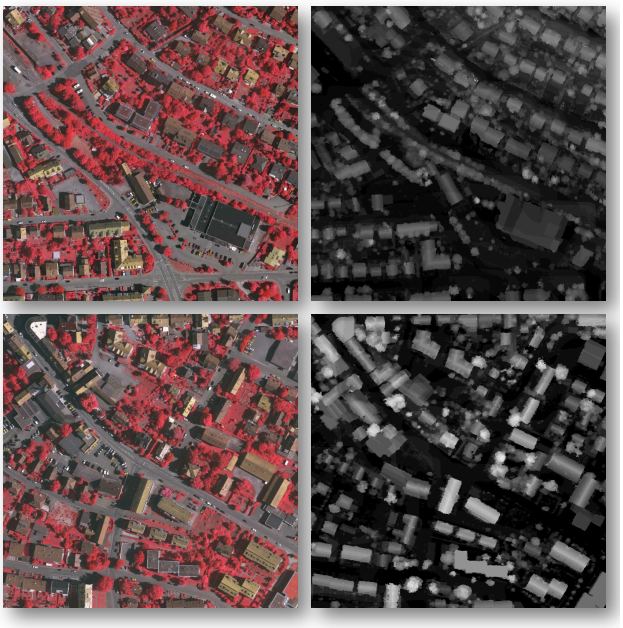


Figure 3: False-color image tiles (left column) and corresponding normalized DSMs (right column).

average, $250 \times n_\alpha \approx 2500$ (with n_α the number of α-shapes per connected component) candidates are generated per 1000×1000 tile.

We use a k-shortest path variation of Fast Marching for *path* road candidate computation. For each pair of seed points we sample k-mutually exclusive paths. This allows us to cover as much as possible of the roads. In our experiments we compute 2000 *paths* (two paths connect each of 1000 node pairs) and 650 *blobs* per tile.

As baseline (*Winn*) we train and predict multi-class labels using a multi-label Random Forest with 20 trees based on pixel-wise responses to the filterbank of (Winn et al., 2005).

The (unary) classifier (*Context*) alone already achieves per-pixel recall and precision above 80% for buildings and roads (Tab. 4.). It clearly outperforms the *Winn* baseline for all classes.

We note that $> 80\%$ recall at $> 80\%$ precision is sometimes quoted as the necessary performance to make automatic methods practically useful (Mayer et al., 2006).

Class-specific prior construction further improves the F1-scores for roads and buildings. The slightly lower performance of the other classes is due to the asymmetric nature of our potentials. It should be noted that visually significant improvements of building shapes (Fig.5 and road network topology (Fig. 6) result in only marginal improvements of pixel-wise scores, due to the small number of affected pixels (relative to the image size). In Fig. 5, right frame, false negatives are significantly reduced and building shapes resemble ground truth more closely. However, in the left frame we can also observe that, although false positives are significantly reduced, very small buildings can be lost completely if their unaries are weak, because the prior then votes for a different class which has more support in the clique. Roads extraction is improved, too (Fig. 6). False positives (e.g., left frame) and false negatives (three right frames) are both reduced. In the three right frames, where road is confused with building, the class-specific building and road priors fruitfully cooperate. On the one hand, the *path* cliques close road network gaps, while on the other hand α-shape cliques repair building shapes.

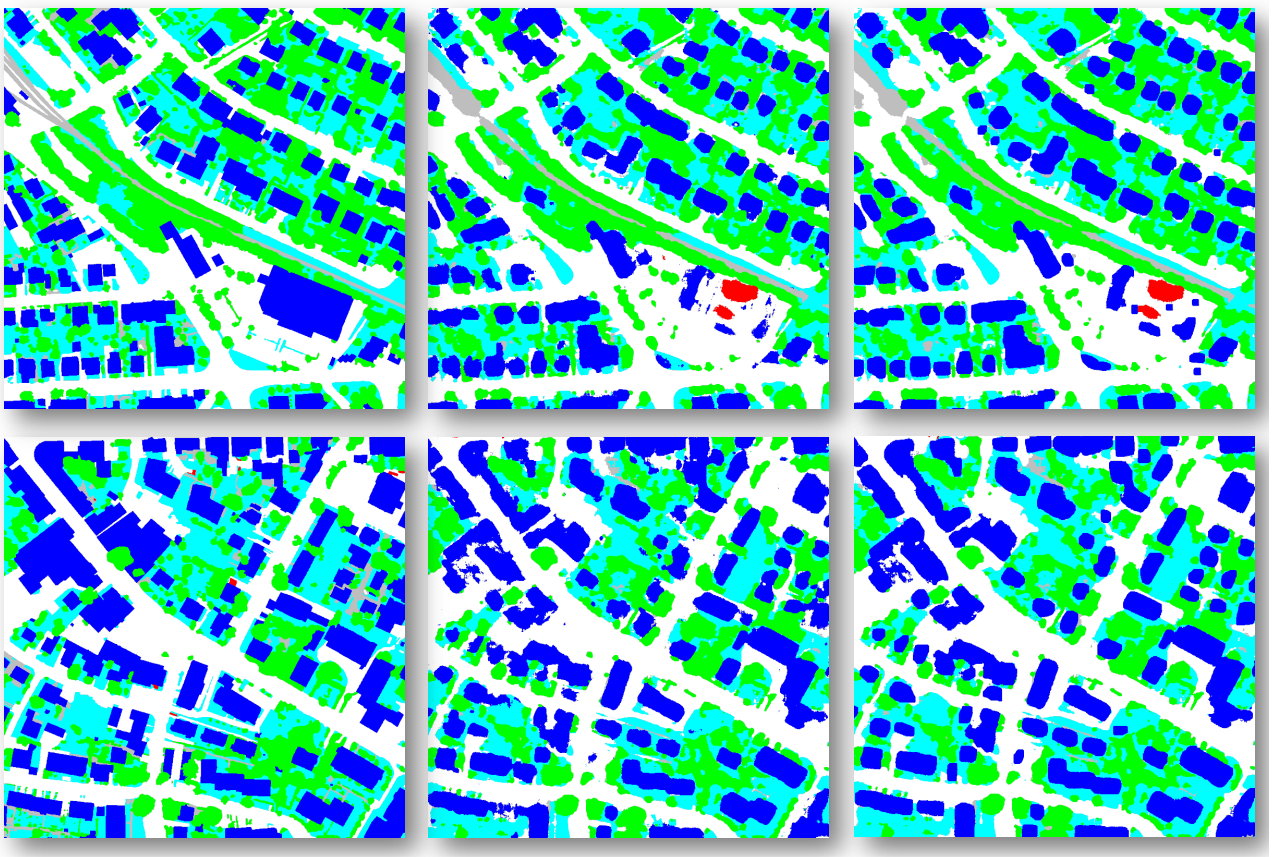


Figure 4: Multi-class semantic segmentation results for the two tiles shown in Fig. 3: ground truths (left column), predictions from the unary classifier of Sec. 3.1 (center column), and final results after CRF inference with class-specific priors for buildings and roads (right column).

	Buildings			Roads			Asphalt			Grass			Trees			Background		
Method	*F1*	*Rec.*	*Prec.*	*F1*	*Rec.*	*Prec.*	*F1*	*Rec.*	*Prec*	*F1*	*Rec.*	*Prec.*	*F1*	*Rec.*	*Prec.*	*F1*	*Rec.*	*Prec.*
Winn	74	72	77	76	73	80	3	14	2	59	63	57	79	77	81	38	55	33
Context	84	83	86	85	82	89	13	23	11	**69**	76	64	**84**	82	86	**46**	47	45
CRF	**86**	87	85	**86**	83	89	**14**	21	12	68	71	65	83	81	86	45	46	45

Table 1: Pixelwise performance for multi-label classification on the Vaihingen dataset. The **overall accuracy** for Winn, Context, and CRF are respectively: **73.69**, **82.35**, and **82.42**. All numbers are percentages.

5. CONCLUSIONS AND FUTURE WORK

We have proposed a multi-label classification with class-specific priors for buildings and roads. In addition to the road network prior of (Montoya-Zegarra et al., 2014), we have added a second higher-order potential for cliques custom-tailored to buildings. Experiments show that the road and building layers can be jointly improved with these class-specific priors within a CRF framework.

At present, clique sampling is done in a data-driven, rather heuristic way. Moreover, object candidate generation (a large set of alpha-shapes per building and the LoG scale-space volume for road widths) is cast as a (discrete) classification task. However, shape parameters for buildings and road widths are continuous variables and it seems more intuitive to directly formulate parameter estimation as regression. A future idea would thus be to *perform classification of object categories and regression of their shape parameters jointly* within a unified framework. A natural starting point are structured prediction methods such as for instance Hough Forests (Gall et al., 2011).

REFERENCES

Chai, D., Förstner, W. and Lafarge, F., 2013. Recovering Line-networks in Images by Junction-Point processes. In: CVPR.

Deschamps, T. and Cohen, L. D., 2001. Fast extraction of minimal paths in 3d images and applications to virtual endoscopy. Medical Image Analysis 5(4), pp. 281–299.

Edelsbrunner, H., Kirkpatrick, D. and Seidel, R., 1983. On the shape of a set of points in the plane. IEEE Transactions on Information Theory 29(4), pp. 551–559.

Fischer, A., Kolbe, T. H., Lang, F., Cremers, A. B., Förstner, W., Plümer, L. and Steinhage, V., 1998. Extracting buildings from aerial images using hierarchical aggregation in 2d and 3d. Computer Vision and Image Understanding 72(2), pp. 185–203.

Fischler, M., Tenenbaum, J. and Wolf, H., 1981. Detection of roads and linear structures in low-resolution aerial imagery using a multisource knowledge integration technique. Computer Graphics and Image Processing 15, pp. 201 – 223.

Fua, P. and Hanson, A. J., 1987. Using generic geometric models for intelligent shape extraction. In: Proceedings of the Sixth National Conference on Artificial Intelligence, pp. 706–709.

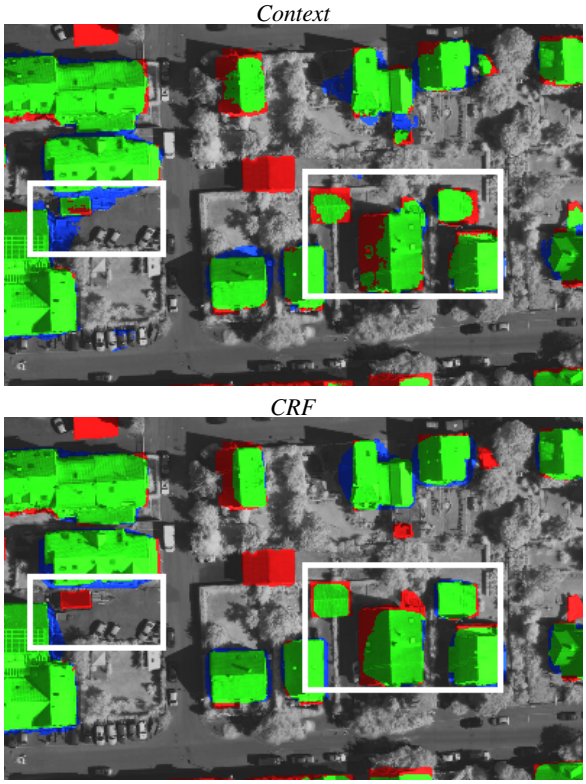


Figure 5: Example of extracted buildings. True positives are displayed green, false positives blue, and false negatives red. White frames highlight significant differences/improvements.

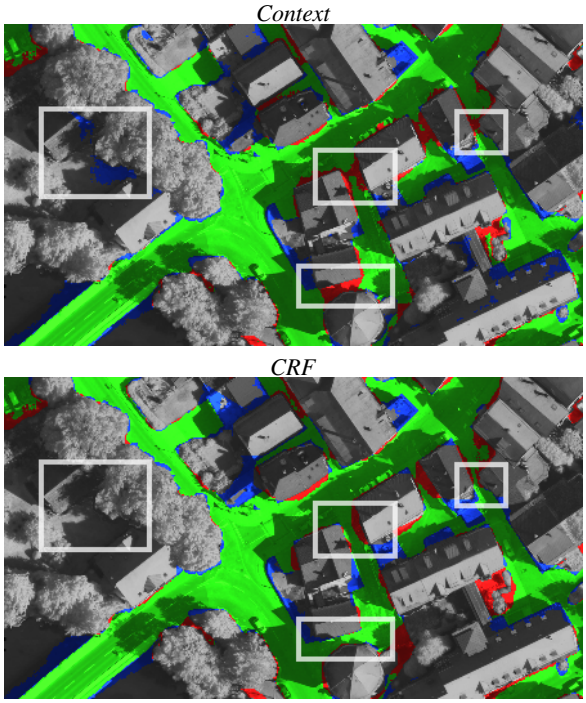


Figure 6: Example of extracted roads. True positives are displayed green, false positives blue, and false negatives red. White frames highlight significant differences/improvements.

Gall, J., Yao, A., Razavi, N., van Gool, L. and Lempitsky, V., 2011. Hough Forests for Object Detection, Tracking, and Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(11), pp. 2188 – 2202.

Herman, M. and Kanade, T., 1984. Image understanding workshop. chapter The 3D MOSAIC scene understanding system: Incremental reconstruction of 3D scenes from complex image, pp. 137–148.

Hinz, S., Baumgartner, A., Steger, C., Mayer, H., Eckstein, W., Ebner, H. and Radig, B., 1999. Road extraction in rural and urban areas. In: Semantic Modeling for the Acquisition of Topographic Information from Images and Maps, pp. 133–153.

Hoberg, T., Rottensteiner, F. and Heipke, C., 2010. Classification of multitemporal remote sensing data using conditional random fields. In: 6th IAPR Workshop on Pattern Recognition in Remote Sensing.

Hussain, S. u. and Triggs, B., 2012. Visual recognition using local quantized patterns. In: European Conference on Computer Vision.

Kluckner, S. and Bischof, H., 2010. Image-based building classification and 3D modeling with super-pixels. In: International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. 38(3A), pp. 233 – 238.

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

Kohli, P., L'ubor Ladický and Torr, P. H. S., 2008. Robust higher order potentials for enforcing label consistency. In: CVPR.

Kohli, P., L'ubor Ladický and Torr, P. H. S., 2009. Robust higher order potentials for enforcing label consistency. IJCV 82(3), pp. 302–324.

Lindeberg, T., 1994. Scale-space theory: A basic tool for analysing structures at different scales. Journal of Applied Statistics pp. 224–270.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision.

Malik, J., Belongie, S., Leung, T. and Shi, J., 2001. Contour and texture analysis for image segmentation. International Journal of Computer Vision.

Mayer, H., Hinz, S., Bacher, U. and Baltsavias, E., 2006. A test of automatic road extraction approaches. In: IAPRS, Vol. 36(3), pp. 209 – 214.

Meyer, F. and Beucher, S., 1990. Morphological segmentation. Journal of Visual Communication and Image Representation 1(1), pp. 21–46.

Mohan, R. and Nevatia, R., 1989. Using perceptual organization to extract 3d structures. Transactions on Pattern Analysis and Machine Intelligence 11(11), pp. 1121–1139.

Montoya-Zegarra, J. A., Wegner, J. D., L'ubor Ladický and Schindler, K., 2014. Mind the gap: modeling local and global context in (road) networks. In: German Conference on Pattern Recogntion, pp. 212–223.

Ortner, M., Descombes, X. and Zerubia, J., 2007. Building Outline Extraction from Digital Elevation Models Using Marked Point Processes. IJCV 72(2), pp. 107 – 132.

Richards, J. A., 2013. Remote Sensing Digital Image Analysis: An Introduction. 5th edn, Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Roscher, R., Waske, B. and Förstner, W., 2010. Kernel discriminative random fields for land cover classification. In: 6th IAPR Workshop on Pattern Recognition in Remote Sensing.

Rottensteiner, F., Sohn, G., Gerke, M., Wegner, J. D., Breitkopf, U. and Jung, J., 2014. Results of the isprs benchmark on urban object detection and 3d building reconstruction. ISPRS Journal of Photogrammetry and Remote Sensing 93(0), pp. 256–271.

Shechtman, E. and Irani, M., 2007. Matching local self-similarities across images and videos. In: Conference on Computer Vision and Pattern Recognition.

Shotton, J., Winn, J., Rother, C. and Criminisi, A., 2006. TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: ECCV.

Steger, C., Glock, C., Eckstein, W., Mayer, H. and Radig, B., 1995. Model-based road extraction from images. In: Automatic Extraction of Man-Made Objects from Aerial and Space Images, Birkh auser Verlag Basel, Birkhauser Verlag, pp. 275–284.

Stilla, U., 1995. Map-aided structural analysis of aerial images. ISPRS Journal of Photogrammetry and Remote Sensing 50(4), pp. 3–10.

Stoica, R., Descombes, X. and Zerubia, J., 2004. A Gibbs Point Process for road extraction from remotely sensed images. IJCV 57(2), pp. 121 – 136.

Verdié, Y. and Lafarge, F., 2014. Detecting parametric objects in large scenes by Monte Carlo sampling. IJCV 106, pp. 57 – 75.

L'ubor Ladický, Russell, C., Kohli, P. and Torr, P. H., 2010. Graph cut based inference with co-occurrence statistics. In: ECCV.

L'ubor Ladický, Russell, C., Kohli, P. and Torr, P. H. S., 2009. Associative hierarchical CRFs for object class image segmentation. In: ICCV.

Wegner, J. D., Montoya-Zegarra, J. A. and Schindler, K., 2013. A higher-order CRF model for road network extraction. In: CVPR.

Weidner, U., 1997. Digital surface models for building extraction. In: Automatic Extraction of Man-Made Objects from Aerial and Space Images (II), pp. 193–202.

Winn, J., Criminisi, A. and Minka, T., 2005. Object categorization by learned universal visual dictionary. In: CVPR.

Zhong, P. and Wang, R., 2007. A Multiple Conditional Random Fields Ensemble Model for Urban Area Detection in Remote Sensing Optical Images. IEEE Transactions on Geoscience and Remote Sensing 45(12), pp. 3978 – 3988.