3D Classification of Crossroads from Multiple Aerial Images using Conditional Random Fields

Sergey Kosov, Franz Rottensteiner, and Christian Heipke
Institute of Photogrammetry and GeoInformation (IPI), Hanover, Germany
kosov@ipi.uni-hannover.de, rottensteiner@ipi.uni-hannover.de, heipke@ipi.uni-hannover.de

Abstract

In this paper we apply Conditional Random Fields to the classification of scenes containing crossroads. We use a simple appearance-based model in combination with a probabilistic model of the co-occurrence of class labels at neighbouring image sites to distinguish different classes that are relevant for scenes containing crossroads. We make use of multiple overlap aerial images to derive a digital surface model and a true orthophoto without dynamic objects such as cars. Our approach is evaluated on a data set of airborne photos of an urban area by a comparison of the results to reference data and it is shown to produce promising results.

1 Introduction

The automatic reconstruction of roads has been an important topic of research in Remote Sensing for a long time. Whereas road extraction methods are reliable under favourable conditions, e.g. in rural areas, they are far from being practically relevant in more challenging environments, e.g. in suburban regions. Failure is often related to crossroads, where model assumptions about roads are violated [9]. Thus, specific models for the extraction of crossroads from images have to be developed. This has been tackled in [1], where neural networks are used for a supervised per-pixel classification of greyscale orthophotos to detect areas corresponding to crossroads. However, only examples for rural areas are shown. In [11], a model based on snakes is used to reconstruct crossroads. The main reasons for failure of that method were occlusion of the road surface by cars and a complex 3D geometry, e.g. at motorway interchanges. To overcome these problems, 3D information, e.g. in the form of a *Digital Surface Model* (DSM), has to be used. Furthermore, *context* should be considered in the classification of the image content. This can be achieved by Markov Random Fields (MRF), which have been used for some time as a probabilistic model for local context [7]. More recently, Conditional Random Fields (CRF) were introduced to avoid the problems of MRFs with oversmoothing in areas where the image content changes abruptly [6]. In remote sensing, CRF have been used for the detection of buildings in optical and SAR images [14], for the classification of optical satellite images [4], and for the generation of a Digital Terrain Model (DTM) from airborne laserscanner data [8]. In this paper we propose a new method for the classification of scenes containing crossroads as a first step of a 3D reconstruction. CRF are employed for a raster-based classification. We use multiple-overlap aerial images in order to derive a DSM that is used in classification to make it more robust with respect to ambiguities of the appearance of objects in a 2D projection of the scenes. Due to the multiple overlap images, we can solve the problem of occlusions of the road surface by moving cars. Our method is evaluated using the Vaihingen data set of the German Society of Photogrammetry, Remote Sensing and Geoinformation (DGPF) [2], comparing three different variants of the classifier to assess the impact of considering context in classification.

2 Conditional Random Fields (CRF)

CRF are undirected graphical models that can be used to consider context for the image labelling problem by modelling statistical dependencies between the labels and the data at neighbouring image sites [6]. Given image data $\mathbf{y}$ consisting of M image sites $i \in \mathbb{S}$ with observed data y_i, i.e., $\mathbf{y} = (y_1, y_2, \ldots, y_M)^T$, where $\mathbb{S}$ is the set of all sites, we want to assign a discrete class label x_i from a given set of classes $\mathbb{C}$ to each site i (e.g., an individual pixel or a segment). Collecting the class labels x_i in a vector $\mathbf{x} = (x_1, x_2, \ldots, x_M)^T$, we want to find the label configuration $\hat{\mathbf{x}}$ that maximises the posterior probability of the labels given the data $p(\mathbf{x}|\mathbf{y})$, thus $\hat{\mathbf{x}} = \arg\max_x p(\mathbf{x}|\mathbf{y})$. CRF are discriminative models that directly model the posterior probability $p(\mathbf{x}|\mathbf{y})$ [6]:

$$p(\mathbf{x}|\mathbf{y}) = \frac{1}{Z} \cdot \exp\Big[\sum_{i \in \mathbb{S}} \varphi_i(x_i, \mathbf{y}) + \sum_{i \in \mathbb{S}} \sum_{j \in \mathcal{N}_i} \psi_{ij}(x_i, x_j, \mathbf{y})\Big]. \tag{1}$$

In Eq. 1, Z is a normalization constant and $\mathcal{N}_i$ is the neighbourhood of data site i (thus, j is a neighbour of i). The *association potential* φ_i links the class label x_i of image site i to the data $\mathbf{y}$, whereas the *interaction potential* ψ_{ij} models the dependencies between the labels $(x_i,\ x_j)$ of neighbouring sites i and j and the data $\mathbf{y}$. The model is very general in terms of the definition of the functional models for both φ_i and ψ_{ij}. Our definitions of the image sites, the neighbourhood $\mathcal{N}_i$ and the potentials φ_i and ψ_{ij} are described in Section 3.

3 Method

The primary input of our method consists of multiple aerial colour infrared (CIR) images and their orientation data. We require at least fourfold overlap from two different image strips for each crossroads to avoid occlusions. Using the OpenCV implementation of semi-global block matching [10], we obtain a raw DSM from each possible image pair. These raw DSMs are combined to a joint DSM, and remaining void areas are filled by an in-painting algorithm [5]. Using the DSM, a true orthophoto is generated from each original image. Merging the resulting raw orthophotos to a combined orthophoto, we eliminate moving cars by taking the median color vectors of the multiple raw images [5].

Both the DSM and the combined true orthophoto are used as the input to the CRF-based classifier. In the classification process, we choose the image sites and, thus, the nodes of the graphical model, to correspond to image patches of $n \times n$ pixels of the true orthophoto. The neighbourhood $\mathcal{N}_i$ of an image site i in Eq. 1 (which defines the edges of the graphical model) is chosen to consist of the four direct neighbours of site i in the image grid. We define six classes that are characteristic for scenes containing crossroads, namely *asphalt* (*asp.*), *building* (*bld.*), *tree* (*tr.*), *grass* (*gr.*), *agricultural* (*agr.*) and *car*. From the orthophoto and the DSM we extract the feature vectors for classification. In a training phase we determine the parameters of the association and interaction potentials in Eq. 1, which requires fully labelled training images. Then, the classification of new images can be carried out by maximizing the posterior probability in Eq. 1.

The association potential $\varphi_i(x_i, \mathbf{y})$ in Eq. 1 is related to the probability of a label x_i taking a value $c \in \mathbb{C}$ given the data $\mathbf{y}$ by $\varphi_i(x_i, \mathbf{y}) = \log p(x_i = c \mid \mathbf{f}_i(\mathbf{y}))$ [6], where the image data are represented by site-wise feature vectors $\mathbf{f}_i(\mathbf{y})$ that may depend on all the observed data $\mathbf{y}$. Note that both the definition of the features and the dimension of the feature vectors $\mathbf{f}_i(\mathbf{y})$ may vary with the dataset. We use a simple model for $p(x_i = c \mid \mathbf{f}_i(\mathbf{y}))$ that is based on a Bayesian classifier with uniform prior on the class labels, thus $p(x_i = c \mid \mathbf{f}_i(\mathbf{y})) \propto p(\mathbf{f}_i(\mathbf{y}) \mid x_i = c)$ and, neglecting terms that are constant over the classes, $\varphi_i(x_i, \mathbf{y}) = \log p(\mathbf{f}_i(\mathbf{y}) \mid x_i = c)$. In the training phase, for each class we generate histograms of all features. These histograms are smoothed and normalised, and the smoothed and normalised histograms are used as probability density functions (pdf) $p(f_{ij} \mid x_i = c) \equiv p_c(f_{ij} \mid x_i)$ for the class c, where f_{ij} is the j^{th} component of $\mathbf{f}_i$. Neglecting the statistical dependencies between the individual features f_{ij}, the association potential becomes:

$$\varphi(x_i = c, \mathbf{y}) = \sum_{j=1}^{N} \log \Big[p_c(f_{ij} \mid x_i)\Big]. \tag{2}$$

In Eq. 2, N is the dimension of the feature vectors $\mathbf{f}_i(\mathbf{y})$. This is a very simplistic model, which is to be replaced by more appropriate ones in the future. Its advantage is that it is very fast to determine in training.

The interaction potential $\psi_{ij}(x_i, x_j, \mathbf{y})$ in Eq. 1 describes how likely x_i is to take the value c given that the label x_j from the neighbouring data site $j \in \mathcal{N}_i$ takes the value c' and given the data: $\psi_{ij}(x_i, x_j, \mathbf{y}) = \log p(x_i = c \mid x_j = c', \mathbf{y})$ [6]. We generate a 2D histogram $h'(x_i,\ x_j)$ of the co-occurrence of labels at neighbouring image sites from the training data. That is, $h'(x_i = c, x_j = c')$ is the number of occurrences of the classes $(c,\ c')$ at neighbouring pixels i and j. After that, the rows of $h'(x_i,\ x_j)$ are scaled so that the largest value in a row will be one, resulting in a matrix $h(x_i,\ x_j)$ that is the basis for the interaction potential. Scaling is necessary to avoid a bias for classes covering a large area in the training data. We determine the Euclidean distance between the feature vectors $\mathbf{f_i}$ and $\mathbf{f_j}$ at the neighbouring image sites i and j, $d_{ij} = \|\mathbf{f}_i(\mathbf{y}) - \mathbf{f}_j(\mathbf{y})\|$. Our definition of $\psi_{ij}(x_i, x_j, \mathbf{y}) \equiv \psi_{ij}(x_i, x_j, d_{ij})$ is obtained by multiplying the diagonal elements of $h(x_i,\ x_j)$ by a weight depending on d_{ij} and taking the logarithms:

$$\psi_{ij}(x_i, x_j, \mathbf{y}) = \begin{cases} \log\Big[\frac{2\lambda}{\sqrt{\lambda^2 + d_{ij}^2}} \cdot h(x_i,\ x_j)\Big] & \text{if } x_i = x_j \\ \log\Big[h(x_i,\ x_j)\Big] & \text{otherwise} \end{cases} \tag{3}$$

In Eq. 3, the parameter λ determines the relative weight of the interaction potential compared to the association potential. As the largest entries of $h(x_i,\ x_j)$ are usually found in the diagonals, a model without the weight factor in Eq. 3 would favour identical class labels at neighbouring image sites and, thus, result in a smoothed label image. This will still be the case if the feature vectors at neighbouring image sites are identical. However, large differences between the features will reduce the impact of this smoothness assumption and make a class change between neighbouring image sites more likely.

Exact probabilistic methods for training of a CRF are computationally intractable [6, 13]. Thus, approximate solutions have to be used for training. We determine the parameters of the association and interaction potentials separately. Given the training data (fully labelled images), the probabilities $p_c(f_{ij}|x_i)$ are determined from smoothed histograms of the features f_{ij} of each class as described above. The interaction potentials are derived from scaled versions of the 2D histograms of the co-occurrence of classes at neighbouring image sites in the way described above. The parameter λ in Eq. 3 is set to a value $\lambda = 2$, determined empirically. Exact inference is also computationally intractable for CRFs. We use Loopy Belief Propagation, a standard technique for probability propagation in graphs with cycles [13].

4 Features

We derive a feature vector $\mathbf{f}_i(\mathbf{y})$ for each image site i that consists of features derived from the orthophoto and features derived from the DSM. For numerical reasons, all features are scaled linearly into the range between 0 and 255 and then quantized by 8 bit.

In total, we determine $N = 18$ features. The first three features are the *normalized difference vegetation index* (*NDVI*), derived from the near infrared and the red band of the CIR orthophoto, the *saturation* (*sat*) component after transforming the image to the LHS colour space, and *image intensity* (*int*), calculated as the average of the two non-infrared channels. These features are derived at three different scales, namely for the individual pixels and taking the average over 10×10 and 100×100 pixels, respectively; this results in altogether nine features ($NDVI_1$, sat_1, int_1, $NDVI_{10}$, sat_{10}, int_{10}, $NDVI_{100}$, sat_{100}, int_{100}). We also make use of the *variance of intensity* (var_{int}), the *variance of saturation* (var_{sat}) and the *variance of gradient* (var_{grad}) determined from a local neighbourhood of each pixel (7×7 pixels for var_{int}, 13×13 pixels for var_{sat} and var_{grad}). The 13^{th} feature (*dist*) models the fact that road pixels are usually found in a certain distance either from road edges or road markings. We generate an edge image by thresholding the intensity gradient of the input image. Then, we determine a distance map from this edge image. The *dist* feature is the distance of an image site to its nearest edge pixel.

The next group of features is based on histograms of oriented gradients (HOG) [3]. We calculate the HOG descriptors for cells consisting of 7×7 pixels, using blocks of 2×2 cells for normalization. Each histogram consists of 9 orientation bins (20° per bin). The gradient directions are related to the main direction of the entire scene, supposed to correspond to the direction of one of the intersecting roads. We extract three features from the HOG descriptor, namely the value corresponding to the main direction (HOG_0) and the values at its two neighbouring bins (HOG_{-1}, HOG_{+1}).

Finally, we determine a coarse DTM from the DSM by applying a morphological opening filter with a structural element whose size corresponds to the size of the largest off-terrain structure in the scene, followed by a median filter with the same kernel size. The last two features are the difference $nDSM$ between the DSM and the DTM, which describes the relative elevation of objects above ground, and the gradient strength of the DSM ($||\nabla DSM||$).

5. Experiments

For evaluation, we used a part of the Vaihingen data set of the DGPF [2]. We selected 81 crossroads visible in at least four of the CIR images. For each of them, we generated a DSM and a true orthophoto covering an area of 80×80 m^2 with a GSD of 8 cm. The image patches were squares of 5×5 pixels, so that each graphical model consisted of 200×200 nodes. The reference was generated by manually labeling the orthophotos using the 6 classes defined in Section 3. We used cross validation in our evaluation procedure: In each test run, 80 images were used for training, and the remaining one for testing. This was repeated so that each image was used as a test image once. The classification results were compared with the reference; we report the completeness and the correctness of the results per class as well as the overall classification accuracy [12].

We carried out three different experiments. In the first experiment (*NoEdge*), each node was classified solely based on the association potentials, thus setting the interaction potential $\psi_{ij}(x_i, x_j, \mathbf{y}) \equiv 0$. In the second experiment (*MRF*) we emulated a MRF using the Potts model [7] by defining the interaction potential to be $\psi_{ij}(x_i, x_j, \mathbf{y}) = \alpha$ if $i = j$ and $\psi_{ij}(x_i, x_j, \mathbf{y}) = 0$ otherwise, using $\alpha = 4.6$. In the third experiment (*CRF*) we use our CRF model with the interaction potential defined in Eq. 3. This comparison should show the impact of the respective formulation of the interaction terms. Fig. 1 shows the results for one crossroad.

The completeness and the correctness of the results achieved in the three experiments are shown in Tab. 1. In the first experiment (*NoEdge*), the overall accuracy of the classification was 66.3%. Fig. 1 also shows the local variation of the class labels, caused by a similar appearance of the classes in the labels that is not compensated by a smoothness term. For the second experiment (*MRF*), the overall accuracy was 70.2%, a value that could be increased to 72.0% in the third experiment (*CRF*). Obviously, the smoothing achieved by the Potts model (*MRF*) already has a positive impact on the classification accuracy, also expressed in the completeness and correctness values of all classes. Considering the data in the interaction terms in (*CRF*) improves the overall accuracy even further by avoiding oversmoothing at region boundaries having sufficient contrast. In particular, the classification of class *asphalt*, the one most relevant for the goal of the classification of crossroads, is improved considerably by avoiding confusions with class *car*. The use of CRF has doubled the *car* class correctens (please, refere to the Tab. 1) what has visibly improved the overall classification quality at Fig. 1. The main error source was a confusion of buildings with asphalt and of trees with grass due to errors in the DSM caused by areas with hardly any texture (buildings) or abrupt height changes (trees).

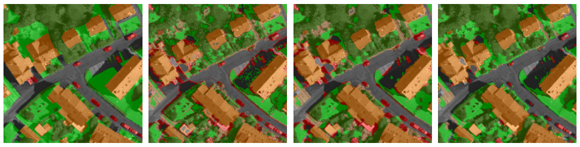

Figure 1. Left to right: reference; NoEdge; MRF; CRF. Grey: asp.; orange: bld.; dark green: tr.; green: gr.; beige: agr.; red: car.

	NoEdge		*MRF*		*CRF*	
	Cm.	*Cr.*	*Cm.*	*Cr.*	*Cm.*	*Cr.*
asp.	70.2	84.8	72.5	86.1	81.3	84.2
bld.	72.0	84.9	76.7	87.1	81.1	82.6
tr.	74.8	62.2	81.7	64.3	80.5	61.2
gr.	51.5	70.7	53.4	77.5	59.6	67.8
agr.	65.3	51.4	71.7	59.0	49.3	69.0
car	73.7	7.8	83.0	9.5	54.6	19.2
avg.	**66,3**		**70,2**		**72,0**	

Table 1. Completeness (Cm.) and Correctness (Cr.) [%] of the results.

Finaly, we measure the perormance of our CRF engine. In Tab. 2 timings for the different steps of our algorithms are given. The training step includes training on 80 crossroads, as other speps are given for one crossroad.

6 Conclusions

In this paper, a method for the classification of crossroads using CRF was proposed. It considered 3D information in the form of a DSM generated from multiple overlapping aerial images and free from dynamic objects. Distinguishing 6 classes relevant in the context of crossroads, an overall accuracy of about 72.0% could be achieved. The main error sources were the errors in DSM generation. The method described here is only a first step in a project aiming at a precise delineation of 3D road outlines. In the future we want to improve our method by using an improved CRF structure, considering the 3D structure of the scene also in the structure of the CRF graph. Furthermore, better models for the association and interaction potentials are required, as well as an integrated method for training.

step	*NoEdge*	*MRF*	*CRF*
training	5,7	5,7	9,0
buiding the graph	0,3	0,4	0,4
decoding	0	13,3	13,4
total	6,0	19,4	22,8

Table 2. Timings [sec] for an Intel® Core™ i7 CPU 950 with 3,07 GHz.

Acknowledgements

The Vaihingen data set was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) [2]: http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html.

References

[1] A. Barsi and C. Heipke. Artificial neural networks fort the detection of road junctions in aerial images. In *Int. Arch. Photogrammetry, Remote Sensing & Geoinformation Sc.*, volume XXXIV-3/W8, pages 18–21, 2003.

[2] M. Cramer. The DGPF test on digital aerial camera evaluation - overview and test design. *Photogrammetrie-Fernerkundung-Geoinformation*, 2(2010):73–82, 2010.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, pages 886–893, 2005.

[4] T. Hoberg, F. Rottensteiner, and C. Heipke. Classification of multitemporal remote sensing data of different resolution using conditional random fields. In *IEEE ICCV Workshops*, pages 235–242, 2011.

[5] S. Kosov, F. Rottensteinter, C. Heipke, J. Leitloff, and S. Hinz. 3d classification of crossroads from multiple aerial images using markov random fields. In *Proc. 22nd ISPRS Congress*, 2012.

[6] S. Kumar and M. Hebert. Discriminative Random Fields. *International Journal of Computer Vision*, 68(2):179–201, 2006.

[7] S. Z. Li. *Markov random field modeling in image analysis*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2001.

[8] W.-L. Lu, K. P. Murphy, J. J. Little, A. Sheffer, and H. Fu. A hybrid conditional random field for estimating the underlying ground surface from airborne Lidar data. *IEEE-TGARS*, 47(8/2):2913–2922, 2009.

[9] H. Mayer, S. Hinz, U. Bacher, and E. Baltsavias. A test of automatic road extraction approaches. In *Int. Arch. Photogrammetry, Remote Sensing & Geoinformation Sc.*, volume XXXVI-3, pages 209–214, 2006.

[10] OpenCV. Camera calibration and 3D reconstruction. http://opencv.itseez.com/modules/calib3d/doc/calib3d.html, Apr. 2012.

[11] M. Ravanbakhsh, C. Heipke, and K. Pakzad. Road junction extraction from high resolution aerial imagery. *Photogrammetric Record*, 23(2):405–423, 2008.

[12] M. Rutzinger, F. Rottensteiner, and N. Pfeifer. A comparison of evaluation techniques for building extraction from airborne laser scanning. *IEEE-JSTARS*, 2(1):11–20, 2009.

[13] S. V. N. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. P. Murphy. Accelerated training of conditional random fields with stochastic gradient methods. In *Proc. 23rd ICML*, pages 969–976, 2006.

[14] J. D. Wegner, R. Hänsch, A. Thiele, and U. Sörgel. Building detection from one orthophoto and high-resolution InSAR data using conditional random fields. *IEEE-JSTARS*, 4(1):83–91, 2011.