THE APPLICATION OF A CAR CONFIDENCE FEATURE FOR THE CLASSIFICATION OF CROSS-ROADS USING CONDITIONAL RANDOM FIELDS

S. G. Kosov[a], F. Rottensteiner[a,*], C. Heipke[a], J. Leitloff[b], and S. Hinz[b]

[a] Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Germany
{kosov, rottensteiner, heipke}@ipi.uni-hannover.de
[b] Institute of Photogrammetry and Remote Sensing, Karlsruhe University of Technology, Germany
{Jens.Leitloff,stefan.hinz@kit.edu}@kit.edu

Commission III WG III/4

KEY WORDS: Conditional Random Fields, Contextual, Classification, Crossroads

ABSTRACT:

The precise classification and reconstruction of crossroads from multiple aerial images is a challenging problem in remote sensing. We apply the Conditional Random Fields (CRF) approach to this problem, a probabilistic model that can be used to consider context in classification. A simple appearance-based model is combined with a probabilistic model of the co-occurrence of class label at neighbouring image sites to distinguish classes that are relevant for scenes containing crossroads. The parameters of these models are learnt from training data. We use multiple overlap aerial images to derive a digital surface model (DSM) and a true orthophoto without moving cars. From the DSM and the orthophoto we derive feature vectors that are used in the classification. Within our framework we make use of a car detector based on support vector machines (SVM), which delivers car probability values. These values are used as additional feature to support the classification when the road surface is occluded by static cars. Our approach is evaluated on a dataset of airborne photos of an urban area by a comparison of the results to reference data. The evaluation is performed for images of different resolution. The method is shown to produce promising results when using the car probability values and higher image resolution.

1 INTRODUCTION

The automatic detection and reconstruction of roads has been an important topic of research in Photogrammetry and Remote Sensing for several decades. Considerable progress has been made, but the problem has not been finally solved. The EuroSDR test on road extraction has shown that road extraction methods are mature and reliable under favourable conditions, in particular in rural areas, but they are far from being practically relevant in more challenging environments as they exist in urban or suburban areas (Mayer et al., 2006). One of the main reasons for failure of road extraction algorithms in that test was the existence of crossroads, due to the fact that model assumptions about roads (e.g., the existence of parallel edges delineating a road) are hurt there. For this reason, specific models for the extraction of crossroads from images have been developed. (Barsi and Heipke, 2003) used neuronal networks for a supervised per-pixel classification of greyscale orthophotos in order to detect areas corresponding to crossroads, combining radiometric and geometric features. However, only examples for rural areas were shown. (Ravanbakhsh et al., 2008b, Ravanbakhsh et al., 2008a) used a model based on snakes to delineate outlines of road surfaces at crossroads, including the delineation of traffic islands. 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. Occlusions were also a major problem in (Grote et al., 2012), which also gives an overview over other current road detection techniques. The problem of occlusion by cars could be overcome if the position of cars were known in the images.

Conditional Random Fields (CRF) can be used for a raster-based classification of images (Kumar and Hebert, 2006). CRF offer probabilistic models for including context in the classification process by considering the statistical dependencies between the class labels at neighbouring image sites. Nevertheless, their application is restricted because of oversmoothing (Schindler, 2012), which is most likely to occur with small classes such as cars. In our previous work (Kosov et al., 2012) we tried to overcome this problem by integrating a *car confidence feature* into a CRF-based classification of image data together with a digital surface model (DSM). This feature was based on a probabilistic car detector, but the use of this feature did not contribute very much to improve the classification of cars because there were too many false positive car detections. It is one of the goals of this paper to overcome these problems by applying a more advanced car detector. Most recent approaches for car detection from aerial imagery use implicit models. In (Grabner et al., 2008) rotational invariant *Histogram of Oriented Gradients* (HOG), local binary pattern and Haar-like features are utilized. They apply an online boosting procedure for efficient training data collection. Another interesting approach is show in (Kembhavi et al., 2011), where new types of image features for vehicle detection are introduced. The feature includes color probability maps and pairs of pixels. The latter are used to extract symmetric properties of image objects. In this paper we propose a method to predict probabilties for vehicles based on rotation invariant features and Support Vector Machines. Thus, the number of false positives can be reduced. The second problem to be tackled in this paper is occlusion. We will address this problem by building a *twin CRF*, introducing two layers of class labels for each pixel. Partially occluded objects were also detected in (Leibe et al., 2008). The objects in the scene are represented as an assembly of parts. The method is robust to the cases where some parts are occluded and, thus, can predict labels for occluded parts from neighbouring unoccluded sites. However, it can only handle small occlusions, and it does not consider the relations between the occluded and the occlusion objects. Methods including multiple layers of class labels in a CRF mostly use part-based models, where the additional layer does not explicitly refer to occlusions, but encodes another label structure. In (Kumar and Hebert, 2005) and (Schnitzspan et al., 2009), multiple layers represent a hierachical object structure,

[*] Corresponding author

i.e. each object on higher level interacts with its smaller parts on lower level. In (Winn and Shotton, 2006), the part-based model is motivated by the methods potential to incorporate information about the relative alignment of object parts and to model longe-range interactions. However, occluded objects are not explicitly reconstructed. The spatial structure of such part-based models is not rotation-invariant and, thus, requires the availability of a reference direction (the vertical in images with a horizontal viewing direction), not available in aerial imagery. In (Wojek and Schiele, 2008), a CRF having several layers is used, but the additional layer is related to a label for object identity, used to track an object detected by a specific object detector over several images.

In (Kosov et al., 2013) we did already propose a two-layer CRF to deal with occlusions, but the classifier used for the association potentials was based Gaussian mixture models and no car confidence feature was applied. The method presented in this paper applies a better base classifier for the association potentials, namely Random Forests (RF), and again includes the car confidence features. The main advantage of separating two class labels is a better potential for correctly classifying partly occluded areas while maintaining the occluding objects such as cars or trees. Our method is evaluated using 90 crossroads of the Vaihingen data set of the German Society of Photogrammetry, Remote Sensing and Geoinformation (DGPF). We use image and DSM data having a ground sampling distance (GSD) of 8 cm. The focus of the evaluation is on the impact of the car confidence feature, the context model, and the image resolution on the results.

2 CONDITIONAL RANDOM FIELDS (CRF)

We assume an image $\mathbf{y}$ to consist of M image sites (pixels or segments) $i \in \mathbb{S}$ with observed data $\mathbf{y}_i$, i.e., $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_M)^T$, where $\mathbb{S}$ is the set of all sites. With each site i we associate a class label x_i from a given set of classes $\mathbb{C}$. Collecting the labels x_i in a vector $\mathbf{x} = (x_1, x_2, \dots, x_M)^T$, we can formulate the classification problem as finding the label configuration $\hat{\mathbf{x}}$ that maximises the posterior probability of the labels given the observations, $p(\mathbf{x}|\mathbf{y})$. A CRF is a model of $p(\mathbf{x}|\mathbf{y})$ with an associated graph whose nodes are linked to the image sites and whose edges model interactions between neighbouring sites. Restricting ourselves to a pairwise interactions, $p(\mathbf{x}|\mathbf{y})$ can be modelled by (Kumar and Hebert, 2006):

$$p(\mathbf{x} \mid \mathbf{y}) = \frac{1}{Z} \prod_{i \in \mathbb{S}} \left[\varphi_i(x_i, \mathbf{y}) \prod_{j \in \mathcal{N}_i} \psi_{ij}(x_i, x_j, \mathbf{y}) \right]. \qquad (1)$$

In Eq. 1, $\varphi_i(x_i, \mathbf{y})$ are the *association potentials* linking the observations to the class label at site i, $\psi_{ij}(x_i, x_j, \mathbf{y})$ are the *interaction potentials* modelling the dependencies between the class labels at two neighbouring sites i and j and the data $\mathbf{y}$, $\mathcal{N}_i$ is the set of neighbours of site i (thus, j is a neighbour of i), and Z is a normalizing constant. Applications of the CRF model differ in the way they define the graph structure, in the observed features, and in the models used for the potentials. Our adaptations of the framework will be explained in Section 3.

3 METHOD

The goal of our method is the pixel-based classification of urban scenes containing crossroads. The primary input consists of multiple aerial images and their orientation data. We require at least fourfold overlap of each crossroads from two different image strips in order to avoid occlusions as far as possible. In a preprocessing stage, these multiple images are used to derive a DSM by dense matching. The DSM is used to generate a true orthophoto from all input images, taking advantage of the multiple views to eliminate moving cars. More details about the preprocessing stage can be found in (Kosov et al., 2012). The DSM and the combined orthophoto are the input for extracting the features, which provide the input to the CRF-based classifier.

3.1 Twin CRF

In this paper we split objects corresponding to the *base level*, i.e. the most distant objects that cannot occlude other objects but could be occluded, and objects corresponding to the *occlusion level*, i.e. all other objects. This implies that, two class labels $x_i^b \in \mathbb{C}^b$ and $x_i^o \in \mathbb{C}^o$ are determined for each image site i. They correspond to the base and occlusion levels, respectively; $\mathbb{C}^b$ and $\mathbb{C}^o$ are the corresponding sets of class labels with $\mathbb{C}^b \bigcap \mathbb{C}^o = \emptyset$. In our application, $\mathbb{C}^b$ consists of classes such as *road* or *building*, whereas $\mathbb{C}^o$ includes classes such as *car* and *tree*. $\mathbb{C}^o$ includes a special class *void* $\in \mathbb{C}^o$ to model situations where the base level is not occluded. We model the posterior probabilities $p(\mathbf{x}^b \mid \mathbf{y})$, $p(\mathbf{x}^o \mid \mathbf{y})$ directly, expanding the model in Eq. 1:

$$p(\mathbf{x}^b, \mathbf{x}^o|\mathbf{y}) = \frac{1}{Z} \prod_{l \in \{o,b\}} \left\{ \prod_{i \in \mathbb{S}} \left[\varphi_i^l(x_i^l, \mathbf{y}) \prod_{j \in \mathcal{N}_i} \psi_{ij}^l(x_i^l, x_j^l, \mathbf{y}) \right] \right\} \qquad (2)$$

In Eq. 2, the association potentials $\varphi_i^l, l \in \{o, b\}$ link the data $\mathbf{y}$ with the class labels x_i^l of image site i at level l. The interaction potentials ψ_{ij}^l, $l \in \{o, b\}$, model the dependencies between the data $\mathbf{y}$ and the labels at two neighbouring sites i and j at each level. This model implies that the two levels do not interact. Training the parameters of the potentials in Eq. 2 requires fully labelled training images. The classification of new images is carried out by maximizing the probability in Eq. 2.

3.1.1 Association Potential:

Omitting the superscript indicating the level of the model, the association potentials $\varphi_i(x_i, \mathbf{y})$ are related to the probability of a label x_i taking a value c given the data $\mathbf{y}$ by $\varphi_i(x_i, \mathbf{y}) = p(x_i = c \mid \mathbf{f}_i(\mathbf{y}))$ (Kumar and Hebert, 2006), where the image data are represented by site-wise feature vectors $\mathbf{f}_i(\mathbf{y})$ that may depend on all the observations $\mathbf{y}$. Note that the definition of these feature vectors may vary with the dataset. We use a Random Forest (*RF*) (Breiman, 2001) in the implementation of (OpenCV, 2012) for the association potentials both of the base and for the occlusion levels, i.e. $\varphi_i^b(x_i^b, \mathbf{y})$ and $\varphi_i^o(x_i^o, \mathbf{y})$. A RF consists of N_T decision trees that are generated in the training phase. In the classification, each tree casts a vote for the most likely class. If the number of votes cast for a class c is N_c, the probability underlying our definition of the association potentials is $p(x_i = c \mid \mathbf{f}_i(\mathbf{y})) = N_c/N_T$.

3.1.2 Interaction Potential:

This potential describes how likely a pair of neighbouring sites i and j is to take the labels $(x_i, x_j) = (c, c')$ given the data: $\psi_{ij}(x_i, x_j, \mathbf{y}) = p(x_i = c, x_j = c'|\mathbf{y})$ (Kumar and Hebert, 2006). We generate a 2D histogram $h'_\psi(x_i,\ x_j)$ of the co-occurrence of labels at neighbouring sites from the training data; $h'_\psi(x_i = c, x_j = c')$ is the number of occurrences of the classes $(c,\ c')$ at neighbouring sites i and j. We scale the rows of $h'_\psi(x_i,\ x_j)$ so that the largest value in a row will be one to avoid a bias for classes covering a large area in the training data, which results in a matrix $h_\psi(x_i,\ x_j)$. We obtain $\psi_{ij}(x_i, x_j, \mathbf{y}) \equiv \psi_{ij}(x_i, x_j, d_{ij})$ by applying a penalization depending on the Euclidean distance $d_{ij} = \|\mathbf{f}_i(\mathbf{y}) - \mathbf{f}_j(\mathbf{y})\|$ of the feature vectors $\mathbf{f}_i$ and $\mathbf{f}_j$ to the diagonal of $h_\psi(x_i,\ x_j)$:

$$\psi_{ij}(x_i, x_j, d_{ij}) = \begin{cases} \lambda_1 \cdot e^{-\lambda_2 \cdot d_{ij}^2} \cdot h_\psi(x_i,\ x_j) & \text{if } x_i = x_j \\ h_\psi(x_i,\ x_j) & \text{otherwise} \end{cases} \qquad (3)$$

In Eq. 3, λ_1 and λ_2 determine the relative weight of the interaction potential compared to the association potential. As the largest entries of $h_\psi(x_i,\ x_j)$ are usually found in the diagonals, a model without the data-dependent term 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 $\mathbf{f}_i$ and $\mathbf{f}_j$ 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. This model differs from the contrast-sensitive Potts model (Boykov and Jolly, 2001) by the use of the normalised histograms $h_\psi(x_i,\ x_j)$ in Eq. 3. It is also different from methods such as those described in (Rabinovich et al., 2007), who use the co-occurrence of objects in a scene to define a *global* prior to make the detection of small objects in a scene more likely if related larger objects are found. We use the co-occurrence of *neighbouring* objects to favour *local* label transitions that occur more frequently in the training data. Again, the training of the models for the base and the occlusion levels, $\psi_{ij}^b(x_i^b, x_j^b, \mathbf{y})$ and $\psi_{ij}^o(x_i^o, x_j^o, \mathbf{y})$, respectively, are carried out independently from each other using fully labelled training data.

3.2 Car Detection

The presence of vehicles in optical images is a strong indicator for roads. Thus a seperate classifcation of cars seem to be very useful for reconstruction of crossroads. A very similar idea was already shown in (Hinz, 2004). There, hierachical wire-frame models were used for the verification of already detected roads. In general, vehicle detection is performed either using implicit or explicit models. Extensive overviews of previous work can be found in (Stilla et al., 2004) and (Hinz et al., 2006).

The directions of the roads are unknown in advance. Thus, we also use HOG features. These image features can be calculated very efficiently by integral histograms (Porikli, 2005) for the sliding classification windows. The window size is 80 × 80 pixels. We calculate histograms with 9 bins for 100 non-overlapping blocks of 8 × 8 pixels each. Training and classification is performed using nonlinear Support Vector Machines (SVM) with soft margins and radial basis functions as kernel. The kernel parameter and error weight of slack variables is determined by cross-validation on the training data. The membership of each pixel i to class car given its feature vector $\mathbf{y_i}$ is calculated by

$$f\left(\mathbf{y}_i\right) = sign\left(\mathbf{w}^T\varphi\left(\mathbf{y_i}\right)\right) \tag{4}$$

where $\mathbf{w}$ is the normal vector and b the vertical distance to feature space origin of the seperating hyperplane in the tranformed feature space. Transformation of feature vectors is given by the tranform $\varphi\left(\mathbf{y_i}\right)$. This function only gives a binary decision, which is not suitable as an input for the CRF. Thus, posteriori probabilities $P\left(x_i|\mathbf{y}_i\right)$ for each pixel i are estimated. For that purpose, the posterior is approximeted by a sigmoid function as proposed by (Platt, 2000):

$$P\left(x_i = car|\mathbf{y}\right) \approx P_{A,B}\left[f\left(\mathbf{y}_i\right)\right] = \frac{1}{1+\exp\left[A\left(\mathbf{y}_i\right)+B\right]} \tag{5}$$

The parameters A and B are estimated by the algorithm given in (Lin et al., 2007), which is more robust than the original algorithm of (Platt, 2000).

3.3 Definition of the Features

As stated in Section 3.1.1, we derive a feature vector $\mathbf{f}_i(\mathbf{y})$ for each image site i that consists of seven features derived from the orthophoto (image features) collected in a vector $\mathbf{f}_{img}$, a feature derived from the DSM (f_{DSM}) and, optionally, the car confidence feature (f_{car}), defined as the posterior in Eq. 5. We also make use of multi-scale features, collected in a vector $\mathbf{f}_{MS}$. The site-wise feature vectors are $\mathbf{f}_i(\mathbf{y})^T = (\mathbf{f}_{img}^T, f_{DSM}, \mathbf{f}_{MS}^T)$ or $\mathbf{f}_i(\mathbf{y})^T = (\mathbf{f}_{img}^T, f_{DSM}, \mathbf{f}_{MS}^T, f_{car})$, depending on whether the car confidence feature is used or not. For numerical reasons all features are scaled linearly into the range between 0 and 255 and then quantized by 8 bit.

We do not use the colour vectors of the images directly to define the site-wise image feature vectors $\mathbf{f}_{img}$. 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. We also make use of the variance of intensity (var_{int}) and the variance of saturation (var_{sat}), determined from a local neighbourhood of each pixel (7 × 7 pixels for var_{int}, 13 × 13 pixels for var_{sat}). The sixth image feature ($dist$) represents the relation between an image site and its nearest edge pixel; this feature should model 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 feature used in classification is the distance of an image site to its nearest edge pixel, taken from the distance map. Thus, the image feature vector for each pixel is $\mathbf{f}_{img} = (NDVI, sat, int, var_{sat}, var_{int}, dist)^T$.

A coarse Digital Terrain Model (DTM) is generated 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 DSM feature is the difference between the DSM and the DTM, i.e., $f_{DSM} = DSM - DTM$. This feature describes the relative elevation of objects above ground such as buildings, trees, or bridges. The multi-scale features $\mathbf{f}_{MS}$ comprise the $NDVI$, f_{DSM} and sat features, calculated at two coarser different scales as average values in squares of 21 × 21 and 49 × 49 pixels, respectively.

3.4 Training and Inference

Training of a CRF is computationally intractable if to be carried out in a probabilistic framework (Kumar and Hebert, 2006). Thus, approximate solutions have to be used for training. In our application, we determine the parameters of the association and interaction potentials separately based on fully labelled training images. The RF classifier used in the association potentials are trained using the site-wise feature vectors of the training images. The interaction potentials are derived from scaled versions of the 2D histograms of the co-occurrence of class labels at neighbouring image sites in the way described in Sec. 3.1.2, taking into account all image sites in the training data. The parameters λ_1 and λ_2 in the Eq. 3 are set manually to values 2.0 and 0.01, respectively. Exact inference is also computationally intractable for CRFs. We use Loopy Belief Propagation (LBP), a standard technique for probability propagation in graphs with cycles that has shown to give good results in the comparison reported in (Vishwanathan et al., 2006).

4 EXPERIMENTS

4.1 Experimental Setup

To evaluate our model we used a part of the aerial images of the Vaihingen data set (Cramer, 2010). We selected 90 crossroads

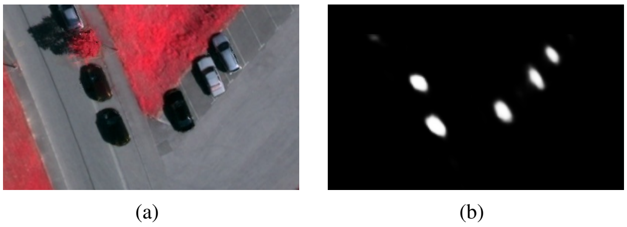

Figure 1: Posterior probability from SVM classification. (a) original image, (b) classification result.

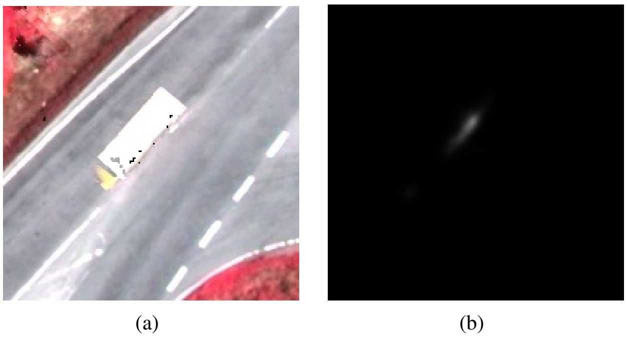

Figure 2: Results of blurred vehicle caused by median filtering. (a) original image, (b) classification result.

for our experiments. For each crossroads, a true orthophoto and a DSM were available, each covering an area of 80 $\times$ 80 m^2 with a GSD of 8 cm. The DSM and the orthophoto were generated from multiple aerial CIR images in the way described in (Kosov et al., 2012). They provide the original input to our CRF-based classifier. We defined each image site to correspond to image pixels, thus in the full resolution each graphical model consisted of 1000×1000 nodes. The neighbourhood $\mathcal{N}_i$ of an image site i in Eq. 1 is chosen to consist of the direct neighbours of i in the data grid.

We defined six classes that are characteristic for scenes containing crossroads, namely *asphalt* (*asp.*), *building* (*bld.*), *tree*, *grass* (*gr.*), *agricultural* (*agr.*) and *car*, so that $\mathbb{C}^b = \{asp., bld., gr., agr.\}$ and $\mathbb{C}^o = \{tree, car, void\}$. The two-level reference was generated by manually labeling the orthophotos using these 6 classes, using assumptions about the continuity of objects such as road edges in occluded areas to define the reference of the base level.

For the evaluation we used cross validation. In each test run, 45 images were used for training, and the remaining 45 for testing. This was repeated two times so that each image was used first for training and second for testing. The results were compared with the reference; we report the completeness and the correctness of the results per class as well as the overall accuracy (Rutzinger et al., 2009).

4.2 Car Detection

Classification gives the probability for vehicles for each pixel. In case of cleary seperated cars, the approach delivers results as illustrated in Fig. 1. During image generation moving vehicles should be eliminated. Still, several ”blurred” vehicles are still visible. These vehicles also give response during classification, even so, the probabilties are smaller than 1 due to low contrast. An example is given in Fig. 2. Furthermore, objects of similar dimension recieve high probalities as it can be seen in Fig. 3.

In Fig. 4 the completeness versus correctness for different thresholds on the estimated vehicle probalities are shown. For this evaluation, the centre point of connected pixel having a larger value than the threshold is compared to the regions of the reference (e.g. first row of Fig. 5). Thus, connected regions which cover multiple vehicles (e.g. last row of Fig. 5) are only counted once and lead to a signifcant reduction of completeness. Therefore, the given value for completeness in Fig. 4 are quite pessimistic. Nevertheless, the overall correctness still needs further improvement, which could be achieved by additional features and an additional classification of the connected regions. This is planed for future work.

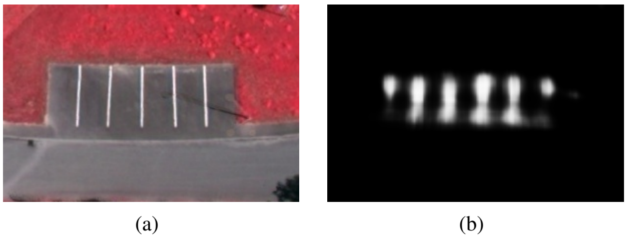

Figure 3: Results of vehicle-like image parts. (a) original image, (b) classification result.

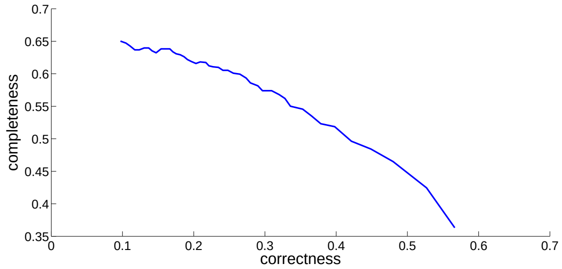

Figure 4: ROC for variying thresholds of probability.

4.3 Results and Discussion

We carried out eight experiments. In the first four experiments (RF^5_{car}, RF^5, CRF^5_{car}, CRF^5) we used a version of the Vaihingen dataset with a reduced GSD of 40 cm (corresponding to 5×5 pixels of the original images), so that the CRF only consisted of 200×200 nodes. In the second set of experiments (RF^1_{car}, RF^1, CRF^1_{car}, CRF^1) we used the images at their full resolution of 8 cm. In the experiments RF^1_{car} and RF^5_{car}, we only used the Random Forest classifier for a local classification of each node, neglecting the interaction potentials. In the experiments CRF^1_{car} and CRF^5_{car}, the twin CRF model in Eq. 2 was used, including the interactions. The experiments RF^5_{car}, CRF^5_{car}, RF^1_{car} and CRF^1_{car} were performed using the car confidence feature, while for the experiments RF^5, CRF^5, RF^1 and CRF^1 the car confidence feature was not applied. The completeness and the correctness of the results achieved in these experiments are shown in Tab. 1 and 2. For the occlusion layer we also report the quality (Rutzinger et al., 2009), which is a measure for the trade-off between completeness and correcntess.

In Tab. 1 the overall accuracy for the base layer does not differ much between the experiments. Considering the interactions increases the overall accuracy by slightly more than 1% in the full resolution and slightly less in the lower resolution experiments. Partly this may be explicable by a good performance of the RF classifier and the inclusion of multiscale features, but a stronger setting of the weights for the interaction potentials might have lead to a larger differces. Using the car feature leads to an even lower increase in the overall accuracy in all experiments, which is, however, to be expected because only a very small area is covered by cars, and the car confidence is low in most of the areas where cars occur.

		asp.	*bld.*	*gr.*	*agr.*	**OA**
RF^5_{car}	$Cm.$	80.1	82.6	82.7	56.3	**78.5**
	$Cr.$	84.6	78.4	79.7	62.2	
CRF^5_{car}	$Cm.$	82.2	76.5	89.0	42.2	**79.2**
	$Cr.$	83.7	87.9	75.3	78.8	
RF^5	$Cm.$	79.8	83.8	82.9	52.9	**78.3**
	$Cr.$	85.5	77.8	79.0	61.1	
CRF^5	$Cm.$	81.7	77.0	88.7	41.8	**79.0**
	$Cr.$	83.7	87.2	75.3	77.1	
RF^1_{car}	$Cm.$	79.8	83.7	82.9	54.9	**78.5**
	$Cr.$	85.5	77.6	79.4	62.0	
CRF^1_{car}	$Cm.$	80.7	84.5	84.7	54.7	**79.6**
	$Cr.$	86.4	78.9	79.6	68.8	
RF^1	$Cm.$	79.8	83.8	82.9	52.9	**78.3**
	$Cr.$	85.5	77.8	79.0	61.1	
CRF^1	$Cm.$	80.8	84.6	84.9	52.5	**79.5**
	$Cr.$	86.5	79.0	79.1	68.5	

Table 1: Completeness ($Cm.$), Correctness ($Cr.$), overall accuracy (OA) [%] for the base layer.

		void	*tree*	*car*	**OA**
RF^5_{car}	$Cm.$	77.8	85.5	75.8	**79.1**
	$Cr.$	95.4	56.2	10.9	
	$Q.$	75.0	51.3	10.5	
CRF^5_{car}	$Cm.$	94.3	50.4	9.0	**84.9**
	$Cr.$	87.8	67.7	75.5	
	$Q.$	83.4	40.6	8.7	
RF^5	$Cm.$	76.5	85.5	72.7	**78.2**
	$Cr.$	95.3	55.6	9.4	
	$Q.$	73.7	50.8	9.1	
CRF^5	$Cm.$	93.9	51.8	3.5	**84.8**
	$Cr.$	88.0	66.7	55.9	
	$Q.$	83.2	41.2	3.4	
RF^1_{car}	$Cm.$	77.5	85.6	77.6	**79.0**
	$Cr.$	95.5	56.3	10.8	
	$Q.$	74.8	51.4	10.5	
CRF^1_{car}	$Cm.$	84.0	87.2	34.2	**84.1**
	$Cr.$	95.5	57.1	41.6	
	$Q.$	80.8	52.7	23.1	
RF^1	$Cm.$	76.0	86.3	75.1	**77.9**
	$Cr.$	95.5	56.2	8.9	
	$Q.$	73.4	51.6	8.6	
CRF^1	$Cm.$	83.5	87.8	32.9	**83.8**
	$Cr.$	95.6	56.8	31.7	
	$Q.$	80.4	52.6	19.3	

Table 2: Completeness ($Cm.$), Correctness ($Cr.$), Quality ($Q.$), overall accuracy (OA) [%] for the occlusion layer.

Tab. 2 shows that the occlusion layer, containing the class *car*, shows a larger variation of the quality metrics between the different experiments. The most obvious improvement is achieved by considering local context: the overall accuracy achieved in the experiments based on CRF is 5%-6% better than the one achieved in the RF experiments. This is mainly due to an improvement of the completeness of class $void$, an indicator that in the RF scenario there are more false positive car and, in the lower resolution, $tree$ objects, which is confirmed by the correctness numbers of these objects in the RF setting. Whereas the overall accuracy is similar between the experiments at full resolution and those at a reduced resolution, it becomes evident that the oversmoothing in the latter leads to a particularly poor performance for the smallest objects in our classification schemes, i.e. cars. For these objects, a classification at full resolution seems to be required. Looking at the results achieved for the images at full resolution, in the CRF setting, a better trade-off between completeness and correctness is achieved for the class *car*, indicated by the higher quality scores ($Q.$ in Tab. 2) compared to the RF experiments. Tab. 2 also shows that indeed the car feature helps in the classification of cars. Experiment CRF^1_{car} achieves the highest quality score for car, though there is still considerable room for improvement.

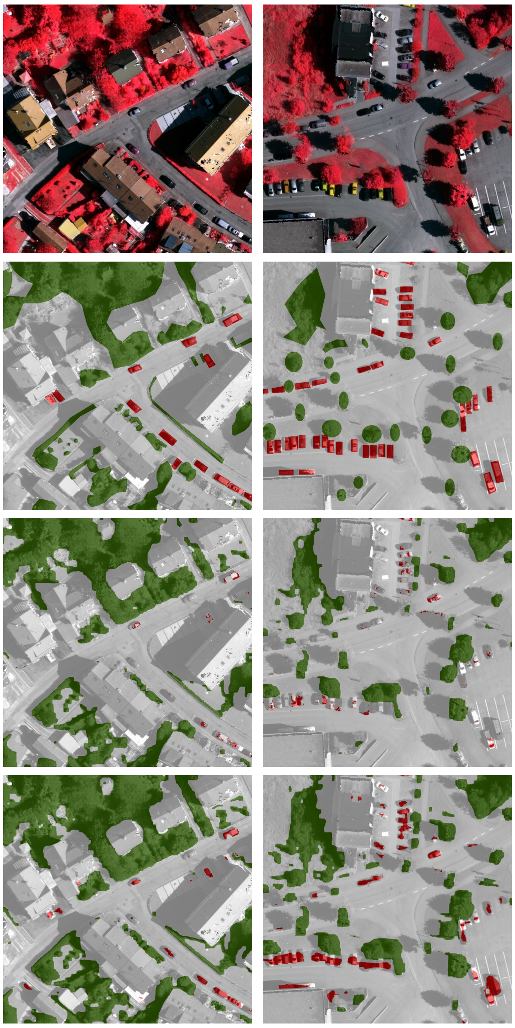
(a) (b)

Figure 5: Classification of the occlusion layer. First row: Original images (GSD 8 cm), second row: reference, third row: CRF, fourth row: CRF_{car}. (a) Scene #23; (b) Scene #36. White: *void*; dark green: *tree*; red: *car*.

Fig. 5 illustrates two scenes with a high number of cars. Its third row presents the results of CRF^1, while the fourth row shows results of the CRF^1_{car} experiment. In these scenes, using the car confidence feature improves the classification rate for cars considerably. In comparison to the reference (second row of Fig. 5), cars are oversmoothed and hardly recognizable in the results of CRF^1. CRF^1_{car} delivers the results with the car regions in the correct positions and nearly without false positives.

5 CONCLUSION

In this paper, a method for the classification of crossroads using CRF was proposed. It considered occlusions explicitly by determining two class labels per pixel. A car confidence feature to avoid problems with occlusions of the road surface by cars.

Distinguishing 7 classes relevant in the context of crossroads, an overall accuracy of about 79 - 85% could be achieved. The car confidence feature, which is based on the output of our car detector, is shown to increase the accuracy of classification especially for the class car. In the future we want to improve our method by integrating more expressive features, e.g. features related to car trajectories. Furthermore, the interactions between the two levels need to be modelled in a way similar to (Kosov et al., 2013).

ACKNOWLEDGEMENTS

This research was funded by the German Science Foundation (DFG) under grants HE 1822/25-1 and HI 1289/1-1. The Vaihingen data set was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) (Cramer, 2010): http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html.

REFERENCES

Barsi, A. and Heipke, C., 2003. Artificial neural networks for the detection of road junctions in aerial images. In: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXIV-3/W8, pp. 18–21.

Boykov, Y. and Jolly, M., 2001. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In: Proc. International Conference on Computer Vision (ICCV), Vol. I, pp. 105–112.

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

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

Grabner, H., Nguyen, T., Gruber, B. and Bischof, H., 2008. Online boosting-based car detection from aerial images. ISPRS Journal of Photogrammetry and Remote Sensing 63(3), pp. 382–396.

Grote, A., Heipke, C. and Rottensteiner, F., 2012. Road network extraction in suburban areas. Photogrammetric Record 27, pp. 8–28.

Hinz, S., 2004. Detection of vehicles and vehicle queues in high resolution aerial images. Photogrammetrie - Fernerkundung - Geoinformation 3, pp. 201–213.

Hinz, S., Bamler, R. and Stilla, U., 2006. Theme issue: Airborne and spaceborne trafc monitoring. ISPRS J. Photogramm. & Rem. Sens. 61(3/4).

Kembhavi, A., Harwood, D. and Davis, L. S., 2011. Vehicle detection using partial least squares. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(6), pp. 1250–1265.

Kosov, S., Rottensteiner, F. and Heipke, C., 2013. Sequential gaussian mixture models for two-level conditional random fields. In: Proceedings of the 35th German Conference on Pattern Recognition (GCPR), LNCS, Vol. 8142, Springer, Heidelberg, pp. 153–163.

Kosov, S., Rottensteiner, F., Heipke, C., Leitloff, J. and Hinz, S., 2012. 3d classification of crossroads from multiple aerial images using markov random fields. In: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXIX-B3, pp. 479–484.

Kumar, S. and Hebert, M., 2005. A hierarchical field framework for unified context-based classification. In: Proc. International Conference on Computer Vision (ICCV), pp. 1284–1291.

Kumar, S. and Hebert, M., 2006. Discriminative Random Fields. International Journal of Computer Vision 68(2), pp. 179–201.

Leibe, B., Leonardis, A. and Schiele, B., 2008. Robust object detection with interleaved categorization and segmentation. International Journal of Computer Vision 77, pp. 259–289.

Lin, H.-T., Lin, C.-J. and Weng, R. C., 2007. A note on platts probabilistic outputs for support vector machines. Machine learning 68(3), pp. 267–276.

Mayer, H., Hinz, S., Bacher, U. and Baltsavias, E., 2006. A test of automatic road extraction approaches. In: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXVI-3, pp. 209–214.

OpenCV, 2012. Machine Learning. http://docs.opencv.org/modules/ml/doc/ml.html.

Platt, J. C., 2000. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: A. Smola, P. Bartlett, B. Schlkopf and D. Schuurmans (eds), Advances in Large Margin Classiers, MIT Press.

Porikli, F., 2005. Integral histogram: A fast way to extract histograms in cartesian spaces. In: Proc. Conf. Computer Vision and Pattern Recognition (CVPR), Vol. 1, IEEE, pp. 829–836.

Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E. and Belongie, S., 2007. Objects in context. In: Proceedings of the International Conference on Computer Vision (ICCV).

Ravanbakhsh, M., Heipke, C. and Pakzad, K., 2008a. Automatic extraction of traffic islands from aerial images. Photogrammetrie Fernerkundung Geoinformation 5(2008), pp. 375–384.

Ravanbakhsh, M., Heipke, C. and Pakzad, K., 2008b. Road junction extraction from high resolution aerial imagery. Photogrammetric Record 23, pp. 405–423.

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

Schindler, K., 2012. An overview and comparison of smooth labeling methods for land-cover classification. IEEE-TGARS 50, pp. 4534–4545.

Schnitzspan, P., Fritz, M., Roth, S. and Schiele, B., 2009. Discriminative structure learning of hierarchical representations for object detection. In: Proc. Conf. Computer Vision and Pattern Recognition (CVPR), pp. 2238–2245.

Stilla, U., Michaelsen, E., Sörgel, U., Hinz, S. and Ender, J., 2004. Airborne monitoring of vehicle activity in urban areas. In: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXV- B3, pp. 973–979.

Vishwanathan, S. V. N., Schraudolph, N. N., Schmidt, M. W. and Murphy, K. P., 2006. Accelerated training of conditional random fields with stochastic gradient methods. In: Proc. 23^{rd} ICML, pp. 969–976.

Winn, J. and Shotton, J., 2006. The layout consistent random field for recognizing and segmenting partially occluded objects. In: Proc. Conf. Computer Vision and Pattern Recognition (CVPR).

Wojek, C. and Schiele, B., 2008. A dynamic conditional random field model for joint labeling of object and scene classes. In: ECCV (4), pp. 733–747.