International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XXXIX-B3, 2012
XXII ISPRS Congress, 25 August – 01 September 2012, Melbourne, Australia

3D CLASSIFICATION OF CROSSROADS FROM MULTIPLE AERIAL IMAGES USING MARKOV RANDOM FIELDS

S. Kosov [a, *], F. Rottensteiner [a], C. Heipke [a], J. Leitloff [b], 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 Institute of Technology, Germany - (jens.leitloff, stefan.hinz)@kit.edu

Commission III, ICWG III/VII

KEY WORDS: Markov 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 Markov Random Fields (MRF) 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 up to 14 different 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. One of the features is a car confidence value that is supposed 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. Whereas the method has problems in distinguishing classes having a similar appearance, it is shown to produce promising results if a reduced set of classes is considered, yielding an overall classification accuracy of 74.8%.

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 noted by (Mayer et al., 2006) is 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. (2008a, 2008b) 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. The problem of occlusion by cars could be overcome if the position of cars were known in the images. Extensive overviews about methods for vehicle detection from optical aerial imagery can be found in (Stilla et. al., 2004) and (Hinz et. al., 2006).

In this paper we propose a new method for the classification of scenes containing crossroads as a first step of a 3D reconstruction. Markov Random Fields (MRF; Geman & Geman, 1984) are employed for a raster-based classification. MRF offer probabilistic models for including context in the classification process by considering the statistical dependencies between the class labels at neighbouring image sites; cf. (Li, 2009) for more details on MRF and their applications in image analysis. We use multiple-overlap aerial images in order to derive a Digital Surface Model (DSM) that is used in the classification process to make it more robust with respect to ambiguities of the appearance of objects in a 2D projection of the scenes. In addition, we include information about cars by integrating the output of a car detector into the process. Our method is evaluated using 55 crossroads of the Vaihingen data set of the German Society of Photogrammetry, Remote Sensing and Geoinformation (DGPF).

2. MARKOV RANDOM FIELDS

Markov random fields (MRF) provide probabilistic models of context for the image labelling problem (Geman & Geman, 1984; Li, 2009). Given image data $\mathbf{y}$ consisting of N image sites $i \in S$ with observed data y_i, i.e., $\mathbf{y} = (y_1, y_2, \ldots, y_N)^T$, where S is the set of all sites, we want to assign a discrete class label x_i from a given set of classes C to each site i. In this context, an image site can correspond to a single pixel or to an image segment. MRF are undirected graphical models that assume the data y_i at image site i to depend on the class label x_i at that site. In addition, the class label x_i is modelled to be statistically

* Corresponding author.

dependent on the class labels of its neighbouring image sites. As a consequence, the individual sites can no longer be labelled independently from each other. Collecting the class labels x_i in a vector $\mathbf{x} = (x_1, x_2, ..., x_N)^T$, we want to find the label configuration $\mathbf{x}^*$ that maximises the posterior probability of the labels given the data $p(\mathbf{x} \mid \mathbf{y})$, thus $\mathbf{x}^* = \arg\max_{\mathbf{x}} p(\mathbf{x} \mid \mathbf{y})$. The posterior probability $p(\mathbf{x} \mid \mathbf{y})$ can be modelled by a Gibbs distribution (Geman & Geman, 1984):

$$p(\mathbf{x}|\mathbf{y}) = \frac{1}{Z} \cdot exp\left(\sum_{i \in S} \varphi_i(x_i, y_i) + \sum_{i \in S} \sum_{j \in N_i} \psi_{ij}(x_i, x_j) \right) \quad (1)$$

In Eq. 1, Z is a normalization constant called the partition function, and N_i is the neighbourhood of data site i (thus, j is a neighbouring data site of i). The *association potential* φ_i links the class label x_i of image site i to the data y_i observed at that site, whereas the *pairwise interaction potential* ψ_{ij} models the dependencies between the labels x_i and x_j of neighbouring sites i and j. The model is very general in terms of the definition of the functional model for both φ_i and ψ_{ij}. Our definitions of the image sites and the neighbourhood N_i (thus, the structure of the graphical model) and the potential functions φ_i and ψ_{ij} used in our application are described in Section 3.

3. METHOD

The goal of our method is the classification of 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 this paper, the images are assumed to be colour infrared (CIR) images, though the methodology can be transferred to other spectral configurations by adapting the definition of the features to be used for classification. In a preprocessing stage, these multiple images are used to derive a DSM by dense matching. After that, the DSM is used to generate a true orthophoto from each input image. As each of these orthophotos will contain void areas due to occlusions, they are all combined to a joint true orthophoto with only few occluded areas left. In this process, we take advantage of the multiple views to also eliminate moving cars.

The DSM and the combined orthophoto are the input to the MRF-based classifier. In the classification process, we choose the image sites and, thus, the nodes of the graphical model, to correspond to small squares of n x n pixels of the joint true orthophoto. The neighbourhood 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 defined 14 classes that are characteristic for scenes containing crossroads both in an urban and in a rural setting, including *road*, *building*, *grass*, *tree*, *car*, but also *sidewalk*, *traffic island*, and *sealed*, the latter corresponding to off-road areas covered by asphalt, e.g. parking lots. Some of these classes have a very similar appearance in the data and are characterised by their relative spatial arrangement; however, it is possible to generate a new set of classes by combining some of the original ones, e.g. by merging all classes covered by asphalt (*road*, *sidewalk*, *traffic island*, *sealed*).

From the orthophoto and the DSM we extract the feature vectors. We use three groups of features, namely image-based features, DSM features, and a specific feature that is used to characterize cars; the use of the latter feature is optional. In a training phase we use images that were labelled manually to determine the parameters of the association and interaction potentials in Eq. 1. Training the parameters of the interaction potentials requires fully labelled images. Once the parameters have been determined, the classification of new test images can be carried out by maximising the posterior probability in Eq. 1 using the trained model.

The individual components of our method, in particular pre-processing, the definition of the potentials, the definition of the features and the methods used for training and inference are described in more detail in the subsequent sections.

3.1 Preprocessing

The first step of preprocessing is the generation of a DSM from the input images. We use the OpenCV implementation (OpenCV, 2012) of semiglobal matching (Hirschmüller, 2008) with the cost function of (Birchfield & Tomasi, 1998) to generate a disparity image for each possible pair of images. For each disparity image thus created, a DSM grid is generated in object space. Due to occlusions and matching errors, these raw DSMs will contain void areas, and there will also be height discrepancies, e.g. at roof overhangs. These raw DSMs are combined to a joint DSM by taking the median of the valid raw DSM heights at each position. Remaining void areas (e.g. caused by problems of the dense matcher in homogeneous image regions) are filled by an in-painting algorithm based on non-linear diffusion that is sensitive to height changes. In this process, we distinguish between void areas where the heights are to be interpolated from their surroundings (largely caused by matching errors) and areas where the heights are to be determined from the lowest surrounding areas (largely caused by occlusion) in a way similar to (Hirschmüller, 2008).

The DSM is the basis for the generation of a true orthophoto from each of the original input images. Ray tracing is used to determine visibility in this process. The resulting raw orthophotos will have void areas caused by occlusion. Finally, these raw orthophotos are merged to a combined orthophoto. For each pixel of the combined orthophoto, the median of all valid colour vectors (i.e. the colour vectors from all raw orthophotos where the respective pixel is not marked as being void) is chosen. Due to the fact that we require at least four-fold overlap, this will result in an elimination of moving cars on the streets, which improves the prospects of automatic classification of road surfaces (Fig. 1).

Figure 1: Detail of a test site. Left: DSM; centre: raw true orthophoto with void areas in black; right: combined true orthophoto.

3.2 Association potential

The association potential $\varphi_i(x_i, y_i)$ in Eq. 1 is related to the probability of observing the image data y_i at data site $i \in S$ given that label x_i takes a value $c \in C$ by

$\varphi_i(x_i,y_i) \propto log\ p(\mathbf{f}_i(\mathbf{y}) \mid x_i = c)$. Thus, $\varphi_i(x_i,y_i)$ is the log likelihood of the generative model with a smoothness prior on the labels expressed in Eq. 1. In this context, the image data are represented by site-wise feature vectors $\mathbf{f}_i(\mathbf{y})$ that may depend on the data observed at site *i* and its local neighbourhood. Both the definition of the features and the dimension of the feature vectors $\mathbf{f}_i(\mathbf{y})$ may vary from dataset to dataset, because the definition of appropriate and expressive features depends on the image resolution and also on the spectral information contained in the images.

We use a simple model for the likelihood $p(\mathbf{f}_i(\mathbf{y}) \mid x_i)$ and, consequently, for the association potential $\varphi_i(x_i,y_i)$. 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) $p(f_{ij} \mid x_i = c) = p_c(f_{ij} \mid x_i)$, where f_{ij} is the j^{th} component of $\mathbf{f}_i$ for the class *c*. Neglecting the statistical dependencies between the individual features f_{ij}, the likelihood $p(\mathbf{f}_i(\mathbf{y}) \mid x_i = c)$ becomes the product of the probability density functions of the individual features, so that the association potential becomes the sum of the logarithms of these functions:

$$\varphi(x_i = c, y_i) = \sum_{j=1}^{M} log\left[\, p_c\left(f_{ij} \mid x_i\right)\right] \qquad (2)$$

In Eq. 2, *M* is the dimension of the site-wise 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.

3.3 Interaction Potential

The pairwise interaction potential $\psi_{ij}(x_i,x_j)$ in Eq. 1 is a measure for the influence of neighbouring labels x_j on the class x_i of image site *i*. The function $\psi_{ij}(x_i,x_j)$ characterizes how likely the variable x_i will take the value *c*, given that the variable x_j from the neighbouring data site $j \in N_i$ takes the value *c'*, that is, $\psi_{ij}(x_i,x_j) \propto log\ p(x_i = c \mid x_j = c')$.

In order to determine this probability, a 2D histogram of the co-occurrence of labels at neighbouring interaction features is generated from the training data. The histogram corresponds to a (symmetric) matrix of dimension n_c x n_c, where n_c is the number of classes to be discerned, and the matrix entry at (*c, c'*) is the number of occurrences of the classes (*c, c'*) at neighbouring pixels *i* and *j*. After generating the histogram matrix, its rows are scaled so that the largest value in a row (usually the diagonal element) will be one. This is done in order to compensate for different numbers of pixels per class in the training data, *i.e.* to guarantee, that $p(x_i = c \mid x_j = c)$ will be the same for all classes $c \in C$. The interaction potentials $\psi_{ij}(x_i,x_j)$ are then defined as the logarithm of the scaled histogram matrix entries. It is a drawback of this type of scaling that $\psi_{ij}(x_i,x_j)$ will no longer be symmetric.

3.4 Definition of the Features

As stated in Section 3.1, we derive a feature vector $\mathbf{f}_i(\mathbf{y})$ for each image site *i* that consists of M_{img} features derived from the orthophoto (image features) collected in a vector $\mathbf{f}_{img}$, a feature derived from the DSM (f_{DSM}) and, optionally, a car confidence feature (f_{car}). Consequently, the total number *M* of features is either $M = M_{img} + 1$ or $M = M_{img} + 2$, corresponding to $\mathbf{f}_i(\mathbf{y})^T = (\mathbf{f}_{img}^T, f_{DSM})$ or $\mathbf{f}_i(\mathbf{y})^T = (\mathbf{f}_{img}^T, f_{DSM}, f_{car})$, depending on whether the car confidence feature is used is or not. In any case, for numerical reasons all features are scaled linearly into the range between 0 and 255 and then quantized by 8 bit.

3.4.1 Image features: We do not use the colour vectors of the images directly to define the site-wise image feature vectors $\mathbf{f}_{img}$. In total, we determine 7 image 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. 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 x 7 pixels for var_{int}, 13 x 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. The last image feature is the local gradient orientation, calculated in respect to the main gradient orientation (ori_{grad}). In order to compute the gradient orientation, we calculate two histograms of oriented gradients (HOG) (Dalal & Triggs, 2005), one considering a local neighbourhood (13 x 13 pixels in our experiments), and one from a larger neighbourhood (101 x 101 pixels). Each histogram consists of 9 orientation bins. The feature is the difference between the angles corresponding to the histogram bins having the maximum entries in the two histograms. Thus, the image feature vector for each pixel is $\mathbf{f}_{img} = (NDVI, sat, int, var_{sat}, var_{int}, dist, ori_{grad})^T$.

3.5.2 DSM feature: 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.

3.5.3 Car confidence feature: This is a feature that is supposed to be particularly useful for classifying cars. We use the output of the car detection algorithm described in (Leitloff et. al. 2010). However, we do not use the binary image of detected cars, but the confidence image derived by that method. The calculation of these confidence values uses an extended set of Haar-like features (Lienhart et. al., 2003) as input for pixel-wise classification. The number of possible features depends on the size of image samples used for training the classifier. Even for a reduced GSD of 20 cm and the resulting image patch size of 30 by 30 pixels more than 800,000 features exist. It is not possible to calculate all those features during classification. Thus, the number of features has to be reduced significantly during training. The idea of using Adaptive Boosting (Friedmann et. al. 2000) for feature reduction has been introduced by Tieu & Viola (2004). Boosting is an ensemble learning method, which combines a set of simple classifiers to generate a strong classifier. The output of each base (weak) learner is a confidence value. The final classification is obtained from the

sum of confidence values of all weak learners. Generally stumps or classification trees are used as base classifier. Each node of a regression tree applies a threshold to only one feature. The thresholds and features are chosen so that the training error becomes minimal. Thus, the most distinguishing features are found during training. In our case only 350 features have been selected, which makes the final classification suitable for large datasets. More details about training the Boosting classifier can be found in our previous work (Leitloff et. al. 2010). The feature f_{car} is defined as the combined confidence value of the classifier.

3.5 Training and Inference

Training of MRF is complex if it is to be carried out in a probabilistic framework, mainly due to the fact that it requires an estimate for the partition function Z in Eq. 1, which is computationally intractable. Thus, approximate solutions have to be used for training. In our application, we determine the parameters of the association and interaction potentials separately. That is, given the training data (fully labelled images), the probabilities $p_c(f_{ij} \mid x_i)$ are determined from histograms of the features f_{ij} (which are quantized by 8 bit for that purpose) of each class and smoothing, in the way described in Section 3.3. In a similar way, the interaction potentials are scaled versions of the 2D histograms of the co-occurrence of classes at neighbouring image sites in the way described in Section 3.4. Exact inference is also computationally intractable for MRF's. For inference, we use a message passing algorithm, namely 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 Test Data and Test Setup

Under the auspices of the DGPF a test data set over Vaihingen (Germany) was acquired in order to evaluate digital aerial camera systems (Cramer, 2010). It consists of several blocks of vertical images captured by various digital aerial camera systems at two resolutions. We used one of the DMC blocks to test our approach. The images are 16 bit pan-sharpened colour infrared images with a ground sampling distance (GSD) of 8 cm (flying height: 800 m, focal length: 120 mm). For our experiments, the radiometric resolution of the images had to be converted to 8 bit. The georeferencing accuracy is about 1 pixel. The nominal forward and side laps of the images are 65% and 60%, respectively. As a consequence, each crossroads in the block is visible in at least four images.

For our experiments, we selected 55 crossroads by digitizing their approximate centres. The set of crossroads contained examples from densely built-up urban and suburban as well as rural areas. For each crossroads, we generated a DSM and a true orthophoto, both with a GSD of 8 cm in the way described in Section 3.2; the size of the orthophotos used in our process was 1000 x 1000 pixels, thus corresponding to 80 x 80 m^2. In the training phase we use the original orthophotos (1000 x 1000 pixels); for inference, squares of 5 x 5 pixels were used as nodes of the graphical model; thus, each graphical model consisted of 200 x 200 nodes. For the car confidence feature we used a classifier trained on data of DLR's 3K-system (Kurz et. al. 2011). The sample images have a resolution of 20 cm. Thus, the Vaihingen dataset is resampled to this resolution for classification. Due to different radiometric properties, the Haar-like features are only calculated from intensity values. Both, resampling and exclusive use of intensity values limit the classification performance in this context.

The ground truth was generated by manually labelling the image areas using altogether 14 classes (cf. Figure 2). We use the ground truth for the algorithm's training phase and for the evaluating the classification accuracy. In order to have a sufficient amount of training data, we had to use cross validation in our evaluation procedure: in each experiment, all images except one were used for training, and the remaining image served as a test image; this procedure was repeated 55 times, each time using a different test image, so that in the end each image was used as a test image once. In all experiments, confusion matrices were determined from a comparison of the test images with the ground truth, as well as the completeness and the correctness of the results for each class and the overall classification accuracy (Rutzinger et al., 2009).

We carried out four different experiments. In the first two experiments, we tried to separate all 14 classes; the only difference is the number of features we used. In the first experiment we used all features described in Section 3.5, including the car feature, whereas the second experiment was carried out without the car feature. In the third and the fourth experiments we reduced the set of classes to eight by merging classes having a similar appearance in the data. Again, the two experiments differ by the use of the car feature.

4.2 Evaluation

The confusion matrices as well as the completeness and the correctness of the results achieved in the first two experiments (the ones using 14 classes) are shown in Tables 1 and 2; an example for the classification result is shown in Fig. 2. The overall accuracy of the classification was 63.5% if the car feature was used and 63.3% if it was not used. Thus, the overall accuracy, while being relatively poor in both cases, was hardly improved by that feature. The relatively poor overall accuracy is caused by the fact that some of the classes have a very similar appearance in the data, e.g. *sealed*, *road*, *sidewalk*, and *traffic islands*. Reasonable values of completeness and correctness could be achieved for buildings (> 80%). For trees, the completeness is also larger than 80%, but the correctness is much lower (62%). Both, for buildings and trees the main error source was errors in the DSM caused by areas with hardly any texture (buildings) or abrupt height changes (trees). One of the problems was the information reduction caused by the conversion of the images to 8 bit, but apparently the openCV matcher also had problems with non-fronto-parallel surfaces and with different illumination. The main impact of the car confidence feature was a considerable reduction of the false positive car detections, though the correctness of 28% achieved with this feature is still not satisfactory.

The evaluation of the experiments carried out with the reduced set of classes is presented in Tables 3 and 4. The overall accuracy increased to about 75%, which indicates that our classification scheme is reasonable, though there is room for improvement. The main error source is the confusion between *trees*, *grass*, and *agriculture*, again partly caused by DSM errors. In this setting, the impact of the car confidence feature is similar to its impact in the first group of experiments.

Class / Ref.	Road	Isl. V	Sidew.	Build.	Grass	Agr.	Water	Sealed	Isl. A	Beach	Railw.	Tree	Car	Bridge	Comp.
Road	10.59	0.06	0.47	0.19	0.39	0.08	0.75	2.49	--	0.01	--	0.12	0.21	--	**68.73**
Isl. V	0.10	0.01	--	--	--	--	0.01	--	--	--	--	--	--	--	**4.32**
Sidew.	1.28	--	0.35	0.13	0.32	0.05	0.21	0.69	--	--	--	0.09	0.04	--	**11.00**
Build.	0.38	--	0.04	13.02	0.48	0.06	0.18	1.30	--	--	--	0.16	0.18	0.23	**81.17**
Grass	0.30	--	0.14	0.57	14.48	1.03	0.16	0.94	0.05	0.40	0.06	4.61	0.07	--	**63.51**
Agr.	0.04	--	0.17	0.25	0.74	7.03		0.85	0.25	0.04	0.01	3.18	0.05	--	**55.70**
Water	0.08	--	--	0.04	0.02	--	0.21	0.06	--	--	--	--	0.02	--	**49.47**
Sealed	3.67		0.52	0.81	1.08	0.28	0.40	3.19	--	--	0.01	0.33	0.23	0.01	**30.22**
Isl. A	0.01	--	--	--	0.01	--	--	--	0.00	--	--	--	--	--	--
Beach	0.01	--	--	--	--	--	--	0.09	--	0.00	0.01	--	--	--	--
Railw.	0.05	--	0.09	--	0.10	0.05	--	0.04	--	--	0.01	0.04	--	--	**2.64**
Tree	0.21	--	0.01	0.12	2.57	0.14	0.01	0.10	--	--	--	14.29	0.01	--	**81.80**
Car	0.05	--	0.02	0.02	0.03	0.01	--	0.19	--	--	--	0.01	0.33	--	**49.96**
Bridge	0.08	--	--	0.09	--	--	--	0.04	--	--	--	--	0.02	0.00	--
Corr.	**62.84**	**6.74**	**19.10**	**85.45**	**71.54**	**80.51**	**11.07**	**31.98**	--	--	**10.60**	**62.59**	**28.14**	--	

Table 1: Confusion matrix for the experiment using 14 classes and the car confidence feature. All values are given in [%]. Overall accuracy: 63.50%. Abbreviations: Ref.: Reference; Isl. V: *traffic island with vegetation*; Sidew.: *sidewalk*; Build.: *building*; Agr.: *agricultural*; Isl. A: *traffic island with asphalt*; Railw.: *railway*; Comp. / Corr.: Completeness / Correctness.

Class / Ref.	Road	Isl. V	Sidew.	Build.	Grass	Agr.	Water	Sealed	Isl. A	Beach	Railw.	Tree	Car	Bridge	Comp.
Road	10.25	0.08	0.50	0.18	0.35	0.12	0.78	2.56	--	0.01	--	0.12	0.43	0.02	**66.56**
Isl. V	0.10	0.01	--	--	--	--	0.01	--	--	--	--	--	--	--	**5.57**
Sidew.	1.22	--	0.39	0.13	0.30	0.08	0.21	0.65	--	--	--	0.10	0.10	--	**12.22**
Build.	0.23	--	0.02	13.18	0.43	0.09	0.17	1.20	--	--	--	0.17	0.35	0.20	**82.18**
Grass	0.14	--	0.12	0.56	14.06	1.41	0.15	0.90	0.04	0.40	0.04	4.84	0.13	--	**61.68**
Agr.	0.03	--	0.15	0.25	0.68	7.21		0.84	0.16	--	0.01	3.19	0.09	--	**57.11**
Water	0.07	--	--	0.04	0.02	--	0.21	0.05	--	--	--	--	0.03	--	**49.39**
Sealed	3.47	0.01	0.55	0.80	0.97	0.43	0.40	3.10	--	--	0.01	0.34	0.46	0.01	**29.36**
Isl. A	0.01	--	--	--	0.01	--	--	--	0.00	--	--	--	--	--	--
Beach	--	--	--	--	--	--	--	0.10	--	0.00	0.01	--	--	--	--
Railw.	0.04	--	0.09	--	0.09	0.07	--	0.04	--	--	0.01	0.04	--	--	**2.80**
Tree	0.01	--	0.02	0.12	2.38	0.26	0.01	0.09	--	--	--	14.57	0.02	--	**83.43**
Car	0.03	--	0.02	0.02	0.03	0.02	--	0.19	--	--	--	0.01	0.33	--	**50.02**
Bridge	0.08	--	--	0.09	--	--	--	0.03	--	--	--	--	0.03	0.00	--
Corr.	**65.37**	**6.92**	**20.85**	**85.78**	**72.75**	**74.45**	**11.02**	**31.80**	--	--	**13.33**	**62.29**	**16.75**	--	

Table 2: Confusion matrix for the experiment using 14 classes without the car confidence feature. All values are given in [%]. Overall accuracy: 63.32%. Abbreviations: see caption of Table 1.

5. CONCLUSION

In this paper, a method for the classification of crossroads using MRF was proposed. It considered 3D information in the form of a DSM generated from multiple overlapping aerial images, as well as a car confidence feature to avoid problems with occlusions of the road surface by cars. Distinguishing 14 classes relevant in the context of crossroads, an overall accuracy of about 63.5% could be achieved. The main error sources were the confusion of object classes that are only distinguished by their relative alignment, but not by their appearance in the data (*road*, *sidewalk*, *sealed*) and errors in the DSM generation process. After merging the classes that are most similar in appearance, the overall accuracy was increased to 74.8%. In the future we want to improve our method by integrating more expressive features, e.g. HOG features or features related to car trajectories, and we will also integrate multi-scale features. Furthermore, we want to build a more sophisticated model of context based on Conditional Random Fields (Kumar & Hebert, 2006), also using improved models for the association potentials linking the class labels to the data. Finally, we will investigate whether the results can be improved by using image segments as the nodes of the graphical model.

ACKNOWLEDGEMENT

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.

Class / Ref.	Asph.	Build.	Grass	Agr.	Beach	Tree	Car	Bridge	Comp.
Asph.	25.14	1.31	1.76	0.48	0.01	0.53	0.44	0.04	**84.62**
Build.	1.80	13.14	0.48	0.08	--	0.16	0.24	0.13	**81.93**
Grass	1.54	0.63	14.65	1.21	0.41	4.65	0.09	--	**63.21**
Agr.	0.98	0.31	0.75	7.32	--	3.23	0.03	--	**58.02**
Beach	0.10	--	--	--	0.00	--	--	--	--
Tree	0.31	0.12	2.57	0.18	--	14.28	0.01	--	**81.73**
Car	0.26	0.02	0.03	0.01	--	0.01	0.31	--	**47.94**
Bridge	0.14	0.06	--	--	--	--	0.02	0.00	--
Corr.	**83.02**	**84.24**	**72.37**	**78.94**	--	**62.48**	**27.22**	--	

Table 3: Confusion matrix for the experiment using 8 classes and the car confidence feature. All values are given in [%]. Overall accuracy: 74.84%. Asph: *asphalt*.

Class / Ref.	Asph.	Build.	Grass	Agr.	Beach	Tree	Car	Bridge	Comp.
Asph.	24.63	1.28	1.60	0.72	0.01	0.56	0.88	0.03	**82.90**
Build.	1.53	13.26	0.44	0.11	--	0.16	0.40	0.13	**82.70**
Grass	1.30	0.62	14.21	1.60	0.41	4.87	0.17	--	**61.30**
Agr.	0.96	0.27	0.68	7.38		3.25	0.07	--	**58.53**
Beach	0.10	--	--	--	0.00	--	--	--	--
Tree	0.10	0.12	2.36	0.27	--	14.58	0.02	--	**83.48**
Car	0.25	0.02	0.03	0.02	--	0.01	0.32	--	**49.33**
Bridge	0.12	0.07	--	--	--	--	0.03	0.00	--
Corr.	**84.95**	**84.79**	**73.49**	**72.99**	--	**62.24**	**17.01**	--	

Table 4: Confusion matrix for the experiment using 8 classes without car confidence feature. All values are given in [%]. Overall accuracy: 74.39%. Asph: *asphalt.*

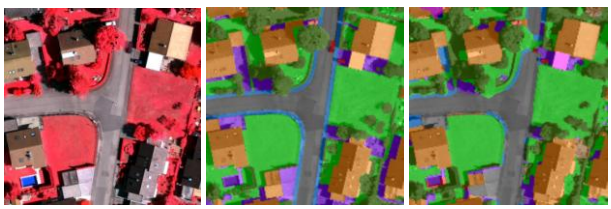

Figure 2: Classification results. Left: original orthophoto; centre: ground truth superimposed to the intensity mage; right: results achieved with 14 classes and the car feature superimposed to the intensity image.

REFERENCES

Barsi, A., Heipke, C., 2003. Artificial neural networks fort the detection of road junctions in aerial images. In: *IAPRSIS* XXXIV (3/W8), pp. 18-21.

Birchfield, S., Tomasi, C., 1998. A pixel dissimilarity measure that is insensitive to image sampling. *IEEE-TPAMI* 20(4), pp. 401-406.

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

Dalal, N. and Triggs, B., 2005. Histograms of Oriented Gradients for Human Detection. *Proc. of IEEE Conference Computer Vision and Pattern Recognition*: 886-893.

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: A statistical view of boosting. *Ann. Stat.* 28(2):337–407.

Geman, G. and Geman, D., 1984. Stochastic relaxation, Gibbs distribution and Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6(6): 721-741.

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

Hirschmüller, H., 2008. Stereo processing by semiglobal matching and mutual information. *IEEE TPAMI* 30(2):328-341.

Kumar, S. and Hebert, M., 2006. Discriminative Random Fields. *Int'l. J. Computer Vision*, 68(2): 179-201.

Kurz, F., Rosenbaum, D., Leitloff, J., Meynberg, O., Reinartz, P., 2011. *In: Proceedings of International Conference on SMPR 2011* (on CD-ROM).

Leitloff, J., Hinz, S., Stilla, U., 2010. Vehicle extraction from very high resolution satellite images of city areas. *IEEE Trans. on Geoscience and Remote Sensing*, 48(7): 2795-2806

Li, S. Z., 2009. Markov Random Field modeling in image analysis. 3rd ed., Springer, London, 357 p.

Lienhart, R., Kuranov, A., Pisarevsky, V., 2003. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. *Pattern Recognition*, vol. 2781, pp. 297–304

Mayer, H., Hinz, S., Bacher, U., and Baltsavias, E., 2006. A test of automatic road extraction approaches. In: *IAPRSIS* XXXVI-3, pp. 209-214.

OpenCV, 2012.
http://opencv.itseez.com/modules/calib3d/doc/calib3d.html

Ravanbakhsh, M.; Heipke, C.; Pakzad, K., 2008a. Road junction extraction from high resolution aerial imagery. *Photogrammetric Record* 23 (124):405-423.

Ravanbakhsh, M.; Pakzad, K.; Heipke, C., 2008b. Automatic extraction of traffic islands from aerial images. *Photo-grammetrie–Fernerkundung–Geoinformation* 5(2008):375-384.

Rutzinger, M., Rottensteiner, F. Pfeifer, N., 2009. A comparison of evaluation techniques for building extraction from airborne laser scanning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 2(1):11-20.

Stilla, U., Michaelsen, E., Sörgel, U., Hinz, S., Ender, J., 2004. Airborne monitoring of vehicle activity in urban areas. *IAPRSIS* XXXV- B3, pp. 973–979.

Tieu, K., Viola, P., 2004. Boosting image retrieval. *Int. J. Comput. Vis.*, vol. 56, no.1-2, pp. 17–36

Viola, P., Jones, M. J., 2004. Robust real-time face detection. *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154

Vishwanathan, S., Schraudolph, N. N., Schmidt, M. W., Murphy, K. P., 2006. Accelerated training of conditional random fields with stochastic gradient methods. *23rd Int. Conf. on Machine Learning*: 969-976.