A Two-Layer Conditional Random Field for the Classification of Partially Occluded Objects

Sergey Kosov[1], Pushmeet Kohli[2], Franz Rottensteiner[1] and Christian Heipke[1]

[1]Institute of Photogrammetry and GeoInformation, Leibniz Universitat Hannover, Germany
[2]Microsoft Research Cambridge, UK

Abstract

Conditional Random Fields (CRF) are among the most popular techniques for image labelling because of their flexibility in modelling dependencies between the labels and the image features. This paper proposes a novel CRF-framework for image labeling problems which is capable to classify partially occluded objects. Our approach is evaluated on aerial near-vertical images as well as on urban street-view images and compared with another methods.

1. Introduction

Labeling of image pixels is a classical problem in pattern recognition. Probabilistic models of context such as Markov Random Fields (MRF) [15] or Conditional Random Fields (CRF) [13] have been increasingly used to model dependencies between labels and/or data at neighbouring image sites. This results in smoothed label images compared to local classifiers. A recent comparison of smooth labelling techniques [22] has shown that smoothing is essential in this context, with CRF performing best among the compared techniqes.

Labelling techniques usually determine a class label for each pixel of an image. This causes problems if the objects to be detected are partially occluded. For instance, the appearance of streets, sidewalks and buildings may not be clear for a computer if they are largerly occluded by objects such as cars or trees. In remote sensing images, characterized by near-vertical views, this has been known to be a problem for a long time, in particular in the context of automated road extraction. Model-based techniques have tried to overcome this problem by treating such objects as context objects in an ad-hoc manner [8, 6], but a systematic statistical model for dealing with occlusions is still missing. Whereas CRF have been applied successfully to many labelling tasks in computer vision, pattern recognition and remote sensing [13, 22, 23, 27], they also have problems with proper labelling of partially occluded objects, in particular if the occluded objects are those one is actually interested in. In this paper we introduce a two-layered Conditional Random Field (*tCRF*), which can handle this problem by explicitly modelling *two* class labels for each image site, one for the occluded object and one for the occluding one; in this way, the 3D structure of the scene is explicitly considered in the structure of the CRF. Labelling might also be supported by depth information obtained from image matching.

Previous work on the recognition of partially occluded objects includes [14], where 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. There have been a few attempts to include multiple layers of class labels in CRFs [12, 23, 27]. However, all these papers also use part-based models where the additional layer does not explicitly refer to occlusions, but encodes another label structure. In [12] and [23], multiple layers represent a hierachical object structure, *i.e.* each object on higher level interacts with its smaller parts on lower level. In [27], the part-based model is motivated by the method's potential to incorporate information about the relative alignment of object parts and to model longe-range interactions. However, occluded objects are not explicitly reconstructed. Such a part-based approach is not applicable to objects such as roads in near-vertical views. Roads do not consist of parts having a specific appearance and appearing in a fixed spatial structure. Besides, 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), which is not available in remote sensing imagery. As a consequence, methods relying on such

a reference direction are not applicable to this class of images. In this respect, the method described in this paper is more general and can be applied to both near-vertical images and images with a horizontal viewing direction. Information about the vertical structure of a scene can be incorporated in scenarios where it makes sense to do so, but it is not a preriquisite for our method to work.

In [29], MRFs are also expanded by additional layers in the temporal domain, related to the previous and subsequent frames in a video sequence. The interactions between these temporal layers are designed for the detection of moving objects. Occlusions are not dealt within this publication. In [7], occluded areas are recovered by fitting geometrical primitives to large background objects using their visible parts, so the whole classification process is supported with additional frameworks, namely contextual prediction [24], and non-parametric label transfer [16]. This work directly adresses the problem of recovering the occluded areas and we show that our method, which consists of only one CRF framework can outperform the reported in [7] method by accuracy of classification for some classes. It is worth nothing that none of the cited publications use depth information as an additional cue to deal with occlusions.

We solve the problem of labelling partially occluded objects by explicitly considering the 3D structure of the scene. For each image site we have two class labels, one corresponding to an occluded object and the other to the occluding one, using a specific class label to encode that no occlusion occurs. The relations between the two class labels per site and the mutual dependencies between class labels at neighbouring sites in each of the two layers are explicitly modelled. Thus, the information from neighbouring unoccluded objects as well as information from the occluding layer will contribute to an improved labelling of occluded objects. Two layers are sufficient for applications which we focus on, though the principle may be expanded to models with multiple layers. To our knowledge, such a two-layered model has not been applied yet. The interaction model between neighbouring image sites is a contrast-sensitive model which considers the relative frequency of class transitions [19]. The data-dependent terms of our CRF are based on the Random Forest approach [3]. Our method is demonstrated on the task of correctly labelling urban scenes containing crossroads, one of the major problems in road extraction [20], with the main goal of correctly predicting the class labels of image sites corresponding to the road surface. We also evaluate our method on urban street-view images and compare the results with those achieved in [7], though we will also evaluate the quality of detection for the occluding objects.

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., $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M)^T$, where $\mathbb{S}$ is the set of all sites. With each site i we associate a discrete class label x_i from a given set of classes $\mathbb{C}$. Collecting the class labels x_i in a vector $\mathbf{x} = (x_1, x_2, \ldots, x_M)^T$, we can formulate the problem of image classification as finding the label configuration $\hat{\mathbf{x}}$ that maximises the posterior probability of the labels given the observations, $p(\mathbf{x}|\mathbf{y})$, thus $\hat{\mathbf{x}} = \arg\max_x p(\mathbf{x}|\mathbf{y})$. A CRF is a model of $p(\mathbf{x} \mid \mathbf{y})$ with an associated graph whose nodes are linked to the image sites and whose edges model interactions between neighbouring sites. Restricting ourselves to CRFs where only pairs of nodes interact, the joint posterior $p(\mathbf{x}|\mathbf{y})$ can be modelled by [13]:

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

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, 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.

3. Method

3.1. Two-Level Conditional Random Fields

In order to classify partially occluded regions we distinguish objects corresponding to the *base level*, *i.e.* the most distant objects that cannot occlude other objects but could be occluded, from objects corresponding to the *occlusion level*, *i.e.* all other objects. We separate the objects according to the background-foreground principle: the base level consists of objects such as roads, buildings or grass, whereas the occlusion level includes objects such as cars and pedestrians. Consequently, we build a *two-level CRF*. Rather than having one label x_i per image site, we determine two such labels x_i^b and x_i^o, corresponding to the base and occlusion levels, respectively. In general, one occlusion level is sufficient for separating foreground from background. Accordingly, we have two sets of classes, namely $\mathbb{C}^b$ and $\mathbb{C}^o$, corresponding to objects at the base and occlusion levels, respectively, with $x_i^b \in \mathbb{C}^b$ and $x_i^o \in \mathbb{C}^o$. Currently, we model $\mathbb{C}^b$ and $\mathbb{C}^o$ to be mutually exclusive, thus $\mathbb{C}^b \bigcap \mathbb{C}^o = \emptyset$. $\mathbb{C}^o$ includes a special class *void* $\in \mathbb{C}^o$ to model situations where the base level is not occluded (Fig. 1). The goal of classification is to determine the most probable values for both x_i^b and x_i^o given the data $\mathbf{y}$. We

model the posterior probability $p(\mathbf{x}^b, \mathbf{x}^o|\mathbf{y})$ directly, expanding the model in Eq. 1:

$$p(\mathbf{x}^b, \mathbf{x}^o|\mathbf{y}, \theta) = \frac{1}{Z} \prod_{i \in \mathbb{S}} \varphi_i^b(x_i^b, \mathbf{y})^{\theta_1} \cdot \varphi_i^o(x_i^o, \mathbf{y})^{\theta_2} \cdot \prod_{i \in \mathbb{S}} \prod_{j \in \mathcal{N}_i} \psi_{ij}^b(x_i^b, x_j^b, \mathbf{y}, \theta_6, \theta_7)^{\theta_3} \cdot \psi_{ij}^o(x_i^o, x_j^o, \mathbf{y}, \theta_6, \theta_7)^{\theta_4} \cdot \prod_{i \in \mathbb{S}} \xi_i(x_i^b, x_i^o, \mathbf{y})^{\theta_5}. \quad (2)$$

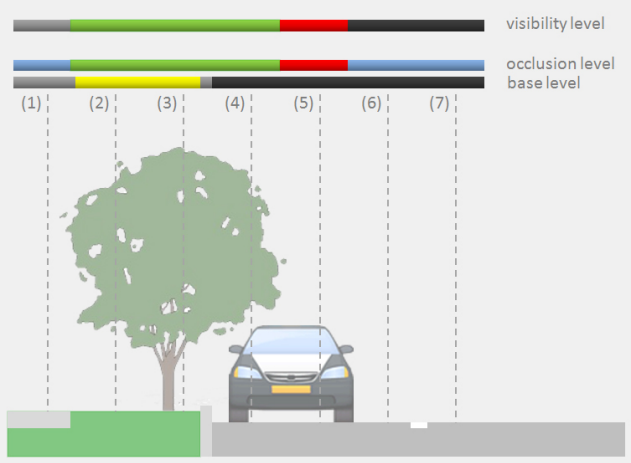


Figure 1. Two-level set of classes, where labels are represented by colours. Base level: black - street; grey - sidewalk; yellow - grass; Occlusion level: green - tree; red - car; blue - *void*. The visibility level depicts the classes as seen for the sensor (aerial view).

In Eq. 2, θ are model parameters: $\theta_1, \theta_2, \ldots, \theta_5 \in \theta$ are weights modulating the influence of the individual terms in the classification and $\theta_6, \theta_7 \in \theta$ are parameters of the potentials ψ_{ij}^b and ψ_{ij}^o. $\mathcal{N}_i$ is the neighbourhood of site i (thus, j is a neighbour of i). The *association potentials* φ_i^b and φ_i^o link the data $\mathbf{y}$ with the class labels x_i^b, x_i^o of image site i. The association potential can be considered as a measure of how likely a site i will take labels x_i^b or x_i^o given all image data $\mathbf{y}$ and ignoring the effects of other sites in the image. The *within-level interaction potentials* ψ_{ij}^b and ψ_{ij}^o model the dependencies between the data $\mathbf{y}$ and the labels at two neighbouring sites i and j at the base and occlusion levels, respectively; these potentials correspond to the interaction potentials in Eq. 1. They are related to the probability of how likely the labels of neighbouring sites from one layer x_i^l and x_j^l, $l \in \{o, b\}$ are to occur at neighbouring sites given the image data $\mathbf{y}$. Finally, in order to link the base and occlusion levels, we define a new *inter-level interaction potential* $\xi(x_i^b, x_i^o, \mathbf{y})$, which models the dependencies between labels from different layers, x_i^b and x_i^o, and the data $\mathbf{y}$. It is a measure of how likely an occlusion of an object at the base level with class label x_i^b by an object from the occlusion level with class label x_i^o is to occur, considering the data $\mathbf{y}$.

Fig. 2 shows the structure of our tCRF model. The dark data nodes represent the input information from sites with occlusion, *i.e.* where only the occluding object is visible. The reason why we have split the levels is to increase the accuracy of the labelling of occluded regions, *i.e.* to reveal the labels of the dark label nodes in Fig. 2, where the association potentials could not provide the corresponding base level nodes with reliable information because the data corresponding to the base level are not observable.

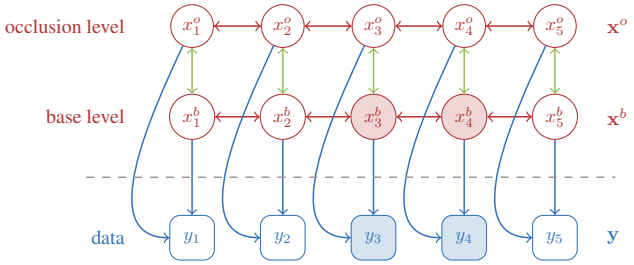


Figure 2. Structure of the tCRF model. The second dimension and additional links between data and labels are omitted for simplicity. Squares and circles correspond to observations and labels, respectively. The dark nodes correspond to a region with occlusion. The graph edges represent dependencies between the nodes.

In a training phase we determine the parameters of the potentials in Eq. 2, which requires fully labelled training images. The classification of new images is carried out by maximizing the posterior probability in Eq. 2. The model is very general in terms of the definition of the potentials φ_i, ψ_{ij} and ξ_i. Our definitions of the potentials as well as the techniques used for training and inference are described in the subsequent sections. For the sake of simplicity, we will omit the indices o or b in the discussion of the association and the within-layer interaction potentials, assuming the same functional model to be valid for both layers.

3.2. 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 $\varphi_i(x_i, \mathbf{y}) = p(x_i = c|\mathbf{f}_i(\mathbf{y}))$ [13], 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 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 Random Forest (*RF*) [3] 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 $p(x_i = c \mid \mathbf{f}_i(\mathbf{y})) = N_c/N_T$.

3.3. Within-Level Interaction Potential

The within-level interaction potential $\psi_{ij}(x_i, x_j, \mathbf{y})$ describes how likely the pair of neighbouring sites i and j is to take the labels $(x_i, x_j) = (c, c')$ given the data: $\psi_{ij}(x_i, x_j, \mathbf{y}) = p(x_i = c, x_j = c'|\mathbf{y})$ [13]. We generate a 2D histogram $h'(x_i,\ x_j)$ of the co-occurrence of labels at neighbouring image sites from the training data, *i.e.* $h'(x_i = c, x_j = c')$ is the number of occurrences of the classes $(c,\ c')$ at neighbouring sites i and j. We scale the rows of $h'(x_i,\ x_j)$ 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 $h(x_i,\ x_j)$. Our contrast-sensitive definition of $\psi_{ij}(x_i, x_j, \mathbf{y}) \equiv \psi_{ij}(x_i, x_j, d_{ij})$ is obtained by applying a penalization depending on the Euclidean distance $d_{ij} = \|\mathbf{f}_i(\mathbf{y}) - \mathbf{f}_j(\mathbf{y})\|$ of the node feature vectors $\mathbf{f}_i$ and $\mathbf{f}_j$ to the diagonal elements of $h(x_i,\ x_j)$:

$$\psi_{ij}(x_i, x_j, \mathbf{y}) = \begin{cases} \theta_6 \cdot e^{-\theta_7 \cdot d_{ij}^2} \cdot h(x_i,\ x_j) & \text{if } x_i = x_j \\ h(x_i,\ x_j) & \text{otherwise} \end{cases} \tag{3}$$

In Eq. 3, the parameter $\theta_6 \in \theta$ modulates the degree to which the within-level interaction potential favours identical classes at neighbouring sites, whereas $\theta_7 \in \theta$ modulates the contrast-sensitive term. The parameters θ_6 and θ_7 are shared by both inter-level potential functions (base and occlusion levels). 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, but 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 [2] by the use of the normalised histograms $h_\psi(x_i,\ x_j)$ in Eq. 3. As a consequence, class transitions become more likely, depending on the frequency with which they occur in the training data. Again, the training of the models for the base and the occlusion levels, $\psi^b_{ij}(x^b_i, x^b_j, \mathbf{y})$ and $\psi^o_{ij}(x^o_i, x^o_j, \mathbf{y})$, respectively, are carried out independently from each other using fully labelled training data.

3.4. Inter-Level Interaction Potential

The inter-level interaction potential $\xi_i(x^b_i, x^o_i, \mathbf{y})$ describes how likely two variables of site i are to take the labels $(x^b_i, x^o_i) = (c, c')$ given the data: $\xi_i(x^b_i, x^o_i, \mathbf{y}) = p(x^b_i = c, x^o_i = c'|\mathbf{y})$. Here $c \in \mathbb{C}^b$ and $c' \in \mathbb{C}^o$. We introduce a new set of class labels $\mathbb{C}^i = \mathbb{C}^b \times \mathbb{C}^o$, which encodes all the possible combinations of two labels $c \in \mathbb{C}^b$ and $c' \in \mathbb{C}^o$ by one label $c'' \in \mathbb{C}^i$. Thus, the potential function becomes $\xi_i(x^b_i, x^o_i, \mathbf{y}) = p(\{x^b_i; x^o_i\} = c'' \mid \mathbf{f}_i(\mathbf{y}))$, which is modelled by the RF approach in the same way as described in Sec. 3.2.

3.5. Training and Inference

Exact probabilistic methods for training of a CRF are computationally intractable [13, 25]. Thus, approximate solutions have to be used. We determine the parameters of the association, within-level interaction and inter-level interaction potentials separately, using only a part of the training data. The association potentials and inter-level interaction potential are trained using the OpenCV implementation of the RF approach [18]. For the each class we use the same amount of training samples $N_{samples}$, which are chosen randomly from the training dataset. This results in a total of $N_{samples} \times N_{classes}$ samples that is used for training of both the association and the inter-level interaction potential. The within-level 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. 3.3, taking into account all image sites in the training data. It is a preriquisite of our method that the training data also have two separate layers of labels, one for the base and one for the occlusion layers, respectively. The parameters $\theta = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7\}$ are trained using the Powell search method [11], an iterative optimisation algortihm that does not require an estimate for the gradient of the objective function. We determine θ by maximising the sum Ω of the diagonal elements of the confusion matrix obtained by classifying the part of the training data that was not used for training the potentials. Exact inference is also computationally intractable for CRFs. We use max-product Loopy Belief Propagation, a standard technique for probability propagation in graphs with cycles [9].

4. Evaluation

4.1. Experiment Setup

As our method requires test data to consist of two separate layers of class labels for the base and occlusion levels, respectively, we only can use datasets providing this information for evaluation. We used the *Vaihingen*[1] and the *StreetScene* [1] datasets for that purpose. The Vaihingen dataset consists of 1440 scenes with a size of 250×250 pixels. Each scene is a colour-infrared (CIR) true orthophoto and a height grid (digital surface model; DSM) generated from wide baseline multiple overlapping airbourne images with a ground sampling distance (GSD) of 8 cm [10]. Both the CIR image and the DSM are geo-coded, and they are defined on the same grid. The reference labels were generated by manually labelling these data in two separate layers. The StreetScene dataset consists of 3547 colour images of 1280×960 pixels and contains a reference in the form of polygons that also consider hidden object parts and hence

[1]The Vaihingen data set was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) [4].

could be used to define the two-layered reference required by our method. The Vaihingen data, based on aerial views, are available in a reference frame aligned with the North direction, which is not helpful to structure the scene because roads and buildings (the dominant objects in these data) are not necessarily aligned in North-South or East-West directions. As the original images were taken at the same flying height, all objects appear at a similar scale. On the other hand, for the StreetScenes data, the vertical (y coordinate axis) provides a physically defined reference direction that is clearly related to the scene structure. Furthermore, the distances at which objects are observed vary considerably, so that the scale of objects varies both within and between different scenes.

For the Vaihingen data we chose the nodes of the graphical model to correspond to single pixels, whereas for the StreetScenes data we used image patches of 5×5 pixels. Thus, each graphical model consisted of 250×250 and 256×192 nodes for the two datasets, respectively. The neighbourhood $\mathcal{N}_i$ of an image site i in Eq. 2 (which defines the red edges of the graphical model in Fig. 2) is chosen to consist of the direct neighbours of site i in the data grid. The reference of the Vaihingen dataset has six classes: *asphalt* (*asp.*), *building* (*bld.*), *tree*, *grass*, *agricultural* (*agr.*) and *car*, so that $\mathbb{C}^b = \{asp., bld., grass, agr.\}$ and $\mathbb{C}^o = \{tree, car, void\}$. The reference of the StreetScenes dataset has 9 classes: *road*, *sidewalk*, (*sdw.*) *bld.*, *store* (*str.*), *tree*, *sky*, *car*, *pedestrian* (*ped.*) and *bicycle* (*bic.*). Since the reference for this dataset is given by polygons, it occures that some image areas are not covered by any polygon. In order to keep our model consistant, we introduce here class *unknown* (*unk.*) and mark with it all the uncovered areas at the base level. At the occlusion level, such areas are marked as *void*, So that $\mathbb{C}^b = \{road, sdw., bld., str., tree, sky, unk.\}$ and $\mathbb{C}^o = \{ped., car, bic., void\}$.

In each test run, 50% of the images were used for RF training. Our RFs consist of $N_T = 100$ trees of maximal depth 25. For the training our RF-based potential functions we used $N_{samples} = 10^5$ samples. We used 8.3% of the images for learning the model parameters θ, and the remaining 41.7% of images for testing. The classification results were compared with the reference; we report the completeness and the correctness (recall and precision) of the results per class as well as the overall classification accuracy [21].

4.2. Features

The site-wise feature vectors $\mathbf{f}_i(\mathbf{y})$ representing the data in the association potentials and inter-level interaction potential depend on the dataset. For the Vaihingen dataset the original data consist of the three colour values of the CIR orthophoto and the associated DSM height for each pixel. The StreetScenes dataset offers only 3-channel colour images. From these original data, we derive the site-wise feature vectors $\mathbf{f}_i(\mathbf{y})$, each consisting of N_f features. For numerical reasons, all features are scaled linearly into the range $[0; 255]$ and then quantized by 8 bit.

The features used for both datasets comprise the *image intensity* (*int*), calculated as the average of non-infrared channels and the *saturation* (*sat*) component after transforming the image to the LHS colour space. 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 site i (7×7 pixels for var_{int}, 13×13 pixels for var_{sat} and var_{grad}, in both cased evaluated at the original resolution).

For the Vaihingen dataset includes CIR images, we make use of the *normalized difference vegetation index* (*NDVI*), derived from the near infrared and the red band of the CIR orthophoto [17]. For Vaihingen we also determine a digital terrain model (DTM) by applying a morphological opening filter to the DSM with a structural element size corresponding to the size of the largest off-terrain structure in the scene, followed by a median filter with the same kernel size. The DTM is used to derive a *normalised DSM* (*nDSM*) [26], *i.e.* a model of the height differences between the DSM and DTM. The nDSM describes the relative elevation of objects above ground and its value at image site i is directly used as a feature. Finally, the 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. The *dist* feature is the distance of an image site to its nearest edge pixel. The last feature used for the Vaihingen data is the gradient strength of the DSM ($||\nabla DSM||$).

The StreetScenes dataset has no infra-red channel and no DSM. Nevertheless the classes in images of this dataset have a strong dependency on the *y image coordiante* that reflects the vertical structure of the scenes. For instance, the sky is usually above road and buildings have vertical structure [28]. Consequently, we use the y coordinate of a node as a feature. This shows that we can incorporate this information when it is helpful (horizontal viewing direction, street scenes), but do not rely on it when it is not available (remote sensing imagery). For similar reasons we make use of histogram of oriented gradients (HOG) features [5] only for the StreetScenes dataset. 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 determined relative to the vertical image axis (which would correspond to a model relative to the North direction in the aerial image case). We extract nine features from the HOG descriptor, namely the value corresponding to each direction bin ($HOG_0, HOG_1, \ldots, HOG_9$).

For both datasets we make use of multiscale features. That is, the features described above are derived at three

different scales. The first scale corresponds to the individual sites, the second and the third are calculated as the average in a local neighbourhoods. For *int*, *sat*, *NDVI* and *nDSM*, these neighbourhoods were chosen to be 45×45 and 91×91 pixels for the second and the third scales, respectively. For var_{int}, var_{sat}, var_{grad}, *dist*, $||\nabla DSM||$ and the HOG features the neighbourhoods were chosen to be 10×10 and 100×100 pixels for scales two and three, respectively.

4.3. Results and Discussion

To assess the tCRF model we carried out a number of different experiments. At the first stage we used the *Vaihingen* dataset and performed two experiments: in the first experiment (*CRF*), each layer was processed independently, thus the inter-level interaction potentials were not considered. In the second experiment (*tCRF*) we use the tCRF model with the inter-level interaction potentials. Fig. 3 shows the convergence behaviour of the Powell method for training the parameters θ in Eq. 2 for both cases. It shows that originally the procedure converges more slowly for the *tCRF* method, probably due to a relatively poor initialisation of some parameters, but in the end-iterated state, a larger value of the objective function Ω can be achieved.

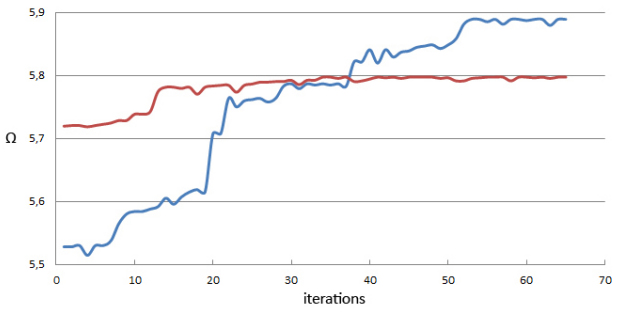
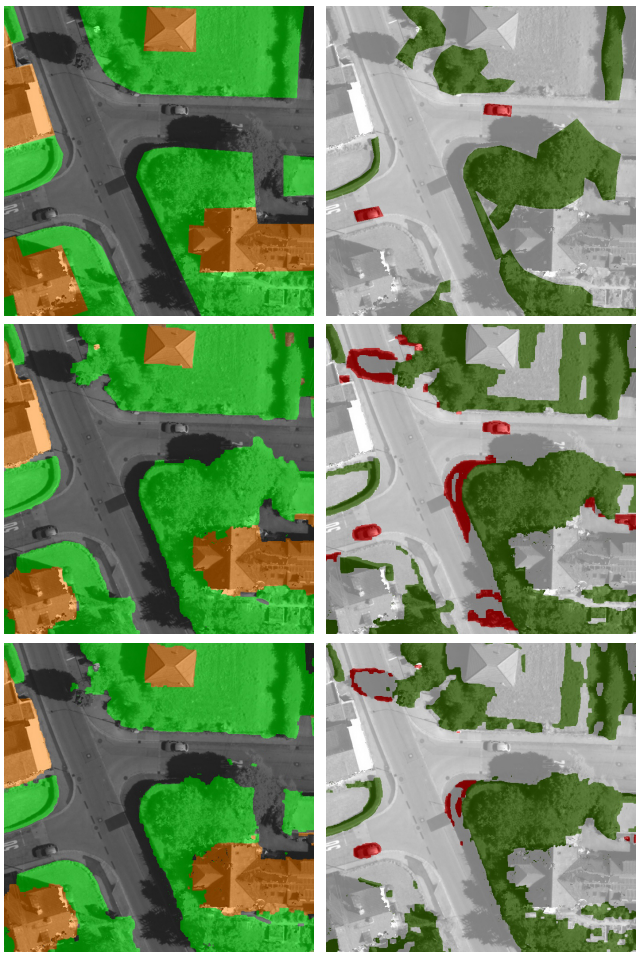


Figure 3. Convergence of the Powell search method: Red curve: CRF; blue curve: $tCRF$.

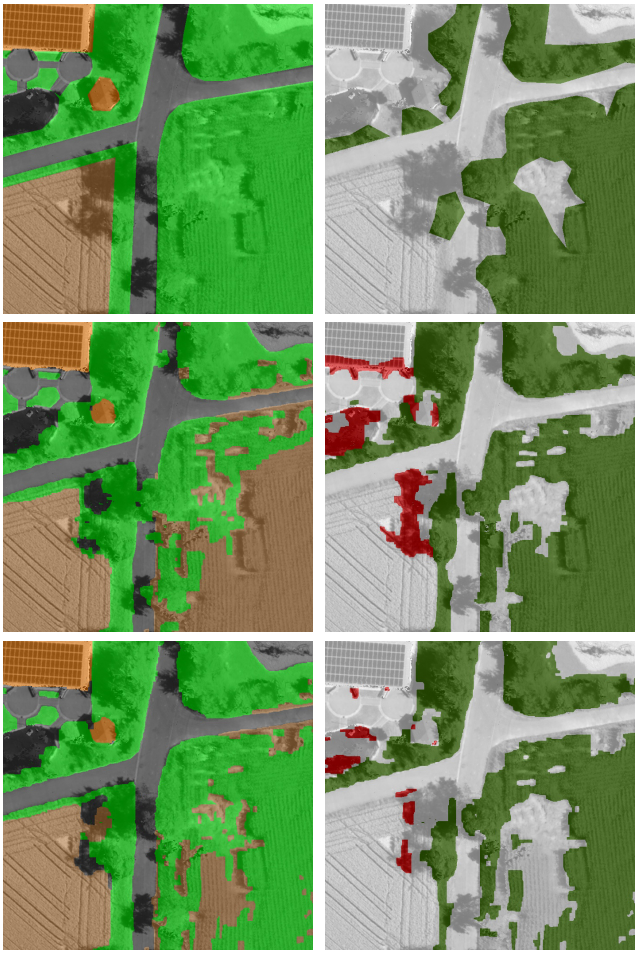
Fig. 4 and 5 show the results of the experiments for two Vaihingen scenes. In both figures we can observe that our two-level model considerable improves the road classification in comparison to the state-of-the-art one-layer model. For example, in the right part of the scene in Fig. 4, the *tCRF* model successfully extracts a road part that is completely occluded with a tree, while *CRF* wrongly labels this area as grass. This improvement is possible because the *tCRF* models explicitly considers occlusion, the results of the base level receiving information from spatially neighbouring image sites, multi-scale features, and the second layer of labels. Fig. 5 also shows how an occluded road can be correctly classified by the *tCRF*. In addition, the grass area in the right lower part of the scene is labelled as agricultural by the *CRF* model, in spite of the occlusion level saying that this region is covered by trees. Agricultural regions are rarely covered by a forest, and the *tCRF* model can use this knowledge (derived from the training data) in order to classify this area correctly. For both scenes we can observe many false positives for the class *car*. Their number is reduced considerably by the *tCRF* model, though at the cost of a few false negative cars (Fig 4). This is also reflected in the quality numbers in Tab. 1.



(a) (b)

Figure 4. Vaihingen (scene 22). First row: reference, second row: CRF, third row: $tCRF$. (a) Base level; (b) Occlusion level. Gray: *asp.*; orange: *bld.*; green: *grass*; beige: *agr.*; white: *void*; darkgreen: *tree*; red: *car*.

The completeness and the correctness as well as the overall accuracy of the results achieved in these two experiments are shown in Tab. 1. Using the *CRF* model, the overall accuracy of the classification was 82.6% for the base level and 80.4% for the occlusion level. In the second (*tCRF*) experiment the overall accuracy for the base level was 86.6%. The improvement can be attributed by more accurate classification in the occlusion areas (cf. Fig. 4 and 5). From the Tab. 1, we can also observe that both the completeness and correctness of car class are still very low. We think that this is due to the fact that cars are relatively small regions and so



(a) (b)

Figure 5. Vaihingen: scene 43. First row: groundtruth, second row: CRF, third row: $tCRF$. (a) Base level; (b) Occlusion level. Gray: *asp.*; orange: *bld.*; green: *grass*; beige: *agr.*; white: *void*; darkgreen: *tree*; red: *car*.

are described with our features not well enough. The outcome of additional car-detector may correct this situation. Nevertheless our $tCRF$ model has almost double correctness value for cars, then CRF model, while having smaller completeness value. As far as completeness and correctness are concerned, the major improvement is an increased correctness for *asp.* and an improved completeness for grass. The class *agr.* has a rather low correctness in the model CRF. For the occlusion level, we observe the best performance when using *tCRF*.

At the second stage of experiments we used the *StreetScene* dataset and also performed two experiments: *CRF* and *tCRF* but this time we compare our method with those, reported in [7], namely *Most Confident* (*MC*) and *Method of Guo and Hoiem* (*GH*). The results are presented in Tab. 2 and some classification examples are depicted in Fig. 6.

	CRF		$tCRF$	
	$Cm.$	$Cr.$	$Cm.$	$Cr.$
asp.	80.2 %	90.3 %	85.0 %	87.7 %
bld.	86.5 %	78.3 %	85.9 %	82.5 %
grass	82.7 %	85.5 %	88.3 %	87.8 %
agr.	84.1 %	64.4 %	85.4 %	84.2 %
OA$_{base}$	**82.6 %**		**86.6 %**	
void	78.1 %	96.9 %	86.8 %	95.7 %
tree	90.4	58.0 %	86.3 %	65.4 %
car	72.7	11.5 %	47.7 %	19.4 %
OA$_{occl}$	**80.4 %**		**86.3 %**	

Table 1. Completeness (Cm.), Correctness (Cr.) and overall accuracy (OA) of the results for Vaihingen dataset.

	CRF	MC	GH	$tCRF$
road	90.9 %	92.5 %	93.0 %	**94.5 %**
sdw.	0.5 %	28.5 %	**52.5 %**	0.3 %
bld.	54.3 %	**90.5 %**	90.0 %	46.6 %
str.	0.0 %	0.5 %	**11.0 %**	0.1 %
tree	92.3 %	69.5 %	73.5 %	**92.9 %**
sky	77.6 %	68.0 %	79.0 %	**80.4 %**

Table 2. Completeness of the results for *StreetScene* dataset. CRF: state-of-the-art 1-layer CRF; MC: Most Confident method and GH: method of Guo and Hoiem, both reported in [7]; $tCRF$ our method.

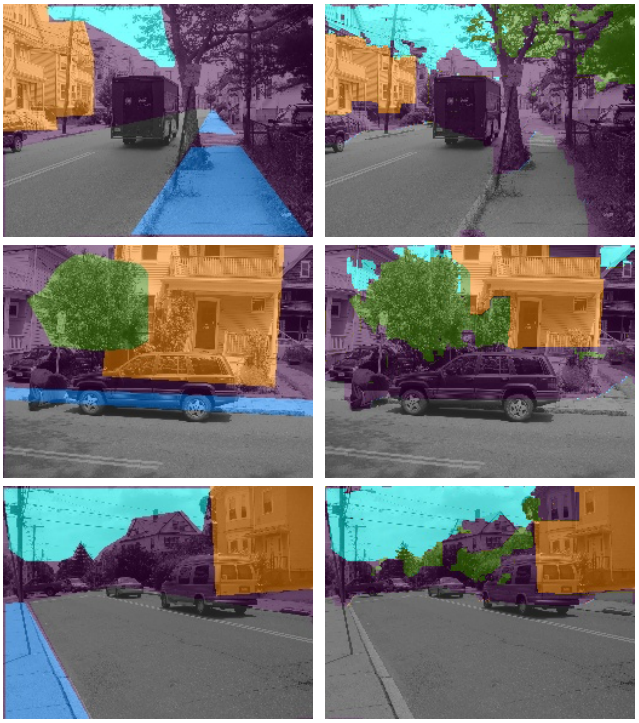
As we can see from Tab. 2 neighter CRF nor $tCRF$ can distinguish *sidewalk* and *store* classes. Newertheless, our $tCRF$ method beats the baseline GH method in terms of classification accuracy for 3 of 6 classes: *road*, *tree*, *sky*.

5. Conclusion

In this paper we have presented a novel approach for considering occlusions in classification based on CRF, the two-level CRF model. Due to its two-level structure it is capable to improve the accuracy of object detection for partially occluded objects. The method was evaluated on the set of airborne- as well as on street-view images and showed a considerable improvement of the overall accuracy in comparison to the classical CRF approach. In the future we want to extend our two-level architecture to n-level architecture and apply it to different classes of data. This will include the removal of the restriction that the sets of Classes corresponding to different layer have an empty intersection ($\mathbb{C}^b \bigcap \mathbb{C}^o = \emptyset$). Furthermore, we want to include additional cues to obtain a better classification accuracy for the occlusion level, in particular for the class car.

References

[1] S. M. Bileschi. Streetscenes: Towards scene understanding in still images. Technical report, PHD DISSERTATION, MASSACHUSETTES INST. OF TECHNOLOGY, 2006. 4



(a) (b)

Figure 6. StreetScenes: examples of base level classification. (a) reference; (b) $tCRF$ classification result. Gray: *road*; blue: *sdw.*; orange: *bld.*; green: *tree*; cyan: *sky*.

[2] Y. Boykov and M. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In *Proc. ICCV*, volume I, pages 105–112, 2001. 4

[3] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. 2, 3

[4] M. Cramer. The DGPF test on digital aerial camera evaluation - overview and test design. *Photogrammetrie-Fernerkundung-Geoinformation*, 2(2010):73–82, 2010. 4

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, pages 886–893, 2005. 5

[6] A. Grote, C. Heipke, and F. Rottensteiner. Road network extraction in suburban areas. *Photogrammetric Record*, 27:8–28, 2012. 1

[7] R. Guo and D. Hoiem. Beyond the line of sight: Labeling the underlying surfaces. In *ECCV*, pages 761–774, 2012. 2, 7

[8] S. Hinz and A. Baumgartner. Automatic extraction of urban road networks from multi-view aerial imagery. *ISPRS J. Photogramm. & Rem. Sens.*, 58:83–98, 2003. 1

[9] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. PAMI*, 28(10):1568–1583, Oct. 2006. 4

[10] S. Kosov, F. Rottensteiner, C. Heipke, J. Leitloff, and S. Hinz. 3d classification of crossroads from multiple aerial images using markov random fields. In *Proc. 22nd ISPRS Congress*, pages XXXIX–B3:479–484, 2012. 4

[11] O. Kramer. Iterated local search with powell's method: a memetic algorithm for continuous global optimization. *Memetic Computing*, 2(1):69–83, 2010. 4

[12] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *Proc. ICCV*, pages 1284–1291, 2005. 1

[13] S. Kumar and M. Hebert. Discriminative Random Fields. *Int. J. Comput. Vis.*, 68(2):179–201, 2006. 1, 2, 3, 4

[14] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *Int. J. Comput. Vis.*, 77:259–289, 2008. 1

[15] S. Z. Li. *Markov Random Field Modeling in Image Analysis*. Springer, 3^{rd} edition, 2009. 1

[16] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(12):2368–2382, 2011. 2

[17] R. B. Myneni, F. G. Hall, P. J. Sellers, and A. Marshak. The interpretation of spectral vegetation indexes. *IEEE-TGARS*, 33:481–486, 1995. 5

[18] OpenCV. Machine Learning. http://docs.opencv.org/modules/ml/doc/ml.html, Apr. 2013. 4

[19] M. Prasad, A. Zisserman, A. W. Fitzgibbon, M. P. Kumar, and P. H. S. Torr. Learning class-specific edges for object detection and segmentation. In *Proc. Indian Conf. on Comp. Vision, Graphics and Image Proc.*, Dec 2006. 2

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

[21] 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. 5

[22] K. Schindler. An overview and comparison of smooth labeling methods for land-cover classification. *IEEE-TGARS*, 50:4534–4545, 2012. 1

[23] P. Schnitzspan, M. Fritz, S. Roth, and B. Schiele. Discriminative structure learning of hierarchical representations for object detection. In *Proc. CVPR*, pages 2238–2245, 2009. 1

[24] Z. Tu and X. Bai. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. *IEEE Trans. PAMI*, 32(10):1744–1757, Oct. 2010. 2

[25] 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. 23^{rd} ICML*, pages 969–976, 2006. 4

[26] U. Weidner and W. Förstner. Towards automatic building reconstruction from high resolution digital elevation models. *ISPRS J. Photogramm. & Rem. Sens.*, 50(4):38–49, 1995. 5

[27] J. Winn and J. Shotton. The layout consistent random field for recognizing and segmenting partially occluded objects. In *Proc. CVPR*, 2006. 1

[28] M. Yang and W. Förstner. Regionwise classification of building facade images. In *Photogrammetric Image Analysis*, volume 6952 of *LNCS*, pages 209–220. Springer, 2011. 5

[29] Z. Yin and R. T. Collins. Belief propagation in a 3d spatio-temporal mrf for moving object detection. In *Proc. CVPR*, 2007. 2