CONTEXT MODELS FOR CRF-BASED CLASSIFICATION OF MULTITEMPORAL REMOTE SENSING DATA

T. Hoberg*, F. Rottensteiner, C. Heipke

IPI, Institute of Photogrammetry and GeoInformation, Leibniz Universitaet Hannover, Germany
(hoberg, rottensteiner, heipke)@ipi.uni-hannover.de

Commission VII, WG VII/4

KEY WORDS: Contextual, Multiresolution, Multitemporal, Land Cover, Classification, Conditional Random Fields

ABSTRACT:

The increasing availability of multitemporal satellite remote sensing data offers new potential for land cover analysis. By combining data acquired at different epochs it is possible both to improve the classification accuracy and to analyse land cover changes at a high frequency. A simultaneous classification of images from different epochs that is also capable of detecting changes is achieved by a new classification technique based on Conditional Random Fields (CRF). CRF provide a probabilistic classification framework including local spatial and temporal context. Although context is known to improve image analysis results, so far only little research was carried out on how to model it. Taking into account context is the main benefit of CRF in comparison to many other classification methods. Context can be already considered by the choice of features and in the design of the interaction potentials that model the dependencies of interacting sites in the CRF. In this paper, these aspects are more thoroughly investigated. The impact of the applied features on the classification result as well as different models for the spatial interaction potentials are evaluated and compared to the purely label-based Markov Random Field model.

1. INTRODUCTION

An increasing number of optical high resolution (HR) remote sensing satellite systems have become available in the last decade. It should thus be possible to improve the classification accuracy and to analyse land cover changes more frequently than this is currently done based on a multitemporal analysis. However, the purchase of HR multitemporal data for these purposes is often not economically viable, especially for large areas. Data having medium resolution do not offer as much detail, but cover a larger area and may often be preferable from an economical point of view. Combining the advantages of both data types requires multiscale and multitemporal analysis.

Up to now most approaches for multitemporal land cover analysis do not make use of temporal dependencies, but derive results by some kind of difference measure between the monotemporal classification results of different epochs (i.e., different acquisition times) (Lu et al., 2004). If data from all epochs are available, it would seem to be advantageous to use the original observations, i.e. the image data, rather than derived data. This has for instance been done in (Feitosa et al., 2009), where a model of temporal dependencies based on Markov chains is applied. As in most techniques for multitemporal classification, each pixel is classified individually without considering spatial context, which leads to a salt-and-pepper-like appearance of the change detection results. Bruzzone et al. (2004) try to overcome this problem by using a cascade of three multitemporal classifiers, one of them considering the k-nearest neighbours of each pixel. A statistical model of spatial context in image classification is given by Markov Random Fields (MRF) (Geman & Geman, 1984), which have also been used for change detection (Melgani & Serpico, 2003), (Moser et al., 2009). In (Melgani & Serpico, 2003), the MRF framework is extended by a temporal energy term based on a transition probability matrix in order to improve the classification results for two consecutive images. Moser et al. (2009) applied the MRF framework to detect changes in optical satellite images based on multiscale features, but without determining the changed object classes.

Using MRF, the interaction between neighbouring image sites (pixels or segments) is restricted to the class labels, whereas the features extracted from different sites are assumed to be conditionally independent. This restriction is overcome by Conditional Random Fields (CRF; Kumar & Hebert, 2006). CRF provide a discriminative framework that can also model dependencies between features from different image sites and interactions between the labels and the features. In remote sensing CRF have been used for monotemporal classification, e.g. of settlement areas in HR optical satellite images (Zhong & Wang, 2007) or crop types and other land cover classes in Landsat data (Roscher et al., 2010). Multitemporal classification based on CRF for improving the overall classification accuracy as well as detecting changes has first been applied in (Hoberg et al., 2010). This method allows for temporal information passing using an extension of the CRF model.

Multiscale analysis is motivated by the fact that the appearance of objects in a scene is a function of the image resolution and because it is capable of providing a more global view on image content and image analysis algorithms (Kato et al., 1993), (Wilsky, 2002). The simplest way of considering multiple scales in classification is to derive the features at multiple scales, e.g. (Kumar & Hebert, 2006), which has been applied for change detection in (Moser et al., 2009). There have also been approaches to combine a multiscale analysis with CRF. In (Schnitzspan et al., 2008), a multiscale CRF is built on an

* Corresponding author.

image grid that in addition to the spatial neighbourhood relations also considers neighbours in scale based on a regular pyramid structure. Different classes are represented at different scale levels by a part-based object model: at finer resolutions, the classes to be discerned correspond to object parts, whereas at coarser resolutions, they correspond to compound objects. In (Yang et al., 2010) this method is extended to an irregular pyramid based on a multi-scale watershed segmentation of the original image.

A combination of multitemporal and multiscale analysis of remote sensing data using CRF is presented by Hoberg et al. (2011). A set of multispectral images of different resolution is classified simultaneously in order to increase the accuracy and reliability of the classification results and to detect land cover changes between the individual epochs. This approach allows to model dependencies between image regions at identical positions in the different epochs that may additionally be characterized by different scales and, hence, by different (though related) class structures.

Unfortunately in publications about CRF there is only little information about feature selection and the influence of different features on the classification result. Moreover in most cases only one model for the interaction potential is applied, without justification of the choice of the particular model. These issues are investigated in this paper. We compare different context models with different subsets of features that are extracted at different scales. First, to find the best subset of features depending on the maximum scale we apply a feature selection process. Next the best feature subset is selected for the association potential. Based on the selected association potential, we investigate three different context models for the spatial interaction potential, again comparing different feature subsets. Finally the results of these investigations are applied in a multitemporal CRF-based classification approach. Tests are performed using two set-ups, one of them using images having identical resolution and one with images of different resolution.

The remainder of this paper is structured as follows. In Section 2, the principles of CRF and the extensions for the classification of multitemporal and multiscale data are presented. Section 3 focuses on the description of the features and on feature selection. In Section 4, the test site is described. A qualitative analysis of the different ways of modelling context is given in Section 5, followed by quantitative results in Section 6. Conclusions and an outlook are given in Section 7.

2. MULTITEMPORAL AND MULTISCALE CRF

In many classification algorithms the decision for a class at a certain image site is just based on information derived at the regarded site (i.e., a pixel, a square block of pixels in a regular grid, or a segment). In fact, the class labels and also the data of spatially and temporally neighbouring sites are often similar or show characteristic patterns, which can be modelled using CRF. In monotemporal classification, we want to determine the vector of class labels **x** whose components x_i correspond to the classes of image sites $i \in S$ and S being the set of all sites for given image data **y** by maximizing the posterior probability $P(\mathbf{x} \mid \mathbf{y})$ (Kumar & Hebert, 2006):

$$P(\mathbf{x}|\mathbf{y}) = \frac{1}{Z} exp\left(\sum_{i \in S} A_i(x_i, \mathbf{y}) + \sum_{i \in S} \sum_{j \in N_i} I_{ij}\left(x_i, x_j, \mathbf{y}\right) \right) \tag{1}$$

In (1), N_i is the spatial neighbourhood of image site i (thus, j is a spatial neighbour of i), and Z is a normalization constant called the *partition function*. The *association potential* A_i links the class label x_i of image site i to the data **y**, whereas the term I_{ij}, called *interaction potential*, models the dependencies between the labels x_i and x_j of neighbouring sites i and j and the data **y**. The model is very general in terms of the definition of the functional model for both A_i and I_{ij}.

In the multitemporal case, we have M co-registered images. In addition to the interactions of spatial neighbours, the temporal neighbourhood is taken into account. Each node is only linked to its direct temporal neighbours at its spatial position (Figure 1). The components of the image data vector **y** are site-wise data vectors $\mathbf{y}_i^t$, with $i \in S$ and S being the set of sites of *all* images (i.e., i does not refer to a particular spatial position, but it refers to one spatial position in one of the images). The index t indicates the membership of image site i to the related epoch $t \in T$ and $T = \{1,\dots\ M\}$. The components of **x** are the class labels of the image sites i, x_i^t, also with epoch index $t \in T$. For each image site we want to determine the class x_i^t from a set of pre-defined classes. The class structure and thus the number of classes are dependent on t. In order to model the mutual dependency of the class labels at an image site at different epochs, the model for $P(\mathbf{x} \mid \mathbf{y})$ in (1) has to be extended:

$$P(\mathbf{x}|\mathbf{y}) = \frac{1}{Z} exp\left[\sum_{i \in S} A(x_i^t, \mathbf{y}^t) + \sum_{i \in S} \sum_{j \in N_i} IS\left(x_i^t, x_j^t, \mathbf{y}^t\right) + \right.$$
$$\left. + \sum_{i \in S} \sum_{k \in E_t} \sum_{l \in L_i^k} IT^{tk}\left(x_i^t, x_l^k, \mathbf{y}^t, \mathbf{y}^k\right) \right] \tag{2}$$

As the different functional models for the potential functions A, IS, and IT^{tk} are shift-invariant, the subscripts of the potential functions in (1) have been omitted in (2). In (2), A is the association potential, IS the *spatial interaction potential* that corresponds to the interaction potential I_{ij} in (1), and IT^{tk} the *temporal interaction potential*. In IT^{tk}, $\mathbf{y}^t$ and $\mathbf{y}^k$ are the images observed at epochs t and k, respectively. E_t is the set of epochs in the temporal neighbourhood of the epoch to which image site i belongs, thus k is the time index of an epoch in temporal neighbourhood of t. The set of image sites at epoch $k \in E_t$ that are temporal neighbours of the image site i is denoted by L_i^k, thus $l \in L_i^k$ is an image site that is a temporal neighbour of i in epoch k. The temporal interaction potential models the dependency between the class labels and the observed data at consecutive epochs. The image sites are chosen to be individual pixels and thus are arranged in a regular grid for each image. Figure 1 shows the spatial and temporal neighbourhood for images having identical or different resolutions.

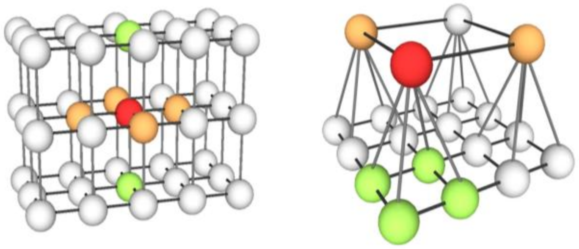


Figure 1. Multitemporal graph structure. Left: images having the same resolution. Right: images having different resolutions. Red nodes: processed primitives; orange / green nodes: spatial / temporal neighbours.

2.1 Association potential

The association potential $A(x_i^t, \mathbf{y}^t)$ in (2) is related to the probability of label x_i^t taking a value c given the image $\mathbf{y}^t$ at epoch t by $A(x_i^t, \mathbf{y}^t) = \log\{P[x_i^t = c \mid \mathbf{f}_i^t(\mathbf{y}^t)]\}$. The image data are represented by site-wise feature vectors $\mathbf{f}_i^t(\mathbf{y}^t)$ that may depend on the entire image at epoch t, e.g. by using features at different scales (Kumar & Hebert, 2006). We use a multivariate Gaussian model for $P[x_i^t = c \mid \mathbf{f}_i^t(\mathbf{y}^t)]$ (Bishop, 2006):

$$P\left[x_i^t = c \mid \mathbf{f}_i^t\left(\mathbf{y}_i^t\right)\right] = \\ = \frac{1}{\sqrt{(2\pi)^n \, det\left(\boldsymbol{\Sigma}_{fc}^t\right)}} e^{-\frac{1}{2}\left[\mathbf{f}_i^t\left(\mathbf{y}^t\right) - \mathbf{E}_{fc}^t\right]^T \cdot \left(\boldsymbol{\Sigma}_{fc}^t\right)^{-1} \cdot \left[\mathbf{f}_i^t\left(\mathbf{y}^t\right) - \mathbf{E}_{fc}^t\right]} \qquad (3)$$

In (3), $\mathbf{E}_{fc}^t$ and $\boldsymbol{\Sigma}_{fc}^t$ are the mean and co-variance matrix of the features of class c, respectively. It is important to note that both the definition of the features and the dimension of the feature vectors $\mathbf{f}_i^t(\mathbf{y}^t)$ may vary from image to image, because the definition of appropriate and expressive features depends on the image resolution and also on the spectral information contained in the images (see also Section 3).

2.2 Spatial interaction potential

The spatial interaction potential $IS(x_i^t, x_j^t, \mathbf{y}^t)$ in (2) is a measure for the influence of the data $\mathbf{y}^t$ and the neighbouring labels x_j^t on the class x_i^t of image site i at epoch t. In this potential, the data are represented by site-wise vectors of *interaction features* $\boldsymbol{\mu}_{ij}^t(\mathbf{y}^t)$. In this work we compare three different models for the spatial interaction potential. The first model only depends on the labels. It is commonly used with MRF and has a smoothing effect on the labels:

$$IS_1\left(x_i^t, x_j^t, \mathbf{y}^t\right) = \begin{cases} \beta & if \quad x_i^t = x_j^t \\ 0 & if \quad x_i^t \neq x_j^t \end{cases} \qquad (4)$$

The second model is based on (Shotton et al., 2007):

$$IS_2\left(x_i^t, x_j^t, \mathbf{y}^t\right) = \begin{cases} \beta \cdot \exp\left[-\frac{\left\|\boldsymbol{\mu}_{ij}^t\left(\mathbf{y}^t\right)\right\|^2}{R}\right] & if \quad x_i^t = x_j^t \\ 0 & if \quad x_i^t \neq x_j^t \end{cases} \qquad (5)$$

The third model is used by Hoberg et al. (2010):

$$IS_3\left(x_i^t, x_j^t, \mathbf{y}^t\right) = \begin{cases} \beta \cdot \exp\left[-\frac{\left\|\boldsymbol{\mu}_{ij}^t\left(\mathbf{y}^t\right)\right\|^2}{R}\right] & if \quad x_i^t = x_j^t \\ \beta \cdot \left\{1 - \exp\left[-\frac{\left\|\boldsymbol{\mu}_{ij}^t\left(\mathbf{y}^t\right)\right\|^2}{R}\right]\right\} & if \quad x_i^t \neq x_j^t \end{cases} \qquad (6)$$

In (5) and (6), $\|\boldsymbol{\mu}_{ij}^t(\mathbf{y}^t)\|$ denotes the Euclidean norm of $\boldsymbol{\mu}_{ij}^t(\mathbf{y}^t)$ and β is a weighting factor for the influence of the spatial interaction potential in the classification process. We use the component-wise differences of the feature vectors $\mathbf{h}_i^t(\mathbf{y}^t)$ for the interaction features $\boldsymbol{\mu}_{ij}^t(\mathbf{y}^t)$, i.e. $\boldsymbol{\mu}_{ij}^t(\mathbf{y}^t) = [\mu_{ij1}^t, \dots \mu_{ijR}^t]^T$, where R is the dimension of the vectors $\mathbf{h}_i^t(\mathbf{y}^t)$ that may vary with t.

Note that the feature vector $\mathbf{h}_i^t(\mathbf{y}^t)$ used for the interaction potential might differ from the feature vector $\mathbf{f}_i^t(\mathbf{y}^t)$ used for the association potential (Kumar & Hebert, 2006). Denoting the m^{th} component of $\mathbf{h}_i^t(\mathbf{y}^t)$ by $h_{im}^t(\mathbf{y}^t)$, the m^{th} component of $\boldsymbol{\mu}_{ij}^t(\mathbf{y}^t)$ is $\mu_{ijm}^t = |h_{im}^t(\mathbf{y}^t) - h_{jm}^t(\mathbf{y}^t)|$. Division by the number of features R in (5) and (6) guarantees an identical influence of the spatial interaction potentials for all images. In IS_2 a potential of zero is assigned in the case of two sites have different labels. Differing labels at neighbouring sites are penalized unless the features of the sites are also very different. IS_3 penalizes both local changes of the class labels if the data are similar and also identical class labels if the features are different.

2.3 Temporal interaction potential

The temporal interaction potential $IT^{tk}(x_i^t, x_l^k, \mathbf{y}^t, \mathbf{y}^k)$ models the dependencies between the data $\mathbf{y}$ and the labels x_i^t and x_l^k of site i at epoch t and site l of epoch k. In principle, IT^{tk} could be modelled similarly to IS by penalizing temporal change of labels unless it is indicated by differences in the data. However, a more sophisticated functional model would be required to compensate for atmospheric effects and varying illumination conditions, different resolutions, and seasonal effects of the vegetation. We use a simple model for the temporal interaction potential that neglects the dependency of IT^{tk} of the data:

$$IT^{tk}\left(x_i^t, x_l^k, \mathbf{y}^t, \mathbf{y}^k\right) = IT^{tk}\left(x_i^t, x_l^k\right) = \frac{\gamma \cdot \mathbf{TM}^{s(t)s(k)}\left(x_i^t, x_l^k\right)}{Q_i^k} \qquad (7)$$

In (7), γ is a weight factor. $\mathbf{TM}^{s(t)s(k)}$ is a temporal transition matrix similar to the transition probability matrix in (Bruzzone et al., 2004). The elements of $\mathbf{TM}^{s(t)s(k)}(x_i^t, x_l^k)$ can be seen as conditional probabilities $P(x_i^t = c^t \mid x_l^k = c^k)$ of an image site i belonging to class c^t at epoch t if the image site l that occupies the same spatial position as i in epoch k belongs to class c^k in that epoch. Q_i^k is the number of elements in L_i^k and acts as a normalization factor ensuring an identical influence of the sum of all temporal interaction potentials in any epoch, no matter how many temporal neighbours exist. The scales $s(t)$ and $s(k)$ of the data at epochs t and k may differ; there is one matrix $\mathbf{TM}^{s(t)s(k)}$ for each combination of scales available in the data. For further information we refer to (Hoberg et al., 2011).

2.4 Training and Inference

Exact training and inference is computationally intractable for CRF (Kumar & Hebert, 2006). In our application, we only train the parameters of the association potentials, i.e. the mean $\mathbf{E}_{fc}^t$ and the co-variance matrix $\boldsymbol{\Sigma}_{fc}^t$ of the features of each class c. They are determined from the features $\mathbf{f}_i^t(\mathbf{y}^t)$ in training sites individually for each epoch t and each class c. The other model parameters, i.e. the weighting factors β and γ of the spatial and temporal interaction potentials and the elements of the transition matrices $\mathbf{TM}^{s(t)s(k)}$, were found empirically. For inference, we use Loopy Belief Propagation (LBP) (Nocedal & Wright, 2006), 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).

3. FEATURES AND FEATURE SELECTION

In order to apply the CRF framework, the site-wise feature vectors $\mathbf{f}_i^t(\mathbf{y}^t)$ for the association and $\mathbf{h}_i^t(\mathbf{y}^t)$ for the spatial interaction potentials for each epoch t must be defined. Both must consist of appropriate features that can help to discriminate the individual classes. In our application, we used several groups of features, namely colour-based, textural and structural features. All features are computed at five different scales λ_{d}, with *d* indicating the scale. Whereas in λ_1 only individual pixels are taken into account, in λ_2 to λ_5 the features are extracted in a square window of size 3, 5, 9, and 13 pixels, respectively, centred at the centre of image site *i*. Hence we do not only consider information derived at site *i* for the site-wise feature vectors $\mathbf{f}_i(\mathbf{y})$ and $\mathbf{h}_i(\mathbf{y})$, but we also model dependencies between the image information of neighbouring sites.

The colour-based features are directly derived from the pixel values of the spectral channels, four in our case. We used the mean and variance of the red ($E^{\mathrm{d}}_{\mathrm{r}}$, $V^{\mathrm{d}}_{\mathrm{r}}$), green ($E^{\mathrm{d}}_{\mathrm{g}}$, $V^{\mathrm{d}}_{\mathrm{g}}$), blue ($E^{\mathrm{d}}_{\mathrm{b}}$, $V^{\mathrm{d}}_{\mathrm{b}}$), and near infrared ($E^{\mathrm{d}}_{\mathrm{nir}}$, $V^{\mathrm{d}}_{\mathrm{nir}}$) channel, the variance of the hue ($V^{\mathrm{d}}_{\mathrm{hue}}$), and the mean of the difference of red and green ($E^{\mathrm{d}}_{\mathrm{r\text{-}g}}$), near infrared and red ($E^{\mathrm{d}}_{\mathrm{nir\text{-}r}}$), and near infrared and green ($E^{\mathrm{d}}_{\mathrm{nir\text{-}g}}$). Moreover the mean and variance of the normalized difference vegetation index ($E^{\mathrm{d}}_{\mathrm{ndvi}}$, $V^{\mathrm{d}}_{\mathrm{ndvi}}$) and the relational vegetation index ($E^{\mathrm{d}}_{\mathrm{rvi}}$, $V^{\mathrm{d}}_{\mathrm{rvi}}$) were computed.

The textural features consist of contrast (con^{d}), correlation (cor^{d}), energy (ene^{d}), homogeneity (hom^{d}), and entropy (ent^{d}) as defined by Haralick et al. (1973). They are all derived from the gray-level co-occurence matrix that represents the distribution of co-occurring values at a given offset (1 in our case).

The structural features are derived from a weighted histogram of oriented gradients (HOG) (Dalal & Triggs, 2005). Each histogram has 30 bins, so that each bin corresponds to an orientation interval of 6° width. Each bin contains the sum of the magnitudes of all gradients having an orientation that is within the interval corresponding to the bin. Summing over the magnitudes and not just counting the numbers of gradients falling into each bin is done to take into account the impact of strong magnitudes. From the histogram we derive five features: The mean of all gradient magnitudes ($E^{\mathrm{d}}_{\mathrm{grad}}$) the variance of the histogram entries ($V^{\mathrm{d}}_{\mathrm{grad}}$), the number of bins with magnitudes above the mean magnitude (num^{d}), the value of the maximum histogram entry (mag^{d}) and the angle between the first two maxima (ang^{d}). All the features are normalised so that the values are in the interval [0, 1].

We define the feature vectors corresponding to a maximum scale λ_{max} to consist not only of the features extracted at λ_{max}, but also of all features of lower scales. Hence, for instance the feature vector corresponding to $\lambda_{\mathrm{max}}=\lambda_5$ contains 113 elements, nine of them extracted at λ_1 and 26 features extracted at each additional scale. Using the large number of features just described makes the classification quite time consuming for two reasons: All the features have to be extracted and all have to be considered for determining the potentials. As many of the features are highly correlated or may only marginally support the classification, we apply a feature selection procedure to find out which features are relevant for our aims and to reduce the number of features accordingly. For that purpose we use the correlation-based feature selection approach by Hall (1999). First, the single feature which best classifies the data set is determined. After that, other features are chosen according to criteria that ensure the selection of a subset that contains features that are highly correlated with the classes, yet uncorrelated with each other.

4. TEST SITE AND DATA

Our test area is situated near Herne, Germany, and covers an area of 8.6 x 5.9 km². We used multispectral Ikonos data with 4 m ground sampling distance (GSD) acquired in 2005 and 2007, and Landsat data of 30 m GSD acquired in 2010. All images were recorded in summer. The area was split into 54 sections, which were processed separately. Seven sections served as training data, the rest as test sites. Ground truth was obtained by manually labelling the images at pixel level. The classes to be distinguished with Ikonos imagery are residential areas (*res*), industrial areas (*ind*), forests (*for*), and cropland (*crp*). Because there is no clear distinction of the classes *res* and *ind i*n the medium resolution Landsat imagery they are fused to a new class built-up areas (*bui*) in that resolution.

5. FEATURE AND MODEL SELECTION

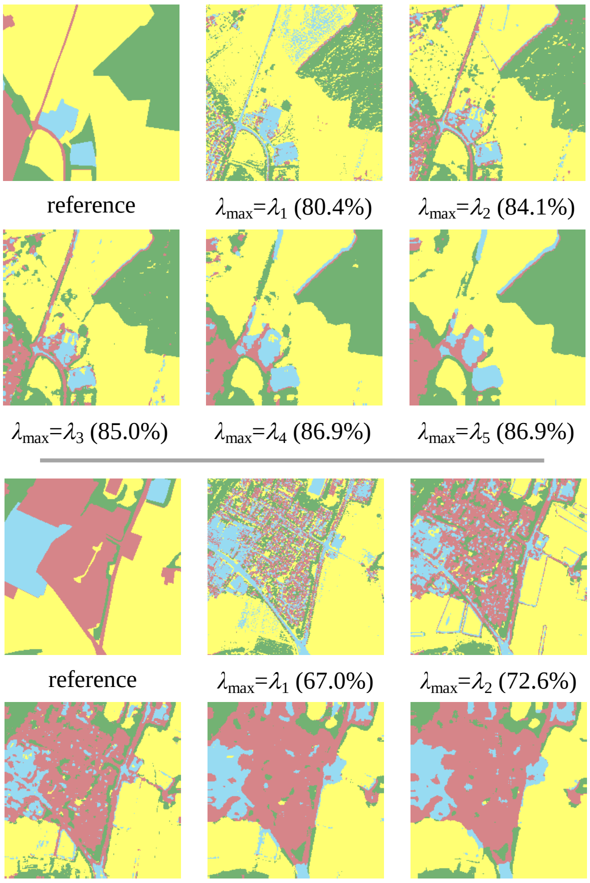
In this section the impact of using features at different scales and of different context models on the classification result is investigated. We try to find a suitable subset of features for each maximum scale λ_{max} and then analyse the results to find the best maximum scale and, thus, the optimal feature subset for the association potential. Then we compare different context models for the spatial interaction potential, using the optimal feature subsets for each maximum scale λ_{max}.

To investigate how many features should be used for our CRF-classification we applied a standard maximum likelihood (ML) classification in subsets with features derived at an increasing number of scales up to a maximum scale λ_{max}, ordering the features according to the results of the feature selection process described above. The ML-classification was chosen because its model is also used for the association potential. For all values of λ_{max} we found that using the six best features was sufficient. Additional features did not further increase the classification accuracy. Hence each of the feature vectors $\mathbf{f}_i^t(\mathbf{y}^t)$ and $\mathbf{h}_i^t(\mathbf{y}^t)$ was reduced to just six features depending on λ_{max}:

$\lambda_{\mathrm{max}}=\lambda_1$: E^1_{r}, E^1_{g}, E^1_{b}, E^1_{nir}, E^1_{ndvi}, E^1_{rvi}
$\lambda_{\mathrm{max}}=\lambda_2$: E^2_{nir}, V^2_{nir}, V^2_{hue}, $E^2_{\mathrm{nir\text{-}r}}$, V^2_{ndvi}, E^2_{grad}
$\lambda_{\mathrm{max}}=\lambda_3$: E^3_{nir}, V^3_{nir}, V^3_{hue}, $E^3_{\mathrm{nir\text{-}r}}$, E^3_{grad}, ent^3
$\lambda_{\mathrm{max}}=\lambda_4$: E^4_{g}, E^4_{nir}, V^4_{hue}, E^4_{grad}, ent^4, V^3_{hue}
$\lambda_{\mathrm{max}}=\lambda_5$: E^5_{nir}, V^5_{hue}, E^5_{grad}, hom^5, E^4_{g}, ent^4

It is obvious that in each subset the features extracted in the largest scale are dominant. The impact of using features extracted at different maximum scales λ_{max} on the association potential was evaluated by comparing the results of ML classification obtained for the selected subsets for each value of λ_{max}. Figure 2 shows exemplary results for two of the sections using Ikonos imagery; the highest overall accuracy is achieved with $\lambda_{\mathrm{max}}=4$. Nevertheless, by visual interpretation most users would consider the result of $\lambda_{\mathrm{max}}=3$ to be best, because many finer structures (for instance the road in the upper example of figure 2) are much better preserved. Because information that is lost at this stage cannot be re-introduced in further processing steps, we decided to apply the feature vector $\mathbf{f}_i^t(\mathbf{y}^t)$ for $\lambda_{\mathrm{max}}=3$ for the association potential of our further computations.

The three context models for the spatial interaction potential (Section 2.2) are evaluated by a monotemporal classification on Ikonos imagery. For the two data-dependent models we used the feature vectors selected for the association potentials in the maximum scales $\lambda_{\mathrm{max}}=\lambda_2$, λ_3 and λ_4 (see above) for $\mathbf{h}_i^t(\mathbf{y}^t)$. In general, the purely label-based model IS_1 results in strong smoothing, while the data-dependent models preserve finer structures better, e.g. the road passing through cropland in Figure 3. However, this does not necessarily lead to a higher overall accuracy. In all scales IS_2 performs slightly better than IS_3, which favours additional class transitions if the features at neighbouring sites are different. The maximum scale of the



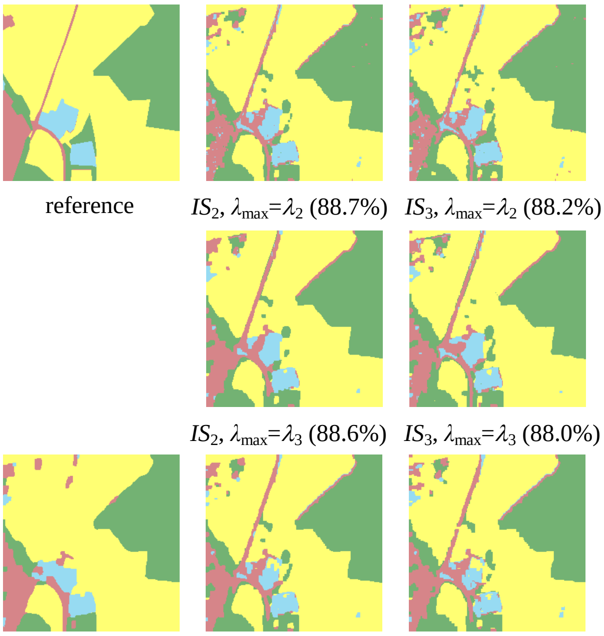
reference — $\lambda_{max}=\lambda_1$ (80.4%) — $\lambda_{max}=\lambda_2$ (84.1%)

$\lambda_{max}=\lambda_3$ (85.0%) — $\lambda_{max}=\lambda_4$ (86.9%) — $\lambda_{max}=\lambda_5$ (86.9%)

reference — $\lambda_{max}=\lambda_1$ (67.0%) — $\lambda_{max}=\lambda_2$ (72.6%)

$\lambda_{max}=\lambda_3$ (77.4%) — $\lambda_{max}=\lambda_4$ (81.8%) — $\lambda_{max}=\lambda_5$ (81.4%)

Figure 2. Overall accuracy of ML-classification in dependence on applied maximum scale λ_{max} for feature extraction. Red: *res*; blue: *ind*; green: *for*; yellow: *crp*.



reference — IS_2, $\lambda_{max}=\lambda_2$ (88.7%) — IS_3, $\lambda_{max}=\lambda_2$ (88.2%)

IS_2, $\lambda_{max}=\lambda_3$ (88.6%) — IS_3, $\lambda_{max}=\lambda_3$ (88.0%)

IS_1 (90.7%) — IS_2, $\lambda_{max}=\lambda_4$ (88.4%) — IS_3, $\lambda_{max}=\lambda_4$ (87.3%)

Figure 3. Overall accuracy of CRF-classification with different context models and varying scale λ_{max} for the spatial interaction potential.

features in $\mathbf{h}_i^t(\mathbf{y}^t)$ only has a minor effect on the results. Using $\lambda_{max}=\lambda_2$, some salt-and-pepper effects remain, whereas the other scales lead to stronger smoothing. Overall, using IS_2 with $\mathbf{h}_i^t(\mathbf{y}^t)$ from $\lambda_{max}=\lambda_3$ delivers the best trade-off of overall accuracy and preservation of details assessed by visual impression, which is why this combination is applied in our experiments (cf. Section 6). Hence, in these experiments, $\mathbf{f}_i^t(\mathbf{y}^t)$ and $\mathbf{h}_i^t(\mathbf{y}^t)$ are identical.

6. QUANTITATIVE EVALUTION

We tested our multitemporal approach for two data set-ups: Set-up I has only one scale and consists of two Ikonos images. In the multiscale set-up II we combined one Ikonos and one Landsat scene. For the Ikonos scenes we used the features as defined in Section 5, for the Landsat scene they were extracted only in the original resolution.

The temporal transition matrix TM between Ikonos and Landsat used in our experiments is shown in Table 1. A similar matrix was defined for the transition between the two HR images in set-up I. The choice of these values is dependent on the land cover structure and the assumed changes. We assume that it is most likely to have no changes in any region. Nevertheless each class transition might happen, but with different probability.

	$x_i^{t+1}=bui$	$x_i^{t+1}=for$	$x_i^{t+1}=crp$
$x_i^t=res$	1	0.05	0.05
$x_i^t=ind$	1	0.05	0.05
$x_i^t=for$	0.2	1	0.1
$x_i^t=crp$	0.2	0.1	1

Table 1: Temporal transition matrix; *t* corresponds to the Ikonos image, *t+1* corresponds to the Landsat image.

For both set-ups, we compared our method (scenario CRF_{multi}) to a Maximum Likelihood classification using the Gaussian model in (3) (scenario *ML*) and to a multitemporal MRF-classification (scenario *MRF*) using the same graph structure as for our CRF_{multi} approach, but applying IS_1. For these three scenarios, the overall classification accuracy and the kappa coefficients are compared for all epochs in Table 2. In both set-ups we achieved an overall accuracy of over 80% for all images with CRF and MRF, which is an increase of about 8% compared to the monotemporal ML-classification for the Ikonos images and even 15% for the Landsat scene (Figure 4). The impact of the multi-temporal approach is highlighted by the overall accuracy achieved in the scenario CRF_{multi} in comparison with the results of a monotemporal CRF classification (CRF_{mono}) for the Landsat scene. Using CRF_{mono} only leads to an accuracy of 72%, which is 12% lower than with $CRF_{multi.}$ The higher information content of the HR images clearly propagates to the medium resolution scene and yields a significant increase. Nevertheless the accuracy of the HR image also increases. There was hardly any difference between scenarios *MRF* and CRF_{multi}. Only in a few regions finer structures are better preserved by the CRF-approach.

S/E	*ML*	CRF_{multi}	*MRF*
I / t_1	73.7% / 0.57	80.8% / 0.72	81.3% / 0.73
I / t_2	72.8% / 0.61	81.1% / 0.72	80.6% / 0.72
II / t_1	73.7% / 0.57	81.8% / 0.73	81.9% / 0.73
II / t_2	69.6% / 0.53	84.3% / 0.74	84.1% / 0.74

Table 2: Overall classification accuracy / kappa coefficients; S/E: Set-up/epoch; set up I: t_1: Ikonos, 2005; t_2: Ikonos, 2007; set up II. t_1: Ikonos, 2005; t_2: Landsat, 2010.

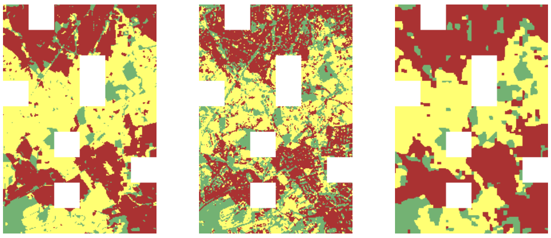


Figure 4. left: Reference of full test scene (Landsat); middle: Results of *ML*; right: Results of CRF_{multi}: yellow: *crp*; green: *for*; dark red: *bui;* white patches: training areas.

7. CONCLUSION

In this work, we evaluated two possibilities for modelling spatial context within a CRF-framework. First the impact of using features extracted at different scales for the association potential was investigated. Neighbourhood dependencies are already taken into account in this step. Large scales result in a severe smoothing, while tiny structures are lost. Furthermore, different context models for the spatial interaction potential were compared. It could be shown that data-dependent models as used for CRF have a better ability to preserve fine structures. The results of these investigations were applied in a CRF-based approach for multitemporal and multiscale image classification. Besides incorporating spatial context, this method uses a model of temporal context by introducing a temporal interaction potential. The overall classification accuracy of all images was improved by at least 8%. The effect of the multitemporal interaction was highlighted in a set-up of an Ikonos and a Landsat image. The overall accuracy of CRF_{multi} in comparison to CRF_{mono} for the Landsat scene increased at about 12%.

Further research will concentrate on an improvement of the model for the temporal interaction potential, which was kept quite simple in this work. Moreover, tests on different data sets with a focus on the ability of the method for change detection will be carried out.

ACKNOWLEDGEMENT

The research was funded by the German Science Foundation (DFG) under grant HE 1822/22-1. Our CRF-classification is based on the UGM toolbox by Mark Schmidt:
http://people.cs.ubc.ca/~schmidtm/Software/UGM.html.

REFERENCES

Bishop, C. M., 2006. *Pattern recognition and machine learning*. 1st edition, Springer New York.

Bruzzone, L., Cossu, R., Vernazza, G., 2004. Detection of land-cover transitions by combining multidate classifiers. *Pattern Recognition Letters,* 25(13): 1491-1500.

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

Feitosa, R. Q., Costa, G. A. O. P., Mota, G. L. A., Pakzad, K., Costa, M. C. O., 2009. Cascade multitemporal classification based on fuzzy Markov chains. *ISPRS J. Photogrammetry Remote Sens.* 64(2): 159-170.

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.

Hall, M. A. (1999). *Correlation-based Feature Subset Selection for Machine Learning*, PhD dissertation, Department of Computer Science, University of Waikato.

Haralick, R.M., Shanmugam, K. und Dinstein, I. (1973). Texture features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(6): 610–622.

Hoberg, T., Rottensteiner, F. and Heipke, C. 2010. Classification of Multitemporal Remote Sensing Data Using Conditional Random Fields. *6th IAPR TC 7 Workshop Pattern Recognition in Remote Sens.*: 4p.

Hoberg, T., Rottensteiner, F. und Heipke, C. (2011). Classification of multitemporal remote sensing data of different resolution using conditional random fields. *1st IEEE/ISPRS Workshop on Computer Vision for Remote Sensing of the Environment*, IEEE ICCV Workshops, Barcelona, 235-242.

Kato, Z., Berthod, M., Zerubia, J., 1993. Multiscale Markov random field models for parallel image classification. *Proc. Fourth Int. Conference on Computer Vision*: 253-257.

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

Lu, D., Mausel, P., Brondizio, E., Moran, E., 2004. Change detection techniques. *Int. J. Remote Sensing,* 25(12): 2365-2401.

Melgani, F. and Serpico, S. B., 2003. A Markov Random Field approach to spatio-temporal contextual image classification. *IEEE-TGARS,* 41(11): 2478-2487.

Moser, G., Angiati, E., Serpico, S. B., 2009. A contextual multiscale unsupervised method for change detection with multitemporal remote-sensing images. *Proc. 9th Conf. Intelligent Systems Design & Applications*: 572-577.

Nocedal, J. and Wright, S. J., 2006. *Numerical Optimization.* 2nd edition, Springer New York.

Roscher, R., Waske, B., Förstner, W., 2010. Kernel discriminative random fields for land cover classification. *6th IAPR TC 7 Workshop Pattern Recognition in Remote Sens.*: 5p.

Schnitzspan, P., Fritz, M., Schiele, B., 2008. Hierarchical support vector random fields: joint training to combine local and global features. *Proc. ECCV II*: 527-540.

Shotton, J., Winn, J., Rother, C., Criminisi, A., 2007. TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context. *International Journal of Computer Vision*, 81(1): 2-23.

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.

Wilsky, A. S., 2002. Multiresolution Markov models for signal and image processing. *Proc. IEEE*, 90(8): 1396-1458.

Yang, M. Y., Förstner, W., Drauschke, M., 2010. Hierarchical conditional random field for multi-class image classification. *Proc. Int'l. Conf. Computer Vision Theory and Applications (VISAPP)*: 464-469.

Zhong, P. and Wang, R., 2007. A multiple conditional random fields ensemble model for urban area detection in remote sensing optical images. *IEEE-TGARS,* 45(12): 3978-3988.