Classification of Multitemporal Remote Sensing Data of Different Resolution using Conditional Random Fields

Thorsten Hoberg, Franz Rottensteiner and Christian Heipke
Institute of Photogrammetry and GeoInformation
Leibniz Universität Hannover, Nienburger Str. 1, 30167 Hannover, Germany
{hoberg, rottensteiner, heipke}@ipi.uni-hannover.de

Abstract

The increasing availability of multitemporal optical remote sensing data offers new potentials for land cover analysis. We present a novel approach for enhancing the classification accuracy of medium resolution data by combining them with high resolution data of an earlier acquisition time, thus saving data acquisition costs. Our approach uses Conditional Random Fields to model both spatial and temporal dependencies. Temporal context is considered by a novel extension of the CRF concept by an additional temporal interaction potential, which can model dependencies between identical regions in images of different acquisition times and scales. The model also considers different levels of abstraction in the class structures at different scales. The approach is tested with two set-ups of Ikonos, RapidEye, and Landsat imagery.

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 analyze land cover changes at a higher frequency 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. It is thus the goal of this paper to present a method capable of combining HR images with data of lower resolution for increasing the classification accuracy and detecting land cover changes. This is achieved by a combined classification using medium resolution data (e.g. 30 m Ground Sampling Distance - GSD), available for one or more points in time, together with data of high resolution (GSD ≤ 5 m) acquired at the beginning of the observation period. In this way it should be possible to benefit from the higher information content of HR imagery while performing change detection in data of lower resolution.

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) [1]. 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 [2], where a model of temporal dependencies based on Markov chains is applied for detecting land cover changes in Landsat images (30 m GSD). As in most techniques for multitemporal classification found in the literature, 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. [3] try to overcome this problem by using a cascade of three multitemporal classifiers, one of them considering the k-nearest neighbors of each pixel. A systematic statistical model of spatial context in image classification is given by Markov Random Fields (MRF) [4], which have also been used for change detection [5], [6]. In [5], 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. In [6], the MRF framework is applied to detect changes in optical satellite images based on multiscale features. The change detection problem is formulated as a hypothesis test, leading to a binary map of changes between the two given images. The method works in an unsupervised way, but it does not distinguish the changed object classes.

Using MRF, the interaction between neighboring 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), originally introduced for classifying one-dimensional data in [7] and later also applied to the classification of terrestrial images [8], [9]. 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 the classification of settlement areas in HR optical satellite images [10], [11], for

generating a digital terrain model from LiDAR data [12], and for classifying crop types and other land cover classes in Landsat data [13]. All these approaches use monotemporal data. The image sites that are to be classified are either the image pixels or small squares of image pixels. In contrast, [14] use an irregular graph derived from mean-shift segmentation for CRF-based building detection from aerial images and airborne InSAR data. Such a definition would not seem to be appropriate for change detection, because image segmentation cannot be expected to yield coincident segment boundaries (and, thus, 1:*N* or *N*:1 relations between segments) in images from different epochs. Multitemporal classification based on CRF for improving the overall classification accuracy as well as detecting changes has only been applied in [15]. Unlike most state-of-the-art methods for multitemporal classification, the method in [15] allows for temporal information passing in both directions, using an expansion of the CRF model by temporal interactions. However, all the input images are assumed to have approximately the same geometrical resolution.

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/or image analysis algorithms [16], [17]. The simplest way of considering multiple scales in classification is to derive the features at multiple scales, e.g. [9], which has been applied for change detection in [6]. Multiscale random fields were defined in [18] as a computationally more tractable alternative to MRF. The image is represented by a pyramidal structure based on an octree whose leaves are the individual pixels. Each node of the octree is modeled to be only dependent on its predecessor on the coarser level, but not on its spatial neighbors, resulting in a series of Markov chains in scale. The smoothing effect of MRF is achieved by octree nodes at coarser scales propagating their information to the leaves. A similar approach was used for image classification in [19], where it was also noted that a multiscale approach requires a redefinition of the class structure at the coarser scales because some classes might be extinct. In [20], such a multiscale representation is combined with spatial interactions at each scale of the representation, but no results are presented.

There have also been approaches to combine a multiscale analysis with CRF. In [21], a multiscale CRF is built on an image grid that in addition to the spatial neighborhood relations also considers neighbors in scale based on a regular pyramid structure. This paper also considers the fact that different classes are represented at different scale levels by defining a part-based object model: at finer resolutions, the classes to be discerned correspond to object parts, whereas at the coarser resolutions, they correspond to compound objects. This method is applied to detect objects such as motorbikes in terrestrial images. In [22], this method is expanded to an irregular pyramid based on a multi-scale watershed segmentation of the original image. The class structure seems to be constant over scale in this model.

It has to be noted that except for [6] most of the multiscale methods are based on monotemporal images. A few of them consider the fact that the class structure changes with scale [19], [21]. Nearly all of them require the images at full resolution to have the same scale, exceptions for the monotemporal case being given in [17]. To our knowledge, there is no method for multitemporal classification and change detection that uses images of different resolutions at the individual epochs and that can handle class transitions between the different scales.

In this paper, a novel approach for the classification of multitemporal remote sensing data is presented that is designed to achieve this very goal. A set of multispectral images of different resolution is classified simultaneously in order to increase of the accuracy and reliability of the classification results and to detect land cover changes between the individual epochs. No existing land cover map is required. This method is based on an extension of the CRF concept by an additional temporal interaction potential in a similar way as [15]. Using this potential it is possible 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.

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 multiscalar data are presented. Section 3 focuses on the description of the features and the class structure. In Section 4, the method is evaluated based on Ikonos, RapidEye, and Landsat data. Conclusions and an outlook are given in Section 5.

2. Conditional Random Fields

In many classification algorithms the decision for a class at a certain image site is just based on information derived at the regarded site, where a site might be 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 neighboring sites are often similar or show characteristic patterns, which can be modeled using CRF. In monotemporal classification, we want to determine the vector of class labels $\mathbf{x}$ whose component x_i corresponds to the class of image site $i \in S$ for given image data $\mathbf{y}$ by maximizing the posterior probability $P(\mathbf{x} \mid \mathbf{y})$ [9]:

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

In (1), N_i is the spatial neighborhood of image site i (thus, j is a spatial neighbor to 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 $\mathbf{y}$, whereas the term I_{ij}, called *interaction potential*, models the dependencies between the labels x_i and x_j of neighboring sites i and j and the data $\mathbf{y}$. The model is very general in terms of the definition of the functional model for both A_i and I_{ij}. For instance, [9] use generalized linear models for both potentials.

2.1. Multitemporal Approach

In the multitemporal case, we have M co-registered images. In addition to the interactions of spatial neighbors, the temporal neighborhood is taken into account. The left part of Figure 1 shows the multitemporal graph structure for images having the same scale. Each node is only linked to its direct temporal neighbors at its spatial position.

The components of the image data vector $\mathbf{y}$ 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 $\mathbf{x}$ are the class labels of image site 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 position at different epochs, the model for $P(\mathbf{x} \mid \mathbf{y})$ in (1) has to be expanded:

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

As we use the same functional model for the potential functions A, IS, and IT^{tk} for all image sites (thus applying a homogeneous CRF model), the subscripts of the potential functions in (1) have been omitted in (2). In (2), the association potential A corresponds to A_i in (1) for the labels and the image data in a specific epoch t and can be modeled in the same way. The second term of (2), IS, corresponds to the interaction potential I_{ij} in (1) for the labels and the image data in a specific epoch t, but it is called *spatial interaction potential* in order to distinguish it from the third term, the *temporal interaction potential* IT^{tk}. 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 neighborhood of the epoch to which image site i belongs, thus k is the time index of an epoch in temporal neighborhood of t. The set of image sites at epoch $k \in E_t$ that are temporal neighbors 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 neighbor to 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. As shown in Figure 1, the spatial neighborhood N_i of a pixel i consists of its four direct neighbors in the image grid. The definition of the temporal neighborhood is explained in Section 2.4.

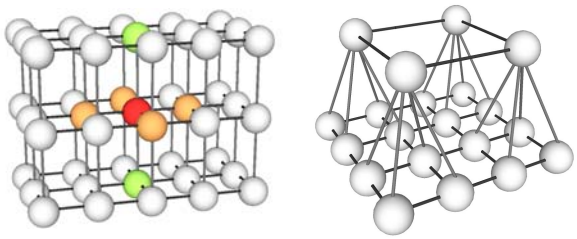

Figure 1: Left: Multitemporal graph structure for images having the same scale. Red node: processed primitive; orange nodes: spatial neighbors; green nodes: temporal neighbors.
Right: Graph structure for images having different scales.

2.2. 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 [9]. We use a simple Gaussian model for $P[x_i^t = c \mid \mathbf{f}_i^t(\mathbf{y}^t)]$ [23]:

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

In (3), $\mathbf{E}_{fc}^t$ and $\mathbf{\Sigma}_{fc}$ 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 over the images, because the definition of appropriate and expressive features depends on the scale and also on the spectral information contained in the images (Section 3).

2.3. 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 neighboring 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* $\boldsymbol{\mu}_{ij}^t(\mathbf{y}^t)$ [9]:

$$IS\left(x_i^t, x_j^t, \mathbf{y}^t\right) = \begin{cases} \beta \cdot \exp\left[-\dfrac{\left\|\boldsymbol{\mu}_{ij}^t\left(\mathbf{y}^t\right)\right\|^2}{R}\right] & if \quad x_i^t = x_j^t \\ \beta \cdot \left\{1 - \exp\left[-\dfrac{\left\|\boldsymbol{\mu}_{ij}^t\left(\mathbf{y}^t\right)\right\|^2}{R}\right]\right\} & if \quad x_i^t \neq x_j^t \end{cases} \quad (4)$$

In (4), $\|\boldsymbol{\mu}_{ij}{}^t(\mathbf{y}^t)\|$ 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 $\mathbf{f}_i^t(\mathbf{y}^t)$ for the interaction features $\boldsymbol{\mu}_{ij}{}^t(\mathbf{y}^t)$, i.e. $\boldsymbol{\mu}_{ij}{}^t(\mathbf{y}^t) = [\mu_{ij1}{}^t, \ldots \mu_{ijR}{}^t]^T$, where R is the dimension of the vectors $\mathbf{f}_i^t(\mathbf{y}^t)$ that may vary with t. Thus, denoting the m^{th} component of $\mathbf{f}_i^t(\mathbf{y}^t)$ by $f_{im}{}^t(\mathbf{y}^t)$, the m^{th} component of $\boldsymbol{\mu}_{ij}{}^t(\mathbf{y}^t)$ is $\mu_{ijm}{}^t = |f_{im}{}^t(\mathbf{y}^t) - f_{jm}{}^t(\mathbf{y}^t)|$. Dividing the Euclidean norm by the number of features R in (4) guarantees an identical influence of the spatial interaction potentials for all scales. This is a very simple model that penalizes local changes of the class labels if the data are similar and also penalizes identical class labels if the features are different.

2.4. 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 modeled 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 the effects of different atmospheric and lighting conditions, different scales, and seasonal effects on the vegetation. We use a simple model for the temporal interaction potential that neglects the dependency of IT^{tk} from the data:

$$IT^{tk}\left(x_i^t, x_l^k, \mathbf{y}^t, \mathbf{y}^k\right) = IT^{tk}\left(x_i^t, x_l^k\right) = \frac{\gamma \cdot \mathbf{TM}^{s(t)s(k)}\left(x_i^t, x_l^k\right)}{Q_i^k} \quad (5)$$

In (5), γ is a weighting factor. $\mathbf{TM}^{s(t)s(k)}$ is a temporal transition matrix similar to the transition probability matrix in [3]. 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. The set of epochs in the temporal neighborhood E_t of x_i^t is chosen to consist of the two epochs t-1 and t+1 if they both exist and of one epoch for the first and the last images of the sequence. In the multiscale case an image site i at epoch t might have more than just one temporal neighbor l in epoch k (right part of Figure 1). The set of temporal neighbors L_i^k consists of all image sites at epoch k that have a spatial overlap with i. The number of elements in L_i^k is denoted by Q_i^k. Q_i^k 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 neighbors exist. Our definition of the temporal interaction potential implies that we apply a bidirectional transfer of temporal information rather than a cascade approach handing information from one image to the next in a sequence as in [3].

As stated in Section 1, not only the appearance of classes, but also the structure of classes that can be discerned is a function of scale. This is considered in our method by defining one set of classes C_s for each group of images having a similar scale s. The set of classes C_s will be used for the images of all epochs having a scale similar to s. In our model for the temporal interaction potentials, this is considered by the superscript in the transition matrix $\mathbf{TM}^{s(t)s(k)}$ in (5), where $s(\cdot)$ denotes the scale of the respective epoch. There is one such matrix for any configuration of scales [$s(t)$, $s(k)$] of epochs t and k linked by the temporal interaction potential IT^{tk}. For example, if there are four epochs (t_1, t_2, t_3, t_4) with $s(t_1) = s(t_2) = s_1$ and $s(t_3) = s(t_4) = s_2$, there will be three such matrices, namely $\mathbf{TM}^{11}$, $\mathbf{TM}^{12}$, and $\mathbf{TM}^{22}$. Note that the transposed matrix can be used for message passing in the other direction: $\mathbf{TM}^{s(t)s(k)} = [\mathbf{TM}^{s(k)s(t)}]^T$. The transition matrices connecting classes at the same scale are square, but they are not symmetric due to the fact that some changes are more likely to occur in one direction than in the other. For instance, farmland is more likely to be changed into settlement as time passes than vice versa. The elements of these matrices also have to model the fact that change is not a very likely event. The transition matrices between different scales are rectangular. In this case, the fact that the most likely event to occur is often that nothing changes may not be modeled as easily because the class label might change simply due to the different class definition. It is relatively simple if there is a 1:N relation between the classes of the coarser scale and those of the finer scale. In this case, the classes at the coarser scale are aggregated classes merging N components that do not occur in any other aggregated class (e.g. the classes *building*, *garden* and *urban road* defined at a GSD of 1 m might be merged to a class *settlement* at a GSD of 30 m). Consequently, all the components will only support one aggregated class and all aggregated classes will only support their components. In case of N:M relations between classes defined at different scales, this is not as easily achieved, because a class defined at a fine scale might give support to more than one class at the coarser scale and vice versa (e.g. if the class *garden* in the previous example is replaced by *grass,* the class *grass* might not only be related to class *settlement* at a GSD of 30 m, but also to a class *pasture*). Currently, we only consider 1:N relations.

2.5. Training and Inference

Exact training and inference is computationally intractable for CRF [9]. In [24], several methods for parameter learning and inference are compared. 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 $\mathbf{\Sigma}_{fc}$ of the features of each class c in (3). 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 are the weighting factors β and γ of the spatial and temporal interaction potentials, respectively, and the elements of the transition matrices $\mathbf{TM}^{s(t)s(k)}$; cf. (4) and (5). Estimating the elements of $\mathbf{TM}^{s(t)s(k)}$ from training data would require a large amount of multitemporal data with a significant number of actual changes, which is not at our disposal. Thus, the transition matrices $\mathbf{TM}^{s(t)s(k)}$ are defined by the user based on expert knowledge about the likelihood of class changes. The weight factors β and γ could be determined from training data in a way similar to [9] if fully labeled training images for all epochs were available, but currently they are defined by the user. The parameter values used in our experiments (cf. Section 4) were found empirically.

For inference, we use Loopy Belief Propagation (LBP) [25], a standard technique for performing probability propagation in graphs with cycles that has shown to give good results in the comparison reported in [24]. In this context, edges linking temporal neighbors are treated in the same way as edges linking spatial neighbors in message passing, except that the messages are different.

3. Features and Class Structure

3.1. Feature Vectors

The site-wise feature vectors $\mathbf{f}_i^t(\mathbf{y}^t)$ used both for the association potential and for deriving the interaction feature vectors must be defined such that they can help to discriminate the different classes. The definition of the features for classification depends on the scale and the spectral configuration of the images. As far as the spectral configuration is concerned, only color infrared images containing a green, a red, and a near infrared band were available for our experiments. We had images at two scales: one group had a GSD of 4-5 m, whereas the GSD of the second group was 30 m (cf. Section 4).

From the HR imagery (GSD 4-5 m), features are extracted at three different scales λ_1, λ_2 and λ_3 for each pixel. In this way, dependencies between the image data of neighboring sites are modeled. The scale λ_1 corresponds to the original resolution, and the only features extracted at this scale are the three color values observed at each pixel. At the scales λ_2 and λ_3, the pixels in a square of size 5 pixels and 11 pixels, respectively, centered at the particular pixel, are taken into account to compute two groups of features. The *color-based features* at scales λ_2 and λ_3 are the mean of the green, red and infrared channels as well as the variance of the red channel. The variances of the other channels were found out empirically not to contain significantly more information. The *gradient-based features* are derived from a weighted histogram of the orientations of the intensity gradient. We use the mean and the variance of these orientations along with the number of bins containing values above the mean. These features allow a good distinction between textured and homogeneous areas. The site-wise feature vectors $\mathbf{f}_i^t(\mathbf{y}^t)$ for the HR data thus consist of 17 elements (3 for λ_1, 4 + 3 for λ_2 and λ_3, respectively). For the medium resolution data (GSD 30 m), the use of gradient based features and the application of larger scales was found not to be suitable. Here the feature vectors $\mathbf{f}_i^t(\mathbf{y}^t)$ just consist of 3 elements, which are the green, red and infrared values at the respective pixel position. The values for each feature in each data set are normalized so that they are in the interval [0, 10] for the training sites. Note that the principles described here could be easily expanded to other spectral configurations and / or images of other spatial resolutions.

3.2. Class Structure

As stated in Section 2.4, the class structure depends on the scale of the images, and we assume 1:N relations between the classes at the coarser scale and the classes at the finer scale. In the medium resolution (GSD 30 m) images used in our experiments, three classes are to be distinguished, namely built-up areas (*bui*), forests (*for*), and cropland (*crp*). In the HR images (GSD 4-5 m), in the built-up areas there is a clear distinction between residential areas (*res*) and industrial areas (*ind*). Thus, the class *bui* in the medium resolution images corresponds to the two classes *res* and *ind* in the HR images.

4. Experiments

4.1. Test Data and Set-up of the Experiments

Our test area is situated near Herne, Germany, and covers an area of 8.6 x 8.6 km² (Figure 2). We used multispectral Ikonos data with 4 m GSD acquired in 2005, RapidEye data with 5 m GSD acquired in 2009, and Landsat data of 30 m GSD acquired in 2010. All images were recorded in summer. About 10% of the scene was used for training, the rest for testing. Ground truth was obtained by manually labeling the images on pixel level.

For change detection analysis some regions containing land cover changes are needed. Unfortunately in our test site there are too few and too small changes to be detected with Landsat imagery. For that reason, we manipulated the Landsat scene, creating 12 new built-up areas by copying

data from another Landsat scene. Some of these new built-up areas are connected to built-up areas in the original scene, whereas others correspond to completely new settlements surrounded by cropland and forests (Figure 3).

Figure 2: Ikonos (2005) image of the entire test site.

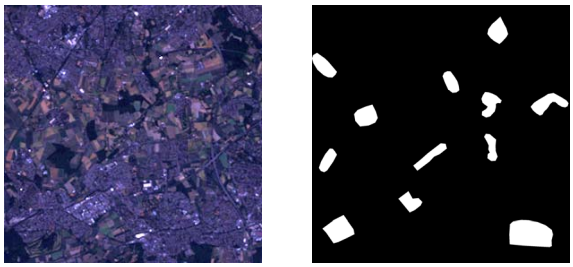

Figure 3: Left: Manipulated Landsat image (2010). Right. Manipulated areas (white).

We used the class structure defined in Section 3.2, the HR data corresponding to the Ikonos and RapidEye images and the medium-resolution data to the Landsat image. We tested our approach for two different data set-ups: Set-up I consists of the Ikonos and the manipulated Landsat scenes, the typical set-up described in Section 1. In set-up II we inserted a RapidEye scene between the two epochs of set-up I to investigate how additional data of a different sensor influences the results.

The temporal transition matrix **TM** between Ikonos / RapidEye 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 II. As learning the parameters of **TM** was impossible given the available data, they were set based on expert knowledge. 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. In our case a transition from forest or cropland to built areas is more likely than vice versa. The factors β and γ used in our models for the spatial and temporal interaction potentials (Equations 4 and 5, respectively) were set to $\beta = 1$ and $\gamma = 1.5$. These values were found empirically.

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, obtained by replacing the exponent in (4) by 0. For these three scenarios, the overall classification accuracy and the kappa coefficients are compared for all epochs. We additionally applied a monotemporal CRF-based classification to the Landsat image (scenario CRF_{mono}) and compared the classification accuracy and the confusion matrices to those achieved for the scenarios CRF_{multi} and *ML*. We also assessed the capability of our method for detecting the actual changes in the Landsat scene.

	$x_i^{t+1} = bui$	$x_i^{t+1} = for$	$x_i^{t+1} = crp$
$x_i^t = res$	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 Ikonos/RapidEye imagery, *t+1* corresponds to Landsat imagery.

4.2. Results and Evaluation

Figure 4 shows the reference for the Landsat scene and some of the classification results. Table 2 shows the overall accuracy and the kappa coefficients achieved for the individual scenes. The scenario *ML* performed worst, with an overall accuracy of only 64% for the Landsat scene. Figure 4 also shows the salt-and pepper-like appearance of the results. Taking into account spatial context in scenario CRF_{mono} improves the overall accuracy for the Landsat scene by 8%. The impact of the multi-temporal approach is highlighted by the overall accuracy achieved in the scenario CRF_{multi}, where over 79% could be achieved in both set-ups. The higher information content of the HR images clearly propagates to the medium resolution scene and yields a significant increase of accuracy of 7% there. 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. It may come as a bit of a surprise that regarding the data in the spatial interaction potential of CRF_{multi} does not improve the results. This may be attributed to the rather simple model that, for instance, weight differences in the multiscalar features in the HR images (which are not likely to change very much between neighboring pixels) in the same way as those for the scale λ_1. Integrating the RadidEye image in set-up II did not affect the overall accuracy either. The poorer *ML* classification results of the RapidEye scene indicate that the model of the association potential does not fit as well as for the Ikonos data.

The confusion-matrices of the different classification scenarios for the manipulated Landsat scene in set-up I are shown in Table 3.

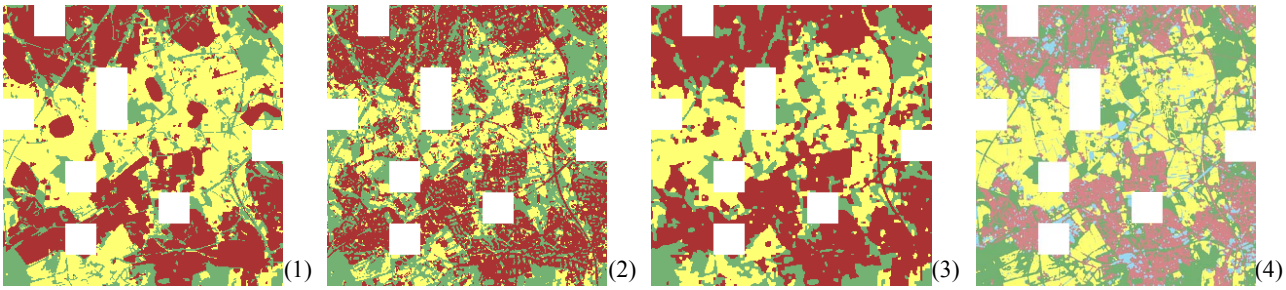

Figure 4: (1) Reference (Landsat); (2) Results of *ML*; (3) Results of CRF_{multi} (both Landsat) (4) Results of CRF_{multi} (Ikonos). Classes: yellow: *crp*; green: *for*; dark red: *bui* (Landsat); light red: *res*; light blue: *ind* (Ikonos); white: training areas.

S/E	CRF_{multi}	CRF_{mono}	*ML*	*MRF*
I / t_1	77.6% / 0.68	-	74.3% / 0.63	77.8% / 0.68
I / t_2	79.4% / 0.68	72.2% / 0.58	64.4% / 0.49	79.2% / 0.68
II / t_1	78.2% / 0.69	-	74.3% / 0.63	78.1% / 0.68
II / t_2	77.9% / 0.68	-	68.2% / 0.55	77.8% / 0.68
II / t_3	79.2% / 0.68	72.2% / 0.58	64.4% / 0.49	79.2% / 0.68

Table 2: Overall classification accuracy / kappa coefficients; S/E: Set-up / epoch; set-up I: t_1: Ikonos, 2005 ; t_2: Landsat, 2010 ; set-up II. t_1: Ikonos, 2005 ; t_2: RapidEye, 2009, t_3: Landsat, 2010.

The results for set-up II and for *MRF* are not shown because there is hardly any difference to those for CRF_{multi} of set-up I. Again there is a clear trend that the quality of the results is improved by considering the spatial context and even more so by the temporal interactions. The completeness and the correctness (producer's and user's accuracy) are improved going from *ML* via CRF_{mono} to CRF_{multi}. The only exceptions are the correctness of class *bui*, which is marginally higher in CRF_{mono} than in CRF_{multi}, an effect compensated by an increase in completeness by 5%, and the completeness of *for*, which is lower in CRF_{mono} than in *ML*. The biggest improvements are achieved in the completeness of class *crp* and the correctness of class *for*. The class *crp* has a very inhomogeneous appearance, which may be the reason why temporal context is important for distinguishing it from *for*. A problem for all scenarios is the relatively high rate of false positive *bui* pixels.

Figure 5 shows the classification results of the Landsat scene for the changed areas in CRF_{mono} and CRF_{multi}. In both set-ups, 70 % of the changed pixels were correctly classified as built-up areas in CRF_{multi} (18% forest, 12% cropland). Using CRF_{mono}, even 87% of the changes could be detected. Whereas in general the temporal interaction term improves the classification considerably, it also oversmoothes some areas of actual change. It has to be noted, though, that in 10 out of 12 changed areas (83%), the majority of the pixels in the respective areas correctly indicated a change. To improve the method's capabilities for change detection, the model for the temporal interactions could be augmented to consider a dependency of the class transitions from the data.

	x_i^{ref}	x_i = *bui*	x_i = *for*	x_i = *crp*	***Comp.***
CRF_{multi}	*bui*	26931	2041	956	**90.0%**
	for	3951	11115	898	**69.6%**
	crp	4086	3448	21326	**73.9%**
	Corr.	**77.0%**	**66.9%**	**92.0%**	
CRF_{mono}	*bui*	25382	3344	1202	**84.8%**
	for	3544	10089	2331	**63.2%**
	crp	3312	6966	18582	**64.4%**
	Corr.	**78.7%**	**49.5%**	**84.0%**	
ML	*bui*	22660	5662	1606	**75.7%**
	for	3669	11087	1208	**69.5%**
	crp	4701	9764	14395	**49.9%**
	Corr.	**73.3%**	**41.8%**	**83.6%**	

Table 3: Confusion matrices for the Landsat scene (pixel numbers in set-up I). *Comp.* / *Corr.*: completeness / correctness. Classes: *bui* = built area; *for* = forest; *crp* = cropland.

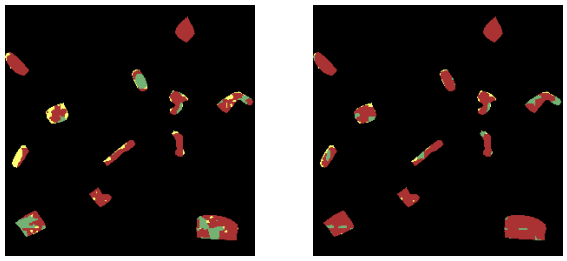

Figure 5: Results of classification in the changed areas in the Landsat scene. Left: CRF_{multi} of set-up I; right: CRF_{mono}.

5. Conclusions

We have presented a novel CRF-based approach for multitemporal and multiscalar image classification with the goal of detecting changes in and improving the classification accuracy of medium resolution data by combining them with HR data of an earlier acquisition time. Besides incorporating spatial context, our method uses a model of temporal context by introducing a temporal interaction potential in order to take into account dependencies between regions at identical positions in images acquired at different times and scales along with a scale-dependent definition of the class structure.

It was shown that the overall classification accuracy of the medium resolution image was improved by about 8% by including spatial context and by another 7% by considering the temporal interactions. These results are quite promising, even more so because they were achieved with a simple set of features. The overall accuracy could still be improved by using better features, but this was not the focus of this work; the impact of the spatial and temporal interaction potentials is quite impressive. The method is capable to detect most of the changes in the scene despite the smoothing effect of the temporal interactions, though one has to note that there is a loss of accuracy in comparison to monotemporal classification. Nevertheless, we think that the multitemporal / multiscale graph structure can be applied to many other tasks in image classification.

The fact that the CRF model did not yield better results than the MRF approach indicates that in the future, the model for the spatial interaction potential should be improved, e.g. by a better selection of the interaction features and by applying a generalized linear model [9], whose parameters can be learnt from the training data. Apart from that, further research will concentrate on an improvement of the model for the temporal interaction potential to improve the completeness of the detected changes. Learning of at least some of the parameters of the temporal interaction potential is also of interest.

Acknowledgement

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

References

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

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

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

[4] G. Geman, D. Geman. Stochastic relaxation, Gibbs distribution and Bayesian restoration of images. *IEEE-TGARS*, 6(6):721-741, 1984.

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

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

[7] J. Lafferty, A. McCallum, F. Pereira. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. *Proc. Int. Conf. Machine Learning*: 8p, 2001.

[8] S. Kumar, M. Hebert. Discriminative Random Fields: A discriminative framework for contextual interaction in classification. *Proc. Int.'l Conf. Computer Vision*, 2: 1150-1157, 2003.

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

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

[11] T. Hoberg, F. Rottensteiner. Classification of settlement areas in remote sensing imagery using Conditional Random Fields. *Int'l. Arch. Photogrammetry, Remote Sens., SIS* XXXVIII (7), 2010.

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

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

[14] J.D. Wegner, U. Sörgel, B. Rosenhahn. Segment-based building detection with Conditional Random Fields, *Proc. 6th Joint Urban Remote Sensing Event*: 205-208, 2011.

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

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

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

[18] C. A. Bouman, M. Shapiro. A multiscale random field for Bayesian image segmentation. *IEEE-TIP,* 3(2):162-177, 1994.

[19] J. Kersten, M. Gähler, S. Voigt. A general framework for fast and interactive classification of optical VHR satellite imagery using hierarchical and planar Markov Random Fields. *PFG* 6(2010):439–449, 2010.

[20] M. J. Choi, V. Chandrasekaran, D. M. Malioutov, J. K. Johnson, A. S. Willsky. Multiscale stochastic modeling for tractable inference and data assimilation. *Comput. Methods Appl. Mech. Engrg*. 197 (2008): 3492–3515, 2008.

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

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

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

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

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