Conditional Random Fields for Multitemporal and Multiscale Classification of Optical Satellite Imagery

Thorsten Hoberg, Franz Rottensteiner, Raul Queiroz Feitosa, and Christian Heipke

***Abstract*—In this paper, we present a method for the multitemporal and contextual classification of georeferenced optical remote sensing images acquired at different epochs and having different geometrical resolutions. The method is based on Conditional Random Fields (CRFs) for contextual classification. The CRF model is expanded by temporal interaction terms that link neighboring epochs via transition probabilities between different classes. In order to be able to deal with data of different resolution, the class structure at different epochs may vary with the resolution. The goal of the multitemporal classification is an improved classification performance at all individual epochs, but also the detection of land-cover changes, possibly using lower resolution data. This paper also contains a comparison of the performance of different models for the interaction potentials. Results are given for two different test sites in Germany, where Ikonos, RapidEye, and Landsat images are available. Our results show that the multitemporal classification does indeed increase the overall accuracy of all epochs compared to a monotemporal classification and to a state-of-the-art multitemporal classification method, and that it is feasible to detect changes in lower resolution images.**

***Index Terms*—Change detection, conditional random field (CRF), Markov random field (MRF), multiscale, multitemporal classification.**

I. Introduction

AN INCREASING number of optical high-resolution (HR) remote sensing satellite systems, offering multispectral images at a Ground Sampling Distance (GSD) of 5 m or below, have become available, e.g., Ikonos, Quickbird, WorldView-1, and WorldView-2, to name just a few. As a consequence of the higher availability of data and the higher quality of these images, it should 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, acquiring multitemporal HR data may not always be economically viable, particularly for large areas. Data having medium resolution (i.e., a GSD of 30 m) do not offer as much detail, but cover a larger area and may often be preferable from an economical point of view. It would be desirable to have a method capable of combining HR images with data of lower resolution and acquired at different epochs of arbitrary order for classification and for detecting land-cover changes. 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. Recent work on image classification has emphasized the importance of considering local context [1], [2], but only in a monotemporal setting. In order to achieve the general goals described earlier, in this paper, we want to extend such a framework to multitemporal classification and change detection, taking into account interactions between images acquired at different epochs and considering the fact that these images may have different geometrical resolutions.

A. Related Work

The detection of land use or land-cover changes based on multitemporal classification is a rather broad topic; overviews can be found in [3]–[7]. There are three groups of approaches [7]. The first group directly compares the data acquired at different epochs by computing differences or ratios of features and applies thresholds to decide whether there is a change or not. Apart from problems in threshold selection, the strongest weakness of this strategy is based on the fact that the radiometric appearance of certain land-cover types may be rather different at different points in time [8]. This is obvious in case of agricultural areas, where the appearance of crops varies considerably throughout a year. Furthermore, such approaches only allow a decision that there has been some change, but it is not possible to determine the type of change [5]. The second group extracts land use or land cover at different epochs and then compares the resultant land-use or land-cover maps. Such methods are more robust with respect to the radiometric variability of different land-cover classes, and they also deliver the information about the actual type of change. A sufficient amount of training data for classification at all epochs is required for that purpose, because any classification error will directly lead to an error in change detection [5]. Up to now, most

Manuscript received July 10, 2013; revised November 29, 2013 and April 10, 2014; accepted May 16, 2014. Date of publication June 20, 2014; date of current version August 12, 2014. This work was supported by the German Science Foundation under Grant HE 1822/22-1 and the EU-FP7-project "Tools for Open Multi-Risk Assessment using Earth Observation Data" (TOLOMEO) of the Marie Curie International Research Staff Exchange Program.

T. Hoberg was with the Institute of Photogrammetry and GeoInformation (IPI), Leibniz Universität Hannover, 30167 Hannover, Germany. He is now with the State Agency of Geoinformation and Land Development of Lower Saxony, Germany (e-mail: thorsten.hoberg@lgln.niedersachsen.de).

F. Rottensteiner and C. Heipke are with the Institute of Photogrammetry and GeoInformation (IPI), Leibniz Universität Hannover, 30167 Hannover, Germany (e-mail: rottensteiner@ipi.uni-hannover.de; heipke@ipi.uni-hannover.de).

R. Q. Feitosa is with the Department of Electrical Engineering, Pontifical Catholic University of Rio de Janeiro, 22451-900 Rio de Janeiro-RJ, Brazil, and also with the Department of Computer Engineering, State University of Rio de Janeiro, 22451-900 Rio de Janeiro-RJ, Brazil (e-mail: raul@ele.puc-rio.br).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2014.2326886

approaches for multitemporal land-cover analysis correspond to these first two groups.

The third group of approaches analyzes all data simultaneously in a combined model. Such a strategy is advantageous because the original observations (i.e., the image data) are used for the combination rather than derived data, whereas it is still possible to model the actual type of change. This has, for instance, been done in [9], where a model of temporal dependencies based on Markov chains is applied. Whereas the authors say that their method could be applied to both a pixelwise classification and a segmentwise classification, they only show results for the segmentwise approach. Segmentation errors cannot be overcome by the classification in such a setting. In [10], a cascade of three multitemporal classifiers is combined for the detection of land-cover changes. One of the three classifiers considers the k-nearest neighbors of each pixel in feature space, but does not model the spatial dependencies between neighboring pixels in image space.

A statistical model of the spatial dependencies between the class labels at neighboring pixels and, thus, of spatial context, in image classification is given by Markov Random Fields (MRFs) [11], which have been used for change detection in [12] and [13]. In [12], 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.* [13] applied the MRF framework 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 (CRFs; [1]). CRFs 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, CRFs have, for instance, been used for the classification of settlement areas in HR optical satellite images [14], for generating a digital terrain model from LiDAR data [15], and for classifying crop types and other land-cover classes in Landsat data [16]. 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, Wegner *et al.* [17] use an irregular graph derived from a segmentation for CRF-based building detection from airborne optical and 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. A recent overview on and comparison of MRF, CRF, and other smooth labeling methods for remote sensing applications can be found in [2]. The author of that paper concludes that any smooth labeling technique outperforms a purely local classfier, with CRF performing best among the compared techniques by a small margin. Unfortunately, except for the overview [2], which compares an MRF with a contrast-sensitive Potts model, in most cases only one model for the interaction potential is applied, without justification of the choice of the particular model. As far as multitemporal classification is concerned, CRFs have first been applied in our previous work [18]. In that contribution, all the input images are assumed to have approximately the same geometrical resolution. This is also true in [19], where changes are detected in a pair of aerial images by combining a mixed Markov model with a conditionally independent random field. The method can only deal with two images of the same geometrical resolution and only delivers a binary result (*change* versus *no change*).

In order to be able to deal with images at different resolutions, a multiscale analysis is required. 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 [20], [21]. The simplest way of considering multiple scales in classification is to derive the features at multiple scales, e.g., [1], which has been applied for change detection in [13]. Multiscale random fields were defined in [22] as a computationally more tractable alternative to MRF. The image is represented by a pyramidal structure based on an quadtree whose leaves are the individual pixels. Each node of the quadtree 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 quadtree nodes at coarser scales propagating their information to the leaves. A similar approach was used for image classification in [23], where it was also noted that a multiscale approach requires a redefinition of the class structure at the coarser scales, because in different resolutions, different classes are discernable. In [24], such a multiscale representation is combined with spatial interactions at each scale of the representation, but no results are presented. In [25], we expanded our work in [18] to the multiscale case, although no comparison of the effects of different models of the interaction potentials was carried out.

There have also been approaches to combine a multiscale analysis with CRF. In [26], 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 coarser resolutions, they correspond to compound objects. This method is applied to detect objects such as motorbikes in terrestrial images. In [27], this method is extended to an irregular pyramid based on a multiscale 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 [13], most of the multiscale methods are based on monotemporal images. A few of them consider the fact that the class structure changes with scale [23], [26]. Nearly all of them require the images at full resolution to have the same scale, with exceptions for the monotemporal case being given in [21].

B. Contribution

In this paper, we present an approach for the classification of multitemporal and multiscale optical remote sensing data. A set of multispectral images of different resolution is classified simultaneously, in order to achieve two goals, i.e., to increase the accuracy of the classification results compared to a monotemporal setting and, at the same time, 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 that we first proposed in [25]. 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, thus, by different (although related) class structures. This goes beyond existing work on context-based monotemporal image classification, e.g., [1], [2], and [17]. We evaluate our method on two multitemporal data sets consisting of Ikonos, RapidEye, and Landsat images. In our experiments, we compare three different spatial context models. Further tests assess the impact of the temporal interaction potentials on the classification accuracy, using both images having identical resolutions and images having different resolutions. We thus show that, in addition to increasing the overall accuracy of classification, our method also has a high potential in detecting changes in the medium-resolution data. Our method is also compared to a state-of-the-art multitemporal classification technique [28], to show the improvement in classification performance achieved by considering spatial interactions.

In Section II, we summarize the principles of CRF. This is followed by a presentation of the extensions for the classification of multitemporal and multiscale data in Section III. In Section IV, we describe our test sites and the experimental setup, whereas a thorough quantitative evaluation of the results of our method is presented in Section V. Conclusions and an outlook are given in Section VI.

II. CRFs

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 CRFs. In monotemporal classification, we want to determine the vector of class labels $\mathbf{x}$, whose component x_i corresponds to the class label of image site $i \in S$, where S is the set of all image sites. The class labels $\mathbf{x}$ shall be derived from the image data $\mathbf{y}$, which are composed of sitewise feature vectors $\mathbf{y}_i$, i.e., $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, where N is the number of image sites. The most probable label configuration $\mathbf{x}$ can be obtained by maximizing the posterior probability $P(\mathbf{x} \mid \mathbf{y})$ of the labels given the data as follows [1]:

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

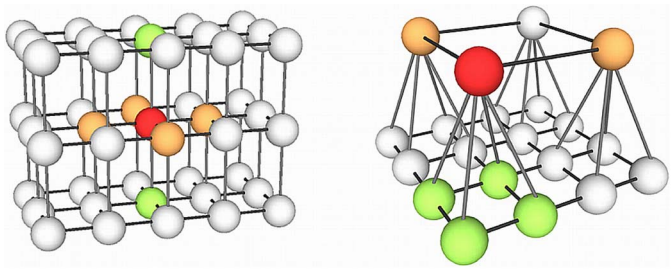


Fig. 1. Multitemporal graph structure. (Left) Images having the same resolution, corresponding to three epochs. (Right) Images having different resolutions, corresponding to two epochs. Red nodes: processed node; orange/green nodes: spatial/temporal neighbors of the red nodes.

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}, which is 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, Kumar and Hebert [1] use generalized linear models for both potentials.

III. Multitemporal CRFs

In the multitemporal case, we have M coregistered images. In addition to the interactions of spatial neighbors, the temporal neighborhood is taken into account. Each node is only linked to its direct temporal neighbors at its spatial position (see Fig. 1). The components of the image data vector $\mathbf{y}$ are sitewise 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 position in object space, 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 $T = \{1, \ldots, M\}$. The components of $\mathbf{x}$ are the class labels of the image sites i, i.e., $\mathbf{x}_i^t$, also with epoch index $t \in T$. For each image site, we want to determine the class x_i^t from a predefined set of classes. In order to be able to handle data of different resolution, the class structure and, thus, the number of classes may depend on t. In order to model the mutual dependencies of the class labels at an image site at different epochs, the model in (1) has to be extended as follows:

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

Since 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 is the *spatial interaction potential* that corresponds to the interaction potential I_{ij} in (1), and IT^{tk} is 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 neighborhood of the epoch to which image site i belongs; thus, k is the time index of an epoch in a 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 of i in epoch k. The temporal interaction potential models the dependence between the class labels and the observed data at consecutive epochs. The image sites are chosen to be individual pixels; Fig. 1 shows the structure of the resultant graph. The model in (2) provides a general framework for multitemporal classification based on CRF. Our definitions of the potentials are given in the subsequent sections.

A. 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 sitewise 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. Basically, any local classifier with a probabilistic output can be used in this context [1]. We use a multivariate Gaussian model

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

In (3), $\mathbf{E}_{fc}^t$ and $\mathbf{\Sigma}_{fc}^t$ are the mean and the covariance matrix of the features of class c, respectively, whereas n denotes the dimension of the feature vectors. The definition of the features $\mathbf{f}_i^t(\mathbf{y}^t)$ may vary from image to image; our definition is given in Section IV-B.

B. 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 label x_j^t on the class x_i^t of image site i at epoch t. In this potential, the data are represented by *interaction features* $\mathbf{g}_{ij}^t(\mathbf{y}^t)$ that are determined independently for each pair of neighboring sites i and j, but may again depend on the entire image at epoch t. In this paper, we compare three different models for the spatial interaction potential. The first model is a Potts model that only depends on the labels. It is commonly used with MRF and has a smoothing effect on the labels. It can be modelled as follows [2]:

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

The second model is a modified version of the contrast-sensitive Potts model described in [29] and formulated as follows:

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

The third model is an extension of the contrast-sensitive Potts model that was introduced in [18] as follows:

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

In (4)–(6), $\|\mathbf{g}_{ij}^t(\mathbf{y}^t)\|$ denotes the Euclidean norm of $\mathbf{g}_{ij}^t(\mathbf{y}^t)$, and β is a weighting factor for the influence of the spatial interaction potential in the classification process. The parameter η modulates the impact of the data-dependent terms in IS_2 and IS_3. For $\eta = 0$, IS_2 and IS_3 become equivalent to IS_1. IS_2 penalizes a class change between neighboring nodes if the data observed at these nodes are similar, but does not favor any configuration if they are different. In contrast, IS_3 actively encourages a class change if the features at the neighboring sites are different. R is the dimension of the interaction feature vectors, thus $\mathbf{g}_{ij}^t = [g_{ij_1}^t, \ldots, g_{ij_R}^t]^T$, and may vary with t. Division by the number of features R in (5) and (6) guarantees an identical influence of the spatial interaction potentials for all images. Our definition of the interaction feature vectors $\mathbf{g}_{ij}^t$ is based on the differences of sitewise feature vectors for the interaction potentials $\mathbf{h}_i^t(\mathbf{y}^t)$; thus, $\mathbf{g}_{ij}^t = \mathbf{h}_j^t(\mathbf{y}^t) - \mathbf{h}_i^t(\mathbf{y}^t)$. In this context, the vectors $\mathbf{h}_i^t(\mathbf{y}^t)$ may be different from the feature vectors $\mathbf{f}_i^t(\mathbf{y}^t)$ used for the association potentials [1].

C. 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. It 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 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 dependence of IT^{tk} of the data, which is formulated as follows:

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

In (7), γ is a weighting factor. $\mathbf{TM}^{s(t),s(k)}$ is a temporal transition matrix similar to the transition probability matrix in [10]. The elements of $\mathbf{TM}^{s(t),s(k)}(x_i^t, x_l^k)$ are related to conditional probabilities $P(x_i^t = c^t | x_l^k = c^k)$ 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. 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; otherwise, it only consists of one epoch. In the multiscale case, an image site i at epoch t might have more than just one temporal neighbor l in epoch k (i.e., right part in Fig. 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 results in a bidirectional transfer of temporal information rather than in a sequential approach as in [10]. Consequently, images acquired at different epochs mutually support each other in the classification.

As noted earlier, the structure of classes that can be discerned is a function of scale. Thus, we define one set of classes C_s for each group of images having a similar scale s. This is considered by the superscript in the transition matrix $\mathbf{TM}^{s(t),s(k)}$ in (7), 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. The transposed matrix can be used for message passing in the other direction, i.e., $\mathbf{TM}^{s(k),s(t)} = [\mathbf{TM}^{s(t),s(k)}]^T$. The transition matrices connecting classes at the same scale are square; however, they are not symmetric because some changes are more likely to occur in one direction than in the other. 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 class label might change simply due to a different class definition, so that it is not as straightforward to model the fact that the most likely event to occur is that nothing changes. 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., *building*, *garden*, and *urban road* defined at a GSD of 1 m might be merged to *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 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 *garden* in the previous example is replaced by *grass*, it might be related not only to class *settlement*, at a GSD of 30 m, but also to *pasture*). Currently, we only consider $1 : N$ relations.

D. Training and Inference

Exact training and inference is computationally intractable for CRF, except for special cases in binary classification [1], [30], [31]. We only train the parameters of the association potentials, i.e., the mean $\mathbf{E}_{fc}^t$ and the covariance matrix $\mathbf{\Sigma}_{fc}$ of the features of each class c [cf. (3)]. They are determined independently for each epoch t and each class c. The other model parameters, i.e., the weighting factors β, η, 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 (cf. Section IV-C). For inference, i.e., the determination of the optimal label configuration $\mathbf{x}$ based on our model in (2), we use Loopy Belief Propagation (LBP) [32], which is a standard technique for probability propagation in graphs with cycles that has shown to give good results in the comparison reported in [33]. See also [31] for a detailed comparative study on different methods for finding the optimal label configuration in CRF.

Fig. 2. Test area in Herne (Ikonos, 2005).

IV. Experimental Setup

A. Test Data

We used two test areas for the evaluation of our method. The first test area is situated near Herne, Germany, and covers an area of 8.6×5.9 km^2 (see Fig. 2). We used multispectral Ikonos data with 4-m GSD acquired in 2005 and 2007, a multispectral RapidEye image acquired in 2009 with an original GSD of 5 m, and Landsat data of 30-m GSD acquired in 2010. All images were recorded in summer. We produced orthophotos with 4-m GSD from the Ikonos and RapidEye images and with 30-m GSD from the Landsat images. The classes to be distinguished with Ikonos and RapidEye imagery are residential areas (res), industrial areas (ind), forests (for), and cropland (crp). Since there is no clear distinction of the classes res and ind in the Landsat imagery, they are fused to a new class built-up areas (bui) in that resolution. Reference data were obtained by manually labeling the images at pixel level in the Ikonos scene from 2005. The reference for res contained roads inside settlements, gardens attached to buildings, but also roads having a width larger than 8 m outside settlements. The percentage of the area covered by the four classes was 30% (res), 5% (ind), 22% (for), and 43% (crp), respectively.

The second test area is located near Husum, Germany, and covers an area of 16.8×9.6 km^2 (see Fig. 3). In Husum, we had RapidEye images acquired at four different epochs (4/2009, 7/2009, 3/2010, and 7/2010) and a Landsat image acquired in 7/2009. The orthophotos, which formed the basis for all our experiments, were generated with a GSD of 5 m for RapidEye and 30 m for Landsat. The set of classes was identical to the one distinguished in Herne, but the reference was based on the German Authoritative Topographic Cartographic Information System (ATKIS) [34]. Outside the settlement areas, major roads, as indicated by ATKIS, are included in class res. Since roads are contained in ATKIS as linear objects, a constant road width of 10 m was assumed. The percentage of the area covered by the four classes was 10% (res), 1.5% (ind), 8.5% (for), and 80% (crp), respectively. Both test sites are flat, and

Fig. 3. Test area in Husum (RapidEye, 7/2009).

TABLE I
FEATURE POOL. THE SUPERSCRIPT d INDICATES THE SCALE. r, g, b, nir: RED, GREEN, BLUE, AND NEAR-INFRARED BANDS; $ndvi$: NORMALIZED DIFFERENCE VEGETATION INDEX; rvi: RELATIONAL VEGETATION INDEX. HOG: HISTOGRAM OF ORIENTED GRADIENTS

Color-based features		
1-9	E_c^d	mean value of feature c with $c \in \{r, g, b, nir, r-g, nir-r, nir-g, ndvi, rvi\}$
10-16	V_c^d	variance of feature c with $c \in \{r, g, b, nir, hue, ndvi, rvi\}$
Texture features		
17	con^d	contrast
18	cor^d	correlation
19	ene^d	energy
20	hom^d	homogeneity
21	ent^d	entropy
Structural features		
22	E_{grad}^d	mean gradient magnitude
23	V_{grad}^d	variance of the HOG entries
24	num^d	number of HOG bins with magnitudes above E_{grad}^d
25	mag^d	maximum HOG entry
26	ang^d	angle between the first two HOG maxima

the agricultural areas show a characteristic pattern for Central Europe with rather small fields of heterogeneous appearance. The main difference between the test sites is that, in Herne, there is a more homogeneous distribution of the individual classes than in Husum, where the most dominant class (crp) accounts for 80% of the area.

B. Features

In order to apply the CRF framework, the sitewise 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. We define a pool of altogether 26 features belonging to three different groups (cf. Table I). All features are determined in different scales s_d, where a scale is characterized by the local neighborhood from which the features are determined (squares of $d \times d$ pixels). The color-based features are directly derived from the pixel values of a multispectral image. We restrict ourselves to the four bands available for Ikonos, in all cases. They comprise the mean values and variances of the four bands, the variance of hue, the mean differences of some pairs of channels, and the mean values and variances of two vegetation indexes. The five textural features are a subset of the Haralick features derived from the gray-level co-occurrence matrix at an offset of one pixel [35]. The structural features are derived from a weighted histogram of oriented gradients (HOG) [36]. Each histogram has 30 bins corresponding to orientation intervals of 6° width. Each bin contains the sum of the magnitudes of all gradients having an orientation within the interval corresponding to the bin. From the histogram, we derive five features (see Table I).

In scale s_1, only individual pixels are taken into account, and thus, only the first nine features in Table I can be computed. For the Landsat images, we only determine these nine features. For the Ikonos and RapidEye images, we compute the nine features for s_1 and the 26 features for scales s_3 and s_5 (considering square windows of 3-and 5-pixel side lengths, respectively). This results in a pool of altogether 61 features for these images from which to choose the sitewise feature vectors $\mathbf{f}_i^t(\mathbf{y}^t)$. All the features are normalized by a linear transformation, so that the values lie in the interval [0, 1].

The neighborhood from which these features are computed, corresponds to 20×20 m^2 and 25×25 m^2 in the Ikonos and RapidEye images, respectively, whereas the features for Landsat are representative for the area covered by one Landsat pixel $(30 \times 30\,\text{m}^2)$. Choosing the maximum scale s_5 for Ikonos and RapidEye is motivated by the fact that some of the classes, in particular res, are characterized by the texture and the structural features rather than by the color bands themselves. The minimum size of a local window for which these types of features can be computed is 3×3 pixels; choosing a slightly larger neighborhood allows capturing the typical structure of overland roads in these features. Extending the region of influence for the sitewise feature vectors might lead to oversmoothing. Since a certain degree of smoothing is to be expected as a consequence of the spatial interaction potential anyway [2], there is no need to further expand the feature pool by considering additional (coarser) scales.

The interaction features $\mathbf{g}_{ij}^t(\mathbf{y}^t)$ are based on the differences of the features $\mathbf{h}_i^t(\mathbf{y}^t)$, $\mathbf{h}_j^t(\mathbf{y}^t)$ at neighboring sites (cf. Section III-B). The coarser the scale, the larger is the overlap between the neighborhoods from which the features at two neighboring sites are determined. Thus, for the Ikonos and RapidEye images, we restrict the feature pool for vectors $\mathbf{h}_i^t(\mathbf{y}^t)$ to the 35 features at scales s_1 and s_3. We also consider s_3 for the classification of the Ikonos and RapidEye images, so that differences in texture features have an influence on the interaction potentials. For the Landsat images, the original pixel size is such that we only consider the nine features that can be determined from a single pixel.

Using a relatively large number of features makes the classification rather time consuming. We applied the greedy correlation-based feature selection approach suggested by Hall [37] to a subset of the data. This approach basically tries to find features that are highly correlated with the class labels but uncorrelated with the other features. This resulted in three ordered lists of features (depending on the image and potential type). We decided to use the six most relevant features for all three types of feature vectors in our experiments; they are shown in Table II. We selected the first six features because a set of preliminary test runs on a subset of the data showed that using additional features did not lead to an improvement of the classification. Looking at Table II, it is obvious that, whenever

TABLE II
FEATURE VECTORS USED IN OUR EXPERIMENTS. $\mathbf{f}_i^t(\mathbf{y}^t)$ (IR), $\mathbf{h}_i^t(\mathbf{y}^t)$ (IR): FEATURES USED FOR THE ASSOCIATION AND INTERACTION POTENTIALS FOR IKONOS AND RAPIDEYE SCENES, RESPECTIVELY. $\mathbf{f}_i^t(\mathbf{y}^t)$ (LS): FEATURES FOR ALL POTENTIALS FOR THE LANDSAT SCENES. TOP ROW: RANK OF THE RESPECTIVE FEATURE

rank	1	2	3	4	5	6
$\mathbf{f}_i^t(\mathbf{y}^t)$ (IR)	E^5_{nir}	V^5_{nir}	V^5_{hue}	E^5_{nir-r}	E^5_{grad}	ent^5
$\mathbf{h}_i^t(\mathbf{y}^t)$ (IR)	E^3_{nir}	V^3_{nir}	V^3_{hue}	E^3_{nir-r}	V^3_{ndvi}	E^3_{grad}
$\mathbf{f}_i^t(\mathbf{y}^t)$ (LS)	E^1_r	E^1_g	E^1_b	E^1_{nir}	E^1_{ndvi}	E^1_{rvi}

TABLE III
TRANSITION MATRIX $\mathbf{TM}^{HH}$ FOR TWO HR IMAGES

	$x_i^{t+1} = res$	$x_i^{t+1} = ind$	$x_i^{t+1} = for$	$x_i^{t+1} = crp$
$x_i^t = res$	1.00	0.05	0.05	0.05
$x_i^t = ind$	0.05	1.00	0.05	0.05
$x_i^t = for$	0.10	0.10	1.00	0.10
$x_i^t = crp$	0.20	0.20	0.10	1.00

multiscale features are considered, the features extracted in the coarsest scale are dominant. It is also evident that, for Ikonos and RapidEye, features from each of the three groups are selected.

C. Parameters and Test Setup

Our method requires a few parameters to be set by the user. They could be derived by a procedure such as cross-validation [29], but in our experiments, we used values that were found empirically. As far as the spatial interaction potentials are concerned, the weight β and the parameter η describing the impact of the data-dependent term have to be defined. We used $\beta = 0.9$ for IS_1 and $\beta = 0.7$ for both IS_2 and IS_3. The parameter η was set to $\eta = 8.0$ for IS_2 and to $\eta = 5.0$ for IS_3. These parameter values were selected as a tradeoff between classification accuracy and the preservation of small details in a series of test runs.

The second set of parameters is related to the temporal interaction potential. In order for the temporal and spatial interaction potentials to have a similar impact on the results, the weight γ in (7) should be about twice the weight β of the spatial interaction potential; we use $\gamma = 1.5$. We also need to define the values for the transition matrices IT^{tk}. The Ikonos and RapidEye images are considered to correspond to the same scale. Thus, we need only one transition matrix $\mathbf{TM}^{HH}$ between images of high resolution. The Landsat images have a different scale. Since there is only one Landsat scene per test site, we need a transition matrix $\mathbf{TM}^{HL}$ to relate HR images to Landsat images, but not a third one relating pairs or Landsat images. Determining the transition matrices by training would require an enormous amount of training data with a sufficient number of changes, which is not feasible in our case. The transition matrix $\mathbf{TM}^{HH}$ that we use in our experiments is shown in Table III. Its values were found empirically; note the dominant values at the main diagonal, which model the fact that the most likely event is that nothing changes. The values of the off-diagonal elements indicate that it is more likely for forest or cropland to be changed to a residential or industrial area than vice versa.

TABLE IV
TRANSITION MATRIX $\mathbf{TM}^{HL}$ FROM A HIGH-RESOLUTION IMAGE (EPOCH t) TO AN IMAGE AT MEDIUM RESOLUTION (EPOCH $t+1$)

	$x_i^{t+1} = bui$	$x_i^{t+1} = for$	$x_i^{t+1} = crp$
$x_i^t = res$	1.00	0.05	0.05
$x_i^t = ind$	1.00	0.05	0.05
$x_i^t = for$	T_{fc-b}	1.00	0.10
$x_i^t = crp$	T_{fc-b}	0.10	1.00

The transition matrix $\mathbf{TM}^{HL}$ relating HR data and Landsat images is shown in Table IV. Since the class structures at the two different scales are different, this matrix is not square. Its coefficients are set so that any class transition from res at the high resolution to another class at the low resolution is just as likely as a transition from ind to that class. The event that nothing changes is modeled as a situation where res and ind at epoch t correspond to the merged class bui in epoch $t+1$. Again, these values were found empirically. In order to assess the impact of these parameters on the classification results, particularly in the presence of actual change, we exemplarily will use different values for the value T_{fc-b} related to the transition from for or crp to bui. In most experiments involving multiscale data, we use $T_{fc-b} = 0.2$. For the experiments on change detection in multiscale data, we compare the results thus achieved to those achieved using $T_{fc-b} = 0.4$.

In our experiments, we divided the data into tiles of about 300 $\times$ 300 pixels and used approximately 10% of the tiles for training, whereas the other tiles were used for testing. For the evaluation, we used quality measures derived from the confusion matrices, i.e., *producer's* and *user's accuracy*, *quality*, and the *overall accuracy* [38], [39]. The confusion matrices were determined by a pixelwise comparison of the classification results with the reference.

V. EVALUATION

Here, we present the results of a quantitative evaluation of our methodology. If not otherwise noted, we use the parameter settings described in Section IV-C. We start with a comparison of the different models of the spatial interaction potentials in monotemporal classification (see Section V-A). These results, in particular those achieved for IS_1, also serve as a baseline for assessing the impact of our multitemporal model on the classification accuracy. The multitemporal model is evaluated for the case of using images of the same resolution in Section V-B and for images having different resolutions in Section V-C. We assess the potential of using multiscale data for detecting changes in Section V-D. Finally, a comparison to a state-of-the-art multitemporal classification technique is presented in Section V-E.

A. Monotemporal Classification

Table V gives the overall accuracy (OA) [%] for a monotemporal classification of all test scenes using maximum-likelihood classification (ML) and CRF-based classification using the spatial interaction models IS_1, IS_2, and IS_3. In most cases, the overall accuracy is better in Husum than it is in Herne.

TABLE V
OVERALL ACCURACY FOR MONOTEMPORAL CLASSIFICATION FOR ALL TEST SCENES. ML: MAXIMUM LIKELIHOOD CLASSIFICATION; IS_1, IS_2, IS_3: MONOTEMPORAL CRF CLASSIFICATION BASED ON THE RESPECTIVE SPATIAL INTERACTION POTENTIAL

Area	Scene	ML	IS_1	IS_2	IS_3
Herne	Ikonos 2005	74.2%	80.7%	79.8%	78.0%
	Ikonos 2007	73.2%	78.7%	78.9%	77.3%
	RapidEye 2009	64.4%	70.5%	68.6%	66.2%
	Landsat 2010	71.7%	76.0%	75.9%	76.1%
Husum	RapidEye 4/2009	79.9%	88.0%	86.4%	84.8%
	RapidEye 7/2009	80.6%	87.0%	85.7%	84.4%
	RapidEye 3/2010	79.8%	87.0%	86.5%	85.5%
	RapidEye 7/2010	86.1%	90.1%	89.8%	89.2%
	Landsat 7/2009	71.7%	81.8%	80.2%	79.4%

This is caused by the fact that Husum is dominated by class crp, which can be detected more reliably than other classes. In Herne, the poor ML results for the RapidEye image from 2009 (OA 64.4%) compared to the Landsat image are striking. Nevertheless, considering spatial context increases the overall accuracy by up to 6%; the largest increase occurs when the MRF model (IS_1) is used. This is somewhat surprising given the tendency of MRF to oversmooth the results. However, in all cases, except the 2009 RapidEye scene, the difference in the performance of the three spatial interaction terms is considerably smaller than the difference between any context-based methods and the ML classification, and for the Landsat scene, the maximum difference in OA between the spatial interaction models of 0.2% is negligible. Similar observations can be made for Husum: the overall accuracy is improved even more (up to 10%), the most striking improvement occurring for the Landsat scene. Again, for the Landsat scene, the performance of the three interaction models is very similar, but also for the other scenes, the impact of using spatial context is larger than the differences between the interaction models.

As stated earlier, the MRF setting (IS_1) achieves the best overall accuracy for all test scenes. However, the question arises whether this is the best criterion for assessing the quality of the classification results. The left column in Fig. 4 shows the results of CRF-based classification using the three models for the interaction potentials. A visual inspection gives the impression of rather strong oversmoothing in the results based on MRF [see Fig. 4(a)]. In order to characterize this impression by a numerical evaluation, we focus on the detectability of roads. We generated a reference of rural roads (i.e., roads outside the settlement areas). Roads, in general, are assumed to belong to class res. Fig. 4(a) gives the impression that more rural roads are classified as crp than in Fig. 4(c) and (e). In all classification results, roads are largely assigned to the classes res and ind; as we shall see in Section V-B, these classes cannot be separated very well. However, if we merge these classes to a joint class bui and compare the classification results to the reference, we see that the producer's accuracy of the detected roads (PA, i.e., the detection rate of road pixels) is increased by using the contrast-sensitive versions of the interaction potentials. This is shown in the right column in Fig. 4, where detected road pixels are displayed in green, whereas red indicates missed road pixels. The ranking of the three interaction terms is inversed, with the model based on IS_3 (which encourages a class change if the feature vectors are different) delivering 95.5% of the road pixels [see Fig. 4(f)], i.e., an increase of 6% over the MRF model [IS_1, Fig. 4(b)]. In Husum, the corresponding detection rates are 67.4% (IS_1), 83.6% (IS_2), and 88.8% (IS_3), respectively, showing a similar tendency although at a slightly lower general level. This preservation of fine detail comes at the cost of false positives in other areas, which, however, might be removed by a subsequent high-level analysis, e.g., based on criteria such as minimum area [40].

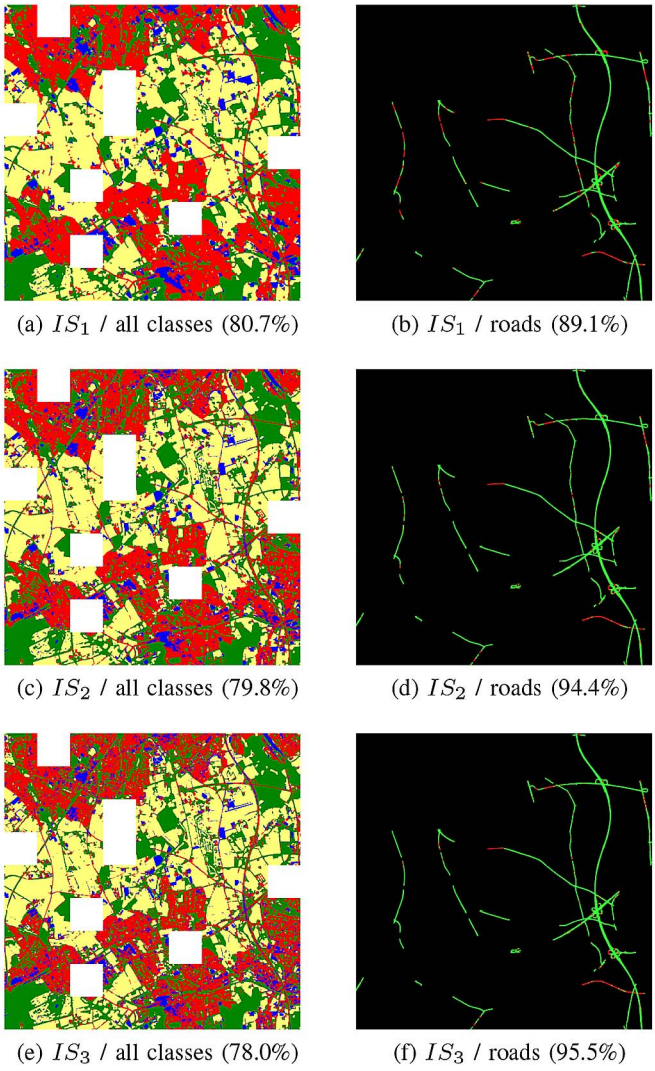
B. Multitemporal and Monoscale Classification

In the second set of experiments, we tested our multitemporal classification method using only images of the same resolution. In Herne, we thus combined the two Ikonos scenes, whereas in Husum, we simultaneously classified the four RapidEye scenes. We used the transition matrix $\mathbf{TM}^{HH}$ from Table III for the temporal interactions. For each data set, the classification was carried out three times, varying the model for the spatial interaction potentials. The overall accuracies of the classification results are shown in Table VI.

Comparing the results in Table VI to those of monotemporal classification in Table V, one can observe that, in all cases, the overall accuracy is improved. The improvement in overall accuracy is on the order of 2% in most cases, but can be up to 5.5% (Husum, RapidEye 7/2009). It is, in general, relatively small for the variant based on IS_1 but larger for the other ones, so that as a consequence, the differences between the results obtained by the three different interaction models nearly vanish (max. 0.4%). The differences between the scenes also become smaller (e.g., 0.2% in OA for Ikonos 2005 in Herne for IS_1 versus 2.0% in the monotemporal case for IS_1). The information from the scenes that could be classified more accurately in the monotemporal scenario propagates to the other scenes. One could assume that the interactions between the individual epochs could decrease the accuracies of the scenes achieving the best monotemporal results, but this is not the case. The propagation of information works in both ways: the scenes originally classified more accurately in the monotemporal setting also achieve a marginally better OA in the multitemporal scenario.

In order to highlight the effect of using multitemporal classification, we determined the producer's accuracy, the user's accuracy, and the quality of the results achieved for the four classes for the RapidEye 7/2009 scene from Husum, using ML classification and both monotemporal and multitemporal CRF-based classification with IS_3 (cf. Table VII; the results for the other scenes are in a similar range, but not reported here for lack of space). A small subset of the corresponding classification results is shown in Fig. 5.

The results in Table VII show that, in all variants, the class crp is the one that is classified most accurately. It provides the largest contribution to the overall accuracy in Husum because it is the most dominant land cover. Using ML classification, all the other classes either have a rather low user's accuracy, in particular class ind, where the classification produces 85% false positives. Obviously, this class cannot be distinguished very well from the others, given the observed feature vectors.



(a) IS_1 / all classes (80.7%) (b) IS_1 / roads (89.1%)
(c) IS_2 / all classes (79.8%) (d) IS_2 / roads (94.4%)
(e) IS_3 / all classes (78.0%) (f) IS_3 / roads (95.5%)

Fig. 4. (Left) Results of the monotemporal classification for the Herne Ikonos 2005 image using the three different interaction potentials. Colors: yellow: crp; red: res; blue: ind; green: for; white: training areas. (Right) Assessment of the detection rate for rural roads. Green: correctly detected road pixels; red: missed road pixels. The numbers in parentheses are overall accuracies (left column) and PA for roads (right column), respectively.

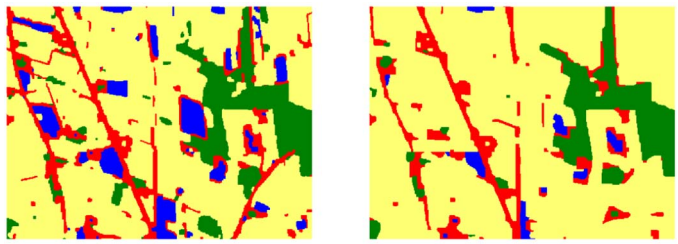
TABLE VI
OVERALL ACCURACY FOR MULTITEMPORAL CLASSIFICATION OF IMAGES HAVING THE SAME RESOLUTION. IS_1, IS_2, IS_3: MULTITEMPORAL CRF CLASSIFICATION BASED ON THE RESPECTIVE SPATIAL INTERACTION POTENTIAL. HMM: HIDDEN MARKOV MODEL [28]. THIS COLUMN IS DISCUSSED IN SECTION V-E

Area	Scene	IS_1	IS_2	IS_3	HMM
Herne	Ikonos 2005	80.9%	81.3%	80.9%	77.0%
	Ikonos 2007	80.7%	80.9%	80.4%	76.2%
Husum	RapidEye 4/2009	89.5%	89.8%	89.6%	85.3%
	RapidEye 7/2009	89.9%	90.1%	89.9%	85.5%
	RapidEye 3/2010	90.1%	90.5%	90.4%	86.5%
	RapidEye 7/2010	90.1%	90.4%	90.3%	87.0%

Considering context in the monotemporal setting helps, the most obvious improvement occurring with the user's accuracy of class for (14.2%). This is mainly caused by a smaller number of crp pixels that are classified as for. As shown in Section V-A, the main contribution of the interaction terms is a certain amount of smoothing of the results, and the smaller amount of confusion between for and crp is the major impact of smoothing, given that the largest portion of the scene is covered by crp. In the multitemporal scenario, we can observe another increase of the user's accuracy of for, but this time, the user's accuracy of ind is also increased by about 10%, largely due to a reduction of crp areas erroneously classified as ind (cf. the blue areas in Fig. 5). Nevertheless, it becomes clear that the separation of class ind remains a problem: the quality indexes in Table VII indicate that the number of false positives for this class (in equal proportion pixels that actually correspond to res and to crp) is too large. The classification accuracy for this class could be improved by using different models for the association potentials (e.g., [41]), but this is beyond the scope of this paper.

TABLE VII
COMPARISON OF PRODUCER'S ACCURACY (PA), USER'S ACCURACY (UA), AND QUALITY (Q) OF THE RESULTS FOR THE RAPIDEYE SCENE OF HUSUM FROM 7/2009

Model		res	ind	for	crp
ML	PA	72.3%	47.0%	85.9%	81.6%
	UA	48.3%	14.8%	52.4%	97.9%
	Q	41.0%	12.7%	48.2%	80.2%
CRF $Mono$ (IS_3)	PA	79.2%	51.4%	86.5%	87.0%
	UA	56.9%	17.0%	66.6%	98.3%
	Q	49.5%	14.6%	60.3%	85.7%
CRF $Multi$ (IS_3)	PA	74.9%	51.7%	84.4%	93.2%
	UA	63.7%	27.6%	80.4%	97.3%
	Q	52.5%	21.9%	70.0%	90.9%



(a) IS_3, monotemporal (b) IS_3, multitemporal

Fig. 5. Results for monotemporal and multitemporal classification using IS_3 for a subset of the Husum RapidEye 7/2009 scene. For the color code, cf. Fig. 4.

C. Multitemporal and Multiscale Classification

In order to evaluate the capabilities of our method to deal with multitemporal images of different resolutions, we applied it to the three scenes Ikonos 2005, RapidEye 2009, and Landsat 2010 of Herne and to two RapidEye scenes (4/2009 and 7/2009) and the Landsat scenes of Husum. The transition matrix $\mathbf{TM}^{HH}$ in Table III could be used for the interactions between the RapidEye and Ikonos images. For the temporal interactions with the Landsat scenes, we used $\mathbf{TM}^{HL}$ in Table IV, using $T_{fc-b} = 0.2$. Again, we computed three different variants, using the spatial interaction models IS_1, IS_2, and IS_3. Note that no relevant land-cover changes could be found in both test areas over the observed period. The overall accuracies of the classification results are shown in Table VIII. Some of the

TABLE VIII
OVERALL ACCURACY FOR MULTITEMPORAL CLASSIFICATION OF IMAGES HAVING DIFFERENT RESOLUTIONS. IS_1, IS_2, IS_3: MULTITEMPORAL CRF CLASSIFICATION BASED ON THE RESPECTIVE SPATIAL INTERACTION POTENTIAL. HMM: HIDDEN MARKOV MODEL [28]. THIS COLUMN IS DISCUSSED IN SECTION V-E

Area	Scene	IS_1	IS_2	IS_3	HMM
Herne	Ikonos 2005	80.8%	81.5%	81.3%	77.3%
	RapidEye 2009	79.8%	80.6%	80.2%	75.9%
	Landsat 2010	80.1%	80.4%	80.3%	79.2%
Husum	RapidEye 4/2009	89.7%	90.1%	89.7%	85.3%
	RapidEye 7/2009	89.7%	89.4%	88.9%	84.5%
	Landsat 7/2009	89.2%	88.2%	88.7%	83.1%

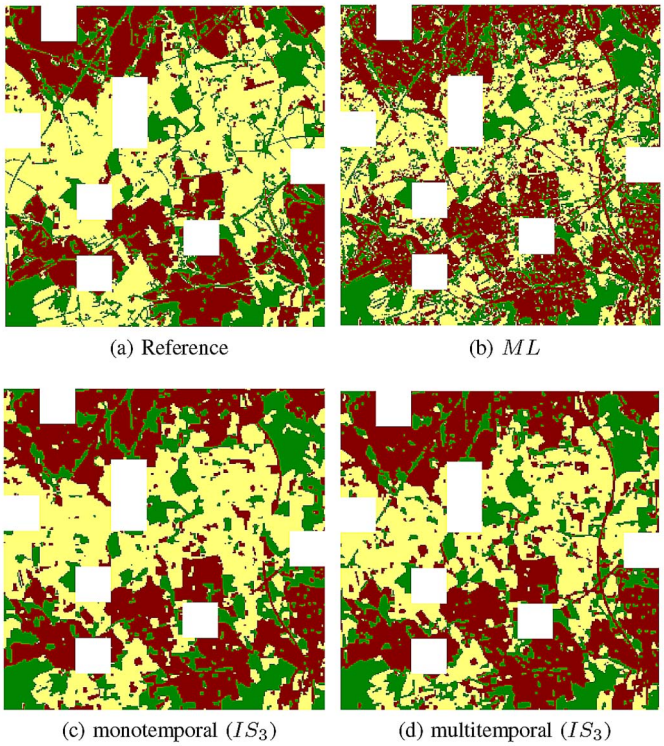


Fig. 6. Reference and classification results for the Landsat scene of Herne using different classification models. ML: Maximum likelihood. Colors: yellow: crp; brown: bui; green: for. The white rectangles are training areas.

results for the Landsat scenes in Herne and Husum are shown in Figs. 6 and 7, respectively.

Compared to the results of the multitemporal classification using images of the same scale in Table VI, the overall accuracies of the HR images (Ikonos 2005 in Herne and the two RapidEye scenes in Husum) are very similar. In some cases, they are larger for the monoscale case (e.g., RapidEye 2007 in Husum), but the difference is never larger than 1%. The observation made in the monoscale case that the overall accuracies for the three models for the interaction potentials are very similar also applies to the results in Table VIII, the maximum difference being 1%. The main effect of the multiscale and multitemporal classification can be seen by comparing the results for the Herne RapidEye and the Landsat scenes in Table VIII with those achieved in monotemporal classification (cf. Table V). The multitemporal setting achieves an increase in the overall accuracy of the Landsat scenes by

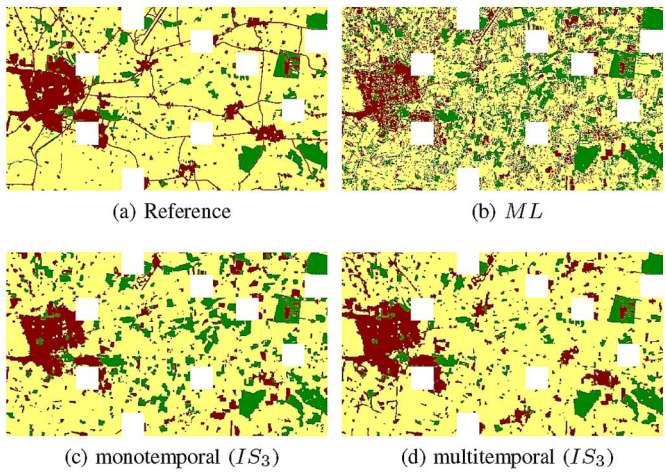


Fig. 7. Reference and classification results for the Landsat scene of Husum using different classification models. ML: Maximum likelihood. The color code is the same as in Fig. 6.

TABLE IX
PRODUCER'S ACCURACY (PA), USER'S ACCURACY (UA), AND QUALITY (Q) OF THE RESULTS FOR THE LANDSAT SCENE OF HUSUM FROM 7/2009

Model		bui	for	crp
ML	PA	49.3%	89.7%	74.6%
	UA	51.6%	32.1%	89.6%
	Q	33.7%	31.0%	68.7%
CRF $Mono$ (IS_3)	PA	51.2%	87.8%	84.3%
	UA	64.8%	42.6%	89.9%
	Q	40.1%	40.2%	77.0%
CRF $Multi$ (IS_3)	PA	70.0%	89.0%	92.4%
	UA	71.2%	72.2%	94.2%
	Q	54.6%	66.3%	87.4%

about 5% and lifts the quality of the results to the level that could be achieved for the HR images. Even more strikingly, in the Herne data set, the overall accuracy of the RapidEye scene is improved by 10%–15%, depending on the type of interaction potential that is used. Obviously, the radiometric information of the RapidEye scene and the spatial interaction model were not sufficient for a classification at the same quality level as for the Ikonos image. The situation could be improved by adding the temporal component. Again, the improvement did not only occur in one direction: in particular for the interaction models IS_2 and IS_3, the temporal interactions also improved the quality of the classification of the HR images by up to 4% (Husum, RapidEye 4/2009, IS_3). Figs. 6 and 7 show the effects of the different types of interaction models on the classification results for the Landsat scenes. ML classification [see Figs. 6(b) and 7(b)] leads to a very noisy result in comparison to the reference [see Figs. 6(a) and 7(a)]. The spatial interactions yield a smoother version of the results, perhaps a bit oversmoothed [see Figs. 6(c) and 7(c)]. The temporal interactions help to alleviate the oversmoothing to a certain degree [see Figs. 6(d) and 7(d)].

Table IX shows the producer's accuracy, the user's accuracy, and the quality of the results for the Landsat scene in Husum (the numbers for Herne are in a similar range). If local context is not considered (ML), the user's accuracy for for is very

low, which is largely due to a similar spectral appearance of for and crp. The classification performance of bui is only about 50%, in both user's and producer's accuracy. Fig. 7(b) shows many crp pixels inside the large settlement in the western part of the scene, and many false-positive bui pixels in the large agricultural area. The situation is somewhat improved if spatial context is taken into account: both the user's accuracy for for and the producer's accuracy for crp are increased by about 10%, which is largely caused by a reduced amount of confusion between these two classes. The smoothing effect of the CRF-based classification also leads to an increased user's accuracy for bui, in accordance with the smaller number of spurious bui pixels in Fig. 7(c). However, the evidence in the data for bui does not seem to be strong enough to improve the producer's accuracy by more than 2%, and as a consequence, nearly 50% of the bui pixels remain undetected. Similarly, the number of false-positive for areas is still nearly 60%. Using the temporal interactions does not only further reduce the confusion between for and crp, but also and most notably between bui and crp. As a consequence, the user's accuracy for for is increased by 30%, and both the user's and the producer's accuracy of bui are raised to values slightly larger than 70%. Class bui is rather heterogeneous in the Landsat imagery, which is probably the reason why this class receives a particularly high benefit from the multitemporal classification. However, the improvement is also striking for the two other classes in the scene. Our experiment clearly shows that our multitemporal and multiscale CRF model can help to raise the classification accuracy to acceptable levels for medium-resolution imagery.

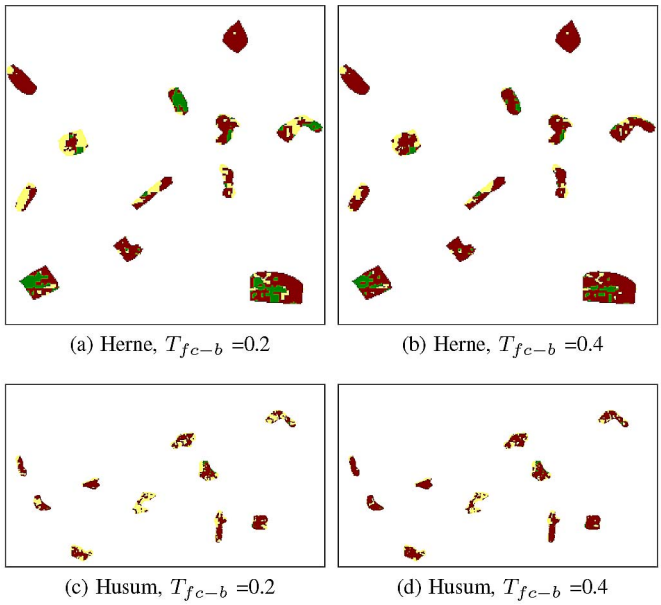
D. Change Detection

The results described in the previous section were based on multitemporal data of areas without changes in land cover between the individual epochs. The model used for the temporal transition matrices was that change was a very unlikely event, with the most likely change (from for or crp to bui) being five times less likely than the event that no change occurs. This raises the question how our method will behave if there are real changes in the scenes. Would the information from the earlier epochs (when HR images are available) just propagate to the later epoch (with the low-resolution images), overriding any indicator for a change?

In order to answer this question, we simulated changes in the Landsat images of Herne and Husum, copying image blocks from the settlement areas into regions formerly covered by forest or cropland. These are the most typical land-cover changes to be expected in Central Europe, given our set of classes. Then, we repeated the multitemporal analysis in the way described in Section V-C, but using two different values for the parameter T_{fc-b} related to the likelihood of a change from for or crp to bui. The value $T_{fc-b} = 0.2$ was also used in the experiments in Section V-C and corresponds to a setting where a change from for or crp to bui is relatively unlikely; we compare the results thus obtained to those achieved using a more loose setting of $T_{fc-b} = 0.4$, making a transition twice as likely as in the first case. The overall accuracies for these experiments, again using different spatial interaction potentials, are shown in Table X.

TABLE X
OVERALL ACCURACY FOR MULTITEMPORAL CLASSIFICATION OF IMAGES HAVING DIFFERENT RESOLUTIONS WITH ARTIFICIAL CHANGES IN THE LANDSAT SCENES. IS_1, IS_2, IS_3: MODEL FOR THE SPATIAL INTERACTION POTENTIALS USED IN THE RESPECTIVE EXPERIMENT. T_c: THE PARAMETER T_{fc-b} RELATED TO THE LIKELIHOOD OF A CHANGE FROM for OR crp TO bui BETWEEN TWO EPOCHS. HMM: HIDDEN MARKOV MODEL [28]. THIS COLUMN IS DISCUSSED IN SECTION V-E

Area	Scene	T_c	IS_1	IS_2	IS_3	HMM
Herne	Ikonos 2005	0.2	80.7%	81.3%	81.1%	76.5%
		0.4	80.7%	81.3%	81.0%	76.2%
	RapidEye 2009	0.2	79.6%	80.3%	80.0%	74.2%
		0.4	79.8%	80.4%	80.0%	73.5%
	Landsat 2010	0.2	78.7%	79.4%	79.1%	74.0%
		0.4	79.0%	79.3%	79.5%	73.1%
Husum	RapidEye 4/2009	0.2	89.7%	90.0%	89.7%	85.4%
		0.4	89.7%	90.1%	89.6%	84.9%
	RapidEye 7/2009	0.2	89.7%	89.2%	88.8%	84.7%
		0.4	89.7%	89.2%	88.8%	83.6%
	Landsat 7/2009	0.2	87.1%	86.8%	87.2%	82.0%
		0.4	87.4%	86.1%	87.0%	81.3%



(a) Herne, T_{fc-b} =0.2 (b) Herne, T_{fc-b} =0.4

(c) Husum, T_{fc-b} =0.2 (d) Husum, T_{fc-b} =0.4

Fig. 8. Results of the multitemporal classification of the changed areas of the Landsat scenes for using IS_3 and two different values for T_{fc-b}. Colors: yellow: crp; brown: bui; green: for.

The classification results for the changed areas in the Landsat scenes are displayed in Fig. 8. Table XI shows the percentage of bui pixels in the changed areas and, thus, the producer's accuracy of change detection. In order to obtain this table, we also applied monotemporal classification to the Landsat scenes of Herne and Husum. Table XII exemplarily shows the user's and producer's accuracy, as well as the *quality* of the results, for the three classes discerned in the Landsat scene in Husum.

There is only a small difference in overall accuracy in Tables VIII and X. The differences between the quality indexes that were achieved using different values for T_{fc-b} is also relatively small. For Herne, the difference in overall accuracy is below 0.5% in all cases; for Husum, the maximum difference is 0.7% (Landsat, IS_2). This difference is much smaller than the amount of change in the Landsat scenes: in Herne and

TABLE XI
PERCENTAGE OF *bui* PIXELS DETECTED IN THE SIMULATED AREAS OF CHANGE IN THE LANDSAT SCENES. ML: MAXIMUM-LIKELIHOOD CLASSIFICATION

Herne							
T_{fc-b}	multitemporal			monotemporal			
	IS_1	IS_2	IS_3	ML	IS_1	IS_2	IS_3
0.2	59%	63%	61%	75%	80%	83%	82%
0.4	75%	79%	77%				
Husum							
T_{fc-b}	multitemporal			monotemporal			
	IS_1	IS_2	IS_3	ML	IS_1	IS_2	IS_3
0.2	43%	63%	58%	75%	72%	81%	81%
0.4	66%	74%	71%				

TABLE XII
PRODUCER'S ACCURACY (PA), USER'S ACCURACY (UA), AND QUALITY (Q) OF THE RESULTS FOR THE LANDSAT SCENE OF HUSUM WITH ARTIFICIAL CHANGES FOR TWO VALUES OF T_{fc-b}. THE SPATIAL INTERACTION MODEL WAS IS_3

T_{fc-b}		*bui*	*for*	*crp*
0.2	PA	67.5%	86.8%	92.5%
	UA	74.5%	75.0%	91.7%
	Q	54.8%	67.4%	85.4%
0.4	PA	71.7%	86.8%	91.1%
	UA	71.4%	75.4%	92.6%
	Q	55.7%	67.6%	85.0%

Husum, about 7.4% and 4.2% of the pixels are affected by a simulated change, respectively. However, looking at Fig. 8, it is shown that not all changed areas are correctly assigned to class *bui* in the multitemporal case, and it would seem that the amount of change that is correctly detected does depend on the value of the parameter T_{fc-b}. This is confirmed by the quality indexes in Table XI. If we apply ML classification, 75% of the pixels in the changed areas are correctly classified as *bui*, i.e., 75% of the changes are correctly detected. Additionally, considering context in the classification, in general, increases the number of detected *bui* pixels by 5%–8%, with the interesting exception of the MRF-based setting in Husum (IS_1), where it is reduced due to oversmoothing effects. In the multitemporal setting, the number of correct *bui* pixels is, in general, lower than in monotemporal classification, which indicates that the propagation of information from previous epochs does, in fact, oversmooth the results to a certain degree. However, this has to be seen in relation to the results achieved for monotemporal classification, which could be seen as an upper limit for what can be achieved in areas where, due to a change, the information from previous epochs is of no value for determining the correct class label. In any case, the degree of oversmoothing depends on the parameter T_{fc-b}. For the more conservative setting $T_{fc-b} = 0.2$, only 43% of *bui* pixels are correctly detected in Husum, if the interaction potential IS_1 is used. However, as we have seen earlier, this is also the setting that performs worst in the monotemporal case. In all the other cases, at least 58% of the changes are correctly detected. For the setting $T_{fc-b} = 0.4$, the percentage of detected changes is still lower in the multitemporal case than in the monotemporal case, but except for the variant based on the interaction potential IS_1 in Husum, more than 70% of the *bui* pixels could correctly be detected in the changed areas. We interpret the relatively poor performance of IS_1 in Husum, in the multitemporal setting, as an indicator that strong spatial smoothing in combination with the temporal smoothing effects may hamper change detection, so that the contrast-sensitive versions of the interaction potentials seem to be preferable. Table XII shows that the producer's and user's accuracies per class are at a similar level as those for the case without simulated changes (see Table IX, version multitemporal CRF). The PA for *bui* is 2.5% lower for $T_{fc-b} = 0.2$, which corresponds well with the observation that 58% of the changed *bui* pixels are detected in that scene and using this configuration; the quality for *bui* is, nevertheless, slightly higher in Table XII than in the corresponding version in Table IX because there are fewer false positives. Increasing the value of T_{fc-b} to 0.4 results in an increase of PA by about 4%, although at the cost of an additional 3% of false-positive *bui* pixels. Class *for* achieves a very similar tradeoff between false positives and false negatives, as expressed by the *quality* in Tables IX and XII in the related experiments, whereas for *crp*, there is a slightly lower quality for the case with simulated changes (difference in quality of about 2.0%–2.5% between Tables IX and XII), which is caused by an increased number of false-positive building pixels erroneously classified as *crp*.

Up to this point, we have only considered area-based quality metrics. We also carried out an object-based analysis. For that purpose, we consider a changed region (i.e., a connected region of changed pixels) to be correctly detected, if the majority of its pixels was detected as *bui*. In this case, we see a clear improvement of the results when using $T_{fc-b} = 0.4$ rather than $T_{fc-b} = 0.2$. In Herne, using $T_{fc-b} = 0.2$ results in 9 detected regions of change (out of 12), whereas all but one changed regions could be detected with $T_{fc-b} = 0.4$. Similarly, when using $T_{fc-b} = 0.4$, we can detect 9 out of 10 changed regions in Husum, whereas that number was reduced to 6, 7, and 8 for IS_1, IS_2, and IS_3, respectively, when using $T_{fc-b} = 0.2$. Obviously, if a significant amount of change is present in a scene, it is advantageous to use a looser setting for the transition matrices. We think that the experiments presented here give a clear indication that medium-resolution images can be very useful to give hints for changes in a scenario, where the time between two HR acquisitions is bridged by data of lower resolution to reduce costs. However, currently, this process would have to be embedded in a semiautomatic scenario, as in [40], because the error rate in the classification process would indicate too high a number of land-cover changes for a fully automated process.

E. Comparison to the State-of-the-Art Method

Here, we compare our results to a state-of-the-art method in multitemporal classification, i.e., the method described in [28]. This method is based on a hidden Markov model, describing temporal dependencies among phenological stages of different crop types by a first-order Markov chain. One of its objectives is to determine the sequence of stages that best explains the features measured at multiple epochs. Essentially, the same method has been used to model the temporal behavior of pixels in our problem domain. The difference lies mostly on the image primitives being classified, the interpretation of some model

elements, and on how parameters have been set. Whereas, in [28], the image primitives to be classified were segments, here, they are defined to correspond to individual pixels. That is, each set of pixels at the same spatial position but acquired at different epochs is classified individually based on a HMM model. The phenological stages of the original method correspond to the class labels of the epochs for which images are available. In [28], the transition probabilities are defined based on a mixture of expert knowledge and training. We do not have training samples for class transitions, so that the transition matrices $\mathbf{TM}^{HH}$ and $\mathbf{TM}^{HL}$ in Tables III and IV, respectively, are used here as well; they were just normalized so that the sum of elements along each row is one. Similar to our method, the likelihood linking the features observed at epoch t, i.e., $\mathbf{f}_i^t(\mathbf{y}^t)$, to the class labels x_i^t is modeled by a multivariate Gaussian [i.e., (3)]. The method in [28] can cope with different class structures at different epochs, but it is not able to cope with images having different resolutions. As a consequence, we had to resample the Landsat images in our experiments to the same resolution as the HR images. Bilinear resampling was applied in this context. The feature vectors used in classification are those defined in Section IV-B (see Table II), with the exception that the Landsat features are determined for the pixels of the resampled image (i.e., after resampling). We refer to this adapted version of the method described in [28] as HMM (hidden Markov model) in the remainder of this section.

First, we applied the HMM method for multitemporal and monoscale classification, using the same configuration of images as in Section V-B. The overall accuracies achieved are shown in column HMM inTable VI. Considering temporal interactions using the HMM model increases the overall accuracy, by about 3% compared to a standard ML classification in Herne and by about 5%–6% for all scenes, except the last one, in Husum (cf. column ML in Table V). However, comparing the results in column HMM to those achieved by applying our method in Table VI, it becomes obvious that considering the spatial interaction results in another increase of overall accuracy by a margin of 3%–5.5% in all scenes. The HMM model achieves results whose accuracy is somewhere in the middle between a simple ML classification and our approach.

We also applied the HMM model to multitemporal classification combining images having different original resolutions, using the same scenes as for the experiments in Section V-C, i.e., combining Ikonos and RapidEye images with the Landsat scenes without synthetic changes. We changed the GSDs of the Landsat scenes to 4 m in Herne and to 5 m in case of Husum. The evaluation for the Landsat images was also carried out at the higher resolution. The overall accuracies achieved by the HMM model for all scenes in both test sites are shown in column HMM in Table VIII. We can observe a similar trend as in the monoscale case: the accuracies achieved by the HMM model are somewhere in the middle between those achieved by a simple ML classification (cf. column ML in Table V) and those obtained using our model (cf. Table VIII). The improvement caused by the HMM model over ML is on the order of 4%–6% for most scenes, but can reach 11%–12% (Herne/RapidEye and Husum/Landsat). In Husum, the accuracies are all raised to a similar level, whereas in Herne, the overall accuracy for the Landsat scene is better by 5% than the one achieved for the RapidEye scene. Considering the spatial interactions in our model adds another 1%–4% and 4%–6% in overall accuracy in Herne and Husum, respectively. We think that the improvement is higher in Husum than in Herne because land cover is more homogeneous in Husum, so that smoothing by considering the spatial interactions has a higher impact.

Finally, we repeated the experiments in Section V-D, replacing the original Landsat images by those with simulated changes. As in the previous example, we changed the GSD of the Landsat scenes for the HMM method to be applicable. We also carried out two tests, using two different values for the parameter T_{fc-b}. The resulting overall accuracies are shown in column HMM in Table X. Again, we observe that the accuracy of the results of the HMM method does not achieve the level obtained by our method, with the margin being on the order of 3%–5%. The improvement for the Landsat scenes is about 5% both in Herne and in Husum, which shows that our method does not lose accuracy compared to the state-of-the-art method in the presence of land-cover changes. Comparing the results achieved for different settings of T_{fc-b} in column HMM in Table X, it would seem that loosening the smoothing effects of the temporal interactions by setting $T_{fc-b} = 0.4$ has a slightly negative effect on the classification, but the maximum difference is only 1.1% (Husum, RapidEye 7/2009). We think that the comparison carried out here clearly shows the improvement over state-of-the-art multitemporal classification that can be achieved by incorporating spatial interactions and the option to classify images having different resolutions simultaneously, as proposed in this paper.

VI. Conclusion

We have presented a supervised method for multitemporal and multiscale classification of remote sensing images that also considers local spatial context. It is an extension of the concept of CRF by multitemporal terms, with the latter being modeled by transition matrices related to the probabilities of certain changes between the classes. In a multiscale setting, the method can deal with different class structures for images having different spatial resolutions. The data terms of the CRF were determined by training, whereas the parameters of the (spatial and temporal) interaction terms were found empirically. We compared the results for different models of the spatial interaction potentials. In a monotemporal setting, we could show that the smoothing effects introduced by the spatial interactions increase the classification accuracy considerably, most notably if using a non-data-dependent interaction term, although this may come at the cost of loosing relevant details. The multitemporal setting improved the classification accuracy for all images involved, but the most striking improvement could be observed for medium-resolution images. Although this improvement could, to a certain degree, be attributed to a smoothing over time achieved by incorporating the HR images of previous epochs, it could be shown based on simulated images that, nevertheless, the majority of the changes could be detected in the medium-resolution images. A comparison to a state-of-the-art method for multitemporal classification

has shown that a considerable portion of the improvement in classification accuracy can be attributed to the consideration of spatial interactions in our model. We think that, using this kind of analysis, it may become possible to use HR images for change detection at a lower frequency, bridging the time in between two HR acquisitions by medium-resolution images and thus reducing the overall costs of data acquisition and, therefore, database updating.

The pool of features used in this work is rather limited, particularly for the medium-resolution images. In the future, an expanded feature pool could help to further improve the classification results. The Gaussian model used for the association potentials of the CRF is rather simplistic and should be replaced by more sophisticated ones, e.g., by state-of-the-art discriminative classifiers such as random forests [2]. The interaction terms selected for our work are also relatively simple, depending only on the Euclidean distance of the feature vectors at neighboring sites. In the context of airborne laser scanning, more complex interaction terms have been shown to improve the classification accuracy for classes that do not occur very frequently [41]; this observation is still to be verified in the context of satellite imagery. In this paper, we also neglected the possibility of data-dependent temporal interaction potentials, although our framework is suited to such an extension in principle. Finally, the incorporation of a geospatial database into the model, which would be related to the problem of change detection in the context of map updating more directly, but could also provide a strong prior for all subsequent epochs, still remains to be investigated.

ACKNOWLEDGMENT

Our CRF classification is based on the undirected graphical models toolbox by Mark Schmidt: http://people.cs.ubc.ca/~schmidtm/Software/UGM.html (visited 19 May 2014).

REFERENCES

[1] S. Kumar and M. Hebert, "Discriminative random fields," *Int. J. Comput. Vis.*, vol. 68, no. 2, pp. 179–201, Jun. 2006.
[2] K. Schindler, "An overview and comparison of smooth labeling methods for land-cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4534–4545, Nov. 2012.
[3] A. Singh, "Digital change detection techniques using remotely-sensed data," *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, Jun. 1989.
[4] D. Li, "Remotely sensed images and GIS data fusion for automatic change detection," *Int. J. Image Data Fusion*, vol. 1, no. 1, pp. 99–108, Mar. 2010.
[5] D. Lu, P. Mausel, E. Brondizio, and E. Moran, "Change detection techniques," *Int. J. Remote Sens.*, vol. 25, no. 12, pp. 2365–2401, Jun. 2004.
[6] R. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: A systematic survey," *IEEE Trans. Image Process.*, vol. 14, no. 3, pp. 294–307, Mar. 2005.
[7] J. Gong, H. Sui, G. Ma, and Q. Zhou, "A review of multi-temporal remote sensing data change detection algorithms," *Int. Archives Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 37, pt. B7, pp. 757–762, 2008.
[8] J. F. Mas, "Monitoring land-cover changes: A comparison of change detection techniques," *Int. J. Remote Sens.*, vol. 20, no. 1, pp. 139–152, Jan. 1999.
[9] R. Q. Feitosa, G. A. O. P. Costa, G. L. A. Mota, K. Pakzad, and M. C. O. Costa, "Cascade multitemporal classification based on fuzzy Markov chains," *ISPRS J. Photogramm. Remote Sens.*, vol. 64, no. 2, pp. 159–170, Mar. 2009.
[10] L. Bruzzone, R. Cossu, and G. Vernazza, "Detection of land-cover transitions by combining multidate classifiers," *Pattern Recognit. Lett.*, vol. 25, no. 13, pp. 1491–1500, Oct. 2004.
[11] G. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution and Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.
[12] F. Melgani and S. B. Serpico, "A Markov random field approach to spatio-temporal contextual image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 11, pp. 2478–2487, Nov. 2003.
[13] G. Moser, E. Angiati, and S. B. Serpico, "A contextual multiscale unsupervised method for change detection with multitemporal remote-sensing images," in *Proc. 9th Conf. Intell. Syst. Des. Appl.*, 2009, pp. 572–577.
[14] P. Zhong and R. Wang, "A multiple conditional random fields ensemble model for urban area detection in remote sensing optical images," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 3978–3988, Dec. 2007.
[15] W.-L. Lu, K. P. Murphy, J. J. Little, A. Sheffer, and H. Fu, "A hybrid conditional random field for estimating the underlying ground surface from airborne lidar data," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2913–2922, Aug. 2009.
[16] R. Roscher, B. Waske, and W. Förstner, "Kernel discriminative random fields for land cover classification," in *Proc. 6th IAPR TC 7 Workshop Pattern Recog. Remote Sens.*, 2010, pp. 1–5, [CD-ROM].
[17] J. Wegner, U. Sörgel, and B. Rosenhahn, "Segment-based building detection with conditional random fields," in *Proc. 6th IEEE/GRSS/ISPRS Joint Urban Remote Sens. Event*, 2011, pp. 205–208.
[18] T. Hoberg, F. Rottensteiner, and C. Heipke, "Classification of multitemporal remote sensing data using conditional random fields," in *Proc. 6th IAPR TC 7 Workshop Pattern Recog. Remote Sens.*, 2010, pp. 1–4, [CD-ROM].
[19] C. Benedek and T. Szirányi, "Change detection in optical aerial images by a multi-layer conditional mixed Markov model," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 10, pp. 3416–3430, Oct. 2009.
[20] Z. Kato, M. Berthod, and J. Zerubia, "Multiscale Markov random field models for parallel image classification," in *Proc. 4th IEEE ICCV*, 1993, pp. 253–257.
[21] A. S. Wilsky, "Multiresolution Markov models for signal and image processing," *Proc. IEEE*, vol. 90, no. 8, pp. 1396–1458, Aug. 2002.
[22] C. A. Bouman and M. Shapiro, "A multiscale random field for Bayesian image segmentation," *IEEE Trans. Image Process.*, vol. 3, no. 2, pp. 162–177, Mar. 1994.
[23] J. Kersten, M. Gähler, and S. Voigt, "A general framework for fast and interactive classification of optical VHR satellite imagery using hierarchical and planar Markov random fields," *Photogrammetrie-Fernerkundung-Geoinformation*, vol. 6, pp. 439–449, Dec. 2010.
[24] M. J. Choi, V. Chandrasekaran, D. M. Malioutov, J. K. Johnson, and A. S. Willsky, "Multiscale stochastic modeling for tractable inference and data assimilation," *Comput. Methods Appl. Mech. Eng.*, vol. 197, no. 43/44, pp. 3492–3515, Aug. 2008.
[25] T. Hoberg, F. Rottensteiner, and C. Heipke, "Classification of multitemporal remote sensing data of different resolution using conditional random fields," in *Proc. 1st IEEE/ISPRS Workshop Comput. Vis. Remote Sens. Environ.*, Barcelona, Spain, 2011, pp. 235–242.
[26] P. Schnitzspan, M. Fritz, and B. Schiele, "Hierarchical support vector random fields: Joint training to combine local and global features," in *Proc. ECCV*, 2008, pp. 527–540.
[27] M. Y. Yang, W. Förstner, and M. Drauschke, "Hierarchical conditional random field for multi-class image classification," in *Proc. Int. Conf. Comput. VISAPP*, 2010, pp. 464–469.
[28] P. B. C. Leite, R. Q. Feitosa, A. R. Formaggio, G. A. O. P. Costa, and K. Pakzad, "Hidden Markov models for crop recognition in remote sensing image sequences," *Pattern Recognit. Lett.*, vol. 32, no. 1, pp. 19–26, Jan. 2011.
[29] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "TextonBoost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *Int. J. Comput. Vis.*, vol. 81, no. 1, pp. 2–23, Jan. 2009.
[30] Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images," in *Proc. 8th IEEE ICCV*, 2001, vol. 1, pp. 105–112.
[31] R. Szeliski *et al.*, "A comparative study of energy minimization methods for Markov random fields with smoothness-based priors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 1068–1080, Jun. 2008.
[32] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York, NY, USA: Springer-Verlag, 2006.
[33] S. Vishwanathan, N. Schraudolph, M. W. Schmidt, and K. P. Murphy, "Hierarchical conditional random field for multi-class image classification," in *Proc. 23rd ICML*, 2006, pp. 969–976.

[34] Arbeitsgemeindschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland. (1997) ATKIS—Amtlich Topographisch-Kartographisches Informationssystem. [Online]. Available: http://www.atkis.de
[35] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SCM-3, no. 6, pp. 610–622, Nov. 1973.
[36] N. Dalal and B. B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2005, pp. 886–893.
[37] M. A. Hall, "Correlation-based feature subset selection for machine learning," Ph.D. dissertation, Dept. Comput. Sci., Univ. Waikato, Hamilton, New Zealand, Apr. 1999. [Online]. Available: http://www.cs.waikato.ac.nz/ml/publications/1999/99MH-Thesis.pdf
[38] G. Foody, "Status of land cover classification accuracy assessment," *Remote Sens. Environ.*, vol. 80, no. 1, pp. 185–201, 2002.
[39] M. Rutzinger, F. Rottensteiner, and N. Pfeifer, "A comparison of evaluation techniques for building extraction from airborne laser scanning," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 2, no. 1, pp. 11–20, Mar. 2009.
[40] P. Helmholz *et al.*, "Semi-automatic quality control of topographic data sets," *Photogramm. Eng. Remote Sens.*, vol. 78, no. 9, pp. 959–972, 2012.
[41] J. Niemeyer, F. Rottensteiner, and U. Sörgel, "Classification of urban lidar data using conditional random fields and random forests," in *Proc. 7th IEEE/GRSS/ISPRS Joint Urban Remote Sens. Event*, 2013, pp. 139–142.

Thorsten Hoberg received the Diplom-Ingenieur degree in surveying from the Leibniz Universität Hannover, Hannover, Germany, in 2000 and the Second State degree to become an Assessor in surveying from the Federal State of Lower Saxony, Germany, in 2004.

From 2005 to 2012, he was a Researcher with the Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Hannover, Germany. Since 2012, he has been with the State Agency of Geoinformation and Land Development of Lower Saxony, Germany.

Franz Rottensteiner received the Dipl.-Ing. degree in surveying and the Ph.D. degree and *venia docendi* in photogrammetry, all from Vienna University of Technology, Vienna, Austria.

He was a Postdoctoral Researcher with the Vienna University of Technology; the University of New South Wales, Kensington, N.S.W., Australia; and the University of Melbourne, Parkville, Vic., Australia. He is currently with the Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Hannover, Germany, where he is an Associate (apl.) Professor, leading the Photogrammetric Image Analysis Group. His research interests include photogrammetry, automated extraction of topographic objects, processing of lidar data, and sensor orientation.

Raul Queiroz Feitosa received the B.Sc. degree in electronic engineering and the M.Sc. degree in engineering from the Technological Aeronautics Institute (ITA), São José dos Campos, Brazil, in 1979 and 1983, respectively, and the Dr.-Ing. degree in computer architecture from the University of Erlangen-Nürnberg, Erlangen, Germany, in 1988.

Since then, he has been with the Department of Electrical Engineering, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil, and also with the Department of Computer Engineering, State University of Rio de Janeiro, Rio de Janeiro, where he is an Associate Professor. His research interest includes image analysis and its applications in remote sensing, biometrics, and medicine.

Christian Heipke received the Dipl.-Ing. degree in surveying and the Ph.D. degree and *venia legendi* in photogrammetry, all from the Technical University of Munich, Munich, Germany.

Since 1998, he has been a Professor of photogrammetry and remote sensing with the Leibniz Universität Hannover, Hannover, Germany, where he leads a group of about 25 researchers. He has authored or coauthored more than 300 scientific papers, more than 70 of which appeared in peer-reviewed international journals. His professional interests comprise all aspects of photogrammetry, remote sensing, image understanding, and their connection to computer vision and Geographical Information Systems (GIS).

Dr. Heipke served as the Vice President of European Spatial Data Research (EuroSDR) from 2004 to 2009. He currently serves as the International Society of Photogrammetry and Remote Sensing (ISPRS) Secretary General and chairs the German Geodetic Commission (DGK).