Building Detection From One Orthophoto and High-Resolution InSAR Data Using Conditional Random Fields

Jan Dirk Wegner, Ronny Hänsch, Antje Thiele, and Uwe Soergel, *Member, IEEE*

***Abstract*—Today's airborne SAR sensors provide geometric resolution in the order well below half a meter. Many features of urban objects become visible in such data. However, layover and occlusion issues inevitably arise in urban areas complicating automated object detection. In order to support interpretation, SAR data may be analyzed using complementary information from maps or optical imagery. In this paper, an approach for building detection in urban areas based on object features extracted from high-resolution interferometric SAR (InSAR) data and one orthophoto is presented. Features describing local evidence as well as context information are used. Buildings are detected by classification of those feature vectors within a Conditional Random Field (CRF) framework. Although as graphical model similar to Markov Random Fields (MRF), CRFs have the advantage of incorporating global context information, of relaxing the conditional independence assumption between features, and of a more general integration of observations. We show that, first, CRFs perform well in comparison to Maximum Likelihood classifiers and MRFs. Second, the combined use of optical and InSAR features may improve detection results.**

***Index Terms*—Conditional random fields, data fusion, interferometry, object detection, synthetic aperture radar.**

I. Introduction

SYNTHETIC APERTURE RADAR (SAR) has become a very important remote sensing technique in the last two decades. Two key features of SAR are that it is almost independent of the weather conditions and its data acquisition capability independent of daylight, because of the larger signal wavelength (usually 3 to 25 centimeters) compared to the visible spectrum and the active sensor principle, respectively. Operating space borne systems like ERS-2 and ENVISAT provide rather coarse spatial resolution (e.g., 25 m ground sampling distance). Information extraction from those images is often restricted to radiometric properties; a typical application is land cover classification. The structure of settlement areas can usually be characterized only in a rather generalized manner, for instance, inner city areas and suburbs may be distinguished. In SAR data of one meter geometric resolution collected by modern space borne sensors such as TerraSAR-X and Cosmo-SkyMed, the geometric extent of individual objects like bridges, buildings, and roads becomes visible. Airborne sensors image the urban scene with even more detail. Using interferometric image pairs from airborne SAR sensors, buildings in urban areas may be detected [1], [2]. Tison *et al.* [3] achieve good results by classifying interferometric SAR data and estimating a height model using a Markov Random Field (MRF) approach. However, shadowing and layover effects, typical for SAR image acquisitions in urban areas, always complicate interpretation. Small buildings are often occluded by higher ones, while façades overlap with trees and cars on the streets. In addition, the appearance of a building in the image highly depends on the sensor's aspect. Buildings that are not oriented in azimuth direction with respect to the sensor are often hard to detect. This drawback can be partly overcome with SAR images from multiple aspects [4]. Especially the building recognition task is simplified if InSAR data are available, which were acquired from two orthogonal flight directions [5] or even more aspects [6].

Automatic SAR data analysis may be further facilitated by incorporating complimentary data, for example, GIS databases or high-resolution optical imagery. Optical images have the advantage of being widely available, whereas this is often not the case with GIS information. Additionally, the level of detail of GIS data may sometimes not be sufficient, for example, if roads are only marked with their centre line. Soergel *et al.* [7] combine high-resolution airborne InSAR data with an optical aerial image in order to reconstruct bridges over water three-dimensionally. Tupin and Roux [8] propose an approach to automatically extract footprints of large flat-roofed buildings based on line features. They use a SAR amplitude image and an optical aerial image as input for their approach. In [9] Tupin and Roux first segment homogeneous regions in an aerial photo. They then set up a region adjacency graph and the segments are used to regularize elevation data derived from a radargrammetrically analyzed SAR image pair by means of a MRF.

One drawback of MRFs is that they generatively model the joint distribution of labels and data. This implies that the distribution of the data has to be modeled which is often hard to accomplish.

Conditional Random Fields (CRF), introduced in [10] for segmenting and labeling 1-D sequence data, are discriminative

Manuscript received October 31, 2009; revised February 02, 2010, April 16, 2010; accepted May 26, 2010. Date of publication August 16, 2010; date of current version March 23, 2011.

J. D. Wegner and U. Soergel are with the Institute of Photogrammetry and GeoInformation, Leibniz University of Hannover, 30167 Hannover, Germany (e-mail: wegner@ipi.uni-hannover.de; soergel@ipi.uni-hannover.de).

R. Hänsch is with the Department of Computer Vision and Remote Sensing, Berlin Institute of Technology, 10587 Berlin, Germany (e-mail: rhaensch@fpk.tu-berlin.de).

A. Thiele is with the Institute of Photogrammetry and Remote Sensing (IPF), Karlsruhe Institute of Technology (KIT), 76128 Karlsruhe, Germany (e-mail: antje.thiele@kit.edu).

Digital Object Identifier 10.1109/JSTARS.2010.2053521

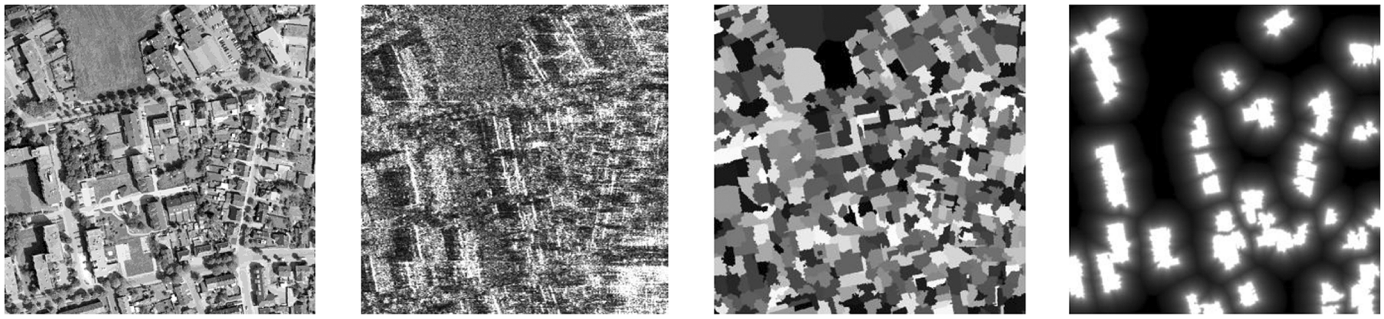

Fig. 1. (a) Intensity image of an optical test region, and (b) corresponding SAR magnitude image, (c) normalized cuts segments of the optical image, and (d) distance map of InSAR corner lines.

models and hence the data distribution does not have to be modeled. Kumar and Hebert [11] extended CRFs to label 2-D images and they have since then been applied successfully to various computer vision tasks [22], [26]–[30].

However, CRFs have only very rarely been used to classify remotely sensed data, yet. Zhong *et al.* [12] set up multiple CRFs to detect settlement areas in optical satellite images. Lu *et al.* [13] use CRFs to extract a digital elevation model from an airborne LiDAR digital surface model. In [14] He *et al.* apply a CRF to SAR data with the goal of building extraction, which, to our knowledge, has been the only time a CRF has been used to classify SAR data so far.

In this paper, for the first time, a CRF is used for combining optical and SAR data. We present an approach for building detection in an urban area that combines features from an orthophoto with line features from mono-aspect InSAR data. Both feature sets are introduced to a single feature vector which is then used to distinguish two classes, buildings and non-buildings, in a CRF framework. We compare the CRF to two other probabilistic methods: a standard Maximum Likelihood classifier (non-contextual) and a state-of-the-art Markov Random Field (contextual). We present, first, that contextual classifiers are advantageous for our task of building detection and, second, that optimal results integrating context are achieved with the CRF.

The paper is organized as follows. Section II explains the selected optical features and InSAR features in detail. Thereafter, classification with CRFs is described in detail in Section III. Results are presented and discussed in Section IV while Section V concludes this paper and gives an outlook.

II. Feature Extraction

In a first step, features that are good hints for buildings in urban areas are extracted from both the orthophoto and the InSAR data. A sub-region of the orthophoto and one corresponding magnitude image of the InSAR image pair are shown in Fig. 1(a), (b). We use rather simple features because the focus of this paper is on the investigation of the CRF framework and not on sophisticated feature selection. Such features are then introduced to a common feature vector, which is used to classify the data into building and non-building sites.

A. Optical Features

A first impression of suitable features was achieved by comparing the marginal distributions of each single feature of the classes building and non-building. Features showing very different marginals for the two classes were taken as basic features. The difference of the marginals is a rather weak feature selection technique because it does not take into account joint distributions of different features. Therefore, those basic features were combined with various other features that did not perform well regarding the marginal difference. Lots of configurations were tested within the CRF framework. According to this rather experimental and simple feature selection technique we selected features based on color and on texture. The mean of the red and the green channel (normalized by the length of the RGB vector), hue mean and hue standard deviation and the saturation mean were found to be descriptive color features. Furthermore, the co-occurrence matrices of all three RGB channels were set up and various measures were tested. We found the Haralick features homogeneity and correlation to be valuable building hints.

Additional features are calculated based on the gradient orientation histogram of the intensity image as already used for building detection in terrestrial images by Kumar and Hebert [15] and Korč and Foerstner [16]. However, we derive slightly different features because we deal with aerial imagery and thus no façades with characteristic horizontal and vertical gradients appear. We use the mean, the mean difference to a uniform distribution and the maximum value of the gradient orientation histogram as features.

Each feature of a particular pixel is calculated within a window. The size of the window strongly influences the final classification result. If the window size is chosen too small, feature extraction becomes unstable, whereas a too big window leads to a lack of detail. The appropriate window size depends on the size of the smallest object we want to detect, which is a function of the pixel size on the ground. In order to provide context information for the classification task, each feature is calculated in three different scales (i.e., three different window sizes). We tested various numbers of scales and window sizes. Best results were achieved with features calculated in window sizes 10×10, 15×15, and 20×20 pixels.

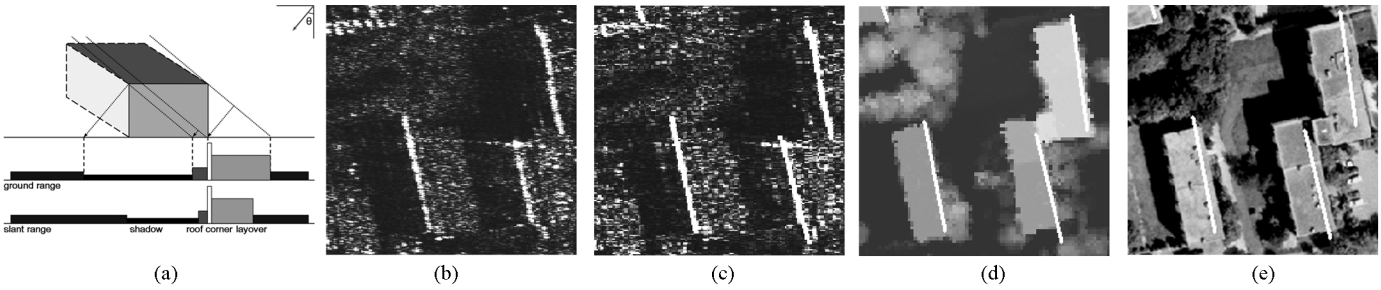

Fig. 2. (a) Schematic view of magnitude signature in SAR data, (b) magnitude data, (c) overlaid with corner lines, (d) LIDAR data overlaid with corner lines, and (e) optical data overlaid with extracted corner lines.

B. InSAR Features

The analysis of the InSAR data is focused on the detection of stable building features and their recognition. Magnitude and phase signatures of buildings are analyzed considering varying viewing directions and geometric building conditions. The signature of buildings is characterized by a layover area (direct reflection), the corner line (double-bounce reflection), the roof area (direct reflection), and the shadow area (no return), which is shown in Fig. 2(a). Flat-roofed building signatures are given in Fig. 2(b). In [5] the change of building signatures due to different viewing directions and building dimensions was discussed in detail. It was pointed out that the most reliable feature is the building corner line, especially if we deal with small buildings in the reconstruction task. Furthermore, the InSAR phase distribution along the corner lines facilitates the separation of these lines from other bright lines in the scene. All signal at the corner position caused by double bounce reflection between ground and wall is summed up resulting in a stable phase that corresponds to the local terrain height (refer to [19] for more details).

The line extraction procedure requires some few thresholds to be set. These thresholds have been trained *a priori* based on a set of high-resolution SAR images of different scenes. At runtime the line detector template and thresholds defined in unit meter are adapted automatically to the given ground sampling distance of the current image. The extraction of the building corner lines starts with the segmentation of bright lines in the magnitude data. Thereby, a slightly modified version of the adapted ratio line detector according to [23] is applied to the original magnitude image (Fig. 2(b)). The template detector determines the probability of a pixel of belonging to a line by considering eight different template orientations. Furthermore, based on a magnitude threshold (10 dB) and a probability threshold (0.5), clusters of bright pixels are extracted. Then, straight lines are fitted to the pixel clusters and lines shorter than two meters are suppressed. Next, short approximately collinear line segments are joined if the gap between them is smaller than two meters and the orientation difference is below 0.2 rad. The magnitude image overlaid with results of the line detection is shown in Fig. 2(c). Thereafter, potential building corner lines are distinguished from other bright lines in the scene by considering the interferometric phases. Basically, a potential building corner line has to show a mean phase value, which corresponds to the local terrain height. The extracted corner lines are projected from slant into ground geometry using the InSAR heights at the line locations in order to enable joint feature classification with the optical features. Fig. 2(d) compares the geo-coded corners to reference LIDAR data. The same corner lines are displayed in Fig. 2(e) overlaid to the orthophoto. Depending on object height and position relative to the optical sensor's nadir point, the buildings fall over the projected corner lines (refer to [20] for a comparison of optical and SAR viewing geometries).

Since feature extraction in the current state of the approach is performed on pixel level, the corner lines may not be used directly (as it would be possible on a higher semantic level). We solve this issue by, first, segmenting the optical intensity image. Usually, strong gradients occur between rather homogeneous building roofs and their surroundings. Thus, we segment the optical intensity image with the multi-resolution Normalized Cuts method of Cour *et al.* [31] (Fig. 1(c)). Second, all segments that are located behind a corner line (in range direction) and that overlap with it are extracted. Segments that do not coincide with a corner line are discarded. We then calculate linear (inverse) distance maps outside the previously extracted segments. More practically speaking, all pixels inside an extracted region are set to one and adjacent pixels are given a value according to their distance to the segment boundary (high values if they are close to the segment boundary, see Fig. 1(d)). Some corner lines may be located right in front of a building without overlapping due to small projection errors of the corner lines and the perspective distortion of the buildings in the optical image. Therefore, a tolerance threshold is set. All segments located up to two meters away from the corner line in range direction are still considered as overlapping segments. The maximum and the minimum distances are then determined for each image patch and used as features. Only patches at the scale with the highest resolution are considered.

III. Classification With Conditional Random Fields

In this paper we adapt the main ideas of Kumar and Hebert proposed in [15] to the task of combining the previously described optical and InSAR features for building detection.

The basic idea is to extract characteristic features for buildings in both optical and InSAR data, to insert both feature sets into a single feature vector, and to finally classify the data based on this feature vector using a CRF. Like all graphical models, CRFs have the advantage of assigning probabilities to the final labeling instead of only providing crisp decisions. Those probabilities are very useful for post-processing or decision making. They also have several advantages over MRFs (1). $\boldsymbol{y}$ contains

all labels (binary, either -1 or 1), $\boldsymbol{x}$ all data, i is an image site out of the set of all image sites S of a single image, j is an image site of the 4-connectivity neighborhood N_i of site i, and $Z(\boldsymbol{x})$ is the partition function. Parameter β is a positive real-valued weighting parameter of the Ising model (for binary classification).

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\left(\sum_{i \in S} \log P_i(x_i|y_i) + \sum_{i \in S} \sum_{j \in N_i} \beta y_i y_j \right) \quad (1)$$

MRFs can be seen as an extension of Naïve Bayes. They are generative models and hence estimate the joint distribution $P(\boldsymbol{x}, \boldsymbol{y})$ of data $\boldsymbol{x}$ and labels $\boldsymbol{y}$, which can be decomposed into a product of factors $P(\boldsymbol{x}|\boldsymbol{y})P(\boldsymbol{y})$. In the Bayesian context the first term in (1) can be viewed as the likelihood term, whereas the second one is a prior over labels $\boldsymbol{y}$. The likelihood $P_i(x_i|y_i)$ uses data only from a single site i and not from all sites (like CRFs). The prior term only compares adjacent labels y_j with the investigated label y_i, usually using (for binary classification tasks) the Ising model $\beta y_i y_j$. Partition function $Z(\boldsymbol{x})$, which can be interpreted as the distribution $P(\boldsymbol{x})$ of data $\boldsymbol{x}$ in the Bayesian framework, acts as a normalization constant (for a given data set). It can be expressed as sum over all possible label configurations of the product $P(\boldsymbol{x}|\boldsymbol{y})P(\boldsymbol{y})$.

In contrast, CRFs are based on the maximum entropy approach, which is known to be able to provide accurate and robust classification results. As opposed to MRFs, CRFs are discriminative models and therefore model only the posterior distribution $P(\boldsymbol{y}|\boldsymbol{x})$ of the labels $\boldsymbol{y}$ given data $\boldsymbol{x}$. The most common approach is based on sufficient statistics of the exponential family and the posterior distribution can be modeled as:

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\left(\sum_{i \in S} A_i(\boldsymbol{x}, y_i) + \sum_{i \in S} \sum_{j \in N_i} \boldsymbol{I}_{ij}(\boldsymbol{x}, y_i, y_j) \right) \quad (2)$$

The association potential $A_i(\boldsymbol{x}, y_i)$ measures how likely a site i is labeled with y_i given the data $\boldsymbol{x}$, while the interaction potential $I_{ij}(\boldsymbol{x}, y_i, y_j)$ describes how two sites i and j interact. It should be noted that both potentials have access to the whole image. In particular the interaction potential $I_{ij}(\boldsymbol{x}, y_i, y_j)$ is not only a function of adjacent labels y_i and y_j in the local neighborhood (like in case of MRFs, compare (1)), but of all data $\boldsymbol{x}$, too. Neighborhood N_i of site i may potentially be the entire image. This becomes convenient if we want to compare labels based on underlying data. In addition, both the association potential and the interaction potential are defined over all data, the entire orthophoto and all InSAR data in our case. Hence, we may introduce both local and global context knowledge which is a major advantage concerning automatic analysis of high-resolution remote sensing data of urban areas. In order to obtain a posterior probability $P(\boldsymbol{y}|\boldsymbol{x})$ of labels $\boldsymbol{y}$ conditioned on data $\boldsymbol{x}$ the exponential of the sum of association potential and interaction potential is normalized by division through the partition function $Z(\boldsymbol{x})$ (3), which is a constant for a given data set.

$$Z(\boldsymbol{x}) = \sum_{y} \exp\left(\sum_{i \in S} A_i(\boldsymbol{x}, y_i) + \sum_{i \in S} \sum_{j \in N_i} I_{ij}(\boldsymbol{x}, y_i, y_j) \right) \quad (3)$$

Our modeling of the association potential $A_i(\boldsymbol{x}, y_i)$ and the interaction potential $I_{ij}(\boldsymbol{x}, y_i, y_j)$ is closely related to the approach proposed by Kumar and Hebert in [15]. For our binary classification task of distinguishing building sites from non-building sites, labels y_i may either become 1 or -1, respectively.

A. Association Potential

The association potential $A_i(\boldsymbol{x}, y_i)$ measures how likely it is that a site i takes label y_i given all data $\boldsymbol{x}$ (see (4)). Data $\boldsymbol{x}$ in our case are the orthophoto and the InSAR data. We use a generalized linear model to distinguish building and non-building sites in the association potential.

$$A_i(\boldsymbol{x}, y_i) = \exp(y_i \boldsymbol{w}^T \boldsymbol{h}_i(\boldsymbol{x})) \quad (4)$$

Vector $\boldsymbol{h_i}(\boldsymbol{x})$ contains all previously described node features of three different scales (refer to Section II. for details). A quadratic expansion of all node features is conducted [15] in order to introduce a more accurate quadratic decision surface instead of a linear one. Vector $\boldsymbol{w^T}$ contains the weights of the features in $\boldsymbol{h_i}(\boldsymbol{x})$ that are tuned during the training process.

B. Interaction Potential

The interaction potential $I_{ij}(\boldsymbol{x}, y_i, y_j)$ determines how two sites i and j should interact regarding all data $\boldsymbol{x}$ (see (5)). Vector $\boldsymbol{v^T}$ contains the weights of the features, which are adjusted during the training process. y_i is the label of the site of interest and y_j the label it is compared to. Unlike clique potentials in MRFs, label y_j does not necessarily have to be a label of a site j in the local neighborhood of y_i. The comparison of labels y_i and y_j follows the Ising model $\beta y_i y_j$. With $\beta = 1$, the product $y_i y_j$ becomes -1 if labels y_i and y_j do not belong to the same class, whereas their product is 1 in case both labels are equal. As already stated in [16], vector $\boldsymbol{\mu_{ij}}(\boldsymbol{x})$ enables to support or suppress this term and ignoring $\boldsymbol{\mu_{ij}}(\boldsymbol{x})$ would lead to the classical MRF smoothing potential, the traditional Ising model.

$$I_{ij}(\boldsymbol{x}, y_i, y_j) = \exp(y_i y_j \boldsymbol{v^T} \boldsymbol{\mu_{ij}}(\boldsymbol{x})) \quad (5)$$

Usually, the feature vector $\boldsymbol{\mu_{ij}}(\boldsymbol{x})$ is simply calculated by either subtracting or concatenating $\boldsymbol{h_i}(\boldsymbol{x})$ and $\boldsymbol{h_j}(\boldsymbol{x})$ [11], [15], [18]. In general, $\boldsymbol{\mu}_{ij}(\boldsymbol{x})$ can also be chosen based on other features than such already used for the association potential and other methods of comparing the features are possible, too. We make use of this flexibility provided by CRFs in order to propose a new design of the interaction potential for our task of building detection.

There are mainly two reasons for a specific design. First, the classical interaction potential is appropriate for typical computer vision tasks, which usually consist of detecting few instances of an object class in an image. Most times, only one instance per object class appears in an image and it often covers a rather large homogeneous area. We face another situation. Many instances of the same object class appear in various locations within an image. Small gaps appear in-between the buildings and the area of a single building is small compared to the image size. Second, various tests showed that simply relying on differences or a concatenated vector for comparing adjacent image

patches did not sufficiently distinguish building patches from adjacent non-building patches.

Therefore, several functions for comparing the features of adjacent patches were tested and an explicit discontinuity constraint was introduced. We achieve the best results if we use a bounded ratio of the feature vectors $\boldsymbol{h_i}(\boldsymbol{x})$ and $\boldsymbol{h_j}(\boldsymbol{x})$ instead of the classical ones. Furthermore, building roofs and neighboring streets or parking lots often show very similar features and hence are hard to distinguish considering only the corresponding feature vectors. However, we may assume that patches with similar features belong to different object classes if they are separated by a high gradient in the optical image (e.g., roof edges or narrow shadows). Therefore, we add the mean of the gradient between two patches as an explicit discontinuity constraint.

For ease of notation we define $x_{nij} = f_{nij}(h_{ni}(x_i), h_{nj}(x_j))$ where $h_{ni}(x_i)$ is the n-th feature at image site i and $h_{ni}(x_i)$ the one of the neighboring site j, respectively. Then, the feature vector $\boldsymbol{\mu_{ij}}(\boldsymbol{x})$ of the edge between nodes (image sites) i and j can be expressed as

$$\boldsymbol{\mu_{ij}}(\boldsymbol{x}) = \begin{pmatrix} x_{1ij} & x_{2ij} & \cdots & x_{nij} \end{pmatrix} \cdot w_{\mathrm{disc},ij} \tag{6}$$

where $w_{\mathrm{disc},ij}$ is a scalar weight derived from the gradient between both sites. Edge feature x_{nij} is determined by the ratio r of the single site features $h_{ni}(x_i)$ and $h_{nj}(x_j)$. It is scaled between zero and one (see (7), (8)). All ratios above a certain bound b are set to the maximum value (one) in order to avoid numerical instability.

$$x_{nij} = \begin{cases} r_1 \text{ if } h_{ni}(x_i) \geq h_{nj}(x_j) \text{ and } \frac{h_{ni}(x_i)}{h_{nj}(x_j)} \leq b \\ r_2 \text{ if } h_{ni}(x_i) < h_{nj}(x_j) \text{ and } \frac{h_{nj}(x_j)}{h_{ni}(x_i)} \leq b \\ 1 \text{ if } \frac{h_{ni}(x_i)}{h_{nj}(x_j)} > b \text{ or } \frac{h_{nj}(x_j)}{h_{ni}(x_i)} > b, \end{cases} \tag{7}$$

$$\begin{aligned} \mathrm{r}_1 &= \left(\frac{h_{ni}(x_i)}{h_{nj}(x_j)} - 1 \right) / (b-1), \\ \mathrm{r}_2 &= \left(\frac{h_{nj}(x_j)}{h_{ni}(x_i)} - 1 \right) / (b-1) \end{aligned} \tag{8}$$

Weights $w_{\mathrm{disc},ij}$ scale the edge features as a function of the mean of the norm gradient g_{ij} between two adjacent sites i and j. Gradient g_{ij} is calculated between adjacent patches. Considering the regular grid structure of the image sites, this configuration would naturally prefer horizontal or vertical gradients. Thus, the width of the area in which the mean gradient is computed has to be chosen carefully in order to also allow for diagonal gradients. In our case of 10×10 pixels wide image patches we achieve best results with two pixel wide areas on each side of a site, which leads to a 4×10 pixels area for a site pair.

An investigation of the histogram of g_{ij} of the entire image suggests that in general values above 0.5 indicate that two sites belong to different objects. Therefore, we do not use the mean gradient g_{ij} directly as weighting parameter (which would correspond to a linear weighting) but introduce it into a sigmoid function with the inflexion position at $\kappa = 0.5$ ((9), Fig. 3). Additionally, we shift the sigmoid function in y-direction in order to still allow for discontinuities between two adjacent sites i and j even if no gradient occurs between them.

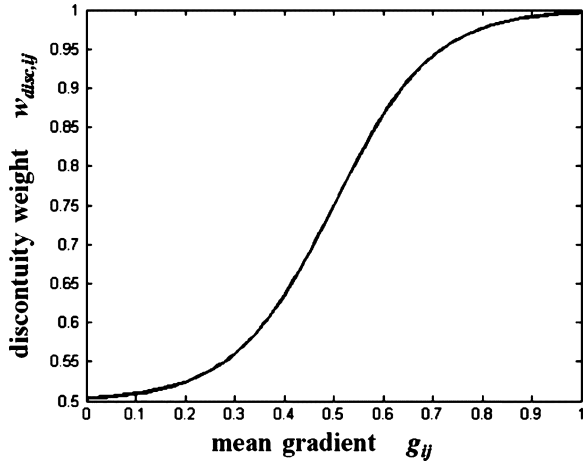

Fig. 3. Sigmoid discontinuity weighting function with parameters $\alpha = 10$ and $\kappa = 0.5$ as used for the presented results (note: y axis begins at 0.5).

$$w_{\mathrm{disc},ij} = \left(1 + \left(\frac{1}{1 + \exp(-\alpha(g_{ij} - \kappa))} \right) \right) / 2 \tag{9}$$

In (9) the two parameters κ (curve inflexion position) and α (curve inflexion) have to be set. Due to the previously mentioned investigation of the mean gradient histogram we set κ to 0.5. Various settings of α were accomplished resulting in an optimal value of $\alpha = 10$. The resulting discontinuity preserving weighting function (9) is shown in Fig. 3.

Only single scale features are used for the interaction potential (i.e., features computed from the patches with the highest resolution 10×10 pixels) because lower resolution patches would overlap. This would lead to highly correlated features of adjacent sites being compared. Again, a quadratic expansion of the feature vector is done as in case of the association potential.

IV. Results and Discussion

We use an optical orthophoto and one mono-aspect InSAR image pair of the city Dorsten, Germany, as test data. The orthophoto was acquired with an analogue aerial camera Zeiss RMK and scanned (© Geoinformation NRW). Pixel size on the ground is 0.31 m. The single-pass X-band InSAR data (wavelength $\lambda = 3.14$ cm) were acquired by the AeS sensor of Intermap Technologies [21]. Spatial data resolution of the original single-look data is 38 cm in range and 18 cm in azimuth with a baseline of 2.4 m. For cross-validation purposes we subdivided the test scene into four non-overlapping parts of equal size of 1000×1000 pixels (corresponding to 310 m $\times$ 310 m).

Several tests and comparisons are done. First, we evaluate the contribution of the InSAR corner lines to the overall building detection quality. Second, we compare the CRF results to those achieved with a Maximum Likelihood (ML) classifier in order to evaluate the benefits of integrating context information. More important, ML is a simple, easy to implement baseline approach

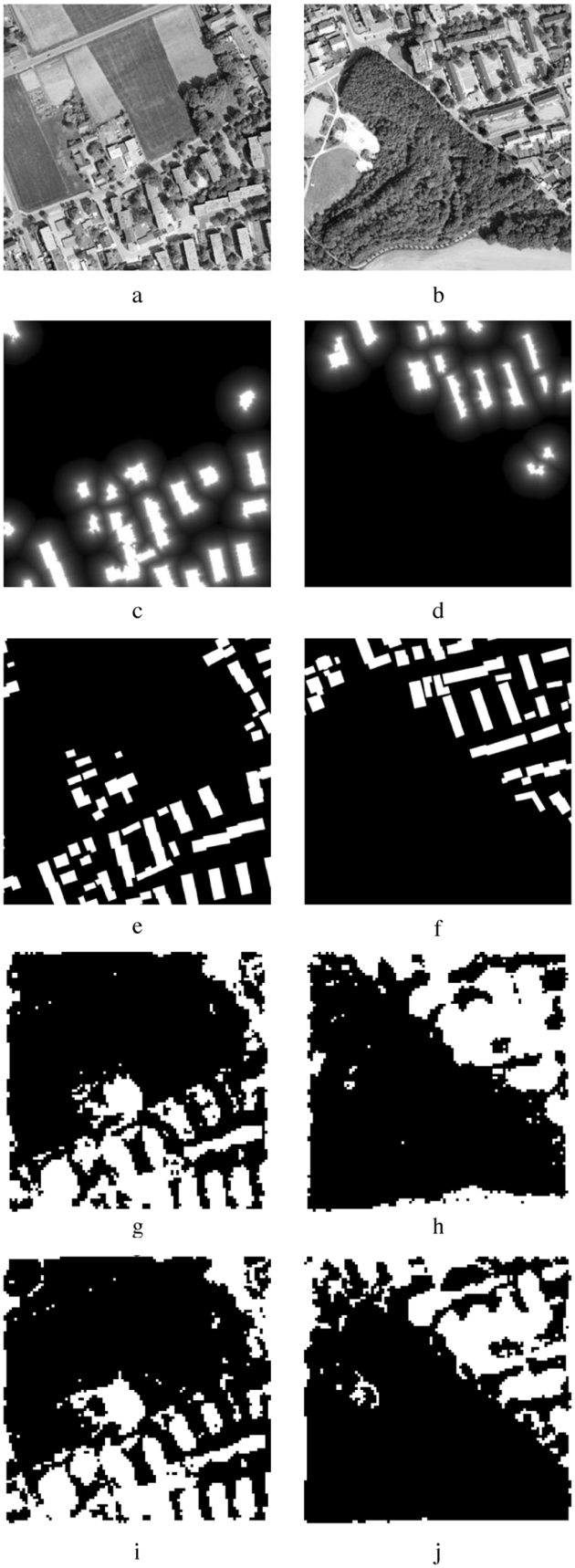

Fig. 4. Results of CRF building detection of two test scenes: (a), (b) intensity images of test regions of the orthophoto containing buildings, agricultural areas, streets, and trees, (c), (d) distance maps of InSAR corner lines of the test regions above, (e), (f) ground truth, (g), (h) CRF results using only optical features, (i), (j) CRF results based on optical and corner line features.

TABLE I
COMPARISON OF CRF BUILDING DETECTION RESULTS USING ONLY THE OPTICAL FEATURES VERSUS THE COMBINATION OF OPTICAL AND CORNER FEATURES (MEAN AND STANDARD DEVIATION OF TRUE POSITIVE RATE (TPR) AND FALSE POSITIVE RATE (FPR) FOR L-BFGS TRAINING AND LBP INFERENCE)

CRF (optical features)				CRF (optical + corner features)			
TPR		FPR		TPR		FPR	
μ	σ	μ	σ	μ	σ	μ	σ
0.83	0.10	0.26	0.08	0.85	0.04	0.24	0.06

often applied for land cover classification still today. Its performance for the task at hand signifies the result achievable with the simplest method. Thus, it indicates the degree of difficulty of the problem to be solved. Our focus is on the degree of improvement achieved with CRFs. Third, we compare CRFs to MRFs, which are the current state-of-the-art contextual classifiers.

We use the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) [17] method as optimizer to train the association potential and the interaction potential because it has been shown to provide good results for training CRFs by Vishvanathan *et al.* [18]. Both potentials are learned simultaneously because we can no longer assume them to be independent like in case of MRFs [15]. For inference we tested Loopy Belief Propagation (LBP) [24] and the Pseudo-likelihood (PL) approach [25] (also see [18] for a comparison of both methods). LBP is a standard technique for approximate inference of graphs with cycles. In contrast, PL is an exact inference technique but with the drawback of changing the log-likelihood objective function (simply the log of (2)) by only summing over all possible label combinations in the local neighborhood of a particular node in the partition function $Z(\boldsymbol{x})$ (2). This leads to an over-smoothing effect resulting in entire settlements being detected instead of single buildings. We found the combination of L-BFGS training and LBP inference to deliver the best results and thus we use them for all further tests.

A. Contribution of SAR Corner Lines to Building Detection Within CRF Framework

First, we want to evaluate experimentally whether the InSAR corner lines lead to an improvement of the overall building detection quality or not. Therefore, we compare building detection results of the CRF framework using merely the optical features to the combination of optical and InSAR features. Mean and standard deviation of true positive rate (TPR) and false positive rate (FPR) estimated by 4-fold cross validation for L-BFGS training and LBP inference are shown in Table I. Two test images, the corresponding InSAR corner line distance maps, ground truth, and results with and without the corner line features are shown in Fig. 4.

The TPR using only the optical features (see Fig. 4(g), (h)) is slightly lower than the combination with InSAR corner line features (Fig. 4(i), (j)) (TPR: 0.83 versus 0.85). In addition, the corner lines reduce the FPR from 0.26 to 0.24. This effect may well be noticed if we compare Fig. 4(h) and (j) top right. The corner lines lead to discontinuities and narrow gaps between buildings being better preserved. In general, false positives of the CRF results are usually due to small gaps between neighboring buildings being classified as building. This effect occurs

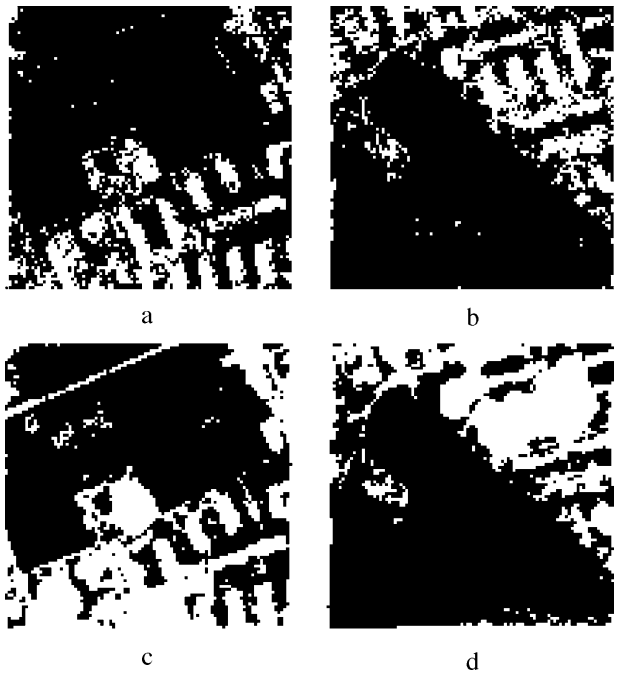

Fig. 5. Building detection results of the same test scenes as shown in Fig. 4 based on the combination of optical and corner line features: (a), (b) with ML, and (c), (d) with MRF.

if building roof and surroundings have similar features (i.e., the edge features x_{nij} of two adjacent sites are very small) and no high gradient appears between them. Another reason for narrow gaps being misclassified is the limited minimum window size, a compromise between the smallest possible window size and stable features. In our experience, a window size smaller than ten pixels leads to very unstable features of the given data. This issue could possibly be solved by applying the CRF framework to a region adjacency graph of image segments. In general, better results could probably be achieved based on more sophisticated optical features and the inclusion of heights derived from the InSAR data.

B. CRF versus ML and MRF

In order to assess the quality of the CRF results, we compare them to two other probabilistic supervised learning methods: to ML and to MRFs. ML is a generative standard approach for classification based on the Bayesian theorem. The main difference of ML compared to MRFs is that it does not model local context information through the prior term. ML simply assumes a uniformly distributed prior and maximizes the likelihood term. We use a multivariate Gaussian function, the classical approach, to model the likelihood term. It may thus be seen as a baseline approach of probabilistic classification without using prior (context) information. MRF is a state-of-the-art contextual classification approach. Context is considered through the prior term (refer to Section III. for more details). We use exactly the same combined optical and InSAR features as input to ML and MRF as for the CRF (Section II.). Table II summarizes the building detection results of ML and MRF that are shown in Fig. 5(a), (b) and (c), (d), respectively.

TABLE II
COMPARISON OF ML AND MRF: MEAN AND STANDARD DEVIATION OF TRUE POSITIVE RATE (TPR) AND FALSE POSITIVE RATE (FPR) OF BUILDING DETECTION USING THE COMBINATION OF OPTICAL AND INSAR CORNER LINE FEATURES

ML				MRF			
TPR		FPR		TPR		FPR	
μ	σ	μ	σ	μ	σ	μ	σ
0.78	0.06	0.22	0.06	0.89	0.03	0.32	0.10

In the following, we compare ML and MRF results to those achieved with the CRF shown in Table I (see Fig. 4(i), (j)). ML correctly detects 78% of all building sites and only misclassifies 22% of the non-building sites. Compared to ML, the MRFs show a better building detection rate of 0.89 with the drawback of a much higher FPR (0.32). This high FPR compared to such of ML already hints at the principle drawback of MRFs if dealing with building detection in urban areas: they work very well as long as buildings sparsely distributed but fail at dense building groups.

Compared to ML (Fig. 5(a), (b)) the CRF (Fig. 4(i), (j)) achieves a much higher detection rate (TPR 0.85 versus 0.78) and a FPR on the same level (0.24 versus 0.22). Visual comparison of CRF results to ML reveals different characteristics. ML usually detects at least one image patch of each building but rarely the entire building. However, CRF often detects all image patches of a building while completely missing out on some very small ones. If we compare ML and CRF results on object level, the CRF detects more patches per building. This is an advantage of the CRF and important for further processing on object level.

The MRF achieves a slightly higher TPR than the CRF (0.89 versus 0.85). Due to the aforementioned smoothing effect of MRFs regarding dense building groups, the MRF FPR is significantly higher than the CRF one (0.32 versus 0.24). This can be explained by the simple Ising model of the MRF. Unlike the CRF the MRF does not consider observations $\boldsymbol{x}$ in the prior term (compare (1) and (2)). Only labels are compared regardless of the edge features as in case of the CRF. Furthermore, the CRF facilitates the integration of an explicit discontinuity constraint based on the observations (the gradient in our case).

Finally, the standard deviations of ML, MRF, and CRF are on the same relatively low level (0.03 to 0.10 for both TPR and FPR) indicating that the classifier performances show limited dependence on the choice of the training samples.

V. CONCLUSION AND OUTLOOK

In this paper we investigated the potential of CRFs for the task of building detection merging optical and InSAR features. Combining features of different data sources (i.e., sensor types in our case) may easily be accomplished by simply introducing them into a common feature vector. A major advantage is that the interaction potential is also defined over the observations instead of simply comparing labels. This advantage over MRFs was clearly highlighted in Section IV.B. CRFs perform well for single building detection compared to the probabilistic standard approach ML.

However, an issue to deal with is that although the standard interaction potential was replaced with a specifically designed

one, still many narrow gaps are misclassified. The current approach is limited by the regular grid structure of image patches. In order to benefit from the inherent structure of the image and to allow more expressive modeling of context, the CRF should be set up on an irregular graph of image segments (taking the CRF to a higher semantic level). Inspired by [9], the CRF could be applied to a region adjacency graph of segments of the optical image. This would also enable a more elegant integration of the gradient discontinuity constraint and the InSAR corner lines possibly increasing their impact on the overall building detection results.

Furthermore, buildings may appear differently particularly in optical data and, second, different building parts like facades and roofs may lead to a variety of characteristic features. Hence, it becomes difficult to treat buildings as a single class because features get inconsistent. An idea would be to break down the building class into further sub-categories by introducing hidden states to the CRF [22].

Finally, more elaborated features, for example textons and InSAR heights will be included in future work. However, the main challenge of future work will be how to introduce (and learn) more sophisticated contextual knowledge in the interaction potential based on a region adjacency graph of irregularly distributed image segments.

REFERENCES

[1] P. Gamba, B. Houshmand, and M. Saccani, "Detection and extraction of buildings from interferometric SAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 1, pp. 611–617, 2000, part 2.

[2] U. Soergel, U. Thoennessen, and U. Stilla, "Reconstruction of buildings from interferometric SAR data of built-up areas," in *Proc. PIA, Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2003, vol. 34, pp. 59–64, part 3/W8.

[3] C. Tison, F. Tupin, and H. Maitre, "A fusion scheme for joint retrieval of urban height map and classification from high-resolution interferometric SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 2, pp. 496–505, 2007.

[4] F. Xu and Y.-Q. Jin, "Automatic reconstruction of building objects from multiaspect meter-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 7, pp. 2336–2353, 2007.

[5] A. Thiele, E. Cadario, K. Schulz, U. Thoennessen, and U. Soergel, "Building recognition from multi-aspect high-resolution InSAR data in urban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 11, pp. 3583–3593, 2007.

[6] R. Bolter, "Buildings From SAR: Detection and reconstruction of buildings from multiple view high resolution interferometric SAR data," Ph.D. thesis, Univ. of Graz, Graz, Austria, 2001.

[7] U. Soergel, E. Cadario, A. Thiele, and U. Thoennessen, "Feature extraction and visualization of bridges over water from high-resolution InSAR data and one orthophoto," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 1, no. 2, pp. 147–153, 2008.

[8] F. Tupin and M. Roux, "Detection of building outlines based on the fusion of SAR and optical features," *ISPRS J. Photogramm. Remote Sens.*, vol. 58, pp. 71–82, 2003.

[9] F. Tupin and M. Roux, "Markov random field on region adjacency graph for the fusion of SAR and optical data in radargrammetric applications," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 8, pp. 1920–1928, 2005.

[10] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. Int. Conf. Machine Learning*, 2001, 8 pp.

[11] S. Kumar and M. Hebert, "Discriminative random fields: A discriminative framework for contextual interaction in classification," in *Proc. IEEE Int. Conf. Computer Vision*, 2003, vol. 2, pp. 1150–1157.

[12] 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, 2007.

[13] 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, 2009, Part 2.

[14] W. He, M. Jäger, A. Reigber, and O. Hellwich, "Building extraction from polarimetric SAR data using mean shift and conditional random fields," in *Proc. Eur. Conf. Synthetic Aperture Radar*, 2008, vol. 3, pp. 439–443.

[15] S. Kumar and M. Hebert, "Discriminative random fields," *Int. J. Comput. Vision*, vol. 68, no. 2, pp. 179–201, 2006.

[16] F. Korč and W. Förstner, "Interpreting terrestrial images of urban scenes using discriminative random fields," *Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 37, pp. 291–296, 2008, part B3a.

[17] D. C. Liu and J. Nocedal, "On the limited memory BFGS for large scale optimization," *Math. Programm.*, vol. 45, no. 1–3, pp. 503–528, 1989.

[18] S. V. N. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. P. Murphy, "Accelerated training of conditional random fields with stochastic gradient methods," in *Proc. Int. Conf. Machine Learning*, 2006, pp. 969–976.

[19] A. Thiele, J. D. Wegner, and U. Soergel, "Building Reconstruction From Multi-Aspect InSAR Data," in *Radar Remote Sensing of Urban Areas*, U. Soergel, Ed., 1st ed. New York: Springer, 2010, ch. 8, pp. 187–214, ISBN: 978-9048137500.

[20] J. D. Wegner and U. Soergel, "Bridge height estimation from combined high-resolution optical and SAR imagery," *Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 37, pp. 1071–1076, 2008, part 7/3.

[21] M. Schwaebisch and J. Moreira, "The high resolution airborne interferometric SAR AeS-1," in *Proc. 4th Int. Airborne Remote Sensing Conf. and Exhibition*, 1999, pp. 540–547.

[22] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Trans. Pattern Analysis Machine Intell.*, vol. 29, no. 10, pp. 1848–1853, 2007.

[23] F. Tupin, H. Maitre, J.-F. Mangin, J.-M. Nicolas, and E. Pechersky, "Detection of linear features in SAR images: Application to road network extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 36, no. 2, pp. 434–453, 1998.

[24] B. J. Frey and D. J. C. MacKay, "A revolution: Belief propagation in graphs with cycles," in *Adv. Neural Inf. Process. Syst.*, M. I. Jordan, M. J. Kearns, and S. A. Solla, Eds. Boston, MA: MIT Press, 1998, vol. 10.

[25] J. Besag, "On the statistical analysis of dirty pictures," *J. Royal Statist. Soc.*, vol. 48, no. 3, pp. 259–302, 1986.

[26] A. Quattoni, M. Collins, and T. Darrell, "Conditional random fields for object recognition," in *Proc. IEEE Conf. Neural Information Processing Systems*, 2004, 8 pp.

[27] A. Torralba, K. Murphy, and W. Freeman, "Contextual models for object detection using boosted random fields," in *Proc. IEEE Conf. Neural Information Processing Systems*, 2004, 8 pp.

[28] X. He, R. S. Zemel, and M. Á. Carreira-Perpinán, "Multiscale conditional random fields for image labeling," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004, 8 pp.

[29] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006, 7 pp.

[30] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *Proc. IEEE Int. Conf. Computer Vision*, 2007, 8 pp.

[31] T. Cour, F. Bénézit, and J. Shi, "Spectral segmentation with multiscale graph decomposition," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2005, 8 pp.

Jan Dirk Wegner was born in Oldenburg, Germany, in 1982. He received the Diploma degree (Dipl.-Ing.) in geodesy and geoinformatics from the Leibniz Universität Hannover, Hannover, Germany, in 2007. In 2005 he studied two semesters at the Geomatics Department of the University of Melbourne, Australia, specializing in close range photogrammetry. From 2006 to 2007 he worked one year as an intern at the French space agency CNES in Toulouse, France, mainly dealing with automatic co-registration of very high-resolution airborne optical and SAR data. Since October 2007 he has been working towards the Ph.D. at the Institute of Photogrammetry and GeoInformation, Faculty of Civil Engineering and Geodesy, Leibniz Universität Hannover, Germany.

He is interested in various research fields with focus on automatic analysis of urban areas combining optical and SAR data particularly using context-based probabilistic approaches.

Ronny Hänsch was born in Görlitz, Germany, in 1982. In 2007 he finished his studies of computer science at the Technische Universität Berlin, Germany and graduated with a Diploma degree (Dipl.-Inform.). He worked as research assistant at the Fraunhofer IPK Berlin in the years 2005 and 2006. His work there was mainly concerned with object recognition from images and digital image processing. Since July 2007 he has been working on his Ph.D. thesis which discusses generic object recognition from polarimetric SAR data at the department of Computer Vision and Remote Sensing of the Technische Universität Berlin, Germany. He also worked at Datenmeer GmbH in 2010 to improve basic methods of probability density estimation.

His main scientific interest lies with methods of artificial intelligence and automatic image analysis in general as well as probabilistic methods for generic object recognition in particular.

Antje Thiele received a diploma degree (Dipl.-Ing.) of Geodesy in 2004, from Dresden University of Technology, Germany. From 2005 to 2010 she was research associate with the Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB former FGAN-FOM). In 2009 she moved to the Institute of Photogrammetry and Remote Sensing (IPF) of Karlsruhe Institute of Technology (KIT). Her current research topics are pattern recognition and reconstruction of man-made objects from InSAR data, supporting of forest management by analyzing multi-temporal SAR data, and monitoring of atmospheric water vapor by InSAR data.

Uwe Soergel (M'06) was born in Zell (Mosel), Germany, in 1969. He received the Diploma degree (Dipl.-Ing.) in electrical engineering from the University of Erlangen-Nuremberg, Erlangen, Germany, in 1997, and the Doctoral degree from the Leibniz Universität, Hannover, Germany, in 2003.

From fall 1997 to the end of 2005, he was a Research Associate with the Institute for Optronics and Pattern Recognition, Research Establishment for Applied Sciences, a German research establishment for defense-related studies. He dealt mainly with pattern recognition of manmade objects from remote-sensing imagery, with an emphasis on SAR data. Since January 2006, he has been an Assistant Professor in radar remote sensing at the Institute of Photogrammetry and GeoInformation, Faculty of Civil Engineering and Geodesy, Leibniz Universität, Germany; since January 2010, he has been a Full Professor for radar remote sensing and active systems at the same institute.