SAR-Based Terrain Classification Using Weakly Supervised Hierarchical Markov Aspect Models

Wen Yang, *Member, IEEE*, Dengxin Dai, Bill Triggs, and Gui-Song Xia, *Member, IEEE*

***Abstract*— We introduce the hierarchical Markov aspect model (HMAM), a computationally efficient graphical model for densely labeling large remote sensing images with their underlying terrain classes. HMAM resolves local ambiguities efficiently by combining the benefits of quadtree representations and aspect models—the former incorporate multiscale visual features and hierarchical smoothing to provide improved local label consistency, while the latter sharpen the labelings by focusing them on the classes that are most relevant for the broader local image context. The full HMAM model takes a grid of local hierarchical Markov quadtrees over image patches and augments it by incorporating a probabilistic latent semantic analysis aspect model over a larger local image tile at each level of the quadtree forest. Bag-of-word visual features are extracted for each level and patch, and given these, the parent–child transition probabilities from the quadtree and the label probabilities from the tile-level aspect models, an efficient forwards–backwards inference pass allows local posteriors for the class labels to be obtained for each patch. Variational expectation-maximization is then used to train the complete model from either pixel-level or tile-keyword-level labelings. Experiments on a complete TerraSAR-X synthetic aperture radar terrain map with pixel-level ground truth show that HMAM is both accurate and efficient, providing significantly better results than comparable single-scale aspect models with only a modest increase in training and test complexity. Keyword-level training greatly reduces the cost of providing training data with little loss of accuracy relative to pixel-level training.**

***Index Terms*— Hierarchical Markov aspect model (HMAM), probabilistic latent semantic analysis (PLSA), scene labeling, synthetic aperture radar.**

I. Introduction

THE LAST decade has witnessed an explosion in the number and throughput of airborne and spaceborne terrain sensors using modalities such as the synthetic aperture radar (SAR) [1]. Overwhelming quantities of high-resolution satellite imagery are now available to support accurate earth observations and topographic measurements. Even with modern computers, it is a daunting task to densely label such images with the underlying terrain-type classes. There are three main reasons for this.

1) *Complex and ambiguous image appearance:* Within a single terrain class, objects of different materials or layouts or observed from different perspectives often produce markedly different images. Sensor artifacts such as SAR "speckle" make the interpretation even more difficult, as does the fact that, locally, small regions of imagery are often highly ambiguous. For example, a homogeneous dark region in a SAR image may be calm water, a road surface, or a radar shadow, Fig. 1.
2) *The need for high throughput:* To process the huge quantities of data that are available, very efficient visual features and classifiers are needed. Stringent accuracy requirements and the incorporation of local context to mitigate aperture effects both tend to increase the computational complexity.
3) *The scarcity of labeled training data:* System performance is critically dependent on the amount and accuracy of the available training data. Producing suitable human-supplied annotations can be prohibitively expensive, dangerous, or even impossible. This is especially true when training requires detailed pixel-level labelings.

This paper addresses the challenges of capturing context and simplifying annotation in a high-throughput framework. Regarding context, there are several intuitive desiderata.

1) Images contain information at multiple scales so the method should exploit multiscale visual cues.

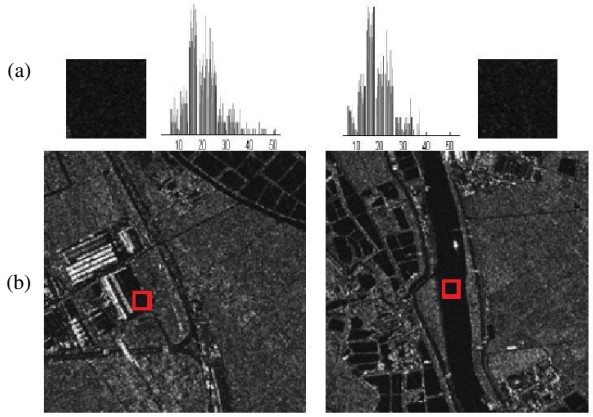

Fig. 1. Ambiguity in SAR images. (a) Two patches of similar appearance and the corresponding intensity histograms. (b) Images containing the patches. One patch is the radar shadow of a building, the other is water.

Manuscript received October 17, 2011; revised March 18, 2012; accepted April 20, 2012. Date of publication May 14, 2012; date of current version August 22, 2012. This work was supported in part by the National Natural Science Foundation of China under Grant 40801183 and the European Commission IST Unit E5 Cognition Research Project 027978 CLASS. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Erhardt Barth.

W. Yang is with the Signal Processing Laboratory, School of Electronics Information and LIESMARS, Wuhan University, Wuhan 430072, China (e-mail: yangwen@whu.edu.cn).

D. Dai is with the Computer Vision Laboratory, Zurich CH-8092, Switzerland (e-mail: dai@vision.ee.ethz.ch).

B. Triggs is with the Laboratoire Jean Kuntzmann, CNRS, Grenoble 38041, France (e-mail: Bill.Triggs@imag.fr).

G.-S. Xia is with the CNRS and CEREMADE, Paris-Dauphine University, Paris 75016, France (e-mail: gsxia.lhi@gmail.com).

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/TIP.2012.2199127

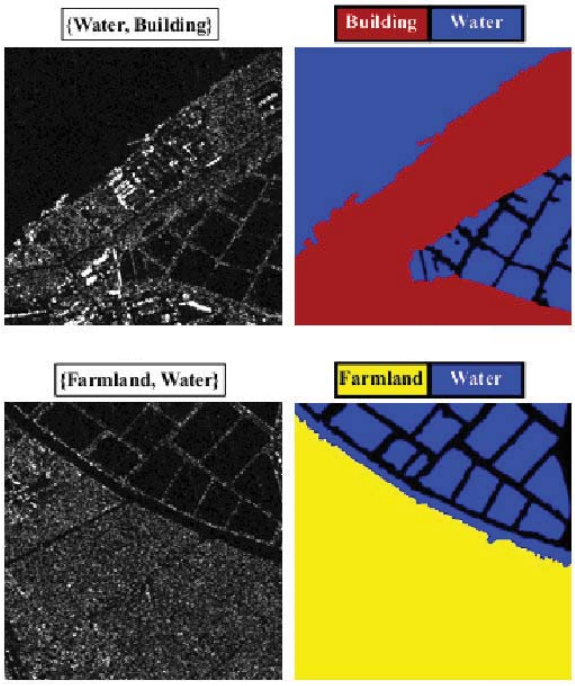

Fig. 2. Two training tiles with (a) keyword-level labeling and (b) corresponding pixel-level labeling.

2) Many scenes are comprised of contiguous regions so adjacent image patches or pixels should be encouraged to take the same label, at least if they have similar appearances.
3) Certain combinations of scene classes tend to co-occur over wider regions, so higher-level contextual cues that capture these relationships should be incorporated.

To our knowledge, the method described here is one of the few to exploit all three sources of consistency.

Regarding annotation, traditional statistical learning methods require a large amount of densely labeled training data to produce an effective pixel-level terrain classifier. Such labeling is very labor intensive and typically requires expert knowledge to compensate for phenomena such as speckle. This limits the availability of training data and hence the quality of the results, and hinders the production of custom labelings for new applications, special needs, etc. Our method allows more cost-effective forms of labeling to be exploited when pixel-level labels are not available. As shown in Fig. 2, we can use "tile-level keywords"—the image is partitioned into tiles of convenient size and for each tile the content classes that appear in it are listed manually, but no pixel-level labelings or tile-level numerical class proportions are needed. We will show that training on weak labelings of this form can provide comparable performance to training on full pixel-level labelings.

To achieve these goals, we augment a conventional quadtree-based hierarchical Markov model with local aspect models. Aspect models [2], [3] are simple statistical "bag-of-words" methods which were originally developed to characterize document collections in terms of a set of hidden underlying "aspects" or "topics" that capture semantic coherencies within the collection (e.g., in a newspaper collection, there might be topics relating to finance, sports, etc.). Aspect models are frequently also used to summarize image content, where the "documents" become images (here, image tiles) and the "words" become "visual words"—vector-quantized visual features over image patches. Aspect models can be viewed as nonnegative probabilistic analogs of principal component analysis (PCA): like PCA, they are usually trained without supervision and they provide both dimensionality reduction and a form of regularization that is useful for subsequent processing. Although we will use supervised forms of training here in which the set of available topics coincides exactly with the set of underlying terrain classes, the fact that aspect models can handle weak supervision will allow us to learn effective models from keyword-level training data.

Our hierarchical Markov aspect model (HMAM) provides dense and efficient terrain-class labeling by exploiting both high-level context (via the tile-level aspect models) and multiscale features and local spatial consistency (via the local quadtrees). It works as follows. The satellite image is partitioned into tiles which are treated as the documents in a document collection. The tiles are divided into square subregions, each of which is partitioned recursively into smaller squares (patches) using a multilevel local quadtree (e.g., [4]). For each patch at each level of a quadtree, a visual descriptor is evaluated and vector-quantized using a level-specific global "visual dictionary" trained using unsupervised k-means. HMAM infers quadtree-leaf-level terrain-class labels from these local quadtrees of visual words by applying conventional hierarchical Markov modeling within each quadtree and regularizing it using tile-wide aspect models defined separately at each quadtree level. The resulting graphical model supports efficient variational inference, using expectation-maximization (EM) for training. A postprocessor takes the posterior class (aspect) distributions for the quadtree-leaf patches, interpolates them to pixel level, and uses the results to provide dense pixel-level image labelings.

We will only evaluate HMAM on SAR-based terrain classification, but we stress that it can be applied to any image labeling problem. Only the visual features and their corresponding dictionaries are SAR-specific.

The rest of this paper is organized as follows. Section II reviews related work. Section III summarizes aspect models and scene-labeling methods based on them. Section IV presents our quadtree-based hierarchical Markov model and shows how we integrate aspect models into it. Section V reports on experiments comparing the accuracies of six classification methods using three types of visual features. Section VI concludes this paper and presents some perspectives for future work.

II. Related Work

The growing practical importance of large-scale terrain mapping is illustrated by the spate of recent publications in this area. Much of the early work focused on modeling low-level pixel distributions, e.g., Gamma, Weibull, and K distributions [5]–[8] and related mathematical techniques [9]–[12].

These approaches are quite effective for low-resolution imagery but they are less well adapted to higher-resolution images because smaller pixels tend to be more uniform and more subject to aperture ambiguities. The remedy is to incorporate visual features and label inference methods that capture more of the local image structure. Regarding features, a number of recent works have adapted computer vision texture descriptors to SAR imagery, often with promising results [13]–[16]. Regarding inference methods, almost every modern machine learning algorithm seems to have been applied to SAR image labeling including neural networks [17], [18], AdaBoost [19], [20], support vector machines [18], [21], and random forests [22], [23]. All of these methods have the potential to provide state-of-the-art labeling accuracies, but, without additional contextual information, both they and the pixel-level statistical models tend to produce noisy labelings with many "salt-and-pepper" inconsistencies. Moreover, such methods typically require a large set of examples with pixel-level labelings for training.

To counter the salt-and-pepper inconsistencies, many methods now incorporate some form of Markov random field (MRF) to promote spatial consistency, often with markedly improved results. For example, Tison *et al*. [24] proposed a classifier for high-resolution SAR images of urban areas, based on a Markov approach that uses a new statistical model based on the Fisher function. Deng *et al*. [25] produced accurate unsupervised segmentation results for SAR sea-ice images by using a function-based parameter to weight the two components of an MRF model. Yang *et al*. [26] developed a hierarchical MRF model based on watershed oversegmentation which gives better results than the corresponding pixel-based MRF. Xia *et al*. [27] produced accurate segmentations using an MRF incorporating a region-adjacency graph, and later [28] proposed a rapid clustering method based on an MRF in the clustering space that uses graph cuts for optimization. Tison *et al*. [29] employed a Markov framework to jointly retrieve a height map and an improved classification from high-resolution interferometric SAR images. Wu *et al*. [30] labeled polarimetric SAR images using a Wishart MRF model to reduce the degree to which speckle produces isolated pixels and small regions. Li *et al*. [31] presented a very promising method for segmenting SAR intensity images that uses Voronoi tessellations and reversible jump Markov chain Monte Carlo Bayesian inference to produce a region-based representation. Ersahin *et al*. [32] used a graph partitioning algorithm to implicitly capture some of the underlying contextual information. However, none of these methods is capable of exploiting high-level context [33] such as the relationships between different kinds of objects and scenes (e.g., the fact that ships usually co-occur with water).

This paper aims to segment and classify SAR terrain images accurately by directly learning category models from keyword-based training data. It is most closely related to the following two approaches. 1) Liénou *et al*. [34] used a latent Dirichlet allocation (LDA) aspect model to semantically annotate large high-resolution satellite images. Their results on panchromatic 60-cm resolution QuickBird images with relatively simple features such as pixel means and standard deviations demonstrated that LDA could lead to satisfactory labeling results. However, although this represents a considerable advance, it only produces coarse labelings covering relatively large image patches (256×256 pixels in their experiments), and its reliance on simple patch overlap to account for spatial context also limits its accuracy. 2) Verbeek and Triggs [35] developed an MRF-based extension to the probabilistic latent semantic analysis (PLSA) aspect model for segmenting personal snapshots captured with hand-held cameras. This method achieves state-of-the-art performance on its datasets, but the application is very different and it incorporates neither multiscale visual cues nor multiscale inference. In related work in image-driven data mining, [36], [37] presented preliminary results on using mined feature-thematic layers for higher level scene understanding, although their main focus was on reducing the computation and memory costs of pixel-block similarity search using σ-tree data structures.

III. Aspect Models and Image Labeling

This section reviews the formulation of aspect models and briefly describes a nonhierarchical variant of our approach that uses a PLSA aspect model. Our full hierarchical model will be described in Section IV.

A. Aspect Models

Aspect models are generative statistical bag-of-words methods that were originally developed for content analysis in textual document collections. They represent documents as combinations of latent "aspects" that denote possible underlying themes or topics. In bag-of-words models, the order of the words in the document is ignored so the document is characterized solely by its distribution of word probabilities over the corpus-wide dictionary. Aspect models regularize this representation by associating a word-distribution to each aspect and forcing the document word distributions to be mixtures of the aspect word distributions. Technically, let variables $z = \{1, \ldots, K\}$ enumerate the K latent "aspects" and $d = \{1, \ldots, N\}$ enumerate the N documents of the corpus D. Given a document d, let w_j denote its jth word, e.g., as an entry number in the dictionary of all words in the corpus. The model supposes that each word w_j of each document d has an associated hidden aspect variable z_j and that the document can be generated by independently sampling each z_j from a document-specific "mixing distribution" over the aspects, $P(z|d)$, and then sampling a word w_j of the dictionary from the z_j member of a corpus-wide set of aspect-specific word distributions $P(w|z)$. Each document can thus be viewed as a "bag" of words sampled independently with probabilities

$$P(w|d) = \sum_z P(w|z)\, P(z|d) \tag{1}$$

and the total log-probability of the document collection D is

$$\log P(D) = \sum_{w,d} n_{wd} \log \left(\sum_z P(w|z)\, P(z|d) \right) \tag{2}$$

where n_{wd} is the number of occurrences of dictionary-word w in document d. In vision applications, the "documents"

become images (here image tiles) and the "words" become "visual words" (vector-quantized visual descriptors of image patches) whose "dictionary" is the k-means codebook used for descriptor quantization.

The document-specific mixture probabilities $P(z|d)$ must be estimated for each incoming document. In the most basic aspect model PLSA [2], [38], these mixtures are unconstrained, whereas in the Bayesian variant LDA [3] they are generated by a common Dirichlet prior over the mixing weights. This provides additional regularization, typically favoring sparser aspect distributions and hence crisper word labeling. However, LDA is computationally demanding and in practice it only improves the accuracy for collections that have small documents and many aspects [39]. This is not the case here, so we will prefer PLSA for its simplicity and computational efficiency. Many other forms of aspect model exist including Harmoniums based on undirected graphs [40], [41] and Pachinko allocation based on hierarchical aspects [42], but variants of our model built on these are beyond the scope of this paper.

PLSA is usually trained using a variant of the EM algorithm [38], [43], alternately estimating the "responsibilities" (E-step)

$$P(z|w,d) = \frac{P(z,w|d)}{P(w|d)} = \frac{P(w|z)\,P(z|d)}{\sum_{z'} P(w|z')\,P(z'|d)} \tag{3}$$

and the factors $P(w|z)$ and $P(z|d)$ (M-step). If we define

$$n_{zwd} \equiv P(z|w,d)\,n_{wd} \tag{4}$$

the M-step becomes simply

$$P(w|z) = \frac{\sum_d n_{zwd}}{\sum_{w',d'} n_{zw'd'}} \tag{5}$$

$$P(z|d) = \frac{\sum_w n_{zwd}}{\sum_{z',w'} n_{z'w'd}}. \tag{6}$$

Note that, if we uniformly select random document words from the training corpus, the resulting empirical joint density is $P_{\text{emp}}(w,d) \propto n_{wd}$, so the n_{zwd} can be viewed as unscaled empirical estimates of $P(z,w,d) = P(z|w,d)\,P(w,d)$.

EM is also used for inference on new documents, alternately estimating the responsibilities $P(z|w,d)$ and the mixing proportions $P(z|d)$, with the $P(w|z)$ being held to their previously trained values. The resulting model both smoothes the empirical word frequencies by forcing them to be representable as mixtures of the $P(w|z)$, and focuses the estimated mixing proportions $P(z|d)$ on the aspects that are most useful for describing the document as a whole.

The discussion above presents unsupervised training, a process similar to clustering in which the K aspects represent latent factors of unknown meaning, with K being an important model parameter. Our application is slightly simpler in that we will always take the aspects to correspond one to one with the scene content classes into which the image is being segmented (building, water, etc.), allowing us to use partially or fully supervised training. In fully supervised training, each image patch ("word") of the training set is labeled with its content class ("aspect") in advance, so the n_{zwd} are given directly and do not need to be estimated with (4), making (5) and (6) closed-form formulae so that EM is not needed during training. In keyword-style weakly supervised training, for each training tile ("document"), the list of "keywords" (content classes/aspects that occur somewhere in the tile) is specified in advance, so we know which elements of $P(z|d)$ are nonzero, but we do not know either their numerical values or the exact locations at which each aspect occurs in the tile. This is handled by running the unsupervised EM algorithm with the known-zero elements of $P(z|d)$ clamped to zero for each document d, i.e., n_{zwd} is set to zero for the z that are known not to occur in d.

B. Aspect Model-Based Labeling of SAR Images

Owing to their huge size (here $48\,189 \times 25\,255$ pixels), it is not usually possible to process complete SAR images as single entities. We handle this by partitioning the image into conveniently sized tiles (e.g., 800×800 pixels), which we treat as independent "documents" in the overall collection (image). In the flat (single-scale nonhierarchical) variant of our aspect models, each tile is simply subdivided into a grid of small nonoverlapping patches, with each patch being represented by the visual word corresponding to its vector-quantized appearance descriptor. We use the terrain categories as the topics, learning the word-to-topic correspondences either explicitly from patch-level training labels or implicitly from keyword-level labels. Whichever training method is used, label inference on new tiles is performed at the patch (visual word) level using aspect mixing proportions $P(z|d)$ obtained by running EM at the tile (document) level with the $P(w|z)$ held fixed. Patch-level labelings can then be produced by posterior likelihood maximization, $\arg\max_z\{P(z|w,d)\}$. To obtain pixel-level labelings, we use bilinear interpolation to propagate the $P(z|w,d)$ to pixel level before the maximization. (Labeling could also be performed directly at the pixel level by including overlapping patches, one based at each pixel: this would probably give slightly better results but it would be very expensive computationally.) The various stages of the training and inference process are illustrated in Fig. 3.

IV. HMAM

Even when tile-level aspect information is included, additional local context often improves the results. In particular, multiscale cues can help to resolve patch-level ambiguities and propagate correlations across the local region. This section describes our local quadtree-based hierarchical Markov model and shows how we combine it with tile-level aspect models to form our full HMAM framework.

A. Hierarchical Markov Models on Quadtrees

Our quadtree representation is similar to that of Laferté *et al.* [4] except that it is built over image patches instead of pixels for computational efficiency. The finest scale (level L) consists of a grid of nonoverlapping $S \times S$ pixel patches. At level $L-1$, squares of 2×2 "child"

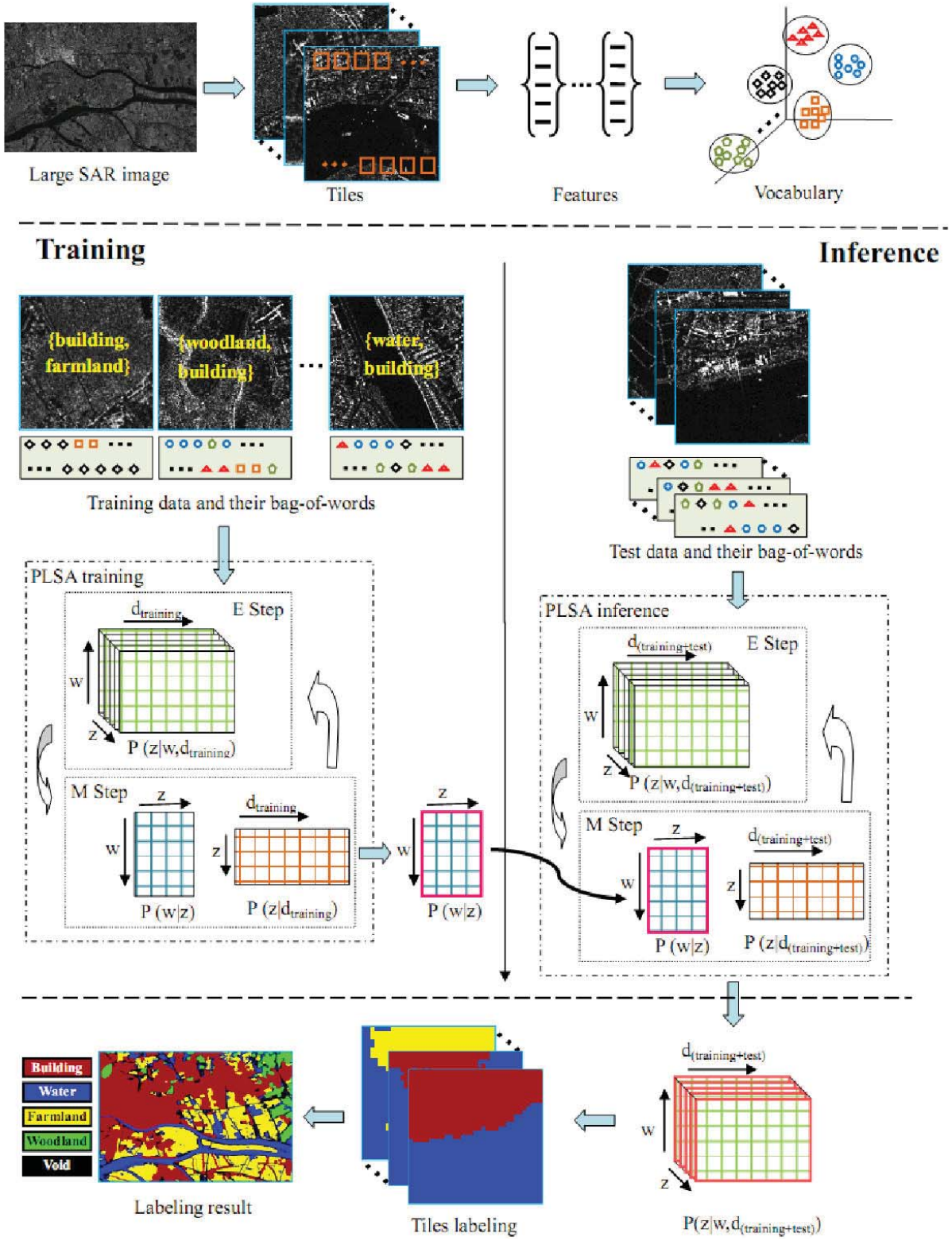

Fig. 3. Outline of the flat (single-scale) variant of our weakly supervised SAR image segmentation and labeling method. The image is partitioned into tiles and each tile is partitioned into local patches that are characterized by "visual words" (vector-quantized local visual features). A PLSA aspect model over the words is trained from tile-level keywords using EM. Given a new tile, the model summarizes its content as a mixture of latent aspects (scene classes) which is then used to determine the most likely label for each individual patch in the tile given its visual word.

patches are merged and subsampled by a factor of two to produce $S \times S$ pixel "parent" patches. This process continues recursively until level 0 is reached, thus producing $L + 1$ levels of quadtree in all. For each patch in each level, a vector of local visual descriptors is extracted and vector-quantized using a level-specific dictionary to produce a visual codeword. The dictionaries are learned using k-means over every patch of that level in the training

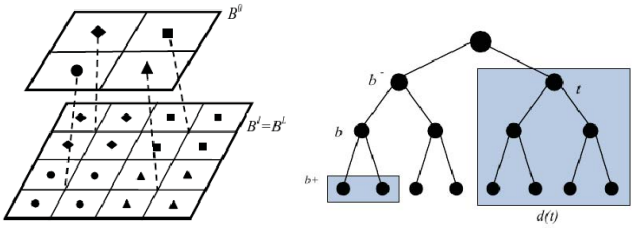

Fig. 4. Local quadtree structure with the corresponding notation.

corpus. The resulting patch-based quadtree is illustrated in Fig. 4.

Let B denote the set of all nodes in the quadtree, and given a node $b \in B$, let b^- denote its unique parent (if it exists), b^+ denote its set of four children (if these exist), $\mathbf{d}(b)$ denote the set of all of its descendants including b itself, and $\bar{\mathbf{d}}(b) = B \setminus \mathbf{d}(b)$ denote all of the remaining nodes in the quadtree. Let $\boldsymbol{w}$ be a vector containing the visual words of all of the patches of the quadtree, with elements w_b indexed by $b \in B$. Let $\boldsymbol{z}$ be a corresponding vector of unknown patch aspect (class) labels, with elements z_b again indexed by $b \in B$.

We assume a hierarchical generative Markov model of the form

$$P(\boldsymbol{z}) = \prod_{b \in B} P(z_b|z_{b^-}) \tag{7}$$

$$P(\boldsymbol{w}|\boldsymbol{z}) = \prod_{b \in B} P(w_b|z_b) \tag{8}$$

where the transition probabilities are free parameters shared across all quadtrees. We will use a default Potts model for the $P(z_b|z_{b^-})$ and learn the $P(w_b|z_b)$ as part of the aspect model.

Inference of $\boldsymbol{z}$ is performed using standard tree-structured belief propagation [44], which finds the marginals $P(z_b|\boldsymbol{w})$ for all $b \in B$ in a single up-and-down pass. Noting that $P(\boldsymbol{w})$ is constant given the observed features $\boldsymbol{w}$ and that $P(z_b, z_{b^-}) = P(z_b|z_{b^-})\, P(z_{b^-})$ and using Bayes rule

$$P(z_{b^-}|\boldsymbol{w}) \propto P(\boldsymbol{w}|z_{b^-}) P(z_{b^-}) \tag{9}$$

$$P(z_b, z_{b^-}|\boldsymbol{w}) \propto P(\boldsymbol{w}|z_b, z_{b^-}) P(z_b|z_{b^-}) P(z_{b^-}). \tag{10}$$

Knowledge of z_b or z_{b^-} is sufficient to factor the $\boldsymbol{w}$'s at and below z_b from all the rest in (8)

$$P(\boldsymbol{w}|z_{b^-}) = P(\boldsymbol{w}_{\mathbf{d}(b)}|z_{b^-})\, P(\boldsymbol{w}_{\bar{\mathbf{d}}(b)}|z_{b^-}) \tag{11}$$

$$P(\boldsymbol{w}|z_b, z_{b^-}) = P(\boldsymbol{w}_{\mathbf{d}(b)}|z_b)\, P(\boldsymbol{w}_{\bar{\mathbf{d}}(b)}|z_{b^-}). \tag{12}$$

Using the factored structure, $P(\boldsymbol{w}_{\mathbf{d}(b)}|z_b)$ can be defined recursively in a bottom-up fashion

$$P(\boldsymbol{w}_{\mathbf{d}(b)}|z_b) = P(w_b|z_b) \prod_{c \in b^+} P(\boldsymbol{w}_{\mathbf{d}(c)}|z_b) \tag{13}$$

where the "message" sent from child c to its parent b is

$$P(\boldsymbol{w}_{\mathbf{d}(c)}|z_b) \equiv \sum_{z_c} P(\boldsymbol{w}_{\mathbf{d}(c)}|z_c)\, P(z_c|z_b). \tag{14}$$

At the top node, this allows us to calculate $P(\boldsymbol{w}|z_0)$ and hence $P(z_0|\boldsymbol{w}) \propto P(\boldsymbol{w}|z_0)\, P(z_0)$. We take the prior $P(z_0)$ to be uniform. From (11) and (9)

$$P(\boldsymbol{w}_{\bar{\mathbf{d}}(b)}|z_{b^-}) P(z_{b^-}) = \frac{P(\boldsymbol{w}|z_{b^-}) P(z_{b^-})}{P(\boldsymbol{w}_{\mathbf{d}(b)}|z_{b^-})} \propto \frac{P(z_{b^-}|\boldsymbol{w})}{P(\boldsymbol{w}_{\mathbf{d}(b)}|z_{b^-})}$$

and combining this with (10) and (12) gives

$$P(z_b|\boldsymbol{w}) \propto \sum_{z_{b^-}} P(z_b, z_{b^-}|\boldsymbol{w}) \tag{15}$$

$$\propto P(\boldsymbol{w}_{\mathbf{d}(b)}|z_b) \sum_{z_{b^-}} \frac{P(z_{b^-}|\boldsymbol{w})}{P(\boldsymbol{w}_{\mathbf{d}(b)}|z_{b^-})} P(z_b|z_{b^-}).$$

Propagating the messages $P(z_{b^-}|\boldsymbol{w})/P(\boldsymbol{w}_{\mathbf{d}(b)}|z_{b^-})$ downward allows all of the $P(z_b|\boldsymbol{w})$ to be calculated. Note that the upward message $P(\boldsymbol{w}_{\mathbf{d}(b)}|z_{b^-})$ needs to be divided out of the naive downward one $P(z_{b^-}|\boldsymbol{w})$ to avoid double-counting the influence of $\boldsymbol{w}_{\mathbf{d}(b)}$.

B. HMAM

Now we incorporate aspect model ideas into the above hierarchical Markov model. At each level l of the quadtree we include a separate PLSA model, having its own dictionary and visual words and hence its own global aspect-to-word probability vectors $P_l(w|z)$ and tile-specific mixture probabilities $P_l(z|d)$, where d denotes the associated tile (i.e., "document"). These models act as priors on $\boldsymbol{z}$ in addition to the existing tree priors $P(z_b|z_{b^-})$. The net effect is to convert the data terms used for tree propagation from $P_l(w_b|z_b)$ to $P_l(w_b|z_b)\, P_l(z_b|d)$, i.e., the tile-specific mixtures regularize the labeling by acting as a form of pseudo-observation on each of the z_b's in their level and tile. With this modification, inference in the tree can still be run as above. During EM, given the current estimates of the tile-specific mixing proportions $P_l(z|d)$ and the global feature distributions $P_l(w|z)$, we estimate the local posteriors

$$P_l(z_b|w_b, d) = \frac{P_l(w_b, z_b|d)}{P_l(w_b|d)} = \frac{P_l(w_b|z_b)\, P_l(z_b|d)}{\sum_{z_b'} P_l(w_b|z_b')\, P_l(z_b'|d)} \tag{16}$$

and use inference in the tree to find the corresponding full posterior marginals $P_l(z_b|\boldsymbol{w}, d)$ (E-step). We then use these as responsibilities to re-estimate the $P_l(w|z)$ and $P_l(z|d)$ (M-step). To speed the convergence, before starting to work with the tree we initialize the $P_l(w|z)$ estimates by fitting flat PLSA models independently at each level l. As before, we take our aspects to correspond one to one to the scene content classes. The final output labels are given by the leaf-patch-level posterior marginals $P_L(z_b|\boldsymbol{w}, d)$, and linear interpolation is again used to refine these to pixel level.

Overall, the local labels z_b in HMAM are thus influenced by three sources of information: 1) the local observation w_b; 2) the surrounding observations $\boldsymbol{w}_{B \setminus b}$ in z_b's quadtree (the multiscale local context), via the Markov propagation; and 3) the complete set of observations in z_b's tile, via the EM based self-consistency iteration that focuses the tile-level "priors" $P_l(z|d)$ on the labels most needed to describe the tile. Note that the various aspect model and quadtree priors are modeled as independent sources of information. This is a variational approximation. In reality these sources are correlated, but the training procedure minimizes the potential overcounting by finding joint parameter settings that reproduce the input as well as it can under the full coupled model.

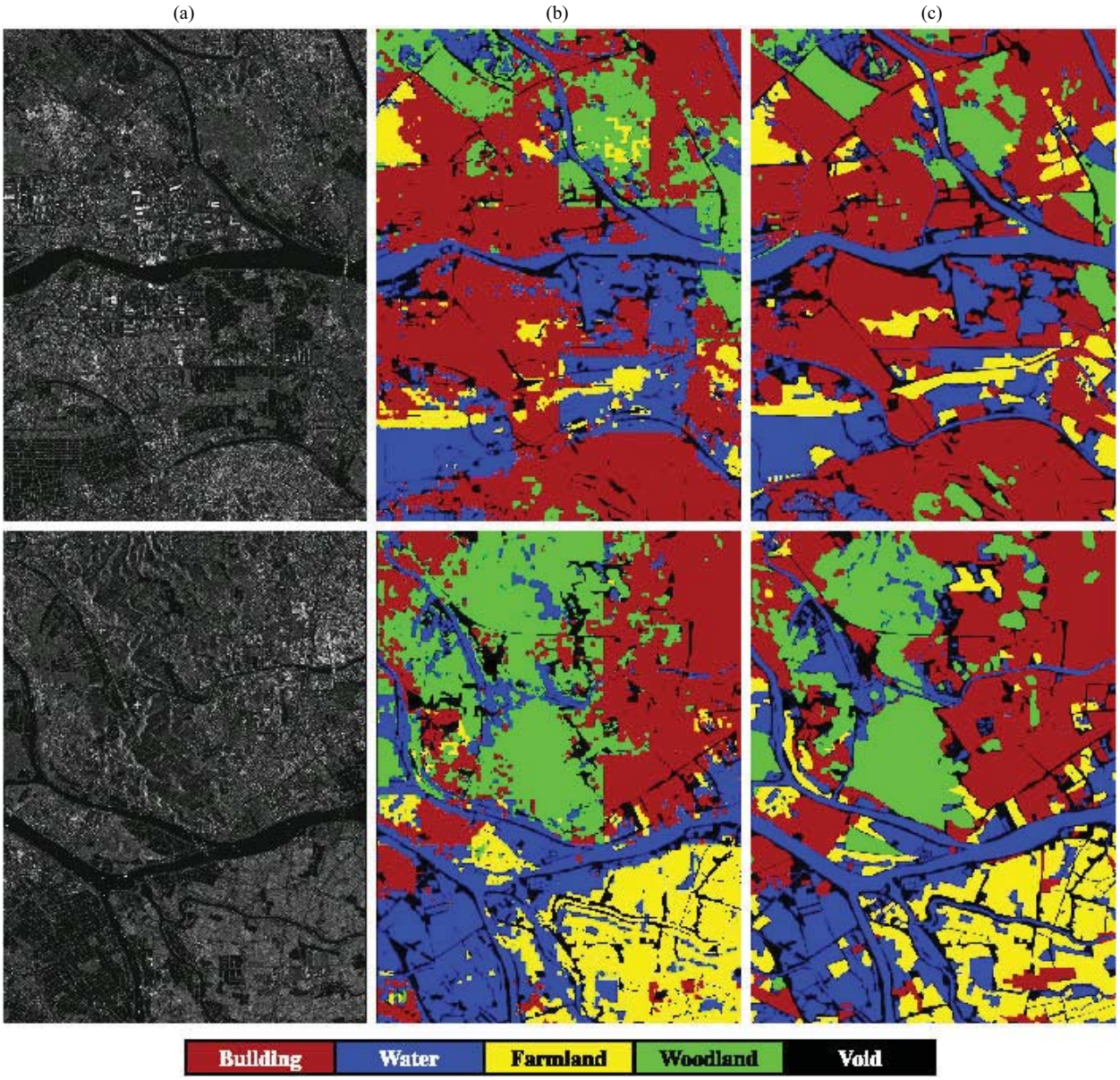

Fig. 5. TerraSAR-X image labeling results. (a) Two 8800 × 6400 pixel regions of the original image. (b) and (c) KHMAM labeling and the ground-truth labeling. To facilitate comparison, labels that are void in the ground truth are also masked out in the KHMAM result.

V. Experiments and Discussion

This section describes our dataset and investigates the accuracy and speed of the proposed labeling methods on it. The flat (AM) and hierarchical (HMAM) aspect models trained on pixel-level labels will be referred to, respectively, as pixel-labeled training based aspect model (PAM) and pixel-labeled training based hierarchical Markov aspect model (PHMAM), while those trained on keyword-annotated tiles will be referred to as keywords-labeled training based aspect model (KAM) and keywords-labeled training based hierarchical Markov aspect model (KHMAM).

A. Datasets and Experimental Settings

Our experiments are based on a 48189 × 25255 pixel whole-scene TerraSAR-X image of Foshan in central Guangdong province, China, acquired on May 24, 2008, in Stripmap mode (Infoterra GmbH/DLR). The spatial resolution is about 3 × 3 m. The pixel-level ground truth was provided manually using associated geographic information. Pixels are assigned to four classes: building, woodland, farmland, and water. Ambiguous pixels (about 13% of the total) are labeled as "void." We ignore the void pixels during both training and evaluation.

In our experiments, the image is partitioned into 1800 nonoverlapping 800 × 800 pixel tiles. In each test, the tiles are randomly partitioned, with 20% being used as training data and the remaining 80% for evaluation. We report average results over 10 random partitions. When training from pixel-level labels, the ground-truth patch label is taken to be the most frequent pixel label within the patch. Test results are propagated to pixel level by linearly interpolating

TABLE I
LABELING ACCURACIES (%) AND THEIR STANDARD DEVIATIONS FOR THE DIFFERENT CLASSIFIERS AND FEATURE SETS

Feature	SVM	PAM	KAM	PHMAM	KHMAM
Gabor	71.2 (3.5)	71.0 (3.6)	69.7 (4.2)	**74.8** (3.6)	73.6 (4.2)
GMRF	75.6 (3.1)	74.2 (3.2)	73.8 (3.9)	**78.4** (3.4)	77.5 (4.0)
Hist	81.3 (3.2)	82.1 (3.3)	80.2 (4.0)	**84.9** (3.5)	83.7 (4.1)

TABLE II
CONFUSION MATRICES (%) FOR KAM AND KHMAM

	True\Est.	Building	Woodland	Farmland	Water
KAM	Building	**84.3**	6.5	4.6	4.6
	Woodland	20.0	**63.8**	13.0	3.2
	Farmland	8.0	6.6	**83.2**	2.2
	Water	7.5	1.5	3.0	**88.0**
KHMAM	Building	**89.1**	3.6	3.9	3.4
	Woodland	20.5	**64.9**	12.1	2.7
	Farmland	8.1	4.9	**83.3**	3.7
	Water	6.3	1.5	3.3	**87.9**

the four neighboring patch-level posteriors to the specific pixel coordinates and taking the most probable class of the resulting posterior as the label.

We tested three visual feature sets that have been widely used for SAR image labeling: Gabor texture descriptors [45], Gauss Markov random fields (GMRF) [46], and 32-bin local pixel amplitude histograms. For GMRF, we used the parameter settings from [47]. Our six-scale, eight-orientation Gabor descriptor is based on the efficient "simple Gabor feature space" of [48].

In each case, descriptors are quantized to visual words using 400 center codebooks obtained by running k-means over all of the training patches at the given quadtree level, with separate codebooks being learned for each level. Three levels ($L = 2$) used for the quadtrees as additional levels produce little further improvement. Among the patch sizes tested, 20×20 pixel leaf patches gave the best overall performance for both the flat and hierarchical models. Larger patches produce more robust global label frequencies at the expense of reduced local accuracies, and vice versa.

The quadtree transition probabilities [4] are taken to be

$$P(z_b = j \mid z_{b^-} = i) = \begin{cases} \alpha, & \text{if } i = j \\ \frac{1-\alpha}{K-1}, & \text{otherwise} \end{cases} \quad (17)$$

where K is the number of scene content classes (here 4) and α is the Potts prior weight. Large α values encourage children and their parents to have the same label, and hence provide more spatial smoothing. We tested values in the range $[0.6, 0.95]$, finally adopting $\alpha = 0.8$.

For comparison, we also give results for two competing methods. In the "Gamma" method, a Gamma distribution is used to model the pixel intensity histogram of each scene content class, and each pixel is classified independently using maximum likelihood. In the "SVM" method, we use the same 20×20 pixel patches and unquantized visual features as our flat methods, with each patch being classified independently using a radial basis function kernel SVM [49]. The model was trained with LibSVM [50],[1] using parameter values $C = 0.4$ and $\sigma = 1.0$ found by cross-validation grid search.

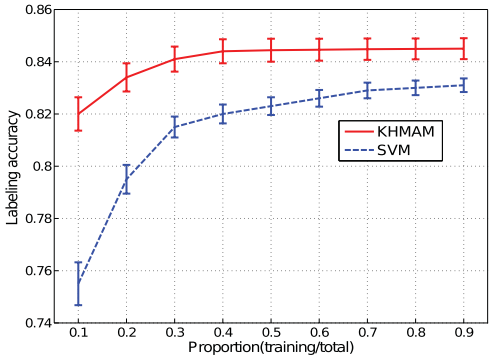

Fig. 6. Labeling accuracy of KHMAM and SVM versus the proportion of the dataset used for training. For easier illustration, the variances shown in this figure are 1/10 of the true values.

B. Qualitative Results on the TerraSAR-X Dataset

Fig. 5 shows some typical labeling results from KHMAM, with the corresponding ground truth for comparison.[2] Each test image has size 8800×6400 and contains 88 (11×8) 800×800 pixel tiles. The results are generally good, but small regions and fine boundary details are sometimes lost and there is occasional evidence of blocking artifacts. These effects can be attributed to two sources. 1) The underlying aspect models strive to interpret each tile with as few aspects as possible. This suppresses many incorrect labels, but it also tends to suppress correct ones when they are rare in the tile. The effect can be reduced by decreasing the tile size, at the expense of noisier labelings (and if keywords are used, increased manual labor when labeling training data). 2) The quadtree-over-patch representation introduces some blocking artifacts. These could be reduced by applying the algorithm at pixel level (with a patch around each pixel) rather than at patch level, and/or by switching to a more uniform spatial representation such as an MRF, but either change would be computationally expensive. The occasional suppression of narrow regions such as canals could be alleviated by incorporating river detectors. However, we emphasize that our main goal here is to recover the major scene categories in large SAR images with minimal human input and low computational cost. Our method's quadtree architecture makes it straightforward to implement and very efficient, and it avoids the need for a prior image segmentation. It would also be possible to run our inference algorithm on more irregular hierarchical graphs obtained from prior segmentations and this might improve the precision, especially for narrow structures. However, it would come at a significantly increased computational cost [51], [52], so we will not test this possibility here.

We now discuss various aspects of our quantitative results.

[1] Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[2] Owing to the random training selection, 20% of the tiles shown in Figs. 5–7 belong to the training set. These tiles are classified and shown in the figures, but excluded from the performance tables.

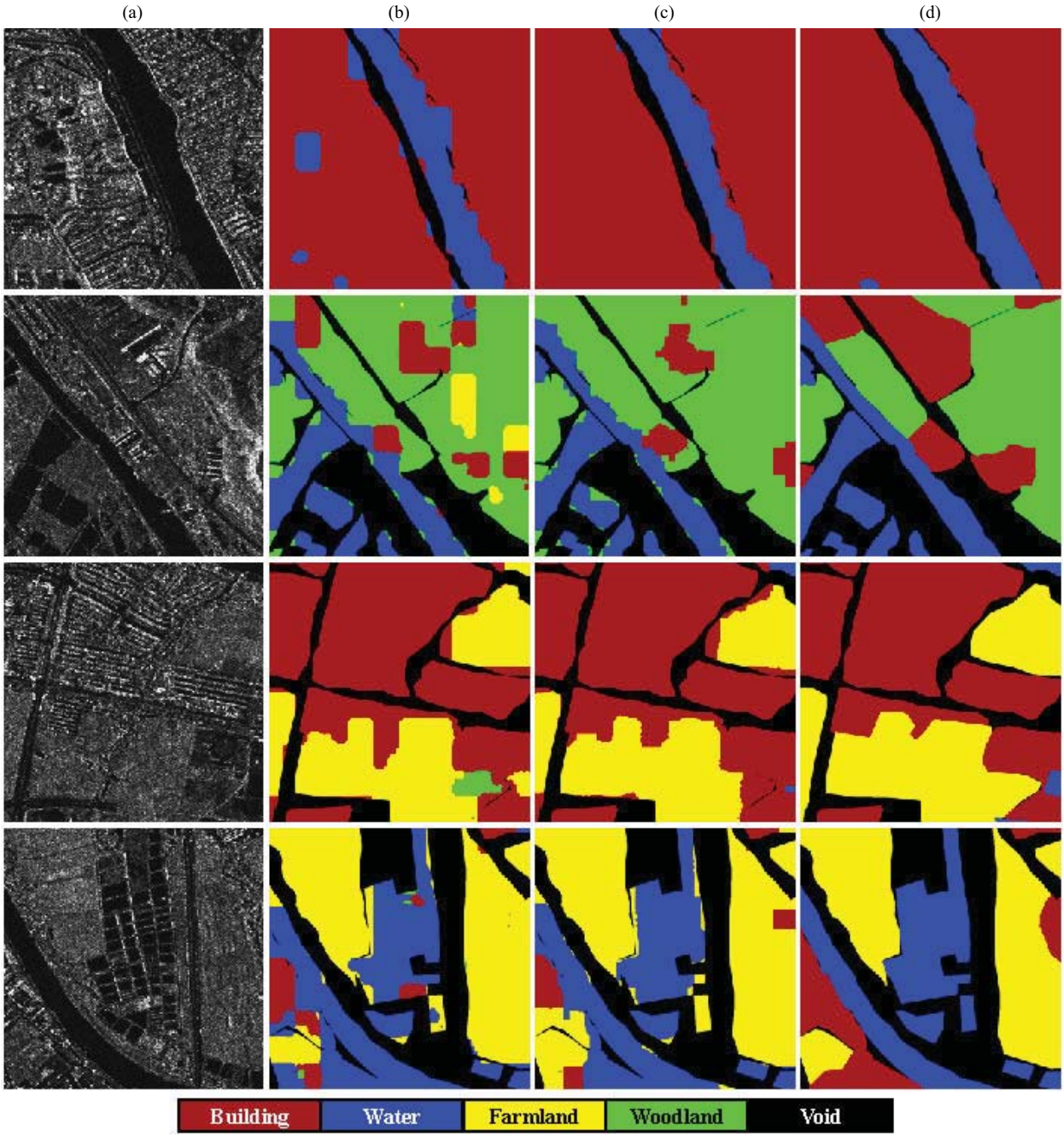

Fig. 7. Comparison of KAM and KHMAM. (a) Four raw 800×800 pixel tiles. (b)–(d) KAM, KHMAM, and ground-truth labelings. For easier comparison, labels that are void in the ground truth are also masked out in the labeling results.

1) Labeling Accuracy: Table I compares the labeling accuracies of the five classifiers for the three feature sets. It reports percentages of image pixels assigned to the correct class, ignoring the ones labeled as void in the ground truth, with standard deviations over the 10 random training–test partitions shown in parentheses. The results show the benefits and complementarity of aspect models and quadtrees. Although the flat aspect models (PAM, KAM) capture enough context to resolve many cases of ambiguity, their HMAM variants (PHMAM, KHMAM) are about 3% better, showing the benefits of including additional local and multiscale context. Keyword-based training (KAM, KHMAM) also does well, giving accuracies that are typically within 1%–2% of the corresponding fully supervised method (PAM, PHMAM), albeit with slightly greater variance. Regarding feature sets, local histograms of pixel amplitudes turn out to be very effective for SAR image segmentation, with 5–10% better accuracy than the Gabor and GMRF texture-based features tested, which are also more complex and slower to compute. We therefore use the amplitude histogram features for the remaining results below.

Overall, the average labeling accuracy obtained is 84.9% for PHMAM with histogram features. In contrast, the pixel-level "Gamma" classifier achieves only 55.6%: the scattering

TABLE III
RUN TIMES FOR THE DIFFERENT ALGORITHMS

Method	Training time (s/tile)	Test time (s/tile)
Gamma	-	<0.01
SVM	15	<0.01
PAM	-	<0.1
KAM	< 0.1	<0.1
PHMAM	<0.05	0.3
KHMAM	0.3	0.3

intensities of individual pixels are too local to provide good classification of high-resolution SAR imagery.

2) Sources of Confusion: Table II gives the interclass confusion matrices for KAM and KHMAM. The visually complex classes appear to be the main source of confusion. In particular, patches containing buildings are difficult to identify reliably—presumably due to their highly variable appearance and tendency to cause radio shadows—and woodland and farmland are sometimes confused—perhaps because there is a continuum between these classes in reality. The incorporation of multiscale information reduces the extent to which building regions are labeled incorrectly, but has little effect on woodland [Table II and Fig. 7(b) and (c)].

3) Ability to Learn from Keyword-Only Data: The fact that keyword-based training achieves accuracies within 1%–2% of pixel-based training is of great practical significance because the manual construction of pixel-level training labels for SAR images is extremely labor-intensive. Keyword-based training will ensure that our methods can be applied to large-scale SAR image classification, and it will also allow more interactive styles of processing where scene classes are defined on the fly by keywords in response to changing user needs.

4) Speed Comparisons: Table III gives timings for our methods and the two benchmarks for amplitude histogram features in our unoptimized MATLAB implementation on a 2.4-GHz Pentium machine with 4 Gb of memory. The results for the other feature sets are similar as all of them use 400-word dictionaries. The table shows that our methods are efficient during both training and testing. PHMAM and KHMAM require about three times more computation than PAM and KAM, but their cost still remains reasonable.

5) Accuracy Versus the Amount of Training Data: Fig. 6 shows how the accuracies of KHMAM and SVM change as the fraction of the dataset used for training varies from 10% to 90%, always using the remaining 10% of the data for testing. The standard deviation of the results over 10 experiments is also shown. For KHMAM, we see that the results saturate at around 40%, but still remain good (about 2.5% less accuracy and a little more variance) if only 10% of the data is used. The curves for PAM, KAM, and PHMAM (not shown) are very similar. Overall, our methods are superior to SVM in both accuracy and stability, especially when small training sets are used. This can probably be attributed to the aspect models' ability to discover and exploit latent structure without supervision: it allows the training data to be used mainly to establish the link between the underlying aspects of the imagery (which exist even without labeling) and the desired scene content classes, so less of it is needed than for training a structure-free discriminant such as SVM.

VI. CONCLUSION

We have addressed the challenge of labeling an entire SAR image with terrain-usage classes, proposing an efficient statistically based method that combines the advantages of two complementary context models: 1) *aspect models*, which provide cleaner labelings by focusing on the classes that are locally most relevant, and which also allow the model to be trained from much weaker data (local keywords rather than detailed pixel-level labelings) and 2) *hierarchical Markov models*, which incorporate multiscale features and adjacency constraints to reduce ambiguity and improve the local contiguity of the labelings. The particular model that we tested—the HMAM—is based on Markov quadtrees over local image patches, with PLSA aspect models over larger image tiles applied to each layer of the quadtrees.

Our experiments tested several types of classifier, including keyword-trained and pixel-trained HMAMs and single-level aspect models over three visual feature sets, on four-class segmentation of a TerraSAR-X image of Foshan in China. Overall, the pixel- and keyword-trained HMAM models over local histogram of pixel amplitude features gave the best results, with keyword-level training giving comparable accuracies to pixel-level training despite its much less labor-intensive labeling requirements.

While the results presented here are encouraging, there is still a need for further improvements. To reduce the blocking artifacts that sometimes appear at tile borders, it may help to use smaller and/or more densely sampled patches or an oversegmentation-based scheme for the final patch-to-pixel label mapping. To further reduce the cost of supplying manual training labels, active learning could be introduced to choose the pixels, patches, or tiles that can most usefully be labeled.

REFERENCES

[1] U. Soergel, *Radar Remote Sensing of Urban Areas*, 1st ed. New York: Springer-Verlag, 2010, pp. 16–277.
[2] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 1999, pp. 50–57.
[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent direchlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. 3, pp. 993–1022, 2003.
[4] J. M. Laferté, P. Pérez, and F. Heitz, "Discrete Markov image modeling and inference on the quadtree," *IEEE Trans. Image Process.*, vol. 9, no. 3, pp. 390–404, Mar. 2000.
[5] C. Oliver, "A model for nonrayleigh scattering statistics," *Opt. Acta, Int. J. Opt.*, vol. 31, no. 6, pp. 701–722, 1984.
[6] A. Lopès, H. Laur, and E. Nezry, "Statistical distribution and texture in multilook and complex SAR images," in *Proc. IEEE Int. Geosci. Remote Sensing Symp.*, May 1990, pp. 2427–2430.
[7] C. Oliver, "Optimum texture estimators for SAR clutter," *J. Phys. D, Appl. Phys.*, vol. 26, no. 11, pp. 1824–1835, 1993.
[8] Y. Delignon, R. Garello, and A. Hillion, "Statistical modeling of ocean SAR images," in *Proc. IEEE Int. Electr. Eng., Radar Sonar Navigat.*, vol. 144, no. 6, pp. 348–354, Dec. 1997.
[9] W. Szajnowski, "Estimators of log-normal distribution parameters," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 13, no. 5, pp. 533–536, Sep. 1977.
[10] J. Sijbers, A. J. den Dekker, P. Scheunders, and D. Van Dyck, "Maximum likelihhod estimation of Rician distribution parameters," *IEEE Trans. Med. Imag.*, vol. 17, no. 3, pp. 357–361, Jun. 1998.

[11] E. E. Kuruoglu and J. Zerubia, "Modelling SAR images with a generalization of the Rayleigh distribution," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 527–533, Apr. 2004.
[12] G. Moser, J. Zerubia, and S. B. Serpico, "SAR amplitude probability density function estimation based on a generalized Gaussian model," *IEEE Trans. Image Process.*, vol. 15, no. 6, pp. 1429–1442, Jun. 2006.
[13] D. A. Clausi, "Comparison and fusion of co-occurrence, Gabor, and MRF texture features for classification of SAR sea ice imagery," *Atmosphere Oceans,* vol. 39, no. 3, pp. 183–194, 2001.
[14] P. Maillard, D. A. Clausi, and H. W. Deng, "Operational map-guided classification of SAR sea ice imagery," *IEEE Trans. Geosci. Remote Sensing*, vol. 43, no. 12, pp. 2940–2951, Dec. 2005.
[15] G. D. De Grandi, J. S. Lee, and D. L. Schuler, "Target detection and texture segmentation in polarimetric SAR images using a wavelet frame: Theoretical aspects," *IEEE Trans. Geosci. Remote Sensing*, vol. 45, no. 11, pp. 3437–3453, Nov. 2007.
[16] V. V. Chamundeeswari, D. Singh, and K. Singh, "An analysis of texture measures in PCA-based unsupervised classification of SAR images," *IEEE Geosci. Remote Sensing Lett.*, vol. 6, no. 2, pp. 214–218, Apr. 2009.
[17] J. A. Karvonen, "Baltic sea ice SAR segmentation and classification using modified pulse-coupled neural networks," *IEEE Trans. Geosci. Remote Sensing*, vol. 42, no. 7, pp. 1566–1574, Jul. 2004.
[18] M. Shimoni, D. Borghys, R. Heremans, C. Perneel, and M. Acheroy, "Fusion of PolSAR and PolInSAR data for land cover classification," *Int. J. Appl. Earth Observat. Geoinf.*, vol. 11, no. 3, pp. 169–180, Jun. 2009.
[19] X. L. She, J. Yang, and W. J. Zhang, "The boosting algorithm with application to polarimetric SAR image classification," in *Proc. 1st Asia-Pacific Conf. Synth. Aperture Radar*, Nov. 2007, pp. 779–783.
[20] J. Chen, Y. Chen, and J. Yang, "A novel supervised classification scheme based on Adaboost for Polarimetric SAR signal processing," in *Proc. 9th Int. Conf. Signal Process.*, Oct. 2008, pp. 2400–2403.
[21] C. Lardeux, P.-L. Frison, C. Tison, J.-C. Souyris, B. Stoll, B. Fruneau, and J.-P. Rudant, "Support vector machine for multifrequency SAR polarimetric data classification," *IEEE Trans. Geosci. Remote Sensing*, vol. 47, no. 12, pp. 4143–4152, Dec. 2009.
[22] W. Yang, T. Zou, D. Dai, and Y. Shuai "Supervised land-cover classification of TerraSAR-X imagery over urban areas using extremely randomized forest," in *Proc. Joint Urban Remote Sensors Event*, Shanghai, China, May 2009, pp. 1–6.
[23] T. Y. Zou, W. Yang, D. X. Dai, and H. Sun, "Polarimetric SAR image classification using multifeatures combination and extremely randomized clustering forests," *EURASIP J. Adv. Signal Process.*, vol. 2010, no. 1, pp. 465612-1–465612-9, 2010.
[24] C. Tison, J.-M. Nicolas, F. Tupin, and H. Maitre, "A new statistical model for Markovian classification of urban areas in high-resolution SAR images," *IEEE Trans. Geosci. Remote Sensing*, vol. 42, no. 10, pp. 2046–2057, Oct. 2004.
[25] H. W. Deng and D. A. Clausi, "Unsupervised segmentation of synthetic aperture radar sea ice imagery using a novel Markov random field model," *IEEE Trans. Geosci. Remote Sensing*, vol. 43, no. 3, pp. 528–538, Mar. 2005.
[26] Y. Yang, H. Sun, and C. He, "Supervised SAR image MPM segmentation based on region-based hierarchical model," *IEEE Geosci. Remote Sensing Lett.*, vol. 3, no. 4, pp. 517–521, Oct. 2006.
[27] G.-S. Xia, C. He, and H. Sun, "Integration SAR image segmentation method using MRF on region adjacency graph," *IET Radar, Sonar Navigat.*, vol. 1, no. 5, pp. 348–354, Oct. 2007.
[28] G.-S. Xia, C. He, and H. Sun, "A rapid and automatic MRF-based clustering method for SAR images," *IEEE Geosci. Remote Sensing Lett.*, vol. 4, no. 4, pp. 596–600, Oct. 2007.
[29] 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 Sensing*, vol. 45, no. 2, pp. 496–505, Feb. 2007.
[30] Y. H. Wu, K. Ji, W. Yu, and Y. Su, "Region-based classification of polarimetric SAR images using wishart MRF," *IEEE Geosci. Remote Sensing Lett.*, vol. 5, no. 4, pp. 668–672, Oct. 2008.
[31] Y. Li, J. Li, and M. A. Chapman, "Segmentation of SAR intensity imagery with a Voronoi tessellation, Bayesian inference, and reversible jump MCMC algorithm," *IEEE Trans. Geosci. Remote Sensing*, vol. 48, no. 4, pp. 1872–1881, Apr. 2010.
[32] K. Ersahin, I. G. Cumming, and R. K. Ward, "Segmentation and classification of polarimetric SAR data using spectral graph partitioning," *IEEE Trans. Geosci. Remote Sensing*, vol. 48, no. 1, pp. 164–174, Jan. 2010.
[33] M. Datcu, S. D'Elia, R. L. King, and L. Bruzzone, "Introduction to the special section on image information mining for earth observation data," *IEEE Trans. Geosci. Remote Sensing*, vol. 45, no. 4, pp. 795–798, Apr. 2007.
[34] M. Liénou, H. Maître, and M. Datcu, "Semantic annotation of satellite images using latent Dirichlet allocation," *IEEE Geosci. Remote Sensing Lett.*, vol. 7, no. 1, pp. 28–32, Jan. 2010.
[35] J. Verbeek and B. Triggs, "Region classification with Markov field aspect models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Minneapolis, MN, Jun. 2007, pp. 1–8.
[36] C. F. Barnes, H. Gritz, and J. Yoo, "Hurricane disaster assessments with image-driven data mining in high-resolution satellite imagery," *IEEE Trans. Geosci. Remote Sensing*, vol. 45, no. 6, pp. 1631–1640, Jun. 2007.
[37] C. F. Barnes, "Image-driven data mining for image content segmentation, classification, and attribution," *IEEE Trans. Geosci. Remote Sensing*, vol. 45, no. 9, pp. 2964–2978, Sep. 2007.
[38] T. Hofmann, "Unsupervised learning by probabilistic latent semantic anaylysis," *Mach. Learn.*, vol. 42, nos. 1–2, pp. 177–196, 2001.
[39] F.-F. Li and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, Jun. 2005, pp. 524–531.
[40] E. Xing, R. Yan, and A. Hauptmann, "Mining associated text and images with dual-wing harmoniums," in *Proc. 21th Annu. Conf. Uncertainty Artif. Intell.*, 2005, pp. 633–641.
[41] J. Yang, R. Yan, Y. Liu, and E. P. Xing, "Harmonium models for video classification," *Stat. Anal. Data Mining*, vol. 1, no. 1, pp. 23–37, Feb. 2008.
[42] W. Li and A. McCallum, "Pachinko allocation: DAG-structured mixture models of topic correlations," in *Proc. 23th Int. Conf. Mach. Learn.*, 2006, pp. 577–584.
[43] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc. Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.
[44] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann, 1988.
[45] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using Gabor filters," *Pattern Recognit.*, vol. 24, no. 12, pp. 1167–1186, 1991.
[46] R. Chellappa, "Two-dimensional discrete Gaussian Markov random field models for image processing," in *Proc. Mach. Intell. Pattern Recognit. Progr. Pattern Recognit.*, 1985, pp. 79–112.
[47] D. A. Clausi, "Comparison and fusion of co-occurrence, Gabor, and MRF texture features for classification of SAR sea ice imagery," *Atmosp. Oceans*, vol. 39, no. 4, pp. 183–194, 2001.
[48] V. Kyrki, J. K. Kämäräinen, and H. Kälviäinen, "Simple Gabor feature space for invariant object recognition," *Pattern Recognit. Lett.*, vol. 25, no. 3 pp. 311–318, Feb. 2004.
[49] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–291, Sep. 1995.
[50] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *J. ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–39, Apr. 2011.
[51] A. Katartzis, I. Vanhamel, and H. Sahli, "A hierarchical Markovian model for multiscale region-based classification of vector-valued images," *IEEE Trans. Geosci. Remote Sensing*, vol. 43, no. 3, pp. 548–558, Mar. 2005.
[52] S. Martinis and A. Twele, "A hierarchical spatio-temporal Markov model for improved flood mapping using multi-temporal X-band SAR data," *Remote Sensing*, vol. 2, no. 9, pp. 2240–2258, 2010.

Wen Yang (M'09) received the B.Sc. degree in electronic engineering, the M.Sc. degree in computer science, and the Ph.D. degree in communication and information systems from Wuhan University, Wuhan, China, in 1998, 2001, and 2004, respectively.

He was a Visiting Scholar with Apprentissage et Interfaces Team, Laboratoire Jean Kuntzmann, Grenoble, France, from 2008 to 2009. He is currently an Associate Professor with the School of Electronic Information, Wuhan University. His current research interests include image segmentation and classification, target detection and recognition, machine learning, and data mining with applications to remote sensing.

Dengxin Dai received the B.Sc. degree in optical information science and technology and the M.Sc. degree in signal and information processing from Wuhan University, Wuhan, China, in 2008 and 2010, respectively. He is currently pursuing the Ph.D. degree with the Computer Vision Laboratory, ETH Zurich, Zurich, Switzerland.

His current research interests include procedural methods for large-scale city creation and reconstruction, image classification, and semantic image labeling.

Bill Triggs is a CNRS Senior Researcher and an Ex Mathematical Physicist, currently working with the Laboratoire Jean Kuntzmann, Grenoble, France. He coordinated the European Commission (EC) Research Project CLASS on unsupervised image and text understanding. He is one of the founding members the EC's PASCAL networks of excellence on machine learning and statistical modeling. His current research interests include machine-learning-based approaches to understanding images and other sensed data.

Gui-Song Xia (M'10) received the B.Sc. degree in electronic engineering and the M.Sc. degree in signal processing from Wuhan University, Wuhan, China, in 2005 and 2007, respectively, and the Ph.D. degree in image processing and computer vision from the CNRS LTCI, TELECOM ParisTech, Paris, France, in 2011.

He is currently a Post-Doctoral Researcher with the CNRS CEREMADE Laboratory, Paris-Dauphine University, Paris. His current research interests include mathematical image modeling, texture synthesis, image indexing and content-based retrieval, perceptual grouping, and remote sensing imaging.