

A Tutorial on Modeling and Inference in Undirected Graphical Models for Hyperspectral Image Analysis

Utsav B. Gewali and Sildomar T. Monteiro

Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology,
Rochester, NY
Email: ubg9540@rit.edu

Abstract

Undirected graphical models have been successfully used to jointly model the spatial and the spectral dependences in earth observing hyperspectral images. They produce less noisy, smooth, and spatially coherent land cover maps and give top accuracies on many datasets. Moreover, they can easily be combined with other state-of-the-art approaches, such as deep learning. This has made them an essential tool for remote sensing researchers and practitioners. However, graphical models have not been easily accessible to the larger remote sensing community as they are not discussed in standard remote sensing textbooks and not included in the popular remote sensing software and toolboxes. In this tutorial, we provide a theoretical introduction to Markov random fields and conditional random fields based spatial-spectral classification for land cover mapping along with a detailed step-by-step practical guide on applying these methods using freely available software. Furthermore, the discussed methods are benchmarked on four public hyperspectral datasets for a fair comparison among themselves and easy comparison with the vast number of methods in literature which use the same datasets. The source code necessary to reproduce all the results in the paper is published on-line to make it easier for the readers to apply these techniques to different remote sensing problems.

Index terms— Spatial-spectral classification; Graphical models; Markov random fields; Conditional random fields; Hyperspectral imaging; A tutorial

1 Introduction

Land cover mapping (also called land cover classification and land cover segmentation) is the process of identifying the materials under each pixel of a spaceborne or an airborne image to create a map showing the spatial distribution of materials over the imaged geographical region. Hyperspectral imaging is an important technology for land cover mapping as it allows for the separation of scene materials into finer classes, compared to other sensing modalities, such as panchromatic, synthetic aperture radar, and multispectral imaging. This is because hyperspectral image at each pixel captures more information about the chemical properties of the materials by recording the reflectance/radiance at hundreds of narrow contiguous bands over the visible and infrared region. Due to its advantages, hyperspectral land cover mapping have been applied to a variety of problems such as separating various species of trees in the forest [16], identifying roads, buildings, trees, and other land covers in urban areas [32], mapping the presence of minerals in soil and rocks [58], differentiating weeds from crops in agricultural land [50], and studying the change in an area by comparing land cover maps at multiple dates [62].

The radiance or reflectance spectrum at each pixel of a hyperspectral image captures the interaction between light and the material, which is dependent on the atomic and the molecular structure of the material and can be used as a signature to discriminate different classes of materials. Traditional classifiers only utilized the spectrum at the pixel to determine the class of the pixel. These are called pixel-wise classifiers. Pixel-wise classifiers can be hand-designed by human experts, such as USGS's tetracoder [15], or be statistical and machine learning based, such as support vector machines (SVM) [51], and random forests [30]. Classifiers can also be supervised (requiring a set of labeled sample spectra belonging to each materials in the scene *a priori*) or unsupervised (not requiring any labeled spectra *a priori*). Pixel-wise classifiers tend to produce noisy and spatially incoherent land cover maps due to the spatial variations in

illumination, shadows, purity of pixels, viewing geometry, atmospheric conditions, and noise across the image. This problem can be alleviated by combining spatial contextual information with the spectral information.

1.1 Spatial-spectral classification

Spatial information can be utilized together with spectral information to produce more accurate and spatially coherent land cover maps [24]. Land covers in the environment tend to be much larger than the ground pixel size of the sensors leading to regions of pixels belonging to a common material class. Additionally, some land cover classes are more likely to exist in close vicinity than others and some land cover classes are highly unlikely to occur together. This leads to strong relationships between the neighboring pixel labels in an image. For example, if a pixel belongs to a class, say building, there is a high probability that the surrounding pixels also belong to the same class, building. Similarly, the probability of neighboring pixels of a building pixel belonging to the road class or the parking lot class is typically much higher than them belonging to the forest class or the bare soil class in an urban scene. These kind of relationships can be exploited by spatial-spectral classifiers. Even though there are exceptions, e.g., [12], [82], and [45], the vast majority of spatial-spectral classifiers can be categorized into two distinct groups—methods that perform spatial-spectral feature extraction followed by pixel-wise classification and methods that combine undirected graphical model and pixel-wise classification.

Spatial-spectral feature extraction utilizes the spectra of the neighborhood of pixels around the pixel to compute feature for the pixel. The features are designed to simultaneously capture the spatial context and be highly discriminative. Classical approach used hand-designed spatial-spectral features, such as co-occurrence matrices [31] and extended morphological features [6]. Modern methods utilized feature representation which is learned from the data itself in supervised or unsupervised manner using sparse dictionary learning and deep learning approaches, e.g., [23], [14], and [88].

Undirected graphical models (UGMs) [36] are powerful and flexible probabilistic models that can represent the complex relationships occurring between the different scene elements in hyperspectral images [75, 95]. They can be combined with pixel-wise classifiers (with or without spatial-spectral features) to enforce spatial contexts. There are other ad-hoc methods, such as majority voting over segmented image [74], that can be used for spatial contexts. However, compared to them UGMs are more principled approach which can model more complex relationships and has sound theory, theoretical guarantees, and efficient inference algorithms. In this tutorial, we provide an introduction to the theory of undirected graphical models, review graphical model based spatial-spectral classification methods published in literature, show how to implement graphical model based spatial-spectral classifiers using freely available toolboxes, and benchmark the discussed methods on four public hyperspectral datasets. We experiment with different supervised machine learning based classifiers and two classes of UGMs. Since popular remote sensing software, such as ENVI, do not include UGM based processing, we have published the code necessary to reproduce all the results in this paper here¹.

This tutorial is organized as follows. Section 2 reviews UGM based methods in remote sensing literature, Section 3 provides a brief background on the UGM theory and inference algorithms, Section 4 discusses the benchmarking experiments and results, and Section 5 summarizes the tutorial.

2 Undirected graphical models in remote sensing

Since their introduction to remote sensing in [69] and [68], undirected graphical models have been widely used to model spatial dependencies in remotely sensed images for land cover mapping. Classical approaches utilize a grid-structured pairwise pixel-based Markov random field (MRF) to model the pixel dependencies and use optimization algorithms, such as iterated conditional mode (ICM), simulated annealing, and graph cuts for inference, see the reviews by [66] and [56]. These models are defined over a grid-structured graph with each node representing a pixel label and edges present between 4-connected neighboring pixels. The unary potentials are defined at each node and captures the spectral information while the pairwise potentials are defined at edges connecting neighboring pixels and captures the spatial information. Since these models only contain unary and pairwise interactions, they are called pairwise models. Unary potentials are derived from pixel-wise classifiers, such as Gaussian maximum likelihood classifier [34], logistic regressions [41], probabilistic support vector machines [75], Gaussian mixture models [44], Gaussian processes [87], and ensemble methods [53, 86, 39]. Potts model (including

¹<https://github.com/UBGewali/tutorial-UGM-hyperspectral>

its contrast sensitive version) is the most popular pairwise potential function. However, an edge based Potts potential function [75], which only encourages neighboring pixels to have same label only if there is no edge in the intensity between them, has shown to produce better results. The hyperparameters of potential functions are mostly chosen by grid search over validation set. However, optimization schemes such as genetic algorithm [77], HoKashyap method [55], and Bayesian optimization [27] have shown to perform better than grid search.

Conditional random fields (CRF) are type of MRF that model conditional distributions by making potential functions dependent on the input features. [95] first utilized a CRF to map the land covers in hyperspectral images. The standard CRF uses with log-linear potential functions which have many parameters that have to be learned from the data. Hence, large number of training pixels is required while using CRF for land cover mapping. This problem has been tackled by modifying the CRF to use simpler functions for pairwise potentials. Some of the proposed pairwise potential functions are based on Euclidean distance between spectra [89], Mahalanobis distance between spectra [97], and Euclidean distance between pixel coordinated and spectra combined with class label cost [93]. These methods utilize grid-structured graph and learn parameters by maximum likelihood estimation. These methods utilize loopy belief propagation during learning and inference.

Though not as common as supervised classification, undirected graphical models have been combined with unsupervised classifiers, such as k-means and ISODATA, for mapping ground covers without any user supplied ground truth [43]. Tree-structured pairwise Markov random field which performs binary segmentation recursively at each level of the tree has been proposed for unsupervised classification [17]. Active learning strategies have also been developed for UGM based spatial-spectral land cover mapping. These methods iteratively select pixels from a set of unlabeled pixels based on some strategy and ask user to manually label them so that those pixels can be added to the training set to further improve the model. [71] proposed selecting the unlabeled pixels which are differently labeled by the pixel-wise classifier and the combination of pixel-wise classifier and UGM. Similarly, [40, 42] defined heuristics, such as breaking ties, over the posterior class probabilities predicted by the random field to select the appropriate unlabeled pixels. Methods that combine results from ensembles of UGMs trained on different features [94] or ensembles of UGMs with different forms of unary potential functions [98] also been published.

Higher order graphical models contain potential functions defined over more than two nodes and are more expressive compared to the grid-structured pairwise models with unary and pairwise potentials. Since, they include potential functions over group of pixels rather than just individual pixels, they can model complex dependencies between various regions, structures, and objects in the image. [96] proposed using robust P^n model for higher order modeling of hyperspectral images. In this method, the image is first segmented using a clustering (unsupervised classification) algorithm. Then, higher order potentials are defined over each segment to encourage all the pixels in each segment to be assigned the same label, which are used along with the unary and pairwise potentials. The P^n model has been widely used for land cover mapping using other remote sensing modalities as well [54, 83, 38, 59]. Similarly, [91] proposed a novel higher order potential over each segment based on the distance between the segments and the similarity between the pixels in the segments. Higher order relationships can also be modeled the simple pairwise model if each node of the graph is representing a label for the entire segment (group of pixels) rather than a single pixel [65, 90, 79]. These methods are less expressive than higher order potential based methods in that all the pixels in each segment are strictly assigned to the same label. In a different application, [2] used two CRF layers, one to model land cover (broader classes such as buildings, grasses etc.) and another to model land use (finer classes such as residential buildings, non-residential buildings, urban grass, grass land etc.) simultaneously in aerial imagery using intra-layer and inter-layer interactions.

Undirected graphical models have also been used for sub-pixel mapping of remote sensing images [35, 92, 81]. These methods produce land cover maps at a scale smaller than the size of a image pixel. This is done by estimating the contents of pixels using classification or unmixing, using the derived material proportions as the unary potentials of UGM at finer resolution, and using pairwise potentials at finer resolution to enforce spatial contexts.

Various studies [69, 47, 52] have also utilized three-dimensional grid-structured UGMs for modeling spatio-temporal dependencies in a time-series of multispectral satellite images for land cover classification and change detection. Apart from land cover mapping, undirected graphical models have also been used to model the spatially dependent parameters of Bayesian hyperspectral unmixing frameworks [3, 4, 21, 20]. Continuous Markov random fields have been successfully used to model textures in hyperspectral images, e.g. [64], however they are beyond the scope of this tutorial.

3 Background

Undirected graphical models define the joint distribution of a set of variables over the structure of an undirected graph [36, 60]. The nodes of the undirected graph represent the variables while the edges between the nodes express the conditional independence relationship between the variables.

Let $\mathbf{y} = [y_1, \dots, y_N]$ be a vector of N variables whose joint probability distribution is defined over an undirected graph G such that following conditional independence relationships are true.

Local Markov property Each node is conditionally independent of all of the other nodes given its neighboring nodes.

$$p(y_i | \mathbf{y}_{\setminus i}) = p(y_i | \mathcal{N}(y_i)), \quad (1)$$

where $\mathcal{N}(y_i)$ is the set of neighbors of y_i .

Global Markov property Two nodes are conditionally independent if all the path between them along the edges in the graph is blocked by an observed node.

$$p(y_i | y_j, y_S) = p(y_i | y_S), \quad (2)$$

where y_S are the set of nodes separating y_i and y_j in G .

Then, the Hammersley-Clifford theorem [7] states that the joint distribution of the variables y_1, \dots, y_N can be factorized as

$$p(\mathbf{y}) = \frac{1}{Z} \prod_{C \in \mathcal{C}(G)} \psi_C(\mathbf{y}_C), \quad (3)$$

where $\mathcal{C}(G)$ is the set of all the cliques of G . A clique (also called maximal subgraphs) is a subset of nodes of a graph that has an edge between every pair of nodes. \mathbf{y}_C denotes a vector of all of the nodes inside the clique C . The functions $\psi_C(\mathbf{y}_C)$ are arbitrary non-negative functions that define the interaction between the variables inside the clique C and are called potential functions. Z is a normalizing constant given by

$$Z = \sum_{\mathbf{y} \in \mathcal{Y}} \prod_{C \in \mathcal{C}(G)} \psi_C(\mathbf{y}_C), \quad (4)$$

and is called a partition function. The partition function computes the sum of the product of potential functions over the set of all possible configurations of the variables y_1, \dots, y_N , denoted by \mathcal{Y} . Division by the partition function makes the product of potential functions a valid probability that sums to one.

In this way, depending upon the structure of the undirected graph G and the form of the potential functions $\psi_C(\mathbf{y}_C)$, UGMs can model a wide variety of families of probability distributions over the variables y_1, \dots, y_N . UGMs can represent the probability distribution of both real and discrete variables, however this tutorial only focuses on discrete UGMs. Also, though we include a broad introduction to UGMs in the tutorial, only pairwise models are explored in detail in explanation and experimentation. It is because pairwise models are the most widely tested and established models in remote sensing. Also, a good understanding of pairwise models is essential to understand the newer, more complex, higher order models.

3.1 Markov random fields

Markov random fields (MRF) is another name for UGMs [57, 9]. However, in this tutorial, similar to many literature, the term MRF is primarily used to denote models representing unconditional distributions (models not conditioned on input features), in order to contrast them with the conditional random fields (CRF).

It is very common to represent (3) in terms of energies by choosing the potentials to be of exponential family, $\psi_C(\mathbf{y}_C | \mathbf{w}) = \exp(-E_C(\mathbf{y}_C | \mathbf{w}))$ where $E_C(\mathbf{y}_C; \mathbf{w}) = -\log(\psi_C(\mathbf{y}_C | \mathbf{w}))$ is the clique energy function. Since, the potential functions contain free parameters in practice, the potential functions and the energy functions have been parameterized with parameter vector \mathbf{w} . Then, the joint probability of the MRF is

$$p(\mathbf{y} | \mathbf{w}) = \frac{1}{Z(\mathbf{w})} \exp(-E(\mathbf{y}; \mathbf{w})), \quad (5)$$

where $E(\mathbf{y}; \mathbf{w}) = \sum_{C \in \mathcal{C}(G)} E_C(\mathbf{y}_C; \mathbf{w})$ is the total energy and $Z(\mathbf{w}) = \sum_{\mathcal{Y}} \exp(-E(\mathbf{y}; \mathbf{w}))$.

Pairwise MRF is the simplest MRF formulation which expresses the total energy as the sum of unary energies and pairwise energies. The unary energy is defined for all the nodes and the pairwise energy is defined for all the edges in the graph. Each node has a different unary energy based on the value assigned to it, with the likely assignments having lower energy. Similarly, each edge exhibits different pairwise energy for different configuration of possible values of the two nodes at its ends, with likely configurations having lower energy. The total energy in a pairwise MRF is

$$E(\mathbf{y}; \mathbf{w}) = \sum_{i \in V} E_i(y_i; \mathbf{w}_1) + \sum_{(i,j) \in D} E_{ij}(y_i, y_j; \mathbf{w}_2), \quad (6)$$

where $E_i(y_i; \mathbf{w}_1)$ is the unary energy of the i^{th} variable when its value is y_i and $E_{ij}(y_i, y_j; \mathbf{w}_2)$ is the pairwise energy between the i^{th} and the j^{th} variables when their values are y_i and y_j respectively. V is the set of all nodes and D is the set of all edges in the graph G . $\mathbf{w} = [\mathbf{w}_1, \mathbf{w}_2]$ are the parameters of the energy functions.

3.2 Conditional random fields

The conditional random field (CRF) [72] is a type of MRF whose clique potentials are conditioned on input features. It is a discriminative version of MRF that models $p(\mathbf{y}|\mathbf{x})$ instead of $p(\mathbf{y})$, where $\mathbf{x} = [x_1, \dots, x_N]$ are the input features for $\mathbf{y} = [y_1, \dots, y_N]$, as

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}) = \frac{1}{Z(\mathbf{x}, \mathbf{w})} \prod_{C \in \mathcal{C}(G)} \psi_C(\mathbf{y}_C | \mathbf{x}, \mathbf{w}), \quad (7)$$

where $Z(\mathbf{x}, \mathbf{w}) = \sum_{\mathbf{y} \in \mathcal{Y}} \prod_{C \in \mathcal{C}(G)} \psi_C(\mathbf{y}_C | \mathbf{x}, \mathbf{w})$ and \mathbf{w} is the vector of potential function parameters. The potential functions of CRFs are most commonly represented by the log-linear function as $\psi_C(\mathbf{y}_C | \mathbf{x}, \mathbf{w}) = \exp(\mathbf{w}_C^T \phi(\mathbf{x}_C, \mathbf{y}_C))$, where $\phi(\mathbf{x}_C, \mathbf{y}_C)$ is a feature function and \mathbf{w}_C is a weight vector. In terms of energies, the clique energy $E_C(\mathbf{y}_C; \mathbf{x}, \mathbf{w}) = -\mathbf{w}_C^T \phi(\mathbf{x}_C, \mathbf{y}_C)$. The feature function produces an arbitrary length vector of features dependent on \mathbf{x}_C and \mathbf{y}_C .

Similar to MRF, a pairwise CRF model can be defined as

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}) = \frac{1}{Z(\mathbf{x}, \mathbf{w})} \exp \left(\sum_{i \in V} \mathbf{w}_1^T \phi_1(x_i, y_i) + \sum_{(i,j) \in D} \mathbf{w}_{y_i, y_j}^T \phi_2(x_i, x_j) \right), \quad (8)$$

where $\mathbf{w} = \{\mathbf{w}_1, \mathbf{w}_{y_i, y_j}\}$ are the parameters and $Z(\mathbf{x}, \mathbf{w})$ is the partition function. $\phi_1(x_i, y_i)$ is the unary feature function for i^{th} variable as a function of the input feature x_i and label y_i . $\phi_2(x_i, x_j)$ is the pairwise feature function between the i^{th} and j^{th} variables, y_i and y_j , as a function of inputs features x_i and x_j . Separate parameter vectors are defined for each possible combinations of y_i and y_j , represented by \mathbf{w}_{y_i, y_j}^T . The pairwise energy is obtained by multiplying the pairwise feature vector by appropriate pairwise weight vector based on the values of y_i and y_j . The weight vector and the feature functions can be function can be function of the node index and be different at different nodes, however in the above formulation it is assumed that they are same across the graph.

The main advantages of CRFs over MRFs is that being discriminative model rather than generative model they can better use data for classification and that potentials in CRFs can be made more data-dependent than MRFs due to the use of input features, while the main disadvantage of CRF is that they require larger training data and longer training time [57].

3.3 Parameter learning

Since MRF typically have very few parameters it is very common to tune the parameters of MRF by grid-search over validation set. The parameters of CRF cannot be tuned in this manner as they are substantially large in number. Parameters of CRF can be learned by maximum likelihood estimation [60]. The log-likelihood of the CRF is

$$l(\mathbf{w}) = \frac{1}{N} \sum_i p(\mathbf{y}^{(i)} | \mathbf{x}^{(i)}, \mathbf{w}), \quad (9)$$

where N is the number of samples and $(\mathbf{y}^{(i)}, \mathbf{x}^{(i)})$ is the i^{th} training pair. The gradient of log-likelihood with respect to clique parameters [57] is given by

$$\frac{\partial l(\mathbf{w})}{\partial \mathbf{w}_C} = \frac{1}{N} \sum_i \left[\psi_C(\mathbf{y}_C^{(i)} | \mathbf{x}^{(i)}, \mathbf{w}) - \mathbb{E}_{\mathbf{y}_C} [\psi_C(\mathbf{y}_C | \mathbf{x}^{(i)}, \mathbf{w})] \right]. \quad (10)$$

The log-likelihood function can be maximized using a gradient based optimizer. The second term in the derivative calculates expectation over marginal probability of the clique. Marginal probabilities can be estimated by the inference methods discussed below. Since, each iteration of gradient optimization requires performing inference once, maximum likelihood parameter estimation is computationally expensive for graphical models. There are other alternatives for parameter learning, such as maximizing pseudo-likelihood [8] and maximum margin learning [76]. Methods like maximum likelihood and maximum pseudo-likelihood can be used for parameter estimation in MRF but maximum margin learning is only for CRFs.

3.4 Inference

The size for the solution space \mathcal{Y} for a discrete undirected graphical model of N variables where each of the N variables can take M distinct value is M^N . Hence, brute-force inference by enumerating cost of all configurations of the variables is not computationally feasible, unless the graph is very small.

There are two popular inference approaches for undirected graphical models—maximum a posteriori (MAP) inference and probabilistic inference (also called marginal inference). Both inference approaches are NP-hard for general graphs and arbitrary potential functions however for restricted graph structure and potential function, exact inference is tractable. For example, for graphs with no loops such as chains and trees exact inference is possible by method called belief propagation [61]. Similarly, Graph-cuts can be used to efficiently find the exact MAP inference in pairwise graphical model of binary variables if the total energy function is sub-modular [28, 37]. For cases where exact inference is not tractable, a variety of efficient approximate algorithms have been developed. The readers are encouraged to read [60] for in-depth coverage of exact and approximate inference algorithms.

MAP inference The MAP inference finds the configuration of \mathbf{y} that maximizes the joint probability or equivalently minimizes the total energy as

$$\mathbf{y}^* = \arg \max_{\mathbf{y} \in \mathcal{Y}} p(\mathbf{y}; \mathbf{x}, \mathbf{w}) = \arg \min_{\mathbf{y} \in \mathcal{Y}} E(\mathbf{y}; \mathbf{x}, \mathbf{w}). \quad (11)$$

Some of the common algorithms used for MAP inference are iterated conditional mode (ICM), simulated annealing, graph cuts and move-making algorithms, belief propagations (including loopy and tree reweighted belief propagations), Markov chain Monte Carlo, and linear programming relaxations.

Probabilistic inference The probabilistic inference finds the value of the log partition function and the marginal probabilities of the cliques:

1. $\log Z(\mathbf{x}, \mathbf{w})$
2. $p(\mathbf{y}_C | \mathbf{x}, \mathbf{w}), \forall C \in \mathcal{C}(G), \forall \mathbf{y}_C \in \mathcal{Y}_C$.

Once the marginal probability of individual variables is calculated, the label with the highest probability is generally assigned to the variable, which is sometimes called the maximum of marginals inference. This is equivalent to minimizing the expected Hamming loss while MAP inference is equivalent to minimizing the expected 0/1 loss [60]. Apart from being an important inference technique, probabilistic inference is essential for maximum likelihood and other parameter estimation techniques as probabilistic inference is performed once per gradient calculation in these methods. Some of the common algorithms for probabilistic inference are belief propagations (including loopy and tree reweighted belief propagations), mean field inference, and Markov chain Monte Carlo.

Both inference techniques can be applied for CRFs and MRFs, however the equations in this Subsection are particularly for CRFs as the terms are conditioned for input features. The input features should be neglected when using them to denote MRFs.

4 Experimental evaluation

In this section, we apply pairwise MRFs and CRFs on hyperspectral images for spatial-spectral classification. First we compare different grid-structured pairwise models defined over pixel labels, and in the second part we compare these grid-structured pixel-based pairwise models against pairwise models defined over superpixels (image segments).

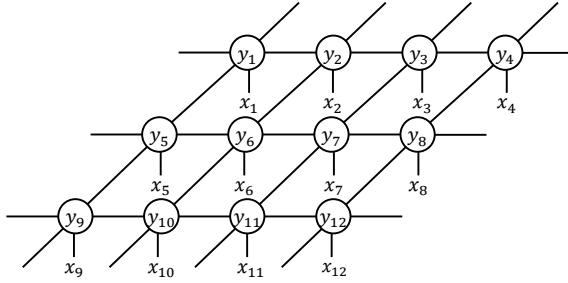


Figure 1: Grid-structured graph used for MRFs and CRFs.

4.1 Hyperspectral datasets

We experiment on four widely used public hyperspectral datasets—(a) Indian pines [5], (b) Salinas [29], (c) University of Pavia [18], and (d) Pavia Center [46]². The Indian Pines dataset contains an image collected by the Airborne Visible/Infrared Imaging Spectrometer (AVRIS) and a corresponding ground truth map, with the identity of the materials. The Indian Pines image captures an area of 2×2 miles, covering agricultural land and forest in Northwest Tippecanoe County, Indiana. It is 145×145 pixels in size and its pixel diameter is around 4 m. In our experiments, only the ground cover classes which were present in 200 or more pixels were used, bring the total number of classes from 16 to 12. The Salinas dataset also consist of an AVRIS image and a ground truth map. It was collected in the south of the city of Greenfield in the Salinas Valley in California and images a 512×217 farming area scene with 16 ground covers. Both images contain 220 spectral bands, with wavelengths ranging from 400 nm to 2500 nm. Twenty water absorption bands were removed from the images as pre-processing. Both of the images are not atmospherically compensated, with the pixels measured in spectral radiance.

The University of Pavia and the Pavia Center datasets were collected by the Reflective Optics System Imaging Spectrometer (ROSIS) over city of Pavia in northern Italy. The ROSIS sensor has 115 bands over the visible and near-infrared spectral range, with wavelengths ranging from 430 nm to 860 nm. Each pixel has a ground sampling distance of 1.3 m. Twelve noisy bands were removed from the University of Pavia image and thirteen noisy bands were removed from the Pavia Center image. The University of Pavia is 610×340 pixels in size and the Pavia center image is 1096×715 pixels in size. There are nine different material classes in both the images and ground truth land cover maps of the images are available. Both the images have been atmospherically compensated, with the spectra being measured in terms of reflectance.

4.2 Grid-structured MRFs and CRFs

Grid-structured pairwise model is the simplest and the most widely used undirected graphical model for land cover classification. In this model, each node represents a pixel label and there is an edge between nodes representing 4-connected neighboring pixels, as shown in Figure 1. In the figure, y_i represents the class label assigned to the i^{th} pixel and x_i represent the spectrum (or any feature derived from the spectrum) measured at that pixel. There are as many nodes as there are pixels and there are edges between the nodes representing neighboring pixels. Each node can take one of the discrete values representing the class of the material. Its value indicate what material is present at the pixel location whose label is represented by that node.

In the experiments, for the MRFs the unary energy at each pixel is derived from pixel-wise classifiers, by using the negative logarithm of the class-conditional probability, $E_i(y_i = c) = -\ln(P(y_i = c | x_i))$. The Potts model is used for pairwise energy. The Potts model promotes spatial smoothness by penalizing when neighboring pixels are assigned different class labels. It is defined as

$$E_{ij}(y_i, y_j) = \begin{cases} 0, & \text{if } y_i = y_j \\ \beta, & \text{otherwise,} \end{cases} \quad (12)$$

where β is the cost of the labels y_i and y_j being different. There are many other choices for pairwise energy function, however, the Potts model is most popular for spatial-spectral classification in literature. The parameters of the Potts model can be learned using maximum likelihood estimation under probabilistic

²obtained from http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes

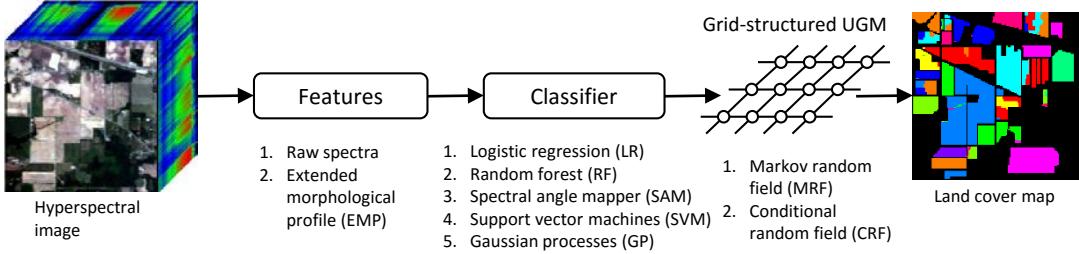


Figure 2: Pipeline for testing grid-structured models.

parameter learning framework. But, it is far more common to tune the parameter using grid search over a validation set. In the experiments, the parameter β was chosen by grid search from $\{0.001, 0.01, 0.1, 1, 10\}$. The MRF inference was performed by 15 iterations of graph cuts with expansion move algorithm [11, 37, 10] using [73]³.

The CRF was implemented using the JGMT⁴ toolbox. The CRF model used is exactly same as described by (8). A vector consisting of the class probabilities for all the M classes predicted by pixel-wise classifiers was used as the unary feature function, $\phi_1(x_i, y_i) = [P(y_i = 1 | x_i), P(y_i = 2 | x_i), \dots, P(y_i = M | x_i)]^T$. The pairwise feature function used was a constant of 1, i.e., the pairwise energy for each configuration of the two nodes was simply equal to an element in the weight vector. The truncated fitting with the clique logistic loss based on tree re-weighted belief propagation with five iterations was used for parameter learning and inference [19]. Typically, when CRFs are used in computer vision, the parameters are learned using many labeled images in the training set. But, in remote sensing, most of the time a large number of labeled training images are not available and parameters have to be learned from few labeled pixels of the test image. This is modeled during learning by making unlabeled pixels of the test image as latent nodes and the training pixels as observed nodes during training.

4.2.1 Results

The land cover maps were generated using methods consisting of feature extraction, pixel-wise classification, and the MRF or the CRF, one after another. We experiment with two features and various pixel-wise classifiers in order to perform a comprehensive analysis. The workflow used for the experimentation is shown in Figure 2.

The features used in the experiments were the raw spectra and the spatial-spectral extended morphological features [6] (EMP). The EMP features were obtained by applying the principal component analysis (PCA) on the image, retaining the relevant PCA components, and then applying a series of opening and closing morphological operators with circular structured elements of increasing size. The size of the first structuring element was 2 pixels in diameter, and the size of the subsequent ones were increased at a fixed step. The fraction of variance preserved after PCA, the number of morphological operations applied, and the increment in the size of the structuring elements are the hyperparameters of the EMP features. These hyperparameters were tuned using grid search over the validation set, which consisted of randomly selected 30% of the training pixels of each class. The remaining 70% was used as classifier's training data. The fraction of variance preserved was chosen from $\{84\%, 89\%, 94\%, 99\%\}$, and the number of morphological operators and the increment in the size of morphological operators were chosen from $\{2, 4, 8\}$. The MRF's parameters were also tuned using this validation set. The CRF was learned using all the training pixels.

The classifiers used were logistic regression (LR), random forest (RF), spectral angle mapper (SAM), support vector machine (SVM), and Gaussian process (GP). These were implemented using the multivariate logistic regression with L2 regularized weights in the LIBLINEAR library [22], MATLAB's random forest, probabilistic multi-output support vector machine in the LIBSVM library [13], and the Gaussian process classifiers in the GPML library [63] respectively. The squared exponential (SE) and the exponential spectral mapper [25] (ESAM) kernel/covariance functions were used with the SVM and the GP. The classifiers were trained on 70% of the pixels remaining in the training set. The C parameter and the kernel scale in the SVM were chosen from $\{10^{-3}, 10^{-2}, \dots, 10^2, 10^3\}$ by training the SVM on 80% of the classifier's training data and validating over the remaining 20%. After validation, the SVM was

³<http://vision.middlebury.edu/MRF/code/>

⁴<https://people.cs.umass.edu/~domke/JGMT/>

CRF, the feature function was obtained by passing the minimum angle through e^{-x} function.

Table 1, Table 2, Table 3, and Table 4 show the performance comparison of different methods on the datasets. A varying number of training pixels per class were randomly selected for training, and the models were tested on a separate set of randomly selected pixels. The testing set consisted of 50 pixels each from all of the classes. This process was repeated for 30 independent trials to obtain the mean, the standard deviation (SD) and the best overall accuracy (OA) over these trials. In the tables, each method's name contains three parts—the first indicates whether raw pixels or EMP features were used, the second contains the classifiers name, and the third indicate whether MRF or CRF was applied. The names of the kernel/covariance function used with the SVMs and the GPs are included in parenthesis.

Table 5: Performance comparison using various metrics.

Indian Pines							
Pixels/class	Methods	OA	κ	Avg. Precision	Avg. Recall	Avg. F1	
20	SVM	65.31±4.45	62.16±4.85	66.04±3.49	65.31±4.45	64.87±4.18	
	SVM-MRF	80.42±5.19	78.64±5.66	82.29±5.43	80.42±5.19	79.74±5.44	
	SVM-CRF	51.95±11.84	47.58±12.92	42.26±14.13	51.95±11.84	43.83±13.28	
	EMP-SVM	89.80±1.89	88.87±2.06	90.47±1.79	89.80±1.89	89.77±1.89	
	EMP-SVM-MRF	90.56±1.77	89.70±1.93	91.30±1.52	90.56±1.77	90.52±1.77	
	EMP-SVM-CRF	57.07±12.95	53.16±14.12	48.14±15.39	57.07±12.95	49.47±14.27	
200	SVM	85.71±1.85	84.13±2.06	85.86±1.88	85.71±1.85	85.67±1.86	
	SVM-MRF	95.01±1.29	94.46±1.44	95.22±1.26	95.01±1.29	94.99±1.30	
	SVM-CRF	97.10±1.05	96.78±1.17	97.28±0.98	97.10±1.05	97.09±1.06	
	EMP-SVM	98.91±0.54	98.79±0.60	98.93±0.53	98.91±0.54	98.91±0.54	
	EMP-SVM-MRF	99.15±0.47	99.05±0.52	99.17±0.46	99.15±0.47	99.15±0.47	
	EMP-SVM-CRF	99.35±0.43	99.27±0.48	99.37±0.41	99.35±0.43	99.34±0.43	
University of Pavia							
Pixels/class	Methods	OA	κ	Avg. Precision	Avg. Recall	Avg. F1	
20	SVM	77.59±3.10	74.78±3.49	78.23±2.95	77.59±3.10	77.21±3.33	
	SVM-MRF	84.40±4.87	82.45±5.48	85.62±4.74	84.40±4.87	84.08±4.97	
	SVM-CRF	21.07±9.18	11.21±10.33	9.06±8.24	21.07±9.18	10.73±8.34	
	EMP-SVM	90.13±3.17	88.89±3.57	90.94±2.92	90.13±3.17	90.12±3.17	
	EMP-SVM-MRF	90.16±3.24	88.92±3.65	90.97±2.99	90.16±3.24	90.15±3.24	
	EMP-SVM-CRF	21.46±9.65	11.64±10.86	9.32±8.05	21.46±9.65	11.27±8.69	
200	SVM	91.47±1.25	90.40±1.41	91.65±1.26	91.47±1.25	91.46±1.26	
	SVM-MRF	96.73±0.87	96.32±0.98	96.88±0.83	96.73±0.87	96.72±0.88	
	SVM-CRF	97.40±1.09	97.08±1.22	97.62±0.93	97.40±1.09	97.41±1.07	
	EMP-SVM	98.96±0.46	98.83±0.52	98.98±0.44	98.96±0.46	98.95±0.46	
	EMP-SVM-MRF	99.04±0.45	98.92±0.50	99.06±0.43	99.04±0.45	99.04±0.45	
	EMP-SVM-CRF	95.89±15.75	95.38±17.72	95.61±17.53	95.89±15.75	95.59±17.35	
Pavia Center							
Pixels/class	Methods	OA	κ	Avg. Precision	Avg. Recall	Avg. F1	
20	SVM	88.19±2.82	86.71±3.18	88.77±2.81	88.19±2.82	88.13±2.83	
	SVM-MRF	91.57±3.05	90.52±3.43	92.19±2.89	91.57±3.05	91.55±3.03	
	SVM-CRF	13.53±6.44	2.72±7.25	2.65±3.75	13.53±6.44	4.05±4.84	
	EMP-SVM	91.89±2.35	90.88±2.64	92.38±2.20	91.89±2.35	91.84±2.41	
	EMP-SVM-MRF	91.98±2.33	90.97±2.62	92.46±2.17	91.98±2.33	91.94±2.39	
	EMP-SVM-CRF	13.52±6.32	2.71±7.11	2.70±4.09	13.52±6.32	4.06±5.00	
200	SVM	94.96±1.00	94.33±1.13	95.11±0.97	94.96±1.00	94.96±1.01	
	SVM-MRF	97.71±0.89	97.42±1.00	97.81±0.82	97.71±0.89	97.71±0.90	
	SVM-CRF	82.95±24.83	80.82±27.93	79.72±27.85	82.95±24.83	80.32±27.41	
	EMP-SVM	98.61±0.73	98.43±0.83	98.64±0.71	98.61±0.73	98.61±0.74	
	EMP-SVM-MRF	98.75±0.71	98.59±0.79	98.78±0.68	98.75±0.71	98.75±0.71	
	EMP-SVM-CRF	78.99±30.93	76.37±34.80	75.39±34.27	78.99±30.93	76.09±33.94	
Salinas							
Pixels/class	Methods	OA	κ	Avg. Precision	Avg. Recall	Avg. F1	
20	SVM	89.96±2.01	89.29±2.14	90.07±2.14	89.96±2.01	89.78±2.10	
	SVM-MRF	93.72±2.95	93.30±3.14	93.19±3.95	93.72±2.95	93.08±3.57	
	SVM-CRF	24.33±8.36	19.29±8.91	11.66±7.34	24.33±8.36	14.07±7.69	
	EMP-SVM	96.11±1.46	95.85±1.55	96.40±1.27	96.11±1.46	96.10±1.44	
	EMP-SVM-MRF	96.46±1.42	96.23±1.52	96.75±1.22	96.46±1.42	96.45±1.42	
	EMP-SVM-CRF	24.46±8.45	19.42±9.02	12.23±7.75	24.46±8.45	14.45±8.04	
200	SVM	95.78±0.89	95.50±0.94	95.82±0.90	95.78±0.89	95.75±0.89	
	SVM-MRF	98.97±0.51	98.90±0.54	99.03±0.47	98.97±0.51	98.96±0.51	
	SVM-CRF	97.83±1.50	97.69±1.60	97.92±2.05	97.83±1.50	97.73±1.91	
	EMP-SVM	99.64±0.22	99.62±0.24	99.65±0.21	99.64±0.22	99.64±0.22	
	EMP-SVM-MRF	99.69±0.20	99.67±0.21	99.70±0.19	99.69±0.20	99.69±0.20	
	EMP-SVM-CRF	99.61±0.27	99.59±0.28	99.63±0.26	99.61±0.27	99.61±0.27	

4.2.2 Discussion

The results demonstrate the benefits of using MRFs and CRFs. The MRF increased the mean accuracy of the pixel-wise classifiers in all cases. However, we obtained mixed results from the CRF. The CRF showed high variance in performance, especially for cases with low number of training example. The performance of the CRF was poor compared to the MRF for smaller training size. However, as the size of the training set was increased the performance of CRF increased, surpassing the performance of MRF for some methods. We did not experiment with training sizes greater than 140, as at that training size other methods were already producing mean accuracies of around 99%. The reason that CRF was not well suited in the experiment is that the training set is limited. In theory, the CRF should produce better result than the MRF when enough training data is available because it is a discriminative model and is also more expressive. The CRF with log-linear potentials is more appropriate for cases where there

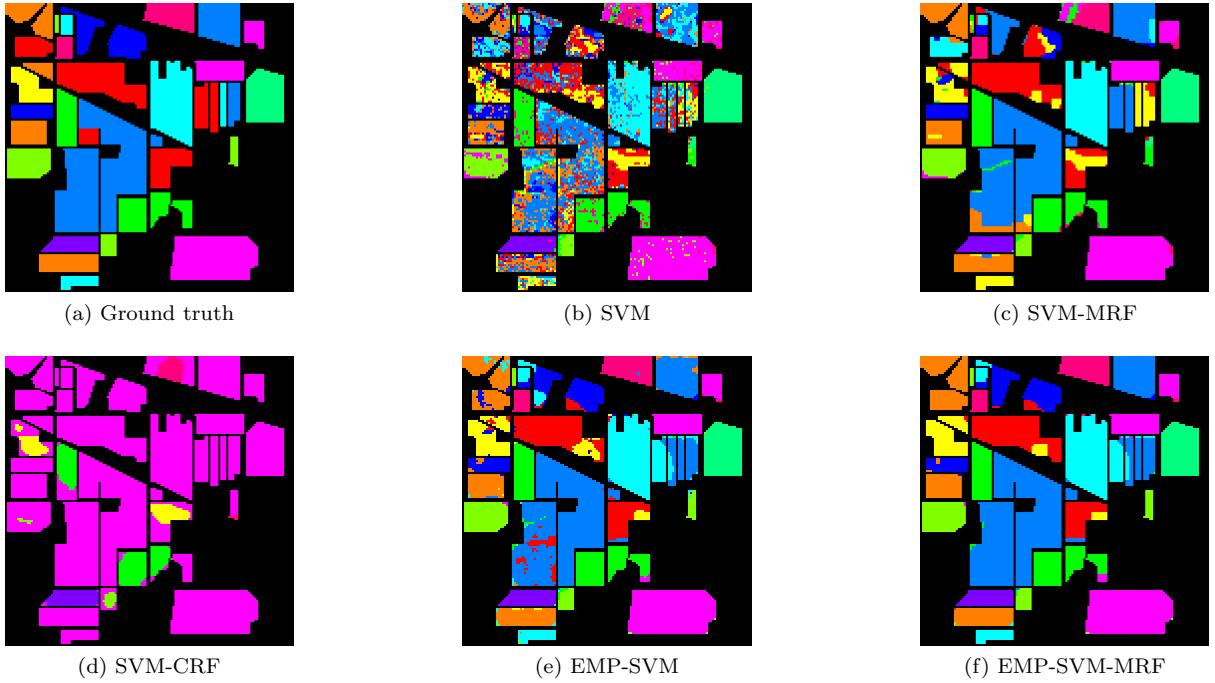


Figure 3: Predicted land cover maps for the Indian Pines image (20 training pixels per class).

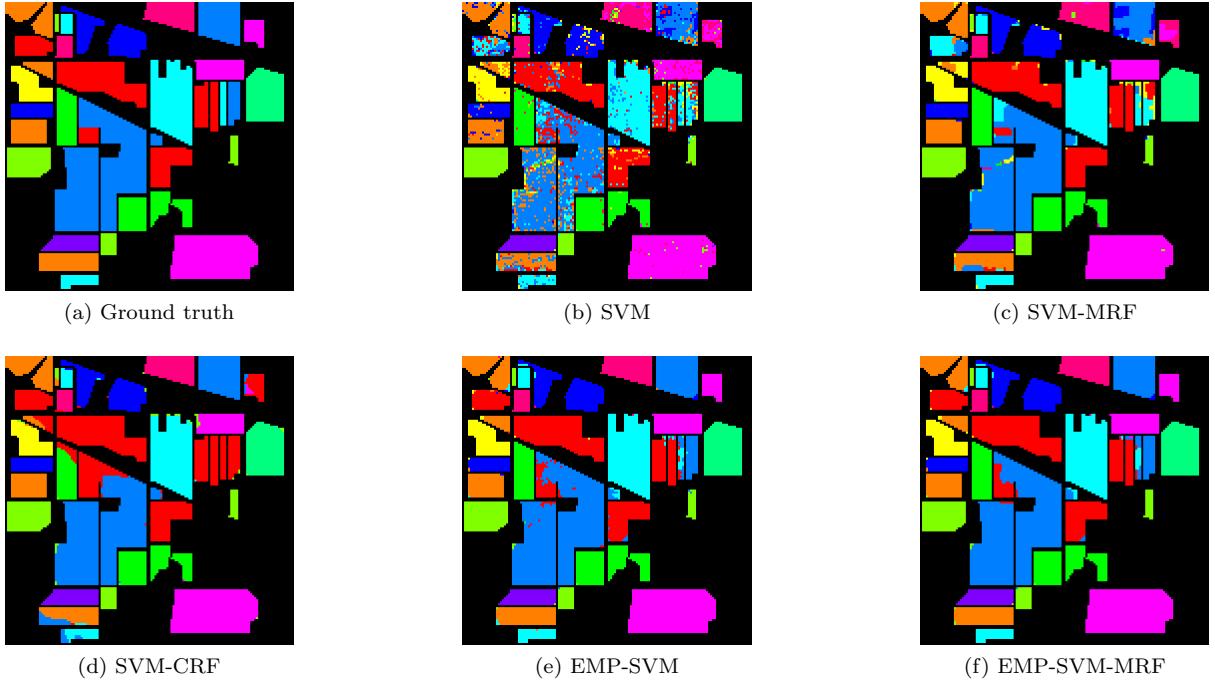


Figure 4: Predicted land cover maps for the Indian Pines image (140 training pixels per class).

are multiple fully labeled training images and the separate test images. However, this is not mostly the case in remote sensing. The dependence of CRF's performance on training set size can be further seen in table 5, where the performance of CRF is even better when the training size is 200 pixels per class. The classification maps in Figure 3, Figure 4, Figure 5, and Figure 6 qualitatively show the same trends as observed in the tables.

All the methods performed better when EMP features were used instead of the raw spectra. The SVM and the random forest classifiers were the best pixel-wise classifiers. The GP produced results comparable to the SVM. For both GP and SVM, better results were obtained with squared exponential

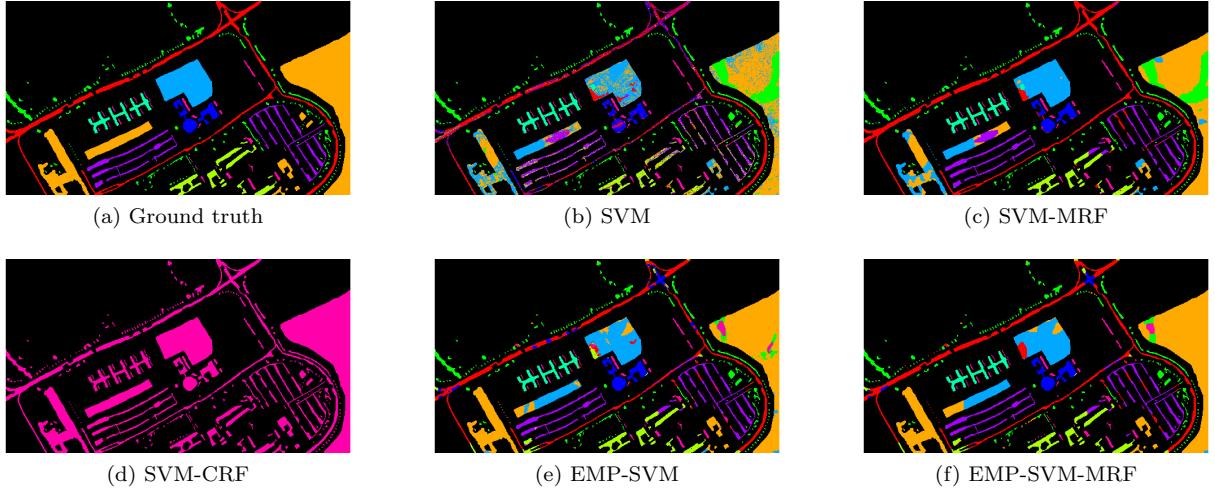


Figure 5: Predicted land cover maps for the University of Pavia image (20 training pixels per class).

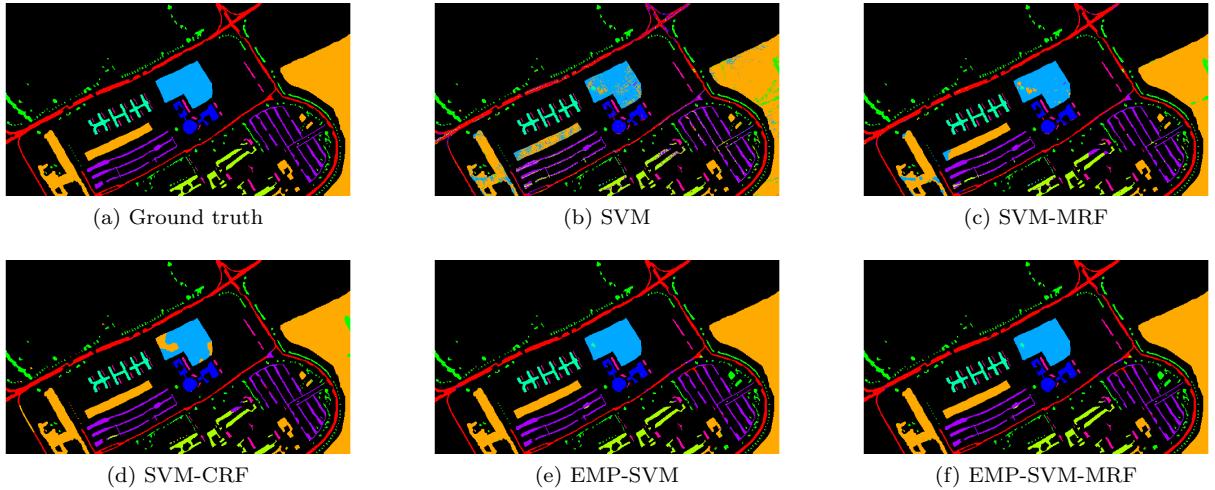


Figure 6: Predicted land cover maps for the University of Pavia image (140 training pixels per class).

kernel/covariance function. This indicate the spectral angle mapper based ESAM is not necessarily better for classification. However, the combination of SAM and MRF produced good surprising results rivaling the performance of SVM and random forest based MRFs and CRFs. SAM is essentially a nearest neighbor search with the angle as the distance metric. It does not require any training and is very fast for datasets with hundreds of training example.

If we are to compare the methods that combine spatial-spectral features and pixel-wise classifier only, such as EMP-SVM(SE), and the methods that combine pixel-wise classifier and graphical models only, such as SVM(SE)-MRF, the former always outperformed. However, methods that utilize all three components—spatial-spectral features, classifier, and UGM, were always the best. This shows an advantage of UGM. They can be added to pre-existing models to further improve the performance. The current trend in hyperspectral remote sensing is to develop more accurate spatial-spectral features with deep learning. The performance of these methods could be further improved using UGMs.

There are cases where the use of spatial-spectral features is not possible for land cover mapping. If the training set consists of a third-party spectral library or ground spectra collected from the scene, spatial-spectral features are not applicable for mapping. Same is the case when land cover maps are created by using physics bases models or unmixing techniques. In these cases, a principled approach to introduce spatial contexts is the use of UGMs. For example, the Santa Barbara urban spectral library [33] was used to map the land covers in the University of Pavia in Figure 7. The spectral library consists of 27 urban covers but the only land cover classes that occupied more than 5% of the pixel area in the predicted map have been included in the legend. The maps were generated by random forest classifier

and the combination of random forest and MRF with $\beta = 1$.

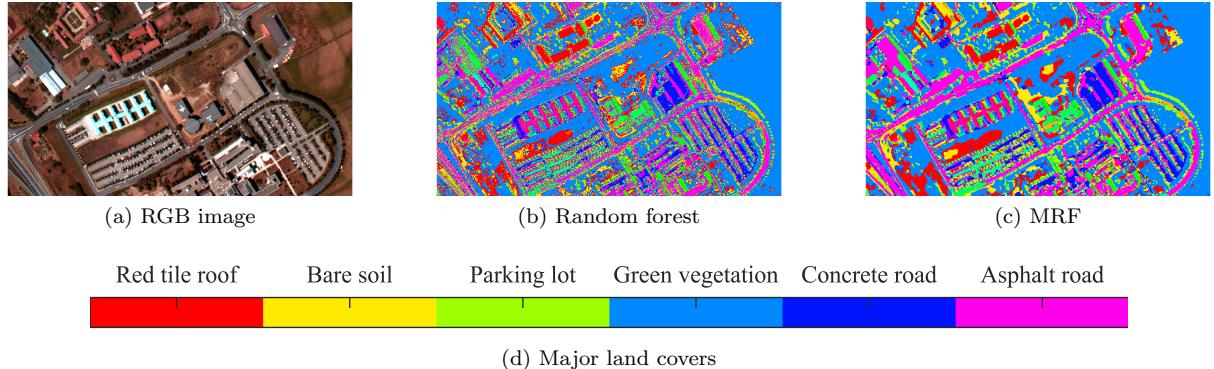


Figure 7: Land cover mapping using third-party spectral library.

Table 5 compares the qualities of land cover maps generated by different methods using a variety of performance metrics used in remote sensing—overall accuracy, kappa coefficient, average precision, average recall, and average F1 score metrics. The training set size is set at 20 and 200 pixels per class and the classifier used is SVM with squared exponential kernel function. As with the previous results, the mean and the standard deviation of 30 trials is reported. The overall accuracy, a metric that is most widely used in remote sensing, measures the fraction of pixels that were correctly classified by the classifier. However, it fails to account if individual classes of materials are accurately classified. For each material class, precision measures the fraction of pixels that were classified to belong to a class that actually belonged to that class while recall measures the fraction of pixels that belonged to a class in the ground truth that were correctly labeled by the classifier to that class. Precision and recall are commonly referred as the user’s accuracy and the producer’s accuracy in the remote sensing literature. F1 score is the harmonic mean of precision and recall. For compactness, we have only included the class averaged values of these metrics in the table. Class averaged recall is sometimes also called average accuracy in literature. Since the number of pixels of each class in the testing set are equal in our experiments, the class averaged recall is equal to the overall accuracy. κ coefficient is similar to overall accuracy and measures statistical agreement between the ground truth labels and the predicted labels.

4.3 Superpixel-based pairwise MRFs

Even though efficient inference algorithms exists for grid-structured models, real-world aerial and satellite images can be enormous and grid-structured models might be slow for time-critical applications. The computational cost of inference is generally a function of the number of nodes in the graph. Therefore, for very large images it is wiser to group pixels into homogeneous regions, called superpixels [70] and use UGMs to model the distribution of superpixels’ labels in order to reduce the total number of nodes. Superpixels are group of similar and connected pixels in the image. Unsupervised segmentation algorithms, such as simple linear iterative clustering (SLIC) [1], can be used to decompose the image into superpixels. All the pixels belonging to a superpixel are assumed to have the same class label. This is a reasonable assumption because pixels which are connected and have similar spectra are likely to be of same material. However, if the size of the superpixels are too big, there is a high chance that some of

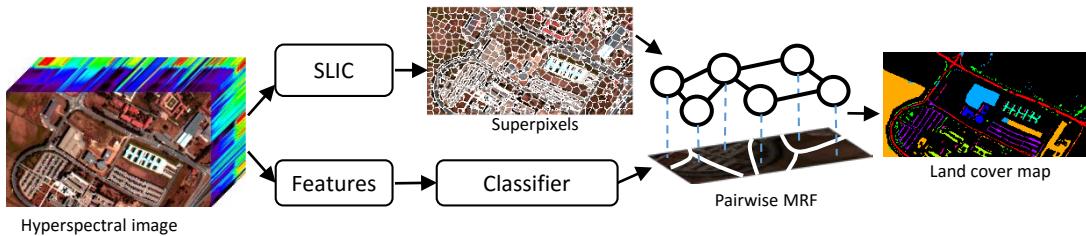


Figure 8: Pipeline for superpixel-based pairwise models.

the pixels in the superpixel will be of different classes.

Larger superpixels leads to fewer total number of superpixels in the image and hence fewer nodes in the graph. This decreases the inference computational cost. On the other hand, larger superpixels decrease the resolution of the predicted map. Any object significantly smaller than the size of a superpixel will be missed. So there is a trade-off between the resolution of the predicted land cover map and the computational complexity, which the users can control. The average size of the superpixels can generally be controlled in the unsupervised segmentation algorithms. Apart from being faster, super-pixel based UGM can model higher level of relationship in the image than pixel-based models because they model relationships between different parts of the image having some semantic meaning rather than just pixels.

Both MRFs and CRFs can be used with superpixels. However, in this section we experiment with only MRFs because as we saw in the previous section CRFs with log-linear potentials require large amount of training data and are better suited for cases where we have multiple training images. Figure 8 shows the workflow of the superpixel-based pairwise MRF used in the experiments. We experiment with raw spectrum and EMP features. Only SVM with squared exponential kernel is used as classifier in this experimentation as it was the one that performed the best previously. The SLIC algorithm was used to segment the images into superpixels. The SLIC algorithm works by assigning each pixel by a feature vector consisting of weighted concatenation of spatial coordinates plus the spectrum of the pixel and using a localized version of k-means clustering in this feature space to generate superpixels. There are three parameters of the SLIC algorithm. The first is called the regularizer and controls the shape of the superpixels. It was set to 100 in our experiments. The second parameter is the initial region size. It controls the final size and the total number of superpixels. The initial region size parameter was calculated in terms of desired number of superpixels in the image. The initial region size is equal to the square root of the total number of pixels in the image divided by the desired number of superpixels. In SLIC algorithm, the number of extracted superpixels is slightly different than the number of superpixels requested to the algorithm. We evaluate the performance of MRFs as a function of the number of superpixels in the experiments. The third parameter is the minimum region size, which was to 9 pixels in the experiments. We used the SLIC implementation in VLFeat library [80] in the experiments. The spectrum was normalized to have zero mean and standard deviation of one at all the wavelengths before using with SLIC algorithm.

In superpixel-based MRF, each superpixel is a node of the undirected graph and there is an edge between the pair of nodes representing the superpixels which touch each other. So, depending upon the superpixel segmentation, the structure of the graph changes. In our experiments, the unary potential at each superpixel was calculated by averaging class probabilities of the pixels inside the superpixel. Averaging is somewhat equivalent to majority voting because if there are many pixels which likely belong to a particular class, the unary potential for the superpixel is higher for that class. If s_1, s_2, \dots, s_K are the K superpixels, the unary energy at the node s_i when it is assigned to class c is given by $E_{s_i}(s_i = c) = -\ln\left(\frac{1}{N_{s_i}} \sum_{y_k \in s_i} P(y_k = c | x_k)\right)$, where N_{s_i} is the number of pixels in s_i . x_k and y_k are the input features and the labels of the pixels in s_i respectively. The class probability at each pixel in the image is estimated by a pixel-wise SVM classifier. The pairwise energy function used is the Potts model (12): $E_{ij}(s_i, s_j) = \beta \mathcal{I}[s_i \neq s_j]$, where $\mathcal{I}[\cdot]$ is an indicator function and β is a parameter. The pairwise MRFs were implemented using UGM library [67]. For fair comparison, the grid-structured pixel-based pairwise models used as baseline in this section were also implemented using the same library. Inference was performed using graph cuts with alpha-expansion.

4.3.1 Results

Table 6, Table 7, Table 8, and Table 9 compare the performance of superpixel-based MRF to the grid-structured pixel-based MRF (referred as pixel-based MRF). The number of superpixels used is varied. The SVM classifier were trained using 50 ground truth pixels per class. The hyperparameters of the SVM and the MRFs were tuned using grid search similarly to previous experiments. The mean and the standard deviation of the performance metrics computed over 30 independent random trials are reported.

Figure 9 and Figure 9 show the quality of land cover maps produce by superpixel-based MRFs on the Indian pines and the University of Pavia images during one of the trials. For visual comparison, a rendered RGB image and the ground truth map of the datasets have been included.

Table 6: Performance of Superpixel-based MRF on the Indian pines dataset.

Features	Superpixels	OA	κ	Avg.Precision	Avg. Recall	Avg. F1	Time (secs)
Spectra	Pixel-based MRF	89.05±2.08	88.05±2.26	89.83±1.90	89.05±2.08	88.93±2.11	19.47±1.16
	50	74.12±2.82	71.77±3.07	72.86±3.36	74.12±2.82	71.92±3.06	11.15±0.12
	100	79.17±2.48	77.27±2.70	82.10±1.86	79.17±2.48	78.76±2.59	11.17±0.12
	200	85.97±2.01	84.69±2.19	87.28±1.78	85.97±2.01	85.95±2.05	11.21±0.12
	400	89.13±1.97	88.15±2.15	90.11±1.69	89.13±1.97	89.05±2.00	11.27±0.12
	800	89.34±1.91	88.37 ±2.08	90.17 ±1.76	89.34 ±1.91	89.22 ±1.96	11.38±0.12
	1600	88.36±2.27	87.30±2.48	89.32±2.21	88.36±2.27	88.20±2.33	11.53±0.14
	3200	86.41±1.95	85.17±2.13	87.28±1.89	86.41±1.95	86.22±2.03	11.59±0.16
EMP	Pixel-based MRF	95.91±1.27	95.53±1.39	96.11±1.17	95.91±1.27	95.89±1.27	374.95±2.89
	50	79.74±1.55	77.90±1.70	78.30±3.22	79.74±1.55	77.75±2.42	386.60 ±2.31
	100	85.28±1.27	83.95±1.39	86.83±1.19	85.28±1.27	84.94±1.38	366.62±2.31
	200	91.81±1.14	91.07±1.24	92.27±1.06	91.81±1.14	91.78±1.16	366.65±2.31
	400	95.82±1.18	95.44±1.29	96.08±1.07	95.82±1.18	95.81±1.18	366.72±2.31
	800	96.06 ±1.12	95.70 ±1.22	96.23 ±1.06	96.06 ±1.12	96.04 ±1.12	366.83±2.31
	1600	95.71±1.11	95.32±1.21	95.91±1.04	95.71±1.11	95.69±1.11	366.98±2.30
	3200	93.67±1.36	93.10±1.49	93.94±1.29	93.67±1.36	93.64±1.38	367.05±2.32

Table 7: Performance of Superpixel-based MRF on the University of Pavia dataset.

Features	Superpixels	OA	κ	Avg.Precision	Avg. Recall	Avg. F1	Time (secs)
Spectra	Pixel-based MRF	91.23±4.06	90.13±4.57	91.69±3.92	91.23±4.06	91.16±4.10	69.58±0.86
	50	64.69±6.18	60.27±6.95	72.54±9.05	64.69±6.18	63.18±7.12	8.29 ±0.71
	100	70.19±4.22	66.46±4.75	69.64±3.64	70.19±4.22	67.43±4.00	8.32±0.71
	200	79.69±5.03	77.15±5.66	83.90±3.75	79.69±5.03	79.78±4.82	8.37±0.71
	400	86.46±3.50	84.77±3.94	88.63±3.14	86.46±3.50	86.62±3.49	8.46±0.71
	800	89.80±3.31	88.53±3.72	90.98±2.81	89.80±3.31	89.83±3.30	8.66±0.71
	1600	91.56±3.87	90.51±4.35	92.40±3.56	91.56±3.87	91.58±3.84	9.05±0.72
	3200	92.83 ±3.67	91.93 ±4.13	93.43 ±3.37	92.83 ±3.67	92.76 ±3.71	9.73±0.72
EMP	4000	92.24±3.38	91.28±3.80	92.77±3.16	92.24±3.38	92.18±3.43	10.78±0.72
	6400	91.26±3.90	90.17±4.39	92.00±3.61	91.26±3.90	91.20±3.98	14.03±0.73
	12800	95.80±1.40	95.42±1.57	96.18±1.28	95.93±1.40	95.94±1.40	267.37±1.25
	Pixel-based MRF	95.81±1.49	95.29±1.68	96.04±1.39	95.81±1.49	95.83±1.48	321.66±0.87
	50	74.30±4.33	71.09±4.87	76.82±3.92	74.30±4.33	72.49±4.72	261.02 ±0.95
	100	81.24±4.38	78.89±4.93	81.36±6.09	81.24±4.38	80.25±5.51	261.05±0.95
	200	88.21±2.75	86.74±3.10	89.24±2.18	88.21±2.75	87.87±2.74	261.10±0.95
	400	90.95±1.60	89.82±1.80	91.63±1.50	90.95±1.60	90.90±1.63	261.18±0.95
	800	94.67±1.67	94.01±1.88	95.08±1.42	94.67±1.67	94.69±1.66	261.39±0.95
	1600	95.01±1.41	94.38±1.59	95.38±1.30	95.01±1.41	95.03±1.41	261.78±0.95
	3200	95.79±1.44	95.27±1.61	96.02±1.37	95.79±1.44	95.80±1.44	262.51±0.91
	6400	96.09 ±1.54	95.60 ±1.73	96.31 ±1.44	96.09 ±1.54	96.10 ±1.54	263.59±0.93
	12800	95.93±1.40	95.42±1.57	96.18±1.28	95.93±1.40	95.94±1.40	267.37±1.25

Table 8: Performance of Superpixel-based MRF on the Pavia center dataset.

Features	Superpixels	OA	κ	Avg.Precision	Avg. Recall	Avg. F1	Time (secs)
Spectra	Pixel-based MRF	95.30±1.60	94.71±1.80	95.59±1.49	95.30±1.60	95.30±1.60	86.87±4.19
	50	61.90±1.57	57.14±2.22	62.72±3.29	61.90±1.97	57.83±2.39	20.36 ±3.80
	100	75.59±1.85	72.54±2.08	74.24±3.66	75.59±1.85	73.16±2.50	20.43±3.80
	200	81.79±2.41	79.51±2.71	83.31±2.74	81.79±2.41	80.73±3.06	20.56±3.79
	400	84.59±1.51	82.67±1.70	86.16±1.55	84.59±1.51	84.17±1.69	20.81±3.79
	800	86.67±1.52	85.00±1.71	87.79±1.24	86.67±1.52	86.39±1.62	21.34±3.79
	1600	90.61±1.64	89.43±1.85	91.14±1.43	90.61±1.64	90.52±1.67	22.37±3.74
	3200	93.18±1.70	92.33±1.91	93.59±1.45	93.18±1.70	93.15±1.72	24.27±3.75
EMP	6400	95.11±1.61	94.50±1.81	95.38±1.48	95.11±1.61	95.11±1.60	28.37±3.73
	12800	95.53 ±1.38	94.97 ±1.55	95.78 ±1.29	95.53 ±1.38	95.53 ±1.37	35.62±3.74
	25600	95.40±1.59	94.83±1.79	95.68±1.45	95.40±1.59	95.40±1.58	48.02±3.87
	Pixel-based MRF	96.00±1.44	95.50±1.62	96.14±1.40	96.00±1.44	96.00±1.43	823.60±3.55
	50	61.64±2.70	56.84±3.04	63.99±3.23	61.64±2.70	57.58±2.93	759.85 ±3.68
	100	77.07±2.85	74.21±3.20	77.39±2.93	77.07±2.85	75.84±3.21	759.93±3.68
	200	82.56±2.65	80.38±2.98	84.57±2.32	82.56±2.65	81.87±3.16	760.05±3.68
	400	86.36±2.04	84.65±2.29	87.63±1.73	86.36±2.04	86.14±2.19	760.30±3.68
	800	89.14±1.86	87.78±2.09	90.00±1.51	89.14±1.86	88.91±1.96	760.83±3.68
	1600	92.30±1.59	91.34±1.79	92.72±1.44	92.30±1.59	92.23±1.63	761.84±3.66
	3200	94.62±1.67	93.95±1.88	94.87±1.56	94.62±1.67	94.60±1.70	763.78±3.72
	6400	95.93±1.58	95.42±1.78	96.10±1.51	95.93±1.58	95.93±1.59	767.98±3.70
	12800	96.05 ±1.48	95.56 ±1.67	96.21 ±1.42	96.05 ±1.48	96.05 ±1.48	773.64±3.16
	25600	95.98±1.73	95.47±1.95	96.12±1.70	95.98±1.73	95.98±1.73	787.12±3.51

4.3.2 Discussion

The results show consistent patterns for all datasets. The performance of the superpixel-based MRF is poor when there are few superpixels. As the number of superpixels is increased, the performance grows reaching a peak performance, which is equal or better than that of pixel-wise MRF. However, after reaching the peak the performance decreases on increasing the number of superpixels. Furthermore, the superpixel-based MRFs are significantly faster than pixel-based MRFs. It should be noted that the time taken to perform the superpixel segmentation is included in the reported method's time in the table.

Pixel-based MRFs learn spatial relationship between the pixels while the superpixel-based MRFs learn spatial relationship between superpixels, which are group of homogeneous pixels and represent higher order structures. Superpixels may represent objects or different parts of objects. Therefore, superpixel-based MRF should perform better than pixel-wise MRF when the size of the superpixels are optimally representing different components in the image. The results confirm this hypothesis. We see that at optimal scale (size of the superpixel), the superpixel-based MRF performs as good as or outperforms the pixel-wise MRF. The optimal scale is different for different images depending upon the

Table 9: Performance of Superpixel-based MRF on the Salinas dataset.

Features	Superpixels	OA	κ	Avg. Precision	Avg. Recall	Avg. F1	Time (secs)
Spectra	Pixel-based MRF	97.42±1.37	97.24±1.46	97.53±1.63	97.42±1.37	97.34±1.60	101.13±2.02
	50	73.26±2.19	71.48±2.33	67.21±4.33	73.26±2.19	67.95±3.39	31.85±1.99
	100	79.59±1.59	78.23±1.69	81.03±2.23	79.59±1.59	76.39±2.03	31.90±1.99
	200	89.31±1.83	88.60±1.95	91.11±1.78	89.31±1.83	89.24±2.12	32.00±1.99
	400	96.84±1.29	96.63±1.38	96.95±1.58	96.84±1.29	96.72±1.61	32.17±1.99
	800	97.19±1.42	97.00±1.51	97.22±1.75	97.19±1.42	97.10±1.69	32.61±1.96
	1600	96.68±1.41	96.46±1.51	96.86±1.64	96.68±1.41	96.57±1.67	33.78±1.94
	3200	97.67±1.54	97.52±1.64	97.75±1.79	97.67±1.54	97.59±1.79	35.96±1.89
	6400	98.04±1.59	97.91±1.70	98.07±1.88	98.04±1.59	97.95±1.85	43.61±1.94
	12800	96.58±1.75	96.35±1.87	96.74±1.97	96.58±1.75	96.46±2.03	44.49±1.96
EMP	Pixel-based MRF	98.78±0.78	98.69±0.83	98.83±0.74	98.78±0.78	98.77±0.78	481.14±2.18
	50	74.76±1.13	73.08±1.21	66.71±2.48	74.76±1.13	68.91±1.76	412.69±2.20
	100	81.07±1.06	79.81±1.13	82.24±1.49	81.07±1.06	78.35±1.26	412.74±2.20
	200	90.89±0.87	90.28±0.93	91.99±0.80	90.89±0.87	90.88±0.88	412.84±2.20
	400	98.11±0.49	97.99±0.53	98.20±0.46	98.11±0.49	98.11±0.50	413.01±2.20
	800	98.62±0.63	98.52±0.67	98.70±0.56	98.62±0.63	98.62±0.63	413.39±2.20
	1600	98.54±0.56	98.44±0.59	98.62±0.53	98.54±0.56	98.54±0.56	414.60±2.23
	3200	98.91±0.52	98.84±0.55	98.96±0.49	98.91±0.52	98.91±0.51	417.13±2.25
	6400	98.93±0.54	98.86±0.57	98.98±0.51	98.93±0.54	98.93±0.54	422.66±2.78
	12800	98.03±0.58	97.89±0.61	98.10±0.56	98.03±0.58	98.03±0.58	424.44±2.33

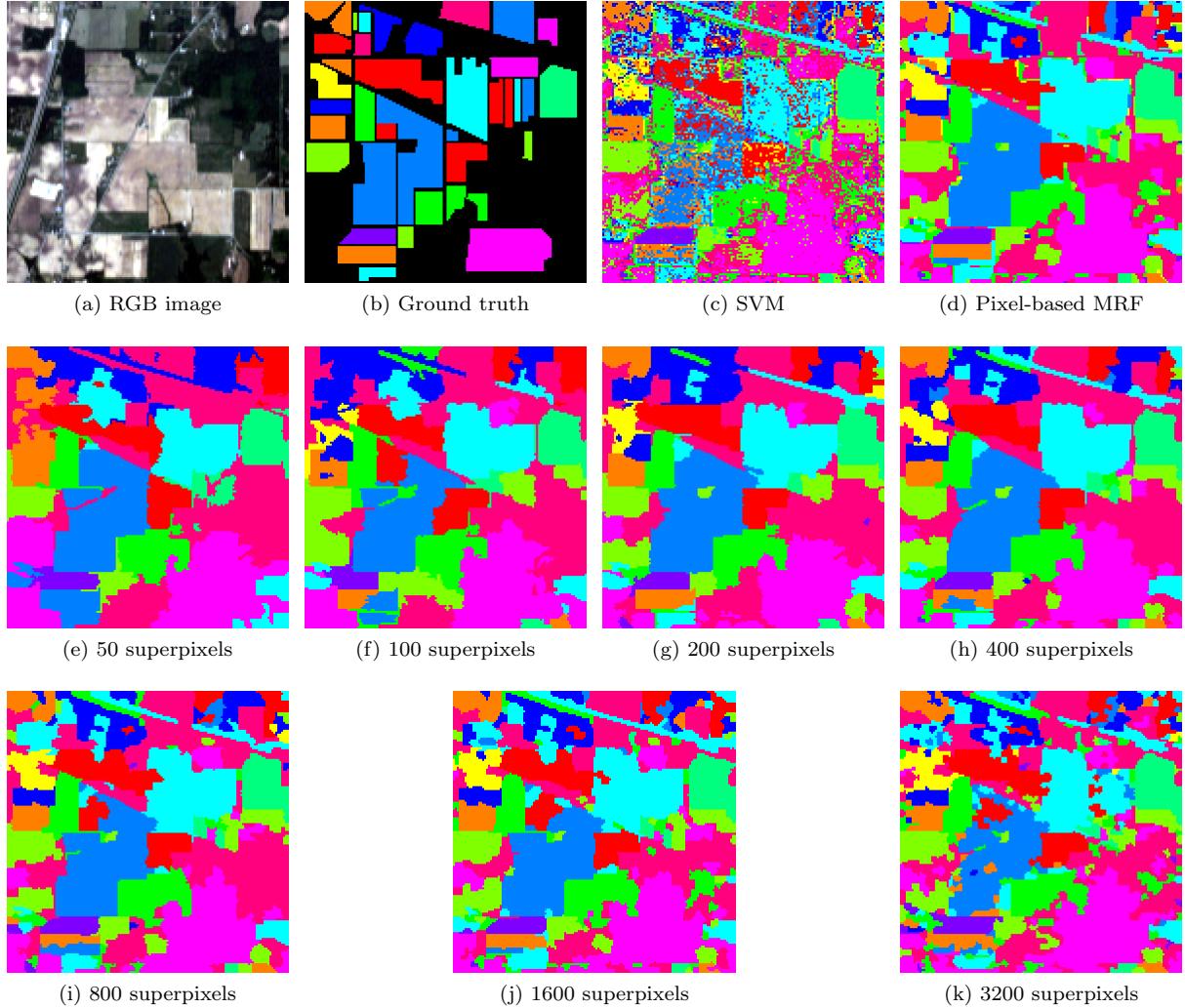


Figure 9: Superpixel-based MRF applied on the Indian pines image.

size of objects in the image. It should be also noted that the performance of superpixel-based MRF is dependent on the performance of the superpixel segmentation algorithm. Additionally, since the number of nodes in superpixel-based MRF is much smaller than pixel-based MRF, which is a great advantage when we are trying to label very large images. The gain in speed of inference in order to get the same or better performance than pixel-based model is much higher in larger images (Salinas and Pavia city) compared to smaller images (Indian pines) indicating superpixel-based UGM become essential while dealing enormous remote sensing images.

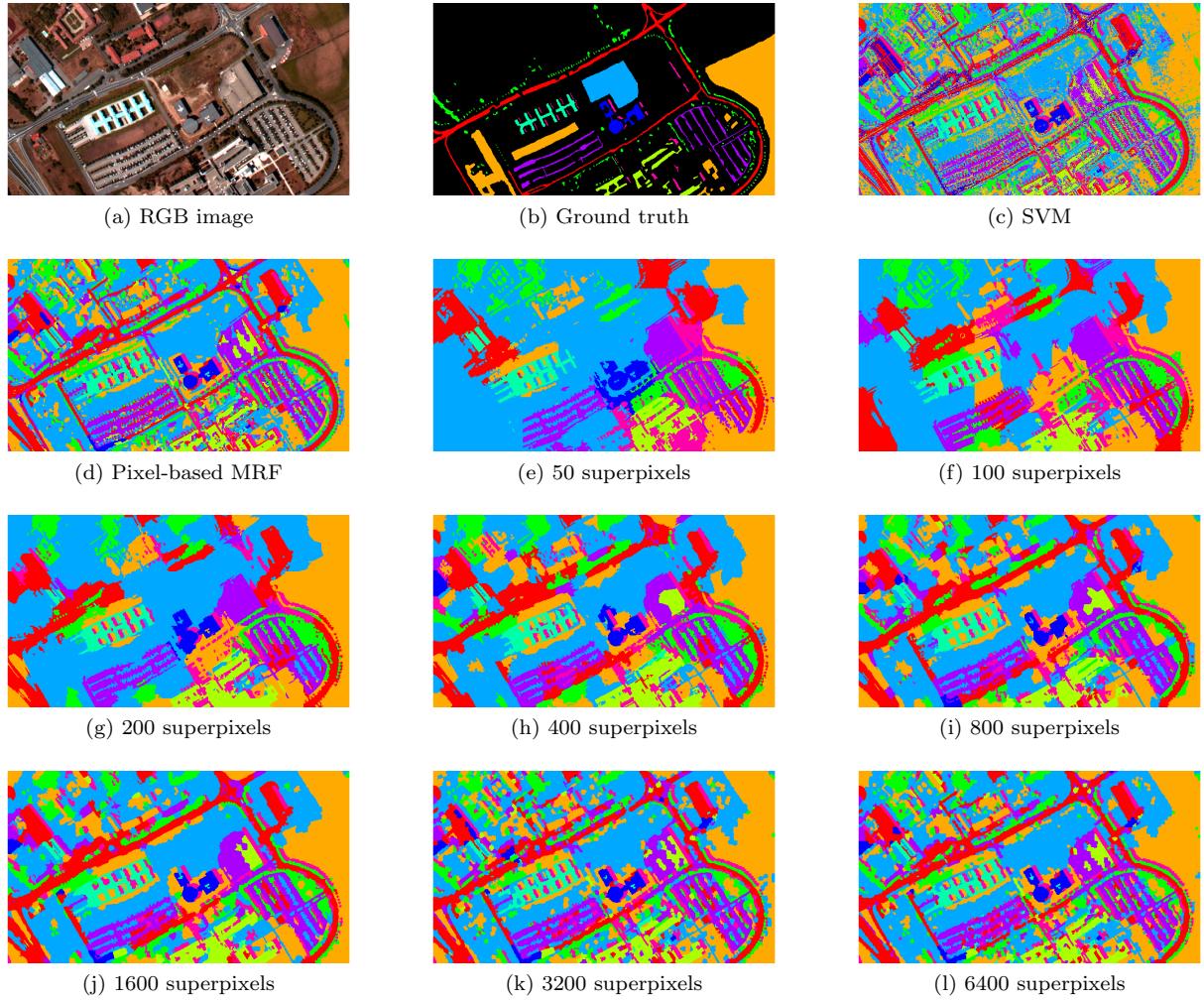


Figure 10: Superpixel-based MRF applied on the University of Pavia image.

In Figure 9 and Figure 10, we see that when fewer superpixels are used the estimated land cover map is of low resolution and small objects in the image are ignored. As the number of superpixels is increased smaller objects become visible in the land cover maps. So superpixel-based MRFs give users an option to control the resolution of the land cover maps and a chance to obtain faster computational time at the cost of decreased resolution.

5 Summary

In this tutorial, we provided a broad introduction to modeling and inference in undirected graphical models (UGMs), reviewed UGM based methods developed for remote sensing applications, explained pixel-based and superpixel-based pairwise UGMs in detail, and experimentally evaluated those models on popular hyperspectral datasets. To make it easier for the readers to have hands-on experience with the discussed models, the source code used to implement the methods for the experiments have been published. The elaborate set of experimental results included in this tutorial can serve as baselines for future UGM-based or any other techniques for spatial-spectral classification.

Pairwise Markov random fields with grid-search for parameter learning seems to be best suited approach for current hyperspectral datasets. Current datasets consists of a moderately sized image with some labeled pixels for training and some for testing. Therefore, only simpler UGMs can be properly trained on them because models with more complex graph and more expressive potential functions have large number of parameters which have to be learned from the data. However, it would be more useful to train and test models on different images, so that once the model is trained it could be applied to any new images obtained from the sensor. Such models would require large amount of labeled images for

training. Researchers have already started to develop this kind of datasets for color and multi-spectral images, for example, SpaceNet, Inria aerial image labeling dataset [49], and 2017 IEEE GRSS data fusion contest dataset [78]. Unfortunately, such datasets are unavailable for hyperspectral imagery and this have limited the growth of development of more sophisticated and robust mapping methods. With enough labeled training example, in theory we should be able to build mapping models that are robust to variation in geographic locations, image acquisition season and time, image resolution, weather conditions, and sensor technologies. Therefore, development of new benchmark datasets should be one of the top priorities of researchers.

Data fusion is a fields where UGMs could have a major impact. They have been successfully used to fuse multi-modal images for land cover classification [48, 84]. However, more interesting applications would arise from the fusion of ground level data, such as digital maps (e.g., OpenStreetMaps) and geotagged data (e.g., photos, online reviews), with the remotely sensed images. The class labels used to describe the points of interest in different sources of data would not have to be the same because UGMs could automatically learn the hierarchical relationships between different them as in [2].

The current popular trend in spatial-spectral classification is to develop deep neural network for feature extraction [99]. UGMs can be used to complement these methods. As we saw in the experiments, UGMs can boost the performance of spatial-spectral feature based classifiers. Moreover, for cases where the training set consists of spectra from third-party spectral library or ground spectra collected from the scene, spectral features either hand-designed or learned using deep network can not be used. In these cases, UGMs become a useful tool to apply spatial contextual information.

References

- [1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.
- [2] Lena Albert, Franz Rottensteiner, and Christian Heipke. A higher order conditional random field model for simultaneous classification of land cover and land use. *ISPRS Journal of Photogrammetry and Remote Sensing*, 130:63–80, 2017.
- [3] Yoann Altmann, Nicolas Dobigeon, Steve McLaughlin, and Jean-Yves Tourneret. Residual component analysis of hyperspectral imagesapplication to joint nonlinear unmixing and nonlinearity detection. *IEEE Transactions on Image Processing*, 23(5):2148–2158, 2014.
- [4] Yoann Altmann, Marcelo Pereyra, and Stephen McLaughlin. Bayesian nonlinear hyperspectral unmixing with spatial residual component analysis. *IEEE Transactions on Computational Imaging*, 1(3):174–185, 2015.
- [5] Marion F. Baumgardner, Larry L. Biehl, and David A. Landgrebe. 220 band aviris hyperspectral image data set: June 12, 1992 indian pine test site 3, Sep 2015.
- [6] Jón Atlí Benediktsson, Jón Aevar Palmason, and Johannes R Sveinsson. Classification of hyperspectral data from urban areas based on extended morphological profiles. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3):480–491, 2005.
- [7] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.
- [8] Julian Besag. Statistical analysis of non-lattice data. *The statistician*, pages 179–195, 1975.
- [9] Andrew Blake, Pushmeet Kohli, and Carsten Rother. *Markov random fields for vision and image processing*. MIT Press, 2011.
- [10] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004.
- [11] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

- [12] Gustavo Camps-Valls, Luis Gomez-Chova, Jordi Muñoz-Marí, Joan Vila-Francés, and Javier Calpe-Maravilla. Composite kernels for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 3(1):93–97, 2006.
- [13] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.
- [14] Yushi Chen, Zhouhan Lin, Xing Zhao, Gang Wang, and Yanfeng Gu. Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6):2094–2107, 2014.
- [15] Roger N Clark, Gregg A Swayze, K Eric Livo, Raymond F Kokaly, Steve J Sutley, J Brad Dalton, Robert R McDougal, and Carol A Gent. Imaging spectroscopy: Earth and planetary remote sensing with the usgs tetracorder and expert systems. *Journal of Geophysical Research: Planets*, 108(E12), 2003.
- [16] Michele Dalponte, Hans Ole Orka, Terje Gobakken, Damiano Gianelle, and Erik Næsset. Tree species classification in boreal forests with hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 51(5):2632–2645, 2013.
- [17] Ciro D’Elia, Giovanni Poggi, and Giuseppe Scarpa. A tree-structured markov random field model for bayesian image segmentation. *IEEE Transactions on Image Processing*, 12(10):1259–1273, 2003.
- [18] Fabio Dell’Acqua, Paolo Gamba, and Alessio Ferrari. Exploiting spectral and spatial information for classifying hyperspectral data in urban areas. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, volume 1, pages 464–466. IEEE, 2003.
- [19] Justin Domke. Learning graphical model parameters with approximate marginal inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(10):2454–2467, 2013.
- [20] Olivier Eches, Jón Atli Benediktsson, Nicolas Dobigeon, and Jean-Yves Tourneret. Adaptive markov random fields for joint unmixing and segmentation of hyperspectral images. *IEEE Transactions on Image Processing*, 22(1):5–16, 2013.
- [21] Olivier Eches, Nicolas Dobigeon, and Jean-Yves Tourneret. Enhancing hyperspectral image unmixing with spatial correlations. *IEEE Transactions on Geoscience and Remote Sensing*, 49(11):4239–4247, 2011.
- [22] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.
- [23] Leyuan Fang, Shutao Li, Xudong Kang, and Jón Atli Benediktsson. Spectral–spatial hyperspectral image classification via multiscale adaptive sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 52(12):7738–7749, 2014.
- [24] Mathieu Fauvel, Yuliya Tarabalka, Jon Atli Benediktsson, Jocelyn Chanussot, and James C Tilton. Advances in spectral-spatial classification of hyperspectral images. *Proceedings of the IEEE*, 101(3):652–675, 2013.
- [25] Utsav B. Gewali and Sildomar T. Monteiro. A novel covariance function for predicting vegetation biochemistry from hyperspectral imagery with gaussian processes. In *International Conference on Image Processing (ICIP)*, pages 2216–2220. IEEE, 2016.
- [26] Utsav B. Gewali and Sildomar T. Monteiro. Spectral angle based unary energy functions for spatial-spectral hyperspectral classification using markov random fields. In *Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pages 1–6. IEEE, Aug 2016.
- [27] Utsav B. Gewali and Sildomar T. Monteiro. Using bayesian optimization to jointly tune the classifier and the random field for spatial-spectral hyperspectral classification. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3692–3695. IEEE, July 2017.
- [28] Dorothy M Greig, Bruce T Porteous, and Allan H Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 271–279, 1989.

- [29] JA Gualtieri, Samir R Chettri, RF Crompt, and LF Johnson. Support vector machine classifiers as applied to aviris data. In *Proc. Eighth JPL Airborne Geoscience Workshop*. JPL, 1999.
- [30] Jisoo Ham, Yangchi Chen, Melba M Crawford, and Joydeep Ghosh. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3):492–501, 2005.
- [31] Robert M Haralick, Karthikeyan Shanmugam, et al. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, 1973.
- [32] Uta Heiden, Karl Segl, Sigrid Roessner, and Hermann Kaufmann. Determination of robust spectral features for identification of urban surface materials in hyperspectral remote sensing data. *Remote Sensing of Environment*, 111(4):537–552, 2007.
- [33] Martin Herold, Margaret E Gardner, and Dar A Roberts. Spectral resolution requirements for mapping urban areas. *IEEE Transactions on Geoscience and Remote Sensing*, 41(9):1907–1919, 2003.
- [34] Qiong Jackson and David A Landgrebe. Adaptive bayesian contextual classification based on markov random fields. *IEEE Transactions on Geoscience and Remote Sensing*, 40(11):2454–2463, 2002.
- [35] Teerasit Kasetkasem, Manoj K Arora, and Pramod K Varshney. Super-resolution land cover mapping using a markov random field based approach. *Remote Sensing of Environment*, 96(3):302–314, 2005.
- [36] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
- [37] Vladimir Kolmogorov and Ramin Zabin. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004.
- [38] Er Li, John Femiani, Shibiao Xu, Xiaopeng Zhang, and Peter Wonka. Robust rooftop extraction from visible band images using higher order crf. *IEEE Transactions on Geoscience and Remote Sensing*, 53(8):4483–4495, 2015.
- [39] Fan Li, Linlin Xu, Parthipan Siva, Alexander Wong, and David A Clausi. Hyperspectral image classification with limited labeled training samples using enhanced ensemble learning and conditional random fields. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(6):2427–2438, 2015.
- [40] Jun Li, José M Bioucas-Dias, and Antonio Plaza. Hyperspectral image segmentation using a new bayesian approach with active learning. *IEEE Transactions on Geoscience and Remote Sensing*, 49(10):3947–3960, 2011.
- [41] Jun Li, José M Bioucas-Dias, and Antonio Plaza. Spectral–spatial hyperspectral image segmentation using subspace multinomial logistic regression and markov random fields. *IEEE Transactions on Geoscience and Remote Sensing*, 50(3):809–823, 2012.
- [42] Jun Li, José M Bioucas-Dias, and Antonio Plaza. Spectral–spatial classification of hyperspectral data using loopy belief propagation and active learning. *IEEE Transactions on Geoscience and Remote Sensing*, 51(2):844–856, 2013.
- [43] Miao Li, Shuying Zang, Bing Zhang, Shanshan Li, and Changshan Wu. A review of remote sensing image classification techniques: The role of spatio-contextual information. *European Journal of Remote Sensing*, 47(1):389–411, 2014.
- [44] Wei Li, Saurabh Prasad, and James E Fowler. Hyperspectral image classification using gaussian mixture models and markov random fields. *IEEE Geoscience and Remote Sensing Letters*, 11(1):153–157, 2014.
- [45] Ying Li, Haokui Zhang, and Qiang Shen. Spectral–spatial classification of hyperspectral imagery with 3d convolutional neural network. *Remote Sensing*, 9(1):67, 2017.

- [46] Giorgio Licciardi, Fabio Pacifici, Devis Tuia, Saurabh Prasad, Terrance West, Ferdinando Giacco, Christian Thiel, Jordi Ingla, Emmanuel Christophe, Jocelyn Chanussot, et al. Decision fusion for the classification of hyperspectral data: Outcome of the 2008 grs-s data fusion contest. *IEEE Transactions on Geoscience and Remote Sensing*, 47(11):3857–3865, 2009.
- [47] Desheng Liu, Kuan Song, John RG Townshend, and Peng Gong. Using local transition probability models in markov random fields for forest change detection. *Remote Sensing of Environment*, 112(5):2222–2231, 2008.
- [48] Yansong Liu, Sankaranarayanan Piramanayagam, Sildomar T Monteiro, and Eli Saber. Dense semantic labeling of very-high-resolution aerial imagery and lidar with fullyconvolutional neural networks and higher-order crfs. In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Honolulu, USA*, 2017.
- [49] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. Can semantic labeling methods generalize to any city? the INRIA aerial image labeling benchmark. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3226–3229, July 2017.
- [50] MP Martin, L Barreto, D Riaño, C Fernandez-Quintanilla, and P Vaughan. Assessing the potential of hyperspectral remote sensing for the discrimination of grassweeds in winter cereal crops. *International Journal of Remote Sensing*, 32(1):49–67, 2011.
- [51] Farid Melgani and Lorenzo Bruzzone. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on geoscience and remote sensing*, 42(8):1778–1790, 2004.
- [52] Farid Melgani and Sebastiano B Serpico. A markov random field approach to spatio-temporal contextual image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 41(11):2478–2487, 2003.
- [53] Andreas Merentitis, Christian Debes, and Roel Heremans. Ensemble learning in hyperspectral image classification: toward selecting a favorable bias-variance tradeoff. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(4):1089–1102, 2014.
- [54] Javier A Montoya-Zegarra, Jan D Wegner, L Ladický, and K Schindler. Semantic segmentation of aerial images in urban areas with class-specific higher-order cliques. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2(3):127, 2015.
- [55] Gabriele Moser and Sebastiano B Serpico. Combining support vector machines and markov random fields in an integrated framework for contextual image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 51(5):2734–2752, 2013.
- [56] Gabriele Moser, Sebastiano B Serpico, and Jon Atli Benediktsson. Land-cover mapping by markov modeling of spatial–contextual information in very-high-resolution remote sensing images. *Proceedings of the IEEE*, 101(3):631–651, 2013.
- [57] Kevin P Murphy. Undirected graphical models (markov random fields). In *Machine learning: a probabilistic perspective*, chapter 19. MIT press, 2012.
- [58] Richard J Murphy and Sildomar T Monteiro. Mapping the distribution of ferric iron minerals on a vertical mine face using derivative analysis of hyperspectral imagery (430–970nm). *ISPRS Journal of Photogrammetry and Remote Sensing*, 75:29–39, 2013.
- [59] Joachim Niemeyer, Franz Rottensteiner, Uwe Sörgel, and Christian Heipke. Hierarchical higher order crf for the classification of airborne lidar point clouds in urban areas. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 41:655, 2016.
- [60] Sebastian Nowozin, Christoph H Lampert, et al. Structured learning and prediction in computer vision. *Foundations and Trends® in Computer Graphics and Vision*, 6(3–4):185–365, 2011.
- [61] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

- [62] Ruiliang Pu, Peng Gong, Yong Tian, Xin Miao, Raymond I Carruthers, and Gerald L Anderson. Invasive species change detection using artificial neural networks and casi hyperspectral imagery. *Environmental monitoring and assessment*, 140(1):15–32, 2008.
- [63] Carl Edward Rasmussen and Hannes Nickisch. Gaussian processes for machine learning (gpml) toolbox. *Journal of Machine Learning Research*, 11(Nov):3011–3015, 2010.
- [64] Guillaume Rellier, Xavier Descombes, Frederic Falzon, and Josiane Zerubia. Texture feature analysis using a gauss-markov model in hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 42(7):1543–1551, 2004.
- [65] Ribana Roscher and Björn Waske. Superpixel-based classification of hyperspectral data using sparse representation and conditional random fields. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3674–3677. IEEE, 2014.
- [66] Konrad Schindler. An overview and comparison of smooth labeling methods for land-cover classification. *IEEE Transactions on Geoscience and Remote Sensing*, 50(11):4534–4545, 2012.
- [67] Mark Schmidt. Ugm: Matlab code for undirected graphical models. URL <http://www.cs.ubc.ca/~schmidtm/Software/UGM.html>, 2012.
- [68] Michael Schroder, Hubert Rehrauer, Klaus Seidel, and Mihai Datcu. Spatial information retrieval from remote-sensing images. ii. gibbs-markov random fields. *IEEE Transactions on Geoscience and Remote Sensing*, 36(5):1446–1455, 1998.
- [69] Anne H Schistad Solberg, Torfinn Taxt, and Anil K Jain. A markov random field model for classification of multisource satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 34(1):100–113, 1996.
- [70] David Stutz, Alexander Hermans, and Bastian Leibe. Superpixels: An evaluation of the state-of-the-art. *Computer Vision and Image Understanding*, 166:1 – 27, 2018.
- [71] Shujin Sun, Ping Zhong, Huaitie Xiao, and Runsheng Wang. An mrf model-based active learning framework for the spectral-spatial classification of hyperspectral imagery. *IEEE Journal of Selected Topics in Signal Processing*, 9(6):1074–1088, 2015.
- [72] Charles Sutton and Andrew McCallum. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373, 2011.
- [73] Richard Szeliski, Ramin Zabih, Daniel Scharstein, Olga Veksler, Vladimir Kolmogorov, Aseem Agarwala, Marshall Tappen, and Carsten Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):1068–1080, 2008.
- [74] Yuliya Tarabalka, Jón Atli Benediktsson, and Jocelyn Chanussot. Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques. *IEEE Transactions on Geoscience and Remote Sensing*, 47(8):2973–2987, 2009.
- [75] Yuliya Tarabalka, Mathieu Fauvel, Jocelyn Chanussot, and Jón Atli Benediktsson. Svm-and mrf-based method for accurate classification of hyperspectral images. *IEEE Geoscience and Remote Sensing Letters*, 7(4):736–740, 2010.
- [76] Ben Taskar, Carlos Guestrin, and Daphne Koller. Max-margin markov networks. In *Advances in neural information processing systems (NIPS)*, pages 25–32, 2004.
- [77] Brandt CK Tso and Paul M Mather. Classification of multisource remote sensing imagery using a genetic algorithm and markov random fields. *IEEE Transactions on Geoscience and Remote Sensing*, 37(3):1255–1260, 1999.
- [78] Devis Tuia, Gabriele Moser, Bertrand Le Saux, Benjamin Bechtel, and Linda See. 2017 ieee grss data fusion contest: Open data for global multimodal land use classification [technical committees]. *IEEE Geoscience and Remote Sensing Magazine*, 5(1):70–73, 2017.

- [79] Devis Tuia, Michele Volpi, and Gabriele Moser. Getting pixels and regions to agree with conditional random fields. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3290–3293. IEEE, 2016.
- [80] Andrea Vedaldi and Brian Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *International Conference on Multimedia*, pages 1469–1472. ACM, 2010.
- [81] Liguo Wang and Qunming Wang. Subpixel mapping using markov random field with multiple spectral constraints from subpixel shifted remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 10(3):598–602, 2013.
- [82] Zhaowen Wang, Nasser M Nasrabadi, and Thomas S Huang. Spatial-spectral classification of hyperspectral images using discriminative dictionary designed by learning vector quantization. *IEEE Transactions on Geoscience and Remote Sensing*, 52(8):4808–4822, 2014.
- [83] Jan D Wegner, Javier A Montoya-Zegarra, and Konrad Schindler. A higher-order crf model for road network extraction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1698–1705. IEEE, 2013.
- [84] Jan Dirk Wegner, Ronny Hansch, Antje Thiele, and Uwe Soergel. Building detection from one orthophoto and high-resolution insar data using conditional random fields. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 4(1):83–91, 2011.
- [85] Ting-Fan Wu, Chih-Jen Lin, and Ruby C Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5(Aug):975–1005, 2004.
- [86] Junshi Xia, Jocelyn Chanussot, Peijun Du, and Xiyan He. Spectral-spatial classification for hyperspectral data using rotation forests with local feature extraction and markov random fields. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5):2532–2546, 2015.
- [87] Futian Yao, Yuntao Qian, Zhenfang Hu, and Jiming Li. A novel hyperspectral remote sensing images classification using gaussian processes with conditional random fields. In *International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, pages 197–202. IEEE, 2010.
- [88] Jun Yue, Wenzhi Zhao, Shanjun Mao, and Hui Liu. Spectral-spatial classification of hyperspectral images using deep convolutional neural networks. *Remote Sensing Letters*, 6(6):468–477, 2015.
- [89] Guangyun Zhang and Xiuping Jia. Simplified conditional random fields with class boundary constraint for spectral-spatial based remote sensing image classification. *IEEE Geoscience and Remote Sensing Letters*, 9(5):856–860, 2012.
- [90] Guangyun Zhang, Xiuping Jia, and Jiankun Hu. Superpixel-based graphical model for remote sensing image mapping. *IEEE Transactions on Geoscience and Remote Sensing*, 53(11):5861–5871, 2015.
- [91] Ji Zhao, Yanfei Zhong, Hong Shu, and Liangpei Zhang. High-resolution image classification integrating spectral-spatial-location cues by conditional random fields. *IEEE Transactions on Image Processing*, 25(9):4033–4045, 2016.
- [92] Ji Zhao, Yanfei Zhong, Yunyun Wu, Liangpei Zhang, and Hong Shu. Sub-pixel mapping based on conditional random fields for hyperspectral remote sensing imagery. *IEEE Journal of Selected Topics in Signal Processing*, 9(6):1049–1060, 2015.
- [93] Ji Zhao, Yanfei Zhong, and Liangpei Zhang. Detail-preserving smoothing classifier based on conditional random fields for high spatial resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5):2440–2452, 2015.
- [94] Ping Zhong and Runsheng Wang. A multiple conditional random fields ensemble model for urban area detection in remote sensing optical images. *IEEE Transactions on Geoscience and Remote Sensing*, 45(12):3978–3988, 2007.
- [95] Ping Zhong and Runsheng Wang. Learning conditional random fields for classification of hyperspectral images. *IEEE Transactions on Image Processing*, 19(7):1890–1907, 2010.

- [96] Ping Zhong and Runsheng Wang. Modeling and classifying hyperspectral imagery by crfs with sparse higher order potentials. *IEEE Transactions on Geoscience and Remote Sensing*, 49(2):688–705, 2011.
- [97] Yanfei Zhong, Xuemei Lin, and Liangpei Zhang. A support vector conditional random fields classifier with a mahalanobis distance boundary constraint for high spatial resolution remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(4):1314–1330, 2014.
- [98] Yanfei Zhong, Ji Zhao, and Liangpei Zhang. A hybrid object-oriented conditional random field classification framework for high spatial resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 52(11):7023–7037, 2014.
- [99] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017.