Remote Sensing for a Changing Europe
D. Maktav (Ed.)
IOS Press, 2009

doi:10.3233/978-1-58603-986-8-378

Hyperspectral image segmentation using FSMLR with Jeffreys prior

J.S. Borges & A.R.S. Marçal
Faculdade de Ciências, Universidade do Porto, Portugal

J.M. Bioucas-Dias
Instituto de Telecomunicações – Instituto Superior Técnico, Lisboa, Portugal

Keywords: hyperspectral images, image segmentation, feature selection

ABSTRACT: The segmentation of satellite images is a valuable tool to obtain useful information about the spatial distribution of different land cover types. The use of segmentation algorithms instead of the traditional pixel-by-pixel classifiers used to produce land cover maps results on images that exhibit a more homogeneous distribution of classes, showing the piecewise spatial continuity of the real world.
Several segmentation and classification methods are being developed to properly handle the high dimensionality of hyperspectral images. An example is a Bayesian segmentation procedure based on discriminative classifiers with a Multi-Level Logistic Markov-Gibbs prior. This method adopts the Fast Sparse Multinomial Logistic Regression as discriminative classifier, a method that promotes sparsity by including a Laplacian prior. However, the use of this type of prior requires an extensive search to for the best parameter of sparsity. In this work, a modification to this method is introduced. Instead of using the Laplacian Prior to enforce the sparsity of FSMLR classifier, the Jeffreys prior is used. This prior avoids the need to proceed to an extensive search for the best parameter, and also keeps the sparsity of the densities estimators, resulting on a faster and competitive segmentation procedure. The results of the application of this new approach to the benchmarked dataset Indian Pines show the effectiveness of the proposed method when compared with that using the Laplacian prior.

1 INTRODUCTION

The classification of satellite images has been used as a powerful tool to access information about the land cover type distribution on a given scene. Until recently, pixel based classification algorithms were the most popular. These classification algorithms base their decisions in the spectral signature of each individual pixel. Recently, there is been a development of classification methods that use information from adjacent pixels. The use of contextual information together with spectral information results on segmentation algorithms that, when compared with pixel based classification algorithms, produce land cover maps that describe better the spatial homogeneity of the real world. This is one reason why segmentation methods are becoming more popular for the production of land cover maps. Other reason is the fact that adjacent pixels are more likely to have the same label, and so, this information can help to improve classification accuracies.

Image segmentation procedures are widely used for many applications such as remote sensing, medical imaging, face and fingertip recognition, machine vision, etc. Theoretically, we might say that, any segmentation algorithm should be able to give a solution to any segmentation problem. Nevertheless, each application has specific characteristics that lead to the development of algorithms somewhat devoted to a specific problem. Regarding the remote sensing application, namely

the segmentation algorithms for land cover classification of satellite images, there are different approaches such as thresholding, edge or region segmentations (Pal & Pal 1993). Region segmentation algorithms integrate the spatial and spectral information to take advantage of the complementarities that both sources can provide. Segmentation algorithms based on morphological profiles (Benediktsson *et al.* 2003), endmember extraction (Plaza *et al.* 2006), hierarchical segmentation (Tilton *et al.* 2006) and Markov random fields (Li 2001) have shown good results in segmentation of satellite images. One of the most explored approaches is the Markov Random Fields (MRF). This type of model allows contextual constraints to be incorporated by modeling the spatial neighborhood of a pixel as a spatially distributed random process. Hidden MRF were introduced in image segmentation (Marroquin *et al.* 2003) and (Tso & Olson 2005) with different approaches. However, these types of algorithms showed some problems when dealing with high dimensionality datasets.

The possibility of having an almost continuous spectral signature of the image pixels makes hyperspectral sensors powerful tools to better identify the different materials present in the land cover. Hyperspectral satellite images are characterized for having hundreds of spectral bands, which alongside the advantage of characterizing the different materials in more detail, also have the problem of producing high dimensional datasets. When supervised algorithms are considered, the high dimension of these images, together with limited number of training samples, bring about problems related to Hughes phenomenon (Hughes, 1968) or curse of dimensionality. When conventional pixel based classification algorithms are considered, the Hughes phenomenon can be overcome by increasing the number of training samples, when possible, or by reducing dimensionality. When spatial information is added, the problem of dimensionality increases. A solution to circumvent this problem is the discriminative approach. Discriminative approaches hold the state of the art of hyperspectral image segmentation (Camps-Valls & Bruzzone 2005). In this approach the difficulties in learning class densities are overcome by learning directly the densities of the labels, given the features. One of the most consolidated discriminative supervised classification tools is the Support Vector Machines (SVMs). They have been successfully used for hyperspectral data classification due to their ability to deal with large input spaces efficiently and to produce sparse solutions (Camps-Valls & Bruzzone 2005).

Discriminative approaches incorporate neighborhood interactions in the labels as well as the observed data and have being presented in recent years. The Discriminative Random Fields (DRF) framework proposed by Kumar & Herbert (2006) is an example. In the vein of this approach, we present a Bayesian segmentation approach which improves the classification performance of discriminative classifiers by adding contextual information in the form of spatial dependencies. The major difference of our work from the one presented by Kumar & Herbert (2006) is the way that the parameters are learnt. DRFs learn all the model parameters simultaneously, leading to hard and complex procedures. On the contrary, in the proposed approach, the parameters are learnt in two consecutive, but non simultaneous steps. As a consequence, the proposed method leads to much lighter procedures, still displaying very good results.

The Bayesian segmentation method for hyperspectral images here presented learn the class densities in a supervised fashion with a modified discriminative Fast Sparse Multinomial Regression (FSMLR) (Borges *et al.* 2006). The FSMLR is a fast version of the Sparse Multinomial Regression (SMLR) (Krishnapuram *et al.* 2005). This method is a sparse classification algorithm capable of dealing with high dimensional datasets. The FSMLR uses a Laplacian prior to enforce the sparsity on the class parameters. The degree of sparseness of the class densities estimates is controlled by a sparsity parameter, which has to be tuned by the user. When dealing with high dimensional datasets, such as hyperspectral images, this task may become time consuming. This can be circumvent by using a parameter-free prior. In this work we present a modified version of the Bayesian hyperspectral image segmentation with discriminative class learning by introducing the Jeffreys prior (Bioucas-Dias 2006). The Jeffreys prior keeps the sparsity of the FSMLR without the need to tune any parameter, controlling the complexity of the learned classifier and, therefore, achieving good generalization capabilities.

Having learned the class densities, the spatial dependencies are enforced by a Multi-Level Logistic (MLL) Markov-Gibs prior, which favors neighboring labels of the same class. To compute an approximation to the Maximum A Posteriori probability (MAP) segmentation, we adopt the α-Expansion graph cut based algorithm proposed in (Boykov *et al.* 2001). This tool is computationally efficient and yields nearly optimum solutions.

The segmentation method presented is applied to the benchmarked hyperspectral dataset Indian Pines (Landgrebe 2003).

The paper is organized as follows. In section 2 the methods are presented: FSMLR classifier with Jeffreys prior and the segmentation procedure based on MLL Markov Gibs prior. Section 3 presents the results of the application to the hyperspectral image and section 4 the concluding remarks.

2 METHODS

A segmentation can be interpreted as an image of labels $\mathbf{y} = \{y_i\}_{i \in S}$ where $y_i \in L = \{1,2,\ldots,K\}$. Let $\mathbf{x} = \{x_i \in R^d, i \in S\}$ be the observed multi-dimensional image, also known as feature image. The goal of the segmentation is to estimate $\mathbf{y}$, having observed $\mathbf{x}$. Regarding the Bayesian framework, this estimation is done by maximizing the posterior distribution $p(\mathbf{x} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{x}) p(\mathbf{y})$, where $p(\mathbf{y} \mid \mathbf{x})$ is the likelihood function (or the probability of feature image) and $p(\mathbf{y})$ is the prior over the classes.

The approach here presented, makes use of the discriminative FSMLR classifier (Borges *et al.* 2006) to learn the class densities $p(y_i \mid x_i)$. The likelihood is then given by $p(x_i \mid y_i) = p(y_i \mid x_i) p(x_i) / p(y_i)$. Noting that $p(x_i)$ does not depend on the labeling $\mathbf{y}$ and assuming $p(y_i) = 1/K$, we have

$$p(\mathbf{x} \mid \mathbf{y}) \propto \prod_{i \in S} p(y_i \mid x_i) / p(y_i) \tag{1}$$

where conditional independence is understood.

In the following sections, the FSMLR method, yielding the density $p(\mathbf{y} \mid \mathbf{x})$, the MLL prior $p(\mathbf{y})$ and α -Expansion optimization algorithm are briefly described.

2.1 *Class density estimation*

Given the training set $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, the SMLR algorithm learns a multi-class classifier based on the multinomial logistic regression. By incorporating a prior, this method performs simultaneously feature selection, to identify a small subset of the most relevant features, and learns the classifier itself (Krishnapuram *et al.* 2005). The goal is to assign to each site $i \in S$ the probability of $y_i = k$, for *k*=1, …, *K*. In particular, if $y_i = \left[y^{(1)}, \ldots, y^{(K)}\right]^T$ is a 1-of-*K* encoding of the *K* classes, and if $w^{(k)}$ is the feature weight vector associated with class *k*, then the probability of $y_i^{(k)} = 1$ given x_i is

$$P\left(y_i^{(k)} = 1 \mid x_i, w\right) = \frac{\exp\left(w^{(k)^T} h(x_i)\right)}{\sum_{k=1}^{K} \exp\left(w^{(k)^T} h(x_i)\right)} \tag{2}$$

where $w = \left[w^{(1)^T}, \ldots, w^{(K)^T}\right]$ and $h(x) = \left[h_1(x), \ldots, h_l(x)\right]$ is a vector of *l* fixed functions of the input, often termed features. Possible choices for this vector are linear $h(x_i) = \left[1, x_{i,1}, \ldots, x_{i,d}\right]^T$, where $x_{i,j}$ is the *j*th component of x_i), and kernel ($h(x) = \left[1, K(x, x_1), \ldots, K(x, x_n)\right]^T$, where $K(.,.)$ is some symmetric kernel function). The latter nonlinear mapping guarantees that the transformed samples are more likely to be linearly separable. A popular kernel used in image classification is the Gaussian Radial Basis Function (RBF): $K(\mathbf{x} \mid z) = -\exp\left(\left\|\mathbf{x} - z\right\|^2 / 2\sigma^2\right)$.

The MAP estimate of *w* is

$$\hat{w}_{MAP} = \arg\max_{w} L(w) = \arg\max_{w} \left[l(w) + \log p(w)\right] \tag{3}$$

where $l(w)$ is the log-likelihood function and $p(w)$ is a prior on w.

The SMLR presented by Krishnapuram et al. (2005) uses a Laplacian prior on w, to control the degree of sparseness of $\hat{w}_{MAP}$. The Laplacian prior is given by $p(w) \propto \exp\left(-\lambda \|w\|_1\right)$ where λ acts as a tunable regularization parameter. The process of selecting the optimum λ is usually done by cross-validation trough the training process. In the case of high dimensional datasets, such as hyperspectral images, this search often becomes a time consuming task. The process of estimating the class densities must be repeated as many times as the number of λ values to be tested. In order to overcome this problem, we introduce a parameter-free prior in the estimation of class densities: the Jeffreys prior (Bioucas-Dias 2006). The Jeffreys prior is given by $p(w) \propto 1/\|w\|_1$ having no longer a sparsity parameter to tune. Adopting this prior, the optimization process is done in a similar way to that done in Krishnapuram et al.(2005).

The weights w are learned using bound optimization tools (Lange 2004), making possible to perform exact MAP multinomial logistic regression, with the same cost as the original iterative reweighted least squares algorithm for maximum likelihood estimation (see (Krishnapuram *et al.* 2005) for details).

The solution of (2) is then given by the iterative equation:

$$\hat{w}^{t+1} = \Gamma^{(t)}\left(\Gamma^{(t)} B \Gamma^{(t)} - I\right)^{-1} \Gamma^{(t)}\left(B\hat{w}^{(t)} - g(\hat{w}^{(t)})\right), \tag{4}$$

where $\Gamma^{(t)}$ is given by

$$\Gamma^{(t)} = diag\left\{\left|\hat{w}_1^{(t)}\right|,\ldots,\left|\hat{w}_{d(K-1)}^{(t)}\right|\right\}, \tag{5}$$

and

$$B \equiv -\frac{1}{2}\left[I - 11^T/K\right] \otimes \sum_{i=1}^{S} x_i x_i^T, \tag{6}$$

and $1 = [1,1,\ldots,1]^T$, $g(w)$ is the gradient of $l(w)$ given by

$$g(w) = \sum_{i=1}^{S}\left(y'_i - p_i(w)\right) \otimes x_i, \tag{7}$$

with $y'_i = \left[y_i^{(1)},\ldots,y_i^{(K-1)}\right]^T$ and $p_i(w) = \left[p_i^{(1)}(w),\ldots,p_i^{(K-1)}(w)\right]^T$.

As can be observed in equation (4) there is no parameter to be defined by the user. The weights estimation procedure with the inclusion of Jeffreys prior instead of the Laplacian prior becomes independent of the selection of any parameter.

In practice, the computational cost of solving the linear system in equation (4) is often prohibitive. Regarding the application of SMLR to hyperspectral images, this becomes a problem since the cost at each iteration is of the order of $(dK)^3$ where the number of bands d is usually very large.

In order to avoid this problem, a modification to the iterative method used in SMLR can be done. This modification results in a faster and more efficient algorithm: the Fast-SMLR (FSMLR) (Borges *et al.* 2006). FSMLR uses the Block Gauss-Seidel method (Quarteroni *et al.* 2000) to solve the system (4). The modification consists, at each iteration, in solving blocks corresponding to the weights belonging to the same class, instead of computing the complete set of weights. Using this technique, what happens is that, at each iteration, K systems of equal dimension to the number of samples are solved. This results in an improvement in terms of computational effort of the order of K^2.

2.2 *Including spatial information with the Markov-Gibs prior*

The sparsity enforced by the Jeffreys prior on the estimation of class densities with FSMLR is a key step in the image classification process. However, the estimation of class densities does not include any spatial information about the classes' dispersion in the image. The information of each pixel neighborhood is introduced by the MLL prior. The MLL prior is a MRF that favors neighboring labels of the same class. In this way, we will be able to better model the piecewise smooth of real world images.

The prior over classes $p(w)$ in equation (1) will be defined in a way that allows to model the spatial distribution of neighboring pixels.

According to the Hammersly-Clifford theorem, the density associated with a MRF is a Gibb's distribution (Geman & Geman 1984). Therefore, the prior model for segmentation has the structure

$$p(\mathbf{y}) = \frac{1}{Z}\exp\left(-\sum_{c\in C} V_c(\mathbf{y})\right), \tag{8}$$

where Z is the normalizing constant and the sum is over the prior potentials $V_c(\mathbf{y})$ for the set of cliques C over the image, and

$$-V_c(\mathbf{y}) = \begin{cases} \alpha_{y_i} & if \ |c| = 1 \quad \text{(single clique)} \\ \beta_c & if \ |c| > 1 \quad \text{and } \forall_{i,j\in c}\, y_i = y_j \\ -\beta_c & if \ |c| > 1 \quad \text{and } \exists_{i,j\in c}\, y_i \neq y_j \end{cases} \tag{9}$$

where β_c is a non-negative constant.

Equation (8) can be written as

$$p(\mathbf{y}) = \frac{1}{Z} e^{\beta n(\mathbf{y})}, \tag{10}$$

where $n(\mathbf{y})$ denotes the number of cliques having the same label, if we let $\alpha_k = \alpha$ and $\beta_c = \beta/2 > 0$. This choice gives no preference to any label nor to any direction.

The conditional probability $p(y_i = k \mid y_j, j \in S - i)$ is then given by

$$p(y_i = k \mid y_{N_i}) = \frac{e^{\beta n_i(k)}}{\sum_{k=1}^{K} e^{\beta n_i(k)}} \tag{11}$$

where $n_i(k)$ is the number of sites in the neighborhood of site i, N_i, with label k.

2.3 *Segmentation algorithm*

All the process was initialized under the Bayesian framework, leading us to the maximization of the posterior distribution $p(\mathbf{x} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{x}) p(\mathbf{y})$. As described in sections 2.1 and 2.2, the class densities were learned using the FSMLR with Jeffreys prior, while the prior over classes, $p(\mathbf{y})$, was modeled by a MLL Markov-Gibs prior.

This way we have that the MAP segmentation is given by

$$\begin{aligned} \hat{\mathbf{y}} &= \arg\max_{\mathbf{y}} \ p(\mathbf{x} \mid \mathbf{y}) p(\mathbf{y}) = \arg\max_{\mathbf{y}} \sum_{i\in S} \log p(x_i \mid y_i) + \beta n(\mathbf{y}) \\ &= \arg\min_{\mathbf{y}} \sum_{i\in S} -\log p(x_i \mid y_i) + \beta \sum_{i,j\in c} \delta(y_i - y_j), \end{aligned} \tag{12}$$

The minimization of (12) is a hard combinatorial optimization problem. Graph cut techniques from combinatorial optimization are able to find the global minimum for some multi-dimensional

energy functions. The minimization problem in (12) presents characteristics that allow us to apply the α-Expansion algorithm (Boykov *et al.* 2001), which achieves very good approximations for this problem.

The α-Expansion algorithm makes use of a min-cut/max-flow algorithm presented by Boykov & Kolmogorov (2004).

3 APPLICATION TO A HYPERSPECTRAL IMAGE

To evaluate the performance of the segmentation algorithm presented in this paper, the well-known hyperspectral AVIRIS spectrometer Indian Pines 92 from Northern Indiana was used (Landgrebe, 2003). This benchmarked dataset has been frequently used to test several techniques in the processing of hyperspectral images allowing for a good evaluation.

The ground truth data image consists of 145 x 145 pixels of the AVIRIS image in 220 contiguous spectral bands. Experiments were carried out without 20 noisy bands (Landgrebe, 2003). Due to the insufficient number of training samples, seven classes were discarded, leaving a dataset with 9 classes distributed by 9345 pixels. This dataset was randomly divided into a set of 4757 training samples and 4588 validation samples.

The spatial distribution of the class labels is presented in figure 1.

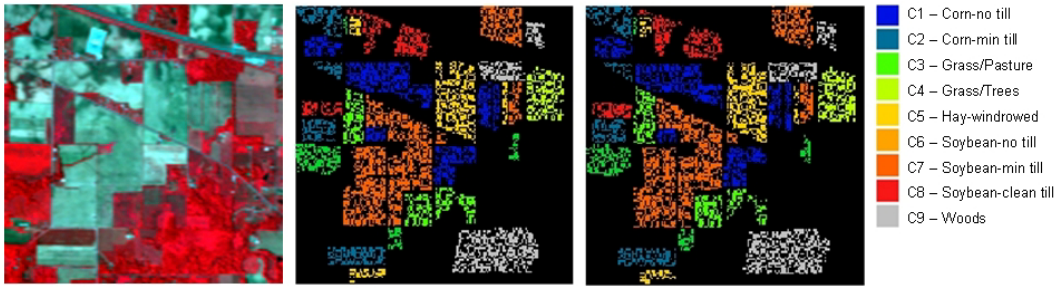

Figure 1. AVIRIS image used for testing. Left: RGB(50, 27, 17) color composite; Centre: training areas; Right: validation areas.

The results presented here are the overall accuracy measured in the independent (validation) dataset with 4588 samples. Experiments were made using 10%, 20% and 50% of the training set with a linear and a RBF kernel to $h(x_i)$. When $h(x_i)$ is set to a linear function the complete training was also used to train the segmentation algorithm. In the GC α-Expansion method a β=1.5 was defined when the complete training set was used, and β =4 for subsets of the training data. The overall accuracy results from are presented in table 1.

Table 1. Overall accuracy of Bayesian segmentation using 10%, 20%, 50% and the complete training set, with $h(x_i)$ Linear and RBF, using a Laplacian and a Jeffreys prior.

		Size of training set			
		10%	20%	50%	100%
h linear	Laplacian prior	86.05%	89.45%	89.69%	95.60%
	Jeffreys prior	86.18%	88.58%	90.43%	95.66%
h RBF	Laplacian prior	92.11%	94.62%	97.86%	---
	Jeffreys prior	89.84%	95.07%	96.71%	---

By the analysis of table 1, it is possible to observe that the Jeffreys prior achieves competitive results with Laplacian prior. The performance of the classifier was found to be nearly independent of the prior used for all training sets tested, the variations on the overall accuracies are minimal. The increase in the size of training set results in better accuracies for all methods. The results presented in table 1 for the Laplacian prior had already been evaluated in other work (Borges et al., 2007) and proved to be very competitive with the state of the art algorithms for hyperspectral image segmentation. The introduction of Jeffreys prior was able to keep the good performance of the Bayesian segmentation method proposed. It should be noted that with this prior there is no need for searching the parameter that best controls sparsity, something that has to be done with the Laplacian prior. This reduces significantly the time needed to classify the image. The reduction is of the order of the number of sparsity parameters to be tested. Moreover, the sparsity achieved by the FSMLR when using a Jeffreys prior is higher than with the Laplacian prior. This can be observed in table2, which presents the number of significant features selected by each prior, considering $h(x_i)$ as a linear function. In both cases there is a significantly reduction in the number of features (bands) used to produce the land cover map. Recall that the number of bands considered in the experiments were 200. Using the Jeffreys prior and only 10% of the training set it is possible to achieve an overall accuracy of 86.18% using the information of only 18 spectral bands. As the size of training set grows, the number of bands selected by the priors also grows as well as the overall accuracies. The best accuracy achieved with a linear kernel (95.66%) is achieved using only 51 spectral bands from the 200 bands considered initially.

Table 2. Number of significant features selected by each prior, with $h(x_i)$ linear.

	Size of training set			
	10%	20%	50%	100%
Laplacian prior	34	49	71	105
Jeffreys prior	18	27	39	51

4 CONCLUSIONS

In this paper we have presented a Bayesian hyperspectral image segmentation algorithm that uses the non-informative parameter-free Jeffreys prior to compute the class densities with the FSMLR algorithm. This is a first step to the whole segmentation process. After estimating the class densities with the discriminative classifier FSMLR, the segmentation method includes spatial information of the neighborhood of each pixel by adopting a MLL Markov-Gibs prior over the classes. The MAP segmentation is carried out using the CG α-expansion algorithm.

The inclusion of the Jeffreys prior instead of the Laplacian prior in the FSMLR method, was able to keep the good performance of the Bayesian segmentation algorithm, while at the same time there was an improvement of the sparsness of the classifier. Moreover, this sparsity improvement was achieved without the need of an extensive search for the parameter that best controls the sparsity, carried out through a cross-validation based model selection. This process, when dealing with high dimensional datasets like hyperspectral images is a challenging problem. The inclusion of the Jefreys prior results on a greatly reduced computational expense mantaining the good performance of the segmentation algorithm

ACKNOWLEDGEMENTS

The first author would like to thank the Fundação para a Ciência e a Tecnologia (FCT) for the financial support (PhD grant SFRH/BD/ 17191/2004). The authors acknowledge Vladimir Kolmo-

gorov for the max-ow/min-cut C++ code made available on the web, see (Boykov et al, 2001) for more details; and David Landgrebe for providing the AVIRIS data.
This work was supported by the Fundação para a Ciência e Tecnologia, under the project PDCTE/CPS/49967/2003, and by the Instituto de Telecomunicações under the project IT/LA/325/2005, and by CICGE under POCI 2010 programme.

REFERENCES

Benediktsson, J.A., Pesaresi, M. & Arnason, K. 2003. Classification and feature extraction for remote sensing images from urban areas based on morphological transformations. *IEEE Trans. Geoscience and Remote Sensing* 41:1940–1949.

Bioucas-Dias, J.M. 2006. Bayesian Wavelet-Based Image Deconvolution: A GEM Algorithm Exploiting a Class of Heavy-Tailed Priors. *IEEE Trans. Image Processing* 15(4):937–951.

Borges, J.S., Bioucas-Dias, J. & Marçal, A.R.S. 2006. Fast Sparse Multinomial Regression Applied to Hyperspectral Data. *In: Campilho, A., Kamel, M. (eds.) ICIAR 2006. LNCS* 4142:700–709.

Borges, J.S., Bioucas-Dias, J. & Marçal, A.R.S. 2007. Bayesian Hyperspectral Image Segmentation with Discriminative Class Learning. *In: Pattern Recognition and Image Analysis: 3rd Iberian Conference, IbPRIA 2007, Lecture Notes in Computer Science, Girona, Spain* I:22–29.

Boykov, Y. & Kolmogorov, V. 2004. An experimental comparison of mincut/max-flow algorithms for energy minimization in vision. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 26:1124–1137.

Boykov, Y., Veksler, O. \& Zabih, R. 2001. Fast Approximate Energy Minimization via Graph Cuts. *IEEE Trans. Pattern Analysis and Machine Intelligence* 23(11):1222–1239.

Camps-Valls, G. & Bruzzone, L. 2005. Kernel-based methods for hyperspectral image classification. *IEEE Trans. Geoscience and Remote Sensing* 43(6):1351–1362.

Geman, S. & Geman, D. 1984. Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 6:721–741.

Hughes, G.F. 1968. On the mean accuracy of statistical pattern recognizers. *IEEE Trans. Information Theory* 14:55–63.

Krishnapuram, B., Carin, L., Figueiredo, M.A.T. & Hartemink, A.J. 2005. Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds. *IEEE Trans. Pattern Analysis and Machine Intelligence* 27(6):957–968.

Kumar, S. & Hebert, M. 2006. Discriminative Random Fields. *International Journal of Computer Vision* 68(2):179–202.

Landgrebe, D.A. 2003. *Signal Theory Methods in Multispectral Remote Sensing*. New Jersey: John Wiley and Sons, Inc..

Lange, K. 2004. *Optimization.* Springer Texts in Statistics, New York: Springer-Verlag.

Li, S.Z., 2001. *Markov Random Field Modeling in Image Analysis*. Tokyo: Springer-Verlag.

Marroquin, J.L., Santana, E.A. & Botello, S. 2003. Hidden Markov Measure Field Models for Image Segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 25(11):1380–1387.

Pal, N.R. & Pal, S.K. 1993. A review on image segmentation techniques. *Pattern Recognition* 26(6):1277–1294.

Plaza, A., Benediktsson, J., Boardman, J., Brazile, J., Bruzzone, L., Camps-Valls, G., Chanussot, J., Fauvel, M., Gamba, P., Gualtieri, A., Tilton, J. & Trianni, G. 2006: Advanced Processing of Hyperspectral Images. *IEEE IGARSS 2006 Proceedings* IV:1974–1979.

Quarteroni, A., Sacco, R & Saleri, F. 2000. *Numerical Mathematics* TAM Series (37). New York: Springer-Verlag.

Tilton, J.C., Lawrence, W.J. & Plaza, A. 2006. Utilizing hierarchical segmentation to generate water and snow masks to facilitate monitoring change with remotely sensed data. *GIScience & Remote Sensing* 43:39–66.

Tso, B. & Olsen R.C. 2005. Combining spectral and spatial information into hidden Markov models for unsupervised image classification. *International Journal of Remote Sensing* 26(10):2113–2133.