SEMI-SUPERVISED HYPERSPECTRAL IMAGE CLASSIFICATION BASED ON A MARKOV RANDOM FIELD AND SPARSE MULTINOMIAL LOGISTIC REGRESSION

Jun Li, José M. Bioucas-Dias

Instituto de Telecomunicações,
Instituto Superior Técnico, TULisbon,
1049-001, Lisboa, Portugal

Antonio Plaza

Department of Technology of Computers and
Communications, University of Extremadura,
E-10071 Caceres, Spain

ABSTRACT

This paper introduces a new semi-supervised classification and segmentation approach tailored to hyperspectral images. The posterior distributions of the classes are modeled by the multinomial logistic regression. The contextual information inherent to the spatial configuration of the image pixels is modeled by a Multi-Level Logistic (MLL) Markov-Gibbs random field. The multinomial logistic regressors, assumed to be random vectors with independent Laplacian components, are learned using the recently introduced LORSAL algorithm. The maximum a posteriori (MAP) segmentation is computed via the α-Expansion algorithm, a powerful graph cut based approach to integer optimization. The effectiveness of the proposed methodology is illustrated by classifying simulated and real data sets. Comparisons with state-of-art methods are also included.

1. INTRODUCTION

In the recent years, new techniques have been developed in the fields of image classification and segmentation. Some of these techniques have been applied to remote sensing images, yielding effective results [1]. However, the classification and segmentation of high dimensional datasets, such as hyperspectral images, is still a challenging endeavor. Hurdles, such as the Hughes phenomenon, come out as the data dimensionality increases; in order to obtain an acceptable classification accuracy, large number of labeled samples are required, which may be difficult, expensive, or sometimes impossible to get. These difficulties foster active research on supervised and semi-supervised algorithms targeted at high dimensional data sets and limited training samples [2, 3].

The discriminative approach to classification circumvents part of the above difficulties by inferring the boundaries between classes in the feature space [4, 5]. Discriminative approaches have shown success in dealing with small class distances, high dimensionality, and limited training sets in hyperspectral classification [6]. The support vector machines (SVMs) [7] and multinomial logistic regression [8] are among the state-of-the-art discriminative techniques to classification. Due to their ability to deal with large input spaces efficiently and to produce sparse solutions, SVMs have been successfully used for hyperspectral supervised classification [6, 9, 10].

The multinomial logistic regression has the advantage of learning the class distributions themselves. Sparse multinomial logistic regression methods are available [11]. More recently, the introduction of the LORSAL (logistic regression via variable splitting and augmented Lagrangian) algorithm [12] has open the door to deal with larger data sets and number of classes. These ideas have been applied to hyperspectral image classification [13].

A recent trend to improve the classification accuracy in hyperspectral classification and segmentation is to include spatial information [9, 10, 13]. These methods exploit, in a way or another, the continuity, in probability sense, of neighboring labels: it is very likely that, in an hyperspectral image, two neighboring pixels have the same label.

In this paper, we present a new semi-supervised algorithm which is an elaboration of [13]. It implements two main steps: (a) learning step, to infer the class distributions; and (b) segmentation, by inferring the labels from a posterior distribution built on the learned class distributions and on a multi-level logistic (MLL) prior. The class distributions are modeled with a multinomial logistic regression, where the regressors are computed with the LORSAL algorithm. The spatial contextual information is exploited both in defining the MLL prior and in the active learning strategy for selecting the unlabeled samples. The maximum a posterior (MAP) segmentation is computed via a min-cut based integer optimization algorithm.

The remainder of the paper is organized as follows. Section 2 formulates the problem. Section 3 describes proposed semi-supervised approach. Section 4 reports segmentation results based on simulated and real hyperspectral datasets, in comparison with state-of-the-art competitors. Finally, section 5 concludes the paper with some remarks.

2. PROBLEM FORMULATION

Let $\mathcal{S} \equiv \{1, \ldots, n\}$ denote a set of integers indexing the pixels of an image, $\mathcal{L} \equiv \{1, \ldots, K\}$ denote a set K labels, $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d \times n}$ denote an image of d-dimensional feature vectors (one per pixel), and finally $\mathbf{y} = (y_1, \ldots, y_n) \in \mathcal{L}^n$ denote an image of labels. With the above definitions in place, the goal of image classification and of image segmentation is to estimate $\mathbf{y}$, having observed $\mathbf{x}$. In a Bayesian framework, this estimation is, usually, carried out by maximizing the posterior distribution $P(\mathbf{y}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{y})P(\mathbf{y})$, where $p(\mathbf{x}|\mathbf{y})$ is the likelihood function (*i.e.*, the probability of feature image given the label image) and $P(\mathbf{y})$ is the prior over the label image. Assuming conditional independency of the features given the labels, *i.e*, $p(\mathbf{x}|\mathbf{y}) = \prod_{i=1}^{i=n} p(x_i|y_i)$, then the posterior $P(\mathbf{y}|\mathbf{x})$ is

This work was supported by the Marie Curie training Grants MEST-CT-2005-021175 and MRTN-CT-2006-035927.

given by

$$\begin{aligned} P(\mathbf{y}|\mathbf{x}) &= \frac{1}{p(\mathbf{x})} p(\mathbf{x}|\mathbf{y}) P(\mathbf{y}) \\ &= \frac{1}{p(\mathbf{x})} \prod_{i=1}^{i=n} p(\mathbf{x}_i|y_i) P(\mathbf{y}) \\ &= \alpha(\mathbf{x}) \prod_{i=1}^{i=n} \frac{P(y_i|\mathbf{x}_i)}{P(y_i)} P(\mathbf{y}), \end{aligned} \tag{1}$$

where $\alpha(\mathbf{x}) \equiv \prod_{i=1}^{i=n} p(\mathbf{x}_i)/p(\mathbf{x})$ is a factor not depending on $\mathbf{y}$. In this paper we assume, without loss of generality, that $P(y_i) = 1/K$. The maximum a posteriori (MAP) segmentation is then given by

$$\widehat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{L}^n} \left\{ \left(\sum_{i=1}^{n} \log P(y_i|\mathbf{x}_i) \right) + \log P(\mathbf{y}) \right\}. \tag{2}$$

The next two subsections addresses aspects of the posterior density $P(y_i|\mathbf{x}_i)$ and of the MLL prior $P(\mathbf{y})$.

2.1. Learning the MLR regressors with the LORSAL algorithm

Let $\mathbf{y}_i = [y_i^{(1)}, ..., y_i^{(K)}]^T$ denote a "1-of-K" encoding of the K classes and $\boldsymbol{\omega}^{(k)} \in \mathbb{R}^d$ be a vector of parameters associated with class k. Note the the variables $\mathbf{y}_i$, just defined, and y_i have different structure but are equivalent [*e.g.*, $(\mathbf{y}_i = [0, 0, 1, 0]^T) \Leftrightarrow (y_i = 3)$]. The multinomial logistic regression gives the probability of class k as

$$P(y_i^{(k)} = 1|\mathbf{x}_i, \boldsymbol{\omega}) = \frac{\exp(\boldsymbol{\omega}^{(k)} \mathbf{h}(\mathbf{x}_i))}{\sum_{k=1}^{K} \exp(\boldsymbol{\omega}^{(k)} \mathbf{h}(\mathbf{x}_i)))}, \tag{3}$$

where $\boldsymbol{\omega} = [{\boldsymbol{\omega}^{(1)}}^T, ..., {\boldsymbol{\omega}^{(K)}}^T]^T$ and $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}), ..., h_l(\mathbf{x})]^T$ is a vector of l fixed functions of the input, often termed features. Since the probability (3) does not depend on translations on the regressors $\boldsymbol{\omega}^{(k)}$, we take $\boldsymbol{\omega}^{(K)} = \mathbf{0}$.

Usual choices for $\mathbf{h}$ are $\mathbf{h}(\mathbf{x}) = [1, x_1, ..., x_d]^T$, thus linear, and Kernels, *i.e.*, $\mathbf{h}(\mathbf{x}) = [1, K_{\mathbf{x},\mathbf{x}_1}, ..., K_{\mathbf{x},\mathbf{x}_l}]^T)$, where $K_{\mathbf{x},\mathbf{x}_j} = K(\mathbf{x}, \mathbf{x}_j)$ and $K(\cdot, \cdot)$ is some symmetric kernel function, often nonlinear of $\mathbf{x}$. Kernels have been largely used because they tend to improve the data separability in the transformed space. In this paper, we use a Gaussian Radial Basis Function (RBF) $K(\mathbf{x}, \mathbf{z}) = \exp[-\|\mathbf{x} - \mathbf{z}\|^2)/2\sigma^2]$, which are widely used in hyperspectral image classification [6]). From now on, d denotes the dimension of $\mathbf{h}(\mathbf{x})$.

The MAP estimate of $\boldsymbol{\omega}$ is

$$\widehat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega}} [l(\boldsymbol{\omega}) + \log p(\boldsymbol{\omega})], \tag{4}$$

where $l(\boldsymbol{\omega})$ is the log-likelihood function of $\boldsymbol{\omega}$ and $p(\boldsymbol{\omega}) \propto \exp(-\lambda \|\boldsymbol{\omega}\|_1)$ is the prior of $\boldsymbol{\omega}$, a Laplacian independent density on each component of $\boldsymbol{\omega}$; λ is a regularization parameter controlling the degree of sparseness of $\widehat{\boldsymbol{\omega}}$. Problem (4), although convex, is hard because the term $l(\boldsymbol{\omega})$ is non-quadratic and the term $\|\boldsymbol{\omega}\|_1$ is non-smooth. The SMLR algorithm introduced in [11] solves this optimization problem with $O((dK)^3)$ complexity, which is unbearable in data sets with a large number of classes or using Kernels.

The LORSAL algorithm introduced in [12] solves (4) with a much lighter computational complexity. This is achieved by replacing a difficult non-smooth convex problem with a sequence of quadratic plus diagonal l_2-l_1 problems, which are solved with a $O(d^2)$ computational complexity. Compared with the SMLR algorithm, the reduction of computational complexity is $O(dK^3)$. For more details see [12].

2.2. The Multi-Level Logistic spatial prior

In order to model the spatial information among the image labels, we adopt an (MLL) prior given by

$$P(\mathbf{y}) = \frac{1}{Z} e^{\mu \sum_{i \sim j} \delta(y_i - y_j)}, \tag{5}$$

where Z is a normalizing constant, $\delta(y)$ is the unit impulse function[1], $\mu > 0$ is a parameter controlling the likelihood that two neighboring pixels belong to the same class, and where $i \sim j$ denotes first-order neighboring sites. Note that the pairwise interaction terms $\delta(y_i - y_j)$ attach higher probability to equal neighboring labels than the other way around. In this way, the MLL prior promotes piecewise smooth segmentations.

3. PROPOSED METHOD

3.1. Supervised Segmentation

Using the LORSAL algorithm to learn $P(y_i|\mathbf{x}_i)$ and the MLL prior $P(\mathbf{y})$, and according to (2), the MAP segmentation is finally given by:

$$\widehat{\mathbf{y}} = \arg\min_{\mathbf{y} \in \mathcal{L}^n} \left\{ \sum_{i \in \mathcal{S}} - \log P(y_i|x_i) - \mu \sum_{i \sim j} \delta(y_i - y_j) \right\}. \tag{6}$$

The minimization (6) is a hard combinational optimization problem. However, given that the pairwise interaction term on the right hand side of (2) is a metric, it is possible to achieve a very good approximation using the α-Expansion graph cut based algorithm [14–17].

3.2. Semi-supervised Segmentation

Very often, the acquisition of labeled samples is very expensive. In these cases, we resort to semi-supervised classification/segmentation approaches, which are based on very few training samples. In our setup, we run a two-step iterative procedure: in the first step, we apply the LORSAL algorithm [12], to learn the logistic regressors; in the second step we compute the MAP segmentation (6) and expand the set of labeled sampled with a subset of the just obtained labeled samples.

This procedure has similarities with that of active data selection [18], [2] in which a given training set $\{(\mathbf{x}_i, y_i), i = 1, \ldots, L\}$ is sequentially increased by including labeled pairs $(\mathbf{x}_{L+1}, y_{L+1})$ selected according to a given criterion. For example, the new feature $\mathbf{x}_{L+1}$ may be chosen to maximize the information about the values that the model parameters should take. In our approach, we also increment the training set iteratively by actively selecting pairs $(\mathbf{x}_j, \widehat{y}_j)$. Note that the label $\widehat{y}_j$ attached to $\mathbf{x}_j$ is not necessarily correct, as in the canonical active learning procedures. Nevertheless, as illustrated in the next section, the proposed selection methodology is very effective. Concerning the selection criteria, we use neighborhood: since we are dealing with images, it is very likely that to neighboring pixels have the same label.

The pseudo-code for the semi-supervised algorithm is presented below, where μ is the regularization parameter, T is the initial labeled set, t is the selected labeled sets from the previous iteration, ϕ is the neighboring selection criterion, l is the log-likelihood (4), $MMiter$ is the maximum number of iterations, and tol is the tolerance to control the log-likelihood.

[1] *i.e.*, $\delta(0) = 1$ and $\delta(y) = 0$, for $y \neq 0$

Algorithm 1 Semi-supervised Segmentation Algorithm

Require: μ, T
  while $k \leq MMiter$ or $\|l_k - l_{k-1}\| \geq tol$ **do**
    $t \Leftarrow \phi(y^{(k)})$
    $T \Leftarrow T \cup t$
    $\hat{y}^{(k+1)} = \arg\max_y p(x|y)p(y)$
    $k \Leftarrow k+1$
  end while

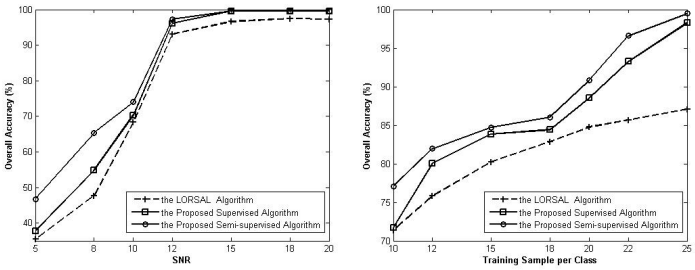

Fig. 1. Experiments with simulated data. Left: Overall accuracy as a function of SNR with 3 training samples. Right: Overall accuracy as a function of the number of training sample with SNR = 5 dB.

4. EXPERIMENT RESULTS

In this section, we evaluate the proposed algorithms with simulated and real hyperspectral data sets. In all experiments, the spectral vectors are scaled by $s > 0$ such that $\|x\|^2/(s^2 nl) = 1$ (l denotes the number of bands). RBF kernels are used as regression functions in (3). The RBF scale parameter is set to $\sigma = 0.6$. The prior regularization parameter is set to $\mu = 2$. Each value of Overall Accuracy (OA) is obtained from 10 Monte Carlo runs. As illustrated below, this setting, although not optimal for all experiments, leads to effective results.

4.1. Simulated Dataset

Simulated scenes of dimension $120 \times 120 \times 221$ (100×100 is the spatial size and 221 is the number of bands), and 10 classes are generated. Each scene is composed of spectral vectors $x = z + n$, where $x \in \mathbb{R}^{221}$, z are spectral signatures obtained from the USGS library[2] according to the MLL prior $P(\mathbf{y})$ given by (5) with $\mu = 2$, and n is zero-mean white Gaussian noise. The signal-to-noise ratio is defined as SNR$= E[\|z\|^2]/E[\|n\|^2]$.

Fig. 1, top, shows the OA as a function of the SNR. The initial training set is just 0.3% of the whole scenario (3 labeled pixels per class) randomly selected; the remaining samples were used as test set. Fig. 1, right, shows the OA results as a function of the size of the labeled samples with SNR = 5 dB. For illustrative purposes, Fig. 2, left, shows the spectral band centered at 0.5 nm of the simulated data set with SNR = 5 dB. The difficulty of the segmentation problem is evident. Fig. 2, right, shows the segmented image obtained with 10 labeled samples per class. Observe that, in all case, the proposed algorithms outperform the LORSAL algorithm, as it does not use contextual information. Moreover, in spite of using a much smaller training set, the semi-supervised algorithm outperforms the supervised counterparts in both plots.

4.2. Real AVIRIS Image

Experiments were also carried out with a real hyperspectral AVIRIS spectrometer image, the Indian Pines 92 from Northern Indiana, taken on June 12, 1992 [1]. The scene is available online[3], with 145×145 pixels and 220 spectral bands. Noisy bands, in number of 20, due to water absorption were removed. Due to the insufficient number of training samples, 7 classes were discarded, leaving a dataset with 9 classes distributed by 9345 elements. This dataset was randomly partitioned into a set of 4757 training samples and 4588 validation samples. The spatial distribution of the ground truth image is presented in Fig. 3 (a).

[2]These signatures were randomly selected from the U.S. Geological Survey (USGS) digital spectral library, available online: http://speclab.cr.usgs.gov.

[3]http://cobweb.ecn.purdue.edu/ biehl/MultiSpec/

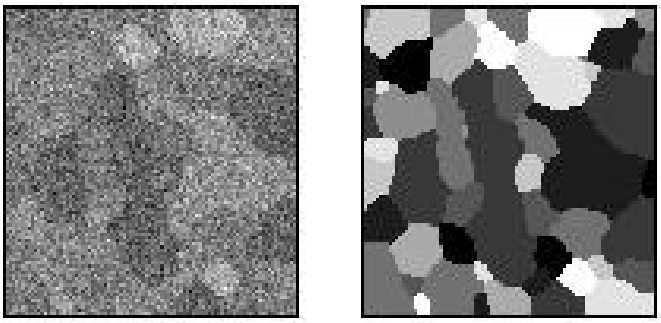

Fig. 2. Simulated scene. Left: Spectral band (0.5 nm) of a simulated data set generated with SNR = 5 dB. Right: Segmentation result obtained with 20 training samples per class.

Table 1 shows segmentation results as a function of the size of the initial labeled samples. These results are compared with those provided by LORSAL and by the state-of-the-art algorithms for semi-supervised hyperspectral classification and segmentation [10, 19, 20]. The proposed semi-supervised algorithm yields effective results, as shown in the segmentation map displayed in Fig. 3 (b), which was obtained with only 20 training samples per class, only 180 labeled samples in total.

With much less labeled samples, the proposed algorithm still produces better results than the compared competitors introduced in [10, 19, 20]. With just 10 labeled samples per class, the results obtained by the proposed supervised and semi-supervised algorithms are 77.98% and 80.15%, respectively. Using 5 labeled samples per class, the best accuracy presented in [20] is 66.04%, whereas proposed semi-supervised approach yields 72.62%, which is 6.58% larger than 66.04% from [20].

Using only spectral information and 25 labeled samples per class, the method [10] yields an OA of 73.41% and 76.20%, which are lower than those achieved by the proposed method. However, this comparison is not fair because we are using spatial information whereas the figures above were achieved without using this advantage. In a fair comparison, *i.e*, both methods using spatial information and 20% of the training set, which is about 120 labeled samples per class, the best results from [10]is 96.53%, which is 0.5% less than the proposed semi-supervised algorithm.

(a) (b)

Fig. 3. AVIRIS image. (a), ground truth image; (b), classification map with 20 labeled samples per class.

5. CONCLUDING REMARKS

The paper introduces a new semi-supervised classification and segmentation algorithm with applications to hyperspectral imagery. The algorithm is a development of the supervised segmentation method proposed in [13]: The probability of the classes are learned discriminatively with the LORSAL algorithm [12] and the spatial contextual information is modeled by a multilevel-logistic Markov random field. Unlabeled samples are actively selected and used to recursively learn the class densities.

The effectiveness of the proposed semi-supervised segmentation algorithm was illustrated with simulated data and in a comparison with state-of-the-art competitors.

6. REFERENCES

[1] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*, John Wiley, Hoboken, NJ, 2003.

[2] B. Krishnapuram, D. Williams, Y. Xue, A. Hartemink, L. Carin, and M. Figueiredo, "On semi-supervised classification," in *Proc. 18th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, 2004.

[3] O. Chapelle, M. Chi, and A. Zien, "A continuation method for semi-supervised svms," in *Proceedings of the 23rd International Conference on Machine Learning*. 2006, pp. 185–192, ACM Press.

[4] Andrew Y. Ng and Michael I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Proc. 16th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, 2002.

[5] V. Vapnik, *Statistical Learning Theory*, John Wiley, New York, 1998.

[6] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, pp. 1351–1362, 2005.

[7] B. Scholkopf and A. Smola, *Learning With KernelsSupport Vector Machines, Regularization, Optimization and Beyond*, MIT Press Series, Cambridge, MA, 2002.

[8] D. Böhning, "Multinomial logistic regression algorithm," *Annals of the Institute of Statistical Mathematics*, vol. 44, pp. 197–200, 1992.

[9] M. Fauvel, J.A. Benediktsson, J. Chanussot, and J.R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 11, pp. 3804–3814, 2008.

Table 1. Results for the AVIRIS Indiana image with the proposed algorithm. Overall accuracy (OA [%]) as a function of the number of labeled samples. [S] denotes the proposed supervised algorithm. [Semi-S] denotes the proposed semi-supervised algorithm . [L] denotes the LORSAL algorithm [12]. [10]-1 used the spectral classifiers. [10]-2 used the spectral and spatial classifiers. [20] used graph-based semi-supervised algorithm. Best results (Bold) are highlighted for each problem.

	number of labeled samples (N) per class					
methods	5	10	15	20	25	120
[S]	66.83	77.98	81.31	84.46	89.33	96.71
[Semi-S]	**72.62**	**80.15**	**83.50**	**87.14**	**91.54**	**96.99**
[L]	58.10	66.56	70.24	72.84	78.88	84.79
[10]-1 ($N \approx 25$) 73.41(SVMs) 76.20(TSVMs)						
[10]-2 ($N \approx 120$) 95.97(mean -weighted)						
[10]-2 ($N \approx 120$) 96.53(mean and standard deviation - weighted)						
[19] ($N \approx 22$) 56.42(RLDA) 45.67 (Lin-SVM) 59.16 (RBF-SVM)						
[20] ($N = 5$) 66.04(cross) 59.35(weighted)						

[10] A. Plaza, J. A. Benediktsson, J. W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, M. Marconcini, J. C. Tilton, and G. Trianni, "Recent advances in techniques for hyperspectral image processing," *Remote Sensing of Environment, in press*, 2009.

[11] B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 957–968, 2005.

[12] J. Bioucas-Dias and M. Figueeredo, "Logistic regression via variable splitting and augmented lagrangian tools," Tech. Rep., Instituto Superior Técnico, TULisbon, 2009.

[13] J. Borges, J. Bioucas-Dias, and A. Marçal, "Evaluation of Bayesian hyperspectral imaging segmentation with a discriminative class learning," in *Proc. IEEE International Geoscience and Remote sensing Symposium*, Barcelona, Spain, 2007.

[14] Y. Boykov, O. Veksler, and R. Zabih, "Efficient approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1222–1239, November 2001.

[15] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision.," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, September 2004.

[16] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 147–159, February 2004.

[17] Shai Bagon, "Matlab wrapper for graph cut," December 2006.

[18] D. Mackay, "Information-based objective functions for active data selection," *Neural Computation*, vol. 4, pp. 590–604, 1992.

[19] Tatyana V. Bandos, Lorenzo Bruzzone, and Gustavo Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Transactions on Geoscience and Remote Sensing*, 2009.

[20] G. Camps-Valls, T. Bandos, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, pp. 3044–3054, Oct 2007.