# Evaluation of Bayesian Hyperspectral Image Segmentation with a Discriminative Class Learning

Janete S. Borges and André R. S. Marçal
Faculdade de Ciências - University of Porto
Rua do Campo Alegre, 687
4169-007 Porto, Portugal
Email: jsborges@fc.up.pt and andre.marcal@fc.up.pt

José M. Bioucas-Dias
Instituto de Telecomunicações
Instituto Superior Técnico
Technical University of Lisbon
Email: bioucas@lx.it.pt

***Abstract*—A Bayesian segmentation approach for hyperspectral images is introduced in this paper. The method improves the classification performance of discriminative classifiers by adding contextual information in the form of spatial dependencies. The technique herein presented builds the class densities based on Fast Sparse Multinomial Logistic Regression and enforces spacial continuity by adopting a Multi-Level Logistic Markov-Gibs prior. State-of-art performance of the proposed approach is illustrated in a set of experimental comparisons with recently introduced hyperspectral classification/segmentation methods.**

## I. Introduction

The wide availability of hyperspectral images led to new developments in the fields of image segmentation and classification. The detailed information about spectral signatures provided by hyperspectral sensors has fostered the development of new algorithms capable of properly handling the high dimensionality of the data. The difficulties in learning high dimensional densities from a limited number of training samples (Hughes phenomenon) is one of the major problems related with this type of data, and although many progresses have been made, it is still an active area of research.

The discriminative approach in classification problems circumvents the difficulties in learning class densities by learning directly the densities of the labels given the features. Discriminative approaches hold the state-of-the art in supervised hyperspectral image classification (see, e.g. [1]). These approaches have proved to be successful in dealing with small class distances, high dimensionality, and limited training sets characteristic of hyperspectral vectors. The Support Vector Machines (SVMs) are one of the most consolidated discriminative supervised classification tools. SVMs have been successfully used for hyperspectral data classification due to their ability to deal with large input spaces efficiently, and to produce sparse solutions. One example of such an application is the work developed by Camps-Valls et. al [1]. More recently, algorithms that integrate spatial and spectral information have been presented. Markov Random Field (MRF) models allow contextual constraints to be incorporated and have been used extensively for various segmentation applications, including hyperspectral data classification (see, e.g. [2]). Other techniques using morphological profiles and segmentation hierarchies have also been proposed and their capacities analised [3].

In this work, a Bayesian segmentation approach for hyperspectral images is introduced. The method improves the classification performance of discriminative classifiers by adding contextual information in the form of spatial dependencies. Our approach is in the vein of the Discriminative Random Fields (DRF) framework introduced in [4]. The major difference concerns the way the model parameters are learnt: in DFRs, all model parameters are learnt simultaneously, leading to hard and complex procedures still under investigation; on the contrary, in the proposed approach, the multinomial and the MLL 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 Hyperspectral Image Segmentation with Discriminative Class Learning methodology used here enforces spatial dependencies by a Multi-Level Logistic (MLL) Markov-Gibs prior. This density favors labeling in which neighboring sites belong to the same class. The class densities are build on the Fast Sparse Multinomial Regression (FSMLR) [5], learned in a supervised fashion. The FSMLR is a sparse classification algorithm capable of dealing with high dimensionality datasets. In experiments performed with a benchmarked dataset (Indian Pines from the AVIRIS spectrometer), the method proved to be fast and efficient for the classification of hyperspectral data [5]. To efficiently estimate the optimal segmentation, the $\alpha$-Expansion graph cut based algorithm [6] is used. This algorithm is capable of achieving nearly optimum solutions for the discrete optimization problem given by the Maximum A posteriori Probability (MAP) segmentation.

In this work, the proposed Bayesian segmentation method with discriminative class learning is applied to urban hyperspectral data from the town of Pavia, Italy, collected by the ROSIS sensor in the framework of HySens project managed by the DLR (German Aerospace Agency) [7]. The results are compared with those obtained by Plaza et. al [3].

The paper is organized as follows. Section II presents the

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).

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.

methods used: the FSMLR classifier and the segmentation based on the MLL Markov Gibs prior. The results are presented in section III and the concluding remark in section IV.

## II. Bayesian Segmentation Methodology

Let $\mathbf{y} = \{y_i\}_{i\in\mathcal{S}}$, where $y_i \in \mathcal{L} = \{1, 2, \ldots, K\}$, be an image of labels and let $\mathbf{x} = \left\{x_i \in \mathbb{R}^d, i \in \mathcal{S}\right\}$ be the observed multi-dimensional images. The goal of the segmentation is to estimate $\mathbf{y}$, having observed $\mathbf{x}$. In a Bayesian framework, this estimation is done 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 (or the probability of feature image given the labels) and $p(\mathbf{y})$ is the prior over the classes.

In the approach here presented, the class densities $p(y_i|x_i)$ are learned by the discriminative FSMLR classifier [5]. The likelihood is then given by $p(x_i|y_i) = p(y_i|x_i)p(x_i)/p(y_i)$. Since $p(x_i)$ does not depend on the labeling $\mathbf{y}$, we have

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

where conditional independence is understood.

In this approach, the classes are assumed as likely probable: $p(y_i) = 1/K$ . Although this assumption may not be the ideal, it still leads to very good results. Anyway, the class densities can be tilted towards other distribution by using the method described in [8].

### A. Class Density Estimation

The estimation of class densities $p(\mathbf{y}|\mathbf{x})$ consists in assigning to each $x_i$ the probability of belonging to each of the $K$ classes, yielding $K$ sets of feature weights, one for each class. If $y_i = [y^{(1)}, \ldots, y^{(K)}]^T$ is a 1-of-K encoding of the $K$ classes, and if $w^{(k)}$ is the feature weight vector associated with class $k$, the multinomial logistic regression gives us the probability of $y_i^{(k)} = 1$ given $x_i$:

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

where $w = [w^{(1)^T}, \ldots, w^{(K)^T}]^T$ and $h(x) = [h_1(x), \ldots, h_l(x)]^T$ is a vector of $l$ fixed functions of the input, often termed features. Possible choices for this function are linear (*i.e.*, $h(x_i) = [1, x_{i,1}, \ldots, x_{i,d}]^T$, where $x_{i,j}$ is the $j^{th}$ component of $x_i$) and kernel (*i.e.*, $h(x) = [1, K(x, x_1), \ldots, K(x, x_n)]^T$, where $K(\cdot, \cdot)$ is some symmetric kernel function). Kernels are nonlinear mappings, thus ensuring 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): $K(\mathbf{x}, \mathbf{z}) = -\exp(\|\mathbf{x} - \mathbf{z}\|^2 / 2\sigma^2)$.

The Sparse Multinomial Logistic Regression (SMLR) algorithm incorporates a Laplacian prior to promote the sparsity in the estimate of $w$. The inclusion of the Laplacian prior lead us to the MAP estimate of $w$:

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

where $l(w)$ is the log-likelihood function and $p(w) \propto \exp(-\lambda\|w\|_1)$; $\lambda$ is a regularization parameter controlling the degree of sparseness of $\hat{w}_{MAP}$. The weights $w$ are learned using bound optimization tools [9], which allow to perform exact MAP multinomial logistic regression under a Laplacian prior, with the same cost as the original iterative reweighted least squares algorithm for maximum likelihood estimation (see [10]).

The SMLR reveals a weak point when applied to large datasets like hyperspectral images. Its practical application to this kind of data is often computationally prohibitive. The FSMLR algorithm tackles this limitation by replacing the solution of a sequence large linear system of equations with a sequence of block Gauss-Seidel iterations [10]. More specifically, in each iteration, instead of solving the complete set of weights, only blocks corresponding to the weights belonging to the same class are solved [5]. The gain in number of floating point operations is of the order of $O(K^2)$, where $K$ is the number of classes.

### B. Segmentation procedure

The segmentation process includes the spatial information so that the piecewise smooth of real world images can be considered. The MLL prior is a MRF that favors neighboring labels of the same class.

According to the Hammersly-Clifford theorem, the density associated with a MRF is a Gibb's distribution [11]. Therefore, the prior model for segmentation has the structure

$$p(\mathbf{y}) = \frac{1}{Z} \exp\left(-\sum_{c\in C} V_c(\mathbf{y})\right), \tag{4}$$

where $Z$ is the normalizing constant and the sum is over the prior potentials $V_c(\mathbf{y})$ for the set of cliques[1] $C$ over the image, and

$$-V_c(\mathbf{y}) = \begin{cases} \alpha_{y_i} & if \quad |c| = 1 \quad \text{(single clique)} \\ \beta_c & if \quad |c| > 1 \quad \text{and } \forall_{i,j\in c}\, y_i = y_j \\ -\beta_c & if \quad |c| > 1 \quad \text{and } \exists_{i,j\in c}\, y_i \neq y_j \end{cases} \tag{5}$$

where $\beta_c$ is a nonnegative constant.

Equation (4) can be written as

$$p(\mathbf{y}) = \frac{1}{Z} e^{\beta n(\mathbf{y})} \tag{6}$$

where $n(\mathbf{y})$ denotes the number of cliques having the same label, if we let $\alpha_k = \alpha$ and $\beta_c = \frac{1}{2}\beta > 0$. This choice gives no preference to any label nor to any direction.

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

$$p(y_i = k|y_{\mathcal{N}_i}) = \frac{e^{\beta n_i(k)}}{\sum_{k=1}^{K} e^{\beta n_i(k)}}, \tag{7}$$

where $n_i(k)$ is the number of sites in the neighborhood of site $i$, $\mathcal{N}_i$, having the label $k$.

[1] A clique is a set of pixels that are neighbours of one another.

The MAP segmentation is given by

$$\begin{aligned}\hat{\mathbf{y}} &= \arg\max_{\mathbf{y}} p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) \\ &= \arg\max_{\mathbf{y}} \sum_{i\in\mathcal{S}} \log p(x_i|y_i) + \beta n(\mathbf{y}) \\ &= \arg\min_{\mathbf{y}} \sum_{i\in\mathcal{S}} -\log p(x_i|y_i) - \beta \sum_{i,j\in c} \delta(y_i - y_j), \quad (8)\end{aligned}$$

where $p(\mathbf{x}|\mathbf{y}) \propto \prod_i p(y_i|x_i)$ was learned using the FSMLR algorithm. The minimization of (8) is a hard combinatorial optimization problem. However, it is possible to achieve a very good approximation using the graph cut $\alpha$-Expansion based algorithm [6]. This algorithm can be applied because the pairwise interaction term on the right hand side of (8) is equivalent to a metric[2].

## III. Experimental Results

The proposed MAP segmentation was applied to three urban hyperspectral images over the town of Pavia, Italy. This section describes the datasets and the experiments performed. the results are presented as well as a discussion and a comparison with the results presented in [3].

### A. Data Description

The data used in this work was collected by the ROSIS sensor in the framework of HySens project managed by DLR (German Aerospace Agency) [7]. The images have 115 spectral bands with a spectral coverage from 0.43 to 0.86 $\mu m$, and a spatial resolution of 1.3m. Two scenes over Pavia were made availabe, a scene over the city centre and another over Pavia University. We consider three different subsets of the full data similarly to the work presented in [3]:

- *Dataset 1* - Image over Pavia city centre with 492 by 1096 pixel in size (Fig.1(a)), 102 spectral bands (without the noisy bands) and nine ground-truth classes distributed by 5536 training samples and 103539 validation samples.
- *Dataset 2* - Image over Pavia University with 310 by 340 pixel in size (Fig.1(b)), 103 spectral bands (without the noisy bands) and nine ground-truth classes distributed by 3921 training samples and 42776 validation samples.
- *Dataset 3* - Superset of the scene over Pavia city centre, including a dense residential area, with 715 by 1096 pixel in size (Fig.1(c)) and nine ground-truth classes distributed by 7456 training samples and 148152 validation samples.

### B. Experimental Setup and Results Discussion

Experiments were carried out to access the efficiency of the presented segmentation procedure when compared to recent algorithms developed for processing hyperspectral imagery, presented in [3].

The class densities estimation described in section II.*A*. involves the choice of a kernel function. In this work, linear and RBF kernels were used in different conditions.

Linear kernels were used in the segmentation of *Dataset 1* and *Dataset 3* using the complete training set to learn the segmentation algorithm, and the complete set of validation samples was used to estimate the Overall Accuracies (OA) (Table I).

[2]A metric is obtained by adding $\beta$ to terms $-\beta\delta(y_i - y_j)$

(a) Dataset 1 (b) Dataset 2 (c) Dataset 3

Fig. 1. Datasets used

TABLE I
Overall accuracies of the MRF segmentation with linear mapping, and the results from [3], using the complete training set.

| | *Dataset 1* | *Dataset 3* |
|---|---|---|
| MRF*Seg* | 98.18% | 98.46% |
| Results from [3] | 96.03% | 97.27% |

TABLE II
Overall accuracies of the MRF segmentation with linear mapping, using different subsets of the training set.

| Training set size | 10% | 20% | 40% | 60% | 80% |
|---|---|---|---|---|---|
| Overall Accuracy | 94.03% | 96.14% | 95.85% | 96.16% | 96.75% |

The results from [3] presented in table I for *Dataset 1* were achieved with a SVM with a Poly kernel; results for *Dataset 3* are a product of a MRF-based spatial characterization where a discriminant analysis feature extraction was applied before in order to increase spectral separability. The application of our segmentation method with a linear mapping managed to improve the results under the same conditions, without any pre-processing to increase the spectral separability.

In order to evaluate the performance of our segmentation method when small training samples are considered, we randomly selected 5 subsets from the training set of *Dataset 1*, with 10%, 20%, 40%, 60% and 80% of each class to learn the segmentation algorithm, using a five-fold cross-validation method to access the parameters of the FSMLR algorithm for class densities estimation and the segmentation method. The OAs were evaluated on the complete validation set (Table II).

From tables I and II we can observe that using 60% of

TABLE III
OVERALL ACCURACIES (%) OF THE MRF SEGMENTATION WITH DIFFERENT MAPPINGS, USING DIFFERENT SUBSETS OF THE TRAINING SET, AND RESULTS FROM [3].

| Training set size | 10 | 20 | 40 | 60 | 80 | 100 |
|---|---|---|---|---|---|---|
| MRF*Seg* RBF | 97.04 | 96.33 | 96.54 | 97.37 | 97.97 | 97.90 |
| SVM-RBF [3] | 93.85 | 94.51 | 94.51 | 94.71 | 95.36 | 95.29 |

the training set, the MRF segmentation method achieved the same OA than the SVM used by [3], and using only 10% of the training set and a linear kernel we manage to get an OA above 94%.

When a RBF kernel is considered, the computational complexity increases and the process of finding the kernel parameters that gives the highest OA becomes a very slow task when a large training set is used. Considering the *Dataset 2*, we randomly selected a subset with 10% of each class present in the training samples to learn the segmentation algorithm, and measured the OA over the complete validation set. With 10% of the training set, we achieved an OA of 91.81%, 6.59% more than the OA from [3] over the same dataset, using a method that also includes spectral and spatial information (Extended Morphological Profile).

The MRF segmentation method proposed using RBF kernels in the class density estimation, was also evaluated using the *Dataset 1*. Subsets with 10, 20, 40, 60, 80 and 100 samples of each class were randomly selected from the training set, and the OAs were calculated over the complete test set. The results are presented on table III, where it is possible to observe that, regardless of the size of the training set, the MRF-Segmentation outperforms the SVM-RBF algorithm used in [3]. The advantage of using a method that includes spatial information is well shown by the comparison of the OAs achieved by both methods: with only 90 samples, the MRF-Segmentation yielded an OA of 97.04%, while the SVM-RBF with the complete training set (5536 samples) achieved an OA of 96.45%.

## IV. CONCLUSION

This paper proposes a segmentation method that uses the FSMLR method to estimate the class densities used to perform a MLL Markov-Gibs prior based segmentation.

Benchmarked datasets were used to access the performance of the segmentation method, both in terms of accuracy as well as in terms of generalization capacity. Results were compared with the results from [3], where recent classification and segmentation techniques were applied to the same datasets. The proposed segmentation method outperformed the results presented in [3] in all cases, with the experiments carried out in similar conditions.

It is well known that one of the major problems in dealing with hyperspectral imagery is the high dimensionality of the data to be processed, leading us the Hughes phenomenon. The segmentation method proposed gives excellent OA results when small training samples are used, showing high generalization capacity.

The possibility of choosing different kernels to estimate the class densities gives the user the possibility to better handle the computational expense of processing hyperspectral images. The use of linear kernels result in less computational demanding learning algorithms and are capable of achieving very good accuracies. It is possible to choose between a simpler model and use all the available training samples to learn the algorithm, or adopt a more complex model with RBF kernels, using a small set of training samples. In both ways high accuracies in the segmentation of the hyperspectral image are yielded.

## ACKNOWLEDGMENT

The authors gratefully thank Paolo Gamba for providing the ROSIS data over Pavia, Italy, along with the training and test sets, and Vladimir Kolmogorov for the max-flow/min-cut C++ code made available on the web (see [12] for more details).

## REFERENCES

[1] Camps-Valls, G.; Bruzzone, L. : Kernel-based methods for hyperspectral image classification. IEEE TGRS, Vol. 43, Issue 6. (2005) 1351–1362.
[2] Marroquin, J. L.; Santana, E. A.; Botello, S. : Hidden Markov Measure Field Models for Image Segmentation. IEEE TPAMI, Vol. 25, Issue 11. (2003) 1380–1387.
[3] Plaza, A. et. al: Advanced Processing of Hyperspectral Images. IEEE IGARSS Proceedings, Vol. IV (2006) 1974–1979.
[4] Kumar, S. ; Hebert, M. : Discriminative Random Fields. Int. Journal of Computer Vision, Vol. 68, Issue 2. (2006) 179–202.
[5] Borges, J.S.; Bioucas-Dias, J.; Marçal, A.R.S.: Fast Sparse Multinomial Regression Applied to Hyperspectral Data. Image Analysis and Recognition. Lecture Notes in Computer Science, Vol. 4142. Springer Berlin / Heidelberg (2006) 700–709
[6] Boykov, Y.; Veksler, O.; Zabih, R.: Fast Approximate Energy Minimization via Graph Cuts. IEEE TPAMI, Vol. 23 Issue 11. (2001) 1222–1239
[7] Dell'Acqua et. al: Exploiting spectral and spatial information in hyperspectral urban data with high resolution. Geoscience and Remote Sensing Letters, IEEE, Vol. 1, Issue 4. (2004) 322–326
[8] McLachlan, G. J. : Discriminant Analysis and Statistical Pattern Recognition John Wiley & Sons (1992)
[9] Lange, K.: Optimization. New York: Springer Verlag (2004)
[10] Quarteroni, A., Sacco, R. and Saleri, F.: Numerical Mathematics. Springer-Verlag, New-York. (2000) TAM Series n. 37.
[11] Geman, S. and Geman, D.: Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. IEEE TPAMI, Vol. 6. (1984) 721–741
[12] Boykov, Y. and Kolmogorov V.: An experimental comparison of mincut/max-flow algorithms for energy minimization in vision. IEEE TPAMI, Vol. 26, Issue 9 (2004) 1124–1137