Bayesian Active Remote Sensing Image Classification

Pablo Ruiz, *Student Member, IEEE,* Javier Mateos, Gustavo Camps-Valls, *Senior Member, IEEE,* Rafael Molina, *Member, IEEE,* and Aggelos K. Katsaggelos, *Fellow, IEEE*

***Abstract*—In recent years kernel methods and in particular support vector machines (SVMs) have been successfully introduced to remote sensing image classification. Their properties make them appropriate for dealing with high number of image features and low number of available labeled spectra. The introduction of alternative approaches based on (parametric) Bayesian inference has been quite scarce in the more recent years. Assuming a particular prior data distribution may lead to poor results in remote sensing problems because of the specificities and complexity of the data. In this context, the emerging field of non-parametric Bayesian methods constitutes a proper theoretical framework to tackle the remote sensing image classification problem.**

This paper exploits the Bayesian modeling and inference paradigm to tackle the problem of kernel-based remote sensing image classification. This Bayesian methodology is appropriate for both finite and infinite dimensional feature spaces. The particular problem of active learning is addressed by proposing an incremental/active learning approach based on three different approaches: the maximum differential of entropies, the minimum distance to decision boundary, and the minimum normalized distance. Parameters are estimated by using the evidence Bayesian approach, the kernel trick, and the marginal distribution of the observations instead of the posterior distribution of the adaptive parameters. This approach allows us to deal with infinite dimensional feature spaces. The proposed approach is tested on the challenging problem of urban monitoring from multispectral and synthetic aperture radar (SAR) data and in multiclass land cover classification of hyperspectral images, in both purely supervised and active learning settings. Similar results are obtained when compared to SVMs in supervised mode, with the advantage of providing posterior estimates for classification and automatic parameter learning. Comparison with random sampling, and standard active learning methods, such as margin sampling and entropy-query-by-bagging reveal a systematic overall accuracy gain and faster convergence with the number of queries.

***Index Terms*—Supervised classification, incremental/active learning, multispectral image segmentation, Bayesian inference**

P. Ruiz, J. Mateos and R. Molina are with Dpt. Ciencias de la Computación e I. A. E.T.S. Ing. Informática y Telecomunicación. Universidad de Granada, 18071 Granada, Spain. (e-mail: mataran@decsai.ugr.es, jmd@decsai.ugr.es, rms@decsai.ugr.es).

G. Camps-Valls is with the Image Processing Laboratory (IPL), University of Valencia, Parc Cientific Universitat de València, C/ Cat. A. Escardino, 46980 Paterna, València, Spain. (e-mail: gustavo.camps@uv.es).

A. K. Katsaggelos is with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA (e-mail: aggk@eecs.northwestern.edu).

This paper has been partially supported by the Spanish Ministry for Education and Science under projects, TIN2010-15137, EODIX/AYA2008-05965-C04-03 and CONSOLIDER/CSD2007-00018, and a grant from the U.S. Department of Energy (DE-NA0000457).

Manuscript received March, 2012; revised April, 2013.

I. Introduction

CURRENTLY, kernel methods in general and support vector machines (SVMs) in particular dominate the field of *discriminative* data classification models [1]. During the last years, the methods have been successfully introduced in the field of remote sensing image classification [2], [3]. Kernel methods deal efficiently with low-sized datasets of potentially high dimensionality, as in the case of hyperspectral images. The use of the kernel trick [4], as is known in the literature, allows kernel methods to work in higher dimensional (possibly infinite-dimensional) spaces requiring the knowledge of only a kernel function which calculates an inner product in the new space using the original data. Also, since kernel methods do not assume an explicit prior data distribution but are inherently non-parametric models, they cope well with remote sensing data specificities and complexities. Alternative Bayesian approaches to remote sensing processing problems also exist and have been introduced as well to Earth observation applications. For example, the relevance vector machine (RVM) [5] assumes a Gaussian prior over the weights to enforce sparsity and uses expectation-maximization to infer the parameters. In [6], [7], the RVM was used for multispectral image segmentation and landmine detection using ground penetrating radar, while in [8] the model was used for adaptive biophysical parameter retrieval. Lately, Gaussian Processes [9] have received much attention in the field of machine learning, and some applications and developments have been introduced in remote sensing data processing as well, both for classification [10], [11] and parameter retrieval [12] settings.

In this paper, we restrict ourselves to the classification problem. Due to the particular characteristics of remote sensing data, namely potentially high-dimensionality, low number of labeled samples and different noise sources, assuming a particular prior data distribution may lead to poor classification results. In this context, the emerging field of *non-parametric Bayesian methods* constitutes a proper theoretical framework to tackle the problem [13], [9], [14][1]. This paper follows a Bayesian modeling and inference paradigm to tackle the problem of *kernel-based* remote sensing image classification. This Bayesian methodology is appropriate for both finite and infinite dimensional feature spaces, and hence robustness to the aforementioned problems in remote sensing is achieved. In two-class classification problems, the goal is to estimate a function and use as decision boundary the points where

[1] Excellent online lectures are available at: http://videolectures.net/mlss09uk_teh_nbm/ and http://videolectures.net/mlss09uk_orbanz_fnbm/

the function is zero, to decide whether a sample belongs to a given class. In its simplest form, and given a training set, this is equivalent to estimate a linear function on a transformed feature space to separate samples from both classes. SVMs approach this problem through the concept of margin which is defined as the smallest distance between the decision boundary and any of the samples. On the other hand, Bayesian modeling and inference approach the problem by introducing information on the hyperplane coefficients using a prior model which in combination with the likelihood of the labeled samples leads to both a posterior distribution of the hyperplane coefficients and a Bayesian classification procedure. The use of the Bayesian paradigm allows for the calculation of the uncertainty of the estimated parameters and also the determination of the certainty of the estimated label for a given sample. It also allows for the estimation of all the model parameters in a rigorous and sound manner.

Relations between SVMs and Bayesian inference is not new. Note that adopting the least-squares SVM formulation may alternatively allow to perform Bayesian inference on SVMs [15]. Bayesian inference on these machines yields some relevant benefits: hyperparameters may be learned directly from data using a consistent theoretical framework, and posterior probabilities for the predictions can be obtained. Consequently, non-parametric Bayesian methods may deal with uncertainties in the data and naturally allow us developing intuitive incremental/active learning methods. The presented Bayesian kernel-based classifier permits to derive efficient closed-form expressions for parameter estimation, as well as to perform incremental, adaptive and active learning in a consistent, principled way.

While kernel-based classification in *static* scenarios has been extensively studied, the problem of *on-line* and *incremental* classification is still unsolved. The most effective schemes so far make use of both incremental and online SVMs [16], [17], [18]. Most of these approaches are based on growing and pruning strategies to create and update a *dictionary* of (representative) support vectors. Unfortunately, the algorithms require tuning several heuristic parameters. Alternatively, Bayesian kernel machines, such as Gaussian processes, have been successfully reformulated to deal with online and sparse settings [19], [20]. These methods typically rely again on a sequential generation of datasets of relevant samples. Nevertheless the framework nicely allows for both a propagation of predictions and Bayesian error estimates.

The previous online/incremental approaches are actually related to the emerging field of *active* learning [21]. Active learning aims at building efficient training sets by *iteratively* improving the model performance through sampling. Many query strategies have been devised in the literature, which are based on different heuristics: 1) large margin, 2) expert committee, and 3) posterior probability (see [21] for a comprehensive review). The first approach typically exploits SVM methods, while the second one can be adopted by any classifier. The latter requires classifiers that can provide posterior probabilities. While Platt's solution [22] of including a sigmoid link in SVMs could do the job, some theoretical concerns have been raised about the true meaning of such posteriors. In Bayesian active learning, the prior over the hypotheses space is updated after seeing new data. For example, in [23], the expected Kullback-Leibler divergence between the current and the revised posterior distributions is maximized, while in [24], the authors proposed a Bayesian framework to tackle the active learning problem, which is utilized in Remote sensing in [25]. In [26], a Bayesian framework is also used and the posterior distribution is obtained as a Multinomial Logistic Regression model. Other basis selection techniques make explicit use of the response functions [27], [28], [29]. See also [30] for the basis selection general theory and [31] for the use of the approach in compressive sensing.

The field of remote sensing image classification has experienced a growing interest in active learning. Most of the introduced methods rely on smart sampling strategies over the SVM margin [32], [33], [34]. Some alternative approaches to work with batches of selections per iteration have been presented, and mainly rely on the concept of *diversity* between candidate pixels [35], [33] or with respect to the current model [36], or both [37]. Recent papers deal with new applications of active learning algorithms: in [38], [39], active learning is used to select the most useful unlabeled pixels to train a *semisupervised* classifier, while in [11], [40] active queries are used to correct for dataset shift in different areas of images. A complete review of the field of active learning in remote sensing can be found in [41].

In this paper, the Bayesian modeling and inference paradigm is applied to kernel-based classifiers. This paradigm is used to tackle both passive and active learning, as well as to address the problem of parameter estimation for infinite dimensional feature spaces, and consequently for problems where basis selection cannot be carried out explicitly. The current work presents the novel introduction of nonparametric Bayesian learning for remote sensing image classification both in purely supervised and active learning settings. This approach proposes an iterative procedure to maximize the marginal of the observations and, to the best of our knowledge, this is the first paper where nonparametric Bayesian methods are used in Active Remote Sensing Images Classification. The presented methods actually go one step further by extending standard nonparametric large margin techniques, such as SVM, which are typically used for image segmentation applications. Nonparametric Bayesian modeling and inference paradigms are introduced here to tackle the problem of kernel-based remote sensing image classification with the resulting major advantage of automatically learning the values of the (hyper)parameters from the data and thus no ad hoc cross-validation tuning schemes are necessary. This Bayesian methodology is appropriate for both finite and infinite dimensional feature spaces. The particular problem of active learning is addressed by proposing an incremental/active learning approach based on three different approaches: the maximum differential of entropies, the minimum distance to decision boundary, and the minimum normalized distance. Comparison with random sampling and standard active learning methods, such as margin sampling, or entropy-query-by-bagging, reveals a systematic overall accuracy gain and faster convergence with the number of queries.

The remainder of the paper is organized as follows. Section II introduces the basic notation to perform Bayesian modeling. Section III presents the Bayesian inference framework proposed in this paper. We first introduce the basic tools and then the novel formulations for parameter estimation, active learning data classification and prediction. Section V illustrates the performance of the proposed method in multispectral image segmentation. Conclusions are outlined in Section VI.

II. Problem statement and Bayesian modeling

Let us introduce the basic problem formulation and notation. Let n be the number of pixels of a d-dimensional hyperspectral image, $\{\mathbf{x}_i | i = 1, \ldots, n\}$, $\mathbf{x} \in \mathbb{R}^d$ we want to classify. The general two-class supervised classification problem we tackle here defines a classification function of the form

$$y(\mathbf{x}) = \boldsymbol{\phi}^\top(\mathbf{x})\mathbf{w} + b + \epsilon, \tag{1}$$

where the mapping $\boldsymbol{\phi} : \mathcal{X} \to \mathcal{H}$ maps the observed data point (samples, spectra) $\mathbf{x} \in \mathcal{X}$ into a higher L-dimensional (possibly infinite) Hilbert feature space $\mathcal{H}$. Note that for a K-class problem, the decision function implies K independent classification functions of the form $y_k(\mathbf{x}) = \boldsymbol{\phi}^\top(\mathbf{x})\mathbf{w}_k + b_k + \epsilon_k$, $k = 1, \ldots, K$ [4].

For the sake of simplicity of the notation, we will focus here on the binary case. However, its extension to multiclass scenarios is straightforward[2]. Therefore, for a data point, $\mathbf{x}$, the output $y(\mathbf{x}) \in \{0, 1\}$ consists of a binary coding representation of its classification as belonging to class $\mathcal{C}_0$ or $\mathcal{C}_1$, respectively, $\mathbf{w}$ is a vector of size $L \times 1$ of adaptive parameters to be estimated, b represents the bias in the classification function, and ϵ is an independent realization of the Gaussian distribution $\mathcal{N}(0, \sigma^2)$.

For a training set, we already know the classification output $y(\mathbf{x}_i)$ associated with the feature samples $\boldsymbol{\phi}(\mathbf{x}_i), i = 1, \ldots, M$, with M the number of samples, and therefore we can write

$$\mathrm{p}(\mathbf{y}|\mathbf{w}, b, \sigma^2) = \prod_{i=1}^{M} \mathcal{N}(y(\mathbf{x}_i)|\boldsymbol{\phi}^\top(\mathbf{x}_i)\mathbf{w} + b, \sigma^2), \tag{2}$$

where $\mathbf{y} = (y(\mathbf{x}_1), y(\mathbf{x}_2), \ldots, y(\mathbf{x}_M))^\top$. Since $\mathbf{x}_i$, $i = 1, \ldots, M$, will always appear as conditioning variable, for the sake of simplicity, we have removed the dependency on $\mathbf{x}_1, \ldots, \mathbf{x}_M$ in the left-hand side of the equation. We note that, for infinite dimensional feature vectors $\phi(\mathbf{x}_i)$, $\mathbf{w}$ is infinite dimensional.

The Bayesian framework allows us to introduce information about the possible value of $\mathbf{w}$ in the form of a prior distribution. Since the likelihood function defined in Eq. (2) is the exponential of a quadratic function of $\mathbf{w}$, its corresponding conjugate prior should be a Gaussian distribution [4] so that the posterior will also be Gaussian. In this work, we consider a particular form of the Gaussian prior in which each component of $\mathbf{w}$ independently follows a Gaussian distribution $\mathcal{N}(0, \gamma^2)$. Notice that this distribution can also be obtained utilizing the Gaussian Process framework [4]. When the feature vectors are infinite dimensional, we will not make explicit use of this prior distribution but still we will be able to carry out parameter estimation, prediction, and active learning tasks.

III. Proposed Bayesian inference method

Due to the possible use of infinite dimensional feature spaces we will mainly use the marginal distribution of the observations to perform inference tasks, that is, parameter estimation, prediction and active learning and avoid, when possible, the use of the posterior distribution of the adaptive parameters, $\mathbf{w}$, since it cannot be calculated for infinite dimensional spaces. However, when a finite dimensional space is used, we will also calculate the posterior distribution in this section.

A. Marginal Distribution of $\mathbf{y}$

The marginal distribution of $\mathbf{y}$ can be obtained by integrating out the vector of adaptive parameters $\mathbf{w}$. It can easily be shown, see for instance [4], that

$$\mathrm{p}(\mathbf{y}|b, \gamma^2, \sigma^2) = \mathcal{N}(\mathbf{y}|b\mathbf{1}, \mathbf{C}), \tag{3}$$

with

$$\mathbf{C} = \gamma^2 \boldsymbol{\Phi}\boldsymbol{\Phi}^\top + \sigma^2 \mathbf{I}, \tag{4}$$

where $\boldsymbol{\Phi}$ is the design matrix whose i-th row is $\boldsymbol{\phi}^\top(\mathbf{x}_i)$, and $\mathbf{1}$ is a column vector with all its M components equal to 1.

It is important to note that we do not need to know the form of $\boldsymbol{\Phi}$ explicitly to calculate this marginal distribution. We only need to know the Gram matrix $\mathbf{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^\top$, which is an $M \times M$ symmetric matrix with elements $\mathbf{K}_{nm} = k(\mathbf{x}_n, \mathbf{x}_m) = \phi^\top(\mathbf{x}_n)\phi(\mathbf{x}_m)$. It has to be a positive semidefinite matrix (see [1]), i.e., we only need to know the kernel function $k(\cdot, \cdot)$ that represents the inner product in the new feature space to calculate the marginal distribution. This leads to the construction of kernel functions $k(\mathbf{x}, \mathbf{x}')$ for which the Gram matrix $\mathbf{K}$ is positive semidefinite for all possible choices of the set $\{\mathbf{x}_n\}$. Note that, even if $\boldsymbol{\Phi}$ has an infinite number of columns, which corresponds to the case of $\phi(\mathbf{x}_i)$ being an infinite dimensional feature vector, we can still calculate $\mathbf{K}$ of size $M \times M$ by means of the kernel function. Consequently, the new feature space dimension depends of the selected kernel function.

It is also worth noting that the above marginal distribution can be obtained by assuming that $\mathbf{y}$ consists of independent additive noisy observations, with variance γ^2, of a Gaussian process with mean b and covariance $\mathbf{K}$.

For a new sample $\mathbf{x}_*$ the distribution of

$$\mathbf{y}_{M+1} = \begin{pmatrix} \mathbf{y} \\ y(\mathbf{x}_*) \end{pmatrix}, \tag{5}$$

has the form

$$\mathrm{p}(\mathbf{y}_{M+1}|b, \gamma^2, \sigma^2) = \mathcal{N}(\mathbf{y}_{M+1}|b\mathbf{1}_{M+1}, \mathbf{C}_{M+1}), \tag{6}$$

[2] Extension to multiclass problems can be accomplished in many different ways by following standard schemes: one-versus-all, one-versus-one, pure multiclass schemes, or even sophisticated puncturing alternatives. We suggest here the use of a one-versus-all scheme, which typically gives rise to simpler and highly competitive results [42].

with $\mathbf{C}_{M+1} = \gamma^2\mathbf{\Phi}_{M+1}\mathbf{\Phi}_{M+1}^\top + \sigma^2\mathbf{I}_{M+1}$, which can be written as

$$\mathbf{C}_{M+1} = \begin{pmatrix} \mathbf{C} & \mathbf{k} \\ \mathbf{k}^\top & c \end{pmatrix}, \tag{7}$$

where $\mathbf{C}$ has been defined in Eq. (4) and

$$\begin{aligned} \mathbf{k}^\top &= \gamma^2\boldsymbol{\phi}^\top(\mathbf{x}_*)\mathbf{\Phi}^\top, && (8) \\ c &= \gamma^2\boldsymbol{\phi}^\top(\mathbf{x}_*)\boldsymbol{\phi}(\mathbf{x}_*) + \sigma^2. && (9) \end{aligned}$$

Furthermore, the conditional distribution $\mathrm{p}(y(\mathbf{x}_*)|\mathbf{y})$ is a Gaussian distribution with mean $m(\mathbf{x}_*)$ and variance $v(\mathbf{x}_*)$ given by

$$\begin{aligned} m(\mathbf{x}_*) &= b + \gamma^2\boldsymbol{\phi}^\top(\mathbf{x}_*)\mathbf{\Phi}^\top\mathbf{C}^{-1}(\mathbf{y} - b\mathbf{1}), && (10) \\ v(\mathbf{x}_*) &= c - \mathbf{k}^\top\mathbf{C}^{-1}\mathbf{k}. && (11) \end{aligned}$$

B. Posterior Distribution of $\mathbf{w}$

When the feature space is finite dimensional we can also calculate the posterior distribution of $\mathbf{w}$, which is given by (see [4]),

$$\begin{aligned} &\mathrm{p}(\mathbf{w}|\mathbf{y}, b, \gamma^2, \sigma^2) = \\ &\qquad \mathcal{N}(\mathbf{w}|\mathbf{\Sigma}_{\mathbf{w}|\mathbf{y},\gamma^2,\sigma^2}\sigma^{-2}\mathbf{\Phi}^\top(\mathbf{y} - b\mathbf{1}), \mathbf{\Sigma}_{\mathbf{w}|\mathbf{y},\gamma^2,\sigma^2}), \end{aligned} \tag{12}$$

where

$$\mathbf{\Sigma}_{\mathbf{w}|\mathbf{y},\gamma^2,\sigma^2} = (\sigma^{-2}\mathbf{\Phi}^\top\mathbf{\Phi} + \gamma^{-2}\mathbf{I})^{-1}.$$

Notice that $m(\mathbf{x}_*)$ defined in Eq. (10) can be expressed in terms of $\mathbb{E}[\mathbf{w}]$ as

$$m(\mathbf{x}_*) = \phi(\mathbf{x}_*)^\top \mathrm{E}\left[\mathbf{w}\right] + b. \tag{13}$$

C. Parameter Estimation

The last step in the Bayesian inference we are carrying out is the estimation of the parameters involved in the models, that is, the estimation of the values of γ^2, σ^2, and b. The value of b can be easily obtained from Eq. (3) as

$$b = \frac{1}{M}\sum_{i=1}^{M} y(\mathbf{x}_i). \tag{14}$$

To estimate the values of γ^2 and σ^2 we use the Evidence Bayesian approach without any prior information on these parameters. The Evidence Bayesian approach [43], see [44], [45] for other possible names, determines the values of the parameters γ^2 and σ^2 by maximizing the marginal distribution in Eq. (3) obtained by integrating out the vector of adaptive parameters $\mathbf{w}$. Intuitively, by integrating over $\mathbf{w}$ we are searching for the best value of γ^2 and σ^2 for all possible values of $\mathbf{w}$. Differentiating $2\ln\mathrm{p}(\mathbf{y}|b, \gamma^2, \sigma^2)$ with respect to γ^2 and equating the result to zero, we obtain

$$\mathbf{tr}[\mathbf{C}^{-1}\mathbf{\Phi}\mathbf{\Phi}^\top] = \mathbf{tr}[(\mathbf{y} - b\mathbf{1})^\top\mathbf{C}^{-1}\mathbf{\Phi}\mathbf{\Phi}^\top\mathbf{C}^{-1}(\mathbf{y} - b\mathbf{1})]. \tag{15}$$

Diagonalizing $\mathbf{\Phi}\mathbf{\Phi}^\top$, we obtain $\mathbf{U}\mathbf{\Phi}\mathbf{\Phi}^\top\mathbf{U}^\top = \mathbf{D}$, where $\mathbf{U}$ is an orthonormal matrix and $\mathbf{D}$ is a diagonal matrix with entries $\lambda_i, i = 1, \ldots, M$. We can then rewrite the above equation as

$$\sum_{k=1}^{M}\frac{\lambda_k}{\gamma^2\lambda_k + \sigma^2} = \sum_{i=1}^{M} z_i^2\frac{\lambda_i}{(\gamma^2\lambda_i + \sigma^2)^2}, \tag{16}$$

where $\mathbf{U}(\mathbf{y} - b\mathbf{1}) = \mathbf{z}$ with components $z_i, i = 1, \ldots, M$.

Multiplying both sides of the above equation by γ^2 we have

$$\gamma^2 = \sum_{i=1}^{M}\frac{\frac{\lambda_i}{\gamma^2\lambda_i+\sigma^2}}{\sum_{k=1}^{M}\frac{\lambda_k}{\gamma^2\lambda_k+\sigma^2}}\frac{\gamma^2 z_i^2}{\gamma^2\lambda_i + \sigma^2} = \sum_{i=1}^{M}\mu_i\frac{\gamma^2 z_i^2}{\gamma^2\lambda_i + \sigma^2}, \tag{17}$$

where

$$\mu_i = \frac{\frac{\lambda_i}{\gamma^2\lambda_i+\sigma^2}}{\sum_{k=1}^{M}\frac{\lambda_k}{\gamma^2\lambda_k+\sigma^2}}. \tag{18}$$

Note that $\mu_i \geq 0$ and $\sum_{i=1}^{M}\mu_i = 1$.

Similarly, differentiating $2\ln\mathrm{p}(\mathbf{y}|\gamma^2, \sigma^2)$ with respect to σ^2 and equating the result to zero, we obtain

$$\sum_{k=1}^{M}\frac{1}{\gamma^2\lambda_k + \sigma^2} = \sum_{i=1}^{M} z_i^2\frac{1}{(\gamma^2\lambda_i + \sigma^2)^2}. \tag{19}$$

Following the same steps we already performed to estimate γ^2, we obtain

$$\sigma^2 = \sum_{i=1}^{M}\nu_i\frac{\sigma^2 z_i^2}{\gamma^2\lambda_i + \sigma^2}, \tag{20}$$

where

$$\nu_i = \frac{\frac{1}{\gamma^2\lambda_i+\sigma^2}}{\sum_{k=1}^{M}\frac{1}{\gamma^2\lambda_k+\sigma^2}}. \tag{21}$$

Note that, again, $\nu_i \geq 0$ and $\sum_{i=1}^{M}\nu_i = 1$.

Equations (17) and (20) suggest the iterative procedure described in Alg. 1 to estimate the parameters where the old value of the parameters is used in the right hand side of the equations to obtain a new estimate of the parameters in the left hand side of the equations.

Algorithm 1 Parameter estimation

```
Using Eq. (14), compute $b = \frac{1}{M}\sum_{i=1}^{M} y(\mathbf{x}_i)$.
Compute $\mathbf{U}$ and $\lambda_i, i = 1, \ldots, M$, as the eigenvector matrix and eigenvalues of $\mathbf{\Phi}\mathbf{\Phi}^\top$, respectively.
Set $\mathbf{z} = \mathbf{U}(\mathbf{y} - b\mathbf{1})$.
Initialize $\gamma^2 = 1, \sigma^2 = 1$.
repeat
    Set $\gamma_{old}^2 = \gamma^2, \sigma_{old}^2 = \sigma^2$.
    Set $\gamma^2 = \sum_{i=1}^{M}\mu_i\,\gamma_{old}^2\,z_i^2/(\gamma_{old}^2\,\lambda_i + \sigma_{old}^2)$.
    Set $\sigma^2 = \sum_{i=1}^{M}\nu_i\,\sigma_{old}^2\,z_i^2/(\gamma_{old}^2\,\lambda_i + \sigma_{old}^2)$.
until $(\gamma^2 - \gamma_{old}^2)^2/(\gamma_{old}^2)^2 < 10^{-6}$ and $(\sigma^2 - \sigma_{old}^2)^2/(\sigma_{old}^2)^2 < 10^{-6}$.
```

D. Classification

Once the system has been trained, we want to assign a class to a new value of $\mathbf{x}$, denoted by $\mathbf{x}_*$. We already know that the conditional distribution $\mathrm{p}(y(\mathbf{x}_*)|\mathbf{y})$ is a Gaussian distribution with mean $m(\mathbf{x}_*)$ and variance $v(\mathbf{x}_*)$ given in Eqs. (10) and (11). We classify $\mathbf{x}_*$ utilizing $m(\mathbf{x}_*)$ and write

$$\mathbf{x}_* \text{ is assigned to } \begin{cases} \mathcal{C}_1 & \text{if } m(\mathbf{x}_*) \geq 0.5 \\ \mathcal{C}_0 & \text{if } m(\mathbf{x}_*) < 0.5 \end{cases}. \tag{22}$$

Notice that the classification of $\mathbf{x}_*$ is based on the proximity of the mean value of $\mathrm{p}(y(\mathbf{x}_*)|\mathbf{y})$ to the value zero or one that represents the classes $\mathcal{C}_0$ and $\mathcal{C}_1$, respectively.

IV. Proposed active learning method

Active learning starts with a small set of observations whose class is already known. From these observations, the marginal distribution of $\mathbf{y}$, the conditional distribution of $\mathbf{w}$ given $\mathbf{y}$, and the parameters b, γ^2, and σ^2 are estimated using the procedure described in the previous sections. In order to improve the performance of the classifier we want to select a new training sample $\mathbf{x}_+$, whose corresponding $y(\mathbf{x}_+)$ will be learned by querying the oracle. Let us now examine different ways to select the new training sample.

A. Method 1: Maximum differential of entropies

Utilizing Eq. (10) and (11) we observe that, for a sample $\mathbf{x}$ not already present in the training set, the distribution of $y(\mathbf{x})$ given the set of observations $\mathbf{y}$ has variance

$$v(\mathbf{x}) = \gamma^2 \boldsymbol{\phi}^\top(\mathbf{x})\boldsymbol{\phi}(\mathbf{x}) + \sigma^2 - \gamma^4 \boldsymbol{\phi}^\top(\mathbf{x})\boldsymbol{\Phi}^\top \mathbf{C}^{-1}\boldsymbol{\Phi}\boldsymbol{\phi}(\mathbf{x}), \quad (23)$$

and consequently we can select the new training sample as the one maximizing the variance of the prediction, that is,

$$\mathbf{x}_+ = \arg\max_{\mathbf{x}} v(\mathbf{x}). \quad (24)$$

Notice that using this criterion amounts to selecting the sample the classifier is less certain about the class it belongs to.

Let us relate this active method procedure to the one proposed in [24], [31] for finite dimensional feature spaces. The covariance matrix of the posterior distribution of $\mathbf{w}$ when a new $\mathbf{x}$ is added to the training set is given by

$$\boldsymbol{\Sigma}^{\mathbf{x}}_{\mathbf{w}|\mathbf{y},\gamma^2\sigma^2} = (\sigma^{-2}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}^\top(\mathbf{x})) + \gamma^{-2}\mathbf{I})^{-1}. \quad (25)$$

For finite dimensional feature spaces it is proposed in [24], [31] to add to the training set the sample with maximum difference between the entropies of the posterior distribution before and after adding the new sample, that is,

$$\mathbf{x}_+ = \arg\max_{\mathbf{x}} \frac{1}{2}\log|\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y},\gamma^2,\sigma^2}||\boldsymbol{\Sigma}^{\mathbf{x}}_{\mathbf{w}|\mathbf{y},\gamma^2,\sigma^2}|^{-1}. \quad (26)$$

Let us first express this criterion in terms of the marginal distribution of the observations in order to remove the need of using finite dimensional feature spaces. We note that

$$\begin{aligned}\log&|\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y},\gamma^2,\sigma^2}||\boldsymbol{\Sigma}^{\mathbf{x}}_{\mathbf{w}|\mathbf{y},\gamma^2,\sigma^2}|^{-1}\\ &= \log|\mathbf{I} + \sigma^{-2}\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}^\top(\mathbf{x})(\sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \gamma^{-2}\mathbf{I})^{-1}|\\ &= \log(1 + \sigma^{-2}\boldsymbol{\phi}^\top(\mathbf{x})(\sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \gamma^{-2}\mathbf{I})^{-1}\boldsymbol{\phi}(\mathbf{x})), \quad (27)\end{aligned}$$

and using

$$(\sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \gamma^{-2}\mathbf{I})^{-1} = \gamma^2\mathbf{I} - \gamma^4\boldsymbol{\Phi}^\top\mathbf{C}^{-1}\boldsymbol{\Phi}, \quad (28)$$

we can write Eq. (27) in terms of the marginal distribution of the observations as

$$\begin{aligned}&\log|\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y},\gamma^2,\sigma^2}||\boldsymbol{\Sigma}^{\mathbf{x}}_{\mathbf{w}|\mathbf{y},\gamma^2,\sigma^2}|^{-1} =\\ &= \log(1 + \sigma^{-2}\gamma^2\boldsymbol{\phi}^\top(\mathbf{x})\boldsymbol{\phi}(\mathbf{x}) - \sigma^{-2}\gamma^4\boldsymbol{\phi}^\top(\mathbf{x})\boldsymbol{\Phi}^\top\mathbf{C}^{-1}\boldsymbol{\Phi}\boldsymbol{\phi}(\mathbf{x}))\\ &= \log(1 + \sigma^{-2}\gamma^2\boldsymbol{\phi}^\top(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})\\ &\quad - \sigma^{-2}\gamma^4\boldsymbol{\phi}^\top(\mathbf{x})\boldsymbol{\Phi}^\top\boldsymbol{\Sigma}^{-1}_{\mathbf{y}|\gamma^2,\sigma^2}\boldsymbol{\Phi}\boldsymbol{\phi}(\mathbf{x})). \quad (29)\end{aligned}$$

Consequently, all needed quantities to select $\mathbf{x}_+$ can be calculated without knowledge of the feature vectors and the posterior distribution of the possibly infinite dimensional adaptive parameters and using only kernel functions and the marginal distribution of the observations.

Furthermore we have

$$\log|\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y},\gamma^2,\sigma^2}||\boldsymbol{\Sigma}^{\mathbf{x}}_{\mathbf{w}|\mathbf{y},\gamma^2,\sigma^2}|^{-1} = \log(\sigma^{-2}v(\mathbf{x})), \quad (30)$$

and consequently both criteria coincide. Notice that, as we have already mentioned, we have also shown that the maximum differential of entropies criterion can be utilized over infinite dimensional feature spaces.

B. Method 2: Minimum distance to decision boundary

In our classification problem the decision boundary corresponds to the set

$$\boldsymbol{\Pi} = \left\{\mathbf{x} \in \mathcal{X} : \boldsymbol{\phi}^\top(\mathbf{x})\mathbb{E}[\mathbf{w}] + b - 0.5 = 0\right\}. \quad (31)$$

We can then select the next sample to be included in the training set by using

$$\begin{aligned}\mathbf{x}_+ &= \arg\min_{\mathbf{x}} \mathrm{d}^2(\mathbf{x}, \boldsymbol{\Pi})\\ &= \arg\min_{\mathbf{x}} \frac{(\boldsymbol{\phi}^\top(\mathbf{x})\mathrm{E}[\mathbf{w}] + b - 0.5)^2}{\|\mathrm{E}[\mathbf{w}]\|^2}\\ &= \arg\min_{\mathbf{x}} (m(\mathbf{x}) - 0.5)^2. \quad (32)\end{aligned}$$

Note that this method provides a Bayesian formulation of the SVM margin sampling heuristic (see [41]).

C. Method 3: Minimum Normalized Distance

The two active learning methods described above take into consideration only partial aspects of the conditional distribution $\mathrm{p}(y(\mathbf{x}_*)|\mathbf{y})$. While maximum differential of entropies utilizes the variance of this distribution, it does not use the distance to the decision boundary. On the other hand, the minimum distance to the decision boundary criterion is based on the mean of this conditional distribution and does not take into account the uncertainty of the distribution. It is obviously very easy to imagine scenarios where these two criteria will not select the best sample, either because it is too far from the decision boundary and, hence, having large variance does not represent a problem, or because, although the sample is the closest to the decision boundary, its uncertainty is very small and consequently it may not be the best sample to be included in the training set.

We can then use the following active learning procedure which combines precision and proximity to the decision boundary

$$\mathbf{x}_+ = \arg\min_{\mathbf{x}} \mathbb{E}\left[\frac{(y(\mathbf{x}) - 0.5)^2}{v(\mathbf{x})}\right], \quad (33)$$

where the expected value is calculated utilizing the conditional distribution $\mathrm{p}(y(\mathbf{x})|\mathbf{y})$ defined in Eqs. (10) and (11).

Notice that since

$$\mathbb{E}\left[\frac{(y(\mathbf{x}) - 0.5)^2}{v(\mathbf{x})}\right] = 1 + \frac{(m(\mathbf{x}) - 0.5)^2}{v(\mathbf{x})}, \quad (34)$$

we can rewrite this criterion as

$$\mathbf{x}_+ = \arg\min_{\mathbf{x}} \frac{(m(\mathbf{x})-0.5)^2}{v(\mathbf{x})} \tag{35}$$

D. Multiclass Extension of the Active Learning Methods

Here we extend the proposed active learning methods to deal with K-class problems. Recently, arquitectures for multiclass active learning have been proposed. For instance, in [33] authors propose the MCLU technique which selects the most uncertain samples according to a confidence score based on the distances to all separation hyperplanes. Note, however, that this approach is specific to maximum margin algorithms like SVM, which is not our case. In this paper, nevertheless, we will use the classical one-versus-all strategy for tackling multiclass problems. Hence, for each candidate $\mathbf{x}$, K different pair of values $\{m_k(\mathbf{x}), v_k(\mathbf{x})\}_{k=1,\dots,K}$ are obtained. These values are used in Eqs. (24), (32) or (35), depending on the selected method, that is finally optimized with respect to $\mathbf{x}$ and k.

V. Experimental Results

In this section, the proposed method is applied to both purely supervised and active remote sensing image classification settings. The method is compared to the standard SVM algorithm in the case of supervised classification when few labeled samples are available. This problem is typically encountered in remote sensing image classification, in which active learning can improve performance. Comparison to random sampling and standard active learning methods, such as margin sampling and entropy-query-by-bagging is then performed. In all cases we provide the overall accuracy, the estimated Cohen's kappa statistic and Z-score [3] as measures of accuracy and class agreement, respectively. All experiments were implemented using Matlab© and run on an Intel© i7@2.67GHz. The Matlab© source code of the proposed method is available at http://decsai.ugr.es/vip/resources/BAL.html for the interested reader. Additionally, a video demonstration of the method is available at the same location.

A. Study area and data collection

Two multispectral images are used in our experiments for supervised and active learning classification:

- *Supervised classification with Landsat imagery.* The image was acquired in the context of the Urban Expansion Monitoring project [46] over the city of Rome (Italy) by the Landsat TM sensor in 1999. An external Digital Elevation Model (DEM) and a reference land cover map provided by the Italian Institute of Statistics (ISTAT) were also available. The available features were the seven Landsat bands, two SAR backscattering intensities (0–35 days), and the SAR interferometric coherence.

  Since image features come from different sensors, the first step was to perform a specific processing and conditioning of optical and SAR data, and to co-register all images [46]. In particular, the seven bands of Landsat TM were co-registered with the ISTAT classification data, and resampled to 30×30 m with the Nearest-Neighbor algorithm. The registration for the multi-source images was performed at the sub-pixel level obtaining a root-mean-squared error of about 10 m, which potentially enables good urban classification ability. We also appended two SAR features: the estimated coherence, *Co*, and a spatially filtered version of the coherence, *FCo*, which is specially designed to increase the urban areas discrimination [46]. After this preprocessing, all features were stacked at the pixel level, and each feature was standardized. The goal is the discrimination of urban ($\mathcal{C}_1$) versus non-urban ($\mathcal{C}_0$) land-cover classes.
- *Active classification with ROSIS imagery.* The second image was acquired by the DAIS7915 sensor over the city of Pavia (Italy), and constitutes a challenging 9-class urban classification problem dominated by directional features and relatively high spatial resolution (5 meters pixels). We took into account only 40 spectral bands of reflective energy in the range $[0.5, 1.76]$ μm, thus skipping thermal infrared bands and middle infrared bands above 1958 nm. We carried out a Principal Components Analysis (PCA) to reduce the dimensionality of the problem and considered the 10 first components for each pixel that have provided good classification performance in previous works (see, for instance, [47]).

B. Supervised Classification Results

For the case of supervised classification, we report results both on the binary classification problem of the Rome scene and the multiclass classification problem of the Pavia scene and compare the performance of our approach to the standard SVM approach.

From the Rome image, of size 1440×930 pixels, a training set of 500 randomly selected pixels was obtained, and results are given in a representative test set of 10000 samples. To obtain unbiased conclusions from the results, the process was repeated 10 times with different randomly selected training and test sets, and the average accuracies are given. In all cases, a Gaussian kernel was used. Using 3-fold cross-validation with the SVM as classifier, a kernel lengthscale $\sigma = 100$ was selected. Although we could have used Bayesian inference to estimate the kernel parameter (see, for instance, [4]) we decided to use the same kernel parameter on both methods and concentrate on the remaining model parameters. Notice that this decision slightly favors SVM since the kernel parameter is estimated seeking the best SVM performance. For the case of SVMs, the regularization parameter C was tuned by 3-fold cross-validation on the training dataset. Our method does not need any heuristic tuning since hyperparameters are estimated automatically in the training phase. The proposed method needed 0.33 seconds to complete the training while the SVM needed 1.94 seconds.

Table I shows the obtained results in the 10 independent realizations and their average and variance. Although SVM obtains better results in many cases, the differences are not

[3] Z-score is defined as the ratio between the estimated kappa statistic and its standard deviation.

TABLE I
CLASSIFICATION ACCURACY FOR SVM AND THE PROPOSED METHOD IN THE ROME (1999) SCENE. OVERALL ACCURACY, ESTIMATED COHEN'S STATISTIC AND Z-SCORE RESULTS ARE GIVEN FOR ALL 10 REALIZATIONS AND AVERAGED.

	Overall accuracy, OA[%]		**Kappa statistic, κ**		**Z-score, Z**	
Realization	Proposed	SVM	Proposed	SVM	Proposed	SVM
1	96.66	**96.95**	0.895	**0.905**	158.04	**169.50**
2	96.48	**96.61**	0.890	**0.896**	154.57	**161.07**
3	**97.27**	96.96	**0.914**	0.905	**177.60**	168.35
4	96.24	**96.54**	0.883	**0.894**	150.20	**160.00**
5	**97.10**	96.54	**0.909**	0.892	**172.39**	157.37
6	**96.64**	95.99	**0.893**	0.874	**156.69**	142.59
7	**96.86**	96.58	**0.905**	0.898	**170.81**	165.14
8	**96.76**	96.72	0.895	**0.896**	156.32	**158.44**
9	**97.02**	96.71	**0.901**	0.900	**174.92**	167.36
10	96.99	**97.00**	0.906	**0.908**	170.54	**173.04**
Average	**96.80**	96.66	**0.90**	**0.90**	**164.21**	162.29
Variance	0.0957	0.0867	$< 10^{-5}$	$< 10^{-5}$	98.95	74.66

TABLE II
MEAN CONFUSION MATRIX FOR SVM AND THE PROPOSED METHOD (IN BRACKETS). WE SHOW THE AVERAGE KAPPA STATISTIC, ALONG WITH ITS VARIANCE, Z-SCORE AND CONFIDENCE INTERVALS FOR BOTH METHODS.

	$\mathcal{C}_0$	$\mathcal{C}_1$
$\hat{\mathcal{C}}_0$	7802.80 (7846.30)	169.00 (198.30)
$\hat{\mathcal{C}}_1$	165.00 (121.50)	1863.20 (1833.90)

	SVM	Proposed
OA [%]	96.66%	96.80%
κ	0.90	0.90
σ^2_κ	3.07e-05	3.02e-05
Z-score	162.29	164.21
κ CI	[0.886,0.908]	[0.889,0.910]

statistically significant, as assessed by the average values of the three measures. Table II shows the *average* confusion matrices for the 10 realizations, along with its variance, Z-score and confidence intervals for both methods. These results also confirm the numerical and statistical similarity of the results. Finally, Fig. 1 shows the classification maps obtained by SVM and the proposed method in a particular realization. Visual results match the previous numerical accuracies as no difference is obtained. The statistical significance of the kappa statistic also confirms this issue.

A second experiment was performed on the 9-class urban classification problem of the Pavia scene depicted in Fig. 2a, which has 400×400 pixels. Training was done on 1260 randomly selected pixels (140 from each class), and a test set of 13314 representative samples was used. Again, ten different realizations were used to obtain unbiased conclusions from the results. We used a Gaussian kernel, whose lengthscale $\sigma = 500$ was selected using 3-fold cross-validation with the SVM as classifier. As in the previous experiment, the regularization parameter C for the SVM was tuned by 3-fold cross-validation on the training dataset while the proposed method estimated all hyperparameters automatically in the training phase. The proposed method needed 14.32 seconds to complete the training while the SVM needed 9.09 seconds. This is explained by the fact that SVM estimates a single value of C for all classifiers while the proposed method has to estimate the value of the hyperparameters for each classifier.

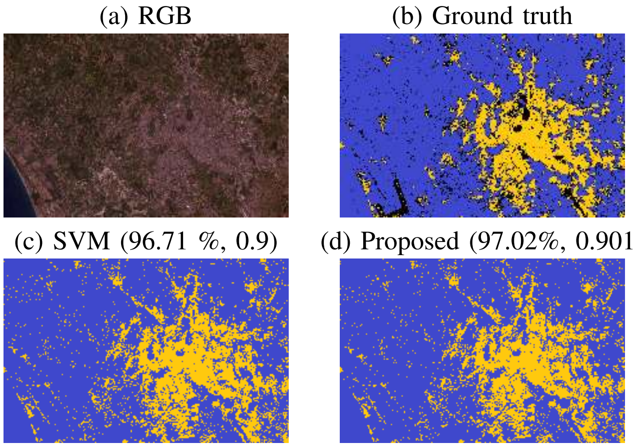

Fig. 1. (a) RGB composite of the Landsat multispectral image, (b) ground truth showing the urban (yellow), non-urban (blue) classes and background (black), (c) classification map with SVMs, and (d) classification map with the proposed method. Overall accuracy and kappa statistic are given in parentheses.

TABLE III
CLASSIFICATION ACCURACY FOR SVM AND THE PROPOSED METHOD IN THE PAVIA SCENE. OVERALL ACCURACY, ESTIMATED COHEN'S STATISTIC AND Z-SCORE RESULTS ARE GIVEN FOR ALL 10 REALIZATIONS AND AVERAGED.

	Overall accuracy, OA[%]		**Kappa statistic, κ**		**Z-score, Z**	
Realization	Proposed	SVM	Proposed	SVM	Proposed	SVM
1	**98.24**	98.10	**0.979**	0.977	**705.28**	678.51
2	97.75	**98.13**	0.973	**0.977**	622.88	**683.29**
3	**98.31**	98.28	**0.979**	**0.979**	**721.51**	712.40
4	**98.42**	**98.42**	**0.981**	**0.981**	743.00	**744.51**
5	**98.46**	98.11	**0.981**	0.977	**754.33**	680.18
6	97.95	**98.36**	0.975	**0.980**	653.39	**730.60**
7	**98.48**	98.27	**0.981**	0.979	**760.01**	710.51
8	**98.29**	98.23	**0.979**	0.978	**714.96**	701.91
9	**98.37**	98.30	**0.980**	0.979	**733.20**	715.85
10	**98.18**	97.87	**0.978**	0.974	**693.48**	640.06
Average	**98.25**	98.21	**0.979**	0.978	**710.20**	699.78
Variance	0.0545	0.0253	$< 10^{-6}$	$< 10^{-6}$	1926.64	906.97

Table III shows the obtained results in the 10 independent realizations and their average and variance. The proposed method provides better results in almost all cases, although the differences are not statistically significant, as assessed by the Z score of the κ statistic for both classifiers. Unlike the overall accuracy, the kappa statistic avoids the chance effect, and a value above 0.8 is typically considered to be a 'very good' agreement. The kappa index confidence interval is $[0.975, 0.980]$ for the proposed method and $[0.975, 0.981]$ for the SVM. These results also confirm the numerical and statistical similarity of the results. Finally, Fig. 2 shows the classification maps obtained by SVM and the proposed method in a particular realization. Visual results match the previous numerical accuracies as no difference is obtained.

C. Active Learning Results

In this second battery of experiments, we illustrate the capabilities of the proposed active learning methods. Classification experiments are conducted using the Rome (Italy) scene acquired in 1999 whose RGB bands are depicted in Fig. 1a. The proposed Bayesian active learning methods are

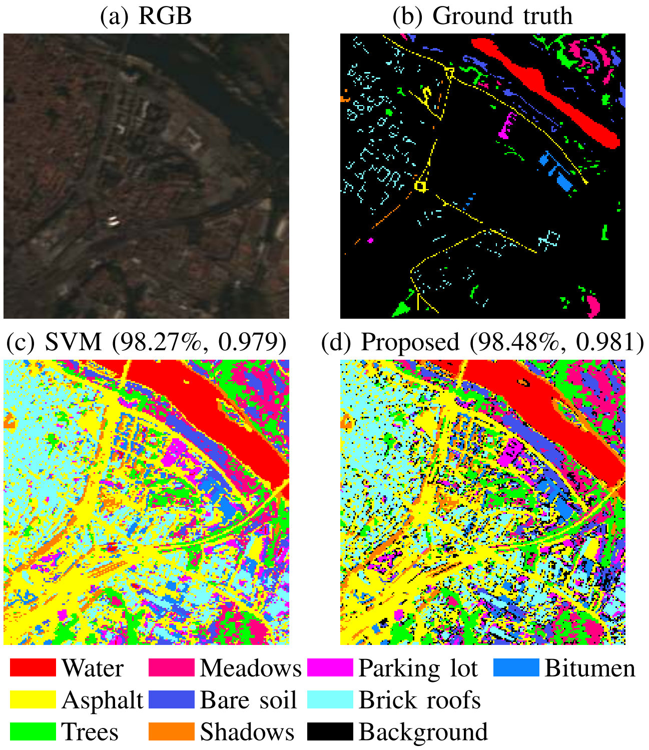

Fig. 2. (a) False color Pavia multispectral image composed by bands [8, 4, 1], (b) ground truth showing classes in colors and background in black, (c) classification map with SVMs, and (d) classification map with the proposed method. Overall accuracy and kappa statistic are given in parentheses.

TABLE IV
FIGURES OF MERIT AT CONVERGENCE IN THE ROME (1999) SCENE (AFTER 100 SAMPLES WERE ADDED) FOR ALL LEARNING METHODS.

Methods	Avg. OA	σ^2_{OA}	Avg. kappa	σ^2_κ	Z-score	κ CI
SVM-RS	95.09	0.7520	0.8467	0.0008	128.79	[0.83,0.86]
SVM-MS	97.08	0.0894	0.9095	0.0001	175.23	[0.90,0.92]
SVM-EQB	97.06	0.1009	0.9094	0.0001	175.53	[0.90,0.92]
BAL-1	96.41	0.1847	0.8869	0.0002	152.87	[0.88,0.90]
BAL-2	97.31	0.0921	0.9166	0.0001	183.82	[0.91,0.93]
BAL-3	**97.34**	**0.0412**	**0.9173**	$< 10^{-4}$	**184.28**	[0.91,0.93]

TABLE V
TOTAL RUNNING TIME IN SECONDS FOR ALL ACTIVE LEARNING METHODS IN THE ROME(1999) SCENE.

SVM-RS	SVM-MS	SVM-EQB	BAL-1	BAL-2	BAL-3
179	185	235	9	9	9

identified as follows: maximum differential of entropies (BAL-1), the minimum distance to decision boundary (BAL-2), and the minimum normalized distance (BAL-3). They are compared to SVM-based approaches following similar heuristics: margin sampling (MS) [21] and entropy-query-by-bagging (EQB) [36]. The naïve (passive) approach of random sampling (RS) is included here as baseline.

Figure 3 shows the average accuracy curves over 10 realizations with different randomly selected training, pool and test sets as a function of the number of training samples. The initial training set is formed by only 7 labeled pixels for each class, while the pool set has 986 spectra, and the test set is formed by 10000 samples. Although the proposed method can be used for the selection of a batch of samples, in the experiments we report results by adding one sample at each iteration (query). At each iteration the SVM model was retrained using 3-fold cross validation on the current training dataset to tune the regularization parameter C. The parameters for the proposed method were automatically estimated using Eqs. (17) and (20). For the EQB method, six classifiers were used. The compared methods perform remarkably differently from the very beginning: while all of them start from approximately $Z = 80$, a fast convergence is observed for all methods but RS, as expected. MS and EQB show very similar performance, and both outperform our proposed BAL-1. The curves also reveal better results at convergence for the BAL-2 and BAL-3 methods. Nevertheless, for a low number of iterations (between 25-50), BAL-3 shows much better results. The dashed line represents the upper bound for OA=97.45 and Z-score=187.62. Table IV gives the accuracy, kappa and Z agreement scores after the full iterative process, when 100 samples were added, and confirms the suitability of the proposed methods, specifically BAL-2 and BAL-3, which show higher accuracies and lower variance. Table V shows the total running time in seconds, after 100 queries, for the compared methods, including the initial learning stage and the parameter estimation at each query. It is worth mentioning that the running time for SVM based methods, MS and EQB, is much higher than the time for the proposed Bayesian active learning methods.

In addition, a multiclass active learning experiment was performed in the Pavia scene. In this experiment, we compare the multiclass extension of the proposed methods with the multiclass versions of RS, MS and EQB. Also, the Multiclass Level Uncertainty method (MCLU) [33] was included in the comparison. Figure 4 shows the average accuracy curves over 10 realizations with different randomly selected training, pool and test sets as a function of the number of training samples. The initial training set is formed by only 5 labeled pixels for each class, while the pool set has 13076 spectra, and the test set is formed by 1453 samples. For the parameter selection we followed the same procedure as in the previous experiment. The proposed methods start with an advantage of 2% with respect to the SVM based methods that, in the case of the proposed BAL-2 and BAL-3 methods, is kept until iteration 40. After that, MS, EQB, MCLU, BAL-2 and BAL-3 have a similar behavior. We think that this is due to the way the parameters are estimated. SVM methods use cross-validation to estimate the parameters and, when the training set is small, it does not provide accurate results. However, the proposed method provides a precise estimation even if the number of training samples is very small. BAL-1 performs similarly to RS which confirms that maximizing the variance of the prediction is not a good selection method by itself but, in some cases, helps when combined with the minimum distance to the decision boundary, as in BAL-3 method. Table VI shows the numerical results when 100 samples were added. From those figures of merit we observe that MCLU provides slightly better results than MS, EQB, BAL-2 and BAL-3. The dashed line represents the upper bound for OA=98.50 and Z-

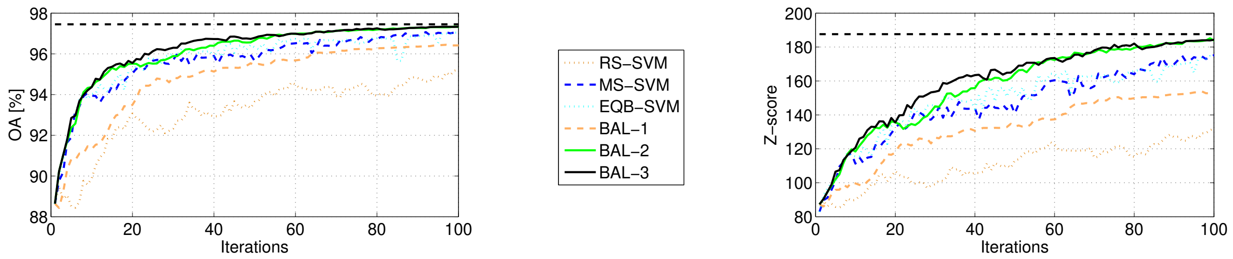

Fig. 3. Average accuracy (left) and Z-score (right) learning curves in the Rome(1999) scene.

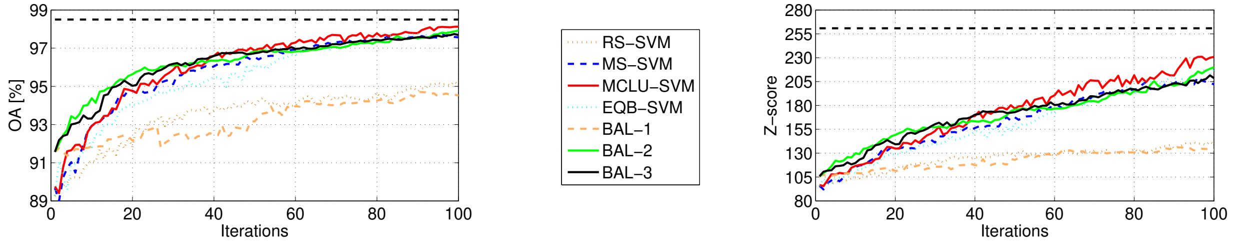

Fig. 4. Average accuracy (left) and Z-score (right) learning curves in the Pavia scene.

TABLE VI
FIGURES OF MERIT AT CONVERGENCE IN THE PAVIA SCENE (AFTER 100 SAMPLES WERE ADDED) FOR ALL ACTIVE LEARNING METHODS.

Methods	Avg. OA	σ^2_{OA}	Avg. kappa	σ^2_κ	Z-score	κ CI
SVM-RS	95.09	0.3812	0.9407	$< 10^{-4}$	139.62	[0.93,0.95]
SVM-MS	97.56	0.2055	0.9706	$< 10^{-4}$	202.10	[0.96,0.98]
SVM-MCLU	**98.12**	**0.0934**	**0.9774**	$< 10^{-4}$	**230.84**	[0.96,0.99]
SVM-EQB	97.90	0.1213	0.9746	$< 10^{-4}$	217.61	[0.96,0.98]
BAL-1	94.51	1.7123	0.9338	0.0003	133.54	[0.92,0.95]
BAL-2	97.92	0.2092	0.9749	$< 10^{-4}$	220.04	[0.97,0.98]
BAL-3	97.69	0.2791	0.9720	$< 10^{-4}$	208.67	[0.96,0.98]

TABLE VII
TOTAL RUNNING TIME IN SECONDS FOR ALL ACTIVE LEARNING METHODS IN THE PAVIA SCENE.

SVM-RS	SVM-MS	SVM-MCLU	SVM-EQB	BAL-1	BAL-2	BAL-3
380	397	401	812	148	165	183

score=260.99. Table VII shows, for the compared methods, the total running time in seconds after 100 queries, including the initial learning stage and parameter estimation at each query. Again the running time for SVM based methods is much higher (from 2 to 5 times depending on the method) than the time required by the proposed Bayesian active learning methods.

VI. CONCLUSIONS

This paper presented a non-parametric Bayesian learning approach based on kernels for remote sensing image classification. The Bayesian methodology efficiently tackles purely supervised and active learning approaches, and shows competitive performance when compared to SVMs and recent active learning approaches. For the latter setting, an incremental learning approach based on three different approaches was presented: the maximum differential of entropies, the minimum distance to decision boundary, and the minimum normalized distance. Automatic parameter estimation is solved by using the evidence Bayesian approach, the kernel trick, and the marginal distribution of the observations instead of the posterior distribution of the adaptive parameters.

The proposed approach was tested in several scenes dealing with the urban monitoring problem from multispectral and SAR data. We observed that, while similar results are obtained by SVMs in supervised mode, an improvement in accuracy and convergence is observed for the active learning scenario. Interestingly our methods do not only provide point-wise class predictions but confidence intervals.

Future work will deal with the application to more challenging multitemporal image segmentation and change detection problems, in which a confidence map could be readily exploited. Also, it is interesting to study the performance of the model in the presence of a reduced number of labeled samples and much higher dimensionality scenarios.

ACKNOWLEDGMENT

The authors would like to thank Dr. J. Muñoz-Marí from the Universitat de València (Spain) and Dr. Devis Tuia at the EPFL (Switzerland) for sharing the code of the RS, MS, and EQB-SVM active learning methods compared in this paper. We would also like to thank Dr. L. Gómez-Chova from the Universitat de València (Spain) for the preprocessing of the images used in this work. Finally, we would like to thanks Dr. J. Malo from the Universitat de València (Spain) for the fruitful discussions and his valuable suggestions.

REFERENCES

[1] B. Schölkopf and A. Smola, *Learning with Kernels – Support Vector Machines, Regularization, Optimization and Beyond*, MIT Press Series, Cambridge, MA, USA, 2002.

[2] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 43, no. 6, pp. 1351–1362, June 2005.

[3] G. Camps-Valls and L. Bruzzone, *Kernel Methods for Remote Sensing Data Analysis*, John Wiley and Sons, 2009.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer, 2007.

[5] M. E. Tipping, "The relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[6] P. Torrione and L.M. Collins, "Texture features for antitank landmine detection using ground penetrating radar," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 45, no. 7, pp. 2374 –2382, July 2007.

[7] F. A. Mianji and Y. Zhang, "Robust hyperspectral classification using relevance vector machine," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 49, no. 6, pp. 2100 –2112, June 2011.

[8] G. Camps-Valls, L. Gómez-Chova, J. Vila-Francés, J. Amorós-López, J. Muñoz-Marí, and J. Calpe-Maravilla, "Retrieval of oceanic chlorophyll concentration with relevance vector machines," *Remote Sensing of Enviroment*, vol. 105, no. 1, pp. 23–33, Nov 2006.

[9] C.E. Rasmussen and C.K. Williams, *Gaussian Processes for Machine Learning*, MIT Press, NY, 2006.

[10] Y. Bazi and F. Melgani, "Gaussian process approach to remote sensing image classification," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 48, no. 1, pp. 186 –197, January 2010.

[11] G. Jun and J. Ghosh, "Spatially adaptive classification of land cover with remote sensing data," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 49, no. 7, pp. 2662 –2673, July 2011.

[12] J. Verrelst, L. Alonso, G. Camps-Valls, J. Delegido, and J. Moreno, "Retrieval of vegetation biophysical parameters using gaussian process techniques," *IEEE Trans. on Geoscience and Remote Sensing*, vol. PP, no. 99, pp. 1 –12, 2011.

[13] A. O'Hagan, *Bayesian Inference*, vol. 2B, chapter 10, Arnold, 1994.

[14] P. Orbanz and Y.-W Teh, *Bayesian Nonparametric Models*, Springer, 2010.

[15] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*, World Scientific, Singapore, 2002.

[16] G. Cauwenberghs and T. Poggio, "Incremental and decremental support vector machine learning," in *NIPS*, 2000, pp. 409–415.

[17] J. Kivinen, A.J. Smola, and R.C. Williamson, "Online learning with kernels," *IEEE Trans. on Signal Proccesing*, vol. 52, no. 8, pp. 2165–2176, 2004.

[18] P. Laskov, C. Gehl, S. Kruger, and K.-R. Müller, "Incremental support vector learning: Analysis implementation and applications," *Journal of Machine Learning Research*, vol. 7, pp. 1909–1936, 2006.

[19] L. Csató and M. Opper, "Sparse on-line Gaussian processes," *Neural Computation*, vol. 14, no. 3, pp. 641–668, 2002.

[20] J. Quiñonero-Candela and O. Winther, "Incremental Gaussian Processes," in *NIPS*, 2002, pp. 1001–08.

[21] B. Settles, "Active learning literature survey," Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[22] J.C. Platt, *Probabilities for SV Machines*, pp. 61–74, MIT Press, 2007.

[23] S. Tong and D. Koller, "Active learning for parameter estimation in Bayesian networks," in *NIPS*, 2000, pp. 647–653.

[24] J. Paisley, X. Liao, and L. Carin, "Active learning and basis selection for kernel-based linear models: A Bayesian perspective," *IEEE Trans. on Signal Processing*, vol. 58, pp. 2686–2700, 2010.

[25] P. Ruiz, J. Mateos, R. Molina, and A.K. Katsaggelos, "A Bayesian Active Learning Framework for a Two-class Classification Problem," in *Lecture Notes in Computer Science series*, 2012, vol. 7252, in press.

[26] Jun Li, J.M. Bioucas-Dias, and A. Plaza, "Hyperspectral image segmentation using a new Bayesian approach with active learning," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 49, no. 10, pp. 3947 –3960, October 2011.

[27] C.J.C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[28] M. Elad, *Sparse and Redundant Representations - From Theory to Applications in Signal and Image Processing*, Springer, 2010.

[29] D. Babacan, R. Molina, and A. Katsaggelos, "Bayesian compressive sensing using Laplace priors," *IEEE Trans. on Image Processing*, vol. 19, no. 1, pp. 53–63, 2010.

[30] D. J. C. MacKay, "Information-based objective functions for active data selection," *Neural Computation*, vol. 4, no. 4, pp. 590–604, 1992.

[31] M. W. Seeger and H. Nickisch, "Compressed sensing and Bayesian experimental design," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 912–919.

[32] P. Mitra, B. Uma Shankar, and S.K. Pal, "Segmentation of multispectral remote sensing images using active support vector machines," *Pattern Recognition Letters*, vol. 25, no. 9, pp. 1067–1074, 2004.

[33] B. Demir, C. Persello, and L. Bruzzone, "Batch mode active learning methods for the interactive classification of remote sensing images," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 49, no. 3, pp. 1014–1032, 2011.

[34] E. Pasolli, F. Melgani, and Y. Bazi, "SVM active learning through significance space construction," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 3, pp. 431–435, 2011.

[35] M. Ferecatu and N. Boujemaa, "Interactive remote sensing image retrieval using active relevance feedback," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 45, no. 4, pp. 818–826, 2007.

[36] D. Tuia, F. Ratle, F. Pacifici, M.F. Kanevski, and W.J. Emery, "Active learning methods for remote sensing image classification," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2218 –2232, July 2009.

[37] M. Volpi, D. Tuia, and M. Kanevski, "Cluster-based active learning for remote sensing image classification," *IEEE Trans. on Geoscience and Remote Sensing*, 2011.

[38] Q. Liu, X. Liao, and L. Carin, "Detection of unexploded ordnance via efficient semisupervised and active learning," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 46, no. 9, pp. 2558 –2567, September 2008.

[39] J. Li, J.M. Bioucas-Dias, and A. Plaza, "Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 48, no. 11, pp. 4085 –4098, November 2010.

[40] D. Tuia, E. Pasolli, and W.J. Emery, "Using active learning to adapt remote sensing image classifiers," *Remote Sensing of Enviroment*, vol. 115, no. 9, pp. 2232–2242, 2011.

[41] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Muñoz-Marí, "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE Journal on Selected Topics in Signal Processing*, vol. 4, pp. 606–617, 2011.

[42] Ryan Rifkin and Aldebaro Klautau, "In defense of one-vs-all classification," *J. Mach. Learn. Res.*, vol. 5, pp. 101–141, Dec. 2004.

[43] R. Molina, A. K. Katsaggelos, and J. Mateos, "Bayesian and regularization methods for hyperparameter estimation in image restoration," *IEEE Transactions on Image Processing*, vol. 8, no. 2, pp. 231–246, 1999.

[44] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag, 1985.

[45] D. J. C. MacKay, "Comparison of approximate methods for handling hyperparameters," *Neural Computation*, vol. 11, no. 5, pp. 1035–1068, 1999.

[46] L. Gomez-Chova, D. Fernández-Prieto, J. Calpe, E. Soria, J. Vila-Francés, and G. Camps-Valls, "Urban monitoring using multitemporal SAR and multispectral data," *Pattern Recognition Letters, Special Issue on "Pattern Recognition in Remote Sensing"*, vol. 27, no. 4, pp. 234–243, 2006.

[47] G. Camps-Valls, D. Tuia, L. Gómez-Chova, S. Jiménez, and J. Malo, Eds., *Remote Sensing Image Processing*, Morgan & Claypool Publishers, LaPorte, CO, USA, Sept 2011, Collection 'Synthesis Lectures on Image, Video, and Multimedia Processing', Al Bovik, Ed.

Pablo Ruiz (S'11) received the M.S. degree in mathematics and the Masters degree in multimedia technologies in 2008 and 2009, respectively. Currently, he is a Ph.D. student of the Visual Information Processing group, at the Department of Computer Science and Artificial Intelligence of the University of Granada, where he enjoys a pre-doctoral fellowship. His research interest include classification, active learning, video retrieval from video databases and computational photography.

Javier Mateos was born in Granada, Spain, in 1968. He received the degree in computer science in 1991 and the Ph.D. degree in computer science in 1998, both from the University of Granada. He was an Assistant Professor with the Department of Computer Science and Artificial Intelligence, University of Granada, from 1992 to 2001, and then he became a permanent Associate Professor. He is coauthor of *Superresolution of Images and Video* (Claypool, 2006) and *Multispectral Image fusion Using Multiscale and Super-resolution Methods* (VDM Verlag, 2011). He was finalist for the IEEE International Conference on Image Processing Best Student Paper Award (2010). He is conducting research on image and video processing, including image restoration, image, and video recovery, super-resolution from (compressed) stills and video sequences, pansharpening and image classification.

Dr. Mateos is a member of the Asociación Espaola de Reconocimento de Formas y Análisis de Imágenes (AERFAI) and International Association for Pattern Recognition (IAPR).

Gustavo Camps-Valls (M'04, SM'07) received a Ph.D. degree in Physics (2002, *summa cum laude*) from the Universitat de València, Spain, where he is currently an Associate Professor in the Electrical Engineering Dep.He teaches time series analysis, image processing, machine learning, and knowledge extraction for remote sensing. His research is conducted as Group Leader of the Image and Signal Processing (ISP) group, http://isp.uv.es, of the same university. He has been Visiting Researcher at the Remote Sensing Laboratory (Univ. Trento, Italy) in 2002, the Max Planck Institute for Biological Cybernetics (Tübingen, Germany) in 2009, and as Invited Professor at the Laboratory of Geographic Information Systems of the École Polytechnique Fédérale de Lausanne (Lausanne, Switzerland) in 2013. His research interests are tied to the development of machine learning algorithms for signal and image processing with special focus on remote sensing data analysis. He conducts and supervises research within the frameworks of several national and international projects, and he is Evaluator of project proposals and scientific organizations. He is the author (or co-author) of 95 international peer-reviewed journal papers, more than 120 international conference papers, 20 international book chapters, and editor of the books "Kernel methods in bioengineering, signal and image processing" (IGI, 2007), "Kernel methods for remote sensing data analysis" (Wiley & Sons, 2009), and "Remote Sensing Image Processing" (MC, 2011). He's a co-editor of the forthcoming book "Digital Signal Processing with Support Vector Machines" (Wiley & sons, 2014). He holds a Hirsch's h index $h = 28$, entered the ISI list of Highly Cited Researchers in 2011, and he is a co-author of the 3 most highly cited papers in relevant remote sensing journals. Thomson Reuters ScienceWatch® identified one of his papers as a Fast Moving Front research. He is a referee of many international journals and conferences, and currently serves on the Program Committees of International Society for Optical Engineers (SPIE) Europe, International Geoscience and Remote Sensing Symposium (IGARSS),Machine Learning for Signal Processing (MLSP), and International Conference on Image Processing (ICIP) among others. In 2007 he was elevated to IEEE Senior Member, and since 2007 he is member of the Data Fusion technical committee of the IEEE Geoscience and Remote Sensing Society, and since 2009 he is member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society. He is member of the MTG-IRS Science Team (MIST) of the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT). He is Associate Editor of the "IEEE Transactions on Signal Processing", "IEEE Signal Processing Letters", "IEEE Geoscience and Remote Sensing Letters", "ISRN Signal Processing Journal", and Guest Editor of "IEEE Journal of Selected Topics in Signal Processing". Visit http://www.uv.es/gcamps for more information.

Rafael Molina (M88) was born in 1957. He received the degree in mathematics (statistics) in 1979 and the Ph.D. degree in optimal design in linear models in 1983. He became Professor of Computer Science and Artificial Intelligence at the University of Granada, Granada, Spain, in 2000. Former Dean of the Computer Engineering School at the University of Granada (1992-2002) and head of the Computer Science and Artificial Intelligence department of the University of Granada (2005-2007).

His research interest focuses mainly in using Bayesian modeling and inference in problems like image restoration (applications to astronomy and medicine), super resolution of images and video, blind deconvolution, computational photography, source recovery in medicine, compressive sensing, low rank matrix decomposition, active learning and classification. See http://decsai.ugr.es/∼rms for publications, funded projects and grants.

Dr. Molina serves the IEEE and other Professional Societies: Applied Signal Processing, Associate Editor (2005-2007), IEEE Trans. on Image Processing, Associate Editor (2010–), Progress in Artificial Intelligence, Associate Editor (2011–) and Digital Signal Processing, Area Editor (2011–). He is the recipient of an IEEE International Conference on Image Processing Paper Award (2007) , an ISPA Best Paper Award (2009) and coauthor of a paper awarded the runner-up prize at Reception for early-stage researchers at the House of Commons.

Aggelos K. Katsaggelos received the Diploma degree in electrical and mechanical engineering from the Aristotelian University of Thessaloniki, Thessaloniki, Greece, in 1979, and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1981 and 1985, respectively.

In 1985, he joined the Department of Electrical Engineering and Computer Science at Northwestern University, Evanston, IL, where he is currently a Professor holder of the AT&T Chair. Before that he was the holder of the Ameritech Chair of Information Technology (19972003). He is also the Director of the Motorola Center for Seamless Communications, a member of the Academic Affiliate Staff, NorthShore University Health System, an affiliated faculty at the Department of Linguistics, and he has an appointment at the Argonne National Laboratory.

He has published extensively (5 books, 180 journal papers, 450 conference papers, 40 book chapters, 20 patents). He is the editor of *Digital Image Restoration* (Springer-Verlag, 1991), co-author of *Rate-Distortion Based Video Compression* (Kluwer, 1997), co-editor of *Recovery Techniques for Image and Video Compression and Transmission* (Kluwer, 1998), and co-author of *Super-Resolution for Images and Video* (Claypool, 2007) and *Joint Source-Channel Video Transmission* (Claypool, 2007).

Dr. Katsaggelos has served the IEEE and other Professional Societies in many capacities; he was, for example, Editor-in-Chief of the IEEE Signal Processing Magazine (19972002), a member of the Board of Governors of the IEEE Signal Processing Society (19992001), and a member of the Publication Board of the IEEE PROCEEDINGS (2003-2007). He is the recipient of the IEEE Third Millennium Medal (2000), the IEEE Signal Processing Society Meritorious Service Award (2001), the IEEE Signal Processing Society Technical Achievement Award (2010), an IEEE Signal Processing Society Best Paper Award (2001), an IEEE International Conference on Multimedia and Expo Paper Award (2006), an IEEE International Conference on Image Processing Paper Award (2007) and an ISPA Best Paper Award (2009). He was a Distinguished Lecturer of the IEEE Signal Processing Society (20072008) and he is a Fellow of IEEE (1998) and SPIE (2009).