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, in particular support vector machines (SVMs), have been successfully introduced to remote sensing image classification. Their properties make them appropriate for dealing with a high number of image features and a 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 nonparametric 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: 1) the maximum differential of entropies; 2) the minimum distance to decision boundary; and 3) 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 images, in both purely supervised and active learning settings. Similar results are obtained when compared to SVMs in the supervised mode, with the advantage of providing posterior estimates for classification and automatic parameter learning. Comparison with random sampling as well as standard active learning methods such as margin sampling and entropy-query-by-bagging reveals a systematic overall accuracy gain and faster convergence with the number of queries.

Index Terms—Bayesian inference, incremental/active learning, multispectral image segmentation, supervised classification.

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 past years, the methods have been successfully introduced into the field of remote sensing image classification [2], [3]. Kernel methods deal efficiently with small-sized datasets of potentially high dimensionality, as in the case of hyperspectral images. The use of the kernel trick [4], as it 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 that 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 nonparametric 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 observation applications. For example, the relevance vector machine and inferring 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. Because of the particular characteristics of remote sensing data, i.e., 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 nonparametric Bayesian methods constitutes a proper theoretical framework to tackle the problem [9], [13], [14]. 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 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 estimating 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 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 the estimation of all the model parameters in a rigorous and sound manner.

Relations between SVMs and Bayesian inference are not new. Note that adopting the least-squares SVM formulation may alternatively allow performing 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, nonparametric Bayesian methods may deal with uncertainties in the data and naturally allow us to develop intuitive incremental/active learning methods. The presented Bayesian kernel-based classifier permits the derivation of efficient closed-form expressions for parameter estimation, as well as performing incremental, adaptive, and active learning in a consistent and principled way.

While kernel-based classification in static scenarios has been extensively studied, the problem of online and incremental classification is still unsolved. The most effective schemes so far make use of both incremental and online SVMs [16]–[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 both the propagation of predictions and estimation of Bayesian errors.

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]–[29]. See also [30] for the general theory of basis selection 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]–[34]. Some alternative approaches to work with batches of selections per iteration have been presented, which mainly rely on the concept of diversity between candidate pixels [33], [35] 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. This paper presents the novel introduction of nonparametric Bayesian learning for remote sensing image classification in both purely supervised and active learning settings. This approach proposes an iterative procedure to maximize the margin 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 image classification. The presented methods actually go one step further by extending standard nonparametric large-margin techniques, such as SVMs, 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 IV 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, \( \{x_i| i = 1, \ldots, n\}, x \in \mathbb{R}^d \) we want to classify. The general two-class supervised classification problem we tackle here defines a classification function of the form

\[
y(x) = \phi^T(x)w + b + \epsilon
\]

where the mapping \( \phi : \mathcal{X} \to \mathcal{H} \) maps the observed data point (samples, spectra) \( x \in \mathcal{X} \) 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 \( y_k(x) = \phi^T(x)w_k + b_k + \epsilon_k \), \( k = 1, \ldots, K \) \[4\].

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 \( x \), the output \( y(x) \in \{0, 1\} \) consists of a binary coding representation of its classification as belonging to class \( C_0 \) or \( C_1 \), respectively, \( w \) is a vector of size \( L \times 1 \) of adaptive parameters to be estimated, \( b \) represents the bias in the classification function, and \( \epsilon \) is an independent realization of the Gaussian distribution \( \mathcal{N}(0, \sigma^2) \).

For a training set, we already know the classification output \( y(x_i) \) associated with the feature samples \( \phi(x_i), i = 1, \ldots, M, \) with \( M \) the number of samples, and therefore we can write

\[
p(y|x, b, \sigma^2) = \prod_{i=1}^{M} \mathcal{N}(y(x_i)|\phi^T(x_i)w + b, \sigma^2)
\]

where \( y = (y(x_1), y(x_2), \ldots, y(x_M))^T \). Since \( x_i, i = 1, \ldots, M \) will always appear as conditioning variable, for the sake of simplicity, we have removed the dependency on \( x_1, \ldots, x_M \) on the left-hand side of the equation. We note that, for infinite-dimensional feature vectors \( \phi(x_i), w \) is infinite dimensional.

The Bayesian framework allows us to introduce information about the possible value of \( w \) in the form of a prior distribution. Since the likelihood function defined in (2) is the exponential of a quadratic function of \( w \), its corresponding conjugate prior should be a Gaussian distribution [4] so that the posterior will also be Gaussian. In this paper, we consider a particular form of the Gaussian prior in which each component of \( w \) 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 we will still be able to carry out parameter estimation, prediction, and active learning tasks.

III. PROPOSED BAYESIAN INFERENCE METHOD

Because of the possible use of infinite-dimensional feature spaces, we will mainly use the marginal distribution of the observations to perform inference tasks, i.e., parameter estimation, prediction, and active learning, and avoid, when possible, the use of the posterior distribution of the adaptive parameters \( w \) 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 \( y \)

The marginal distribution of \( y \) can be obtained by integrating out the vector of adaptive parameters \( w \). It can easily be shown (see, e.g., [4]) that

\[
p(y|b, \gamma^2, \sigma^2) = \mathcal{N}(y|b1, C)
\]

with

\[
C = \gamma^2 \Phi \Phi^T + \sigma^2 I
\]

where \( \Phi \) is the design matrix whose \( i \)th row is \( \phi^T(x_i) \), and \( I \) 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 \( \Phi \) explicitly to calculate this marginal distribution. We only need to know the Gram matrix \( K = \Phi \Phi^T \), which is an \( M \times M \) symmetric matrix with elements \( K_{nm} = k(x_n, x_m) = \phi^T(x_n)\phi(x_m) \). 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 \( k(x, x') \) for which the Gram matrix \( K \) is positive semidefinite for all possible choices of the set \( \{x_n\} \). Note that, even if \( \Phi \) has an infinite number of columns, which corresponds to the case of \( \phi(x_i) \) being an infinite-dimensional feature vector, we can still calculate \( K \) 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 \( y \) consists of independent additive noisy observations, with variance \( \gamma^2 \), of a Gaussian process with mean \( b \) and covariance \( K \).

For a new sample \( x_s \), the distribution of

\[
y_{M+1} = \left( \begin{array}{c} y \\ y(x_s) \end{array} \right)
\]

has the form

\[
p\left(y_{M+1}|b, \gamma^2, \sigma^2\right) = \mathcal{N}(y_{M+1}|b1_{M+1}, CM_{M+1})
\]

\[2\] 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 \( C_{M+1} = \gamma^2 \Phi_{M+1} \Phi_{M+1}^T + \sigma^2 I_{M+1} \), which can be written as
\[
C_{M+1} = \begin{pmatrix} C & k \\ k^T & c \end{pmatrix}
\]
where \( C \) has been defined in (4) and
\[
k^T = \gamma^2 \Phi^T (x_n) \Phi^T \quad \text{and} \quad c = \gamma^2 \Phi^T (x_n) \Phi (x_n) + \sigma^2.
\]
Furthermore, the conditional distribution \( p(y(x_n)|y) \) is a Gaussian distribution with mean \( m(x_n) \) and variance \( v(x_n) \) given by
\[
m(x_n) = b + \gamma^2 \Phi^T (x_n) \Phi^{-1} (y - b1) \quad \text{and} \quad v(x_n) = c - k^T C^{-1} k.
\]

B. Posterior Distribution of \( w \)
When the feature space is finite dimensional, we can also calculate the posterior distribution of \( w \), which is given by (see [4])
\[
p(w|y, b, \gamma^2, \sigma^2) = \mathcal{N}(w|\Sigma_w[y, \gamma^2, \sigma^2 - \sigma^{-2} \Phi^T (y - b1), \Sigma_w[y, \gamma^2, \sigma^2])
\]
where
\[
\Sigma_w[y, \gamma^2, \sigma^2] = (\sigma^{-2} \Phi^T \Phi + \gamma^{-2} I)^{-1}.
\]
Notice that \( m(x_n) \) defined in (10) can be expressed in terms of \( E[w] \) as
\[
m(x_n) = \phi(x_n)^T E[w] + b.
\]

C. Parameter Estimation
The last step in the Bayesian inference we carry out is the estimation of the parameters involved in the model, i.e., the values of \( \gamma^2, \sigma^2, \text{ and } b \). The value of \( b \) can be easily obtained from (3) as
\[
b = \frac{1}{M} \sum_{i=1}^{M} y(x_i).
\]
To estimate the values of \( \gamma^2 \) and \( \sigma^2 \), 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 \( \gamma^2 \) and \( \sigma^2 \) by maximizing the marginal distribution in (3) obtained by integrating out the vector of adaptive parameters \( w \). Intuitively, by integrating over \( w \), we are searching for the best value of \( \gamma^2 \) and \( \sigma^2 \) for all possible values of \( w \). Differentiating \( 2 \ln p(y|b, \gamma^2, \sigma^2) \) with respect to \( \gamma^2 \) and equating the result to zero, we obtain
\[
\text{tr}[C^{-1} \Phi \Phi^T] = \text{tr}[(y - b1)^T C^{-1} \Phi \Phi^T C^{-1} (y - b1)].
\]
Diagonalizing \( \Phi \Phi^T \), we obtain \( U \Phi \Phi^T U^T = D \), where \( U \) is an orthonormal matrix and \( D \) is a diagonal matrix with entries \( \lambda_i, i = 1, \ldots, M \). We can then rewrite the above equation as
\[
\sum_{k=1}^{M} \frac{\lambda_k}{\gamma^2 \lambda_k + \sigma^2} = \sum_{i=1}^{M} \frac{\lambda_i}{\gamma^2 \lambda_i + \sigma^2}
\]
where \( U(y - b1) = z \) with components \( z_i, i = 1, \ldots, M \).

Multiplying both sides of the above equation by \( \gamma^2 \), we have
\[
\gamma^2 = \sum_{i=1}^{M} \frac{\lambda_i}{\gamma^2 \lambda_i + \sigma^2} \quad \text{and} \quad \gamma^2 \lambda_i + \sigma^2 = \sum_{i=1}^{M} \frac{\gamma^2 \lambda_i}{\gamma^2 \lambda_i + \sigma^2}
\]
where
\[
\mu_i = \frac{\lambda_i}{\sum_{k=1}^{M} \lambda_k + \sigma^2}.
\]
Note that \( \mu_i \geq 0 \) and \( \sum_{i=1}^{M} \mu_i = 1 \).
Similarly, differentiating \( 2 \ln p(y|\gamma^2, \sigma^2) \) with respect to \( \sigma^2 \) and equating the result to zero, we obtain
\[
\sum_{i=1}^{M} \frac{1}{\gamma^2 \lambda_i + \sigma^2} = \sum_{i=1}^{M} \frac{1}{(\gamma^2 \lambda_i + \sigma^2)^2}.
\]
Following the same steps we already performed to estimate \( \gamma^2 \), we obtain
\[
\sigma^2 = \sum_{i=1}^{M} \frac{1}{\gamma^2 \lambda_i + \sigma^2}
\]
where
\[
\tau_i = \frac{1}{\lambda_i + \sigma^2}.
\]
Note that, again, \( \tau_i \geq 0 \) and \( \sum_{i=1}^{M} \tau_i = 1 \).
Equations (17) and (20) suggest the iterative procedure described in Algorithm 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 (14), compute \( b = \frac{1}{M} \sum_{i=1}^{M} y(x_i) \).
Compute \( U \) and \( \lambda_i, i = 1, \ldots, M \), as the eigenvector matrix and eigenvalues of \( \Phi \Phi^T \), respectively.
Set \( z = U(y - b1) \).
Initialize \( \gamma^2 = 1, \sigma^2 = 1 \).

repeat
Set \( \gamma^2_{old} = \gamma^2, \sigma^2_{old} = \sigma^2 \).
Set \( \gamma^2 = \sum_{i=1}^{M} \mu_i \gamma_{old}^2 z_i^2 / (\gamma_{old}^2 \lambda_i + \sigma_{old}^2) \).
Set \( \sigma^2 = \sum_{i=1}^{M} \tau_i \gamma_{old}^2 z_i^2 / (\gamma_{old}^2 \lambda_i + \sigma_{old}^2) \).
until \( (\gamma^2 - \gamma_{old}^2)^2 / \gamma_{old}^2 < 10^{-6} \) and \( (\sigma^2 - \sigma_{old}^2)^2 / \sigma_{old}^2 < 10^{-6} \).

D. Classification
Once the system has been trained, we want to assign a class to a new value of \( x \), denoted by \( x_s \). We already know that the conditional distribution \( p(y(x_s)|y) \) is a Gaussian distribution with mean \( m(x_s) \) and variance \( v(x_s) \) given in (10) and (11).

We classify \( x_s \) utilizing \( m(x_s) \) and write
\[
x_s \text{ is assigned to } \begin{cases} C_1 & \text{if } m(x_s) \geq 0.5 \\ C_0 & \text{if } m(x_s) < 0.5 \end{cases}
\]
Notice that the classification of \( x_n \) is based on the proximity of the mean value of \( p(y(x_n)|y) \) to the value 0 or 1, which represents the classes \( C_0 \) and \( C_1 \), 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 \( y \), the conditional distribution of \( w \) given \( y \), and the parameters \( b, \gamma, \) and \( \sigma^2 \) 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 \( x_+ \), whose corresponding \( y(x_+) \) 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 (10) and (11), we observe that, for a sample \( x \) not already present in the training set, the distribution of \( y(x) \), given the set of observations \( y \), has variance

\[
v(x) = \gamma^2 \phi^T(x)\phi(x) + \sigma^2 - \gamma^4 \phi^T(x)\phi(x)\phi^T C^{-1}\phi(x)
\]

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

\[
x_+ = \arg \max_x v(x).
\]

Notice that using this criterion amounts to selecting the sample which 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 \( w \) when a new \( x \) is added to the training set is given by

\[
\Sigma_{w|y,y^2,\gamma^2} = \left( \sigma^{-2}(I + \phi(x)) + \gamma^2 I \right)^{-1}.
\]

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

\[
x_+ = \arg \max_x \frac{1}{2} \log |\Sigma_{w|y,y^2,\gamma^2}|^{-1}.
\]

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

\[
\log |\Sigma_{w|y,y^2,\gamma^2}|^{-1} = \log \left( \sigma^{-2}(I + \phi(x)) + \gamma^2 I \right)^{-1}
\]

and using

\[
(\sigma^{-2}(I + \phi(x)) + \gamma^2 I)^{-1} = \gamma^2 I - \gamma^4 \phi^T C^{-1} \phi
\]

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

\[
\log |\Sigma_{w|y,y^2,\gamma^2}|^{-1} = \log(1 + \sigma^{-2}(I + \phi(x))\phi(x) - \gamma^2 \phi^T C^{-1} \phi(x))
\]

Consequently, all needed quantities to select \( x_+ \) can be calculated without the knowledge of the feature vectors and the posterior distribution of the possibly infinite-dimensional adaptive parameters but using only kernel functions and the marginal distribution of the observations.

Furthermore, we have

\[
\log |\Sigma_{w|y,y^2,\gamma^2}|^{-1} = \log(\sigma^{-2}v(x))
\]

and consequently both criteria coincide. Notice that we have also shown that the maximum differential of entropies criterion can be used over infinite-dimensional feature spaces.

B. Method 2: Minimum Distance to Decision Boundary

In our classification problem, the decision boundary corresponds to the set

\[
\Pi = \{ x \in \mathcal{X} : \phi^T(x)E[w] + b - 0.5 = 0 \}.
\]

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

\[
x_+ = \arg \min_x d^2(x, \Pi)
\]

\[
= \arg \min_x \frac{(\phi^T(x)E[w] + b - 0.5)^2}{\|E[w]\|^2}
\]

\[
= \arg \min_x (m(x) - 0.5)^2.
\]

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 \( p(y(x_n)|y) \). While the 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:

\[
x_+ = \arg \min_x E \left[ \frac{(y(x) - 0.5)^2}{v(x)} \right]
\]

where the expected value is calculated utilizing the conditional distribution \( p(y(x)|y) \) defined in (10) and (11).

Notice that, since

\[
E \left[ \frac{(y(x) - 0.5)^2}{v(x)} \right] = 1 + \frac{(m(x) - 0.5)^2}{v(x)}
\]

we can rewrite this criterion as

\[
x_+ = \arg \min_x \frac{(m(x) - 0.5)^2}{v(x)}.
\]
D. Multiclass Extension of the Active Learning Methods

Here we extend the proposed active learning methods to deal with $K$-class problems. Recently, architectures for multiclass active learning have been proposed. For instance, in [33] authors propose the multiclass level uncertainty (MCLU) technique. In this paper, nevertheless, we will use the classical one-versus-all strategy for tackling multiclass problems. Hence, for each candidate $s$, $K$ different pair of values $(m_s(x), v_s(x))_{k=1,...,K}$ are obtained. These values are used in (24), (32), or (35), depending on the selected method, that is finally optimized with respect to $x$ 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 the Z-score as measures of accuracy and class agreement, respectively. All experiments were run on an Intel i7@2.67-GHz processor. The MATLAB source code of the proposed method is available at http://decsai.ugr.es/vip/resources/BAL.html.

A. Study Area and Data Collection

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

1) 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 synthetic aperture radar (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]. We also appended two SAR features: the estimated coherence $C_0$, and a spatially filtered version of the coherence $FC_0$ that is specially designed to increase the discrimination of urban areas [46]. After this preprocessing, all features were stacked at the pixel level, and each feature was standardized. The goal is the discrimination of urban ($C_1$) versus nonurban ($C_0$) land-cover classes.

2) Active classification with ROSIS imagery: The second image was acquired by the DAIS7915 sensor over the city of Pavia (Italy), and constitutes a challenging nine-class urban classification problem dominated by directional features and relatively high spatial resolution (5-m 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 component analysis (PCA) to reduce the dimensionality of the problem and considered the 10 first components for each pixel that 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 with the standard SVM approach.

From the Rome image, of size $1440 \times 930$ 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 threefold cross-validation with the SVM as classifier, a kernel length scale $\sigma = 100$ was selected. Although we could have used Bayesian inference to estimate the kernel parameter (see [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 threefold 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 s to complete the training while the SVM needed 1.94 s.

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 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 nine-class urban classification problem of the Pavia scene depicted in Fig. 2(a), which has $400 \times 400$ pixels. Training was done on 1260 randomly selected pixels (140 from each class), and a test set of 13 314 representative samples was used. Again, 10
different realizations were used to obtain unbiased conclusions from the results. We used a Gaussian kernel, whose length scale $\sigma = 500$ was selected using threefold cross-validation with the SVM as classifier. As in the previous experiment, the regularization parameter $C$ for the SVM was tuned by threefold cross-validation on the training dataset, whereas all hyperparameters were estimated by the proposed method automatically in the training phase. The proposed method needed 14.32 s to complete the training while the SVM needed 9.09 s. 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.

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 $\kappa$ 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.

### Classification Accuracy Results

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. 1(a). The proposed Bayesian active learning methods are 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 with the SVM-based approaches following

---

### TABLE I

<table>
<thead>
<tr>
<th>Realization</th>
<th>Overall Accuracy, OA[%]</th>
<th>Kappa Statistic, $\kappa$</th>
<th>Z-score, $Z$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Proposed SVM</td>
<td>Proposed SVM</td>
<td>Proposed SVM</td>
</tr>
<tr>
<td>1</td>
<td>96.66</td>
<td>96.95</td>
<td>0.895</td>
</tr>
<tr>
<td>2</td>
<td>96.48</td>
<td>96.61</td>
<td>0.890</td>
</tr>
<tr>
<td>3</td>
<td>97.27</td>
<td>96.96</td>
<td>0.914</td>
</tr>
<tr>
<td>4</td>
<td>96.24</td>
<td>96.54</td>
<td>0.883</td>
</tr>
<tr>
<td>5</td>
<td>97.10</td>
<td>96.54</td>
<td>0.909</td>
</tr>
<tr>
<td>6</td>
<td>96.64</td>
<td>95.99</td>
<td>0.893</td>
</tr>
<tr>
<td>7</td>
<td>96.86</td>
<td>96.58</td>
<td>0.905</td>
</tr>
<tr>
<td>8</td>
<td>96.76</td>
<td>96.72</td>
<td>0.895</td>
</tr>
<tr>
<td>9</td>
<td>97.02</td>
<td>96.71</td>
<td>0.901</td>
</tr>
<tr>
<td>10</td>
<td>96.99</td>
<td>97.00</td>
<td>0.906</td>
</tr>
<tr>
<td>Average</td>
<td>96.30</td>
<td>96.66</td>
<td>0.90</td>
</tr>
</tbody>
</table>

### TABLE II

<table>
<thead>
<tr>
<th></th>
<th>$C_0$</th>
<th>$C_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_0$</td>
<td>7802.80 (7846.30)</td>
<td>169.00 (198.30)</td>
</tr>
<tr>
<td>$C_1$</td>
<td>165.00 (121.50)</td>
<td>1863.20 (1833.90)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SVM</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>OA [%]</td>
<td>96.66%</td>
<td>96.80%</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>3.07e-05</td>
<td>3.02e-05</td>
</tr>
<tr>
<td>Z-score</td>
<td>162.29</td>
<td>164.21</td>
</tr>
<tr>
<td>$\kappa$ CI</td>
<td>[0.886, 0.908]</td>
<td>[0.889, 0.910]</td>
</tr>
</tbody>
</table>

### TABLE III

<table>
<thead>
<tr>
<th>Realization</th>
<th>Overall Accuracy, OA[%]</th>
<th>Kappa Statistic, $\kappa$</th>
<th>Z-score, $Z$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Proposed SVM</td>
<td>Proposed SVM</td>
<td>Proposed SVM</td>
</tr>
<tr>
<td>1</td>
<td>98.24</td>
<td>98.10</td>
<td>0.979</td>
</tr>
<tr>
<td>2</td>
<td>97.75</td>
<td>98.13</td>
<td>0.973</td>
</tr>
<tr>
<td>3</td>
<td>98.31</td>
<td>98.28</td>
<td>0.979</td>
</tr>
<tr>
<td>4</td>
<td>98.42</td>
<td>98.42</td>
<td>0.981</td>
</tr>
<tr>
<td>5</td>
<td>98.46</td>
<td>98.11</td>
<td>0.981</td>
</tr>
<tr>
<td>6</td>
<td>97.95</td>
<td>98.36</td>
<td>0.975</td>
</tr>
<tr>
<td>7</td>
<td>98.48</td>
<td>98.27</td>
<td>0.981</td>
</tr>
<tr>
<td>8</td>
<td>98.29</td>
<td>98.23</td>
<td>0.979</td>
</tr>
<tr>
<td>9</td>
<td>98.37</td>
<td>98.30</td>
<td>0.980</td>
</tr>
<tr>
<td>10</td>
<td>98.18</td>
<td>97.87</td>
<td>0.978</td>
</tr>
<tr>
<td>Average</td>
<td>98.25</td>
<td>98.21</td>
<td>0.979</td>
</tr>
<tr>
<td>Variance</td>
<td>0.0545</td>
<td>0.0253</td>
<td>&lt; 10^-5</td>
</tr>
</tbody>
</table>
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.

Fig. 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 seven labeled pixels for each class, while the pool set has 986 spectra and the test set is formed by 10 000 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 threefold cross-validation on the current training dataset to tune the regularization parameter $C$. The parameters for the proposed method were automatically estimated using (17) and (20). For the EQB method, six classifiers were used. The compared methods perform quite differently from the very beginning; while all of them start from approximately $Z = 80$, a fast convergence is observed for all methods except 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 small number of iterations (between 25 and 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, i.e., MS and EQB, is much higher than that 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 MCLU [33] was included in the comparison. Fig. 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 five labeled pixels for each class, while the pool set has 13 076 spectra, and the test set is formed by 1453 samples. For 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, which, 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 they do 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 similar 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
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 were obtained by SVMs in supervised mode, an improvement in accuracy and convergence was observed for the active learning scenario. Interestingly, our methods provide not only pointwise class predictions but also 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-Márí from the Universitat de València (Spain) and Dr. Devis Tuia from the EPFL (Switzerland) for sharing the code of the RS, MS, and EQB-SVM active learning methods compared in this paper; Dr. L. Gómez-Chova from the Universitat de València (Spain) for the preprocessing of the images used; and Dr. J. Malo from the Universitat de València (Spain) for fruitful discussions and valuable suggestions.

**REFERENCES**


TABLE VI

<table>
<thead>
<tr>
<th>Methods</th>
<th>Avg. OA</th>
<th>Avg. kappa</th>
<th>Z-score</th>
<th>k CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM-RS</td>
<td>95.09</td>
<td>0.3812</td>
<td>0.9407</td>
<td>&lt; 10^{-4}</td>
</tr>
<tr>
<td>SVM-MS</td>
<td>97.56</td>
<td>0.2055</td>
<td>0.9706</td>
<td>&lt; 10^{-4}</td>
</tr>
<tr>
<td>SVM-MCLU</td>
<td>98.12</td>
<td>0.0934</td>
<td>0.9774</td>
<td>&lt; 10^{-4}</td>
</tr>
<tr>
<td>SVM-EQB</td>
<td>97.90</td>
<td>0.1213</td>
<td>0.9746</td>
<td>&lt; 10^{-4}</td>
</tr>
<tr>
<td>BAL-1</td>
<td>94.51</td>
<td>1.7123</td>
<td>0.9338</td>
<td>0.0003</td>
</tr>
<tr>
<td>BAL-2</td>
<td>97.92</td>
<td>0.2092</td>
<td>0.9749</td>
<td>&lt; 10^{-4}</td>
</tr>
<tr>
<td>BAL-3</td>
<td>97.69</td>
<td>0.2791</td>
<td>0.9720</td>
<td>&lt; 10^{-4}</td>
</tr>
</tbody>
</table>

TABLE VII

<table>
<thead>
<tr>
<th>Methods</th>
<th>Total Running Time in Seconds for All Active Learning Methods in the Pavia Scene</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM-RS</td>
<td>380</td>
</tr>
<tr>
<td>SVM-MS</td>
<td>397</td>
</tr>
<tr>
<td>SVM-MCLU</td>
<td>401</td>
</tr>
<tr>
<td>SVM-EQB</td>
<td>812</td>
</tr>
<tr>
<td>BAL-1</td>
<td>148</td>
</tr>
<tr>
<td>BAL-2</td>
<td>165</td>
</tr>
<tr>
<td>BAL-3</td>
<td>183</td>
</tr>
</tbody>
</table>

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

VI. CONCLUSION

This paper presented a nonparametric 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 with SVMs and recent active learning approaches. For the latter setting, an incremental learning approach based on three different approaches was presented. Automatic parameter estimation was 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.


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 currently pursuing the Ph.D. degree with the Visual Information Processing Group, Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain. He was Pre-Doctoral Fellowship. His current research interests 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 from the University of Granada, 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 co-author of Superresolution of Images and Video (Claypool, 2006) and Multispectral Image Fusion Using Multiscale and Super-resolution Methods (VDM Verlag, 2011). Dr. Mateos was a finalist for the IEEE International Conference on Image Processing Best Student Paper Award in 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. He is a member of the Asociación Española de Reconocimiento de Formas y Análisis de Imágenes y Interna-
Rafael Molina (M’88) 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 a Professor of computer science and artificial intelligence with the University of Granada, Granada, Spain, in 2000. He was a Former Dean of the Computer Engineering School, University of Granada, from 1992 to 2002, and the Head of the Computer Science and Artificial Intelligence Department, University of Granada, from 2005 to 2007. His current research interests include 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.

Dr. Molina serves the IEEE and other Professional Societies: Applied Signal Processing, an Associate Editor from 2005 to 2007, the IEEE TRANSACTIONS ON IMAGE PROCESSING, an Associate Editor since 2010, Progress in Artificial Intelligence, an Associate Editor since 2011 and Digital Signal Processing, an Area Editor since 2011. He was a recipient of an IEEE International Conference on Image Processing Paper Award in 2007, an ISPA Best Paper Award in 2009 and Co-author of a Paper Awarded the Runner-Up Prize at Reception for early-stage researchers at the House of Commons.

Aggelos K. Katsaggelos (F’98) 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, GA, USA, in 1981 and 1985, respectively.

He joined the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA, in 1985, where he is currently a Professor holder of the AT&T Chair. He was the holder of the Ameritech Chair of Information Technology from 1997 to 2003. He is the Director of the Motorola Center for Seamless Communications, a member of the Academic Affiliate Staff, NorthShore University Health System, an Affiliated Faculty with the Department of Linguistics, and he has an appointment with the Argonne National Laboratory. 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). He has published over five books, 180 journal papers, 450 conference papers, 40 book chapters, and 20 patents.

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 from 1997 to 2002, a member of the Board of Governors of the IEEE Signal Processing Society from 1999 to 2001, and a member of the Publication Board of the IEEE PROCEEDINGS from 2003 to 2007. He was a recipient of the IEEE Third Millennium Medal in 2000, the IEEE Signal Processing Society Meritorious Service Award in 2001, the IEEE Signal Processing Society Technical Achievement Award in 2010, the IEEE Signal Processing Society Best Paper Award in 2001, the IEEE International Conference on Multimedia and Expo Paper Award in 2006, the IEEE International Conference on Image Processing Paper Award in 2007, and the ISPA Best Paper Award in 2009. He was a Distinguished Lecturer of the IEEE Signal Processing Society from 2007 to 2008 and he has been a Fellow of SPIE since 2009.