2009 Ninth International Conference on Intelligent Systems Design and Applications

Fully Unsupervised Learning of Gaussian Mixtures for Anomaly Detection in Hyperspectral Imagery

Tiziana Veracini, Stefania Matteoli, Marco Diani, Giovanni Corsini
Dipartimento di Ingegneria dell'Informazione
Università di Pisa
Pisa, Italy
{tiziana.veracini, stefania.matteoli, m.diani, g.corsini}@iet.unipi.it

***Abstract*— This paper proposes a fully unsupervised anomaly detection strategy in hyperspectral imagery based on mixture learning. Anomaly detection is conducted by adopting a Gaussian Mixture Model (GMM) to describe the statistics of the background in hyperspectral data. One of the key tasks in the application of mixture models is the specification in advance of the number of GMM components, the determination of which is essential and strongly affects detection performance. In this work, GMM parameters estimation was performed through a variation of the well-known Expectation Maximization (EM) algorithm that was developed within a Bayesian framework. Specifically, the adopted mixture learning technique incorporates a built-in mechanism for automatically assessing the number of components during the parameter estimation procedure. Then, Generalized Likelihood Ratio Test (GLRT) is considered for detecting anomalies. Real hyperspectral imagery acquired by an airborne sensor is used for experimental evaluation of the proposed anomaly detection strategy.**

Keywords- hyperspectral imagery; Gaussian mixture; model selection; Bayesian approach; anomaly detection

I. INTRODUCTION

In recent years, hyperspectral remote sensing has found many applications in earth observation, such as, environmental monitoring, land use management, and wide-area surveillance. Hyperspectral sensors collect data in several narrow and adjacent spectral bands, thus providing a very densely sampled spectrum for each pixel in the scene. Such a high spectral resolution preserves important aspects of the spectrum and makes it possible to reveal even very subtle spectral characteristics. In fact, hyperspectral sensing has proven valuable for discrimination of materials on the basis of their unique *spectral signature*, which is the spectral reflectance as to the Visible/Near InfraRed – Short Wave InfraRed (VNIR-SWIR) range [1].

In this work, we are interested in Anomaly Detection (AD), which aims at detecting targets that are "rare" in the image (i.e. characterized by a low probability of occurrence with respect to background objects), without knowledge of their spectral signature [1]. Therefore, AD algorithms search the image for pixels whose spectral content is significantly different from that of background. Hence, estimating background distribution is an essential step of most of AD algorithms. Recently, parametric models have been used to describe background statistics in hyperspectral imagery [1, 2]. In [1, 3], AD algorithms have been developed that rely upon a parametric Gaussian Mixture Model (GMM) for background characterization.

Mixture models have been successfully applied for modeling large heterogeneous populations [2, 4], and they have been adopted in many applications such as clustering and density estimation [2, 5-7]. Within this framework, the GMM is undoubtedly one of the most widely adopted models for approximating distributions [2, 5, 7]. GMM learning has been typically conducted through the well-known Expectation Maximization (EM) approach [2, 8], which estimates the mixture parameters from the data once the number of GMM components has been specified a-priori. Therefore, a not correct choice of this parameter could strongly degrades the estimation accuracy of the data distribution. This is particularly significant when using mixture learning in AD applications [9-11], where a not reliable background characterization may seriously compromise the target detection outcome.

In the literature, methods for GMM learning that adopt a Bayesian approach for automatically assessing the number of mixture components were developed [2, 4, 12-16]. However, their potential has been shown and exploited restricting to low-dimensional simulated data, artificial texture images, handwritten digits, and natural images.

In this paper, the same Bayesian philosophy to GMM learning is adopted and embodied in an anomaly detection scheme to be applied to hyperspectral data. In this way, the number of background components is automatically determined during the background parameters estimation procedure. Real hyperspectral imagery from an airborne sensor is used to evaluate performance of the proposed strategy. Results obtained are compared to those of an AD approach based on the classical EM approach for mixture learning.

The structure of the paper is organized as follows. In section II, the Bayesian methods for Gaussian mixture learning are described, whereas in section III we illustrate the proposed AD strategy. Section IV describes the hyperspectral data set used in the analysis and the design of experiments. Experimental results and conclusions are discussed in sections V and VI, respectively.

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.220

II. VARIATIONAL BAYESIAN MODEL SELECTION FOR GAUSSIAN MIXTURE DISTRIBUTIONS

Mixture models are flexible and valuable statistical tools for modeling a Probability Density Function (PDF). In particular, Gaussian mixture provides a computationally tractable representation for PDF shape that can be used to model heterogeneous data in high dimension.

A Gaussian Mixture Model (GMM) [2, 4] is a parametric model that assumes the data originate from a weighted sum of several multivariate Gaussian sources. Formally, the finite GMM with J components can be expressed as:

$$f_{\mathbf{X}}(\mathbf{x}) = \sum_{j=1}^{J} \pi_j N\left(\mathbf{x}; \boldsymbol{\mu}_j, \mathbf{T}_j^{-1}\right) \quad (1)$$

where $\mathbf{X}$ is a multidimensional random vector, $N\left(\mathbf{x}; \boldsymbol{\mu}_j, \mathbf{T}_j^{-1}\right)$ is the multivariate normal PDF, characterized by mean vector $\boldsymbol{\mu}_j$, precision (inverse covariance) matrix $\mathbf{T}_j$, and mixing proportion (weight) $\pi_j \geq 0$, which is subject to the constraint $\sum_{j=1}^{J} \pi_j = 1$. It has been shown that by using a sufficient number of Gaussian sources, and by adjusting their means and covariances as well as the weights in the linear combination of equation (1), any continuous PDF can be approximated with high accuracy [1, 2, 4]. The shape of the GMM PDF is hence governed by $\boldsymbol{\pi}=\{\pi_j | j=1,\ldots,J\}$, $\boldsymbol{\mu}=\{\boldsymbol{\mu}_j | j=1,\ldots,J\}$ and $\mathbf{T}=\{\mathbf{T}_j | j=1,\ldots,J\}$. A possible way to set the values of these parameters is by using maximum likelihood criterion, i.e. by maximizing the likelihood function [17]. This can be obtained iteratively through the Expectation Maximization (EM) [2, 4, 8]. Let $\mathbf{X}_N=\{\mathbf{x}_n | \mathbf{x}_n \in \Re^d, n=1,\ldots, N\}$ be a set of N independent and identically distributed (iid) data. Modeling $\mathbf{X}_N$ using $f_X(x)$ assumes that for each observation $\mathbf{x}_n$ there exists a hidden variable $\mathbf{z}_n$ denoting the component that generated $\mathbf{x}_n$. $\mathbf{z}_n$ can be represented as a *J*-dimensional binary vector such that if the j-*th* component is responsible for $\mathbf{x}_n$ then $\mathbf{z}_{nj}=1$, otherwise $\mathbf{z}_{nj}=0$. Let $\mathbf{Z}_N=\{\mathbf{z}_n | n=1,\ldots, N\}$ denote the set of these hidden variables. Hereinafter, $\{\mathbf{X}_N, \mathbf{Z}_N\}$ will be referred to as the complete data set, whereas we will refer to the actual observed data $\mathbf{X}_N$ as incomplete. The EM algorithm provides an iterative computation of maximum likelihood estimation when the observed data are incomplete. However, several limitations of EM approach can be highlighted. First of all, convergence to a global maximum is not guaranteed. In fact, for likelihood functions with multiple maxima, EM may converge, depending on starting values, to a local maximum. Another drawback of this approach for GMM training is that it cannot be used for determining the number of components during the estimation process. Furthermore, collapse of the PDF of one or more components onto a specific data point is likely to happen. When this occur, the component mean vector equals one of the data points, and the corresponding variance along some principal axis tends to zero, thus making the covariance matrix singular. This is the reason why EM is not suitable for assessing the number of components, for example, by starting with a large number of components and deleting the ones whose weights approach zero. In fact, components with low weights are associated to clusters with a few elements, which are likely to lead to singular covariance matrices. Possible solutions to the aforementioned issues may be obtained by adopting a Bayesian framework for the modeling the mixture.

Bayesian analysis treats parameters as random variables with a given prior probability distribution (hereinafter referred to as prior). Bayes's rule provides the framework for combining the prior information with sample data to make inferences about the model. It is worth noting that whereas in classical statistics all inferences are based on the sample data without using prior information, in the Bayesian framework the parameters of the distribution to be fitted are random variables. Therefore, this approach differs from the aforementioned EM approach in that parameters no longer appear because they are now stochastic variables and they are absorbed into latent variables. Typically, due the assumption of GMM for the data, conjugate priors from the exponential family are used for their mathematically tractability. Conjugate priors choice lead to posterior distributions having the same functional form as the prior, and, therefore, lead to a greatly simplified Bayesian analysis. That is why Dirichlet prior is used for $\boldsymbol{\pi}$, whereas an independent Gauss-Wishart prior is assumed for both $\boldsymbol{\mu}$ and $\mathbf{T}$ [2, 4]. The Dirichlet prior for $\boldsymbol{\pi}$ is given by:

$$f_{\pi}(\boldsymbol{\pi}) = Dir(\pi; \alpha_1, \ldots, \alpha_J) = C(\alpha_1, \ldots, \alpha_J) \prod_{j=1}^{J} \pi_j^{\alpha_j - 1} \quad (2)$$

where, by symmetry, the same α_j is chosen for each component, i.e. $\alpha_j=\alpha_0$ for $j=1,\ldots,J$, and $C(\alpha_1, \ldots, \alpha_J)$ is the normalization constant for the Dirichlet distribution. The Gauss-Wishart prior that governs the mean and the precision of each Gaussian component in equation (1) is given by:

$$f_{\mu,T}(\boldsymbol{\mu}, \mathbf{T}) = \prod_{j=1}^{J} N\left(\boldsymbol{\mu}_j | \mathbf{0}, (\beta \mathbf{T}_j)^{-1}\right) W(\mathbf{T}_j | \nu, \mathbf{V}) \quad (3)$$

which is the product of a Gaussian PDF and a Wishart PDF $W(\boldsymbol{T}_j|\nu,\mathrm{V})$, where parameters ν and $\boldsymbol{V}$ denote the degrees of freedom and the scale matrix, respectively. α_0, β, ν and $\boldsymbol{V}$ are called hyperparameters, and they have to be specified in advance. It is worth noting that Bayesian GMM allows for the optimal number of components to be determined: in fact, during the optimization procedure, as soon as one of the mixing coefficients converges to zero, the corresponding component is eliminated from the mixture. However, the Dirichlet prior does not allow the mixing weight of a component to become zero and, hence, the corresponding component to be eliminated from the mixture. Also, the final result depends on the hyperparameters of the priors.

A method that simultaneously trains the mixture, adjusts the number of components, and reduces the sensitivity to $\mathbf{V}$ was proposed in [12, 13]. To address the aforementioned issues, the method follows an incremental structure. It starts with one component and, progressively, adds components to

the model. The procedure for component addition is based on a splitting test applied to each of the existing mixture components. According to this test, a component is split into two sub-components ("free" components) and then variational Bayesian learning is applied to the specific pair of components, while the other components remain "fixed". In order to apply this method, priors on $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\mathbf{T}$ have to be imposed. Specifically, this approach assumes Gaussian and Wishart priors for $\boldsymbol{\mu}$ and $\mathbf{T}$, respectively:

$$f_{\mu}(\boldsymbol{\mu}) = \prod_{j=1}^{J} N(\boldsymbol{\mu}_j \mid \mathbf{0}, \beta\mathbf{I}) \tag{4}$$

$$f_{T}(\mathbf{T}) = \prod_{j=1}^{J} W(\mathbf{T}_j \mid \nu, \mathbf{V}). \tag{5}$$

It also fixes an uniform prior over the "free" components and a Dirichlet prior over the "fixed" components. These choices allow weights of the "free" components to become zero whereas the "fixed" components weights have zero probability to become zero. This Bayesian method for Gaussian mixture learning (hereinafter it will be referred to as Bayesian GMM Split, BGMMS) is fully automatic and does not depend on the initialization. It also allows the number of components for modeling the density shape to be determined automatically and, hence, resolves adequately the model selection problem, i.e. PDF approximation together with an automatic selection of the number of components.

III. ANOMALY DETECTION STRATEGY

The task of hyperspectral anomaly detection is to decide whether a target of interest is present or not in a pixel under test without a priori information about the spectral signature of the target. If the target class and the background class are both characterized by statistical models, the AD problem is typically formulated as a binary hypothesis testing:

$$\mathrm{H}(\mathbf{x}) = \begin{cases} \mathrm{H}_0 : & \mathbf{x} \text{ is a background pixel} \\ \mathrm{H}_1 : & \mathbf{x} \text{ is a target pixel} \end{cases} \tag{6}$$

where $\mathbf{x}$ is a realization of the random vector $\mathbf{X}$ used for modeling the pixel under test, whereas H_0 and H_1 denote the target absent and target present hypothesis, respectively.

The most widely adopted decision strategy derived from (6) is given by a Generalized Likelihood Ratio Test (GLRT) [17] depending on the PDFs conditioned on the two hypotheses. Specifically, the GLRT assumes that PDFs have a parametric form dependent on a set of unknown parameters $\{\theta_i\}$ i=0,1. It has been shown [1] that if θ_0 is estimated from a large sample of reference data, the GLRT is well approximated by the background likelihood [1, 18]:

$$\Lambda_{GLRT}(\mathbf{x}) = -\log\left\{f_{\mathbf{X}|H_0}(\mathbf{x}; \vartheta_0)\right\} \underset{H_1}{\overset{H_0}{\gtrless}} \eta \tag{7}$$

where $f_{\mathbf{X}|H_0}(\mathbf{x}; \vartheta_0)$ is the PDF of $\mathbf{x}$ under the null hypothesis, i.e. the background PDF, and η is the detection threshold. Any specification for the background PDF leads to a different detector.

In this work, the decision rule (7) is adopted for performing AD. Specifically, a GMM PDF with J components (as specified in equation (1)) is assumed for and its estimation is carried out within the Bayesian model selection framework described in section 2.

IV. EXPERIMENTAL DESIGN

In this section, the experiments carried out by applying the proposed AD strategy to a real hyperspectral data set are described.

A. Data set description

For testing and validating the proposed method on real hyperspectral data, an acquired at-sensor radiance image was utilized. The image was collected by the airborne hyperspectral sensor SIM-GA from a flight altitude of about 850 m, resulting in an approximate Ground Instantaneous Field of View (GIFOV) of 0.7 m. The acquired data span the VNIR spectral range (0.4-1 μm), with 512 spectral samples and an average spectral sampling of about 2 nm. In the scene, several different types of land-cover classes were observed, including grass, trees and roads. During measurement campaign, target panels were placed in the scene and a ground truth targets map was constructed. In order to perform experiments, a sub-image of size 560 x 280 pixels was extracted from the original image. A true-color representation of the resulting image can be seen in Fig. 1. A spectral binning, aimed at increasing Signal to Noise Ratio (SNR), was performed. Besides this, water-vapor absorption and noisy bands were discarded. Thus, a total of 89 spectral samples were used in this investigation.

B. Design of the experiments

Anomaly detection was conducted by adopting a GMM PDF, which was estimated through the BGMMS algorithm. Performance was compared to that obtained by employing the well-known EM approach. Actually, the EM approach was applied through a modified sequential version of the EM algorithm [7] (which will be referred to as sequential EM, S-EM). Here, S-EM incorporates an initial run of K-means [19] on few observations, in order to start the on-line estimation

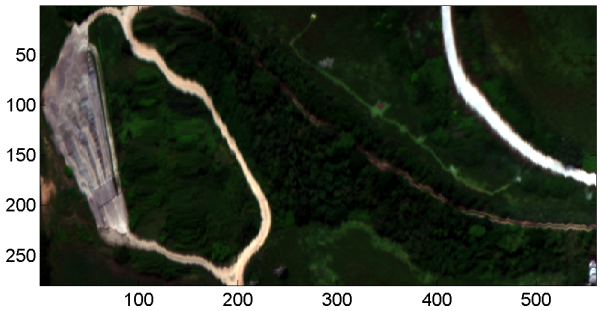

Fig. 1: True-color representation of the hyperspectral data employed.

with suitable initial parameters, and proceeds applying the on-line algorithm sequentially on the rest of observations. Using current parameters and new observations, parameters are updated recursively through the stochastic gradient algorithm.

The experimental analysis performed can be summarized as follows:

1. BGMMS and S-EM algorithms are implemented on a feature-reduced data set, in order to speed up computation. The procedure was performed on the first 10 components, extracted by the Principal Component Analysis (PCA) [10], which address 99.95% of the energy of the image. Results of GMM training are then collected in "cluster maps". In fact, fitting a mixture model to the distribution of the data can be interpreted as identifying clusters within the image. Cluster maps are constructed assigning each pixel to the GMM component with the maximum responsibility, i.e. the probability that one GMM component generated the n-*th* data vector.
2. Mixture model statistics ($\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\mathbf{T}$) are then estimated over all the 89 spectral samples according to the cluster map obtained.
3. Finally, according to equation (7), the detection statistic map is created to be thresholded for determining whether a given pixel is anomalous or not.

V. Experimental Results

In this section we discuss results obtained from the experiments described in the previous section.

In order to evaluate and compare the performance obtained, Receiver Operating Characteristic (ROC) curves [20] are employed, which plot the Fraction of Detected Target (FoDT), versus the False Alarm Rate (FAR), computed by varying the detection threshold. Only the pure target pixels are assumed to be actually the targets to be detected, i.e. in the calculation of the FAR the boundary target pixel mixed with the background were neglected.

BGMMS, applied by making no assumptions regarding the number of GMM components, provided a cluster map with 3 clusters (hence, 3 GMM components), which is shown in Fig. 2 (a), and which was employed for the AD purpose.

As regards S-EM, the number of components is a user-specified parameter that should be set according to the spectral diversity of the scene. Therefore, several configurations for this parameter (from 2 up to 17 components) were tested. All the resulting cluster maps were employed to perform AD, and the one yielding the best performance was selected for comparison with BGMMS-based detection. This was achieved on the basis of the Area Under the ROC curve (AUC), which is generated by calculating the area underneath the ROC curve, so that the larger the area, the better the performance. The AUCs calculated for each configuration are displayed in Fig. 3(a). As is observable, the best S-EM-based detection performance was obtained when using the 2-components cluster map, illustrated in Fig. 2 (b).

The detection maps resulting from the application of the AD strategy to the cluster maps obtained (i.e., the BGMMS map and the 2-components S-EM map) are shown in Fig. 2 (c) and (d), respectively. By thresholding of these detection maps, the ROC curves plotted in Fig. 3(b) were then obtained. These curves show that the proposed AD strategy,

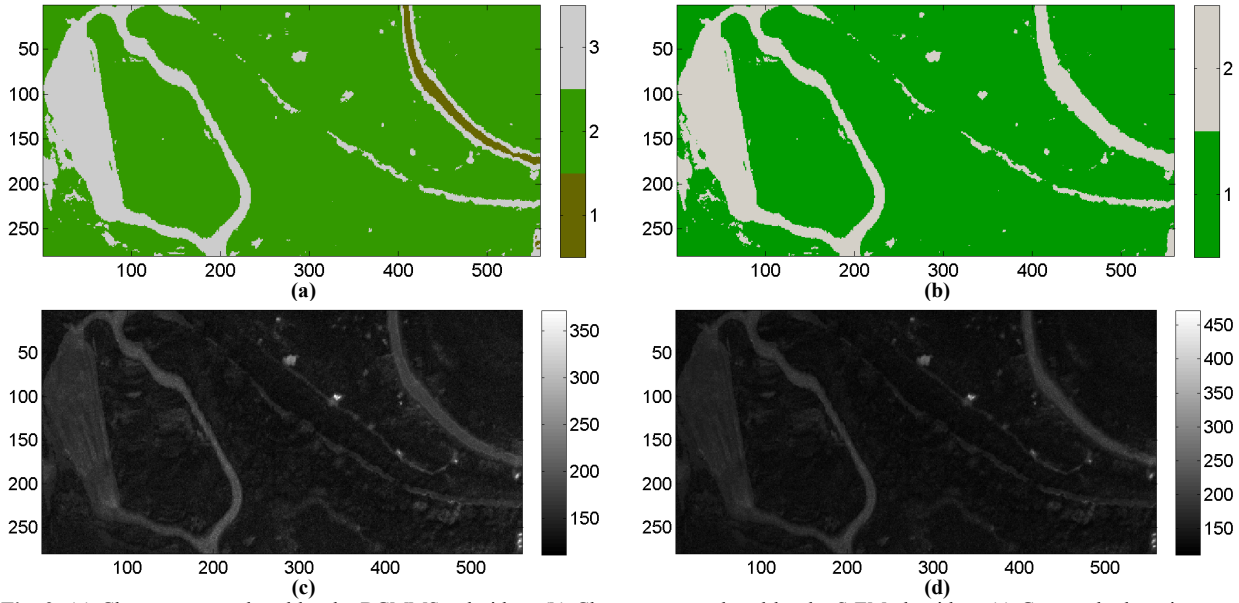

Fig. 2: **(a)** Cluster map produced by the BGMMS aglorithm. **(b)** Cluster map produced by the S-EM algorithm. **(c)** Grey-scale detection map obtained by using the AD algorithm based on BGMMS for mixture learning. **(d)** Grey-scale detection statistical map obtained by using the AD algorithm based on S-EM for mixture learning.

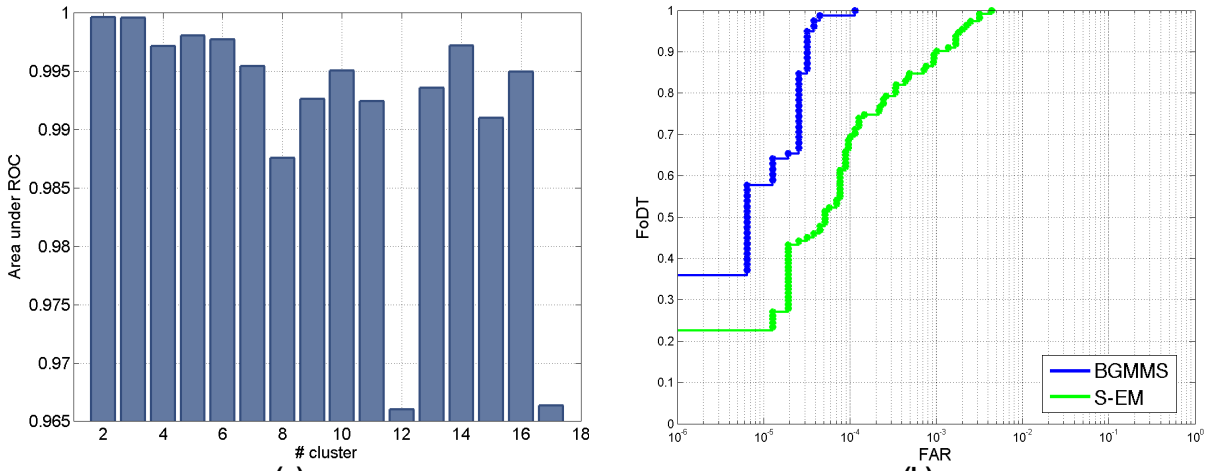

Fig. 3: **(a)** AUC vs number of clusters/GMM components. **(b)** Anomaly detection performace comparison. ROC curves for BGMMS-based AD (in blue) and S-EM-based AD (in green).

conducted by adopting BGMMS for learning the mixture model, performs significantly better than the approach based on classical EM learning. In fact, ROC curves exhibit high detection probabilities with low false alarm rates.

VI. CONCLUSIONS

In this paper, a new anomaly detection strategy for hyperspectral imagery based on a fully unsupervised Gaussian mixture learning has been presented.

The architecture of the proposed strategy combines a GMM learning for the background PDF along with a GLRT decision rule. Specifically, the GMM training is based on a recently proposed Bayesian technique that allows the GMM components to be automatically estimated during the learning procedure. The resulting AD strategy is, therefore, fully automatic and capable to adequately solve the model selection problem within the AD scheme.

During the experimental analysis, the BGMMS algorithm has managed to reliably estimate the background model that has been shown to be effective for detecting rare anomalous objects within the hyperspectral image employed. Furthermore, the algorithm has shown to reasonably estimate the correct number of GMM components, without producing significant over-segmentation. On this data, the proposed BGMMS-based anomaly detector has significantly outperformed the approach based on classical EM learning, even though this latter was tested with several configurations with respect to the number of GMM components. More importantly, the conducted analysis has highlighted how a discrete search over the number of components in a mixture distribution can be avoided by adopting a Bayesian philosophy within AD schemes. The results obtained confirm the benefits of the resulting AD strategy, whose potential deserves further investigation. A more detailed analysis, which includes also testing of different AD strategies, is still subject of ongoing work.

Future research will allow the actual effectiveness of Bayesian learning-based AD strategies to be assessed as regards detecting anomalous objects in hyperspectral images.

REFERENCES

[1] D.W.J. Stein, S.G. Beaven, L.E. Hoff, E.M. Winter, A.P. Shaum, and A.D. Stocker, "Anomaly detection from hyperspectral imagery", IEEE Signal Process. Mag., vol. 19, no.1, 2002, pp. 58-69.

[2] Bishop, C.M., Pattern recognition and machine learning, Springer, US, 2006.

[3] M.J. Carlotto, "A cluster-based approach for detecting man-made objects and changes in imagery", IEEE Trans. Geosci. Remote Sensing, vol. 43, no. 2, 2005, pp. 374–387.

[4] D.G. Tzikas, A.C. Likas, and N.P. Galatsanos, "The variational approximation for Bayesian inference", IEEE Sign. Process. Mag., 2008, pp. 131-146.

[5] P. Bradley, U. Fayyad, and C.R. Reina, "Clustering very large databases using EM mixture models", in Proc. 15th Internat. Conf. Pattern Recognit., vol. 2, 2000, pp. 76–80.

[6] M.A.T. Figueiredo, and A.K. Jain, "Unsupervised learning of finite mixture model", IEEE Trans. Pattern Anal. Mach. Intell, 2002, pp. 381-396.

[7] A. Samè, C. Ambroise, and G. Govaert, "An online classification EM algorithm based on the mixture model", Statistics and Computing, Vol. 17, No. 3 2007, pp. 209-218.

[8] T.K. Moon, "The Expectation-Maximization algorithm", IEEE Sig. Process. Mag., vol. 13, no. 6, 1996, pp. 47–60.

[9] E.A. Ashton, "Detection of subpixel anomalies in multispectral infrared imagery using an adaptive Bayesian classifier, IEEE Trans. Geosci. Remote Sensing, 1998, pp. 506-517.

[10] P. Hytla, R.C. Hardie, M.T. Eismann, and J. Meola, "Anomaly detection in hyperspectral imagery: a comparison of methods using seasonal data", in proc. SPIE, vol. 6565, 2007, pp. 566506-1-11 .

[11] S. Matteoli, F. Carnesecchi, M. Diani, G. Corsini, and L. Chiarantini, "Comparative analysis of hyperspectral anomaly detection strategies on a new high spatial and spectral resolution data set", in proc. SPIE, vol. 6748, 2007, pp. 67480E-1-11

[12] C. Constantinopoulos, and A. Likas, "Unsupervised learning of Gaussian mixtures based on varitional component splitting", IEEE Trans. Neural Netw., vol. 18, issue 3, 2007, pp. 745-755

[13] C. Constantinopoulos, and A. Likas,, "Image modeling and segmentation using incremental Bayesian mixture models", Proc. 12th Int.

Conf. Computer Analysis of Images and Patterns, Springer, 2007, pp. 596-603.

[14] C. Constantinopoulos, M.K. Titsias, and A. Likas, "Bayesian feature and model selection for Gaussian mixture models", IEEE Trans. Pattern Anal. Mach. Intell., 2006, pp. 1013-1018.

[15] A. Corduneanu, and C. Bishop, "Variational Bayesian model selection for mixture distribution", in Proc. 8th Internat. Conf. Artificial Intelligence and Statistics, 2001, pp. 27-34.

[16] M. Sato, "On-line model selection based on the variational Bayes", Neural Comput., vol. 13, no. 7, MIT Press, 2001, pp. 1649–1681.

[17] Kay, S.M., Fundamentals of Statistical Processing: Detection Theory, Prentice Hall, US, 1998.

[18] S. Matteoli, M. Diani, and G. Corsini, "A tutorial overview of anomaly detection in hyperspectral images", IEEE Aerosp. Electron. Syst. Mag., in press.

[19] Richards, J.A., and X. Jia, Remote sensing digital Image analysis: an introduction, Springer, Germany, 2005.

[20] T. Fawcett, "An introduction to ROC analysis", Pattern Recognit. Lett., Elsevier, 2005, pp. 861-874.