ON THE USE OF THE EM ALGORITHM FOR TRAINING A MAP CLASSIFIER

Ş. OPRIŞESCU, V. BUZULOIU *

Segmentarea imaginilor satelitare constă în două etape: segmentarea propriu-zisă, şi îmbunătăţirea segmentării, numită regularizare. Ambele etape sunt realizate utilizând „clasificatori" Bayesieni, care trebuie antrenaţi (parametrii care caracterizează respectivele modele statistice trebuie întâi estimaţi). Algoritmul EM este o tehnică de estimare statistică performantă în prelucrarea imaginilor satelitare, atât pentru zonele urbane cât şi pentru cele rurale, dacă se utilizează ca modele statistice mixturile Gaussiene multidimensionale. Această concluzie este susţinută de o analiză experimentală extinsă, folosind imagini satelitare reale.

Satellite image segmentation consists of two steps: the actual segmentation, and the improvement of the segmented image, called regularization. Both steps are performed by Bayesian "classifiers", which must be trained (that is, the parameters which characterize the corresponding statistical models must be estimated in advance). The EM algorithm is a powerful statistical estimation technique in satellite image analysis both for urban and rural areas, if the multivariate Gaussian mixtures are used as statistical models. This conclusion is supported by an extended experimental analysis using actual satellite images.

Keywords: EM algorithm, Gaussian mixtures, Bayesian segmentation.

Introduction

In image processing, the starting point is an actual image whose accuracy can be affected by several random or deterministic factors. The goal of processing is reaching a better fitting image with respect to its real origin. The satellite Earth observation supplies hyper-spectral images, which can be used for map drawings, landscape analysis, or crop supervision and, in most of the cases, the images arise from mixed zones, including rural and urban areas. For such urban & rural zones, the correct "reading" of the original landscape has the highest importance. Image processing consists of two stages, the segmentation of the actual image and the regularization of the segmented image. Both steps are performed by "classifiers", which provide the solutions of some optimization problems. Of course, these

* Ph.D. student, Prof., The Image Processing and Analysis Laboratory, University "Politehnica" of Bucharest, Romania

classifiers use statistical models, and they must be trained (that is, the parameters which characterize the models must be estimated in advance).

Traditionally, rural zones are modelled through 4-dimensional Gaussian distributions, when working with hyper-spectral images (4 band records) in the state space. Modelling the urban areas is much more difficult, as they consist of several materials representing a large source of variability of the reflectance ([1], [5], [11]). In a previous study ([9]) we have established the fact that, when an urban zone will be treated as a single object, modelling it through a 4-dimensional Gaussian *mixture* is well justified by the texture analysis. That is, the global texture parameters (homogeneity, contrast, uniformity, entropy, correlation, the Gauss-Markov entropy) are similar for the "urban spots" and the corresponding Gaussian mixtures.

Let us denote by $E = \mathbf{R}^N$ the space of states corresponding to N frequency bands. A hyper-spectral image S consisting of M pixels s can be represented as a point in the space E^M. Let us denote by $\mathbf{Y} = (Ys_1,\ldots,Ys_M)$ the random field associated with an image S and by $\mathbf{y} = (y_{s1},\ldots,y_{sM}) \in E^M$ a realization of $\mathbf{Y}$. Let Λ be the set of the K classes which are used as labels for the pixels in S. We denote by $\mathbf{X} = (Xs_1,\ldots,Xs_M)$ the random field which expresses the classification process and by $\mathbf{x} = (x_{s1},\ldots,x_{sM}) \in \Lambda^M$ a configuration (realization of the classification process) of the image S. The probability distribution of $\mathbf{Y}$ conditioned by a specified configuration, denoted $P(\mathbf{Y}|\mathbf{X}=\mathbf{x})$, is given through the conditional density $f(\mathbf{y}|\mathbf{X}=\mathbf{x})$.

The Bayesian segmentation uses the a posteriori probability for the configurations $\mathbf{x}$, given a realization of $\mathbf{Y}$ in the state space,

$$P(\boldsymbol{X}=\boldsymbol{x}\,|\,\boldsymbol{Y}=\boldsymbol{y}) = \frac{P(\boldsymbol{X}=\boldsymbol{x})\cdot f(\boldsymbol{y}\,|\,\boldsymbol{X}=\boldsymbol{x})}{\sum_{\boldsymbol{x}} P(\boldsymbol{X}=\boldsymbol{x})\cdot f(\boldsymbol{y}\,|\,\boldsymbol{X}=\boldsymbol{x})} \tag{1}$$

The MAP segmented image is the solution of the following optimization problem

$$\boldsymbol{x}^* = arg\ max\ P(\boldsymbol{X}=\boldsymbol{x}\,|\,\boldsymbol{Y}=\boldsymbol{y}) = arg\ max\ \{P(\boldsymbol{X}=\boldsymbol{x})\cdot f(\boldsymbol{y}\,|\,\boldsymbol{X}=\boldsymbol{x})\} \tag{2}$$

The basic hypothesis of the MAP method is the conditional independence of the observed states:

$$f(\boldsymbol{y}\,|\,\boldsymbol{X}=\boldsymbol{x}) = \prod_{s\in S} f(y_s\,|\,X_s=x_s)$$

In a first approximation (the segmentation process), the configurations $\mathbf{x}$ have equal probabilities,

$$P(\boldsymbol{X}=\boldsymbol{x}) = \frac{1}{K^M}$$

According to these hypotheses, the MAP configuration $\mathbf{x}^* = (x^*_{s1},\ldots,x^*_{sM})$ is the solution of a local optimization problem. For each pixel *s*, one has to solve the optimization problem

$$x_s^* = arg\ max\{f(y_s \mid X_s = x_1),...,f(y_s \mid X_s = x_K)\} \tag{3}$$

The MAP-Markov regularization is performed under the hypothesis that **X** is a Markov field on *S* with the neighbourhood system *N(S)*. That is, we assume that the following conditions are satisfied:

$$\forall x \in \Lambda^M, P(\boldsymbol{X} = \boldsymbol{x}) > 0$$
$$\forall i \in \{1,...,M\},\ P(X_{s_i} = x_{s_i} \mid \boldsymbol{x}_{(-i)}) = P(X_{s_i} = x_{s_i} \mid x_t,\ t \in N_{s_i}(S))$$

where $\mathbf{x}_{(-i)}$ is the (M-1)- dimensional configuration obtained after removing s_i. According to the Hamersley – Clifford theorem, the random field **X** is a Markov field on *S* with the neighbourhood system *N(S)* if and only if it is a Gibbs field on *S* with respect to *N(S)*. This means that, under the Markov assumption, the distribution P(**X=x**) must be a Gibbs distribution.

Let us denote by *C* the family of cliques corresponding to a neighbourhood system *N(S)*, by $U_c(\mathbf{x})$ the energy of the clique $c \in C$, and by $U(\mathbf{x})$ the total energy of a configuration **x**,

$$U(\boldsymbol{x}) = \sum_{c \in C} U_c(\boldsymbol{x})$$

The Gibbs field associated with a neighbourhood system *N(S)* is the random field **X** characterized by the Gibbs distribution

$$P(\boldsymbol{X} = \boldsymbol{x}) = Z^{-1} \cdot exp(-U(\boldsymbol{x}))$$

where *Z* is the normalization constant.

The MAP-Markov regularized image is the solution of the following optimization problem

$$\boldsymbol{x}^{**} = arg\ max\ \{ exp(-U(\boldsymbol{x})) \cdot f(\boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x})\} \tag{4}$$

Like in the segmentation stage, the solution $\mathbf{x}^{**}=(x^{**}_{s1},\ldots, x^{**}_{sM})$ is obtained by solving the corresponding local optimization problems,

$$x_s^{**} = arg\ max\{exp(-U(x_1))f(y_s \mid X_s = x_1),...,exp(-U(x_K))f(y_s \mid X_s = x_K)\} \tag{5}$$

A rural zone consists of a rather small number of *rural categories*, such as grain, forest, grassland, water etc. Each spot corresponding to a rural category is a homogenous one, and it can be represented by one class (color). Therefore, a rural area will be looked upon as a *composed zone*, consisting of several rural categories. Traditionally, rural categories are modelled through 4-dimensional

Gaussian distributions, when working with hyper-spectral images in the state space.

The urban zones consist of several materials representing a large source of variability of the reflectance. These materials could be typical for certain natural landscape, which means that some urban pixels could have the same reflectance value as the natural ones. Because of that, urban areas cannot be characterized through only grey level information, but they should be treated as textures and accordingly analyzed in order to discriminate between urban areas and non urban ones. In spite of its heterogeneity, an urban zone will be treated as a single object belonging to one and the same category, the *urban category*. There exist several texture parameters, either global or local, either empirical or based on statistical models, which allow differentiating between urban and rural. Any statistical model for urban area (when treated as a single object) should agree with the texture description offered by these parameters.

In [9] we have examined the possibility of modelling an urban category through a mixture of 4-dimensional Gaussian distributions. The positive answer to this issue has been obtained by studying the concordances of the texture parameters for an urban category and the associated Gaussian mixture.

A texture can be defined as "an attribute representing the spatial arrangement of gray levels of the pixels in a region" ([2]). A texture feature (or parameter) is a value, computed from the image of an object, that quantifies some characteristic of the gray-level variation within the object.

One of the most known texture analysis methods, gray level co-occurrence matrix (GLCM), estimates image properties related to second-order statistics. Based on GLCM, one calculates the following texture parameters: homogeneity, contrast, uniformity, entropy, correlation.

The Gauss-Markov model represents another well known approach for the characterization of textures. When considering V4 neighborhoods, the probability distribution of the 5-dimensional vector (one pixel and its neighbors) in the state space is a Gaussian one. The conditional distribution of the grey level of a pixel, given the values for its neighbors is Gaussian, one dimensional. The corresponding conditional variance is used as a texture parameter and it is called "temperature" ([5]). Another texture parameter, based on the Gauss-Markov model is the G-M entropy, which is the continuous entropy of the 5-dimensional Gaussian model.

In [9] we have established that the global texture parameters (homogeneity, contrast, uniformity, entropy, correlation, G-M entropy) are similar for the urban categories and the identified mixtures. This fact strongly supports our approach of modeling an urban area through a Gaussian mixture, when it is treated as a single object. While the variance for each pixel is the same for the real texture and the corresponding mixture, the parameter "temperature" is smaller for

urban textures than for Gaussian mixtures. Since the "temperature" is a local parameter, this mismatch doesn't encroach upon our model.

For a rural & urban satellite image, we assume that (K-1) classes correspond to some rural categories (grain, forest, grassland, water etc.), while the last class corresponds to an urban category. Then, for any pixel s, $f(y|X=x_i)$ are some *Gaussian probability densities* $N(4; \theta_i, \Phi_i)$ for i=1,…,(K-1), while $f(y|X=x_K)$ is *the probability density of a Gaussian mixture with m components*,

$$f(y \mid X = x_K) = \sum_{i=1}^{m} \alpha_i N(4; \mu_i, \Sigma_i)$$
$$0 < \alpha_i < 1 \; \forall i, \; \sum_{i=1}^{m} \alpha_i = 1 \tag{6}$$

In order to process the actual image (through segmentation and regularization), one has to estimate the parameters $\{\theta_i, \Phi_i, i=1,\ldots,(K-1)\}$ and $\{\alpha_i, \mu_i, \Sigma_i, i=1,\ldots,m\}$ in advance. That is, one has to train the classifier. Best training is reached when a ground truth is available but, most often, this is not the case. Therefore, one has to choose some "training windows" and implement some appropriate estimation techniques. The method we discuss is the EM algorithm, which proved itself a very reliable method, leading to accurate MAP-Markov regularized images.

1. The EM Algorithm

The formulation of the EM Algorithm in its present generality is due to Dempster A.P., Laird N.M., Rubin D.B. ([3], [7]). The EM Algorithm is a broadly applicable method that provides an iterative procedure for computing the Maximum Likelihood Estimation (MLE). The observed value of the random vector **Y**, denoted **y**, is viewed as being a vector of *incomplete-data*. Also, it is regarded as an observable function of the so called *complete data*. The notion of *incomplete data* includes the conventional sense of missing data, but it also applies to situations where the *complete data* represent what would be available from some hypothetical experiment. In the latter case, the complete data may contain some variables that are never observable in a data sense.

On each iteration, there are two steps: The **E-step** consists in manufacturing data for the complete-data problem, using the observed data set of the incomplete-data problem and the current value of the parameter. The **M-step** consists in the maximization of the log-likelihood of the complete-data problem. In fact, the log-likelihood is replaced by its conditional expectation given the observed data. Starting from suitable initial parameter values, the E and M steps are repeated until convergence.

Notation:
For the presentation of the algorithm we use the following notations:

- The unknown parameter $\boldsymbol{\psi} = (\psi_1, ..., \psi_d)^t \in \Omega \subseteq \mathbf{R}^d$;
- The incomplete-data random vector $\mathbf{Y}$, with the probability density function $g(\mathbf{y};\boldsymbol{\psi})$;
- For the observed $\mathbf{y}$, the likelihood function $L(\boldsymbol{\psi})=g(\mathbf{y};\boldsymbol{\psi})$;
- The score statistic for the incomplete-data $S(\mathbf{y};\boldsymbol{\Psi}) = \dfrac{\partial\, lnL(\boldsymbol{\Psi})}{\partial \boldsymbol{\Psi}}$;
- The complete-data random vector $(\mathbf{Y}^t,\mathbf{Z}^t)^t$, with the probability density function $g_c(\mathbf{y}, \mathbf{z}; \boldsymbol{\psi})$;
- For specified $(\mathbf{y},\mathbf{z})$, the likelihood function $L_c(\boldsymbol{\psi})=g_c(\mathbf{y}, \mathbf{z}; \boldsymbol{\psi})$;
- The score statistic for the complete-data $S_c(\mathbf{y},\mathbf{z};\boldsymbol{\Psi}) = \dfrac{\partial\, lnL_c(\boldsymbol{\Psi})}{\partial \boldsymbol{\Psi}}$;
- The current value of the parameter $\boldsymbol{\psi}^{(k)}$, k = 0, 1, ….
- The conditional expectation of the complete-data log-likelihood

$$Q\left(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(k)}\right) = E_{\boldsymbol{\Psi}^{(k)}}\left(lnL_c(\boldsymbol{\Psi}) \mid \mathbf{y}\right) \tag{7}$$

The EM Algorithm

- Let $\boldsymbol{\psi}^{(0)}$ be some initial value for $\boldsymbol{\psi}$
- The (k+1)-th iteration (k=1,2,....) consists of the following steps:

- *The E step*
  Estimate the non-observed data by taking
  $$\mathbf{z}^{(k)} = E_{\boldsymbol{\Psi}^{(k)}}\left(\mathbf{Z} \mid \mathbf{y}\right)$$
  Calculate
  $$Q\left(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(k)}\right) = E_{\boldsymbol{\Psi}^{(k)}}\left(lnL_c(\boldsymbol{\Psi}) \mid \mathbf{y}\right)$$
- *The M step*
  Choose $\boldsymbol{\psi}^{(k+1)}$ so that
  $$Q\left(\boldsymbol{\Psi}^{(k+1)},\boldsymbol{\Psi}^{(k)}\right) \geq Q\left(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(k)}\right), \forall \boldsymbol{\Psi} \in \boldsymbol{\Omega}$$

Dempster, Laird and Rubin ([3]) have proved that the incomplete-data likelihood function $L(\boldsymbol{\psi})$ is not decreased after an EM iteration, as well as the convergence towards a stationary point.

Proposition *The incomplete-data likelihood function L(ψ) is not decreased after an EM iteration,*

$$L(\boldsymbol{\Psi}^{(k+1)}) \geq L(\boldsymbol{\Psi}^{(k)}), k = 0,1,2,...$$

Hence, the convergence must be obtained with a sequence of likelihood values that are bounded above.

The E and M steps are alternated repeatedly until the difference $L(\boldsymbol{\psi}^{(k+1)}) - L(\boldsymbol{\psi}^{(k)})$ changes by an arbitrarily small amount.

The nature of the limit is described in the following propositions.

Proposition *Suppose that $Q(\boldsymbol{\psi},\boldsymbol{\varphi})$ is continuous in both $\boldsymbol{\psi}$ and $\boldsymbol{\varphi}$. Then all the limit points of any sequence of EM iterates $(\boldsymbol{\psi}^{(k)})_k$ (for any initial value $\boldsymbol{\psi}^{(0)}$) are stationary points of $L(\boldsymbol{\psi})$, and the sequence $(L(\boldsymbol{\psi}^{(k)}))_k$ converges monotonically to some value $L^*=L(\boldsymbol{\psi}^*)$, for some stationary point $\boldsymbol{\psi}^*$.*

Proposition *Suppose that the likelihood function for the incomplete data, $L(\boldsymbol{\psi})$, is unimodal, with $\boldsymbol{\psi}^*$ being the only stationary point and that $\partial Q(\boldsymbol{\psi},\boldsymbol{\varphi})/\partial\boldsymbol{\psi}$ is continuous in both $\boldsymbol{\psi}$ and $\boldsymbol{\varphi}$. Then, any sequence of EM iterates $(\boldsymbol{\psi}^{(k)})_k$ (for any initial value $\boldsymbol{\psi}^{(0)}$) converges to the unique maximizer $\boldsymbol{\psi}^*$ of $L(\boldsymbol{\psi})$; that is, it converges to the unique MLE of $\boldsymbol{\psi}$.*

The proofs of these propositions can be found in Mc Lachlan and Krishnan ([7]).

The case of multivariate Gaussian mixtures

Let Y be a random vector with the probability density given by a mixture of m Gaussian, N-dimensional distributions $N(\mathrm{N};\ \boldsymbol{\mu}_i,\ \Sigma_i)$. For each component i of the mixture, the mean vector is $\boldsymbol{\mu}_i \in \mathbf{R}^N$, and Σ_i is a symmetrical, positive defined matrix, of dimension N×N.

$$f(y;\boldsymbol{\Psi}) = \sum_{i=1}^{m} \alpha_i f(y;\boldsymbol{\mu}_i,\Sigma_i)$$

$$= \sum_{i=1}^{m} \frac{\alpha_i}{\sqrt{(2\pi)^N \det\Sigma_i}} exp\left(-\frac{1}{2}(y-\boldsymbol{\mu}_i)^t \Sigma_i^{-1}(y-\boldsymbol{\mu}_i)\right),$$

$$\alpha_i \in (0,1), i = 1,...,m, \sum_{i=1}^{m} \alpha_i = 1.$$

The parameter, written in a vector form, is

$$\boldsymbol{\Psi} = (\alpha_1,...,\alpha_m,\boldsymbol{\mu}_1,...,\boldsymbol{\mu}_m,\Sigma_1,...,\Sigma_m)^t.$$

We consider n independent, identical distributed random vectors, $Y_1,...,Y_n$, denote $\mathbf{Y}=(Y_1^t,...,Y_n^t)^t$, and denote by $\mathbf{y} = (y_1^t,...,y_n^t)^t$ the observed data. Then

$$g(\mathbf{y};\boldsymbol{\Psi}) = \prod_{j=1}^{n} f(y_j;\boldsymbol{\Psi})$$

and, for the observed $\mathbf{y}$, the likelihood function is $L(\boldsymbol{\psi})=g(\mathbf{y};\boldsymbol{\psi})$.

The *missing-data* are the indicator variables

$$Z_{ij} = \begin{cases} 1, \text{ if } y_j \text{ arouse from } f(y; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \\ 0, \text{ otherwise} \end{cases} \tag{8}$$

We consider the random vectors $\mathbf{Z}_j = (Z_{1j}, ..., Z_{mj})^t$, $j = 1,...,n$, $\mathbf{Z} = (\mathbf{Z}_1^t, ..., \mathbf{Z}_n^t)^t$.
The *complete-data*, denoted $(\mathbf{y}^t, \mathbf{z}^t)^t$, are the values of the random vector $(\mathbf{Y}^t, \mathbf{Z}^t)^t$.

The EM Algorithm

- The initial value: $\boldsymbol{\psi}^{(0)} = (\boldsymbol{\mu}_1^{(0)t}, ..., \boldsymbol{\mu}_m^{(0)t}, \Sigma_1^{(0)}, ..., \Sigma_m^{(0)}, \alpha_1^{(0)}, ..., \alpha_m^{(0)})^t$
- The (k+1)-th iteration, the E-step:

$$z_{ij}^{(k)} = \frac{\alpha_i^{(k)} \frac{1}{\sqrt{det\left(\Sigma_i^{(k)}\right)}} \cdot exp\left(-\frac{1}{2} \left(y_j - \boldsymbol{\mu}_i^{(k)} \right)^t \left(\Sigma_i^{(k)} \right)^{-1} \left(y_j - \boldsymbol{\mu}_i^{(k)} \right) \right)}{\sum_{i=1}^{m} \alpha_i^{(k)} \frac{1}{\sqrt{det\left(\Sigma_i^{(k)}\right)}} \cdot exp\left(-\frac{1}{2} \left(y_j - \boldsymbol{\mu}_i^{(k)} \right)^t \left(\Sigma_i^{(k)} \right)^{-1} \left(y_j - \boldsymbol{\mu}_i^{(k)} \right) \right)} \tag{9}$$

  for i =1, ..., m, j = 1, ..., n.

- The (k+1)-th iteration, the M-step:

$$\alpha_i^{(k+1)} = \frac{1}{n} \sum_{j=1}^{n} z_{ij}^{(k)}, \ i = 1,...,m, \tag{10}$$

$$\mu_i^{(k+1)} = \frac{1}{\sum_{j=1}^{n} z_{ij}^{(k)}} \sum_{j=1}^{n} z_{ij}^{(k)} y_j, \ i = 1,...,m, \tag{11}$$

$$\Sigma_i^{(k+1)} = \sum_{j=1}^{n} z_{ij}^{(k)} \left(y_j - \boldsymbol{\mu}_i^{(k+1)} \right) \left(y_j - \boldsymbol{\mu}_i^{(k+1)} \right)^t \Big/ \sum_{j=1}^{n} z_{ij}^{(k)}, \ i = 1,...,m. \tag{12}$$

The iterative process stops when

$$L(\boldsymbol{\Psi}^{(k+1)}) - L(\boldsymbol{\Psi}^{(k)}) < threshold$$

or when

$$\left\| \boldsymbol{\Psi}^{(k+1)} - \boldsymbol{\Psi}^{(k)} \right\|^2 < threshold$$

Remark *In the studied case, the probability density $f(y;\boldsymbol{\psi})$ is a finite linear combination of Gaussian densities, hence it is a bounded function. It follows that the sequence $(L(\boldsymbol{\psi}^{(k)}))_k$ is bounded, hence its convergence is ensured.*

$$L(\boldsymbol{\Psi}^{(k)}) = \prod_{j=1}^{n}\left(\sum_{i=1}^{m}\frac{\alpha_i^{(k)}}{\sqrt{(2\pi)^N det\Sigma_i^{(k)}}}exp\left(-\frac{1}{2}\left(y_j-\boldsymbol{\mu}_i^{(k)}\right)^t\left(\Sigma_i^{(k)}\right)^{-1}\left(y_j-\boldsymbol{\mu}_i^{(k)}\right)\right)\right)$$

Remark *The complete-data log-likelihood function has the expression*

$$L_c(\boldsymbol{\Psi}) = \prod_{i=1}^{m}\prod_{j=1}^{n}\left(\alpha_i \cdot f\left(y_j;\boldsymbol{\mu}_i,\Sigma_i\right)\right)^{z_{ij}}$$

The conditional expectation of the complete-data log-likelihood, $Q(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(k)}) = E_{\boldsymbol{\Psi}^{(k)}}\left(lnL_c(\boldsymbol{\Psi}) \mid \boldsymbol{y}\right)$, becomes equal to the expression

$$\sum_{i=1}^{m}\left(\sum_{j=1}^{n}z_{ij}^{(k)}\right)ln\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}z_{ij}^{(k)}\left(ln(det\Sigma_i) + (y_j-\boldsymbol{\mu}_i)^t\Sigma_i^{-1}(y_j-\boldsymbol{\mu}_i)\right) - C,$$

where $C = (1/2)nN \cdot ln(2\pi)$. We notice that $Q(\boldsymbol{\psi},\boldsymbol{\psi}^{(k)})$ is a continuous function with respect to both variables, $\boldsymbol{\psi}$ and $\boldsymbol{\psi}^{(k)}$. This fact implies the convergence of $(L(\boldsymbol{\psi}^{(k)}))_k$ towards $L^*=L(\boldsymbol{\psi}^*)$, for some stationary point $\boldsymbol{\psi}^*$.

Training a MAP Classifier

Training the classification system consists in the estimation of the parameters of the involved statistical models. The estimators of the parameters for urban areas (modelled through Gaussian mixtures) are obtained by the EM algorithm. The estimators of the parameters for rural areas should be constructed by direct statistical estimation, using an appropriate ground truth. When such a ground truth is not available, an appropriate EM algorithm will be used for estimation, on the basis of a rural composed zone.

We have considered several mixed, urban & rural hyper-spectral images, with 4 bands, of free Internet access [12]. The parameters which characterize the statistical models must be estimated by using some appropriate learning windows in order to train the classification system. We have developed an algorithm for training the classification system, by means of the EM Algorithm.

EML algorithm (Estimation-Maximization-Learning)

It estimates (learns) the parameters of an urban or rural zone, by means of an appropriate EM algorithm

- The input image is a portion of a real, 4 bands SPOT 4 satellite image. It consists of *n* pixels;
- Choose the maximal number of classes of the Gaussian mixture;
- Initialize the parameter $\boldsymbol{\psi}$;

- Iteratively compute $\mathbf{Z}^{(k)}, \boldsymbol{\psi}^{(k+1)}$ on using formulae (9), (10), (11) and (12) for equal covariance matrices, $\Sigma_i=\Sigma$, until

$$\left\|\boldsymbol{\alpha}^{(k+1)}-\boldsymbol{\alpha}^{(k)}\right\|^2+\sum_{i=1}^{m}\left\|\boldsymbol{\mu}_i^{(k+1)}-\boldsymbol{\mu}_i^{(k)}\right\|^2<0.001$$

- Output: $\boldsymbol{\psi}^{(k+1)}$.

We exemplify the above algorithms on the image Fig. 1, which involves both rural and urban areas.

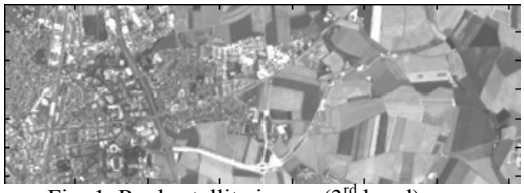

Fig. 1. Real satellite image (3rd band)

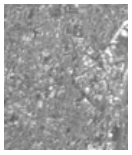

Fig. 2. (a) Urban learning window (b) Rural learning window

Learning the parameters for the urban area has been achieved by the EML algorithm and *eight* components have been retained. The corresponding learning window is presented in Fig. 2. (a). The Table 1 contains some of the estimated parameters (the values α_i, where *i* is the index of the identified component, and the mean vectors $\mu(Bk)$, where *k* is the index of the spectral band).

Table 1

EML-Estimated Parameters for Urban Area

Comp.	α_i	μ(B1)	μ(B2)	M(B3)	μ(B4)
1	0.62	75.3	157.56	145.45	113.46
2	0.1035	104.87	124.23	125.5	112.52
3	0.1026	81.73	192.47	178.66	120
4	0.0411	91.7	181.48	176.38	143.98
5	0.0359	68.63	151.64	138.13	125.28
6	0.0262	83.65	190.23	161.57	121.72
7	0.0216	95.31	221.78	214.43	137.24
8	0.0075	122.34	236.97	240.93	183.95

Learning the parameters for the rural area has been achieved by the EML algorithm applied to a composed rural image, and *seven* different classes have been retained. But, as we mentioned, we don't have a ground truth, so we cannot make the correspondence between these classes and different crop types. The

corresponding learning window is presented in Fig. 2. (b). The *table 2* contains the mean vectors *θ(Bk)*, where *k* is the index of the spectral band.

Table 2

EML-Estimated Parameters for Rural Area

Classes	θ(B1)	θ(B2)	θ(B3)	θ(B4)
1	105.63	231.22	177.79	161.61
2	183.35	87.69	117.21	88.56
3	81.51	174.31	147.68	129.15
4	131.77	250.19	210.26	193.52
5	150.94	114.59	127.09	115.62
6	190.95	77.08	107.42	118.98
7	123.27	157.22	145.21	131.83

Remark *In the EML algorithm, the initial value of the parameter,* $\boldsymbol{\psi}^{(0)}$*, was chosen in the most likely region of the parameter space (*$\alpha_i^{(0)}=1/m$ *for* $i=1,...,m$*,* $\boldsymbol{\mu}_i^{(0)}$ *were chosen such that they cover the interval of grey levels (0,255), and* $\Sigma^{(0)}$ *was taken equal to the covariance matrix calculated for the whole actual image).*

Remark *The convergence of the EM algorithm was obtained in a rather small number of steps (minimum 7 steps, maximum 100 steps).*

The quality of the trained classifier was validated through the MAP segmentation (by solving the optimization problem (3)) and the MAP-Markov regularization (by solving the optimization problem (5)). The final image is presented in Fig. 3.

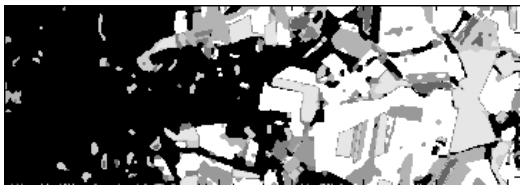

Fig. 3. Final segmented & regularized image.

The experiments have been performed using either the Matlab programming environment or C++.

Conclusions

We have detailed in this paper the construction of the steps of an EM algorithm for training the classification of both rural and urban areas in satellite images.

In [9] we've established that, when an urban zone will be treated as a single object, modeling it through Gaussian mixtures is well justified by the

texture analysis. According to this model, the EM algorithm for a Gaussian mixture is the natural approach for learning the parameters of an urban zone.

For a rural zone, the traditional training process requires a ground truth, which is not available in most of the cases. We use an adapted EM algorithm for learning the parameters of a rural zone and apply it to a representative, composed, rural training window.

The performances of the trained classifier are fully confirmed by the MAP segmentation and the MAP-Markov regularization of the image. On the basis of our study, we can conclude that the EM algorithm is suitable for the training stage. It is a powerful statistical estimation technique in image analysis, which can be successfully used in training the classification system both for urban and rural areas.

As a final conclusion, we consider that the use of the Gaussian mixtures in the modeling process, the use of the EM algorithm in the training stage, and the implementation of MAP segmentation and MAP-Markov regularization offer very good results in the processing of mixed, rural & urban satellite images.

R E F E R E N C E S

1. *A. Baraldi and F. Parmiggiani*, "Urban area classification by multispectral spot images", IEEE Trans. on Geoscience and Remote Sensing, **vol. 28**, No.4, pp. 674-680, 1990
2. *K.R. Castleman*, Digital image processing, New Jersey, Englewood Cliffs, Prentice Hall, 1996.
3. *A.P. Dempster, N.M. Laird and D.B. Rubin*, "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, B, 39, pp. 1-38, 1977
4. *H. Derin and H. Elliott*, "Modeling and segmentation of noisy and textured images using Gibbs random fields", IEEE. Trans. on Pattern Anal. and Mach. Intell., No.9, pp. 721-741, 1987
5. *X. Descombes*, Champs markovien en analyse d'images, Thèse de Doctorat, ENST, 1993
6. *F. Lagona*, "Adjacency selection in Markov random fields for high spatial resolution hyperspectral data", J. Geograph. Syst., **vol.4**, pp. 53-68, 2002.
7. *G. J. Mc Lachlan and T. Krishnan*, The EM Algorithm and Extensions, New York, John Wiley & Sons, 1997.
8. *S. Oprisescu and M. Dumitrescu*, "On the Regularization of Segmented Satellite Images", Proceedings of the 7th IEEE - ISSCS Conference, Iasi, Romania, **Vol. 1**, pp.83-86, 2005
9. *S. Oprisescu, V. Buzuloiu and M. Dumitrescu*, "Statistical modelling of satellite images, for MAP segmentation", to be presented at "Communications 2006" The 6th IEEE Communications International Conference, Bucureşti, June 2006.
10. *D. K. Panjwani and G. Healey*, "Markov random field models for unsupervised segmentation of textured color images", IEEE, Trans. on Pattern Anal. Machine Intell., **Vol.17**, No.10, 1995, pp.939-954.
11. *O. Pony, X. Descombes and J. Zerubia*, "Classification d'images satellitaires hyperspectrales en zone rurale et periurbaine", Research Report, INRIA, Sophia Antipolis, No. 4008, September 2000.
12. http://www.spotimage.fr/education.html