Estimation of Fuzzy Gaussian Mixture and Unsupervised Statistical Image Segmentation

Hélène Caillol, Wojciech Pieczynski, and Alain Hillion, *Associate Member, IEEE*

***Abstract*—This paper addresses the estimation of fuzzy Gaussian distribution mixture with applications to unsupervised statistical fuzzy image segmentation. In a general way, the fuzzy approach enriches the current statistical models by adding a fuzzy class, which has several interpretations in signal processing. One such interpretation in image segmentation is the simultaneous appearance of several thematic classes on the same site. We introduce a new procedure for estimating of fuzzy mixtures, which is an adaptation of the iterative conditional estimation (ICE) algorithm to the fuzzy framework. We first describe the blind estimation, i.e., without taking into account any spatial information, valid in any context of independent noisy observations. Then we introduce, in a manner analogous to classical hard segmentation, the spatial information by two different approaches: contextual segmentation and adaptive blind segmentation. In the first case, the spatial information is taken into account at the segmentation step level, and in the second case it is taken into account at the parameter estimation step level. The results obtained with the iterative conditional estimation algorithm are compared to those obtained with expectation-maximization (EM) and the stochastic EM (SEM) algorithms, on both parameter estimation and unsupervised segmentation levels, via simulations. The methods proposed appear as complementary to the fuzzy C-means algorithms.**

I. Introduction

The statistical approach to the image segmentation problem requires modeling two random fields. For $S = \{1, \cdots, n\}$ the set of pixels, $\zeta = (\zeta_s)_{s\in S}$ is the unobservable random field whose realizations are the true nature of the observed scene, and $X = (X_s)_{s\in S}$ is the observed random field, which is seen as a corrupted version of ζ and corresponds to the intensity of the observation. The random variables ζ_s take their values in a set of thematic classes denoted Ω. This field is usually assumed discrete. As the classes are numbered from 1 to e, we will denote $\Omega = \{\omega_1, \cdots, \omega_e\}$. In the case of satellite data, ζ models the true nature of the ground in such a way that the classes $(\omega_i)_{1\le i\le e}$ are, for instance, water, forest, urban area, and so forth. A more general way to consider this problem is the fuzzy approach. From this point of view, each pixel s is associated with an e-dimensional vector $u(s) = [u_{is}]_{1\le i\le e}$. Roughly speaking, the grade of membership of the sth pixel to the ith class, denoted u_{is}, is the proportion of area s belonging to class ω_i. In [23], Pedrycz wrote a survey on the use of fuzzy representation in pattern recognition with a sizeable bibliography.

When considering the probabilistic framework of interest here, random variables $\zeta = (\zeta_s)_{s\in S}$ become random vectors. Thus, for each pixel s, $\zeta_s = [\zeta_{is}]_{1\le i\le e}$. For the sake of simplicity, we confine our study to the case $e = 2$ and put $\zeta_s = \zeta_{1s} = 1 - \zeta_{2s}$. In the example of satellite data to be treated subsequently, the fuzzy statistical model allows one to take account of pixels in which, for instance, water and forest are simultaneously present. To be more precise, let the class "water" be associated with the numerical value 0 and the class "forest" with the value 1. Since a classical model may assume that $\Omega = \{0, 1\}$, the fuzzy model assumes that two different cases are possible for each pixel s, as follows.

- If s belongs to either of both hard classes, then $\zeta_s = 0$ if s belongs to class "water" or $\zeta_s = 1$ if s belongs to class "forest." This kind of pixel will be called a *pure* pixel.
- If the both hard classes are present in the pixel s, then $\zeta_s = \varepsilon$, where ε is a real value in]0, 1[which represents the degree of membership of s to the class "forest." Thus, $1 - \varepsilon$ is the degree of membership of s to the class "water." In this case, the pixel s will be called a *mixed* pixel.

According to the model proposed in [4], these cases are expressed by two types of components in the distribution of ζ_s: a hard component modeled by two Dirac weights in 0 and 1, and a fuzzy component defined by a density with respect to the Lebesgue measure. Thus, the hard component corresponds to the pure pixels and the fuzzy component corresponds to the mixed pixels. Such a "blind" model, i.e., only using the marginal distributions of the both random fields, can be successfully used to perform the blind unsupervised segmentation. "Blind" means that no spatial information is taken into account, and "unsupervised" means that all parameters needed are estimated from the noisy data.

The present paper extends the work presented in [4] in three directions:

1) we generalize the blind unsupervised statistical fuzzy segmentation to a contextual setting;
2) we adapt iterative conditional estimation (ICE) [3], [25], a recent general method of estimation in the case of hidden data, to the model proposed and show that in some situations the analogous adaptations of expectation-maximization (EM) [10], [27] and stochastic

Manuscript received January 24, 1995; revised March 4, 1996. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. William E. Higgins.

H. Caillol and W. Pieczynski are with the Département Signal et Image, Institut National des Télécommunications, 91011 Evry Cedex, France (e-mail: wojciech.pieczynski@int-evry.fr).

A. Hillion is with the Département Images et Traitement de l'Information, École Nationale Supérieure des Télécommunications, 29285 Brest Cedex, France.

Publisher Item Identifier S 1057-7149(97)00335-7.

EM (SEM) [6], {22], are inefficient and ICE has to be used;

3) we propose adaptive unsupervised segmentation, which is extremely efficient in some situations [24], to our fuzzy model.

Different methods are compared via simulation study and unsupervised fuzzy segmentations of real images are presented. Let us note that the originality of the model we use lies in the inclusion of Dirac weights, which allows the simultaneous existence of pure and mixed pixels. Indeed, different stochastic or deterministic image models using fuzzy membership previously proposed involve only the presence of the mixed pixels. This is not necessarily a serious drawback; in fact, hard pixels can be produced by some "hardening" procedure. However, the conceptual originality of our model implies the complementarity, with respect to the existing methods, of the methods involved.

Our works deal with local segmentation methods and it is well known in the hard framework that global methods, i.e., methods based on hidden Markov models, can be much more efficient. However, we believe that the study of local fuzzy segmentation methods is of interest for two reasons. First, local hard methods can be competitive with respect to global hard ones in several particular situations [3] and, thus, the same is true for fuzzy methods, at least when there are few fuzzy pixels. Second, local fuzzy methods present the following advantage with respect to the corresponding global fuzzy methods recently proposed [26], [28]: The hard local model is really a particular case of the fuzzy one in the sense that it corresponds to a particular value of some parameter. This is not the case in the global framework: A hard hidden Markov field can only be obtained from a fuzzy hidden Markov field when some parameter tends to infinity. Thus, when the real class image is hard, one can use the fuzzy local model because the parameter estimation step should make it hard automatically and such automated adaptation of the fuzzy model to the hard reality is undoubtedly more problematic in the global case.

The paper is organized as follows. Section II explains the principle of the ICE procedure and its implementation in blind, contextual, and adaptive cases. Section III presents numerical comparisons between ICE and the SEM and EM algorithms. Section IV is devoted to unsupervised fuzzy segmentation based on the preceding estimations. The final section contains conclusions and future prospects.

II. The ICE Algorithm

A. Principle of the ICE Algorithm

In a general manner, let us consider a pair of random variables (ζ, X) whose distribution depends on a parameter δ. The problem is to estimate δ from X. The idea behind the ICE procedure is the following: The complexity of the estimation problem is due to the absence of an observation of ζ. If ζ were observable, one could generally use some efficient parameter estimation procedure. Indeed, if the estimation of δ from (ζ, X) is impossible there is no sense in estimating it from the X data alone. So let us suppose temporarily that ζ is observable and let us consider an estimator $\hat{\delta} = \hat{\delta}(\zeta, X)$, defined from (ζ, X), of the parameter δ. In a general manner, if we want to approximate a random variable Z by some function of a random variable W, the best approximation, when the squared error is concerned, is the conditional expectation. To be more precise, if we denote the conditional expectation by $E[Z/W]$, we have

$$E[Z - E(Z/W)]^2 = \min_{\varphi} E[Z - \varphi(W)]^2. \tag{1}$$

Considering the problem of constructing $\hat{\delta}$ using X alone, one can thus consider $E[\hat{\delta}(\zeta, X)/X]$. The problem is that $E[\hat{\delta}(\zeta, X)/X]$ depends, in a general case, on δ, and is then no longer an estimator.

Thus, let us denote $E_\delta[\ /X]$ the conditional expectation based on δ. It is then possible to define an iterative procedure, ICE [3], [25], using an initial value δ^0 of δ and putting

$$\delta^{k+1}(X) = E_{\delta^k}\left[\hat{\delta}(\zeta, X)/X\right]. \tag{2}$$

When $E[\hat{\delta}(\zeta, X)/X]$ is not computable but samplings of ζ according to the distribution conditional on $X = x$ are possible, one can use a stochastic approximation. In fact, the conditional expectation at the point $X = x$ is the expectation according to the distribution of ζ conditional on $X = x$. Thus, it can be approached, by virtue of the law of large numbers, by the empirical mean. After having sampled N realizations $\varepsilon_1, \cdots, \varepsilon_N$ of ζ according to its distribution conditioned on $X = x$, we can put

$$\delta^{k+1} = \frac{1}{N}\,[\hat{\delta}(\varepsilon_1, x) + \cdots + \hat{\delta}(\varepsilon_N, x)]. \tag{3}$$

Thus ICE appears as an alternative to the EM algorithm. Unfortunately, the theoretical study of the ICE seems difficult and no relevant results can be proposed at present. This could be due to the fact that the sequence produced by ICE depends on the parameterization, which means that for a given problem ICE gives a family of different methods. In order to illustrate this fact, let us shortly discuss the differences between the two methods in the case of a simple mixture of Gaussian distributions $N(m_0, 1)$ and $N(m_1, 1)$. The parameter to be estimated is $\delta = (\alpha, m_0, m_1)$, where α and $1 - \alpha$ are priors.

The sample considered is denoted by $x_1, \cdots, x_n$. For $\zeta = (\zeta_1, \cdots, \zeta_n)$ let us put $U = 1_{[\zeta_1 = \omega_0]} + \cdots + 1_{[\zeta_n = \omega_0]}$, $V = X_1 1_{[\zeta_1 = \omega_0]} + \cdots + X_n 1_{[\zeta_n = \omega_0]}$, and $\pi_{1,k}, \cdots, \pi_{n,k}$, the δ^k based distributions $P[\zeta_1 = \omega_0/X_1 = x_1], \cdots, P[\zeta_n = \omega_0/X_n = x_n]$. The EM reestimation formulas are

$$\alpha_{k+1} = \frac{1}{n}\,(\pi_{1,k} + \cdots + \pi_{n,k}) \tag{4}$$

$$m_{0,k+1} = \frac{x_1\pi_{1,k} + \cdots + x_n\pi_{n,k}}{\pi_{1,k} + \cdots + \pi_{n,k}}. \tag{5}$$

If $\zeta = (\zeta_1, \cdots, \zeta_n)$ were observable, α could be estimated by $\hat{\alpha}(\zeta) = U/n$ and m_0 by $\hat{m}_0(\zeta, X) = V/U$. According to the ICE principle, we have to take the conditional expectation of these two estimators. In the case of $\hat{\alpha}(\zeta)$, one obtains (4), i.e., the same formula as in the EM case. Taking the conditional expectation of $\hat{m}_0(\zeta, X) = V/U$ is not feasible and one has to resort to the stochastic approximation. Now, let us consider the changing of parameter $\lambda = \varphi(\delta)$, with $\lambda = (\lambda_1, \lambda_2, \lambda_3) = (\alpha, \alpha/m_0, \alpha/m_1)$. It is possible to show that if $\delta^0, \cdots, \delta^k, \cdots$ is the sequence produced by the EM using δ, then $\lambda^0 = \varphi(\delta^0), \cdots, \lambda^k = \varphi(\delta^k), \cdots$ is the sequence produced by the EM using $\lambda = \varphi(\delta)$. This is not true in the ICE case: taking $\hat{\lambda}_1(\zeta) = U/n$, and $\hat{\lambda}_2(\zeta, X) = V/n$ ICE produces the same sequence as EM. In other words, EM does not depend on the parameter used, while ICE does. Delmas has shown in a recent paper [9] that this holds true in the general exponential structure.

B. Implementation of ICE in a Fuzzy Context

Two tools are thus necessary to implement the algorithm exposed above: an estimator $\hat{\delta}$ from the complete data ζ, X, and the means to calculate $E[\hat{\delta}/X]$. If the latter calculation is not feasible, it is sufficient to dispose of a method of sampling realizations of ζ according to its distribution conditional on X. In the following subsections, we present the ICE algorithm in blind, contextual, and adaptive cases.

1) Fuzzy Blind ICE Algorithm: Before explaining how the blind ICE algorithm runs, let us focus on the statistical modeling of the involved random variables.

As stated in the introduction, for each pixel s the random variable ζ takes its values in [0, 1] and contains two types of components: two hard components and a fuzzy one. Let δ_0, δ_1 be Dirac weights on 0 and 1 and μ the Lebesgue measure on R. By taking $\nu = \delta_0 + \delta_1 + \mu$ as a measure on [0, 1], the *a priori* distribution of each ζ_s can be defined by a density h on [0, 1], with respect to $\nu = \delta_0 + \delta_1 + \mu$. If we assume that ζ is a stationary process and that the distribution of each ζ_s is uniform on the fuzzy class, this density can be written

$$\begin{aligned} h(0) &= P[\zeta_s = 0] = \pi_0 \\ h(1) &= P[\zeta_s = 1] = \pi_1 \\ h(\varepsilon) &= 1 - \pi_0 - \pi_1, \qquad \text{for } \varepsilon \in]0, 1[. \end{aligned} \tag{6}$$

In order to define the distribution of X_s conditional on ζ_s, let us consider two independent Gaussian random variables X_0 and X_1, associated with the two "hard" values 0 and 1, whose densities f_0 and f_1 are, respectively, characterized by (m_0, σ_0^2) and (m_1, σ_1^2). We will assume

$$X_s = \zeta_s X_1 + (1 - \zeta_s) X_0 \tag{7}$$

which means that X_0 models the noise of the class 0, X_1, models the noise of the class 1, and, for $\zeta_s = \varepsilon \in]0, 1[$, $X_s = \varepsilon X_1 + (1 - \varepsilon) X_0$ models the noise of the fuzzy class ε. This is relevant with the view according to which fuzzy class ε contains, in proportion, ε of class 1 and $1 - \varepsilon$ of class 0. Finally, the density defining the distribution of X_s conditional to $\zeta_s = \varepsilon$ is a Gaussian density $f(x/\varepsilon)$ characterized by the mean $m(\varepsilon) = (1 - \varepsilon) m_1 + \varepsilon m_0$ and the variance $\sigma^2(\varepsilon) = (1 - \varepsilon)^2 \sigma_1^2 + \varepsilon^2 \sigma_0^2$.

Finally, for the case considered, the parameters required to be estimated are

$$\begin{aligned} \alpha &= (\pi_0, \pi_1) \\ \beta &= (m_0, \sigma_0^2, m_1, \sigma_1^2). \end{aligned} \tag{8}$$

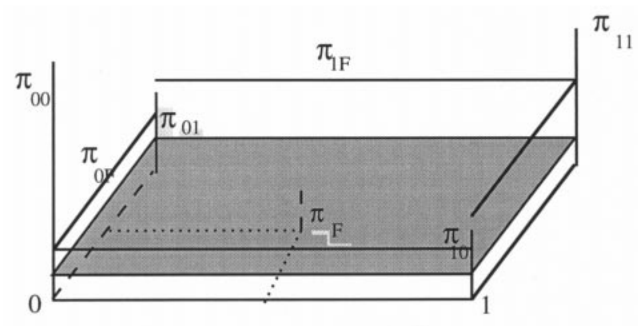


Fig. 1. Density of the distribution of $\zeta_V = (\zeta_s, \zeta_t)$ with respect to $\nu \otimes \nu$.

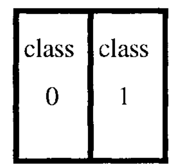


Fig. 2. Example of an image in which $\pi_{0\lambda} \neq \pi_{\lambda 0}$.

Returning to the ICE procedure, let us consider a subsample in S of m sites. First, we have to propose an estimator $\hat{\delta}$ of δ from complete data $\zeta_1, \cdots, \zeta_m$ and $X_1, \cdots, X_m$. We choose the empirical frequencies as estimators of the prior parameters and empirical moments as estimators of the noise parameters. To be more precise

$$\begin{aligned} \hat{\pi}_i &= \frac{1}{m} \sum_{j=1}^{m} 1_{[\zeta_j = i]} \\ \hat{m}_i &= \frac{\sum_{j \in Q_i} X_j}{\text{Card}\,(Q_i)} \\ (\hat{\sigma}_i)^2 &= \frac{\sum_{j \in Q_i} (X_j - \hat{m}_i)^2}{\text{Card}(Q_i)}, \qquad \text{for } i = 0, 1 \end{aligned} \tag{9}$$

with $Q_0 = \{j/\zeta_j = 0\}$, $Q_1 = \{j/\zeta_j = 1\}$, which defines $\hat{\delta}$.

According to the ICE principle, the updated values of the parameters are obtained by taking the expectation conditional to $(X_1, \cdots, X_m) = (x_1, \cdots, x_m)$ based on the current values of δ. This gives

$$\pi_i^{k+1} = \frac{1}{m} \sum_{j=1}^{m} p^k(i/x_j), \qquad \text{for } i = 0, 1 \tag{10}$$

where $p^k(\varepsilon/x_j)$ is the density with respect to ν of the distribution of ζ_j conditional on $X_j = x_j$ and based on the current value δ^k, as follows:

$$p^k(\varepsilon/x_j) = \frac{\pi_\varepsilon^k f(x_j/\varepsilon)}{\pi_0^k f(x_j/0) + \pi_1^k f(x_j/1) + (1 - \pi_0^k - \pi_1^k) \int_0^1 f(x_j/\theta)\, d\theta},$$

$$\text{for } \varepsilon = [0, 1] \tag{11}$$

which are obtained, in practice, by numerical integration. Thus $p^k(0/x_j)$ and $p^k(1/x_j)$ in (10) are the hard components of the distribution of ζ_j conditional on $X_j = x_j$.

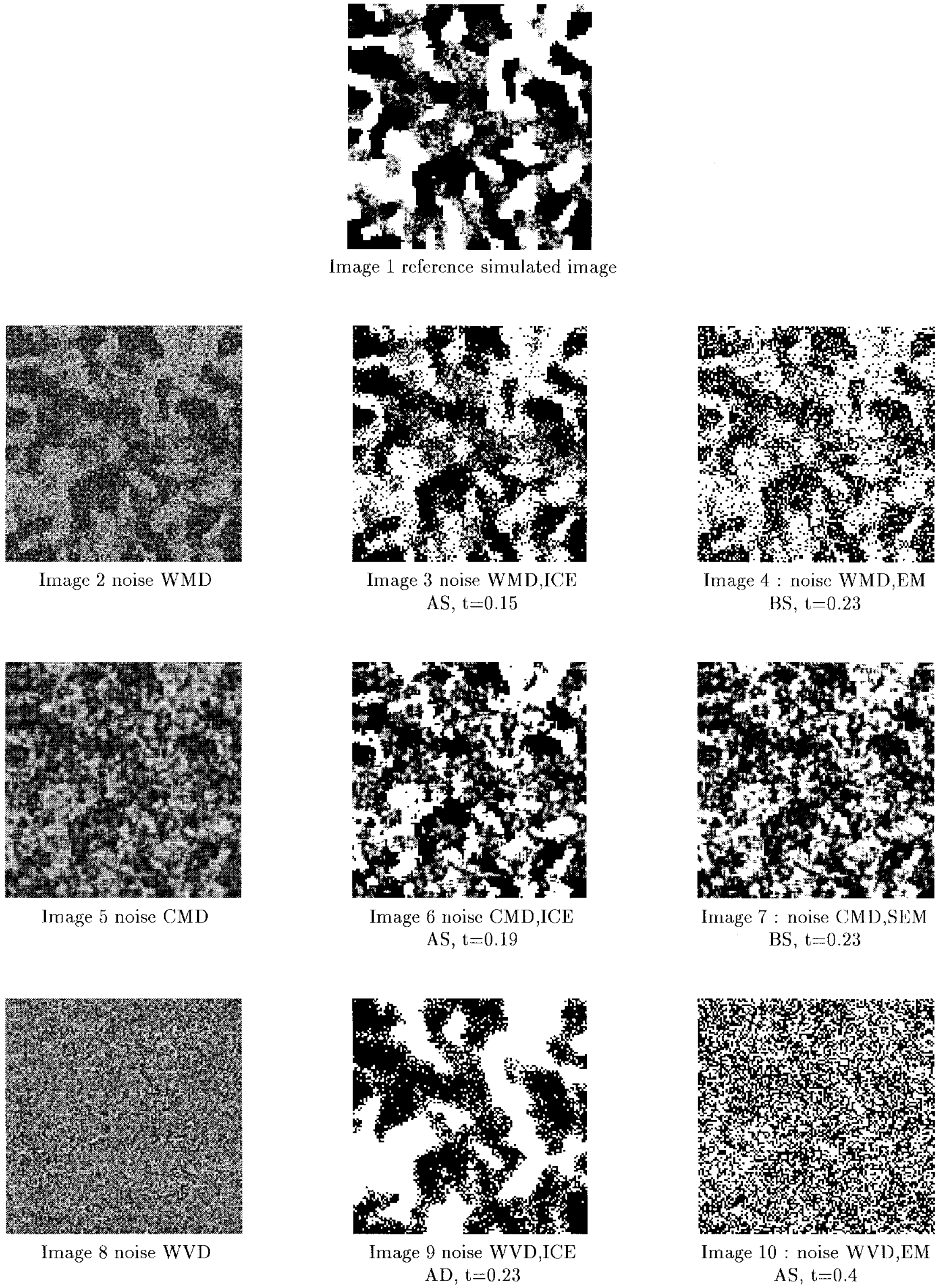


Fig. 3. Images 1 through 10.

Concerning the parameters of the Gaussian densities, the direct computation of conditional expectations of empirical means and variances is not feasible. Thus, we have recourse to a stochastic approximation, in accordance with the law of large numbers. Indeed, simulations of ζ_j realizations according to its posterior distribution are workable.

Finally the fuzzy blind ICE algorithm runs as follows.

- Give an initial value of the parameter $\delta^0 = [\pi_0^0, \pi_1^0, m_0^0, (\sigma_0^0)^2, m_1^0, (\sigma_1^0)^2]$.
- At each step k, δ^{k+1} is obtained from δ^k and the data $x_1, \cdots, x_m$ by

  reestimation of the priors: use (10);

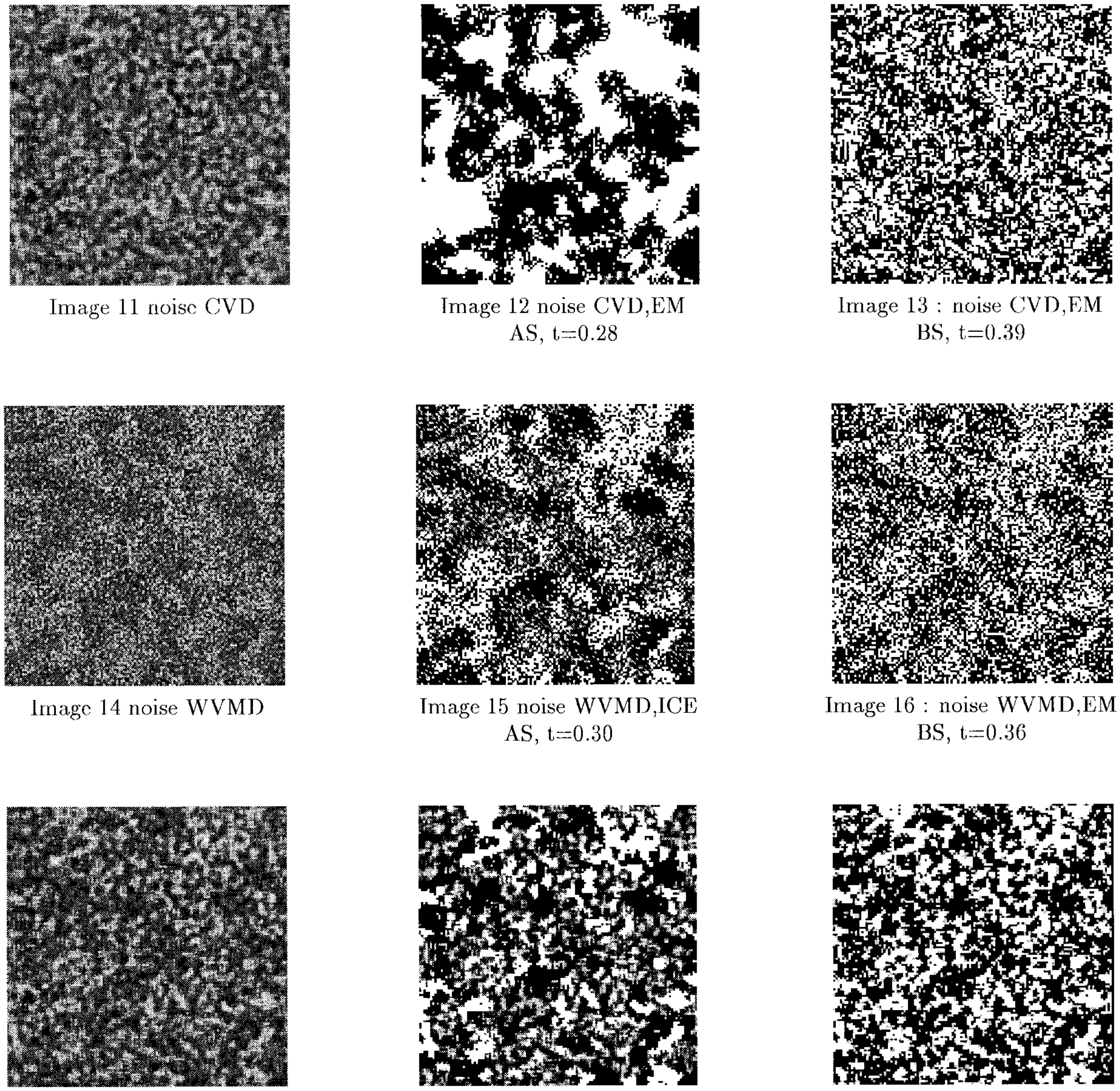


Image 11 noise CVD

Image 12 noise CVD,EM AS, t=0.28

Image 13 : noise CVD,EM BS, t=0.39

Image 14 noise WVMD

Image 15 noise WVMD,ICE AS, t=0.30

Image 16 : noise WVMD,EM BS, t=0.36

Image 17 noise CVMD

Image 18 noise CVMD,EM AS, t=0.34

Image 19 : noise CVMD,ICE CS, t=0.37

Fig. 4. Images 11 through 19.

reestimations of the noise parameters:

a) For each x_j of the sampler $x_1, \cdots, x_m$, compute the *a posteriori* probabilities $p^k(0/x_j)$ and $p^k(1/x_j)$ and sample a value in the set $\{0, 1, F\}$ according to $p^k(0/x_j)$, $p^k(1/x_j)$, and $1 - p^k(0/x_j) - p^k(1/x_j)$ (F representing the fuzzy pixels). Let $\varepsilon_1^k, \cdots, \varepsilon_m^k$ denote the realizations so obtained.

b) Let $Q_0^k = \{j/\varepsilon_j^k = 0\}$, $Q_1^k = \{j/\varepsilon_j^k = 1\}$; reestimate the noise parameters by

$$m_i^{k+1} = \frac{\sum_{j \in Q_i^k} x_j}{\text{Card}(Q_i^k)}$$

$$(\sigma_i^{k+1})^2 = \frac{\sum_{j \in Q_i^k} (x_j - m_i^{k+1})^2}{\text{Card}(Q_i^k)}, \qquad \text{for } i = 0, 1. \quad (12)$$

Let us note that according to the stochastic approximation of ICE, several samplings should be made and the next values of noise parameters would be given by means of different values obtained in the way described above. Simulation studies show that one can use, in general, just one sampling without significant alteration of the efficiency of the method. However, the possibility of regulating the stochastic aspect of the ICE by changing N [see (4)] can have great importance in some particular situations (see Section III-B2).

Remark 1: The fuzzy component of the prior distribution is assumed to be uniform and this assumption could turn out to be strong in some real situations. A generalization is possible; in fact, the uniform distribution can be superseded by a parametric family of distributions h_γ in such a way that $P[\zeta_s \in]a, b[] = (1 - \pi_0 - \pi_1) \int_a^b h_\gamma(\varepsilon)\, d\varepsilon$ for any $0 \le a \le b \le 1$. It is just necessary to propose an estimator

$\hat{\gamma} = \hat{\gamma}(\zeta)$ of γ. If $E[\hat{\gamma}/X]$ is not workable, one will have to resort to simulations.

2) Fuzzy Contextual ICE Algorithm: This section focuses on contextual estimation, which consists of working with spatial information by considering a sequence of contexts $(V_j)_{j=1,\cdots,m}$. In the following, these contexts are site pairs instead of single sites in the blind case exposed above. Thus, each pair of sites will be represented by two pairs of random variables, the unobservable couple $\zeta_{V_j} = (\zeta_{js}, \zeta_{jt})$ and the observation couple $X_{V_j} = (X_{js}, X_{jt})$. As in the blind case we must first define the distribution of (ζ_V, X_V), which will be given by the distribution of $\zeta_V = (\zeta_s, \zeta_t)$ and the distributions of $X_V = (X_s, X_t)$ conditional to $\zeta_V = (\zeta_s, \zeta_t)$.

The distribution of $\zeta_{V_i} = (\zeta_{js}, \zeta_{jt})$ can be defined by a density h on $[0, 1]^2$ with respect to the measure $\nu \otimes \nu$. This density h includes three types of components: four "hard" components corresponding to the case in which the two pixels s and t are "pure," four "combined" components corresponding to the case in which one of the pixels is "pure" and the other one is "mixed," and the last fuzzy component corresponding to the case where both pixels s and t are "mixed." We will suppose in the following that h is constant on each component, as expressed in Fig. 1.

More precisely, the four "hard" components of h can be expressed by $h(i, j) = P[\zeta_s = i, \zeta_t = j] = \pi_i j$ for $i, j = 0, 1$, the four combined components of h become $h(0, \theta) = h(\varepsilon, 0) = \pi_{0F}$, $h(1, \theta) = h(\varepsilon, 1) = \pi_{1F}$, and, finally, $h(\varepsilon, \theta) = \pi_F$ for $\varepsilon, \theta \in]0, 1[$.

Thus the density function h is defined by seven parameters, namely $\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}, \pi_{0F}, \pi_{1F}, \pi_F$. These parameters are bounded by the normalization constraint $\int_{[0,1]^2} h(\varepsilon, \theta)\, d\nu(\varepsilon)\, d\nu(\theta) = 1$, which gives $\pi_{00} + \pi_{01} + \pi_{10} + \pi_{11} + 2\pi_{0F} + 2\pi_{1F} + \pi_F = 1$. Let us note that π_{01} and π_{10} are not necessarily equal: If the pixels s and t are on the same line, then in the case of the example given by the Fig. 2 we have $\pi_{01} \neq 0$ and $\pi_{10} = 0$.

As in the blind case exposed in the preceding subsection, let us define the distributions of the pair (X_s, X_t) conditionally to (ζ_s, ζ_t). As above, we assume that these distributions are normal. In the following, $f(x, y/\varepsilon, \theta)$ denotes the normal density of (X_s, X_t) conditional on $(\zeta_s, \zeta_t) = (\varepsilon, \theta)$. In an analogous way as in the blind case, let us introduce four Gaussian random variables, X_0^1, X_0^2, X_1^1, and X_1^2, and let us assume that the distributions of the four Gaussian vectors (X_0^1, X_0^2), (X_0^1, X_1^2), (X_1^1, X_0^2), and (X_1^1, X_1^2) are the distributions of (X_s, X_t) conditional to $(\zeta_s, \zeta_t) = (0, 0), (0, 1), (1, 0), (1, 1)$, respectively. We will assume that the distribution of (X_s, X_t) conditional on $(\zeta_s, \zeta_t) = (\varepsilon, \theta)$ is defined by

$$(X_s, X_t) = [\varepsilon X_1^1 + (1-\varepsilon)X_0^1, \theta X_1^2 + (1-\theta)X_0^2]. \quad (13)$$

Let us denote by $m_{ij} = (m_i, m_j)$ and $\Gamma_{ij} = \begin{bmatrix} \sigma_i^2 & \rho_{ij} \\ \rho_{ji} & \sigma_j^2 \end{bmatrix}$ the four mean vectors and covariance matrices defining the Gaussian densities $f(x, y/i, j)$ (for $i = 0, 1$ and $j = 0, 1$). The Gaussian density $f(x, y/\varepsilon, \theta)$ of the distribution of (X_s, X_t) conditional on $(\zeta_s, \zeta_t) = (\varepsilon, \theta)$ is then defined by the mean vector $m(\varepsilon, \theta) = [m(\varepsilon), m(\theta)]$ and the covariance matrix $\Gamma(\varepsilon, \theta) = \begin{bmatrix} \sigma(\varepsilon)^2 & \rho(\varepsilon, \theta) \\ \rho(\varepsilon, \theta) & \sigma(\theta)^2 \end{bmatrix}$ with

$$\begin{aligned} m(\varepsilon) &= (1-\varepsilon)m_0 + \varepsilon m_1 \\ m(\theta) &= (1-\theta)m_0 + \theta m_1 \\ \sigma(\varepsilon)^2 &= (1-\varepsilon)^2\sigma_0^2 + \varepsilon^2\sigma_1^2 \\ \sigma(\theta)^2 &= (1-\theta)^2\sigma_0^2 + \theta^2\sigma_1^2 \\ \rho(\varepsilon, \theta) &= (1-\varepsilon)(1-\theta)\rho_{00} + (1-\varepsilon)\theta\rho_{01} \\ &\quad + \varepsilon(1-\theta)\rho_{10} + \varepsilon\theta\rho_{11}. \end{aligned} \quad (14)$$

Finally, all distributions of (X_s, X_t) conditional to (ζ_s, ζ_t) are defined by the seven parameters m_0, m_1, σ_0^2, σ_1^2, ρ_{00}, ρ_{10}, and ρ_{11}.

The distribution of $(\zeta_s, \zeta_t, X_s, X_t)$ on $[0, 1]^2 \times R^2$ admits the following density with respect to the measure $\nu \otimes \nu \otimes \mu \otimes \mu$ (where $\nu = \delta_0 + \delta_1 + \mu$, and μ is the Lebesgue measure)

$$g(\varepsilon, \theta, x, y) = h(\varepsilon, \theta) f(x, y/\varepsilon, \theta). \quad (15)$$

This distribution involves the distributions of (ζ_s, ζ_t) conditional to (X_s, X_t), which will be needed in the ICE procedure, and whose densities with respect to $\nu \otimes \nu$ are

$$g(x, y/\varepsilon, \theta) = \frac{g(\varepsilon, \theta, x, y)}{\int_{[0,1]^2} g(\varepsilon, \theta, x, y)\, d\nu(\varepsilon)\, d\nu(\theta)}. \quad (16)$$

Thus, the distributions of ζ_s conditional to (X_s, X_t), which will be needed in the segmentation step, are given by the densities

$$g(\varepsilon/x, y) = \int_{[0,1]} g(\varepsilon, \theta/x, y)\, d\nu(\theta) \quad (17)$$

with respect to ν.

Let us notice that the integration with respect to ν contains sums and Lebesgue integrals. For instance, the calculation of $w(x, y) = \int_{[0,1]^2} g(\varepsilon, \theta, x, y)\, d\nu(\varepsilon)\, d\nu(\theta)$, which is the density of the distribution of (X_s, X_t), is as follows:

$$\begin{aligned} w(x, y) &= \int_{[0,1]^2} g(\varepsilon, \theta, x, y)\, d\nu(\varepsilon)\, d\nu(\theta) \\ &= \sum_{0 \le i, j \le 1} g(i, j, x, y) \\ &\quad + \int_0^1 [g(0, \theta, x, y) + g(1, \theta, x, y)]\, d\theta \\ &\quad + \int_0^1 [g(\varepsilon, 0, x, y) + g(\varepsilon, 1, x, y)]\, d\varepsilon \\ &\quad + \int_{[0,1]^2} g(\varepsilon, \theta, x, y)\, d\varepsilon\, d\theta \end{aligned}$$

Returning to the ICE algorithm, let us consider a sequence of contexts $(V_q)_{q=1,\cdots,m}$ of two neighbors. We denote by $\zeta_{V_q} = (\zeta_{qs}, \zeta_{qt})$ the restriction of ζ to V_q and by $X_{V_q} = (X_{qs}, X_{qt})$ the restriction of X to V_q.

The parameter is initialized with

$$\begin{aligned} \delta^0 = (&\pi_{00}^0, \pi_{01}^0, \pi_{10}^0, \pi_{11}^0, \pi_{0F}^0, \pi_{1F}^0, \pi_F^0, m_0^0, m_1^0, \\ &\sigma_0^0, \sigma_1^0, \rho_{00}^0, \rho_{10}^0, \rho_{11}^0) \end{aligned} \quad (18)$$

and the problem is to calculate δ^{k+1} from δ^k and $x_{V_1}, \cdots, x_{V_m}$.

The empirical frequency estimators of the *a priori* parameters used are

$$\hat{\pi}_{ij} = \frac{1}{m}\sum_{q=1}^{m} 1_{[\zeta_{V_q}=(i,j)]}, \qquad \text{for } i, j \in \{0, 1, F\}. \tag{19}$$

Thus, the reestimation formulae obtained by computing the expectation of the above estimators conditional to the observations $X_{V_q} = (X_{qs}, X_{qt}) = (x_{qs}, x_{qt})$ are written

$$\pi_{ij}^{k+1} = \frac{1}{m}\sum_{q=1}^{m} g^k(i, j/x_{qs}, x_{qt}) \text{ for } i, j \in \{0, 1\} \tag{20}$$

where g^k is δ^k-based g defined with (16), and

$$\pi_{ij}^{k+1} = \frac{1}{m}\sum_{q=1}^{m} g^k(i, x_{qs}, x_{qt}) \text{ for } i \in \{0, 1\} \text{ and } j = F \tag{21}$$

where g^k is δ^k-based g defined with (17).

Concerning the parameters of the Gaussian densities, as in the blind case, we have recourse to simulations according to the distribution of $\zeta_{V_q} = (\zeta_{qs}, \zeta_{qt})$ conditional to $X_{V_q} = (X_{qs}, X_{qt})$, which are given by (16), based on the current parameter δ^k. In doing so, we obtain $\varepsilon_{V_1}^k, \cdots, \varepsilon_{V_m}^k$, m values in $\{0, 1, F\}^2$. We then define a partition of the sample $x_{V_1}^k, \cdots, x_{V_m}^k$ into nine subsamples by putting $Q_{ij}^k = \{q = 1, \cdots, m/\zeta_{V_q} = (i, j)\}$. The parameters are then estimated by empirical means, standard deviations, and correlations from these subsamples.

To be more precise,

$$m_{ij}^{k+1} = \begin{pmatrix} m_i^{k+1} \\ m_j^{k+1} \end{pmatrix} = \frac{1}{\text{Card}\,(Q_{ij})}\sum_{q \in Q_{ij}} x_{V_q} \tag{22}$$

$$\Gamma_{ij}^{k+1} = \begin{bmatrix} (\sigma_i^{k+1})^2 & \rho_{ij}^{k+1} \\ \rho_{ij}^{k+1} & (\sigma_j^{k+1})^2 \end{bmatrix} = \frac{1}{\text{Card}\,(Q_{ij})}\sum_{q \in Q_{ij}} (x_{V_q} - m_{ij}^{k+1})^t (x_{V_q} - m_{ij}^{k+1}). \tag{23}$$

Finally, (20)–(23) define the next value of the parameters

$$\delta^{k+1} = (\pi_{00}^{k+1}, \pi_{01}^{k+1}, \pi_{10}^{k+1}, \pi_{11}^{k+1}, \pi_{0F}^{k+1}, \pi_{1F}^{k+1}, \pi_F^{k+1}, m_0^{k+1}, m_1^{k+1}, \sigma_0^{k+1}, \sigma_1^{k+1}, \rho_{00}^{k+1}, \rho_{10}^{k+1}, \rho_{11}^{k+1}).$$

3) Fuzzy Adaptive ICE Algorithm: In this section, we will take up the blind segmentation point of view. The blind segmentation proceeds "pixel by pixel" and does not exploit any spatial information. However, in the adaptive unsupervised framework, the spatial information is taken into account through the estimation step. In fact, priors are assumed to depend on pixels, and are estimated from the observations on windows centered on each pixel. In the model we adopt, the noise parameters do not vary with pixels. This approach stays valid in the case of the nonstationary class field. Moreover, it can strongly improve the unsupervised blind segmentation results even in the stationary case, especially when the class field is homogeneous [24]. Thus, the model here is exactly the same as in Section II-B1, with the difference that the parameters defining priors depend on s. The other difference is that the subsample $x_1, \cdots, x_m$ used has to be the whole image, i.e., $m = \text{Card}\,(S)$. In fact, priors for each pixel are required. The parameters needed are

$$\delta = [\pi_0(s), \pi_1(s), m_0, m_1, \sigma_0^2, \sigma_1^2]. \tag{24}$$

The execution of ICE is modified in that that (10) is replaced by

$$\pi_i^{k+1}(s) = \frac{1}{\text{Card}\,[W(s)]}\sum_{t \in W(s)} p^k(i/x_t) \tag{25}$$

where $W(s)$ is a window centered on s.

Remark 2: The choice of the reestimation window size can play an important role in the adaptive framework. Small window sizes yield better local characteristics, but on the other hand, the estimation is less reliable. We have experimentally determined that the optimal size of the reestimation window, as concerns the error rate used, is around 7 × 7 pixels.

III. Numerical Comparisons Between ICE, EM, and SEM

This section is devoted to numerical applications and in particular, to the comparisons between results using ICE, EM, and SEM. First, we specify how the EM and SEM procedures are adapted to the model considered. Then the three algorithms are applied on simulated fuzzy data.

A. The EM and SEM Principles Compared to ICE

The EM algorithm is a classical procedure [10] and [27], which consists in the maximization, with respect to the parameter δ, of the likelihood of the observations. Starting from an initial value δ^0, it generates a deterministic sequence of values δ^k. As explained in Section II-A, the priors reestimation formulae with EM and ICE are the same in the hard case. Thus, we will keep in the "fuzzy" EM the same priors reestimation formulae as that in fuzzy ICE above.

Concerning the noise parameters reestimation, we propose the following adaptations of the hard EM to the fuzzy context:

Blind Case:

$$m_j^{k+1} = \frac{\sum_{i=1}^{n} x_i p^k(j/x_i)}{\sum_{i=1}^{n} p^k\, j/x_i} \tag{26}$$

$$(\sigma_j^{k+1})^2 = \frac{\sum_{i=1}^{n} (x_i - m_j^{k+1})^2 p^k\, j/x_i}{\sum_{i=1}^{n} p^k\, j/x_i}. \tag{27}$$

Contextual Case:

$$m_{ij}^{k+1} = \frac{\sum_{q=1}^{m} x_{V_q} p^k(i, j/x_{V_q})}{\sum_{q=1}^{m} p^k(i, j/x_{V_q})} \tag{28}$$

Fig. 5. Real SPOT image.

Fig. 6. True ground of the image in Fig. 5.

$$\Gamma_{ij}^{k+1} = \frac{\sum_{q=1}^{m} (x_{V_q} - m_{ij}^{k+1})^t (x_{V_q} - m_{ij}^{k+1}) p^k(i, j/x_{V_q})}{\sum_{q=1}^{m} p^k(i, j/x_{V_q})}. \tag{29}$$

The SEM algorithm [6] is a stochastic version of the EM algorithm. The principle of the SEM is as follows: at each step, one draws exactly one sample according to the posterior distribution, in the same way as in the ICE case when estimating the noise parameters. The difference with ICE is that the sample so obtained is also used in order to reestimate priors. This algorithm has already been applied in [4] in a fuzzy framework and, in a hard framework, it has given efficient results [22].

Adaptive versions of EM and SEM are obtained from EM and SEM in the same way that adaptive version of ICE is obtained from ICE.

Remark 3: We have seen in Section II-B1, Remark 1, that ICE can be used when the fuzzy component of priors is not uniform. In fact, it is always possible, using discretization if necessary, to simulate realizations of ζ according to (11). Thus, SEM also can be used in such situations. The adaptation of EM seems much more difficult.

B. Numerical Results

The three algorithms have been applied on a simulated fuzzy image corrupted with different Gaussian noises. The procedure used to sample the fuzzy data proceeds in two steps:

1) considering the fuzzy class as a third hard class, sample a classical three-class Markov field using the Gibbs sampler;

Fig. 7. Fuzzy statistical segmentation of the image in Fig. 5.

2) at each pixel in the fuzzy class sample a value in $]0, 1[$.

Thus, the first step gives a three-class image, each pixel being in $\{0, 1, F\}$. The second step is initialized putting 0.5 in each pixel labeled F. Then we scan the set of pixels "line by line." If the current pixel if hard, nothing is done. If it is fuzzy, we look at the sum Σ of the four neighboring pixels, which is in [0, 4] (0 if they are all hard and 0, 4 if they are all hard and 1). The fuzzy value is then updated sampling in $]0, 1[$ according to the density $f(x) = a(\Sigma - 2)x + b(\Sigma)$,

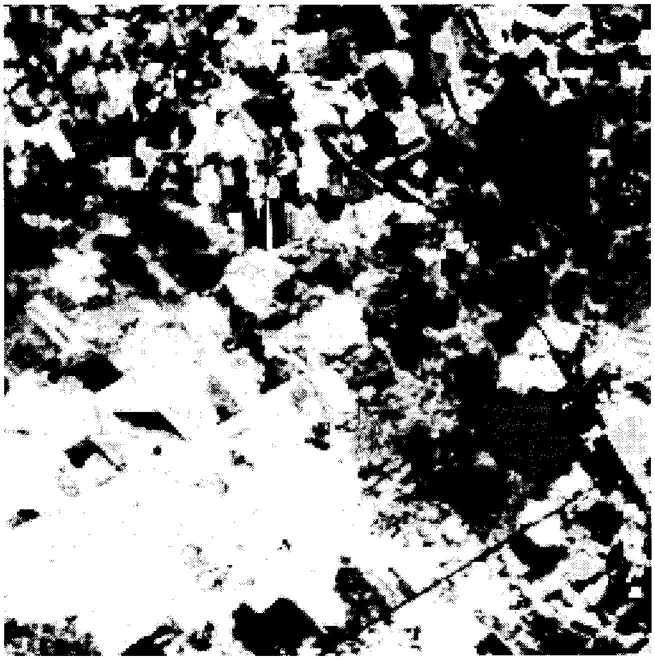


Fig. 8. Fuzzy C-means segmentation of the image in Fig. 5.

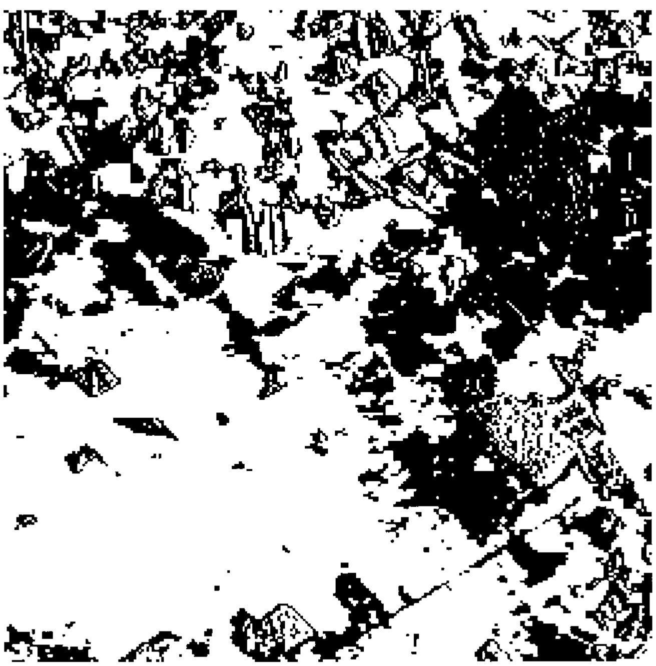


Fig. 9. Hard statistical segmentation of the image in Fig. 5 (two classes).

where $a > 0$ is fixed and $b(\Sigma)$ is calculated from a and Σ to ensure $\int_0^1 f(x)\,dx = 1$. The idea behind this way of sampling is the following: if 0 is dominant in the neighborhood, i.e., $\Sigma < 2$, f gives greater probability to the values near 0, and, if 1 is dominant in the neighborhood, i.e., $\Sigma > 2$, f gives greater probability to the values near 1. The aim of such a procedure is to ensure a visually good gradation when passing from one hard class to another. In the example below we use

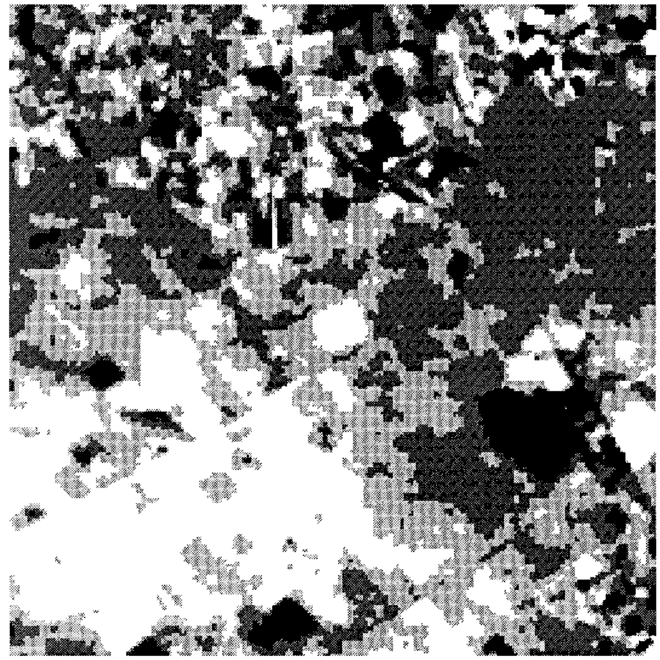


Fig. 10. Hard statistical segmentation of the image in Fig. 5 (four classes).

$a = 2.5$ for step 2). The three-class Markov field used in step 1) is Markovian with respect to the four nearest neighbors, thus its distribution is defined by functions φ on the cliques $\{s, t\}$. The simulated image (Image 1) has been obtained with $\varphi(\varepsilon_s, \varepsilon_t) = -1$ if $\varepsilon_s = \varepsilon_t$, and $\varphi(\varepsilon_s, \varepsilon_t) = 1$ if $\varepsilon_s \neq \varepsilon_t$. Image 1 is of size 128 × 128.

The reference fuzzy image obtained by the simulation procedure above is then corrupted with different Gaussian noises. We distinguish white (W) noises and correlated (C) ones. Each of them can be *means discriminating* (MD), i.e., $m_0 \neq m_1$ and $\sigma_0^2 = \sigma_1^2$, *variances discriminating* (VD), i.e., $m_0 = m_1$ and $\sigma_0^2 \neq \sigma_1^2$, or both *means and variances discriminating* (MVD), i.e., $m_0 \neq m_1$ and $\sigma_0^2 \neq \sigma_1^2$. For instance, WMD denotes white means discriminating noise, CMVD correlated, means and variances discriminating noise, and so on.

Let $(B_s)_{s\in S}$ be independent Gaussian random variables with zero mean and unit variance. Images corrupted with white noises are obtained with

$$X_s = [(1-\zeta_s)\sigma_0 + \zeta_s\sigma_1]B_s + [(1-\zeta_s)m_0 + \zeta_s m_1]. \quad (30)$$

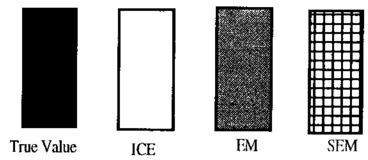
In order to obtain the images corrupted with correlated noises we first use the mobile average: for $(B_s)_{s\in S}$ independent Gaussian random variables with zero mean and unit variance, let

$$W_s = \frac{1}{\sqrt{\text{Card}\,(V_s)}} \sum_{t\in V_s} B_t. \quad (31)$$

Thus, $(W_s)_{s\in S}$ are correlated Gaussian random variables with zero mean and unit variance. Corrupted images are then obtained by (30) with $(W_s)_{s\in S}$ instead of $(B_s)_{s\in S}$. In the experiments below we have taken $\text{Card}\,(V_s) = 9$.

TABLE I

ESTIMATES OF BLIND MIXTURE. $1: m_0$, $2: m_1$, $3: \sigma_0^2$, $4: \sigma_1^2$, $5: \pi_0$, AND $6: \pi_1$



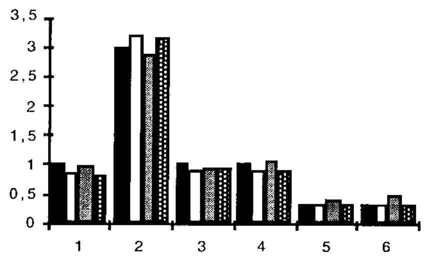


Table I 1: Noise WMD

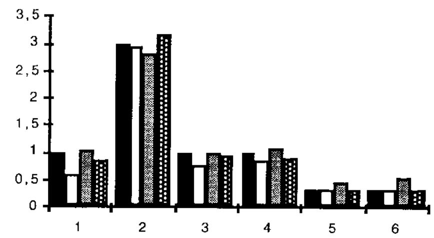


Table I 2: Noise CMD

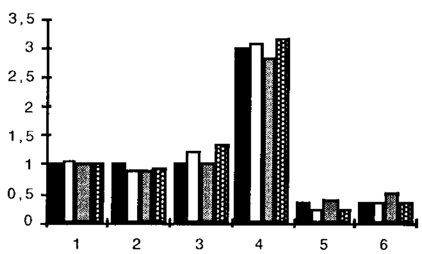


Table I 3: Noise WVD

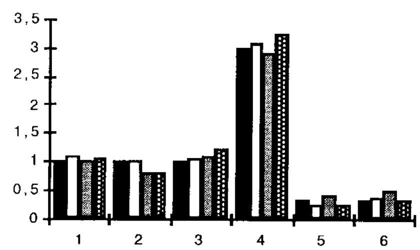


Table I 4: Noise CVD

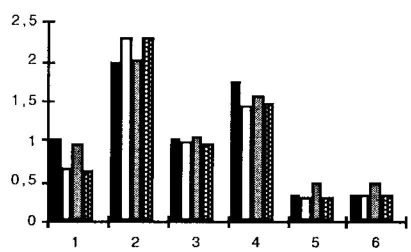


Table I 5: Noise WMVD

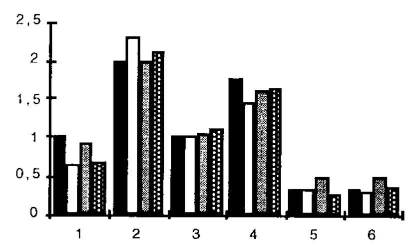


Table I 6: Noise CMVD

The experiments have been organized as follows.

- For the blind and the adaptive estimation procedures, start from an initialization arbitrarily chosen sufficiently apart from the true values in order to test the dependence on the initialization; stop the procedure when the estimated values stabilize, if they stabilize.
- For the contextual estimation procedure, start from the empirical estimates based on the blind unsupervised segmentation.

1) Blind Estimation: Simulations show that in general the EM procedure stabilizes more slowly than the ICE and SEM procedures. This confirms the fact that the EM procedure is more sensitive to initialization. However, we should note that the ICE and the SEM procedures, due to their stochastic properties, fluctuate around their convergence values when the EM procedure converges regularly. The noise parameters are correctly estimated, in most situations, by the three procedures, even though they are sensitive to the correlation of the noise, which was also shown in the hard framework [24]. In this case, the EM procedure seems less sensitive to this correlation. The most important result is that the EM procedure poorly estimates the *a priori* parameters and, in particular, does not recover the fuzzy class. Some results illustrating these conclusions are presented in Table I, and several others can be seen in [5].

In blind cases (classical and adaptive) the initialization of the parameters is as follows: One considers the empirical mean $\hat{m}$ and the empirical variance $\hat{\sigma}^2$ of the sample $(x_1, \cdots, x_n)$ used. The noise parameters are initialized with $m_0^0 = \hat{m} - \hat{\sigma}^2/2$, $m_1^0 = \hat{m} + \hat{\sigma}^2/2$, $\sigma_0^0 = \hat{\sigma} - \eta$, and $\sigma_1^0 = \hat{\sigma} + \eta$,

TABLE II
ESTIMATES OF CONTEXTUAL MIXTURE: $1 : m_0$, $2 : \sigma_0^2$, $3 : \rho_0$, $4 : \rho_{01}$, $5 : \pi_{00}$, $6 : \pi_{01}$, AND $7 : \pi_{0F}$

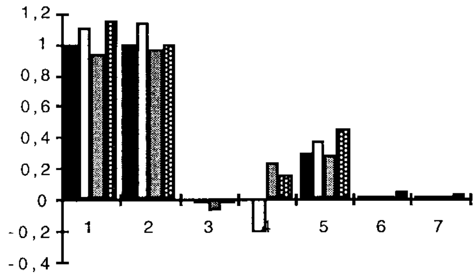


Table II 1: Noise WMD

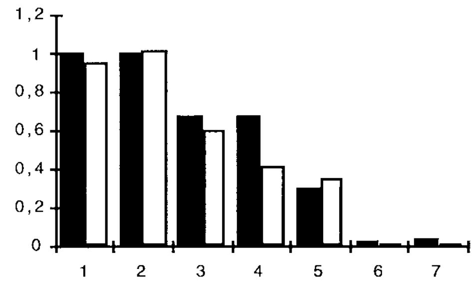


Table II 2: Noise CMD

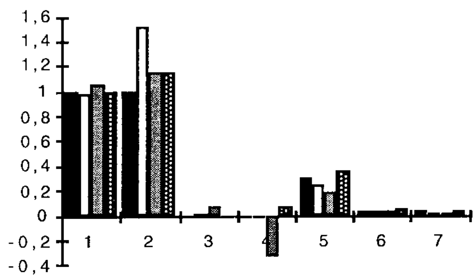


Table II 3: Noise WVD

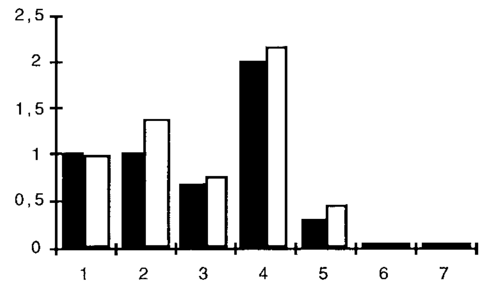


Table II 4: Noise CVD

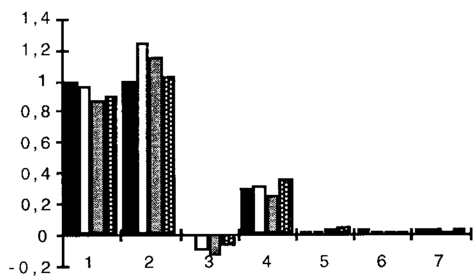


Table II 5: Noise WVMD

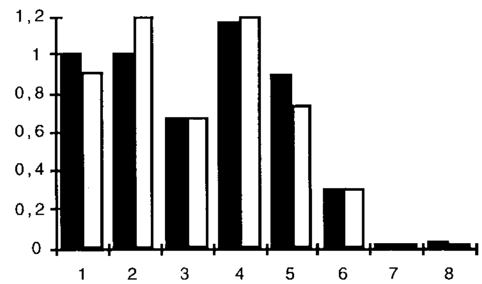


Table II 6: Noise CVMD

η being a value small relative to $\hat{\sigma}$. The starting priors are equal: $\pi_0^0 = \pi_1^0 = \pi_F^0 = \frac{1}{3}$.

2) Contextual Estimation: The prominent remark is that in all correlated situations the SEM and EM procedures do not converge, due to a poor estimation of the prior probabilities, even if the starting prior values are close to the real values. In this case, the ICE algorithm can be stabilized after few iterations of the procedure. Furthermore, the estimation of the noise parameters (particularly correlation) can be improved by increasing the number of the samplers according to the posterior distribution [see (3)]. Finally, the ICE procedure is clearly more reliable than the EM and the SEM ones in the contextual estimation case.

In contextual estimation, the starting values of the parameters are deduced from a blind segmentation.

3) Adaptive Estimation: In this section, we can only compare the estimation of the noise parameters to the true values. There is no significant differences between these results and the blind case, except in the case of variances and means discriminating noise, which seems to perturb the three procedures, and especially the EM.

IV. Fuzzy Statistical Unsupervised Segmentation

A. Segmentation Rule

There are two main approaches for statistical image segmentation: the global approach [1], [3], [7], [11], [12], [14], [17], [18], [20], [21], [25], [26], [28], [29], [31], and the local one [3], [4], [22], and [24]. A global method takes into account the values of X in the entire image. For instance, the MPM algorithm [21] estimates the value of each ζ_s, s being in the set of pixels S, by the class whose probability conditional to $X = x$ is maximal. Another global algorithm, the MAP [14] algorithm, estimates the value of ζ by $\varepsilon \in \Omega^{\mathrm{Card}(S)}$ whose probability conditional to $X = x$ is maximal. Both are Bayesian with two different loss functions. Let us recall that in the local framework, the expected value of each ζ_s is

TABLE III
ADAPTIVE ESTIMATES: $1 : m_0$, $2 : m_1$, $3 : \sigma_0^2$, AND $4 : \sigma_1^2$

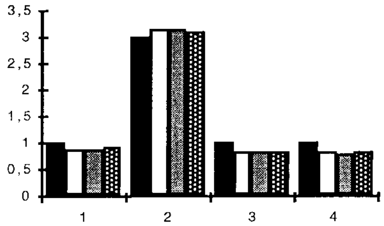


Tableau III 1: Noise WMD

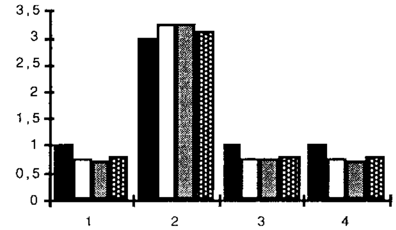


Tableau III 2: Noise CMD

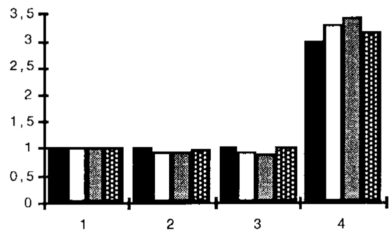


Table III 3: Noise WVD

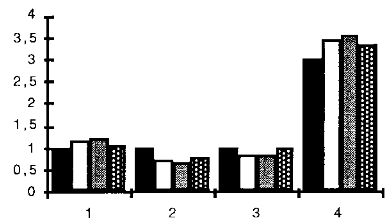


Table III 4: Noise CVD

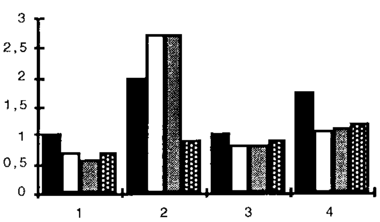


Table III 5: Noise WVMD

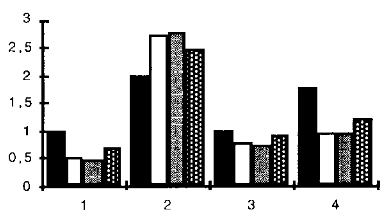


Table III 6: Noise CVMD

estimated from the observed values of $X = x$ restricted to a neighborhood V_s of s. A blind method is a local noncontextual method, i.e., $V_s = \{s\}$.

However, when considering the MPM method, the contextual method, or the blind method, the segmentation step is the same. In fact, these three methods define three different posterior distributions of each ζ_s, which are obtained with the conditioning by $X = x$, $X_{\nu_s} = x_{\nu_s}$, and $X_s = x_s$, respectively, but, once this distribution known, the problem of attributing of a class to s is the same in the three cases. In this paper, we will restrict ourselves to one possible segmentation method, namely the *maximum posterior likelihood* method. This has been successfully compared in [4] to three other methods, and we conjecture that it remains of interest in contextual and adaptive cases. Nevertheless, the other methods presented in [4] can be used.

In the blind case the "maximum posterior likelihood" method is as follows: Let us consider $p(\varepsilon/x_s)$ given by (11). Putting $p(F/x_s) = 1 - p(0/x_s) - p(1/x_s)$, the decision rule $\hat{\zeta}_s = \hat{s}(x_s)$ is

(i) let $\omega = \arg\max_{\lambda \in \{0, 1, F\}} p(\lambda/x_s)$. If $\omega \in \{0, 1\}$ put $\hat{s}(x_s) = \omega$. If $\omega = F$;

(ii) $\hat{s}(x_s) = \arg\max_{t \in]0, 1[} p(t/x_s)$.

Thus, the rule is following: First decide, maximizing the posterior probability, if the pixel is 0, 1, or "fuzzy." If it is 0 or 1, stop. If it is "fuzzy" determine its exact value maximizing the restriction of $p(t/x_s)$ to $]0, 1[$.

In the contextual case the rule is the same with $p(\varepsilon/x_s)$ replaced by $g(\varepsilon/x_s, x_t)$ [see (17)]. In the adaptive case it is still the same with the difference that $p(\varepsilon/x_s)$ also depends on s through priors.

Finally, an unsupervised segmentation method is obtained by adding to the segmentation rule above one of the parameter estimation methods of the previous sections.

Let us briefly discuss the relation of such methods to the fuzzy C-means methods. The fuzzy C-means algorithm was first proposed by Dunn [13] for the case $m = 2$ [see (32)], as an extension of hard classification ($m = 1$) called Isodata. The general form of the fuzzy C-means algorithm, i.e., for any m greater then one, was proposed by Bezdek [2] and studied by Hunstberger, Jacobs, and Canno [16], among others. In the latter methods the fuzzy partition is obtained by maximizing a given objective function Q. Recalling that a fuzzy partition is $\varepsilon = (\varepsilon_s)_{s \in S}$, with $\varepsilon_s = [\varepsilon_{is}]_{1 \le i \le e}$ (cf., Introduction), Q is written

$$Q(\varepsilon) = \sum_{i=1}^{e} \sum_{s=1}^{n} \varepsilon_{is}^{m} |\varepsilon_{is} - m_i|^2 \tag{32}$$

where $m \geq 1$ is a weighting exponent and m_i are the center values of the classes. The weighting exponent controls the magnitude of the fuzzy aspect of the image: The greater its magnitude, the more fuzzy will be the segmented image. If $m = 1$, this reduces to a classical hard segmentation.

The fuzzy C-means algorithm runs as follows:

1) initialize the matrix $\varepsilon = [\varepsilon_{is}]$ with ε^0;
2) update ε^k with

$$\varepsilon_{is}^{k+1} = \sum_{q=1}^{n} \left(\frac{|x_s - m_i^k|}{|x_s - m_q^k|} \right)^{(m-1)/2}$$

where

$$m_i^k = \frac{\sum_{s=1}^{n} (\varepsilon_{is}^k)^m x_s}{\sum_{s=1}^{n} (\varepsilon_{is}^k)^m}. \tag{33}$$

3) calculate

$$||\varepsilon^{k+1} - \varepsilon^k|| = \max_{j,\, s} |\varepsilon_{is}^{k+1} - \varepsilon_{is}^k| \tag{34}$$

4) repeat the Steps 2) and 3) until $||\varepsilon^{k+1} - \varepsilon^k|| \leq \gamma$, with γ a fixed threshold.

The procedure converges to ε^* at which the objective function Q has a local maximum.

We can see that the statistical method we propose and the fuzzy C-means method are very different in their principles, and undoubtedly their behavior can be quite distinct in different situations. In particular, the fuzzy C-means method does not use any probabilistic model and the noise is not explicitly modeled.

C. Segmentation Results

The resulting segmented image is compared with the simulated reference image by computing the following error rate:

$$t = \frac{\sum_{s \in S} |\hat{\zeta}_s - \zeta_s|}{\text{Card}\,(S)}. \tag{35}$$

The results presented in Table IV show that segmentation based on the ICE estimates gives the best results, particularly in the adaptive framework. In the blind case, the EM procedure leads the highest error rates, but the difference is not as appreciable as one would expect, considering the poor estimation of the prior parameters. The adaptive segmentation always improves the blind segmentation, the improvement being less noticeable in the case of correlated noise. This fact was also pointed out in a hard framework in [24]. The contribution of the contextual segmentation is significant in some situations (noise WMD and noise WVMD), but in many cases, principally all correlated noises, contextual segmentation offers little improvement over blind segmentation. The same conclusion is presented in [24] in a hard case, in such a way that we may conclude that fuzzy segmentation has the same properties as hard segmentation.

According to the results of Table IV, the great efficiency of adaptive segmentations (AS) is the prominent conclusion. They are clearly better that the blind segmentations (BS),

TABLE IV
RATE OF WRONGLY CLASSIFIED PIXELS. WMD, CMD, WVD, CVD, WVMD, AND CVMD: DIFFERENT NOISES WITH SAME PARAMETERS AS IN BS: BLIND SEGMENTATION, AS: ADAPTIVE SEGMENTATION, CS: CONTEXTUAL SEGMENTATION, DIV: THE ESTIMATION PROCEDURE DOES NOT STABILIZE

		WMD	CMD	WVD	CVD	WVMD	CVMD
ICE	BS	0,22	0,23	0,38	0,38	0,35	0,36
	AS	0,15	0,19	0,23	0,28	0,30	0,34
	CS	0,18	0,22	0,37	0,39	0,33	0,37
EM	BS	0,23	0,23	0,40	0,39	0,36	0,37
	AS	0,15	0,20	0,29	0,29	0,32	0,34
	CS	0,18	div	0,38	div	0,32	div
SEM	BS	0,22	0,23	0,38	0,38	0,35	0,36
	AS	0,17	0,19	0,29	0,31	0,31	0,34
	CS	0,19	div	0,36	div	0,33	div

consistent with ones intuition. What is more striking, they are also more efficient than contextual segmentations (CS). This would be expected if the image were nonstationary, i.e., if its visual aspect were different according to the place in pixels set. This is clearly not the case in examples we have chosen. The CS would perhaps be better if more than one pixel were used; to do so, however, would be rather tedious in the fuzzy model we consider. When considering the three ICE, EM, SEM-based AS, the first is slightly more efficient than the other two. In fact, calculating for each of them the mean of six error rates corresponding to different noises, we find 0.248 for ICE-based AS, and 0.265, 0.268 for EM and SEM, respectively.

The general conclusion is that the adaptive ICE-based segmentation should be used in the framework considered.

D. Segmented Images

We present in Figs. 3 and 4 some of the more revealing segmented synthetic images. For each type of noise the best segmented image, with respect to the error rate, and the worst segmented image have been selected. For each image, the procedure of estimation (ICE, EM, or SEM) and the type of segmentation method (BS for blind segmentation, AS for adaptive segmentation, and CS for contextual segmentation) are specified.

We observe that the nature of the noise has a noticeable effect on the visual aspect of the noisy images. On the other hand, the difference between the best and the worst segmentation is always visible, and, in some cases, quite significant. The general impression is that EM often makes images lose their homogeneity on the one hand, and their fuzzy aspect on the other.

Other results of similar studies can be seen in [5].

We present finally in Figs. 5–10 examples of unsupervised fuzzy and hard segmentations of a satellite image (SPOT). The parameters have been estimated by the ICE algorithm and the segmentation method is the blind posterior maximum likelihood.

The real image represents an agricultural area which may be considered as containing two classes: cultivated area and uncultivated area. The first class contains different kinds of cultures and the second class contains mainly water. According to the model we propose the fuzzy pixels simultaneously contain cultivated land and water, i.e., cultures on a very damp ground, marsh, etc.

The classical two-class segmentation (see Fig. 9) allows one to distinguish the two hard classes, although a great deal of information is lost. In particular, different plots of the ground in the "cultivated" class cannot be distinguished. Furthermore, different "borders" visible in both real and "true ground" images (see Fig. 6), such as rivers or roads, are lost, even when increasing the number of hard classes (see Fig. 10). These borders are preserved by both C-means (23) and statistical (see Fig. 7) fuzzy segmentations and one can observe that the statistical segmentation is more efficient for this particular problem (see lower left part of the image). Furthermore, the fuzzy statistical segmentation is the only one allowing the detection of some borders invisible in the real image and confirmed by the "true ground" image (see center of the image).

Fuzzy statistical segmentation also seems to better detect different parcels in the class "cultivated" than the C-means segmentation (see lower-left part of the image).

Finally, we put forth the following.

1) Both fuzzy segmentations render fine details better than hard segmentation, even when the hard segmentation uses several classes.
2) Fuzzy statistical segmentation better detects different borders than the C-means method, probably because of the Dirac measures.
3) When the noise is prominent, fuzzy statistical segmentation seems to be more efficient than C-means at the parcel detection level.
4) Fuzzy C-means segmentation is easier to perform and more efficient in terms of computer time.

Let us notice that our conclusions cannot be seen as general ones; indeed, we have presented only one image. Several other results can be seen in [5]. As a general conclusion we may say that both fuzzy segmentation methods are of interest compared to hard segmentation. As the principles of both fuzzy segmentation methods are very different, they should be seen more as complementary than as competing.

E. Extension to C-Class Problems

In this section, we present an extension of our fuzzy statistical modeling to any number of classes. In this framework, ζ_s takes its values in the e-dimensional unit simplex $S_e = \{\varepsilon \in [0, 1]^e / \sum_{i=1}^e \varepsilon_i = 1\}$. Let us consider $B_{e-1} = \{\varepsilon \in [0, 1]^{e-1} / \sum_{i=1}^{e-1} \varepsilon_i \leq 1\}$ and the one-to-one function T_e: $B_{e-1} \rightarrow S_e$ defined by

$$\begin{aligned}(\varepsilon_1, \cdots, \varepsilon_{e-1}) &\rightarrow T_e(\varepsilon_1, \cdots, \varepsilon_{e-1}) \\ &= [\varepsilon_1, \cdots, \varepsilon_{e-1}, 1-(\varepsilon_1 + \cdots + \varepsilon_{e-1})]. \qquad (36)\end{aligned}$$

Thus, we may define the measure ν_e on S_e as the image by T_e of a measure on B_{e-1}. The latter measure includes the Dirac weight on the vector 0 of R^{e-1} in order to weight the summit of $(0, \cdots, 0, 1)$ of S_e, the Lebesgue measure on B_{e-1}, and the measure ν_{e-1} which weighs the boundary S_{e-1} of B_{e-1}.

Finally, $\nu_e = T_e[(\delta_0 + \mu)^{\otimes(e-1)} 1_{B_{e-1}} + \nu_{e-1}]$, where μ is the Lebesgue measure on R. By putting $\nu = \delta_0 + \mu$ we have

$$\nu_e = T_e[\nu^{\otimes(e-1)} 1_{B_{e-1}} + \nu_{e-1}]. \qquad (37)$$

In the case $e = 2$, we have $\nu_2 = T_2(\delta_0 + \mu + \nu_1)$, where the measure ν_1 weighs the borderline $S_1 = \{1\}$ of $B_1 = [0, 1]$, which implies $\nu_1 = \delta_1$. Thus $\nu_2 = T_2(\delta_0 + \mu + \delta_1)$ and we recover the model of this paper.

Finally, $\nu_1 = \delta_1$ and (37) define a sequence of measures, each ν_e being valid in the e-dimensional case. The *a priori* distribution of ζ_s is then defined by a density with respect to the measure ν_e.

At each pixel s the observed field X_s is assumed to be a sum of independent Gaussian variables $X_i \leftrightarrow N(m_i, \sigma_i^2)$, as follows:

$$X_s = \sum_{i=1}^{e} \varepsilon_i X_i \qquad (38)$$

so that $E[X_s] = \sum_{i=1}^e \varepsilon_i m_i$ and $\mathrm{Var}\,[X_s] = \sum_{i=1}^e \varepsilon_i^2 \sigma_i^2$.

F. Relaxing Unit Hypothesis

Insofar as the grade of membership ζ_{is} is interpreted as the proportion of the area of the site s belonging to the class i, the constraint $\sum_{i=1}^e \zeta_{is} = 1$ is well founded. As pointed out by Krishnapuram and Keller [19], this constraint is not suitable in some applications, for instance, when the memberships have typical interpretations. In the same vein, in the original formulation of the fuzzy representation by Bezdek [2], the grades of membership are not relative and there is, thus, no relation between them. In the fuzzy statistical modeling proposed in the present paper, the probabilistic aspect in not connected with the fuzzy representation. Thus, the constraint that each ζ_s takes its values in the e-dimensional unit simplex is easy to relax by defining a suitable measure on the considered subspace of the hypercube $[0, 1]^e$. For instance, the possibilistic approach proposed by Krishnapuram and Keller [19] can be extended to a "possibilistic statistical" approach by considering $[\delta_0 + \mu + \delta_1]^e$ as a measure on the hypercube. Priors would be then defined by some density h with respect to this measure, and the observation process would be defined in the same way as in the previous section [cf., (38)], with $\max_{1 \leq i \leq e} \varepsilon_i \leq 1$ instead of $\sum_{i=1}^e \varepsilon_i = 1$. Let us note that merging Krishnapuram and Keller's approach with our method leads to an original and more complex model. In particular, as our approach allows us to model the noise and Krishnapuram and Keller's method encounters some problems when images are noisy, such a merged new model could turn out to be useful in very noisy image cases.

V. Conclusion

We have proposed in [4] a fuzzy statistical image model that simultaneously takes into account fuzzy and probabilistic aspects. The originality of our approach with respect to the Kent and Mardia method [18] is the simultaneous inclusion of Lebesgue and Dirac measures in priors, which allows the hard model to appear as a particular case of the fuzzy one. We then proposed some unsupervised statistical fuzzy "pixel by pixel" segmentation methods in which the previous estimation of the fuzzy mixture was performed with the SEM algorithm [6], [22]. In this paper, which is an extension of the results described in [4], we have focused on two points. First, we proposed two fuzzy mixture estimation algorithms, which are the ICE algorithm [3], [25], and an adaptation of the EM algorithm [10], [27] to the fuzzy context. The three methods, SEM, ICE, and EM, have been tested in different situations and the results obtained, which attest to their suitability, can be useful in any situation, eventually beyond image processing. Another aspect of this work was to include the contribution of spatial information in the segmentation methods and to compare blind, contextual, and adaptive estimation, and segmentation. With respect to segmentation error rates, the adaptive ICE-based approach provides the best results. The contextual approach, which improves on blind results in some situations, especially in the case of white noise, is not suitable in most cases due to its complexity and the surplus of necessary parameters. This fact was also pointed out in a hard context [22]. We must remark that in the hard case, contextual segmentation with just one neighbor is not relevant and it becomes necessary to consider a larger context [22] to improve noticeably the segmentation, which is not realistic in a fuzzy framework.

The segmentation methods we presented in this paper are local and it is well known, in the hard case, that global hidden Markov model-based methods [1], [7], [8], [11], [12], [14], [17], [18], [20], [21], [25], [29], [31] are much more efficient in several situations. However, it has been established [3] that local methods can be competitive in some situations and we conjecture, as the hard framework can be seen as a particular case of the fuzzy one, that the same is true in the fuzzy context. Otherwise, it is possible to define fuzzy hidden Markov models, which include Lebesgue and Dirac measures in priors and consider the corresponding global methods [26], [28].

The segmentation method we presented is different from the fuzzy C-means algorithm [2], [13], [16] and appears, according to the results of segmentation of a real image, as complementary.

As for topics of further work, let us point out the possibility of merging of our algorithms with Krishnapuram and Keller's approach [19]. This allows one to relax our hypothesis according to which the sum of grades of membership is one.

REFERENCES

[1] J. Besag, "On the statistical analysis of dirty pictures," *J. Royal Stat. Soc., Series B,* vol. 48, pp. 259–302, 1986.
[2] J. C. Bezdek, *Pattern Recognition and Fuzzy Objective Function Algorithm.* New York: Plenun, 1981.
[3] B. Braathen, W. Pieczynski, and P. Masson, "Global and local methods of unsupervised Bayesian segmentation of images," *Machine Graphics Vis.,* vol. 2, pp. 39–52, 1993.
[4] H. Caillol, A. Hillion, and W. Pieczynski, "Fuzzy random fields and unsupervised image segmentation," *IEEE Trans. Geosci. Remote Sensing,* vol. 31, pp. 801–810, 1993.
[5] H. Caillol, "Segmentation statistique floue d'images," thesis, Univ. Paris VI, Jan. 1995.
[6] G. Celeux and J. Diebolt, "L-algorithme SEM: Un algorithme d-apprentissage probabiliste pour la reconnaissance de mélanges de densités,"*Revue de statistique appliquée,* vol. 34, 1986.
[7] B. Chalmond, "An iterative Gibbsian technique for reconstruction of m-ary images," *Pattern Recog.,* vol. 22, no. 6, pp. 747–761, 1989.
[8] R. Chellapa and A. Jain, Ed., *Markov Random Fields.* Academic Press, 1993.
[9] J. P. Delmas, "Relations entre les algorithmes d'estimation itérative EM et ICE avec exemples d'application," in *Quinzième Colloque GRETSI,* Juan-les-Pins, France, Sept. 1995.
[10] M. M. Dempster, N. M. Laird, and D. B. Jain, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc., Series B,* vol. 39, pp. 1–38, 1977.
[11] H. Derin and H. Elliot, "Modeling and segmentation of noisy and textured images using Gibbs random fields," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. PAMI-9, pp. 39–55, 1987.
[12] R. C. Dubes and A. K. Jain, "Random field models in image analysis," *J. Appl. Stat.,* vol. 16, pp. 131–164, 1989.
[13] J. C. Dunn, "A fuzzy relative of the Isodata process and its use in detecting compact well-separated clusters," *J. Cybernet.,* vol. 31, pp. 32–57, 1974.
[14] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. PAMI-6, pp. 721–741, 1984.
[15] A. Hillion, "Les approches statistiques pour la reconnaissance des images de télédétection," in *Atti della XXXVI Riunione Scientifica, SIS,* vol. 1, pp. 287–297, 1992.
[16] T. L. Hunstberger, C. L. Jacobs, and R. L. Cannon, "Iterative fuzzy image segmentation," *Pattern Recog.,* vol. 18, pp. 131–138, 1985.
[17] P. A. Kelly, H. Derin, and K. D. Hartt, "Adaptive segmentation of speckled images using a hierarchical random field model," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 36, pp. 1628–1641, 1988.
[18] J. T. Kent and K. V. Mardia, "Spatial classification using fuzzy membership," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 10, pp. 659–671, 1991.
[19] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Trans. Fuzzy Syst.,* vol. 2, pp. 98–110, 1993.
[20] S. Lakshmanan and H. Derin, "Simultaneous parameter estimation and segmentation of Gibbs random fields using simulated annealing," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 11, pp. 799–813, 1989.
[21] J. Marroquin, S. Mitter, and T. Poggio, "Probabilistic solution of ill-posed problems in computational vision," *J. Amer. Stat. Assoc.,* pp. 76–89, 1987.
[22] P. Masson and W. Pieczynski, "SEM algorithm and the unsupervised segmentation of satellite data," *IEEE Trans. Geosci. Remote Sensing,* vol. 31, pp. 618–633, 1993.
[23] W. Pedrycz, "Fuzzy sets in pattern recognition: Methodology and methods," *Pattern Recog.,* vol. 23, pp. 121–146, 1990.
[24] A. Peng and W. Pieczynski, "Adaptive mixture estimation and unsupervised contextual Bayesian image segmentation," *Graphic. Models Image Proc.,* vol. 57, pp. 389–399, 1995.
[25] W. Pieczynski, "Statistical image segmentation," *Machine Graphics Vis.,* vol. 1, pp. 261–268, 1992.
[26] W. Pieczynski and J. M. Cahen, "Champs de Markov flous cachés et segmentation d'images," *Revue Stat. Appl.,* vol. 42, pp. 13–31, 1994.
[27] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Rev.,* vol. 26, pp. 195–239, 1984.
[28] F. Salzenstein and W. Pieczynski, "Unsupervised Bayesian segmentation using hidden fuzzy Markov fields," in *Proc. ICASSP-95,* Detroit, MI.
[29] A. Veijanen, "A simulation-based estimator for hidden Markov random fields," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 13, pp. 825–830, 1991.
[30] F. Wang, "Fuzzy supervised classification of remote sensing images," *IEEE Trans. Geosci. Remote Sensing,* vol. 28, pp. 194–201, 1990.
[31] L. Younes, "Parametric inference for imperfectly observed Gibbsian fields," *Prob. Theory Related Fields,* vol. 82, pp. 625–645, 1989.

Hélène Caillol was born in France in 1968. She obtained the Doctorat degree from Pierre et Marie Curie University, Paris, France, in 1995.

She is currently a temporary teacher and researcher at the Institute of Statistics, Pierre et Marie Curie University. Her research interests include statistical and fuzzy image processing and mathematical statistics.

Wojciech Pieczynski was born in Poland in 1955. He received the Doctorat d'État degree from Pierre et Marie Curie University, Paris, France, in 1986. He is currently Professor at the Institut National des Télécommunications, Evry, France. His areas of research interest include mathematical statistics, theory of evidence, stochastic processes, and statistical image processing.

Alain Hillion (A'87) was born in Brest, France, in 1947. He received the Agregationd de Mathématique degree from the École Normale Supérieure in 1970, and the Doctorat d'État degree from Pierre et Marie Curie University, Paris, France, in 1980.

He is currently Professor and Deputy Director of Research of the École National Supérieure des Télécommunications de Bretagne, Brest, France. His areas of research interest include mathematical statistics, pattern recognition, decision theory, and signal and image processing.