STATISTICAL IMAGE SEGMENTATION

Wojciech Pieczynski
Groupe Image
Département Systèmes et Réseaux
Institut National des Télécommunications
9, rue Charles Fourier 91011 Evry Cedex France

Abstract. This work deals with the unsupervised Bayesian segmentation of images. We describe a recent general procedure of estimation of components in a mixture of distributions and show some possibilities of its application to the problem of the unsupervised segmentation. As far as modelling by Markov random fields is concerned the procedure presented can generate several particular algorithms of unsupervised segmentation of images.

1. Introduction

Our work deals with the problem of the unsupervised segmentation of images whose importance has been pointed out by several authors ([8], [9], [21]). When considering the statistical approach, with Bayesian segmentation, the modelling by Markov random fields is the generally adopted one. When distributions of corresponding Markov fields are sufficiently known there exist some efficient algorithms (simulated annealing, MPM, ICM) of Bayesian segmentation. When they are not sufficiently known the problem becomes more difficult: one has to estimate the useful parameters, either in a previous step or simultaneously. The statistical problem is that of “lost data”: parameters are estimated from random variables whose distribution is a mixture of distributions.

The aim of this paper is to show how a recent procedure of estimation in the case of lost data, whose generality goes beyond the problems of image processing, can generate numerous algorithms of unsupervised Bayesian segmentation. The development of this paper is as follows:

In the following section we recall the modelling by Markov random fields and three Bayesian methods of segmentation.

The third section is devoted to a recent iterative method of estimation, called iterative conditional estimation (ICE). We introduce it in the case of a finite mixtures, sufficient to deal with the problem of segmentation, but a more general framework can be considered.

In the fourth section we show how the ICE procedure can generate numerous algorithms of unsupervised Bayesian segmentation and give an example.

Machine GRAPHICS and VISION vol. 1, nos. 1/3, 1992, pp. 261–268
(Proceedings of 2nd GRPO'92 Conference, Malbąg, May 18–23, 1992)
The final section comprises the conclusion.

2. Modelling and Bayesian Segmentation

When considering the statistical segmentation of images authors generally suppose the existence of two random fields: the field of "classes" $\zeta = (\zeta_s)_{s \in S}$, and the field of "measurements" $X = (X_s)_{s \in S}$. Each $\zeta_s$ takes its values in a finite set $\Omega = \{\omega_1, \omega_2, \ldots, \omega_m\}$ of classes and $X_s$ in $\mathbb{R}^d$ with $d \in N$. So the problem of segmentation in the problem of the estimation of an "ignored" realization of $\zeta$ from an "observed" realization of $X$. We will suppose that realizations of $X$ depend on realizations of $\zeta$ and a noise containing as well the "natural variability" or "texture" as the disturbances due to the "transmission". The distribution of $(\zeta, X)$ is defined by $P_{\zeta}$, distribution of $\zeta$, and the family $P^X_\zeta$ of distributions of $X$ conditional to $\zeta = \varepsilon$. The field $\zeta$ will be supposed Markovian with all realizations possible; so its distribution is Gibbs distribution:

$$P[\zeta = \varepsilon] = P_{\zeta}[\varepsilon] = ke^{-U_\alpha(\varepsilon)}$$  \hspace{1cm} (1)

where $k$ is unknown, $\alpha$ a parameter and the form of $U_\alpha$ simple enough to make possible the calculation of the spatially conditional distributions. The distribution of $X$ can be defined by $m$ distributions $P^1, P^2, \ldots, P^m$ of $X$ conditional to $m$ uniform realizations $(\zeta_s = \omega_1$ for each $s$, $\zeta_s = \omega_2$ for each $s$, \ldots, $\zeta_s = \omega_m$ for each $s$ respectively) and the hypothesis according to which the random variables $X_i$ (with $X_i$ restriction of $X$ to $Q_i = \{s \in S/ \zeta_s = \omega_i\}$) are independent. In order to allow simulations of $P^1, P^2, \ldots, P^m$ we will suppose that they are Gibbs distributions.

In the most general case Bayesian segmentation consists in estimating the realization of $\zeta_A$ from $X_B$, for $A, B \subseteq S$. More precisely $\zeta_A$ is estimated by $\hat{\zeta}_A$ which maximises the "posterior" (conditional to $X_B = \varepsilon_B$) distribution of $\zeta_A$. Different choices of $A, B$ generate different approaches: our work concerns the "global methods" which correspond to $B = S$ (pixels are classified using the whole information simultaneously), "local methods" corresponding to $B$ of "small size" with $A \subseteq B$ (each pixel is classified from information contained in its neighbourhood). The best known global approaches are:

- MAP corresponds to $A = S$, $B = S$. The problem is resolved by Geman and Geman's ([11]) iterative algorithm (simulated annealing).

- MPM corresponds to $A = \{s\}$, $B = S$. In this case the problem can be treated by Marroquin et al's algorithm ([16]).

- ICM. This deterministic algorithm of Besag ([2]) is rather a local one but we will range it with global methods because it necessitates the same modelling (using Gibbs distributions) that MAP or MPM.
Thus, we can apply one of the above-mentioned methods if we know the corresponding parameters. Let us denote $\alpha$ parameters concerning the distribution of $\zeta$ and $\beta$ parameters corresponding to the conditional distributions of $X$. As we said above, $\alpha$ is a parameter, generally appearing in the function "energy" $U$ of the Gibbs distribution (1), knowing which makes it possible to compute the distribution of each $\zeta$ conditional to $(\zeta, x)$. The use of global methods does not allow the taking of the noise correlation into account: the random variables $(X_i)$ are assumed to be independent conditionally to each realization of $\zeta$. So, if we have $m$ classes, it is sufficient to define $m$ distributions of each $X_i$ (conditional to $\zeta = \omega_1, \ldots, \omega_m$ respectively). Thus $\beta$ will have $m$ components $\beta_1, \ldots, \beta_m$. For instance, if the field $X$ is Gaussian conditionally to $\zeta$ and each $X_i$ real, we have $\beta_i = (\mu_i, \sigma_i^2)$, where $\mu_i$ is the mean and $\sigma_i^2$ the variance of the Gaussian distribution of each $X_i$ conditional to $\zeta = \omega_i$.

In practice we just often have an observation $X = x$.

Thus we have to choose a segmentation method and propose an estimator for the corresponding parameter $(\alpha, \beta)$. The distribution of $X$ is:

$$P_X = \sum_{\zeta} P(\zeta|x)P(\zeta)$$

which is said to be a "finite mixture of distributions". The problem of estimating $(\alpha, \beta)$ (in a certain sense "components of the mixture") is a very general statistical problem which often arises when dealing with Bayesian classification. We describe in the next section a recent general procedure of estimation and describe, in section 4, how this procedure can generate several algorithms of unsupervised segmentation.

3. Iterative Conditional Estimation (ICE)

Let us consider, in a general manner, a couple of random variables $\zeta, X$, $\zeta$ taking its value in a finite set and $X$ in $R^d$. Let us suppose that the distribution of $\zeta, X$ is defined by $P_\zeta$, distribution of $\zeta$, and the family $P_X^\zeta$ ($\zeta$ being a possible realization of $\zeta$) of distributions of $X$ conditional to $\zeta = \zeta$. The distribution $P_\zeta$ depends on a parameter $\alpha$ and the family $P_X^\zeta$ on a parameter $\beta$ (the correspondence between the distributions of $(\zeta, X)$ and $(\alpha, \beta)$ is not necessarily bijective). $X$ is the only observable and the problem is to estimate the parameter $(\alpha, \beta)$.

Let us suppose that we have at our disposal two estimators $\hat{\alpha} = \hat{\alpha}(\zeta) \text{ and } \hat{\beta} = \hat{\beta}(\zeta, X)$, $\hat{\alpha}$ being defined from $\zeta$ and $\hat{\beta}$ from $(\zeta, X)$. The direct use of $\hat{\alpha}$ and $\hat{\beta}$ is impossible because realizations of $\zeta$ are ignored, so, we have to approach $\hat{\alpha}$ and $\hat{\beta}$ by some function of $X$, the only one observable. The best approximation, as far as the square error is concerned, is the conditional expectation. To be more precise, if, for two random variables $Y, Z$, we denote the conditional expectation by $E[Y/Z]$ we can state:
\[ E(Y - E(Y|Z)^2) = \min_{\varphi} E((Y - \varphi(Z))^2) \]

(3)

Thus we are led to condition \( \hat{\alpha} \) and \( \hat{\beta} \) by \( X \): in doing this we lose the quality as "estimators". In fact, \( \alpha \) is independent from the parameter \((\alpha, \beta)\) but \( E(\hat{\alpha}/X) \), which, for each \( X = x \) is the expectation of \( \alpha \) according to the distribution of \( \zeta \) conditioned by \( X = x \) (so called "posterior" distribution) depends on \((\alpha, \beta)\). The same is true for \( E(\hat{\beta}/X) \). Hence, if we want to compute \( E(\hat{\alpha}/X = x) \) and \( E(\hat{\beta}/X = x) \) we have to take a parameter \((\alpha_n, \beta_n)\) previously defined by some way. This defines an iterative procedure which will be called Iterative Conditional Estimation (ICE, [20]). By denoting \( E_n \) the conditional expectation using \((\alpha_n, \beta_n)\) the procedure is:

- one takes an initial value \((\alpha_0, \beta_0)\)
- \((\alpha_{n+1}, \beta_{n+1})\) is computed from \((\alpha_n, \beta_n)\) and \( X = x \) by:

\[
\alpha_{n+1} = E_n[\hat{\alpha}/X = x]
\]
\[
\beta_{n+1} = E_n[\hat{\beta}/X = x]
\]

(4) and (5)

In order to simplify matters let us denote by \( P_n^x \) the distribution of \( \zeta \) conditioned by \( X = x \) (posterior distribution) computed from \((\alpha_n, \beta_n)\). (4) and (5) can be written:

\[
\alpha_{n+1} = \sum_{\epsilon} \hat{\alpha}(\epsilon) P_n^x[\epsilon]
\]
\[
\beta_{n+1} = \sum_{\epsilon} \hat{\beta}(\epsilon, x) P_n^x[\epsilon]
\]

(6) and (7)

(4) and (5) or (6) and (7) are not necessarily workables in practice. As we will see in next section, \( P_n^x \) is not entirely known when the problem of image segmentation is tackled from modelling by Markov random fields, but it is possible to simulate realizations of \( \zeta \) according to \( P_n^x \). In situations like this we can use an "approximative" ICE: (6) and (7) can be approached, in accordance with the law of large numbers, by:

\[
\alpha_{n+1} = \frac{1}{N} \sum_{i=1}^{N} \hat{\alpha}(\epsilon_i)
\]
\[
\beta_{n+1} = \frac{1}{N} \sum_{i=1}^{N} \hat{\beta}(\epsilon_i, x)
\]

(8) and (9)
where $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ are realizations of $\zeta$ according to $P_n^\varepsilon$.

Finally, we can use the ICE procedure as soon as we get:

- An estimator $\hat{\alpha} = \alpha(\zeta)$ of $\alpha$ and another one $\hat{\beta} = \hat{\beta}(\zeta, X)$ of $\beta$.

- An initial value $(\alpha_0, \beta_0)$ not “too far” from the real parameter $(\alpha, \beta)$.

- A way of computing $E_n[\hat{\alpha}/X]$ and $E_n[\hat{\beta}/X]$ or a way of simulating realizations of $\zeta$ according to the posterior distribution $P_n^\varepsilon$.

4. Unsupervised Segmentation of Images

The ICE procedure can generate a great number of non supervised segmentation algorithms (NSSA). $\alpha$ is a parameter, generally appearing in the energy of the Gibbs distribution of $\zeta$, knowing which allows the computing of the conditional distributions. $\beta$ is a parameter defining the distributions, conditional to $\zeta = \varepsilon$, of $X$. The use of ICE is possible as soon as we can simulate realizations of $\zeta$ according to the posterior distribution, the latter being computed with the current parameter $(\alpha_k, \beta_k)$. Thus, the possibility of simulations requires the markovianity of the posterior distribution which is ensured by the conditional independence of the random variables $(X_i)$. Let us suppose $X$ real and conditionally Gaussian. So the parameter $\beta$ is $\beta = (\beta_1, \ldots, \beta_m)$, with $m$ the number of classes, and $\beta_i = (\mu_i, \sigma_i^2), \mu_i, \sigma_i^2$ being respectively the mean and variance of the distribution of each $X_i$ conditional to $\zeta = \omega_i$ (for $1 \leq i \leq m$).

So ICE will define a NSSA as soon as we take $\hat{\alpha}, \hat{\beta}$ estimators of $\alpha, \beta$ respectively (from $(\zeta, X)$) and choose between MAP, MPM, ICM. Let us mention four possibilities for $\hat{\alpha}$:

$\hat{\alpha}_1$: pseudo-maximum likelihood (2)]. The idea is to substitute the likelihood(1), where $k$ is unknown, by the product of conditional distribution. $\alpha$ is then estimated by $\alpha^\ast$ which, for a given realization of $\zeta$, maximises the function thus obtained.

$\hat{\alpha}_2$: least square error method. The description of this method, proposed by Derin and Elliot, can be seen in [9].

$\hat{\alpha}_3$: maximum likelihood. L. Younes proposes ([26]) an iterative procedure defining a sequence converging to the maximum of the likelihood defined by (1). His algorithm, based on the stochastic gradient, allows a rigorous mathematical justification.

$\hat{\alpha}_4$: algorithm of Metropolis. Its description can be seen in [13].
The more natural choice of the estimator \( \hat{\beta} = \hat{\beta}(\zeta, X) \), in the hypothesis when conditional distributions of \( X \) are Gaussian, are the empirical mean and variance \((\mu_i, \sigma_i^2)\) which are defined from \( s \in S \) such that \( \zeta_s = \omega_i \). This can be written:

\[
\mu_i = \frac{\sum_{s \in S} X_s 1_{[\zeta_s = \omega_i]}}{\sum_{s \in S} 1_{[\zeta_s = \omega_i]}} \tag{10}
\]
\[
\sigma_i^2 = \frac{\sum_{s \in S} (X_s - \mu_i)^2 1_{[\zeta_s = \omega_i]}}{\sum_{s \in S} 1_{[\zeta_s = \omega_i]}} \tag{11}
\]

Formulae (4) and (5) are not workable and have to use (6) and (7) the sequence \( \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N \) being simulated by the Gibbs sampler ([10]). Then we can regulate the importance of the "stochastic" aspect of ICE by choosing \( N \): when \( N \) increases the stochastic side of the algorithm decreases. The intentional choice of \( N \) "small" can increase its efficiency: the authors of the SEM algorithm([4]), which can be seen as a particular case of ICE with \( N = 1 \) and the choice of (8) instead of (4), show its superiority over the EM algorithm in a simple case of a mixture of Gaussian densities. It is also possible to make \( N \) depend on the number \( k \) of iteration and introduce a kind of "cooling".

When combining these different possibilities we arrive twelve global unsupervised methods of segmentation. For instance, taking \( \alpha, \beta \) and MAP the proceeding is:

- one takes \( \Theta_0 = (\alpha_0, \beta_0) \)
- \( \Theta_{k+1} = (\alpha_{k+1}, \beta_{k+1}) \) is computed from \( \theta_k = (\alpha_k, \beta_k) \) and \( X = x \) in the following way:
  - using the Gibbs sampler \( N \) realizations \( \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N \) of \( \zeta \) are simulated according to the posterior distribution computed from \( (\alpha_k, \beta_k) \) and \( x \),
  - for each \( \varepsilon_j \) the parameter \( \alpha \) is estimated by the iterative algorithm of L. Younes (which requires \( m \) new simulations by the Gibbs sampler) – let us denote \( \alpha(\varepsilon_j) \) the result obtained,
  - for each \( \varepsilon_j \) the parameter \( \beta \) is estimated by formulae (10) and (11), which gives \( \beta(\varepsilon_j) \),
  - \( \alpha_{k+1}, \beta_{k+1} \) are obtained from \( (\alpha(\varepsilon_j), \beta(\varepsilon_j)) \), \( 1 \leq j \leq N \) by (8) and (9),
- if the sequence \( (\alpha_k, \beta_k) \) becomes steady, the estimation step is finished and one proceeds to the segmentation using the simulated annealing([10]).

A comparative study between this algorithm and the simulated annealing based on parameters estimated by another algorithm of L. Younes ([27]) can be seen in [1].

Machine GRAPHICS and VISION vol. 1, nos. 1/2, 1992, pp. 261-268
(Proceedings of 2nd GKPO'92 Conference, Nalyczów, May 18-23, 1992)
5. Conclusion

We presented some applications of the ICE procedure to unsupervised image segmentation. We have limited our investigations to global methods but the same ICE general method can also generate some local ones ([3],[14],[15],[20]). It is therefore possible to define numerous NSSA: probably none of them is better than any other in the general case. From theoretical point of view the study of their asymptotical behaviour us without doubt, rather tedious: until now we have only had some partial results([6],[13],[19],[20]). Numerous publications concerning the EM algorithm, in the case of independent data, lead us to imagine the extent of the difficulty.

From the point of view of applications there is a wide study of tests to do: possibilities of applications are extremely various and the problem is that of knowing what family of NSSA is well suited to given kind of images. We have begun some studies ([3],[14]) based on simulations of different kinds of noises but the collaboration of experts in the processing of real images would be, undoubtedly, of the utmost importance.

References


Machine GRAPHICS and VISION vol. 1, nos. 1/2, 1992, pp. 361–368
(Proceedings of 2nd GKPO'92 Conference, Nałęczów, May 18–23, 1992)
268

Statistical image segmentation


Machine GRAPHICS and VISION vol. 1, nos. 1/2, 1992, pp. 261–266
(Proceedings of 2nd GKPO’92 Conference, Nidzsków, May 18–23, 1992)