

EMBEDDING DIFFUSION IN VARIATIONAL BAYES: A TECHNIQUE FOR SEGMENTING IMAGES

GIUSEPPE BOCCIGNONE* and PAOLO NAPOLETANO†

*Natural Computation Lab
Dipartimento di Ingegneria dell'Informazione e Ingegneria Elettrica
Università di Salerno, via Ponte Melillo 1
84084 Fisciano (SA), Italy
*boccig@unisa.it
†pnapoletano@unisa.it*

MARIO FERRARO

*Dipartimento di Fisica Sperimentale
Università di Torino
via Pietro Giuria 1, Torino, Italy
ferraro@ph.unito.it*

In this paper, we discuss how image segmentation can be handled by using Bayesian learning and inference. In particular variational techniques relying on free energy minimization will be introduced. It will be shown how to embed a spatial diffusion process on segmentation labels within the Variational Bayes learning procedure so as to enforce spatial constraints among labels.

Keywords: Image segmentation; variational Bayes; diffusion equation; model selection.

1. Introduction

The process of image segmentation is generally understood as the partitioning of the observed data (pixels) into meaningful constituent parts (regions), which can be achieved by assigning region labels to pixels in accordance with a uniformity, or homogeneity criterion (color similarity, spatial proximity, etc.). A large number of techniques have been proposed, over the years, both for grey-level and color images, (for an in-depth survey, see Refs. 9 and 13) but so far there is no satisfactory solution.

Despite its intuitive definition, segmentation is, at the most general level, a hard problem since real world images are fundamentally ambiguous and our perception of an image changes over time. On the one hand, the aggregation of pixels into segments representing meaningful parts is a compelling challenge, such parts are often too complex to be characterized through low-level image features without taking into account prior knowledge of the scene and objects within the scene. On

the other hand, the use of object-based information, requires that objects have been identified, while identification in turn relies upon object segmentation. Further, the semantic interpretation of an image is highly subjective and application dependent. As commonly experienced in computer vision, it is more productive to adopt the minimalist view of segmentation as a process which results in a “reasonable” partitioning of the image, a hypothesis conveniently exploited by other visual routines.¹⁸

This minimalist version of the problem, still challenging but solvable, nevertheless requires that any algorithm, in order to be effective, must cope with uncertainties related to the data, the choice of useful features and the actions to be taken for achieving the proposed partitioning, while exploiting prior knowledge on the data if available.

To this end, probability theory, and, in particular, the Bayesian approach, offers a mathematically consistent way to formulate segmentation algorithms in terms of model based inference.^{6,9,10} The adoption of Bayesian methods is further motivated by the need of learning the parameters of the underlying models.

In this paper we will discuss how the problem of perceptual Bayesian learning and inference for segmentation purposes can be suitably managed by using variational techniques relying on free energy minimization.^{2,14} The use of Variational Bayes (VB) techniques is fairly recent in computer vision (see Ref. 10 for an in-depth discussion), and to the best of our knowledge there are only two attempts to exploit it for segmentation,^{16,22} but with some important limitations such as the lack of spatial constraints¹⁶ or trading off model generality for model tractability.²² Here, we will show how such limitations can be overcome by embedding a spatial diffusion process on segmentation labels within the VB learning procedure.

2. Segmentation in a Bayesian Perspective: Background

Segmentation, from a probabilistic viewpoint, can rather naturally be considered as a missing data problem.⁹ The complete data space is represented by a pair of random fields: $\mathbf{Y} = \{\mathbf{y}_n\}_{n=1}^N$ is the observed random field whose configuration (image) consists of the measurements at each random variable \mathbf{y}_n (pixel), which may be either a scalar or D -dimensional vector-valued; $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N$ represents a configuration of unobservable, hidden variables, where the value (label) of each random variable \mathbf{x}_n indicates to which region or object $k \in \{1, \dots, K\}$ each pixel belongs. Here n indexes the set of sites $\mathbf{S} = \{1, 2, \dots, N\}$, the square lattice domain of the image.

The observed data set \mathbf{Y} is assumed to be generated from hidden states \mathbf{X} and the segmentation process, starting from \mathbf{Y} , aims at estimating for each pixel the hidden object/class it belongs to. This implies *learning* the model, using the model to *infer* the partitioning probability and *deciding* the most reliable partitioning.

In a Bayesian setting, the generative model, indexed by $m \in \mathcal{M}$ within the set of models \mathcal{M} , is specified in terms of both a prior distribution over the causes (\mathbf{X}, Θ) ,

namely $P(X, \Theta|m)$, and the likelihood function $P(Y|X, \Theta, m)$. Thus, hidden and observable data are coupled by the generative model specified through the joint probability distribution $P(Y, X, \Theta|m) = P(Y|X, \Theta, m)P(X, \Theta|m)$.

Learning a generative model corresponds to making the probabilistic distribution of input data, implied by a model of parameters Θ , as close as possible to those actually observed. To this end, it is possible to derive the marginal distribution of the data generated under the model m (evidence) that has to be matched to the input distribution $P(Y)$

$$P(Y|m) = \int_{\mathbf{X}, \Theta} P(Y|X, \Theta, m)P(X, \Theta|m)dXd\Theta. \quad (1)$$

Once the parameters of the generative model have been learned, the recognition model is defined in terms of inverse probability,¹⁴ and *inference* of hidden variables X defining the partitioning of the image, is performed via Bayes' rule:

$$P(X|Y, \Theta, m) = \frac{P(Y|X, \Theta, m)P(X, \Theta|m)}{P(Y|m)}. \quad (2)$$

Finally, for a given pixel configuration Y , the best segmentation estimate \hat{X} can be recovered under some suitable extremum principle (e.g. minimum mean squared error, MMSE or maximum *a posteriori*, MAP) related to the posterior probability $P(X|Y, \Theta, m)$.

However, marginalization in Eq. (1) is often difficult because, in principle, all parameters of the model can be coupled; furthermore, the estimate \hat{X} can be difficult to compute without approximations. Thus, in general, the generative model cannot be easily inverted and it may not be possible to parametrize the posterior distribution.

A variational solution is to posit a simpler approximate distribution $Q(X, \Theta)$ that is consistent (same support) with $P(X, \Theta, Y)$ (in the following we drop model index m for notational simplicity). Any such distribution can be used to provide a lower bound to the evidence $P(Y)$, or equivalently to the log-likelihood $\mathcal{L}(Y) = \log P(Y)$, which can be rewritten as:

$$\begin{aligned} \mathcal{L}(Y) &= \int_{\mathbf{X}, \Theta} Q(X, \Theta) \log \frac{P(X, \Theta, Y)}{Q(X, \Theta)} dXd\Theta + \int_{\mathbf{X}, \Theta} Q(X, \Theta) \log \frac{Q(X, \Theta)}{P(X, \Theta|Y)} dXd\Theta \\ &= F(Q) + \text{KL}(Q||P) \end{aligned} \quad (3)$$

where $\text{KL}(Q||P)$ is the *Kullback–Leibler* divergence between the approximating distribution and the true posterior distribution, while $F(Q)$ represents a lower bound on $\mathcal{L}(Y)$. By definition $\text{KL}(Q||P) \geq 0$, being equal to 0 when $Q(X, \Theta) = P(X, \Theta|Y)$, which implies that $\mathcal{L}(Y) \geq F(Q)$. The “best” approximating distribution $Q^*(X, \Theta)$ is then the one that maximizes $F(Q)$, or equivalently minimizes the Kullback–Leibler divergence between the distribution $Q(X, \Theta)$ and the true joint posterior $P(X, \Theta|Y)$.

For notational simplicity, define the latent variables $Z = \{X, \Theta\}$. It is a common practice to restrict the family of Q so that they comprise only tractable distributions,¹⁴ for instance, those that can be factorized as $Q(Z) = \prod_{i=1}^M Q_i(Z_i)$ with $M = N_p + N$, N_p being the number of parameters in the set Θ .

It has been shown that the free-form variational optimization of $F(Q)$ with respect to the distributions Q_i provides the optimal solution²:

$$Q_j^*(Z_j) = \frac{\exp[I(Z_j)]}{\int \exp[I(Z_i)] dZ_i} \quad (4)$$

with $I(Z_j) = \int \log P(Z, Y) \prod_{i \neq j} Q_i(Z_i) dZ_i$. The variational approximation thus maximizes $F(Q)$ as a functional of the distribution $Q(X, \Theta)$, by iteratively maximizing F , with respect to each Q_j , $\frac{\partial F(Q)}{\partial Q_j} = 0$, $j = 1 \dots M$.

Note that the set of equations used to recover $Q_j^*(Z_j)$ is a set of coupled fixed point equations ($Q_j^*(Z_j)$ is computed in terms of $Q_i(Z_i)$), that require an iterative solution. In particular, for $Q(X, \Theta) = Q(X)Q(\Theta)$ and $Q(X) = \prod_{n=1}^N Q(x_n)$ the following holds.¹

Theorem 1. *Let m be a model with parameters Θ giving rise to an i.i.d. data set $Y = \{y_n\}_{n=1}^N$ with corresponding hidden variables $X = \{x_n\}_{n=1}^N$. A lower bound on the model log marginal likelihood is $F(Q) = \int_{X, \Theta} Q(X)Q(\Theta) \log \frac{P(X, \Theta, Y)}{Q(X)Q(\Theta)} dX d\Theta$, and this can be iteratively optimized by performing the following updates (superscript t denoting iteration number):*

VBE step:

$$Q^{t+1}(x_n) \propto \exp \left[\int_{\Theta} Q^t(\Theta) \log P(x_n, y_n | \Theta, m) d\Theta \right], \quad \forall n \quad (5)$$

VBM step:

$$Q^{t+1}(\Theta) \propto P(\Theta | m) \exp \left[\int_X Q^{t+1}(X) \log P(x_n, y_n | \Theta, m) dX \right]. \quad (6)$$

Moreover, the update rules converge to a local maximum of $F(Q)$.

These steps represent a Bayesian generalization of the E and M steps of the classic Expectation–Maximization (EM) algorithm¹⁴ and in the following it will be referred to as the VBEM algorithm.

As we will see, the distribution $Q(X)$ over hidden variables that approximates the true posterior $P(X|Y, \Theta)$ provides a natural form to estimate the Gaussian component weights (called in this case *variational responsibilities*) when the image is modeled via a Finite Gaussian Mixture (FGM) model, which is widely used in probabilistic image segmentation.^{9,16,22,23}

According to the FGM model, each pixel y_n is generated by one among K Gaussian distributions $\mathcal{N}(y_n; \boldsymbol{\mu}_k, \Lambda_k^{-1})$, with $\boldsymbol{\mu}_k, \Lambda_k$, the means and the precision matrix (inverse covariance) of the k th Gaussian, and likelihood

$$P(y_n|\Theta) = \sum_{k=1}^K \pi_k \mathcal{N}(y_n; \boldsymbol{\mu}_k, \Lambda_k^{-1}). \quad (7)$$

Here $\{\pi_k\}_{k=1}^K$ are the mixing coefficients, with $\sum_{k=1}^K \pi_k = 1$ and $\pi_k \geq 0$ for all k .

Denote $\Theta = \{\boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda\}$ the vector of parameters (random variables), with $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K$, $\boldsymbol{\mu} = \{\boldsymbol{\mu}_k\}_{k=1}^K$, $\Lambda = \{\Lambda_k\}_{k=1}^K$.

Each hidden variable $\mathbf{x}_n \in \mathbf{X}$ related to observation y_n , is a 1-of- K binary vector of components $\{x_{nk}\}_{k=1}^K$, in which a particular element x_{nk} is equal to 1 and all other elements are equal to 0, that is $x_{nk} \in \{0, 1\}$ and $\sum_k x_{nk} = 1$. In other terms, \mathbf{x}_n indicates which Gaussian component is responsible for generating pixel y_n , $P(y_n|x_{nk} = 1, \Theta) = \mathcal{N}(y_n; \boldsymbol{\mu}_k, \Lambda_k^{-1})$.

The FGM generative model (joint probability $P(\mathbf{Y}, \mathbf{X}, \Theta)$) is defined as follows (see Bishop² for details): $P(\mathbf{Y}, \mathbf{X}, \boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda) = P(\mathbf{Y}|\mathbf{X}, \boldsymbol{\mu}, \Lambda)P(\mathbf{X}|\boldsymbol{\pi})P(\boldsymbol{\pi})P(\boldsymbol{\mu}, \Lambda)$, where $P(\mathbf{Y}|\mathbf{X}, \boldsymbol{\mu}, \Lambda) = \prod_{n=1}^N P(y_n|x_n, \boldsymbol{\mu}, \Lambda) = \prod_{n=1}^N \prod_{k=1}^K \mathcal{N}(y_n, \boldsymbol{\mu}_k, \Lambda_k^{-1})^{x_{nk}}$, $P(\mathbf{X}|\boldsymbol{\pi}) = \prod_{n=1}^N P(\mathbf{x}_n|\boldsymbol{\pi}) = \prod_{n=1}^N \prod_{k=1}^K \pi_k^{x_{nk}}$, $P(\boldsymbol{\mu}, \Lambda) = \prod_{k=1}^K \mathcal{N}(\boldsymbol{\mu}_k; \mathbf{m}_0, (\beta_0 \Lambda_k)^{-1})\mathcal{W}(\Lambda_k; W_0, \nu_0)$, $P(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = C(\boldsymbol{\alpha}) \prod_{k=1}^K \pi_k^{\alpha_0 - 1}$.

Here the conjugate priors over model parameters $\boldsymbol{\mu}$, Λ and $\boldsymbol{\pi}$, namely $\mathcal{N}(\boldsymbol{\mu}_k; \mathbf{m}_0, (\beta_0 \Lambda_k)^{-1})$, $\mathcal{W}(\Lambda_k; W_0, \nu_0)$ and $\text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})$, are the Gaussian, Wishart and Dirichlet distributions, respectively,² and $\alpha_0, W_0, \nu_0, \beta_0, \mathbf{m}_0$ are the *hyperparameters* of the model.

VB learning considers the approximating distribution $Q(\mathbf{X}, \boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda)$ factorized as $Q(\mathbf{X})Q(\boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda) = Q(\mathbf{X})Q(\boldsymbol{\pi})Q(\boldsymbol{\mu}, \Lambda)$, and the lower bound $F(Q)$, is maximized by applying Eq. 4. Close form computation results in the following solutions for the factors of the variational posterior^{2,17}:

$$Q(\mathbf{X}) = \prod_{n=1}^N \prod_{k=1}^K q_{nk}^{x_{nk}}, \quad (8)$$

$$Q(\boldsymbol{\pi}) = C(\boldsymbol{\alpha}) \prod_{k=1}^K \pi_k^{(\bar{N}_k + \alpha_0 - 1)}, \quad (9)$$

$$Q(\boldsymbol{\mu}, \Lambda) = \prod_{k=1}^K \mathcal{N}(\boldsymbol{\mu}_k; \mathbf{m}_k, (\beta_k \Lambda_k)^{-1})\mathcal{W}(\Lambda_k; W_k, \nu_k), \quad (10)$$

where $q_{nk} \simeq P(k|y_n, \boldsymbol{\mu}_k, \Lambda_k^{-1})$ denote the *responsibilities*, each representing an approximation to the posterior probability of labeling pixel y_n as belonging to the k th class.

Unfortunately, the FGM model relies upon the assumption of independence of pixel data and class labels, which is inadequate for images where some form of spatial constraints should be introduced. Spatial constraints can be introduced

explicitly but this usually makes very complex the underlying graphical model and the learning/inference procedures.^{11,22,23} In Boccignone *et al.*,⁵ to keep the model structure simple it has been proposed to introduce spatial constraints while performing the VB learning algorithm. A heuristic justification of the method was presented, graphically showing how each step of the resulting algorithm is guaranteed to increase or leave unchanged the lower bound F on the fixed marginal likelihood. The method can be formally derived as follows.

3. A Method for Embedding Spatial Constraints in VBEM for Gaussian Mixture Models: Theory

A property of the variational interpretation of the EM algorithm is that at each step we are allowed to assign any distribution $Q(\mathbf{x}_n)$ to individual pixels as long as this increases the lower bound F .

In order to design such transformation, it is convenient to define the following quantities in analogy with statistical physics, that allow a deeper insight of the physical meaning of the bounding functional $F(Q)$, namely: the Helmholtz free energy

$$F_H = -\log Z = -\mathcal{L}(Y) = -\log \int_{\mathbf{X}, \Theta} P(Y, \mathbf{X}, \Theta) d\Theta d\mathbf{X}; \quad (11)$$

the Gibbs' variational free energy

$$F_G = -F(Q) = -\int_{\mathbf{X}, \Theta} Q(\mathbf{X}, \Theta|Y) \log \frac{P(\mathbf{X}, \Theta, Y)}{Q(\mathbf{X}, \Theta|Y)} d\mathbf{X} d\Theta; \quad (12)$$

the average energy (internal energy);

$$U(Q) = -\int_{\mathbf{X}, \Theta} Q(\mathbf{X}, \Theta|Y) \log P(\mathbf{X}, Y, \Theta|m) d\mathbf{X} d\Theta; \quad (13)$$

the entropy

$$S(Q) = -\int_{\mathbf{X}, \Theta} Q(\mathbf{X}, \Theta|Y) \log Q(\mathbf{X}, \Theta|Y) d\mathbf{X} d\Theta. \quad (14)$$

Then, by taking into account Eq. (3), the following holds:

$$F_G = F_H + \text{KL}(Q||P) = U(Q) - S(Q), \quad (15)$$

which says that the Kullback–Leibler distance will be zero, when the variational Gibbs' free energy F_G is equal to the Helmholtz free energy F_H . From a statistical physics point of view, the problem of learning is thus the problem of minimizing the Gibbs' free energy with respect to the distribution $Q(\mathbf{X}, \Theta)$.

Assume that after a VBE step the new distribution $Q(\mathbf{X})$ has been obtained via Eq. (5). Then we can apply any transformation $\mathcal{G}(Q) \rightarrow \tilde{Q}$ such that the Gibbs' free energy F_G decreases (i.e. $F(Q) = -F_G$ increases). For instance, recalling that

$F_G = U(Q) - S(Q)$, one can choose a mapping $\mathcal{G}(Q)$ such that the entropy $S(Q)$ [Eq. (14)] increases. This can be stated more precisely as follows.

Lemma 1. *Consider the iterative optimization of $F(Q)$ as performed through Eqs. (5) and (6). If a transformation $\mathcal{G}(Q(X))$ is applied such that*

$$S(\mathcal{G}(Q(X))) \geq S(Q(X)), \quad (16)$$

then the update rules converge to a local maximum of $F(Q)$.

Proof. By using factorization $Q(X, \Theta) = Q(X)Q(\Theta)$, the normalization constraints $\int Q(X)dX = 1$ and $\int Q(\Theta)d\Theta = 1$, and assuming Θ fixed, Eq. (14) can be rewritten as:

$$S(Q) = - \int Q(X) \log Q(X) dX - \int Q(\Theta) \log Q(\Theta) d\Theta = S(Q(X)) + \text{const.} \quad (17)$$

Thus, since $F_G = U(Q) - S(Q)$ and $F_G = -F(Q)$, any transformation such that Eq. (16) holds, decreases Gibbs' free energy F_G and increases the lower bound $F(Q)$, i.e. $F(G(Q)) \geq F(Q)$ and the conclusion holds as a direct consequence of Theorem 1. \square

This simple result indicates a viable solution to embed spatial constraints in variational learning. Assume that after a VBE step [Eq. (5)], the new distribution $Q(X)$ has been obtained.

Proposition 1. *Let \mathcal{G}_t be an irreversible transformation parametrized by t , with $t \geq 0$, acting as a translation in the scale space. Then \mathcal{G}_t is instantiated by either isotropic or anisotropic diffusion and the transformation $\mathcal{G}_t : Q(X) \rightarrow \mathcal{G}_t(Q(X))$ decreases the Gibbs' free energy F_G along the variational optimization.*

Proof. We first compute $S(Q(X)) = -E_{Q(X)} [\log Q(X)]$. From Eq. (8), $\log Q(X) = \sum_{n=1}^N \sum_{k=1}^K x_{nk} \log q_{nk}$. Since $E[x_{nk}] = q_{nk}$ (see Ref. 2 for detailed discussion), then the following holds:

$$S(Q(X)) = - \sum_{k=1}^K \sum_{n=1}^N q_{nk} \log q_{nk} = \sum_{k=1}^K S_k(Q(X)), \quad (18)$$

where $S_k(Q(X)) = \sum_{n=1}^N q_{nk} \log q_{nk}$ is a spatial entropy on responsibilities (segmentation labels). Since the total entropy can be written as the sum of the entropies S_k , each related to a label, the treatment can be restricted to the action of \mathcal{G}_t on a single probability q_{nk} . Transformation \mathcal{G}_t is such that, as t grows, probabilities q_{nk} are shifted toward increasingly coarse scales of resolution. To make the argument more precise assume the domain on which q_{nk} is defined to be a continuum \mathcal{D} , in other words replace the discrete variable n with $\mathbf{r} \in \mathcal{D}$; furthermore, \mathcal{G}_t generates a family of functions q_k , and each element of the family will depend also on t , so that, in conclusion q_k is defined on the product space $\mathcal{D} \times T$, $q_k : (\mathbf{r}, t) \rightarrow q_k(\mathbf{r}, t)$.

The action of \mathcal{G}_t on q_k is determined by its differential operator $\frac{\partial}{\partial t}$: to make \mathcal{G}_t a transformation from fine to coarse scales of resolution it is enough to set $\frac{\partial}{\partial t}$ equal to a diffusion operator so that $\frac{\partial}{\partial t} = \text{div}(g(\mathbf{r})\nabla)$, where g is a function that specifies the type of diffusion process under consideration. Then we obtain a system of partial differential equations, one for each k ,

$$\frac{\partial q_k(\mathbf{r}, t)}{\partial t} = \nabla \cdot (g(\mathbf{r})\nabla q_k(\mathbf{r}, t)), \quad (19)$$

where, by virtue of latent variable factorization, each equation is independent of the others.

If g is a constant, Eq. (19) is the usual isotropic diffusion equation, $\frac{\partial q_k(\mathbf{r}, t)}{\partial t} = g \cdot \nabla^2 q_k(\mathbf{r}, t)$, and g is just the diffusion coefficient, whereas anisotropic diffusion is obtained by requiring $g(\cdot)$ to be a monotonically decreasing function of $\|\nabla q_k(\mathbf{r}, t)\|$, the norm of the gradient of q .

Isotropic and anisotropic diffusion increase spatial entropy^{4,8} and it has been shown²⁰ that, in both cases, the functional $-S_k = \sum_{n=1}^N q_{nk} \log q_{nk}$ is a Lyapunov functional, decreasing under the transformation for $t \rightarrow \infty$. In conclusion then, for each component k , Eq. (19) increases the k th entropy S_k , thus giving rise to a growth of the total entropy $S(\mathcal{G}_t Q(X)) = \sum_{k=1}^K S_k(\mathcal{G}_t Q(X))$ as t increases. Therefore, condition specified by Eq. (16) is satisfied, and lower bound optimization is achieved, see Lemma 1 and Theorem 1. \square

It is worth noting that isotropic diffusion is sufficient to guarantee a spatial conditioning of labels pertaining neighboring pixels³ but does not allow selection of the optimal label in that the solution of [Eq. (19)] would be $q_{nk} = \text{constant}$ for all k , meaning that all pixel have the same probability with assigned label k .

Note that neighboring pixels, belonging to the same region, will have the same probability assigned a given label k and that labels at boundaries between regions should be characterized by an abrupt change of probability values. Thus, each q_k should be a piecewise constant function across the image and this result can be achieved²⁰ by a system of k anisotropic diffusion equations. Because of the form of g , small labeling differences of q_k among pixels close to each other are smoothed out, since diffusion is allowed, whereas large variations are preserved.

Summing up, when $Q(X)$ has been modified to account for spatial constraints through a diffusion step (VBD step), it can be used in the VBM step to maximize the negative free energy $F(Q)$ with respect to the parameters. We name this procedure the Variational Bayes Diffused EM (VBDEM).

4. The Learning and Segmentation Algorithm

Standard VB learning of the FGM model^{2,17} amounts to an iterative update of hidden variables and parameters distributions [Eqs. (8)–(10)]. This entails a solution^{2,17} in which the computation of the approximating posteriors q_{nk}

(VBE step)

$$q_{nk} = e^{(-\frac{D}{2} \log 2\pi)} \tilde{\pi}_k \tilde{\Lambda}_k^{1/2} e^{(-\frac{1}{2} \nu_k (y_n - m_k)^T W_k^{-1} (y_n - m_k))} e^{(-\frac{D}{2\beta_k})} \quad (20)$$

and that of hyperparameters (VBM step), obtained by adding data counts to prior counts,

$$\begin{aligned} \alpha_k &= \alpha_0 + \overline{N}_k, \quad \beta_k = \beta_0 + \overline{N}_k, \quad m_k = \frac{\beta_0 m_0 + \overline{N}_k \overline{\mu}_k}{\beta_k}, \\ W_k &= \overline{N}_k \overline{\Sigma}_k + \frac{\overline{N}_k \beta_0}{\beta_k} (\overline{\mu}_k - m_0)(\overline{\mu}_k - m_0)^T + W_0, \quad \nu_k = \nu_0 + \overline{N}_k, \end{aligned} \quad (21)$$

is repeated until convergence.^{2,17}

To compute the hyperparameter update, the following statistics of the observed data with respect to the q_{nk} need to be calculated^{2,17}:

$$\overline{\pi}_k = \frac{1}{N} \sum_{n=1}^N q_{nk}, \quad \overline{N}_k = N \overline{\pi}_k, \quad (22)$$

$$\overline{\mu}_k = \frac{1}{\overline{N}_k} \sum_{n=1}^N q_{nk} y_n, \quad \overline{\Sigma}_k = \frac{1}{\overline{N}_k} \sum_{n=1}^N q_{nk} (y_n - \overline{\mu}_k)(y_n - \overline{\mu}_k)^T. \quad (23)$$

Spatial constraints on the distribution of segmentation labels $Q(X)$ are applied through the discretized version of diffusion equation (19):

$$q_{nk}(\tau + 1) = q_{nk}(\tau) + \lambda(\nabla \cdot (g(\|\nabla q_{nk}\|) \nabla q_{nk}(\tau))). \quad (24)$$

At convergence, segmentation is obtained by setting $y_n = \boldsymbol{\mu}_{k^*}$ where $k^* = \arg \max_k q_{nk} \simeq \arg \max_k P(k|y_n, \boldsymbol{\mu}_k, \Lambda_k^{-1})$.

The VBDEM procedure is summarized in Algorithm 1. The initialization of the conjugate prior parameters $\alpha_0, W_0, \nu_0, \beta_0, m_0$ is performed similarly to Ref. 17, while the negative free energy F is straightforwardly computed via closed-form solution.^{2,17}

The computational complexity of a single iteration of the VBDEM algorithm in the inference/learning stage is determined by the order of complexity of the VBE and VBM steps plus the order of complexity of the VBD step. Recalling that K, D, N denote the number of Gaussian components, the pixel dimension (for color images $D = 3$) and the number of pixels, respectively, the VBE and VBM steps have the same order of complexity of the standard EM algorithm, namely $O(KD^2N)$. For what concerns the VBD step, the order of complexity for diffusing on K components is in principle $O(KTN)$; however, since we are not committed here to achieve the fixed point solution of the equation but rather to a filtering/regularization operation, we set \mathcal{T} to be a constant, small number of iterations, then $O(KTN) \simeq O(KN)$. Thus, the order of complexity of a single inference and learning iteration is $O(KD^2N) + O(KN) = O(KD^2N)$, and if

Algorithm 1 Learning and Segmentation via VBDEM.

{Spatially constrained inference and learning}

Initialize prior parameters:
 $\alpha_0 = 0.001$, $W_0 = 0.01D\mathbf{I}$, $\nu_0 = D$, $\beta_0 = 1$, $m_0 = \frac{1}{N} \sum_n y_n$;
Initialize responsibilities $q_{nk}^{(0)}$ via *k-means* algorithm;
Initialize statistics $\bar{N}_k^{(0)}$, $\bar{\Sigma}_k^{(0)}$, $\bar{\mu}_k^{(0)}$ e $\pi_k^{(0)}$ according to Eq. 22 ;
Initialize hyperparameters $\alpha_k^{(0)}$, $W_k^{(0)}$, $\nu_k^{(0)}$, $\beta_k^{(0)}$, $m_k^{(0)}$ according to Eq. 21;
Initialize lower bound $F^{(0)}$; $F^{new} \leftarrow F^{(0)}$;
 $t \leftarrow 0$;
repeat
 $F^{old} \leftarrow F^{new}$;
 { VBE-step }
 for ($n = 1, \dots, N$) **do**
 for ($k = 1, \dots, K$) **do**
 Compute the posteriors $q_{nk}^{(t)}$ according to Eq. 20;
 { VBD-step }
 for ($k = 1, \dots, K$) **do**
 for ($\tau = 1, \dots, T$) **do**
 for ($n = 1, \dots, N$) **do**
 Diffuse responsibilities $q_{nk}^{(t)}(\tau + 1)$ via anisotropic diffusion Eq. 24;
 { VBM-step }
 for ($k = 1, \dots, K$) **do**
 Compute statistics $\bar{\pi}_k^{(t)}$, $\bar{N}_k^{(t)}$, $\bar{\mu}_k^{(t)}$, $\bar{\Sigma}_k^{(t)}$ according to Eq. 22;
 for ($k = 1, \dots, K$) **do**
 Update hyperparameters $\beta_k^{(t)}$, $m_k^{(t)}$, $W_k^{(t)}$, $\nu_k^{(t)}$ via Eq. 21;
 Compute lower bound $F^{(t)}$; $F^{new} \leftarrow F^{(t)}$;
 $t \leftarrow t + 1$;
until $|F^{new} - F^{old}| < \epsilon$
{ Segmentation }
for ($n = 1, \dots, N$) **do**
 $k^* \leftarrow \arg \max_k q_{nk}^{(t)}$;
 $y_n \leftarrow \mu_{k^*}$;

convergence is reached after a number T of iterations, the overall complexity eventually is $O(TKD^2N)$. It is worth remarking that the marginal real time increase, occurring in practice due to diffusion computations, can be further reduced by adopting more sophisticated discretization schemes in place of the finite difference scheme used in Eq. (24) (e.g. Additive Operator Splitting²¹). Finally, the complexity of the segmentation step is $O(KN)$, the responsibilities q_{nk} being available from previous steps. To give an idea of the actual execution time, for a 256×256 color image ($D = 3$, $N = 65536$), $K = 7$ components, $T = 10$ diffusion steps, and

convergence achieved after $T = 30$ iterations, the elapsed time is 63.6 s; this result is obtained by nonoptimized Matlab code executing under the Mac OS X 10.4.11 operating system running on a 2 GHz Intel Core Duo Processor, 2 GB RAM.

5. Simulation

We have experimented the method on different kinds of sports, natural and medical images. Here, due to space limitations, we present one significant example for each category, namely the *Players*, *Landscape* and *Skin cancer* images, shown in Figs. 1(a), 2(a) and 4(a), respectively. The *Players* image is a complex one due to the variety of colors and shape details present in the original scene; the *Landscape* provides a difficult example of a set of regions subtly distinguished by color shading. Finally, the *Skin cancer* image is part of a set of images for which the ground-truth segmentation is available. For each image we compare results obtained by using the EM, VBEM and VBDEM algorithms.

The input to the algorithms is an RGB image which is converted to the YCrCb color space. Initialization of the VBEM algorithm is the same as the initialization of VBDEM previously described. The approximate posteriors q_{nk} are initialized by using few iterations (5) of the k -means algorithm²; the same number of iterations are used to initialize the EM algorithm. Convergence condition $|F^{\text{new}} - F^{\text{old}}| < \epsilon$, is controlled by setting $\epsilon = 10^{-4}$.

For what concerns the VBD step, the conductance function g can have a quite general form, but it must be such that label boundaries are preserved, and numerical stability guaranteed. Here we set $g(\nabla q_{nk}) = |\nabla q_{nk}|^{-9/5}$, $\lambda = 0.01$ and a number of $\mathcal{T} = 10$ iterations is used. The functions $q_{nk}(\tau)$ are renormalized after each iteration so that their sum is one.

The optimal number K of classes — the model selection problem that for the FGMs corresponds to the selection of the correct number of Gaussians — is determined for each image via cross-validation based on Gibbs' free energy minimization.¹⁷ This procedure is motivated by the fact that the Bayesian Information Criterion for model selection is recovered from the negative free energy.^{1,2,14} In Fig. 3, a demonstration of such technique over 10 classes, for the *Landscape* image, is reported; here the best choice is $K = 6$. By applying the same criterion, $K = 6$ and $K = 5$ were selected for the *Players* and *Skin cancer* images, respectively.



Fig. 1. Segmentation results. (a) The original image of the *Players*; (b) EM; (c) VBEM; (d) VBDEM.

Experimental results obtained with the different methods for these images are reported in Figs. 1(a) and 2(a).

First, and most importantly, it should be noted that, by using the vector μ_k as the color to represent the region of class k , the segmentation obtained with VBDEM is chromatically more coherent with the original image, as it can be seen by comparing the results obtained by standard EM [Figs. 1(b) and 2(b)], VBEM method [Figs. 1(c) and 2(c)], and VBDEM method [Figs. 1(d) and 2(d)]. In fact, it is readily apparent the higher perceptual significance and the reliability of the VBDEM results, [Figs. 1(d) and 2(d)], as regards region classification and spatial uniformity.

In particular, results obtained for the *Players* image illustrate the VBDEM performance with respect to chromatic faithfulness and detail preservation. Further, the results achieved on *Landscape* show that though the VBEM method provides a better performance than classic EM (which merges some regions of different colors), nevertheless it cannot take advantage of spatial constraint propagation controlled in VBDEM by anisotropic diffusion on responsibilities (compare mountain region completion, Figs. 2(c) and 2(d)).

It is worth remarking at this point that the evaluation of segmentation algorithms thus far has been subjective. This is due to image segmentation being an ill-defined problem (see discussion in Sec. 1); except for specific domains where one can resort to domain expert's knowledge, there is no unique ground-truth segmentation of an image against which the output of an algorithm may be compared. In principle, in the absence of a unique ground-truth segmentation, the

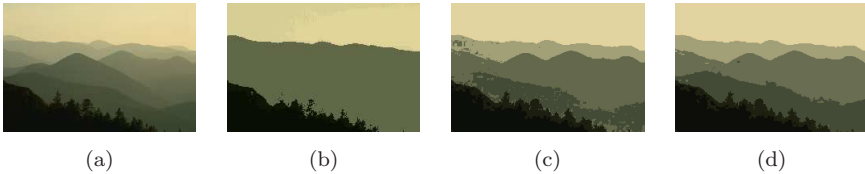


Fig. 2. Segmentation results. (a) The original image of the *Landscape*; (b) EM; (c) VBEM; (d) VBDEM.

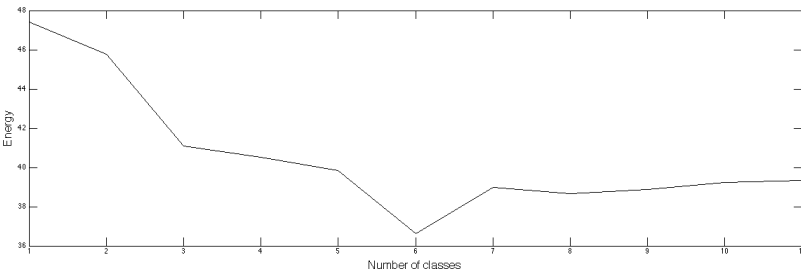


Fig. 3. Model selection for the *Landscape* image. The x -axis represents the number of classes, the y -axis represents the Gibbs' free energy function on a log-scale. In this case, the minimum is achieved for $K = 6$.

comparison should be made against the set of all possible perceptually consistent interpretations of an image; some advance in this direction has been recently reported.¹⁹ However, this endeavor is far from trivial and certainly beyond the scope of this paper, mainly focused on theoretical aspects of segmentation in a Bayesian framework.

Nevertheless, it is possible in our case to give a quantitative insight of the performance of the proposed method by exploiting the *Skin cancer* image. The ground-truth is shown in Fig. 4(b), where the two borders of interest (the cyan outer contour and the yellow inner contour) identified by a dermatologist have been overlapped on the original image.

According to Huang and Dom,¹² segmentation performance can be evaluated in terms of the accuracy of the extracted region boundaries. To this end, for each point \mathbf{r} of a computed boundary, the minimum Euclidean distance from \mathbf{r} to all the points in the corresponding ground-truth boundary is calculated. This provides a distance distribution from which a number of statistics can be derived, such as its mean and standard deviation; a perfect match between two borders should yield zero mean and zero standard deviation. We apply this procedure for both the outer and the inner borders marked in the ground-truth [Fig. 4(b)], and the corresponding borders in the segmented images [Figs. 4(c)–4(e)]. Results are reported in Table 1. The median of the distance distribution is also given since a large standard deviation may reveal the existence of outliers, in which case the median provides a better indication in terms of the accuracy of the segmentation.

The *Skin cancer* image belongs to a data set of 20 dermatologic images for which the ground-truth traced by an expert is available; the results shown in the table are representative of those obtained on the whole data set.

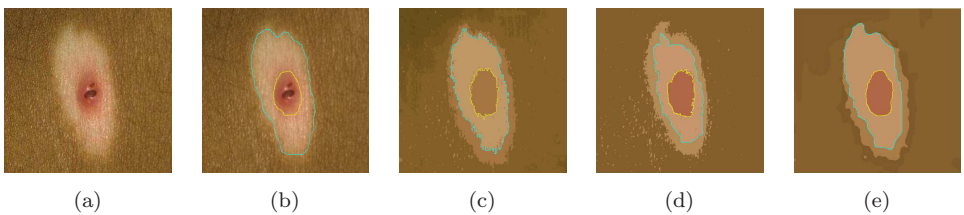


Fig. 4. Segmentation results. (a) The *Skin cancer* original image; (b) the ground-truth traced by a specialist; (c) EM; (d) VBEM; (e) VBDEM.

Table 1. Boundary-based evaluation of the three segmentation algorithms.

	Mean1	Std1	Median1	Mean2	Std2	Median2
EM	53.55	37.27	50.07	60.43	36.64	66.94
VBEM	15.26	5.77	15.03	2.36	1.84	2
VBDEM	5.78	4.30	5	1.77	1.31	1.41

Note: Mean1, Std1, Median1 refer to the outer border distance distribution; Mean2, Std2, Median2 refer to the inner border.

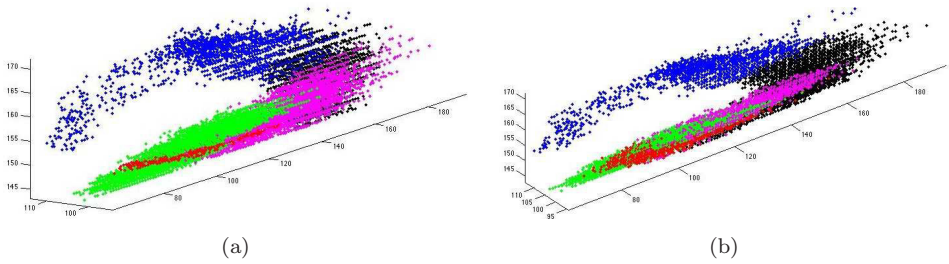


Fig. 5. A clustered representation of pixel label assignment obtained via (a) VBEM and (b) VBDEM, respectively. Each color (red, green, blue, magenta and black) denotes one among the five segmentation classes.

It is worth noting that by inspecting Fig. 4(d), we find that apparently the number of classes shown by the VBEM result seems to be equal to 4, in disagreement with the optimal number ($K = 5$) selected through cross-validation; in contrast, VBDEM segmentation [Fig. 4(e)] agrees with the exact number. This “missing label” phenomenon can be explained by representing the five segmentation labels as colors in a 3D (YCrCb) color space. To this end, each pixel is associated to a point in this space according to its label k (Fig. 5), and thus colored with the color identifying the k th class. As it can be seen from Fig. 5(a), the actual number of classes for VBEM is $K = 5$, but the class represented by “red” is assigned to few pixels, whereas many more are assigned to the overlapping “green” class (in some sense the “green” class wins over the “red” one). By contrast, Fig. 5(b) obtained through VBDEM shows how the anisotropic diffusion acts on class assignments by preserving the “red” class boundaries, thus providing a more balanced distribution of labels with respect to VBEM.

6. Concluding Remarks

This paper contributes a novel approach to image segmentation in which a Variational Bayes technique is spatially constrained in order to overcome drawbacks due to independent pixel labeling.¹⁶

The proposed VBDEM algorithm is somehow related to those approaches in the classic Maximum Likelihood (ML) setting in which prior terms have been incorporated within the EM algorithm so as to maximize a log-posterior probability instead of log-likelihood, see for instance, Ref. 11 or 23. However here, we are not working in a ML setting, but, rather, in a full Bayesian framework where parameters are treated as random variables and a distribution is derived for each of them; this way we avoid the problem of overfitting and achieve a regularized solution. Meanwhile, because of the generality of the proposed method we are not concerned with either designing specific priors or trading off model design for integrability constraints.²² This allows the method to be adopted for a variety of fields and different kinds of images. Promising results have been obtained for medical images, thus indicating

a potential field of application. Nevertheless, by resorting to optimized code (a C language implementation should decrease the actual execution time of an order of magnitude), the algorithm is expected to be suitable for video analysis, where at each frame parameter initialization may be provided by statistics computed from the previous frame.

Interestingly enough, Nasios and Bors¹⁶ have considered the unconstrained VBEM algorithm as a learning procedure for a Gaussian neural network, acting at each pixel as a competitive process among the k different labels. In the algorithm we propose here, competition is integrated with a cooperation in terms of a diffusion step within sites on the same labeling plane; it should be noted that both competitive and cooperative processes occur in nature for the formation of patterns,¹⁵ thus, in some sense, recognition, to be effective, must feature both competitive and cooperative elements.

Finally, the problem of model selection has been addressed here through cross-validation via Gibbs' free energy minimization.¹⁷ However, it should be noted that model selection is naturally handled in the Bayesian framework,^{2,14} and could be straightforwardly incorporated here along the iterations of the learning step, following suggestions proposed by Corduneanu and Bishop.⁷

References

1. M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, The Gatsby Computational Neuroscience Unit, University College London, London, UK (2003).
2. C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, NY, 2006).
3. G. Boccignone and M. Ferraro, An information-theoretic approach to interactions in images, *Spat. Vis.* **12** (1999) 345–362.
4. G. Boccignone, M. Ferraro and T. Caelli, Encoding visual information from anisotropic transformations, *IEEE Trans. Patt. Anal. Mach. Intell.* **23** (2001) 207–211.
5. G. Boccignone, M. Ferraro and P. Napoletano, A variational Bayes approach to image segmentation, *Advances in Brain, Vision, and Artificial Intelligence. Second Int. Symp.*, eds. F. Mele, G. Ramella, S. Santillo and F. Ventriglia, Lecture Notes in Computer Science, Vol. 4729 (2007), pp. 234–243.
6. N. Chater, J. B. Tenenbaum and A. Yuille, Probabilistic models of cognition: conceptual foundations, *Trends Cogn. Sci.* **10** (2006) 287–291.
7. A. Corduneanu and C. M. Bishop, Variational Bayesian model selection for mixture distributions, *Proc. Eight Int. Conf. Artif. Intell. Stat.* (2001), pp. 27–34.
8. M. Ferraro, G. Boccignone and T. Caelli, On the representation of image structures via scale-space entropy conditions, *IEEE Trans. Patt. Anal. Mach. Intell.* **21** (1999) 1199–1203.
9. D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach* (Prentice Hall International, 2002).
10. B. J. Frey and N. Jojic, A comparison of algorithms for inference and learning in probabilistic graphical models, *IEEE Trans. Patt. Anal. Mach. Intell.* **27** (2005) 1392–1416.

11. S. S. Gopal and T. J. Hebert, Bayesian pixel classification using spatially variant finite mixtures and the generalized EM algorithm, *IEEE Trans. Image Proc.* **7** (1998) 1014–1028.
 12. Q. Huang and B. Dom, Quantitative methods of evaluating image segmentation, *IEEE Int. Conf. Image Processing* **3** (1995) 53–56.
 13. L. Lucchese and S. K. Mitra, Color image segmentation: a state-of-the-art survey, *Proc. Indian Nat. Sci. Acad. (INSA-A)* **67** (2001) 207–221.
 14. D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge, UK, 2002).
 15. J. D. Murray, *Mathematical Biology* (Springer-Verlag, Berlin, 2002).
 16. N. Nasios and A. G. Bors, Variational learning for Gaussian mixture models, *IEEE Trans. Syst. Man Cybern.-B* **36** (2006) 849–862.
 17. W. Penny, Variational Bayes for d-dimensional Gaussian mixture models, Technical Report, Wellcome Department of Cognitive Neurology, University College, London, UK (2001).
 18. Z. Tu and S.-C. Zhu, Image segmentation by data-driven Markov chain Monte Carlo, *IEEE Trans. Patt. Anal. Mach. Intell.* **24** (2002) 657–673.
 19. R. Unnikrishnan, C. Pantofaru and M. Hebert, Toward objective evaluation of image segmentation algorithms, *IEEE Trans. Patt. Anal. Mach. Intell.* **29** (2007) 929–944.
 20. J. Weickert, Applications of nonlinear diffusion in image processing and computer vision, *Acta Math. Univ. Comenianae* **70** (2001) 33–50.
 21. J. Weickert, B. M. ter Haar Romeny and M. A. Viergever, Efficient and reliable schemes for nonlinear diffusion filtering, *IEEE Trans. Imag. Proc.* **7** (1998) 398–410.
 22. M. W. Woolrich and T. E. Behrens, Variational Bayes inference of spatial mixture models for segmentation, *IEEE Trans. Med. Imag.* **25** (2006) 1380–1391.
 23. Y. Zhang, M. Brady and S. Smith, Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm, *IEEE Trans. Med. Imag.* **20** (2001) 45–57.
-



Giuseppe Boccignone received the laurea degree in theoretical physics from the University of Turin (Italy) in 1985. In 1986, he joined Olivetti Corporate Research, Ivrea, Italy. From 1990 to 1992, he served as a Chief Researcher of the Computer Vision Lab at CRIAI, Naples, Italy. From 1992 to 1994, he held a Research Consultant position at Research Labs of Bull HN, Milan, Italy, leading projects on biomedical imaging. In 1994, he joined the Dipartimento di Ingegneria dell'Informazione e Ingegneria Elettrica, University of Salerno, Italy, where he is currently an Associate Professor of Computer Science.

He has been active in the field of computer vision, image processing, and pattern recognition. He is a Member of the IEEE Computer Society.

His current research interests lie in active vision, Bayesian models for computational vision, cognitive science and medical imaging.



Paolo Napoletano received the laurea degree in telecommunication engineering from the University of Naples Federico II, Naples, Italy, in 2003, and the Ph.D. degree in information engineering from the University of Salerno, Italy, in 2007. He currently holds

a post-doc position at the Natural Computation Lab, Dipartimento di Ingegneria dell'Informazione e Ingegneria Elettrica, University of Salerno. He is Member of the IEEE Computer Society.

His current research interests lie in active vision, Bayesian models for computational vision and ontology building.



Mario Ferraro received the laurea degree in theoretical physics from the University of Turin (Italy) in 1973. He has worked in various Universities in England, Canada, Germany and United States, focussing his research

on fuzzy sets theory, human vision, invariant pattern recognition and computational vision. Presently, he is an Associate Professor of Physics at the University of Turin. He is a Member of the IEEE Computer Society.

His research interests include image and shape analysis, cellular biophysics and the theory of self-organizing systems.