Diffused expectation maximisation for image segmentation

G. Boccignone, M. Ferraro and P. Napoletano

Diffused expectation maximisation is a novel algorithm for image segmentation. The method models an image as a finite mixture, where each mixture component corresponds to a region class and uses a maximum likelihood approach to estimate the parameters of each class, via the expectation maximisation algorithm, coupled with anisotropic diffusion on classes, in order to account for the spatial dependencies among pixels.

Introduction: Any image can be considered as a set of N unlabelled samples $F=\{f_1, f_2, \ldots, f_N\}$, on a 2-D discrete support $\Omega \subseteq Z^2$, with $N=|\Omega|$. Thus, an image segmentation/classification problem can be defined in probabilistic terms as the problem of assigning a label k to each site i, given the observed data F, and where each label $k \in [1, \ldots, K]$ defines a particular region/model. Different models are selected with probability $P(k)$, and a sample is generated with probability distribution $p(f_i|k, \theta)$ where $\theta=\{\theta_k, k=1, \ldots, K\}$ and θ_k is the vector of the parameters associated with label k. Thus $p(f_i|k, \theta)$ is the probability of f_i given the parameters of all models and the fact that we have selected model (label) k. Each image can be conceived as drawn from a mixture density, so that, for any site (pixel), $p(f_i|\theta)=\sum_{k=1}^{K} p(f_i|k, \theta)P(k)$, and the likelihood of the data is $\mathcal{L}=p(f|\theta)=\prod_{i=1}^{N} p(f_i|\theta)$. For clarity, we define $p(f_i)$ and $\pi(f)$, two probability distributions; the former is the probability that a given grey level f is assigned to pixel i, so that $\sum_i p(f_i)=1$, whereas the latter is the probability that, given any pixel i, it has grey level f.

Image segmentation can be achieved by finding the set of labels that maximise the likelihood $\mathcal{L}=\prod_{i=1}^{N}\sum_{k=1}^{K} p(f_i|k, \theta)P(k)$, or, equivalently, $(1/N)\log \mathcal{L}=(1/N)\sum_{i=1}^{N} \log \sum_{k=1}^{K} p(f_i|k, \theta)P(k)$. By the weak law of large numbers and the ergodic theorem the term to be maximised can be written as $E[\log \sum_{k=1}^{K} p(f_i|k, \theta)P(k)]=\sum_{f=0}^{L-1}(\pi(f)\log \sum_{k=1}^{K} p(f_i|k, \theta)P(k))$, E being the expectation operator and L the number of grey levels (e.g. $L=256$). Simple manipulations lead to

$$\frac{1}{N}\log \mathcal{L}=\sum_{f=0}^{L-1}\pi(f)\log \pi(f)-\sum_{f=0}^{L-1}\left(\pi(f)\log \frac{\pi(f)}{\sum_{k=1}^{K} p(f_i|k, \theta)P(k)}\right) \tag{1}$$

Hence a straightforward maximisation of $\log \mathcal{L}$ can be obtained by minimising the second term of the last expression, namely the Kullback-Leibler (KL) distance $D(\pi(f)\|\sum_{k=1}^{K} p(f_i|k, \theta)P(k))$ between distributions $\pi(f)$ and $\sum_{k=1}^{K} p(f_i|k, \theta)P(k)$, while holding the first term fixed. This is exactly what is performed by the classic expectation maximisation (EM) algorithm [1] which minimises the KL distance between the manifold of the observed data and that of the true distribution.

Alternatively, one could attempt a multistep approach by iteratively minimising the entropy $H(f)=-\sum_{f=0}^{L-1}\pi(f)\log \pi(f)$, while holding $p(f|k, \theta)$, $P(k)$ fixed, and then minimising the KL distance D, while keeping $H(f)$ fixed. In particular, from a segmentation standpoint, it is interesting to reformulate entropy $H(f)$ in terms of spatially dependent conditional probabilities $p(k|f_i)$. To this aim, first note that minimising H corresponds to maximising the spatial entropy $H_s=-\sum_{i=1}^{N} p(f_i)\log p(f_i)$. In fact, probabilities $p(f_i)$ and $\pi(f)$ can be estimated as $p_i=p(f_i)\simeq(f_i/f_{tot})$ and $f_{tot}=\sum_{i=1}^{N} f_i$ and $\pi(f)\simeq(n_f/N)$ where the following relations hold: $f_{tot}=\sum_{i=1}^{N} f_i=\sum_{f=0}^{L-1} n_f f$.

With the approximation introduced above $H(f)$ can be written as $H(f)=-\sum_{f=0}^{L-1}(n_f/N)\log(n_f/N)$ and similarly $H_s=-\sum_{f=0}^{L-1} n_f(f/f_{tot})\log(f/f_{tot})$. By applying the approximation $\log x \simeq (x-1)$ it is not difficult to show that $H(f)\simeq 1-\sum_{f=0}^{L-1}[p(f)]^2$ and $H_s\simeq 1-(N/f_{tot})E[f^2]$. These two relations show that H_s increases when H decreases and vice versa. Next, recall that by Bayes' rule $p(k|f_i)=p(f_i|k)\cdot P(k)/p(f_i)$, where $p(k|f_i)$ is the probability to assign the label k to pixel $i(\theta_k=const)$; then the minimisation of $E[\log p(k|f_i)]=\sum_{i=1}^{N} p(k|f_i)\log p(k|f_i)$ corresponds to the maximisation of $E[\log(p(f_i))]$.

The main idea behind the diffused expectation maximisation (DEM) approach is that maximisation should be attained so that labels cannot be assigned to a pixel independently from others in its neighbourhood; then, a process must be devised that takes into account spatial correlations. It has been proved [2] that $\sum_{i=1}^{N} p(f_i)\log p(f_i)=-H_s$ is a Lyapunov functional decreasing under isotropic diffusion; however, this result *per se* does not allow to select the optimal label. Note that, for each label, neighbouring pixels should have the same probability to be assigned label k, and that labels at boundaries between regions should be characterised by an abrupt change of probability values. Denote for simplicity $h_{ik}=p(k|f_i)$; for each model k, h_{ik} defines a probability field on the image support D. Thus, each h_{ik} field should be a piecewise constant function across the image and indeed this result can be achieved [2] by a system of k anisotropic diffusions $(\partial h_{ik}(t)/\partial t)=\nabla\cdot(g(\nabla h_{ik})\nabla h_{ik}(t))$ each performing on the kth label probability plane, $g(\cdot)$ being a suitable conductance function, monotonically decreasing, and ∇ the gradient operator. Hence, small differences of h_{ik} among pixels close to each other are smoothed out, since diffusion is allowed, whereas large variations are preserved. As in the isotropic case, anisotropic diffusion is proved to increase the spatial entropy H_s [2].

The algorithm: We obtain the maximisation of $\log \mathcal{L}$ by iteratively computing $p(k|f, \theta)$, $p(f|k, \theta)$, $P(k)$ while diffusing on $p(k|f, \theta)$, which in practice regularises each k labelling field by propagating anisotropically such labels. Since, in terms of the mixture model, we are dealing with an incomplete data problem (i.e. we must simultaneously determine the labelling $p(k|f)$ given distribution parameters θk and vice versa), a suitable choice for parameter estimation is the EM algorithm interleaved with diffusion steps. Eventually, the segmentation is performed using the estimated parameters k, θ_k. The probabilistic model is assumed to be a mixture of Gaussians $p(f|k, \mu_k, \sigma_k)=(1/\sqrt{(2\pi)}\sigma_k)\exp(-((x-\mu_k)^2/2\sigma_k^2)$, thus $\theta_k=(\mu_k, \sigma_k)$, μ_k, σ_k being the unknown means and deviations, respectively, weighted by mixing proportions $\alpha_k=P(k)$. Note that we assume K fixed, in that we are not concerned here with the problem of model selection, in which case K may be selected by Bayesian information criterion (BIC). DEM works as follows.

Estimation: repeat for t iterations until $|\log\mathcal{L}^{(t+1)}-\log\mathcal{L}^{(t)}|<\epsilon$:

i) E-step: with fixed parameters $\alpha_k^{(t)}$, $\mu_k^{(t)}$, $\sigma_k^{(t)}$, compute the labelling probabilities at each site i as:

$$h_{ik}^{(t)}=\frac{\alpha_k^{(t)} p(f_i|k, \mu_k^{(t)}, \sigma_k^{(t)})}{\sum_k \alpha_k^{(t)} p(f_i|k, \mu_k^{(t)}, \sigma_k^{(t)})} \tag{2}$$

ii) D-step: propagate h_{ik} by m iterations of the discrete form of anisotropic diffusion

$$h_{ik}^{(t+1)}=h_{ik}^{(t)}+\lambda\nabla\cdot(g(\nabla h_{ik}^{(t)})\nabla h_{ik}^{(t)}) \tag{3}$$

and set $\tilde{h}_{ik}^{(t)}=h_{ik}^{(t+1)}$

iii) M-step: with $\tilde{h}_{ik}^{(t)}$ fixed, calculate the parameters that maximise $\log \mathcal{L}$:

$$\alpha_k^{(t+1)}=\frac{1}{N}\sum_i \tilde{h}_{ik}^{(t)}, \mu_k^{(t+1)}=\frac{\sum_i \tilde{h}_{ik}^{(t)} f_i}{\sum_i \tilde{h}_{ik}^{(t)}}, \sigma_k^{(t+1)}=\frac{\sum_i \tilde{h}_{ik}^{(t)}[f_i-\mu_k^{(t)}]^2}{\sum_i \tilde{h}_{ik}^{(t)}} \tag{4}$$

and calculate $\log \mathcal{L}^{(t+1)}$.

Segmentation: for each site $i \in \Omega$, obtain final labelling via estimated parameters by assigning to i, the label k for which $\max_k\{p(f_i|k, \mu_k, \sigma_k)\}$ hold.

Simulation: We have experimented with the method on different kinds of natural images. The test image used in this Letter is shown in Fig. 1*a*. To demonstrate the segmentation performance of the algorithm, both EM and DEM have been applied by assuming $K=4$ classes. Non-uniform initial estimates were chosen for $\alpha_k^{(0)}$, $\mu_k^{(0)}$, $\sigma_k^{(0)}$ parameters; $\{\mu_k^{(0)}\}$ were set in the range from minimal to maximal values of f_i in a constant increment; $\{\sigma_k^{(0)}\}$ were set in the range from 1 to $\max\{f_i\}$ in a constant increment; $\{\alpha_k^{(0)}\}$ were set from $\max\{f_i\}$ to 1 in a constant decrement and then normalised, $\sum_k \alpha_k^{(0)}=1$. For what concerns the D-step of the DEM algorithm, it is not limited to any specific selection of the conductance g, provided that label boundaries are preserved, numerical stability guaranteed, and probabilities $h_{ik}^{(t)}$ renormalised so that their sum is one after each

iteration [4]; in our experiments we set $g(\nabla h_{ik}) = |\nabla h_{ik}|^{-9/5}$, $\lambda = 0.1$; a number of $m = 10$ iterations of (3) was used. We found that convergence rate is similar for both methods, convergence being achieved after $t = 60$ iterations (with $\epsilon = 0.1$). More important, by comparing the results obtained by standard EM (Fig. 1*b*) and by DEM method (Fig. 1*c*), it is apparent the higher perceptual significance and the reliability of the latter as regards region classification. For visualisation purposes, a pixel i belonging to class k, is coloured as $f_i = \mu_k$.

Fig. 1 *Original image and segmentation results*

a Original image
b Segmented image obtained using EM method
c Segmented image obtained using DEM method

Conclusions: The DEM algorithm is different from related approaches previously proposed. Several methods have tried to incorporate within the EM algorithm a prior term in order to maximise a log posterior probability instead of log-likelihood, thus leading to quite complex EM steps [3]. Alternatively, Haker *et al.* [4] compute an initial posterior probability map, through some kind of classification (e.g. clustering), followed by anisotropic diffusion on such map in order to account for spatial constraints among sites; clearly, in this way final results strictly depend upon the goodness of the initial labelling. Here, we follow a different approach: we operate on the maximisation of the log-likelihood function, and spatial context is implicitly accounted for via label diffusion along maximisation. As a result we obtain a quite simple but effective segmentation algorithm, which can be easily interpreted in terms of a competition/co-operation scheme on the k label probability planes: the E and M steps implement a competition among the different labels at site i, while the D-step can be considered as a co-operation step among sites on the same plane. Its flexibility makes it suitable for any type of application.

10 June 2004
Electronics Letters online no: 20045792
doi: 10.1049/el:20045792

G. Boccignone (*Dipartimento di Ingegneria dell'Informazione e Ingegneria Elettrica, Universitá di Salerno, via Ponte Don Melillo 1, 84084 Fisciano (SA), Italy*)

E-mail: boccig@unisa.it

M. Ferraro (*Dipartimento di Fisica Sperimentale, Universitá di Torino, via Giuria 1, 10125 Torino, Italy*)

P. Napoletano (*Dipartimento di Ingegneria dell'Informazione e Ingegneria Elettrica and INFM, Universitá di Salerno, via Ponte Don Melillo 1, 84084 Fisciano (SA), Italy*)

References

1 Amari, S.I.: 'Information geometry of the EM and em algorithms for neural networks', *Neural Netw.*, 1995, **8**, pp. 1379–1408
2 Weickert, J.: 'Applications of nonlinear diffusion in image processing and computer vision', *Acta Math. Univ. Comenianae*, 2001, **70**, p. 3350
3 Sanjay-Gopal, S., and Hebert, T.J.: 'Bayesian pixel classification using spatially variant finite mixtures and the generalized EM algorithm', *IEEE Trans. Image Process*, 1998, **7**, pp. 1014–1028
4 Haker, S., Sapiro, G., and Tannenbaum, A.: 'Knowledge-based segmentation of SAR data with learned priors', *IEEE Trans. Image Process*, 2000, **9**, pp. 299–301