Appears in Proc. IEEE Conference in Computer Vision and Pattern Recognition, San Juan, Puerto Rico, 1997.

Empirical Bayesian EM-based Motion Segmentation[1]

Nuno Vasconcelos Andrew Lippman
MIT Media Laboratory
20 Ames St, E15-320M, Cambridge, MA 02139
{nuno,lip}@media.mit.edu

March 2, 2004

Abstract

A recent trend in motion-based segmentation has been to rely on statistical procedures derived from Expectation-Maximization (EM) principles. EM-based approaches have various attractives for segmentation, such as proceeding by taking non-greedy *soft decisions* with regards to the assignment of pixels to regions, or allowing the use of sophisticated priors capable of imposing *spatial coherence* on the segmentation. A practical difficulty with such priors is, however, the determination of appropriate values for their parameters. In this work, we exploit the fact that the EM framework is itself suited for *empirical Bayesian* data analysis to develop an algorithm that finds the estimates of the prior parameters which best explain the observed data. Such an approach maintains the Bayesian appeal of incorporating prior beliefs, but requires only a *qualitative* description of the prior, avoiding the requirement of a *quantitative* specification of its parameters. This eliminates the need for trial-and-error strategies for parameter determination and leads to better segmentations in less iterations.

1. Introduction

Image segmentation and motion (or optical flow) estimation have been widely studied in the fields of machine vision and image processing. Due to the difficulty of segmentation, early approaches to optical flow computation simply disregarded this component of the problem, relying on smoothness assumptions and regularization to overcome the ill-posed nature of optical flow estimation. This, however, resulted in poor motion estimates and imposed strong constraints on image analysis. It has been realized more recently that the problem can be solved only by procedures capable of jointly addressing the two components [3, 7]. This has led to a new generation of algorithms which iterate between optic flow estimation and segmentation.

The idea is, for a given set of motion parameters and observed flow, to find the maximum a posteriori (MAP) probability estimate of the segmentation; and, given this segmentation, to find the set of motion parameters which maximizes the likelihood of the measured flow. Because a hard-decision (regarding the membership of each pixel in the image to each of the segmentation classes) is performed for each iteration of these algorithms, they are sometimes referred to as clustering or *hard-decision* algorithms.

From a statistical perspective, such algorithms can be seen as variations of a stochastic optimization procedure known as the *Expectation-Maximization* (EM) algorithm [4]. Under the EM framework, segmentation masks (i.e. which region is responsible for each sample) are seen as hidden (non-observed) variables and the algorithm finds the values of the motion parameters that maximize the likelihood of the observed data by iterating between two steps. The E-step estimates the *expected* values of the hidden variables given the current values of the motion parameters and the observed data. The M-step then uses these expected values to find the set of parameters that maximize the likelihood of the data.

Because the region-assignment variables are binary, and expectations of binary values are equal to the probabilities of the variables being "on"; the estimates computed in the E-step are nothing more than the posterior probability of the region-assignments given the observed optical flow. I.e. EM is similar to the *hard-decision* algorithms above, but proceeds by taking *soft-decisions*, the MAP estimate of the segmentation being taken only upon the convergence of the iterative procedure.

Even though soft-decisions can lead to significantly better performance than hard-decisions [12], there are additional attractives in using EM for segmentation. In particular, because it provides an elegant statistical framework for the segmentation problem, EM allows the use of sophisticated priors, such as *Markov Random Fields* (MRFs) to enforce *spatial coherence* on the segmentation [10, 11]. However, such priors are typically characterized by parameters whose values are difficult to determine a priori. In practice, these parameters are commonly set to arbitrary values, or adapted to the observed data through heuristic procedures.

[1]See [9] for an extended version of this paper.

In this work, we exploit the fact that the EM framework is itself suited for *empirical Bayesian* data analysis [2] and a well known approximation to the likelihood of MRF processes to develop an algorithm that finds the estimates of the prior parameters which best explain the observed data. This eliminates the need for trial-and-error strategies for the determination of these parameters and leads to better segmentations in less EM iterations.

2. Bayesian and empirical Bayesian inference

In this section, we briefly review Bayesian and empirical Bayesian procedures [2, 8] for making inferences about the world, given observed image data. Assume that we are trying to make inferences about the world property Ω, given the image feature ω. Under the Bayesian framework, all inferences are based on the *posteriori* density function

$$P(\Omega|\omega) = \frac{P(\omega|\Omega)P(\Omega|\gamma_0)}{\int P(\omega|\Omega)P(\Omega|\gamma_0)d\Omega}, \quad (1)$$

where γ_0 is a parameter (or set of parameters) which controls the shape of the prior density for the world property. Under the Bayesian philosophy, properties in the world are not unknown deterministic quantities, but random variables characterized by probability densities that express a degree of prior belief in their possible configurations. The ratio between the posterior likelihoods of two configurations is proportional to the ratio of the respective prior likelihoods, the proportionality factor being dependent on the data. I.e. observation of the data merely re-scales prior beliefs [6].

It is therefore important, in Bayesian analysis, to get the prior beliefs right, a task which is generally difficult in practice. Typically, one does not have absolute certainty about the shape of the prior density and the parameters that characterize it which, unless known with certainty, must be regarded as random variables. That is, unless there is absolute certainty regarding the value of γ_0, inferences should be based on

$$P(\Omega|\omega) = \frac{\int P(\omega|\Omega)P(\Omega|\gamma_0)P(\gamma_0)d\gamma_0}{\int\int P(\omega|\Omega)P(\Omega|\gamma_0)P(\gamma_0)d\Omega d\gamma_0}, \quad (2)$$

instead of on equation (1).

While from a perceptual standpoint such a hierarchical structure has the appeal of modeling changes of prior belief according to context (different contexts lead to different values of γ_0, altering the shape of the density which characterizes prior beliefs), from a computational standpoint it significantly increases the complexity of the problem. After all, the parameters of $P(\gamma_0)$ are themselves random variables as well as the parameters of their density functions, and so on. We are therefore caught on a endless chain of conditional probabilities which is computationally intractable.

These issues are generally ignored in practice, where priors are typically chosen in order to minimize computational complexity, or set to arbitrary values. The alternative suggested by the empirical Bayesian philosophy is to replace γ_0 by an estimate $\hat{\gamma}_0$ obtained as the value which maximizes the marginal distribution $P(\omega|\gamma_0)$ as a function of γ_0. Inferences are then based on equation (1) using this estimated value.

While, strictly speaking, this approach violates the fundamental Bayesian principle that priors should not be estimated from data, in practice it leads to more sensible solutions than setting priors arbitrarily, or using priors whose main justification comes from computational simplicity (the so-called *conjugate* priors). More importantly, it provides a way to break the infinite chain of conditional probabilities mentioned above, while still allowing for different priors depending on context. Consider, for example, the task of, given pictures of a tree, to determine the probability of the world property "color" (C) from the image feature "pixel color" (c). The standard Bayesian solution would be to perform inferences based on equation (1) or, in this case,

$$P(C|c) \propto P(c|C)P(C|s),$$

where $P(c|C)$, which is determined by the camera optics and sensor noise, relates world and pixel colors, and $P(C|s)$ expresses prior beliefs in tree colors according to the parameters s. The main limitation of such model is that it fails to capture many factors that have an influence on tree colors, such as geography (leaf colors vary from region to region), seasonality (leaves are green in the Spring and yellow in the Fall), etc. Even though a simple prior may be appropriate to describe the colors of a given type of tree, at a given time of the year, in a given geographical location, no prior will be able to describe the colors of all trees, at all locations, for the entire year. Better models are obviously possible by taking the route of equation (2), i.e. by considering *hyperpriors* for all these factors, at the cost of enduring a significant increase in complexity.

The empirical Bayesian perspective is to avoid this increase by keeping the simple model $P(C|s)$, but choosing the parameters s that best explain the data. In this way, even though not directly, the model can account for the variations above, as the estimated s will be different for pictures taken in different seasons, locations, etc. Choosing the s which maximizes $P(c|s)$ will originate a prior which favors green colors for pictures taken in the Spring, and yellow colors for pictures taken in the Fall. In a sense, the empirical Bayesian approach allows the observer to concentrate on the specification the *qualitative* shape of the prior, letting the *quantitative* computation of prior parameters be inferred from the data.

Computationally, the bulk of work associated with empirical Bayesian procedures relies on the search for the prior parameters that maximize the marginal likelihood $P(\omega|\gamma_0)$. Because these parameters are related to the *observed* image features by the *hidden* world properties,

$$P(\omega|\gamma_0) = \int P(\omega|\Omega)P(\Omega|\gamma_0)d\Omega,$$

the problem fits naturally into an EM framework. Thus, in practice, empirical Bayesian estimates are commonly obtained through EM procedures, which iterate between the computation of the expected values for the world properties and the maximization over prior parameters. Therefore, the empirical Bayesian perspective not only supports the recent trend towards the application of EM for motion (and texture) segmentation, but extends it by providing a meaningful way to *tune* the priors to the observed data.

3. Doubly stochastic motion model

Our approach to image segmentation is based on linear parametric motion models, according to which the motion of the pixels associated with a given object is related to their image coordinates by

$$\mathbf{p}(\mathbf{x}) = \mathbf{\Psi}(\mathbf{x})\,\phi, \tag{3}$$

where $\mathbf{x} = (x, y)^T$ is the vector of pixel coordinates in the image plane, $\mathbf{p}(\mathbf{x}) = (p_x(\mathbf{x}), p_y(\mathbf{x}))^T$ the pixel's motion, and $\phi = (a_1, \ldots, a_P)^T$ the parameter vector which characterizes the motion of the entire object. In this work, we consider the particular case of affine motion where $P = 6$,

$$\mathbf{\Psi}(\mathbf{x}) = \begin{bmatrix} 1 & x & y & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x & y \end{bmatrix}, \tag{4}$$

and equation (3) models each of the components of the motion vector field as a plane in velocity space.

To account for uncertainties due to the imaging process, this motion model is embedded in a probabilistic framework, where pixels are associated with classes that have a one-to-one relationship with the objects in the scene. We assume that, conditional on the knowledge of image $\mathbf{I}_{t-1}(\mathbf{x})$ and the class of pixel $\mathbf{x}$ in image $\mathbf{I}_t(\mathbf{x})$, the observed value of this pixel is the outcome of an independent identically distributed Gaussian random process characterized by

$$P(\mathbf{I}_t(\mathbf{x})|\mathbf{z}(\mathbf{x}) = \mathbf{e}_i, \phi_i, \mathbf{I}_{t-1}(\mathbf{x})) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\{-\frac{1}{2\sigma_i^2}[\mathbf{I}_t(\mathbf{x}) - \mathbf{I}_{t-1}(\mathbf{x} - \mathbf{p}_i(\mathbf{x}))]^2\}, \tag{5}$$

where $\mathbf{p}_i(\mathbf{x})$ is the prediction of the motion of pixel $\mathbf{x}$ according to the class's model, σ_i^2 the variance of the pixels in the class, $\mathbf{z}(\mathbf{x})$ a vector of binary indicator variables, and $\mathbf{z}(\mathbf{x}) = \mathbf{e}_i$ (where $\mathbf{e}_i$ is the i^{th} vector of the standard unitary basis) if and only if pixel $\mathbf{x}$ belongs to object i.

Dependencies between the class-assignment probabilities of adjacent pixels are modeled by introducing a second-order MRF as a segmentation prior

$$\begin{aligned} P(\mathbf{z}(\mathbf{x}) = \mathbf{e}_i|\mathbf{z}) &= P(\mathbf{z}(\mathbf{x}) = \mathbf{e}_i|\mathbf{z}_\eta(\mathbf{x})) \\ &= \frac{1}{Z}\exp\left[\alpha_i + \beta u_i(\mathbf{x})\right], \end{aligned} \tag{6}$$

where $\mathbf{z}$ is the random field of indicator vectors $\mathbf{z}(\mathbf{x})$, $\mathbf{z}_\eta(\mathbf{x})$ the second-order neighborhood of pixel $\mathbf{x}$ (composed by the eight adjacent pixels), $u_i(\mathbf{x})$ the number of neighbors of pixel $\mathbf{x}$ that belong to class i, and Z a normalizing constant or partition function.

This leads to a doubly stochastic motion model. Doubly stochastic random fields using MRFs are the 2-D extension of Hidden Markov Models (HMMs), and have long been used for texture modeling and segmentation. In particular, the prior of equation (6) has been shown to be a good model for segmentation masks (see for example Figure 5 of [5]) and extensively used in the texture analysis literature. It is parameterized by the scalar β and the vector $\alpha = (\alpha_1, \alpha_2, \ldots)^T$. β controls the degree of clustering, i.e. the likelihood of more or less class transitions between neighboring pixels, while the α's control the relative likelihood of each of the segmentation classes.

4. EM-based parameter estimation

For a typical video sequence, the likelihood of the observed image data is a complicated function of the segmentation and motion parameters. This presents a significant challenge to EM-based algorithms since, given a poor initial estimate, EM will get trapped in undesirable local minima. In order to obtain a robust initial segmentation, we rely on a procedure which, starting from a collection of locally-computed motion models, iterates between 1) the merging of models which are likely to be associated with the same object, and 2) the elimination of bad models by cross validation[2].

Given this initial estimate for the segmentation map and the associated motion parameter estimates, the second stage of our algorithm uses the EM-based empirical Bayesian learning approach of section 2 and the doubly stochastic motion model of section 3 to 1) refine these initial estimates, 2) find the MRF prior parameters which best explain the observed motion, and 3) compute the MAP class assignment for each image pixel.

As mentioned in section 2, the fundamental computational problem posed by the empirical Bayesian framework is that of maximizing the marginal likelihood of the ob-

[2] See [9] for a detailed description of the parameter initialization procedure.

served data as a function of the motion and MRF parameters

$$P(\mathbf{I}_t|\mathbf{\Phi},\mathbf{I}_{t-1}) = \sum_{\mathbf{z}} P(\mathbf{I}_t|\mathbf{z},\mathbf{\Phi},\mathbf{I}_{t-1})P(\mathbf{z}|\mathbf{\Phi},\mathbf{I}_{t-1}),$$

where the summation is over all possible configurations of the hidden assignment variables vector $\mathbf{z}$, $\mathbf{\Phi}$ is the vector of all motion and MRF parameters, and $\mathbf{I}_t$ and $\mathbf{I}_{t-1}$ are the observed images. The pair $(\mathbf{I}_t, \mathbf{z})$ is usually referred to as the *complete data* and has log-likelihood

$$\begin{aligned} l_c &= \log P(\mathbf{z}|\mathbf{I}_{t-1},\mathbf{\Phi}) + \\ &+ \sum_{\mathbf{x},i} z_i(\mathbf{x}) \log\left[P(\mathbf{I}_t(\mathbf{x})|\mathbf{z}(\mathbf{x}) = \mathbf{e}_i, \mathbf{\Phi}, \mathbf{I}_{t-1})\right], \end{aligned}$$

where $z_i(\mathbf{x})$ is the i^{th} component of the vector $\mathbf{z}(\mathbf{x})$, and where we have used the class conditional probabilities of equation (5), the conditional independence of the observations given the indicator variables, and the binary nature of $z_i(\mathbf{x})$. The EM algorithm maximizes the likelihood of the incomplete, observed, data by iterating between two steps which act on the log-likelihood of the complete data.

4.1. The E-step

The E-step computes the so-called Q function defined by

$$\begin{aligned} Q(\mathbf{\Phi}'|\mathbf{\Phi}^{(p)}) &= E[l_c|\mathbf{I}_t,\mathbf{\Phi}^{(p)}] = E[\log P(\mathbf{z}|\mathbf{I}_t,\mathbf{\Phi}^{(p)})] + \\ &+ \sum_{\mathbf{x},i} E[z_i(\mathbf{x})|\mathbf{I}_t,\mathbf{\Phi}^{(p)}] \log\left[P(\mathbf{I}_t(\mathbf{x})|\mathbf{z}(\mathbf{x}) = \mathbf{e}_i)\right] \end{aligned} \quad (7)$$

where $\mathbf{\Phi}^{(p)}$ are the parameters obtained in the previous iteration and, for simplicity, we have dropped the dependence on $\mathbf{I}_{t-1}$. Under the MRF assumption for the prior class probabilities, the computation of $E[z_i(\mathbf{x})|\mathbf{I}_t,\mathbf{\Phi}^{(p)}]$ and $E[\log P(\mathbf{z}|\mathbf{I}_t,\mathbf{\Phi}^{(p)})]$ becomes analytically intractable, and can only be addressed through Monte Carlo procedures such as Gibbs sampling. Such procedures are, however, expensive from a computational perspective, and nesting a Gibbs sampler inside the EM iteration would lead to a prohibitive amount of computation. In order to simplify the problem, we rely on the well known approximation first proposed by Besag in his iterated coding mode (ICM) procedure for MAP estimation of MRF parameters [1], and later used by Zhang et al. in the context of EM-based segmentation [12]. This approximation consists of replacing the true likelihood by the pseudo-likelihood

$$P(\mathbf{z}) \approx \prod_{\mathbf{x}} P(\mathbf{z}(\mathbf{x})|\mathbf{z}_\eta(\mathbf{x})). \quad (8)$$

Assuming, further, that the configuration of the MRF does not change drastically from one iteration of the EM algorithm to the next, the pseudo-likelihood can be approximated by

$$\begin{aligned} P(\mathbf{z}) &\approx \prod_{\mathbf{x}} P(\mathbf{z}(\mathbf{x})|\mathbf{z}_\eta^{(p)}(\mathbf{x})) \\ &= \prod_{\mathbf{x},i} [P(\mathbf{z}(\mathbf{x}) = \mathbf{e}_i|\mathbf{z}_\eta^{(p)}(\mathbf{x}))]^{z_i(\mathbf{x})}. \end{aligned}$$

It is straightforward to show [12] that, under such approximation,

$$\begin{aligned} &P(z_i(\mathbf{x}) = 1|\mathbf{\Phi}^{(p)}) \approx \\ &\quad \approx \; P(z_i(\mathbf{x}) = 1|\mathbf{z}_\eta(\mathbf{x}),\mathbf{\Phi}^{(p)}) = \pi_i^{(p)}(\mathbf{x}), \end{aligned} \quad (9)$$

from which

$$\begin{aligned} & h_i(\mathbf{x}) = E[z_i(\mathbf{x})|\mathbf{I}_t,\mathbf{\Phi}^{(p)}] = P(z_i(\mathbf{x}) = 1|\mathbf{I}_t,\mathbf{\Phi}^{(p)}) \\ &= \frac{P(\mathbf{I}_t(\mathbf{x})|\mathbf{z}(\mathbf{x}) = \mathbf{e}_i,\mathbf{\Phi}^{(p)})P(z_i(\mathbf{x}) = 1|\mathbf{\Phi}^{(p)})}{\sum_k P(\mathbf{I}_t(\mathbf{x})|\mathbf{z}(\mathbf{x}) = \mathbf{e}_k,\mathbf{\Phi}^{(p)})P(z_k(\mathbf{x}) = 1|\mathbf{\Phi}^{(p)})} \\ &\approx \frac{P(\mathbf{I}_t(\mathbf{x})|\mathbf{z}(\mathbf{x}) = \mathbf{e}_i,\mathbf{\Phi}^{(p)})\,\pi_i^{(p)}(\mathbf{x})}{\sum_k P(\mathbf{I}_t(\mathbf{x})|\mathbf{z}(\mathbf{x}) = \mathbf{e}_k,\mathbf{\Phi}^{(p)})\,\pi_k^{(p)}(\mathbf{x})}, \end{aligned} \quad (10)$$

where we also used the binary nature of the indicator variables, and Bayes rule. Notice that the $h_i(\mathbf{x})$ are the *posterior* class assignment probabilities given the observed images. Given the current estimate of the prior probabilities $\pi^{(p)} = (\pi_1^{(p)}(\mathbf{x}), \pi_2^{(p)}(\mathbf{x}), \ldots)^T$, and the motion model parameters in $\mathbf{\Phi}^{(p)}$ they are computed by substituting equation (5) in equation (10).

One possible problem with this computation is that a pixel whose motion is poorly explained by all the models in $\mathbf{\Phi}^{(p)}$ will originate zero class-conditional likelihoods and the corresponding $h_i(\mathbf{x})$ will be undefined. To avoid this problem, we rely on the fact that a pixel which cannot be explained by any of the models is an outlier, and set the corresponding $h_i(\mathbf{x})$ to zero. Such a solution has the additional benefit of producing robust estimates without increasing the complexity of the M-step. Once outliers are eliminated, equation (8), and the computed h_i's are substituted in equation (7), and the Q function becomes

$$\begin{aligned} Q(\mathbf{\Phi}'|\mathbf{\Phi}^{(p)}) &= \sum_{\mathbf{x},i} h_i(\mathbf{x}) \log P(\mathbf{I}_t(\mathbf{x})|\mathbf{z}(\mathbf{x}) = \mathbf{e}_i) + \\ &+ \sum_{\mathbf{x},i} h_i(\mathbf{x}) \log P(\mathbf{z}(\mathbf{x})|\mathbf{z}_\eta(\mathbf{x}),\mathbf{\Phi}^{(p)}). \end{aligned} \quad (11)$$

4.2. The M-step

In the empirical Bayesian framework, the M-step maximizes the Q function obtained in the E-step with respect to both the motion and MRF parameters. Substituting equations (5) and (6) in equation (11), we obtain

$$\begin{aligned} Q(\mathbf{\Phi}'|\mathbf{\Phi}^{(p)}) &= -\frac{1}{2}\sum_{\mathbf{x},i} h_i(\mathbf{x}) \log(2\pi\sigma_i^2) \\ &- \frac{1}{2}\sum_{\mathbf{x},i} \frac{h_i(\mathbf{x})}{\sigma_i^2}[\mathbf{I}_t(\mathbf{x}) - \mathbf{I}_{t-1}(\mathbf{x} - \mathbf{p}_i(\mathbf{x}))]^2 \\ &+ \sum_{\mathbf{x},i} h_i(\mathbf{x})[\alpha_i + \beta u_i(\mathbf{x}) - \log Z]. \end{aligned}$$

Since the first two terms on the right hand side of this equation do not depend on α_i or β and the third term does not depend on ϕ_i or σ_i, the maximization can be separated into two sub-problems. The first - maximization of Q with respect to the parameters of the class conditional pdfs - is a variation of the non-linear least-squares problem found in optical flow estimation, and is solvable by non-linear optimization techniques. In our implementation, we use a simplified version of Newton's method leading to the iteration

$$\phi_i^{(k+1)} = \phi_i^{(k)} - \\ - \left[\sum_{\mathbf{x}} h_i(\mathbf{x})\mathbf{\Psi}(\mathbf{x})^T \nabla_x \mathbf{I}'_{t-1}(\mathbf{x}) \nabla_x \mathbf{I}'_{t-1}(\mathbf{x})^T \mathbf{\Psi}(\mathbf{x})\right]^{-1} \\ \times \sum_{\mathbf{x}} h_i(\mathbf{x})[\mathbf{I}_t(\mathbf{x}) - \mathbf{I}'_{t-1}(\mathbf{x})]\mathbf{\Psi}(\mathbf{x})^T \nabla_x \mathbf{I}'_{t-1}(\mathbf{x}),$$

$$(\sigma_i^{(k+1)})^2 = \frac{\sum h_i(\mathbf{x})[\mathbf{I}_t(\mathbf{x}) - \mathbf{I}_{t-1}(\mathbf{x} - \mathbf{\Psi}(\mathbf{x})\phi_i^{(k+1)})]^2}{\sum h_i(\mathbf{x})},$$

where $\nabla_x \mathbf{I}$ is the spatial gradient of $\mathbf{I}$, and $\mathbf{I}'_{t-1}(\mathbf{x}) = \mathbf{I}_{t-1}(\mathbf{x} - \mathbf{\Psi}(\mathbf{x})\phi_i^{(k)})$. The second sub-problem - maximization of Q with respect to MRF parameters - depends only on the third term, and can also be solved through standard non-linear programming methods. In our implementation we have used gradient ascent, under which the MRF parameters are updated in the direction of the gradients of the likelihood function with respect to them

$$\frac{\partial Q}{\partial \alpha_i} = \sum_{\mathbf{x}} [h_i(\mathbf{x}) - \pi_i(\mathbf{x})], \quad (12)$$

$$\frac{\partial Q}{\partial \beta} = \sum_{\mathbf{x}} [\sum_i h_i(\mathbf{x}) u_i(\mathbf{x}) - \sum_i \pi_i(\mathbf{x}) u_i(\mathbf{x})] \\ = \sum_{\mathbf{x}} [E[u(\mathbf{x})]_{post} - E[u(\mathbf{x})]_{prior}], \quad (13)$$

where $E[u(\mathbf{x})]$ is the expected number of neighbors of pixel $\mathbf{x}$ that belong to the same class. Once the new values of the MRF parameters are computed, the prior probabilities $\pi_i^{(p+1)}$ are obtained by applying a single cycle of Besag's ICM procedure: each pixel is visited in a raster scan order and, given the configuration of its neighborhood, the corresponding $\pi_i(\mathbf{x})$ are computed using equation (6). It can be shown that Q is a concave function of α_i and β, guaranteeing the existence of a single global maxima and allowing fast convergence to the optimal value.

It is interesting to analyze the meaning of the equations above. The new motion parameters are what one would obtain by performing a weighted non-linear least-squares fit to the motion field that best aligns the two images. The parameter update does not, however, rely on a greedy binary segmentation mask which is instead replaced by the posterior class assignment probabilities.

The gradient update equations also have a nice intuitive meaning. A step in the direction of equation (12) changes the MRF α parameter so that, at each pixel, the prior class-assignment probabilities move towards the posterior assignment probabilities obtained from the observed motion. Similarly, a step in the direction of equation (13) changes β so that, at each pixel, the expected number of neighbors in the same state as the pixel is equal under both the prior and the posterior distributions. I.e. the EM algorithm sets the model parameters to the values that best explain the observed data, both in terms of class assignment probabilities and average number of neighbors in the same state as the neighborhood's central pixel.

5. Experimental results and conclusions

In this section, we report on simulation results obtained with the "Flower Garden" sequence. Figure 1 presents a pair of frames from the sequence.

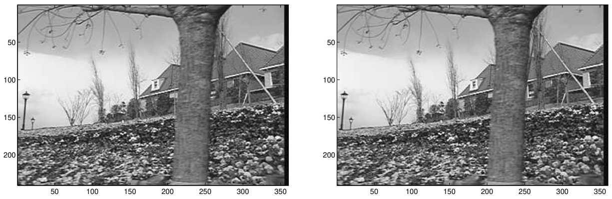

Figure 1: A frame-pair from the input video sequence.

Figure 2 illustrates the benefits of the empirical Bayesian solution to the motion segmentation problem that is now proposed. As can be seen from the figure, when the MRF parameters are set arbitrarily, the segmentation depends critically on the choice of the clustering parameter β. Small values of clustering lead to noisy segmentations such as the one on the top of the figure, while large values of β originate segmentations with weakly defined region boundaries (notice the leakage between the house and sky regions and between the areas of tree detail and sky in the middle picture).

While it may be possible to obtain better results by a trial-and-error strategy for the determination of MRF parameters, we were not able to obtain, in this way, a better segmentation than the originated by the empirical Bayesian approach, which is shown at the bottom of the figure. The better performance of empirical Bayesian estimates can be understood by considering Figure 3, which presents the evolution of the clustering parameter estimate as a function of the iteration number (for two different starting points). Once again, the result of empirical Bayesian parameter updating makes intuitive sense: while in early iterations (where uncertainty is high) clustering is small and pixels are free to wonder between regions, the clustering parameter increases as the EM procedure approaches convergence, and the segmentation "freezes" when this happens.

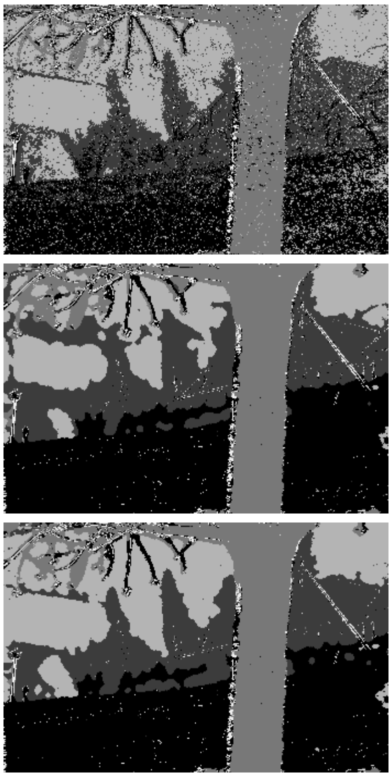

Figure 2: Three EM-based motion segmentations. For the top two, the MRF parameters were set to arbitrary values (top: $\beta = 0.2$, middle: $\beta = 0.7$). The bottom one was obtained with the empirical Bayesian parameter estimates discussed in the text. White pixels are outliers.

Even if such gradual evolution were not required for a good segmentation, it is not clear that the best trial-and-error estimate for a given sequence would be a good estimate for a different one. In fact, a review of the texture segmentation literature reveals a wide range of proposals for the value of β, which did not include the values that worked best for us. The point is that using empirical Bayesian estimates eliminates the need for tedious trial-and-error procedures that are not always guaranteed to provide the best results.

References

[1] J. Besag. On the Statistic Analysis of Dirty Pictures. *J. R. Statistical Society B*, 48(3):259–302, 1986.

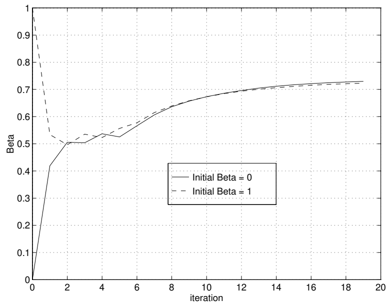

Figure 3: Evolution of the clustering parameter β as a function of iteration number. The two curves correspond to two different initial estimates of the parameter value. Notice that the evolution of β is very insensitive to the initial estimate.

[2] B. Carlin and T. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman Hall, 1996.

[3] T. Darrel and A. Pentland. Cooperative Robust Estimation Using Layers of Support. Technical Report 163, MIT Media Laboratory Perceptual Computing Group, June 1993.

[4] A. Dempster, N. Laird, and D. Rubin. Maximum-likelihood from Incomplete Data via the EM Algorithm. *J. of the Royal Statistical Society*, B-39, 1977.

[5] H. Derin and H. Elliott. Modeling and Segmentation of Noisy and Textured Images using Gibbs Random Fields. *IEEE Trans. on Pattern. Analysis and Machine Intelligence*, vol. PAMI-9, January 1987.

[6] A. Jepson, W. Richards, and D. Knill. Modal Structure and Reliable Inference. In D. Knill and W. Richars, editors, *Perception as Bayesian Inference*. Cambridge Univ. Press, 1996.

[7] D. Murray and B. Buxton. Scene Segmentation from Visual Motion Using Global Optimization. *IEEE Trans. on Pattern. Analysis and Machine Intelligence*, Vol. PAMI-9, March 1987.

[8] H. Robbins. An Empirical Bayes Approach to Statistics. In *Proc. Third Berkley Symposium Math. Statist.*, 1956.

[9] N. Vasconcelos and A. Lippman. Empirical Bayesian EM-based Motion Segmentation. Technical report, MIT Media Laboratory, 1997. Available from http://www.media.mit.edu/˜nuno.

[10] Y. Weiss and E. Adelson. Perceptually Organized EM: A Framework for Motion Segmentation that Combines Information about Form and Motion. Technical Report 315, MIT Media Lab Vision and Modeling Group, 1995.

[11] Y. Weiss and E. Adelson. A Unified Mixture Framework for Motion Segmentation: Incorporating Spatial Coherence and Estimating the Number of Models. In *Proc. Computer Vision and Pattern Recognition Conf.*, 1996.

[12] J. Zhang, J. Modestino, and D. Langan. Maximum-Likelihood Parameter Estimation for Unsupervised Stochastic Model-Based Image Segmentation. *IEEE Trans. on Image Processing*, Vol. 3, July 1994.