A Statistical Model for Contours in Images

François Destrempes and Max Mignotte

Abstract—In this paper, we describe a statistical model for the gradient vector field of the gray level in images validated by different experiments. Moreover, we present a global constrained Markov model for contours in images that uses this statistical model for the likelihood. Our model is amenable to an Iterative Conditional Estimation (ICE) procedure for the estimation of the parameters; our model also allows segmentation by means of the Simulated Annealing (SA) algorithm, the Iterated Conditional Modes (ICM) algorithm, or the Modes of Posterior Marginals (MPM) Monte Carlo (MC) algorithm. This yields an original unsupervised statistical method for edge-detection, with three variants. The estimation and the segmentation procedures have been tested on a total of 160 images. Those tests indicate that the model and its estimation are valid for applications that require an energy term based on the log-likelihood ratio. Besides edge-detection, our model can be used for semiautomatic extraction of contours, localization of shapes, non-photo-realistic rendering; more generally, it might be useful in various problems that require a statistical likelihood for contours.

Index Terms—Contours in images, edge-detection, parameter estimation, unsupervised statistical segmentation, Markov Random Field model.

1 Introduction

Detection of contours is an important problem in Image Processing. We make the distinction between methods of extraction of contours, such as the active snake [1], the live-wire [2] (based on the positions of the two endpoints of a curve), the jetstream [3] (based on the initial position and initial tangent vector of a curve), on one hand, and edge-detection,[1] on the other hand. In the first case, the algorithm finds only the edge points that are located on an optimal curve (in the sense of some *prior* geometric knowledge about the curves sought); in the latter case, one would like to detect all significant points of the image that are located "on edges." A further problem is localization of shapes [4], based on a *prior* knowledge about the shapes sought.

One standard edge-detection algorithm is the Canny edge-detector [5]. This algorithm is based on an optimal linear filter in the sense of three criteria (good detection, good localization, and low multiplicity of the response to a single edge), under a step edge model and for which the gradient of a Gaussian kernel is a good approximation. Using variants of Canny's criteria, one obtains a second-order recursive filter [6], or a first-order recursive filter [7] (also known as exponential or Shen filter). These filters assume a continuous signal. In [8], an optimal edge-detector is developed in the case of discrete signals. The reader can consult [8], [9] for numerous further references on edge-detectors.

The main inconvenience with the Canny edge-detector algorithm and others that have been proposed in the literature is the need for specifying thresholds or internal parameters of the algorithm in order to classify the pixels as "on edges" or "off edges" from the values of the filter response. In [9], a method is proposed for evaluating edge-detectors and selecting optimal parameter settings, based on the Receiver Operating Characteristic (ROC) curves. Roughly speaking, an ROC curve for a given edge-detector consists of the leftmost boundary of the region formed by the points $(\mathrm{TP}, \mathrm{FP})$ corresponding to the various parameter settings, with TP the proportion of true positives (with respect to edge-detection) and FP the proportion of false positives. In particular, this method requires a database of images with a ground truth segmentation.

In [10], various filters are considered for edge-detection and their conditional probability distributions "on edges" and "off edges" are represented nonparametrically. Those distributions are learned on a database of images equipped with a ground truth segmentation, using the Chernoff information and the ROC curves as quality measures. Edge-detection is then formulated as a statistical inference based on the log-likelihood ratio test. This follows [11], where it is pointed out that the log-likelihood ratio can be used to define appropriate energy terms in the active snake algorithm [1] or other methods of extraction of contours.

Relying on a database and ground truth segmentations presents an advantage in modeling various aspects of an image simultaneously (because one then obtains a very complex model). But, one might also be interested in performing *online* estimations. In that case, it seems interesting, in our opinion, to develop a parametric model for each aspect of an image and then perform the fusion of all those aspects at a higher level of decision-making. For instance, edge-detection is one aspect and texture segmentation is another aspect; a final image segmentation would then depend on those two aspects (as well as others); see, for instance, [12]. In the context of edge-detection, one would then be interested in an *online* estimation procedure of the conditional distributions of a filter response "on edges" and "off edges," according to a parametric model for contours. The point of view of using a parametric family of distributions is adopted in [11], in the context of extraction of roads in

1. An "edge" is defined as a boundary point of an object or a point presenting a significant contrast in gray level with some of its neighbors. This definition is incomplete as it stands since we eventually take into account the principle of low multiplicity of the filter-response to a single edge by using the nonmaxima suppression procedure.

• *The authors are with DIRO, Département d'Informatique et de Recherche Opérationnelle, C.P. 6128, Succ. Centre-Ville, Montréal, Canada (Quebec), H3C 3J7. E-mail: {destremp, mignotte}@iro.unmontreal.ca.*

Manuscript received 25 Oct. 2002; revised 03 July 2003; accepted 29 Nov. 2003. Recommended for acceptance by V. Solo.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 117664.

images. In [3], the same point of view is adopted in proposing a statistical model for the norm of the gradient of the gray-level "off edges" and for a related random variable "on edges," in the context of semiautomatic extraction of contours.

One might find it interesting to consider a Markovian framework in order to formulate various estimation and segmentation statistical criteria. For instance, the log-likelihood ratio test corresponds to the Maximum A Posteriori (MAP) criterion in the case of a uniform *prior* distribution. In [13], a Markov model with constraints for segmentation of an image into regions or region boundaries has been presented, but there are no statistical distributions, nor any estimation procedure.

We present, in this paper, a parametric statistical model (similar to [14], [15]) for the norm of the gradient of the gray level for points "on" and "off edges," as well as for the angle between the gradient and the normal to segments "on" and "off edges." We also present a new constrained Markov model for contours that takes into account the statistical distribution of the gradient vector field of the gray level in the image. Note that we do not describe directly the distribution of the angle of the gradient of the gray level in the image, but rather the distribution of the angle formed by the gradient and the normal to a segment. For that reason, we find convenient to consider a presegmentation set T consisting of various paths that contains potentially all edge-points, as well as other points that will eventually be classified as "off edges." In this paper, we define T by the standard nonmaxima suppression procedure (upon using the principle of low multiplicity of the response to a single edge), but see [10] for other suggestions. This constrained model is used for an Iterative Conditional Estimation (ICE) procedure [16], [17], [18], [19], [20], [21], [22] for the estimation of the parameters. Note that the ICE procedure was previously used only for segmentation of an image into regions. Comparison between histograms and estimated distributions suggests that our model is reasonable for the tested images.

As an application of our model, we view edge-detection in an image as a constrained minimization problem. We can find an optimal solution in the sense of the Maximum A Posteriori (MAP) using the Simulated Annealing (SA) algorithm [23]. The Iterated Conditional Modes (ICM) algorithm [24] can also be used to find a satisfactory suboptimal solution. One can use a Monte Carlo (MC) algorithm to find an approximated optimal solution in the sense of the Modes of Posterior Marginals (MPM) [25]. This yields a new unsupervised statistical method for edge-detection in images, with three variants. Our model can also be used for localization of shapes [14], [26] or semiautomatic extraction of contours [14], [15]; in those two applications, we do not use a binary edge-detection but only the estimation of the distributions. Note, however, that, in [14], [15], [26], we have used slightly different versions of the model presented here (for instance, the estimation procedure has been improved in this paper), but we obtain equivalent results (if not better) with the present version.

This paper brings two contributions: 1) Our statistical model is parametric and, hence, its parameters may be estimated *online* and 2) our model yields a probabilistic method for edge-detection, not just a binary detection. Thus, our model might be useful for various problems that require a statistical likelihood for contours, in the same spirit of [10], but with a parametric model and an *online* estimation procedure.

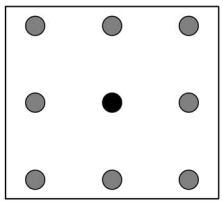


Fig. 1. A pixel s and its eight neighbors (8 pixels) in the graph G.

We have tested our method on the University of South Florida (USF) image database, consisting of 10 aerial images and 50 objects images and on the 100 natural images of the University of California at Berkeley (UCB) image test data set [27]. We think that all of them are optical images obtained by electronic acquisition, though we do not have that information at hand. Other types of images (such as radar, MRI, or PET images) present different kinds of noise and texture, due to other acquisition modes. So, we are inclined to think that, for other types of images, our method would require other edge-detection filters and, presumably, other statistical distributions. Even with optical images, one needs different filters to take into account various textures, together with different statistical distributions. See, for instance, [28] for a parametric model for texture segmentation and an *online* estimation procedure.

The remaining part of this paper is organized as follows: The basic notations and definitions are presented in Section 2. In Section 3, we present the statistical model for the gradient vector field of the gray level. In Section 4, we explain the constrained Markov model for contours. Section 5 presents, in detail, the estimation procedure for the model parameters. In Section 6, we explain three variants of an unsupervised method for edge-detection. In Section 7, we discuss experimental results. Finally, we conclude in Section 8.

2 Basic Notations and Definitions

2.1 Graphs Considered

Given an image of size $M \times N$, $G = (V_G, E_G)$ will denote the nonoriented graph consisting of the MN pixels of the image together with the segments[2] given by the usual 8-neighbors. If s and t are adjacent sites of G, we denote the segment joining s and t by (s,t) (so, $(s,t) \in E_G$).

$V_{G'}$ will denote the set of all sites of G together with the set of its segments ($V_{G'} = V_G \cup E_G$). We form a nonoriented graph $G' = (V_{G'}, E_{G'})$ by making two adjacent sites in G also adjacent in G' and by making (s,t) adjacent to its end points s and t. Thus, $E_{G'} = E_G \cup \{(s,u),(t,u) \ : \ u = (s,t) \in E_G\}$. See Figs. 1 and 2 for an illustration of the graphs G and G'.

2.2 Random Fields Considered

We compute the gradient of the gray-level function I of an image by the second-order approximation $I_x = \frac{1}{2}(I(x+1,y) - I(x-1,y))$ and $I_y = \frac{1}{2}(I(x,y+1) - I(x,y-1))$. A first order approximation seems to be more sensible to noise. For each site $s \in V_G$, the observable random variable Y_s represents the norm of the gradient of the gray level at the corresponding

2. The graph theoretical terminology of "edge" would be unfortunate in the context of edge-detection.

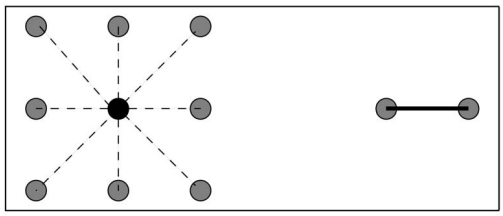


Fig. 2. Left: A pixel s and its 16 neighbors (8 pixels and 8 segments) in the graph G'. Right: A segment (s,t) and its two neighbors (2 pixels) in the graph G'.

Fig. 3. Presegmentation set obtained by nonmaxima suppression for the image of Fig. 6.

pixel. Also, the hidden random variable X_s takes its values in the label set $\{e_1 =$"off edges", $e_2 =$ "on edges"$\}$.

If $(s,t) \in E_G$, we consider the angle (in absolute value) Y_{st} between the mean of the gradient at s and t with the normal to the segment (s,t). The angle is normalized between $-\pi/2$ and $\pi/2$ (before taking its absolute value). Whenever the mean of the gradient at s and t vanishes, we define artificially the angle to be $*$, an abstract value not in the interval $[0, \frac{\pi}{2}]$. We consider the hidden random variable X_{st} taking its values in the label set $\{e_1, e_2\}$. A statistical model for the gradient vector field is presented in Section 3, i.e., a model for the conditional distributions $P(y_s \mid x_s)$ and $P(y_{st} \mid x_{st})$.

For the ICE estimation procedure presented in Section 5, as well as for our unsupervised method for edge-detection, we need a constrained Markov model for contours. Thus, we consider two random fields on the graph G': the observable continuous random field $Y = \{Y_s, Y_{st} : s \in V_G, (s,t) \in E_G\}$ and the hidden discrete random field $X = \{X_s, X_{st} : s \in V_G, (s,t) \in E_G\}$. The joint distribution of the couple of random fields (X,Y) is modeled by a constrained Markov model in Section 4. The likelihood $P(y \mid x)$ is based on the local conditional distributions of Section 3, whereas the prior $P(x)$ is based on a presegmentation set T, that we now describe.

2.3 Presegmentation Set

We consider a presegmentation set $T \subseteq V_G$ in which are confined all points that might be classified as on edges. See Fig. 3 for an illustration of the presegmentation set in the case of the image of Fig. 6.

We define T in an ad hoc manner using the nonmaxima suppression technique. Namely, given an image, consider, at each pixel s, the approximation of the gradient of the gray level by the nearest direction $\hat{d}$ among the directions $d_1, \ldots, d_4$ corresponding to the angles $-\pi/4, 0, \pi/4, \pi/2$. Define T^+ as the set of all pixels for which the norm of the gradient is no less than the norm of the gradient at its two neighbors along $\hat{d}$. The approximate value of the gradient is used only at this step.

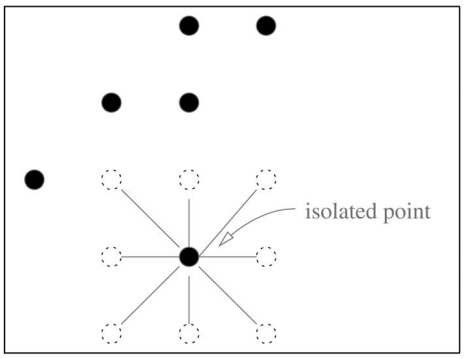


Fig. 4. Illustration of an isolated point of the set T^+.

Next, remove the isolated points of T^+. A point of T^+ is an isolated point if none of its 8-neighbors is in the set T^+ (see Fig. 4 for an illustration). Finally, we remove from T^+ the pixels at which the gradient of the gray level vanishes, thus obtaining a presegmentation set T. We then define T^* to be the set of all segments (s,t) with both end points in T.

Note that the presegmentation set T depends on the observed data of the image (i.e., the gradient of the gray level). So, strictly speaking, it might be preferable to model T as the observed realization of a discrete random field. Yet, for simplicity, we view T as a "metaparameter" which is estimated once and for all by nonmaxima suppression. The role of the presegmentation set T is crucial in what follows and our choice might not be optimal. Nevertheless, other presegmentation sets are presented in [10] and it is reported that they appear to be close to the one obtained by nonmaxima suppression.

2.4 Summary of the Main Technical Notations

For further reference, we collect in Table 1 the main symbols that will be used in the next sections.

3 Statistical Model for the Gradient Vector Field

We present in this section a model for the conditional distributions of the gradient vector field of the gray level on and off edges. These distributions are used to define the likelihood of the Markov model presented in the next section.

3.1 Norm of the Gradient

The distribution $P(y_s \mid x_s = e_1)$ (corresponding to the sites off edges) is modeled in [3] by an exponential distribution. Here, we model $P(y_s \mid x_s = e_1)$ by a shifted Weibull distribution [29],

$$\mathcal{W}(y_s; \min, C, \alpha) = \frac{C}{\alpha}\left(\frac{y_s - \min}{\alpha}\right)^{C-1} \exp\left(-\left(\frac{y_s - \min}{\alpha}\right)^{C}\right)$$

defined for $y_s > \min$. Note that, if $C = 1$, one recovers the exponential distribution. In our tests, we have observed that the estimated value of C varies approximately between 0.4 and 1.4, which justifies our preference for a Weibull distribution. Since the norm is nonnegative, one has to take

TABLE 1
Summary of the Main Technical Notations

$\mathcal{W}(y; \min, C, \alpha)$	shifted Weibull distribution with shape parameter C
k_1	normalizing constant $\{\int_0^\infty \mathcal{W}(y; \min, C, \alpha)\, dy\}^{-1}$
$\mathcal{N}(y; \mu, \sigma)$	normal distribution with mean μ and variance σ
$\mathcal{M}(y; w_j, \mu_j, \sigma_j)$	mixture of normal distribution with proportions $w_1, ..., w_K$
k_2	normalizing constant $\{\int_0^\infty \mathcal{M}(y; w_j, \mu_j, \sigma_j)\, dy\}^{-1}$
q_0	proportion of segments in the image for which the mean of the gradient at the two end points is $\neq 0$
$\mathcal{U}(y; 0, \frac{\pi}{2})$	uniform distribution on the interval $[0, \frac{\pi}{2}]$
$\mathcal{E}(y; \alpha_0)$	truncated exponential distribution on the interval $[0, \frac{\pi}{2}]$
k_0	normalizing constant $\{\int_0^{\frac{\pi}{2}} \frac{1}{\alpha_0} \exp(-\frac{y}{\alpha_0})\, dy\}^{-1}$
Φ	vector of parameters $(q_0, C, \alpha, w_j, \mu_j, \sigma_j, \alpha_0)$ for the likelihood
θ	regularizing parameter for the prior model
$\delta(x_s, x_t)$	equal to 1, if $x_s = x_t$; equal to 0, otherwise
$\eta(x_s, x_t, x_{st})$	equal to 1, if the equivalence $x_s = x_t = e_2 \Leftrightarrow x_{st} = e_2$ holds; equal to 0, otherwise
T	pre-segmentation set obtained by non-maxima suppression
$\chi_T(x)$	constraint function $\prod_{s \notin T} \delta(x_s, e_1) \prod_{(s,t) \in E_G} \eta(x_s, x_t, x_{st})$

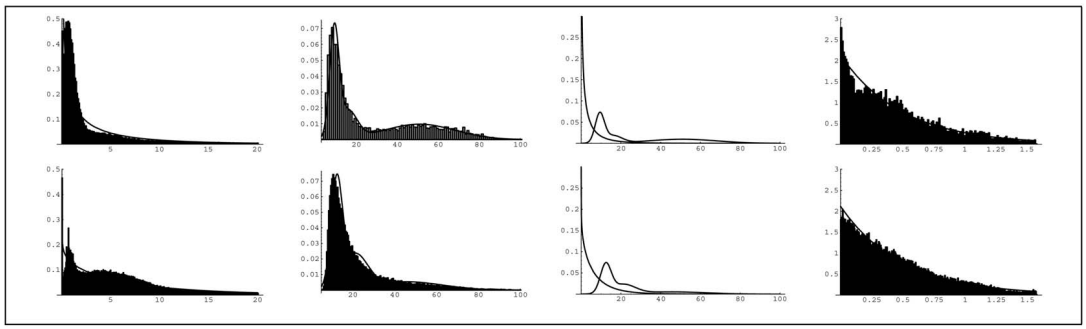


Fig. 5. Examples of distributions for the gradient of the gray level. From left to right: Norm of the gradient for points off edges; norm of the gradient for points on edges; comparison between the two distributions; angle between the gradient and the normal to segments on edges. The rows correspond, from top to bottom, to Figs. 6 and 7, respectively.

$\min < 0$. In our tests, we took, systematically, $\min = -10^{-3}$ after the normalization described in Section 7. We admit that this is quite ad hoc and that the value of $\min$ should be estimated. Nevertheless, the tests reported in Section 7 indicate that our choice is reasonable.

We model the distribution $P(y_s \mid x_s = e_2)$ (corresponding to pixels on edges) by a mixture of Gaussian kernels,

$$\mathcal{M}(y_s; w_j, \mu_j, \sigma_j) = \sum_{j=1}^{K} w_j \mathcal{N}(y_s; \mu_j, \sigma_j),$$

where $\sum_{j=1}^{K} w_j = 1$ and $w_j \geq 0$, for $j = 1, \ldots, K$. Roughly speaking, each Gaussian kernel represents a class of edge-points according to the relative degree of contrast in gray level. As is well-known, a mixture of Gaussian kernels is a good approximation to any given continuous distribution, provided the number of kernels is sufficiently large. Nevertheless, if the sample set is too small, a large number of kernels will cause overfitting. In our tests, a mixture of three Gaussian kernels seems flexible enough (as far as we can tell) to model the wide range of variations in the distribution $P(y_s \mid x_s = e_2)$ from one image to an other. But, in the case of very complex images, one might want to consider greater values of K if the size of the image is sufficiently large. For that matter, one could use the Bayesian Information Criterion (BIC) [30].

Now, since the norm is nonnegative, the distribution $\mathcal{W}(y_s; \min, C, \alpha)$ should be restricted to $y_s \geq 0$ and adjusted by the factor

$$k_1 = \left\{ \int_0^\infty \mathcal{W}(y; \min, C, \alpha)\, dy \right\}^{-1} = \exp\left(\left(\frac{-\min}{\alpha} \right)^C \right).$$

However, if $(\frac{-\min}{\alpha})^C \approx 0$, this point can be ignored for all practical purposes. Similarly, the distribution $\mathcal{M}(y_s; w_j, \mu_j, \sigma_j)$ should be adjusted by the factor $k_2 = \{\int_0^\infty \mathcal{M}(y; w_j, \mu_j, \sigma_j)\, dy\}^{-1}$. Again, this factor can be ignored if the values of μ_j are positive and sufficiently large with respect to σ_j (for instance, $\mu_j > 3\sigma_j$). The experimental results reported in Section 7 indicate that the effect of these simplifications is negligible in the context of our applications.

Our hypothesis is that the norm of the gradient of the gray level tends to be larger for points on edges than for points off edges. This is captured in the distributions adopted here whenever the values of $\mu_1, \ldots, \mu_K$ are sufficiently large with respect to α. We take this hypothesis into account in the initialization step of the estimation procedure explained in Section 5. See Fig. 5.

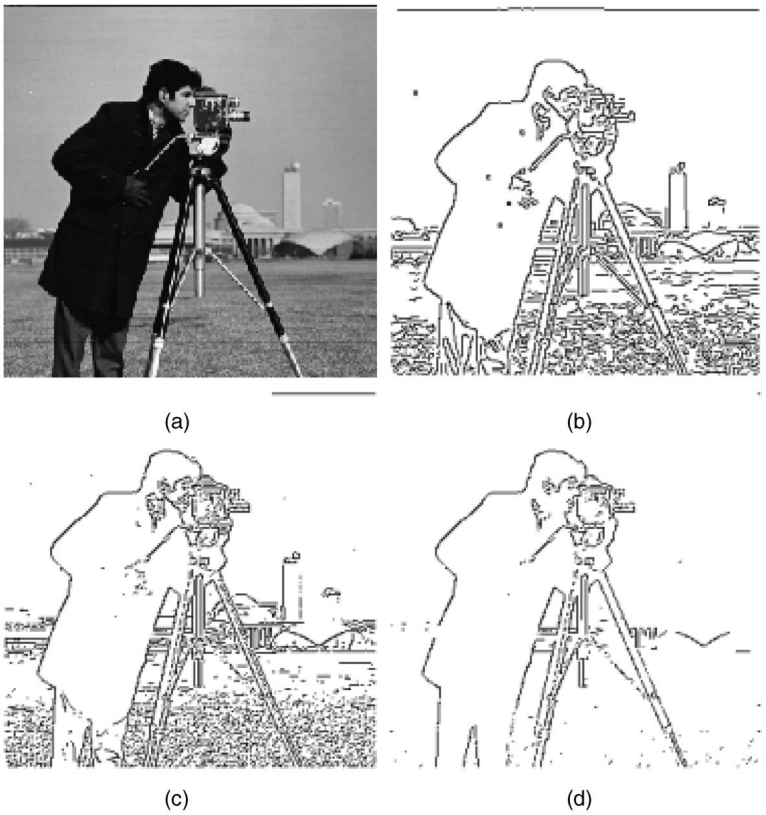


Fig. 6. Top left: Original image. Top right: Edge-detection using the SA algorithm. Bottom left: Segmentation using the Canny edge-detector (lower threshold: 11.5; upper threshold: 23). Bottom right: Segmentation using the Canny edge-detector (lower threshold: 38; upper threshold: 76).

3.2 Angle between the Gradient and the Normal to Segments

The random variable Y_{st} can be viewed as the angle (in absolute value) between a level curve of the gray-level function and the curve going through the segment (s,t). If the segment is off edges, there is no privileged value expected for Y_{st}. On the other hand, if the segment is located on edges, we expect the value of Y_{st} to be near 0, i.e., our hypothesis is that a curve located on edges tends to coincide locally with a level curve of the gray-level function. This hypothesis is implicit in the hysteresis technique of the Canny's edge-detection algorithm.

Corresponding to the case where (s,t) is off edges, we consider the uniform distribution on $[0,\frac{\pi}{2}]$,

$$\mathcal{U}\Big(y_{st};0,\frac{\pi}{2}\Big)=\frac{2}{\pi},\qquad 0\le y_{st}\le \pi/2.$$

Corresponding to the case where the segment (s,t) is located on edges, we consider the truncated exponential law

$$\mathcal{E}(y_{st};\alpha_0)=\frac{k_0}{\alpha_0}\exp\Big(-\frac{y_{st}}{\alpha_0}\Big),\qquad 0\le y_{st}\le \pi/2.$$

The factor k_0 is taken so as to obtain a distribution on the interval $[0,\frac{\pi}{2}]$, i.e.,

$$k_0=\left\{\int_0^{\frac{\pi}{2}}\frac{1}{\alpha_0}\exp\Big(-\frac{y}{\alpha_0}\Big)\,dy\right\}^{-1}=\left\{1-\exp\Big(-\frac{\pi}{2\alpha_0}\Big)\right\}^{-1}.$$

In order to take care of the case where the angle is not defined (i.e., $y_{st}=*$), we consider the Dirac distribution δ_* centered at the value $*$. Altogether, we model the distribution $P(y_{st}\mid x_{st}=e_1)$ by

$$q_0\mathcal{U}\Big(y_{st};0,\frac{\pi}{2}\Big)+(1-q_0)\delta_*(y_{st})$$

and the distribution $P(y_{st}\mid x_{st}=e_2)$ by

$$q_0\mathcal{E}(y_{st};\alpha_0)+(1-q_0)\delta_*(y_{st}),$$

where q_0 is the proportion of segments in the image for which the mean of the gradient at the two end points does not vanish.

Our tests reported in Section 7 suggest that our model might be useful in some applications. See Fig. 5 for examples of the empirical distributions and the estimated distributions.

4 Constrained Markov Model for Contours

We now describe the joint distribution of the couple of random fields (X,Y). For the likelihood, we use the local conditional distributions of Section 3

$$\begin{aligned}
P(y_s\mid x_s=e_1) &\approx \mathcal{W}(y_s;\min,C,\alpha);\\
P(y_s\mid x_s=e_2) &\approx \mathcal{M}(y_s;w_j,\mu_j,\sigma_j);\\
P(y_{st}\mid x_{st}=e_1) &= q_0\mathcal{U}\big(y_{st};0,\tfrac{\pi}{2}\big)+(1-q_0)\delta_*(y_{st});\\
P(y_{st}\mid x_{st}=e_2) &= q_0\mathcal{E}(y_{st};\alpha_0)+(1-q_0)\delta_*(y_{st}).
\end{aligned}$$

We assume, as usual, that all the variables Y_s and Y_{st} are mutually independent conditional to $X=x$ and that

furthermore, $P(y_s|x) = P(y_s|x_s)$ and $P(y_{st}|x) = P(y_{st}|x_{st})$. Thus, the likelihood $P(y|x)$ is given by

$$\prod_{s\in V_G} P(y_s \mid x_s) \prod_{(s,t)\in E_G} P(y_{st} \mid x_{st}).$$

As usual, the dependence structure among pixels is captured by a Markov *prior*. For the *prior* distribution, we consider the Gibbs energy $\theta \sum_{(s,t)\in T^*}\big(1-\delta(x_s,x_t)\big)$ in order to favor a homogeneous segmentation within the set T, where $\delta(.)$ is the Kronecker delta function and $\theta > 0$ is a parameter. We also impose the following two constraints on a realization x of the labeling field X: 1) $x_s = e_1$ for all $s \notin T$ (only pixels in the presegmentation set can be located on edges) and 2) for any $(s,t) \in E_G$, $x_s = e_2$ and $x_t = e_2$ if and only if $x_{st} = e_2$ (a segment is on edges if and only if its two endpoints are on edges). We define a function χ_T on the set of all realizations of X by setting $\chi_T(x) = 1$ whenever x satisfies the two constraints above and 0 otherwise. In detailed form, $\chi_T(x)$ can be expressed as

$$\prod_{s\notin T} \delta(x_s, e_1) \prod_{(s,t)\in E_G} \eta(x_s, x_t, x_{st}),$$

where $\eta(x_s,x_t,x_{st}) = 1$ if the equivalence $x_s = x_t = e_2 \iff x_{st} = e_2$ holds and $\eta(x_s,x_t,x_{st}) = 0$ otherwise. Given this model with constraints, the prior distribution $P(x)$ is equal to

$$k \exp\left\{-\theta \sum_{(s,t)\in T^*} \big(1-\delta(x_s,x_t)\big)\right\} \chi_T(x),$$

where k is a normalizing constant.

Let U be the Gibbs energy[3] defined by

$$U(x,y) = \sum_{s\in V_G} -\ln P(y_s \mid x_s) + \sum_{(s,t)\in E_G} -\ln P(y_{st} \mid x_{st}) + \theta \sum_{(s,t)\in T^*} \Big(1-\delta(x_s,x_t)\Big).$$

The a posteriori distribution can then be expressed as

$$P(x \mid y) \;\propto\; \exp\Big(-U(x,y)\Big)\, \chi_T(x),$$

where the omitted factor depends only on the observed realization y. Due to the constraint expressed by χ_T, the model is called constrained Markov model, rather than Markov model.

5 Estimation Procedure

5.1 The ICE Procedure

Let (X,Y) be the couple of random fields presented in Section 4. The observable random field Y is called the "incomplete data" and $Z = (X,Y)$ the "complete data." As has been presented in Section 4, the likelihood of the model for contours depends on a vector of parameters $\Phi = (q_0, C, \alpha, w_j, \mu_j, \sigma_j, \alpha_0)$, whereas the *prior* distribution depends on a parameter θ. Note that the factor k_0 can be deduced directly from the value of α_0. Various methods have been developed for the estimation of the parameters of a *prior* distribution in the complete data case (see [24], [31], [32]), as well as in the incomplete data case (see, for instance, [33]). Nevertheless, for simplicity, we fix the parameter θ equal to 1 throughout this paper and concentrate solely on the estimation of the likelihood parameters.

In order to obtain an estimation of the likelihood parameter vector Φ from the incomplete data, we resort to the ICE algorithm. This procedure, described in detail in [16] for the segmentation of an image into regions, relies on an estimator $\hat{\Phi}(x,y)$ of Φ for the complete data with good asymptotic properties (such as the Maximum Likelihood (ML) estimator). One starts with an initial estimation $\Phi^{[0]}$ (based on the observed data y) and then considers the sequence defined recursively by $\Phi^{[p+1]} = \int_x \hat{\Phi}(x,y)\, P(x \mid y, \Phi^{[p]})\, dx$. The properties of this sequence are not fully understood as of now, but various experiments [16], [17], [18], [19], [20], [21], [22], [33] indicate its relevance.

The computation of the expectation $\int_x \hat{\Phi}(x,y) P(x \mid y, \Phi^{[p]})\, dx$ is impossible in practice, but we can approximate it by,

$$\frac{1}{n}\,\big[\hat{\Phi}(x_{(1)},y) + \cdots + \hat{\Phi}(x_{(n)},y)\big],$$

where $x_{(i)},\ i = 1,\ldots,n$ are realizations of X drawn according to the posterior distribution $P(x \mid y, \Phi^{[p]})$. One can take $n = 1$ without altering the quality of the estimation in the context of [16]. In our case, taking $n = 5$ seems satisfactory as the tests reported in Section 7 indicate. In the context of region segmentation, an image offers a larger sample of points than in the context of edge-detection because there are fewer edge-points. Hence, one needs in principle more simulations.

Altogether, the ICE procedure can be outlined as follows:

1. **Initialization**: $\Phi^{[0]}$ is obtained from the parameters estimated on each class of an initial segmentation obtained by the K-means algorithm described in [34].
2. Set $p = 0$ and repeat until a stopping criterion is met:
   a. **Simulation**: Using the Gibbs sampler, n realizations $x_{(1)},\ldots,x_{(n)}$ are simulated according to the posterior distribution $P(x \mid y, \Phi^{[p]})$, with parameter vector $\Phi^{[p]}$.
   b. **Estimation**: The parameter vector $\Phi^{[p+1]}$ is approximated by the sum

$$\frac{1}{n}\,\big[\hat{\Phi}(x_{(1)},y) + \cdots + \hat{\Phi}(x_{(n)},y)\big].$$

   Set $p \leftarrow p+1$.

In our tests, the stopping criterion is that $||\Phi^{[p+1]} - \Phi^{[p]}||_\infty \le \varepsilon ||\Phi^{[p]}||_\infty$, where $\varepsilon = 10^{-2}$, with a maximum of 100 iterations. We now explain the estimation procedure in details in the context of our constrained Markov model for contours.

3. Due to the presence of the Dirac distribution δ_* in the definition of $P(y_{st} \mid x_{st})$, the Gibbs energy U can be equal to $-\infty$; however, this fact does not affect any of the algorithms explained in the next two sections.

5.2 Initialization

We initialize the ICE procedure by the parameters estimated on the complete data obtained by a K-means clustering segmentation of T into two classes based on the norm of the gradient of the gray level. The class with smallest mean is labeled $e_1 =$ "off edges."

5.3 Simulation

We simulate a realization of X according to the posterior distribution $P(x \mid y, \Phi^{[p]})$, with parameter vector $\Phi^{[p]}$. Our model for contours has constraints. The constrained stochastic relaxation has been developed in this context [35] and could be used in the simulation of X. However, for computational reasons, we impose directly the constraints by setting $x_s = e_1$ whenever $s \notin T$ and $x_{st} = e_2$ if and only if $x_s = e_2$ and $x_t = e_2$. Thus, we simply use the Gibbs sampler algorithm [23] on $T \cup T^*$. In doing so, we visit, sequentially, the sites of $T \cup T^*$ as follows: If s is a site of G, we let $E(s) = \{t : t \in T, (s,t) \in E_G\}$ (the neighbors of the pixel s which are in the set T). There are two possible labels for s and each one determines the label of (s,t), for $t \in E(s)$, upon using the constraint $\eta(x_s, x_t, x_{st}) = 1$. For $i = 1, 2$, we consider the real number $f(e_i)$ equal to

$$P(y_s \mid x_s = e_i) \prod_{t \in E(s)} P(y_{st} \mid x_{st}) \exp(-\theta(1 - \delta(e_i, x_t)))$$

with each x_{st} adjusted so that $\eta(x_s, x_t, x_{st}) = 1$. We then draw x_s according to the probabilities

$$f(e_i)/(f(e_1) + f(e_2)) = 1/\left(\frac{f(e_1)}{f(e_i)} + \frac{f(e_2)}{f(e_i)}\right),$$

for $i = 1, 2$. Note that, in the case where $y_{st} = *$, one can simply ignore the factor $P(y_{st} \mid x_{st})$ because the ratio $\delta_*(y_{st} = *)/\delta_*(y_{st} = *)$ contributes to 1 in the calculation of $f(e_1)/f(e_2)$ or $f(e_2)/f(e_1)$.

5.4 Estimation

Given a realization x of X, the complete data $Z = (X, Y)$ is known. Henceforth, the parameters of the statistical distribution associated to each class can be computed using the appropriate ML estimators. We use the following approximated estimation procedures:

- We compute the ML estimators of the Weibull distribution as follows: Let $Y_1, \ldots, Y_m$ be m random variables i.i.d. according to a shifted Weibull law $\mathcal{W}(y; \min, C, \alpha)$ and let $y_1, \ldots, y_m$ be a realization of those variables. We assume that the value of $\min$ is fixed and we set $\tilde{y}_i = y_i - \min$. The log-likelihood function is equal to

$$m \ln C - mC \ln \alpha + (C-1) \sum_{i=1}^{m} \ln \tilde{y}_i - \frac{1}{\alpha^C} \sum_{i=1}^{m} \tilde{y}_i^C.$$

  Let $\hat{C}$ and $\hat{\alpha}$ be the ML estimators of C and α. Then, $(\hat{C}, \hat{\alpha})$ is a critical point of the log-likelihood function. In [36], it is proven that such a critical point necessarily satisfies the identities

$$\frac{1}{\hat{C}} = F(\hat{C}), \qquad \hat{\alpha} = \left(\frac{1}{m} \sum_{i=1}^{m} \tilde{y}_i^{\hat{C}}\right)^{\frac{1}{\hat{C}}},$$

  where

$$F(C) = \frac{\sum_{i=1}^{m} \tilde{y}_i^C \ln \tilde{y}_i}{\sum_{i=1}^{m} \tilde{y}_i^C} - \frac{1}{m} \sum_{i=1}^{m} \ln \tilde{y}_i.$$

  In Appendix A, which can be found at http://www.computer.org/tpami/archives.htm, we show that the function $\frac{1}{C} - F(C)$ is decreasing in the interval $(0, \infty)$ and admits a unique root in that interval (except in the degenerate case where all y_i are equal). It follows that the log-likelihood has a unique critical point. Moreover, we can use a dichotomy search algorithm based on the sign of $\frac{1}{C} - F(C)$ in order to efficiently solve the equation $\frac{1}{C} - F(C) = 0$. We stop the procedure when the relative distance between two successive values is less than 5×10^{-3}. In [29], it is suggested to use Newton-Raphson's method; however, there is no guarantee of convergence. In contrast, the dichotomy search always converges, provided there are two values of the function with opposite sign, as is the case here. This allows us to estimate (C, α) on the set $\{y_s : x_s = e_1\}$.
- The ML estimators of the parameters of a mixture of Gaussian distributions cannot be computed directly. We use the SEM algorithm [37], which is a stochastic version of the EM algorithm [38]. Let $Y_1, \ldots, Y_m$ be m random variables i.i.d. according to a mixture of K Gaussian kernels $\mathcal{M}(y; w_j, \mu_j, \sigma_j)$ and let $y_1, y_2, \ldots, y_m$ be a realization of those variables. Consider m hidden random variables $X_1, \ldots, X_m$ taking their values in the set of auxiliary labels $\{f_1, \ldots, f_K\}$. Set $\Psi = (w_j, \mu_j, \sigma_j)$. The SEM algorithm can be outlined as follows:

  1. **Initialization**: Use the K-means algorithm to obtain an initial segmentation and compute the usual ML estimators $(\mu_j^{[0]}, \sigma_j^{[0]})$ of a Gaussian distribution on each class. Next, compute the proportion $w_j^{[0]}$ of each class, thus obtaining $\Psi^{[0]} = (w_j^{[0]}, \mu_j^{[0]}, \sigma_j^{[0]})$ for the initial estimation of the parameters.
  2. Set $p = 0$ and repeat until a stopping criterion is met:
     a. **Simulation**: For $i = 1, \ldots, m$, the label x_i is drawn according to the posterior distribution

$$P(f_j \mid y_i, \Psi^{[p]}) = \frac{w_j^{[p]} \mathcal{N}(y_i; \mu_j^{[p]}, \sigma_j^{[p]})}{\sum_{k=1}^{K} w_k^{[p]} \mathcal{N}(y_i; \mu_k^{[p]}, \sigma_k^{[p]})}.$$

     b. **Estimation**: Use the ML estimators on the resulting segmentation to obtain the parameter vector $\Psi^{[p+1]}$. Set $p \leftarrow p + 1$.

  The initialization step is used only at the first ICE iteration; subsequently, we use the parameters estimated at the previous ICE iteration for the initial value $\Psi^{[0]}$. Since an average over a few simulations is performed within the ICE procedure, we consider only one simulation within the SEM algorithm. The stopping criterion

is as for the ICE procedure, but with a minimum of 10 iterations and a maximum of 500 iterations (in our tests, the algorithm usually stops within 30 iterations). In this manner, we obtain an estimation of (w_j, μ_j, σ_j) from the set $\{y_s : x_s = e_2\}$.

- For the truncated Exponential law, we compute the ML estimator as follows: Let $Y_1, \ldots, Y_m$ be m random variables i.i.d. according to a truncated Exponential law $\{\alpha_0(1 - \exp(-\frac{\pi}{2\alpha_0})\}^{-1} \exp(-\frac{y}{\alpha_0})$. If $y_1, y_2, \ldots, y_m$ is a realization of those variables, the log-likelihood function is equal to

$$-m\left(\ln\{\alpha_0\} + \ln\left\{1 - \exp\left(-\frac{\pi}{2\alpha_0}\right)\right\}\right) - \sum_{i=1}^{m} \frac{y_i}{\alpha_0}.$$

Now, the ML estimator $\hat{\alpha}_0$ is a critical point of the log-likelihood function and, hence, satisfies the equation

$$-\frac{1}{\hat{\alpha}_0} + \frac{\pi \exp(-\frac{\pi}{2\hat{\alpha}_0})}{2\hat{\alpha}_0^2(1 - \exp(-\frac{\pi}{2\hat{\alpha}_0}))} + \frac{\frac{1}{m}\sum_{i=1}^{m} y_i}{\hat{\alpha}_0^2} = 0.$$

Thus, $\hat{\alpha}_0$ is a root of the function $\alpha_0 - H(\alpha_0)$, where

$$H(\alpha_0) = \frac{1}{m}\sum_{i=1}^{m} y_i + \frac{\pi}{2(\exp(\frac{\pi}{2\alpha_0}) - 1)}.$$

In Appendix B, which can be found at http://www.computer.org/tpami/archives.htm, it is shown that the function $\alpha_0 - H(\alpha_0)$ is increasing in the interval $(0, \infty)$ and admits a unique root in that interval, provided that $\frac{1}{m}\sum_{i=1}^{m} y_i < \frac{\pi}{4}$. Thus, in that case, a simple dichotomy search allows us to find $\hat{\alpha}_0$ efficiently. The stopping criterion is as for the Weibull distribution. In the case where $\frac{1}{m}\sum_{i=1}^{m} y_i \geq \frac{\pi}{4}$, the log-likelihood function is maximal at $\alpha_0 = \infty$ and the truncated Exponential law is then the uniform distribution. In our tests, the latter case never occurred. We can thus estimate α_0 on the set $\{y_{st} : x_{st} = e_2\}$.

6 Application to Unsupervised Edge-Detection

We view edge-detection in an image as finding a realization x of X that maximizes the a posteriori distribution $P(x \mid y) \propto P(x, y)$. In doing so, we minimize the average cost function $\int_X C_{\text{MAP}}(X, x) P(X \mid y)\, dX$, where $C_{\text{MAP}}(X, x)$ is the Maximum A Posteriori (MAP) cost function defined by $1 - \delta(X, x)$. Equivalently, we want a realization x that minimizes the Gibbs energy $U(x, y)$ subject to the constraint $\chi_T(x) = 1$, with U and χ_T as in Section 4.

An optimal solution can be found using the Simulated Annealing (SA) algorithm [35], though at the expense of a significant computational cost. The SA depends on a parameter τ called *temperature* which, in principle, should be equal to $\tau_p = \frac{\tau_0}{\ln(p+1)}$ at iteration $p \geq 0$, where τ_0 is a sufficiently large initial temperature. Nevertheless, we take $\tau_p = \tau_0 a^p$, with $0 < a < 1$. This suboptimal version of the SA can be outlined as follows in our context:

1. **Random initialization**: For each site s of V_G, set $x_s = e_1$ whenever $s \notin T$; otherwise, set $x_s = e_1$ or e_2 randomly. Then, for each segment $(s, t) \in E_G$, adjust the label x_{st} so that $\eta(x_s, x_t, x_{st}) = 1$.
2. Set $\tau = \tau_0$ and repeat until τ is numerically small (i.e., $\tau = \tau_f$):
   a. **Stochastic optimization**: Visit sequentially the sites of $T \cup T^*$ as follows: If s is a site of G, let $E(s) = \{t : t \in T, (s, t) \in E_G\}$ be as in Section 5.3. For $i = 1, 2$, let $V(e_i)$ be equal to

$$-\ln\{P(y_s \mid x_s = e_i)\} + \sum_{t \in E(s)} \Big(-\ln\{P(y_{st} \mid x_{st})\} + \theta(1 - \delta(e_i, x_t))\Big)$$

with each x_{st} adjusted so that $\eta(x_s, x_t, x_{st}) = 1$. Draw x_s according to the probabilities[4] $\exp(-\frac{V(e_i)}{\tau})/(\exp(-\frac{V(e_1)}{\tau}) + \exp(-\frac{V(e_2)}{\tau}))$, i.e.,

$$\frac{1}{\exp\left(-\frac{V(e_1) - V(e_i)}{\tau}\right) + \exp\left(-\frac{V(e_2) - V(e_i)}{\tau}\right)},$$

for $i = 1, 2$. After one sweep, set $\tau \leftarrow a\tau$.

In our tests, we take $\tau_0 = 3$, $\tau_f = 0.1$, and $a = 0.995$.

One can also obtain a good suboptimal solution by means of the Iterated Conditional Modes (ICM) algorithm [24]. The ICM in our context can be outlined as follows:

1. **ML initialization**: For each site s of V_G, set $x_s = e_1$ whenever $s \notin T$; otherwise, set $x_s = \arg\min_{i=1,2} -\ln\{P(y_s \mid x_s = e_i)\}$. Then, for each segment $(s, t) \in E_G$, adjust the label x_{st} so that $\eta(x_s, x_t, x_{st}) = 1$.
2. Repeat until no more sites are modified after a complete sweep:
   a. **Greedy optimization**: Visit sequentially the sites of $T \cup T^*$ and take $x_s = \arg\min_{i=1,2} V(e_i)$, where $V(e_i)$ is as for the SA.

In the Markovian framework, one can also consider the Modes of Posterior Marginals (MPM) [25] cost function defined by $C_{\text{MPM}}(X, x) = \sum_{s \in V_{G'}} (1 - \delta(X_s, x_s))$. The computation of an optimal solution x (i.e., a realization that minimizes the average cost function $\int_X C_{\text{MPM}}(X, x)\, P(X \mid y)\, dX$) is intractable, but one can find an approximate value by means of the following Monte Carlo (MC) algorithm:

1. **Random initialization**: As for the SA.
2. **Simulation**: Using the Gibbs sampler, n realizations $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are simulated sequentially according to the posterior distribution $P(x \mid y)$ (see Section 5 for the details of the Gibbs sampler in our context).
3. **Segmentation**: For each site s of V_G, set $x_s = e_1$ whenever $s \notin T$; otherwise, set x_s to be the label appearing the most frequently in the n simulations $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ (after dropping the first few simulations). Then, adjust the labels x_{st} as above.

In our tests, we perform 150 simulations and ignore the first 20 ones. Our tests reported in Section 7 indicate that the MC algorithm yields a solution close to the SA algorithm, even if the MPM and the MAP are different criteria.[5]

4. Again, in the case where $y_{st} = *$, one can simply ignore the term $-\ln P(y_{st} \mid x_{st})$ because the difference $-\ln \delta_*(y_{st} = *) + \ln \delta_*(y_{st} = *)$ contributes to 0 in the calculation of $V(e_1) - V(e_2)$ or $V(e_2) - V(e_1)$.

5. This would not necessarily hold if the *prior* model were changed, for instance, if $\theta \neq 1$.

TABLE 2
Time of Execution of the Algorithms on a Workstation 2.0GHz for the Images of Figs. 6 and 7

Images	Fig.7	Fig.8
size	256×256	400×600
ICE	58.562 sec. (26 iterations)	2 min. 44 sec. (20 iterations)
Viterbi	78 msec.	287 msec.
ICM	259 msec.	1.267 sec.
MC	12.317 sec.	46.662 sec.
SA	55.732 sec.	3 min. 31 sec.

7 Experimental Results

7.1 General Remarks

For the estimation and the segmentation procedures, we first recalibrate the gray level of the image between 0 and 255 and then apply a 3×3 Gaussian mask. Afterward, we multiply the gradient by $100/\nu$, where ν is the maximum of the norm of the gradient in the image. We present in Table 2 the time of execution of the ICE procedure, the Viterbi algorithm, the ICM, the MPM classifier, and the SA on a Workstation 2.0GHz, for the images of Figs. 6 and 7. See Fig. 5 for the distributions corresponding to the images of Figs. 6 and 7.

In our other main tests, we have used the University of South Florida (USF) database, consisting of 10 aerial images and 50 indoor images, as well as the University of California at Berkeley (UCB) test data set of 100 natural images. In Table 3, we show the variation of some of the estimated parameters on these databases. One can appreciate the point of an *online* estimation procedure: One single distribution (learned from a database) might not be fully adapted to a given image, even if that image were taken from a similar database. One can also observe that the factors k_1 and k_2 are virtually equal to 1.

7.2 Evaluation of the Segmentation Methods

We have evaluated four variants of our edge-detection method:

1. the Viterbi algorithm [39] with the MAP criterion on a *simplified version of the model that ignores junction points* (c.f. [15] and [14]),
2. the ICM algorithm with the MAP criterion (c.f. Section 6),
3. the SA algorithm with the MAP criterion (c.f. Section 6),
4. the MC algorithm with the MPM criterion (c.f. Section 6).

We have compared the entropy $-\ln P(x \mid y)$ of the solution obtained by each of the methods 1, 2, 4 with the one obtained by the SA. As expected, the SA yields the lowest entropy. Although the MAP and the MPM are different criteria, the MC yields the entropy closest to the SA. Finally, the ICM and the Viterbi algorithm are comparable; note that the Viterbi algorithm gives the optimal solution to a simplified model, whereas the ICM gives a suboptimal solution to the full model. We have also compared the classification error produced by methods 1, 2, 4 with respect to the solution obtained by the SA. Again, the MC is closest to the SA, whereas the Viterbi algorithm and the ICM are comparable. We have omitted the description of the simplified version of the model in this paper since one might as well use the ICM algorithm on the full model, for a quick algorithm, or else the SA algorithm or the MPM classifier.

The results of the edge-detection using the SA algorithm, as well as the energy map $-\ln \frac{P(y_s \mid e_2)}{P(y_s \mid e_1)}$ for the norm of the gradient, for the databases mentioned above, can be found at http://www.iro.umontreal.ca/~destremp/PAMI02/.

We present in Figs. 6 and 7 the segmentations obtained by our edge-detection method (with the SA variant) and compare it with Canny's edge-detection algorithm. It seems that, in our case, most significant edges can be detected without introducing more noise. Note also that, in the case of Canny's algorithm, one has to specify two thresholds. In contrast, our method is unsupervised, except for the size of the Gaussian mask and the values of $\min$ and θ (though those values have been fixed once and for all in all our tests) since the other parameters are estimated, rather than adjusted manually. But, most importantly, our method allows the use of the statistical distributions "on" and "off edges," rather than the binary segmentation itself. In fact, in our applications [14], we only use the estimation procedure and postpone decision-making to higher level procedures (such as localization of shapes).

7.3 Technical Evaluation of the Estimation Procedure

To evaluate the validity of our model, we have compared, for each image of the databases, the distributions estimated by the ICE procedure with the empirical distributions, using the segmentation obtained by the SA algorithm. Namely, we calculate the Kullback-Leibler distance between the normalized histogram and the estimated distribution. In our tests, we have implemented the following version of the Kullback-Leibler distance:

$$d(H, P) = \sum_{i=1}^{N} \ln(H(t_i)/P(t_i))H(t_i),$$

where H is the histogram and P is the estimated distribution, both of which are defined on a bounded interval I split into $N = 100$ subintervals of equal length, and $t_1, \ldots, t_N$ are the midpoints of those subintervals. Here, H and P are normalized so that $\sum_{i=1}^{N} H(t_i) = 1$ and $\sum_{i=1}^{N} P(t_i) = 1$. Working with the norm for now, this gives two measures that evaluate the absolute errors made in the estimation procedure:

$$\begin{aligned} d(H(Y_s \mid e_1), P(Y_s \mid e_1)) &= \sum_{i=1}^{N} \ln \frac{H(t_i \mid e_1)}{P(t_i \mid e_1)} H(t_i \mid e_1) \\ &\approx \frac{1}{N_1} \sum_{s:\, x_s = e_1} \ln \frac{H(y_s \mid e_1)}{P(y_s \mid e_1)} \\ d(H(Y_s \mid e_2), P(Y_s \mid e_2)) &= \sum_{i=1}^{N} \ln \frac{H(t_i \mid e_2)}{P(t_i \mid e_2)} H(t_i \mid e_2) \\ &\approx \frac{1}{N_2} \sum_{s:\, x_s = e_2} \ln \frac{H(y_s \mid e_2)}{P(y_s \mid e_2)}, \end{aligned}$$

where $N_1 = |\{s :\ x_s = e_1\}|$, $N_2 = |\{s :\ x_s = e_2\}|$. We also consider the two complementary measures defined by

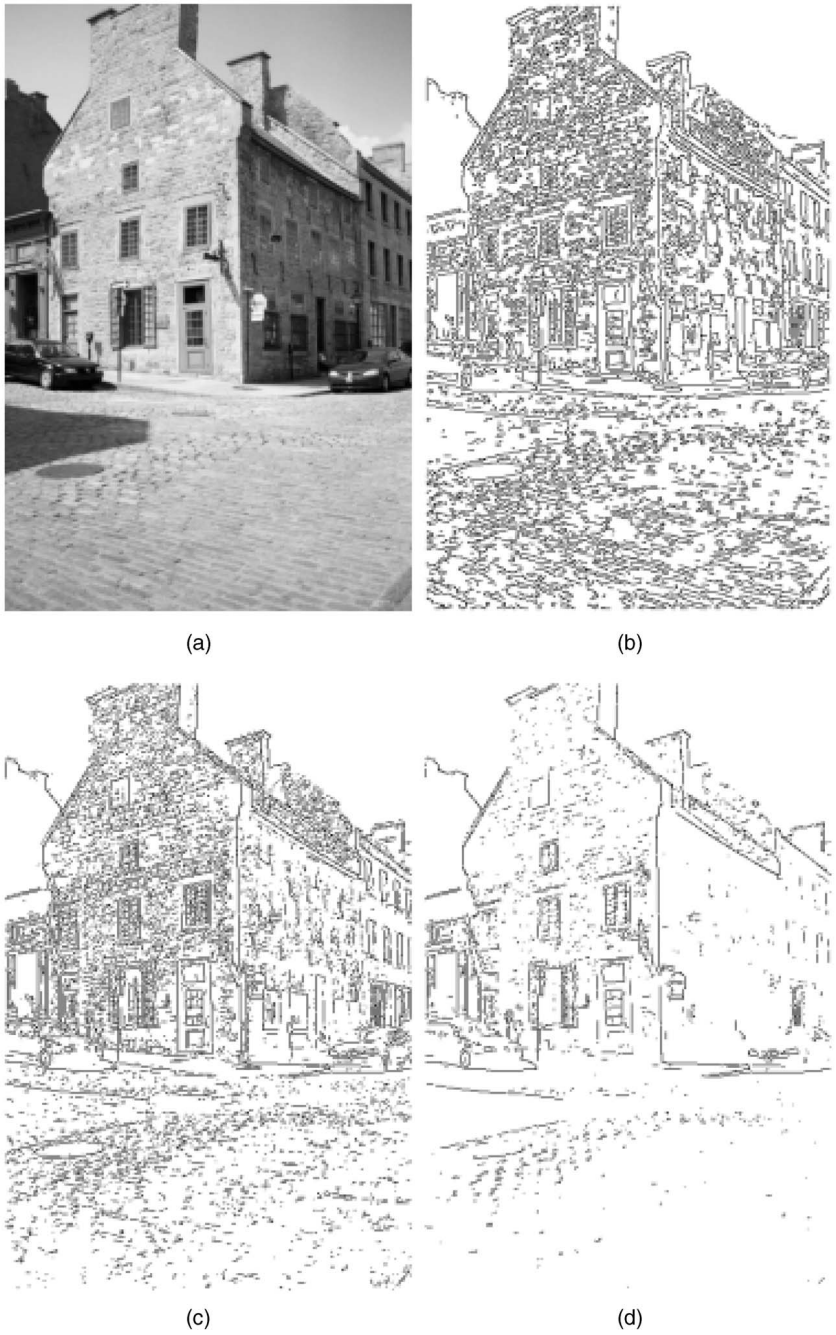


Fig. 7. Top left: Original image. Top right: Edge-detection using the SA algorithm. Bottom left: Segmentation using the Canny edge-detector (lower threshold: 18.5; upper threshold: 37). Bottom right: Segmentation using the Canny edge-detector (lower threshold: 40; upper threshold: 80).

$$\bar{d}(H(Y_s \mid e_1), P(Y_s \mid e_1)) = \sum_{i=1}^{N} -\ln \frac{H(t_i \mid e_1)}{P(t_i \mid e_1)} H(t_i \mid e_2) \approx \frac{1}{N_2} \sum_{s:\, x_s = e_2} -\ln \frac{H(y_s \mid e_1)}{P(y_s \mid e_1)}$$

$$\bar{d}(H(Y_s \mid e_2), P(Y_s \mid e_2)) = \sum_{i=1}^{N} -\ln \frac{H(t_i \mid e_2)}{P(t_i \mid e_2)} H(t_i \mid e_1) \approx \frac{1}{N_1} \sum_{s:\, x_s = e_1} -\ln \frac{H(y_s \mid e_2)}{P(y_s \mid e_2)}.$$

Next, we compute, for each image, the average value of the energy term $-\ln \frac{P(\cdot \mid e_2)}{P(\cdot \mid e_1)}$ over the two classes "off edges" and "on edges," respectively. This gives a measure of classification efficiency of the estimated distributions:

$$\sigma_{\text{norm}} = \left\{ \frac{1}{N_1} \sum_{s:\, x_s = e_1} -\ln \frac{P(y_s \mid e_2)}{P(y_s \mid e_1)} \right\} - \left\{ \frac{1}{N_2} \sum_{s:\, x_s = e_2} -\ln \frac{P(y_s \mid e_2)}{P(y_s \mid e_1)} \right\},$$

and the analogue measure for the histograms:

$$\tilde{\sigma}_{\text{norm}} = \left\{ \frac{1}{N_1} \sum_{s:\, x_s = e_1} -\ln \frac{H(y_s \mid e_2)}{H(y_s \mid e_1)} \right\} - \left\{ \frac{1}{N_2} \sum_{s:\, x_s = e_2} -\ln \frac{H(y_s \mid e_2)}{H(y_s \mid e_1)} \right\}.$$

TABLE 3
Average Value and Standard Deviation of Some of the Estimated Parameters on Three Databases

parameters	USF	USF	UCB
	aerial images	objects images	natural images
C	0.96 ± 0.06 SD	0.79 ± 0.14 SD	0.91 ± 0.22 SD
α	7.034 ± 1.77 SD	3.85 ± 1.32 SD	6.42 ± 2.95 SD
α_0	0.459 ± 0.098 SD	0.497 ± 0.097 SD	0.547 ± 0.061 SD
μ_1	20.59 ± 4.34 SD	11.85 ± 2.96 SD	16.43 ± 5.28 SD
μ_2	37.32 ± 6.33 SD	23.45 ± 5.67 SD	27.9 ± 7.8 SD
μ_3	56.15 ± 10.08 SD	49.71 ± 11.23 SD	50.77 ± 8.8 SD
k_1	1.00024 ± 0.00012 SD	1.00308 ± 0.0042 SD	1.0027 ± 0.0073 SD
k_2	1.00012 ± 0.00021 SD	1.00127 ± 0.0012 SD	1.00074 ± 0.003 SD

TABLE 4
Evaluation Measures for the Estimation of the Various Distributions on Three Databases

measures	USF	USF	UCB
	aerial images	objects images	natural images
$d(H(Y_s \mid e_1), P(Y_s \mid e_1))$	0.075	0.056	0.038
$d(H(Y_s \mid e_2), P(Y_s \mid e_2))$	0.022	0.053	0.049
$d(H(Y_{st} \mid e_1), P(Y_{st} \mid e_1))$	0.0057	0.0077	0.0047
$d(H(Y_{st} \mid e_2), P(Y_{st} \mid e_2))$	0.095	0.024	0.015
Δ_{norm}	0.024	0.086	0.068
Δ_{angle}	0.28	0.15	0.12
Δ_{total}	0.041	0.091	0.071

See the explanations in the text.

Note that the difference between the two efficiency measures is given by

$$\begin{aligned}\delta_{\text{norm}} &= |\tilde{\sigma}_{\text{norm}} - \sigma_{\text{norm}}| \\ &\approx |d(H(Y_s \mid e_1), P(Y_s \mid e_1)) + \bar{d}(H(Y_s \mid e_2), P(Y_s \mid e_2)) \\ &\quad + d(H(Y_s \mid e_2), P(Y_s \mid e_2)) + \bar{d}(H(Y_s \mid e_1), P(Y_s \mid e_1))|,\end{aligned}$$

as a little calculation shows.

Finally, we are interested in the relative measures:

$$\Delta_{\text{norm}} = \frac{\delta_{\text{norm}}}{\sigma_{\text{norm}}}, \quad \Delta_{\text{angle}} = \frac{\delta_{\text{angle}}}{\sigma_{\text{angle}}}, \quad \Delta_{\text{total}} = \frac{\delta_{\text{norm}} + \delta_{\text{angle}}}{\sigma_{\text{norm}} + \sigma_{\text{angle}}},$$

where δ_{angle} and σ_{angle} are defined similarly for the angle distributions.

When $\sigma > 0$, one can use the energy term $-\ln \frac{P(\cdot \mid e_2)}{P(\cdot \mid e_1)}$ in order to distinguish points "on edges" from points "off edges," a lower value being in favor of the former case (c.f. [11], [3], [14], [15], [26]). Furthermore, the measure Δ evaluates the relative error between the previous classification efficiency measure (based on the estimated distributions) and the one obtained using the histograms. In Table 4, the values of the various measures are reported for the USF and UCB databases. We obtained that σ_{total} is always positive and that Δ_{total} is on average less than 0.09. Thus, the parametric model presented here is accurate with less than 9 percent error on average, with respect to the energy term $-\ln \frac{P(y_s \mid e_2)}{P(y_s \mid e_1)} - \ln \frac{P(y_{st} \mid e_2)}{P(y_{st} \mid e_1)}$.

8 Conclusion

In this paper, we have described a new statistical model for the gradient vector field of the gray level in images validated by various experiments. As pointed out in [8], [9], [10], the gradient of a Gaussian kernel is not the optimal filter for edge-detection; nevertheless, we chose it for its simplicity and its usefulness in applications such as localization of shapes. That being said, it would be interesting to find a statistical parametric model for the optimal edge-detector developed in the case of discrete signals [8].

Moreover, we have presented a global constrained Markov model for contours in images that uses the conditional distributions of the gradient vector field on and off edges for the likelihood. Admittedly, the Markovian framework is a simplification of the great complexity of information contained in an image; however, this is a standard hypothesis that appears to be very useful in the context of *online* estimation procedures and unsupervised segmentation methods. We exploit the Markovian framework in the estimation procedure (using the ICE procedure) and in the segmentation method (using a MAP or MPM criterion). This yields an original unsupervised method for edge-detection.

The estimation and the segmentation procedures have been tested on a total of 160 images. For the class of images that we have tested, the experimental results indicate that the model presented in this paper and its estimation procedure are valid for applications that require the energy term based on the log-likelihood ratio. Other kinds of images might require other filters and other statistical distributions. As for image segmentation, other aspects of an image (such as

textures) require other filters as well and we admit that we have treated only the aspect of edges in this paper.

Most importantly, our model can be used to define cost functions suitable for semiautomatic extraction of contours, as in the live-wire algorithm [2] or the jetstream algorithm [3] (see [14] and [15]). As another application, our model can be used for localization of shapes (see [14] and [26]). Our model has also been used in computer graphics in the context of a sketching procedure for non-photo-realistic rendering [40]. In all those applications, the binary edge-detection itself is not used, but only the estimated distributions. For that matter, our model might be useful in various problems requiring a statistical likelihood for contours.

ACKNOWLEDGMENTS

The authors would like to thank FCAR (Fonds formation chercheurs & aide recherche, Qc., Canada) for financial support of this work. They are grateful to all of the anonymous reviewers for their comments that helped them improve both the technical content and the presentation quality of this paper. In particular, they acknowledge the contribution of the reviewer who suggested the use of the MPM classifier for the edge-detection method and the estimation of the *prior* model, as well as the reviewer who suggested the use of the SA algorithm; these two suggestions inclined the authors to stay within the framework of the constrained Markov model for contours presented in Section 4, rather than introducing a simplified model for edge-detection that ignores junction points in the sole purpose of applying the Viterbi algorithm. They would like to thank Professor Jean-François Angers and Professor Louis G. Doray for helpful discussions on ML estimators. They gratefully acknowledge the use of the University of South Florida database. They are grateful to Dr. David Martin for the permission to use the University of California at Berkeley image test data set.

REFERENCES

[1] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active Contour Models," *Int'l J. Computer Vision,* vol. 1, no. 4, pp. 321-331, 1988.

[2] E.N. Mortensen and W.A. Barrett, "Interactive Segmentation with Intelligent Scissors," *Graphical Models and Image Processing,* vol. 60, no. 5, pp. 349-384, Sept. 1998.

[3] P. Pérez, A. Blake, and M. Gangnet, "Jetstream: Probabilistic Contour Extraction with Particles," *Proc. Int'l Conf. Computer Vision,* July 2001.

[4] A.K. Jain, Y. Zhong, and S. Lakshmanan, "Object Matching Using Deformable Templates," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 3, pp. 267-278, Mar. 1996.

[5] J.F. Canny, "A Computational Approach to Edge Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 8, no. 6, pp. 679-698, 1986.

[6] R. Deriche, "Using Canny's Criteria to Derive a Recursively Implemented Optimal Edge Detector," *Int'l J. Computer Vision,* vol. 1, no. 2, pp. 167-187, 1987.

[7] J. Shen and S. Castan, "An Optimal Linear Operator for Step Edge Detection," *Computer Vision, Graphics, and Image Processing,* vol. 54, pp. 112-133, 1992.

[8] D. Demigny, "An Optimal Linear Filtering for Edge Detection," *IEEE Trans. Image Processing,* vol. 11, no. 7, pp. 728-737, 2002.

[9] K.W. Bowyer, C. Kranenburg, and S. Dougherty, "Edge Detector Evaluation Using Empirical ROC Curves," *Computer Vision and Image Understanding,* vol. 84, no. 10, pp. 77-103, 2001.

[10] S. Konishi, A.L. Yuille, J.M. Coughlan, and S.C. Zhu, "Statistical Edge Detection: Learning and Evaluating Edge Cues," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 1, pp. 57-74, Jan. 2003.

[11] D. Geman and B. Jedynak, "An Active Testing Model for Tracking Roads in Satellite Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 1, pp. 1-14, Jan. 1996.

[12] Z.W. Tu and S.C. Zhu, "Image Segmentation by Data-Driven Markov Chain Monte Carlo," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 5, pp. 657-673, May 2002.

[13] D. Geman, S. Geman, C. Graffigne, and P. Dong, "Boundary Detection by Constrained Optimization," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 12, no. 7, pp. 609-628, July 1990.

[14] F. Destrempes, "Détection Non-Supervisée de Contours et Localisation de Formes à l'Aide de Modèles Statistiques," Master Thesis, Université de Montréal, Apr. 2002.

[15] F. Destrempes and M. Mignotte, "Unsupervised Detection and Semi-Automatic Extraction of Contours Using a Statistical Model and Dynamic Programming," *Proc. Fourth IASTED Int'l Conf. Signal and Image Processing,* pp 60-65, Aug. 2002.

[16] B. Braathen, P. Masson, and W. Pieczynski, "Global and Local Methods of Unsupervised Bayesian Segmentation of Images," *Machine Graphics and Vision,* vol. 2, no. 1, pp. 39-52, 1993.

[17] A. Peng and W. Pieczynski, "Adaptive Mixture Estimation and Unsupervised Local Bayesian Image Segmentation," *CVGIP: Graphical Models and Image Processing,* vol. 57, no. 5, pp. 389-399, 1995.

[18] H. Caillol, W. Pieczynski, and A. Hillon, "Estimation of Fuzzy Gaussian Mixture and Unsupervised Statistical Image Segmentation," *IEEE Trans. Image Processing,* vol. 6, no. 3, pp. 425-440, 1997.

[19] N. Giordana and W. Pieczynski, "Estimation of Generalized Multisensor Hidden Markov Chains and Unsupervised Image Segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 5, pp. 465-475, May 1997.

[20] F. Salzenstein and W. Pieczinsky, "Parameter Estimation in Hidden Fuzzy Markov Random Fields and Image Segmentation," *CVGIP: Graphical Models and Image Processing,* vol. 59, no. 4, pp. 205-220, 1997.

[21] Y. Delignon, A. Marzouki, and W. Pieczynski, "Estimation of Generalized Mixture and Its Application in Image Segmentation," *IEEE Trans. Image Processing,* vol. 6, no. 10, pp. 1364-1375, 1997.

[22] W. Pieczynski, J. Bouvrais, and C. Michel, "Estimation of Generalized Mixture in the Case of Correlated Sensors," *IEEE Trans. Image Processing,* vol. 9, no. 2, pp. 308-311, 2000.

[23] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 6, no. 6, pp. 721-741, 1984.

[24] J. Besag, "On the Statistical Analysis of Dirty Pictures," *J. Royal Statistical Soc. B,* vol. 48, pp. 259-302, 1986.

[25] J. Maroquin, S. Mitter, and T. Poggio, "Probabilistic Solution of Ill-Posed Problems in Computation Vision," *J. Am. Statistical Assoc.,* vol. 82, no. 397, pp. 76-89, 1987.

[26] F. Destrempes and M. Mignotte, "Unsupervised Localization of Shapes Using Statistical Models," *Proc. Fourth IASTED Int'l Conf. Signal and Image Processing,* pp. 66-71, Aug. 2002.

[27] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics," *Proc. Eighth Int'l Conf. Computer Vision,* vol. 2, pp. 416-423, July 2001.

[28] F. Destrempes and M. Mignotte, "Unsupervised Texture Segmentation Using a Statistical Wavelet-Based Hierarchical Multi Data Model," *Proc. 10th IEEE Int'l Conf. Image Processing,* vol. II, pp. 1053-1056, Sept. 2003.

[29] S.A. Klugman, H.H. Panjer, and G. E. Willmot, *Loss Models,* Wiley Series in Probability and Statistics, 1998.

[30] G. Schwartz, "Estimating the Dimension of a Model," *Annals of Statistics,* vol. 6, pp. 461-464, 1978.

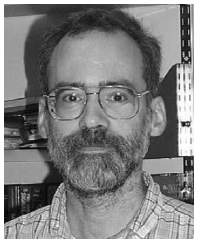
[31] H. Derin and H. Elliott, "Modeling and Segmentation of Noisy and Textured Images Using Gibbs Random Fields," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 9, no. 1, pp. 39-55, 1987.

[32] L Younes, "Parametric Inference for Imperfectly Observed Gibbsian Fields," *Springer-Verlag Probability Theory and Related Fields 82,* pp. 625-645, 1989.

[33] M. Mignotte, C. Collet, P. Perez, and P. Bouthemy, "Sonar Image Segmentation Using a Hierarchical MRF Model," *IEEE Trans. Image Processing,* vol. 9, no. 7, pp. 1216-1231, 2000.

[34] S. Banks, *Signal Processing, Image Processing and Pattern Recognition.* Prentice Hall, 1990.

[35] D. Geman and S. Geman, "Relaxation and Annealing with Constraints," Complex Systems Technical Report 35, Division Applied Math., Brown Univ., 1987.
[36] A.C. Cohen, "Maximum Likelihood Estimation in the Weibull Distribution Based on Complete and Censored Samples," *Technometrics,* vol. 7, pp. 579-588, 1965.
[37] J. Diebolt and G. Celeux, "Asymptotic Properties of a Stochastic EM Algorithm for Estimating Mixing Proportions," *Comm. Statistics Stochastic Models,* vol. 9, no. 4, pp. 599-613, 1993.
[38] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Royal Statistical Soc.,* pp. 1-38, 1976.
[39] A.J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm," *IEEE Trans. Information Theory,* vol. 13, pp. 260-269, 1967.
[40] M. Mignotte, "Unsupervised Statistical Sketching for Non-Photorealistic Rendering Models," *Proc. 10th IEEE Int'l Conf. Image Processing,* vol. III, pp. 573-576, Sept. 2003.



François Destrempes received the BSc degree in mathematics from the Université de Montréal (1985), the MSc degree (1987) and the PhD degree (1990) in mathematics from Cornell University. He was a postdoctoral fellow at the Centre de Recherche Mathématiques (CRM) of the Université de Montréal from 1990-1992. He has taught mathematics at Concordia University, the University of Ottawa, the University of Toronto, and the University of Alberta. He also received a postgraduate degree (2000) in applied computer science and the MSc degree (2002) in computer science from the Université de Montréal, where he is now studying in the PhD program in computer science. His current research interests include statistical methods for image segmentation, parameters estimation, detection of contours, localization of shapes, and applications of stochastic optimization to computer vision.

Max Mignotte received the DEA (postgraduate degree) in digital signal, image, and speech processing from the INPG University, France (Grenoble), in 1993 and the PhD degree in electronics and computer engineering from the University of Bretagne Occidentale (UBO) and the digital signal laboratory (GTS) of the French Naval Academy, France, in 1998. He was an INRIA postdoctoral fellow at the University of Montréal (DIRO), Canada (Québec), from 1998 to 1999. He is currently with DIRO at the Computer Vision & Geometric Modeling Lab as an assistant professor (professeur adjoint) at the University of Montreal. He is also a member of LIO (Laboratoire de Recherche en Imagerie et Orthopédie, Centre de Recherche du CHUM, Hôpital Notre-Dame) and researcher at CHUM. His current research interests include statistical methods and Bayesian inference for image segmentation (with hierarchical Markovian, statistical templates, or active contour models), hierarchical models for high-dimensional inverse problems from early vision, parameters estimation, tracking, classification, shape recognition, deconvolution, 3D reconstruction, and restoration problems.

Appendix A

Let $\tilde{y}_1, \ldots, \tilde{y}_m$ be m positive real numbers, and consider the function

$$F(C) = \frac{\sum_{i=1}^m \tilde{y}_i^C \ln y_i}{\sum_{i=1}^m \tilde{y}_i^C} - \frac{1}{m}\sum_{i=1}^m \ln \tilde{y}_i.$$

The derivative of the function F with respect to C is equal to

$$\frac{\{\sum_{i=1}^m \tilde{y}_i^C \ln^2 \tilde{y}_i\}\{\sum_{i=1}^m \tilde{y}_i^C\} - \{\sum_{i=1}^m \tilde{y}_i^C \ln \tilde{y}_i\}\{\sum_{i=1}^m \tilde{y}_i^C \ln \tilde{y}_i\}}{\left(\sum_{i=1}^m \tilde{y}_i^C\right)^2}.$$

The numerator is equal to $||(\tilde{y}_i^{C/2} \ln \tilde{y}_i)||^2 ||(\tilde{y}_i^{C/2})||^2 - |\langle (\tilde{y}_i^{C/2}), (\tilde{y}_i^{C/2} \ln \tilde{y}_i)\rangle|^2$, which is non-negative by Cauchy-Schwartz inequality (in the context of the Euclidean norm). The denominator being positive, we conclude that $F'(C) \geq 0$ on $(0, \infty)$. Furthermore, $F(0) = 0$, and $\lim_{C\to\infty} F(C) = \max_i \ln \tilde{y}_i - \frac{1}{m}\sum_{i=1}^m \ln \tilde{y}_i$, which is positive unless all $\tilde{y}_i$ are equal.

Thus, except in the degenerate case, we have $-\frac{1}{C^2} - F'(C) < 0$ on the interval $(0, \infty)$; $\lim_{C\to 0} \frac{1}{C} - F(C) = \infty$; and $\lim_{C\to\infty} \frac{1}{C} - F(C) < 0$. We conclude that the continuous function $\frac{1}{C} - F(C)$ is decreasing and has a unique root in the domain $(0, \infty)$.

Appendix B

Let $y_1, \ldots, y_m$ be m real numbers in the interval $[0, \frac{\pi}{2}]$, and consider the log-likelihood function of a truncated Exponential distribution $\{\alpha(1 - \exp(-\frac{\pi}{2\alpha})\}^{-1} \exp(-\frac{y}{\alpha})$:

$$\ln \mathcal{L}(\alpha) = -\ln(\alpha) - \ln(1 - \exp(-\frac{\pi}{2\alpha})) - \frac{\bar{y}}{\alpha},$$

where $\bar{y} = \frac{1}{m}\sum_{i=1}^m y_i$. We have $-\alpha^2 \frac{d \ln \mathcal{L}}{d\alpha}(\alpha) = \alpha - H(\alpha)$, where

$$H(\alpha) = \bar{y} + \frac{\pi}{2(\exp(\frac{\pi}{2\alpha}) - 1)}.$$

We compute: $1 - \frac{dH}{d\alpha}(\alpha) = 1 - \frac{\pi^2}{16\alpha^2 \sinh^2(\frac{\pi}{4\alpha})}$. From the inequality $\sinh^2(t) - t^2 > 0$ holding for $t > 0$, we conclude that $1 - \frac{dH}{d\alpha}(\alpha) > 0$ upon taking $t = \frac{\pi}{4\alpha}$. Thus, the function $\alpha - H(\alpha)$ is increasing on the interval $(0, \infty)$.

Now, assume that $0 < \bar{y} < \frac{\pi}{4}$. We have $\lim_{\alpha \to 0} \alpha - H(\alpha) = -\bar{y} < 0$ and $\lim_{\alpha \to \infty} \alpha - H(\alpha) = -\bar{y} + \frac{\pi}{4} > 0$. Hence, in that case, the function $\alpha - H(\alpha)$ admits a unique root in the interval $(0, \infty)$.

Next, assume that $\frac{\pi}{4} \leq \bar{y}$. In that case, $\lim_{\alpha \to \infty} \alpha - H(\alpha) = -\bar{y} + \frac{\pi}{4} \leq 0$. Thus, $\alpha - H(\alpha) < 0$ on the interval $(0, \infty)$. It follows that $\frac{d \ln \mathcal{L}}{d \alpha}(\alpha) > 0$, and henceforth that the function $\ln \mathcal{L}$ is increasing on $(0, \infty)$. Thus, $\ln \mathcal{L}$ is maximal at $\alpha = \infty$. Note that $\lim_{\alpha \to \infty} \{\alpha(1 - \exp(-\frac{\pi}{2\alpha})\}^{-1} \exp(-\frac{y}{\alpha}) = \frac{2}{\pi}$ for any $y \geq 0$.