

# A Nonparametric Statistical Method for Image Segmentation Using Information Theory and Curve Evolution

Junmo Kim\*, John W. Fisher III, Anthony Yezzi, Jr., Müjdat Çetin, and Alan S.  
Willsky

77 Mass. Ave. Rm 35-425  
Cambridge, MA 02139, USA  
Phone: (617) 253-4874 Fax: (617) 258-8364  
E-mail: junmo@mit.edu

EDICS: 2-SEGM Segmentation

This work was supported by the Office of Naval Research under Grant N00014-00-1-0089, and the Air Force Office of Scientific Research under Grant F49620-00-0362

J. Kim, J. W. Fisher, M. Çetin, and A. S. Willsky are with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: junmo@mit.edu, fisher@ai.mit.edu, mcetin@mit.edu, willsky@mit.edu).

J. W. Fisher is with the Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: fisher@ai.mit.edu).

A. Yezzi is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA (e-mail: ayezzi@ece.gatech.edu).

### Abstract

In this paper, we present a new information-theoretic approach to image segmentation. We cast the segmentation problem as the maximization of the mutual information between the region labels and the image pixel intensities, subject to a constraint on the total length of the region boundaries. We assume that the probability densities associated with the image pixel intensities within each region are completely unknown a priori, and we formulate the problem based on nonparametric density estimates. Due to the nonparametric structure, our method does not require the image regions to have a particular type of probability distribution, and does not require the extraction and use of a particular statistic. We solve the information-theoretic optimization problem by deriving the associated gradient flows and applying curve evolution techniques. We use level set methods to implement the resulting evolution. The experimental results based on both synthetic and real images demonstrate that the proposed technique can solve a variety of challenging image segmentation problems. Furthermore, our method, which does not require any training, performs as good as methods based on training.

### Keywords

Image segmentation, curve evolution, level set methods, nonparametric density estimation, information theory

## I. INTRODUCTION

Image segmentation is an important problem in image analysis, appearing in many applications including pattern recognition, object detection, and medical imaging. Some previous approaches to image segmentation, which provide the basis for a variety of more recent methods, include boundary-based segmentation such as Canny edge detection [5], region-based segmentation such as region growing [2], [25], and global optimization approaches such as those based on the Mumford-Shah functional [28], [9], [39]. Recently there has been a considerable amount of work on image segmentation using curve evolution techniques [6], [7], [8], [9], [27], [31], [34], [35], [43], [44]. Some of these techniques, including the ones in [31], [44] have relations to the approach we present here. In particular, Paragios et al. [31] developed a parametric model for supervised segmentation of textured images. Yezzi et al. [44] developed a segmentation technique using a particular discriminative statistical feature such as the mean or the variance of image regions. These and many other recent methods (such as [35], [38]) have been inspired by the region competition model of Zhu and Yuille [45].

Our strategy is different from those of previous curve evolution-based methods in three major

ways. First, unlike the techniques mentioned above, our approach is based on nonparametric statistics. The performance of parametric methods can be severely affected when the assumed parametric model is not correct. This limits the class of images that can be segmented using such methods with a particular parametric model. In response to the need for robustness and a larger modeling capacity in statistical analysis, nonparametric methods [32] have been widely used in machine learning problems. Nonparametric methods estimate the underlying distributions from the data without making strong assumptions about the structures of the distributions. The nonparametric aspect of our approach makes it especially appealing when there is little or no prior information about the statistical properties of the regions to be segmented. Note that there is a trade-off, namely, with a nonparametric approach we expect some performance loss when the image fits a parametric model. However, we will give examples that clearly make the case that there are rich classes of real images for which our method is advantageous. In particular, we will show a compelling example where two regions of same means and same variances are segmented.

The second aspect of our technique is that no training is required. Again this has advantages and disadvantages. Obviously if one *has* training data from which to learn the distributions of the image regions, one should take advantage of this, as in Paragios et. al. [31]. However, it is also of practical interest to develop methods that do not require prior knowledge. We will see that the method developed here can yield results as good as those of other methods which take advantage of prior training (which our method does not, and simply must perform segmentation based on the image presented to it without *any* prior training.)

The third aspect of our technique is that this is a principled information-theoretic framework (using mutual information) that allows us to understand the several key quantities that drive the resulting curve evolution. In particular, the first such term is a likelihood ratio (LR) term that is similar to that used by Zhu *et al.* [45], the difference being that in [45] LR is computed using parametric distributions whose parameters are estimated at each iteration, while ours uses distributions that are learned and dynamically adapted in a nonparametric way. If the particular parametric model is not well-matched to data, the nonparametric method will outperform the parametric counterpart. Even though the image fits the parametric model, our distribution estimates approach the quality achieved by parametric estimates. The formalism we describe also includes two additional terms which capture the sensitivity of the estimated distributions

(and hence the LR) to changes in the segmenting curve as it evolves.

The technique proposed by Jehan-Besson et al. [21] is related to our work regarding these additional terms. The work in [21] considers general region-based active contours, where the energy functionals to minimize are given as region-integrals of so-called descriptors. In particular, they consider the case where the descriptors themselves depend on the region, and formulate an optimization method. Their formulation can also be applied to our energy functional, which is also region-based. What is new with our method is that our energy functional is based on mutual information and that our “descriptor” involves nonparametric density estimates, whereas they consider means, variances, determinants of covariance matrices, and histograms (in their subsequent work [3]) as the descriptors.

The curve evolution technique in [40], [41] also takes a nonparametric approach to the image segmentation problem with an information-theoretic perspective. However, their approach is different from ours in a number of ways. First, they implement their technique for polygonal contours whereas we implement evolution of continuous curves. Furthermore, their approach only utilizes the first few estimated moments as approximations of the non-parametric distributions whereas our approach uses nonparametric estimates of the entire distributions.

There exists some other work aimed at building a framework for segmenting a large class of images. In particular, the technique proposed by Heiler et al. [19] is motivated by recent work on natural image statistics [33], [20], and is based on parametric modeling of filter responses by generalized Laplacian distributions. Their energy functional involves a Kullback-Leibler divergence between those parametric densities, and their method can segment both gray level natural images and textured images in an unsupervised fashion. In addition, there exists some other work pursuing a similar goal of segmenting multiple types of images, but using a framework other than active contours. For example, Malik et al. [26] have proposed an approach that works on a variety of both gray level images and textured images based on a graph-theoretic framework.

The remainder of this paper is organized as follows. Section II presents the information-theoretic objective functional for two-region image segmentation. Section III contains our curve evolution-based approach to minimizing this objective functional. Section IV presents an extension of the two-region version of the technique to the multi-phase segmentation problem. We then present experimental results in Section V, using both synthetic images with a variety of distributions and real images. Finally, we conclude in Section VI with a summary.

## II. INFORMATION-THEORETIC COST FUNCTIONAL FOR IMAGE SEGMENTATION

### A. Problem Statement

In this section, we consider a two-region image segmentation problem. The two regions are distinct in the sense that they have different probability density functions for the pixel intensities. We assume that the pixel intensities in each region are independent, identically distributed (i.i.d.). The associated probability density functions are unknown, and we impose no constraints on the form of these densities. More formally, the image intensity at pixel  $x$ , denoted by  $G(x)$ , is drawn from the density  $p_1$  if  $x \in R_1$ , and from  $p_2$  if  $x \in R_2$  as follows:

$$\begin{aligned} \{G(x)|x \in R_1\} &\stackrel{i.i.d.}{\sim} p_1 \\ \{G(x)|x \in R_2\} &\stackrel{i.i.d.}{\sim} p_2, \end{aligned} \quad (1)$$

where  $R_1$  and  $R_2$  denote the two regions which are unknown, and the associated densities  $p_1$  and  $p_2$  are also unknown. In other words, we model the observed pixel intensities as a spatial random process  $G(x)$  with pixel index  $x$ . Note that the lower case  $x$  is not a random variable but a pixel index. Later we will introduce a random variable  $X$ , which is written in a capital letter. The left-hand side of Figure 1 illustrates this image model. Note that a region can be composed of several topologically separate components, as shown in this figure. This image model is similar to that of the region competition method of Zhu and Yuille [45] in that both models assume that pixel intensities in each region are i.i.d. The difference is that here the distributions are unknown, whereas the model in [45] uses a family of pre-specified probability distributions.

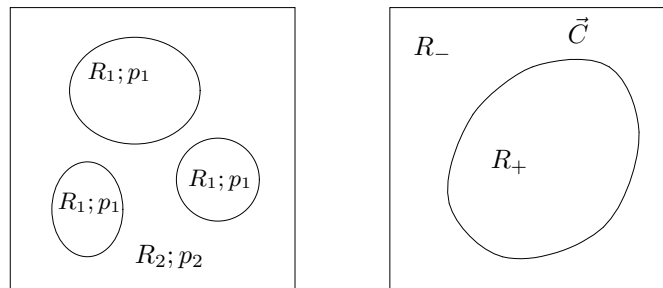


Fig. 1. Left: Illustration of the foreground region ( $R_1$ ), the background region ( $R_2$ ), and the associated distributions ( $p_1$  and  $p_2$ ). Right: Illustration of the curve ( $\vec{C}$ ), the region inside the curve ( $R_+$ ), and the region outside the curve ( $R_-$ ).

The goal of two-region image segmentation by curve evolution is to move a curve  $\vec{C}$  such that

it matches the boundary between  $R_1$  and  $R_2$ , i.e. the region inside the curve  $R_+$  and the region outside the curve  $R_-$  converge to  $R_1$  and  $R_2$  respectively or vice versa. The right-hand side of Figure 1 illustrates two regions,  $R_+$  and  $R_-$ . This partitioning of the image domain by the curve  $\vec{C}$  gives us a binary label  $L_{\vec{C}} : \Omega \rightarrow \{L_+, L_-\}$ , which is a mapping from the image domain  $\Omega$  to a set of two labeling symbols  $\{L_+, L_-\}$  defined as follows:

$$L_{\vec{C}}(x) = \begin{cases} L_+ & \text{if } x \in R_+ \\ L_- & \text{if } x \in R_- \end{cases} \quad (2)$$

By this correspondence between labels and curves, image segmentation is equivalent to the binary labeling problem.

### B. Mutual Information between the Image Intensity and the Label

We now introduce the mutual information (MI) between the image intensity and the label and discuss its properties. Let us initially consider the case where  $p_1$  and  $p_2$  are known. As mentioned before, we have a candidate segmenting curve  $\vec{C}$ , and  $R_1, R_2$  are the true unknown regions. Now suppose that we randomly choose a point  $X$  in  $\Omega$  such that  $X$  is a uniformly distributed random location in the image domain.<sup>1</sup> In this case, the label  $L_{\vec{C}}(X)$  is a binary random variable that depends on the curve  $\vec{C}$ . It takes the values  $L_+$  and  $L_-$  with probability  $\frac{|R_+|}{|\Omega|}$  and  $\frac{|R_-|}{|\Omega|}$  respectively, where  $|R_+|$  denotes the area of the region  $R_+$ .

On the other hand, the image intensity  $G(X)$  is a random variable that depends on the true regions  $R_1$  and  $R_2$ , and has the following density

$$p_{G(X)}(z) = Pr(X \in R_1)p_{G(X)|X \in R_1}(z) + Pr(X \in R_2)p_{G(X)|X \in R_2}(z) \quad (3)$$

$$= \frac{|R_1|}{|\Omega|}p_1(z) + \frac{|R_2|}{|\Omega|}p_2(z), \quad (4)$$

where  $z$  is an argument for the densities. Note that this density  $p_{G(X)}$  is a mixture of  $p_1$  and  $p_2$  due to the randomness of the pixel location  $X$ . As can be seen in (3),  $G(X)$  has two sources of uncertainty, namely the uncertainty of pixel location being in  $R_1$  or  $R_2$ , and the uncertainty of the intensity given the pixel location. The binary label  $L_{\vec{C}}(X)$  contains some information about the former uncertainty, namely  $X$  being in  $R_1$  or  $R_2$ . Therefore, intuitively speaking, the more

<sup>1</sup>This is similar to the work of Viola *et al.* [42], where they measure the amount of dependence between two images  $u(x)$  and  $v(x)$  by mutual information  $I(u(X); v(X))$ , where  $X$  is a random variable, which ranges over the domain of  $u(\cdot)$  and  $v(\cdot)$ .

accurately the label  $L_{\vec{C}}(X)$  can determine whether  $X \in R_1$  or  $X \in R_2$ , the less uncertainty  $G(X)$  has, and the more information about  $G(X)$  the label will have. This motivates using the mutual information  $I(G(X); L_{\vec{C}}(X))$  as a segmentation criterion.

Now let us consider more formally the mutual information  $I(G(X); L_{\vec{C}}(X))$

$$\begin{aligned} I(G(X); L_{\vec{C}}(X)) &= h(G(X)) - h(G(X)|L_{\vec{C}}(X)) \\ &= h(G(X)) - Pr(L_{\vec{C}}(X) = L_+)h(G(X)|L_{\vec{C}}(X) = L_+) \\ &\quad - Pr(L_{\vec{C}}(X) = L_-)h(G(X)|L_{\vec{C}}(X) = L_-) \end{aligned} \quad (5)$$

where the differential entropy  $h(Z)$  of a continuous random variable  $Z$  with support  $S$  is defined by  $h(Z) = -\int_S p_Z(z) \log p_Z(z) dz$ . The three entropies in (5) are functionals of  $p_{G(X)}$ ,  $p_{G(X)|L_{\vec{C}}(X)=L_+}$ , and  $p_{G(X)|L_{\vec{C}}(X)=L_-}$  respectively. The two conditional distributions are given as follows:

$$\begin{aligned} p_{G(X)|L_{\vec{C}}(X)=L_+}(z) &= \sum_{i=1}^2 Pr(X \in R_i | L_{\vec{C}}(X) = L_+) p_{G(X)|X \in R_i, L_{\vec{C}}(X)=L_+}(z) \\ &= \frac{|R_+ \cap R_1|}{|R_+|} p_1(z) + \frac{|R_+ \cap R_2|}{|R_+|} p_2(z) \end{aligned} \quad (6)$$

$$p_{G(X)|L_{\vec{C}}(X)=L_-}(z) = \frac{|R_- \cap R_1|}{|R_-|} p_1(z) + \frac{|R_- \cap R_2|}{|R_-|} p_2(z) \quad (7)$$

Each conditional entropy measures the degree of heterogeneity in each region determined by the curve  $\vec{C}$ . In other words, the more homogeneous the segmented regions, the less the conditional entropies, and the higher the mutual information is, which is a desirable property for segmentation.

We can show that the mutual information  $I(G(X); L_{\vec{C}}(X))$  is maximized if and only if  $\vec{C}$  is the correct segmentation, i.e. if  $R_+ = R_1$ ,  $R_- = R_2$  (or equivalently  $R_+ = R_2$ ,  $R_- = R_1$ ). The proof is given in Appendix-A. This result suggests that mutual information is a reasonable criterion for the segmentation problem we have formulated.

However, in practice, we really cannot compute the mutual information  $I(G(X); L_{\vec{C}}(X))$  for two reasons. First, the computations above involve the regions  $R_1$  and  $R_2$ , which are actually unknown to us (otherwise the segmentation problem would be solved). Second, unlike what we assumed in the above discussion, we would like to solve the segmentation problem when  $p_1$  and  $p_2$  are unknown.

We thus need to estimate the mutual information as follows:

$$\begin{aligned} \hat{I}(G(X); L_{\vec{C}}(X)) &= \hat{h}(G(X)) - Pr(L_{\vec{C}}(X) = L_+) \hat{h}(G(X) | L_{\vec{C}}(X) = L_+) \\ &\quad - Pr(L_{\vec{C}}(X) = L_-) \hat{h}(G(X) | L_{\vec{C}}(X) = L_-) \end{aligned} \quad (8)$$

This in turn requires us to estimate the densities  $p_{G(X)}$ ,  $p_{G(X)|L_{\vec{C}}(X)=L_+}$ , and  $p_{G(X)|L_{\vec{C}}(X)=L_-}$ . The way we estimate these densities are presented in Section III.

### C. The Energy Functional

Finally, we combine the mutual information estimate with the typical regularization penalizing the length of the curve in order to construct our overall energy functional to be used for segmentation. This regularization prevents the formation of a longer jagged boundary. Depending on the prior information one might have about the region boundaries, constraints other than the curve length penalty can also be used in our framework.

In the energy functional, the mutual information should be weighted by the area of the image domain in order to represent the total amount of information between the label and the image, since  $I(G(X); L_{\vec{C}}(X))$  corresponds to the contribution of a single pixel to the total information. The resulting energy functional to minimize is then given by

$$E(\vec{C}) = -|\Omega| \hat{I}(G(X); L_{\vec{C}}(X)) + \alpha \oint_{\vec{C}} ds, \quad (9)$$

where  $\oint_{\vec{C}} ds$  is the length of the curve and  $\alpha$  is a scalar parameter. The statistical interpretation of this energy functional is given in Appendix-B.

## III. NONPARAMETRIC DENSITY ESTIMATION AND GRADIENT FLOWS

This section contains the derivation of the curve evolution formula for minimizing the energy functional  $E(\vec{C})$  of (9) using nonparametric Parzen density estimates. First, we present the way the nonparametric Parzen density estimates are used in estimating the conditional entropy terms in (8). This results in the expression of the energy functional  $E(\vec{C})$  in the form of nested region integrals. We then calculate the gradient flow for  $E(\vec{C})$  and discuss the properties of the curve evolution formula.



### A. Estimation of the Differential Entropy

The expression (8) involves differential entropy estimates, and we use nonparametric Parzen density estimates in order to estimate the differential entropies. For a review of nonparametric entropy estimation, we refer the reader to [4].

Since  $\hat{h}(G(X))$  in (8) is independent of the curve, we just consider  $\hat{h}(G(X)|L_{\vec{C}}(X) = L_+)$  and  $\hat{h}(G(X)|L_{\vec{C}}(X) = L_-)$  as follows:

$$\begin{aligned} & \hat{h}(G(X)|L_{\vec{C}}(X) = L_+) \\ &= -\frac{1}{|R_+|} \int_{R_+} \log \hat{p}_+(G(x)) dx \end{aligned} \quad (10)$$

$$= -\frac{1}{|R_+|} \int_{R_+} \log \left( \frac{1}{|R_+|} \int_{R_+} K(G(x) - G(\hat{x})) d\hat{x} \right) dx, \quad (11)$$

Note that  $h(G(X)|L_{\vec{C}}(X) = L_+)$  involves the expected value of the logarithm of  $p_+ \triangleq p_{G(X)|L_{\vec{C}}(X)=L_+}$ , and we approximate this expected value by the sample mean of  $\log p_+$  in (10). We then use a continuous version of the Parzen density estimate [32] of  $p_+$  in (11). We use the kernel  $K(z) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{z^2}{2\sigma^2}}$ , where  $\sigma$  is a scalar parameter. Similarly, we have:

$$\begin{aligned} & \hat{h}(G(X)|L_{\vec{C}}(X) = L_-) \\ &= -\frac{1}{|R_-|} \int_{R_-} \log \left( \frac{1}{|R_-|} \int_{R_-} K(G(x) - G(\hat{x})) d\hat{x} \right) dx \end{aligned} \quad (12)$$

### B. Gradient Flows for General Nested Region Integrals

Note that (11) and (12) have nested region integrals. Let us consider a general nested region integral of the form

$$\int_R f(\varepsilon(x, t)) dx \quad \text{where} \quad \varepsilon(x, t) = \int_R g(x, \hat{x}) d\hat{x}, \quad \text{and } g \text{ does not depend on } \vec{C}, \quad (13)$$

where  $R$  is the region inside the curve  $\vec{C}$  and  $t$  is a time index for the evolution of  $\vec{C}$  (which we often drop for notational convenience as in  $R = R(t)$  and  $\vec{C} = \vec{C}(t)$ ). For such integrals we have derived the gradient flow (the negative of the gradient so that the region integral decreases most rapidly), which is given by

$$\frac{\partial \vec{C}}{\partial t} = - \left[ f(\varepsilon(\vec{C})) + \int_R f'(\varepsilon(x)) g(x, \vec{C}) dx \right] \vec{N}, \quad (14)$$

where  $\vec{N}$  is the outward unit normal vector. The detailed derivation can be found in Appendix-C.

The second term appears in (14) because the integrand  $f(\varepsilon(x, t))$  in (13) depends on the curve  $\vec{C}$ .

### C. The Gradient Flow for the Information-Theoretic Energy Functional

Now that we have the nonparametric estimates of the mutual information in the form of nested region integrals as in (11) and (12), it is straightforward to calculate the gradient flow for the energy functional  $E(\vec{C})$  using the result of Section III-B. We provide the details of this computation in Appendix-D. Here, we state the main result, namely the overall gradient flow for  $E(\vec{C})$  of (9):

$$\begin{aligned} \frac{\partial \vec{C}}{\partial t} = & \left[ \log \frac{\hat{p}_+(G(\vec{C}))}{\hat{p}_-(G(\vec{C}))} + \frac{1}{|R_+|} \int_{R_+} \frac{K(G(x) - G(\vec{C}))}{\hat{p}_+(G(x))} dx \right. \\ & \left. - \frac{1}{|R_-|} \int_{R_-} \frac{K(G(x) - G(\vec{C}))}{\hat{p}_-(G(x))} dx \right] \vec{N} - \alpha \kappa \vec{N}, \end{aligned} \quad (15)$$

where  $\kappa$  is the curvature of the curve and  $-\alpha \kappa \vec{N}$  is the gradient flow for the curve length penalty, whose derivation can be found in [16]. We implement the curve evolution for the gradient flow in (15) using the level set method [30], [36] together with the narrow band approach [11], [1].

A direct computation of this gradient flow is expensive. In particular, the bottleneck is in the computation of the second and the third terms. If we use a direct computation, it takes  $\mathcal{O}((\# \text{ of pixels})^2)$  time per each iteration, which we now explain. Since the evaluation of the density estimate in the form of  $\hat{p}_+(G(x)) = \frac{1}{N} \sum_{i=1}^N K(G(x) - G(x_i))$  at each pixel  $x$  on the curve takes  $\mathcal{O}(N)$  time, evaluation of  $\hat{p}_+(G(\vec{C}))$  at each pixel on the curve takes  $\mathcal{O}(|R_+|)$  time, where  $|R_+|$  is the number of pixels in region inside the curve. Thus the computation of the first term at all the points on the curve takes  $\mathcal{O}(M(|R_+| + |R_-|))$  time, where  $M$  is the number of pixels along the curve (i.e. the size of the narrow band). In order to compute the second term, we compute and store  $\hat{p}_+(G(x))$  for all  $x \in R_+$ , which takes  $\mathcal{O}(|R_+|^2)$  time and then compute the integral using the stored values of  $\hat{p}_+(G(x))$ . The computation of this integral at all the points on the curve takes  $\mathcal{O}(M|R_+|)$  time. Therefore, the complexity of a direct computation of the gradient flow is  $\mathcal{O}(M(|R_+| + |R_-|) + |R_+|^2 + M|R_+| + |R_-|^2 + M|R_-|) \sim \mathcal{O}((\# \text{ of pixels})^2)$  per each step.

However, we reduce the complexity by using an approximation method based on the fast Gauss transform (FGT) [18], [37], [17]. FGT can evaluate density estimates based on  $N$  data points in the form of  $\hat{P}(x) = \frac{1}{N} \sum_{i=1}^N K(x - x_i)$  at  $M$  different points in  $\mathcal{O}(c(M + N))$  time instead

of  $\mathcal{O}(MN)$  time, where  $c$  is the precision number which grows with the required precision of the approximation. The precision number  $c$  is the order of the Taylor series expansions used in FGT, and  $c$  less than 10 is often sufficient in most cases. Furthermore, in evaluating  $\hat{p}_+$ , we observe that using only a randomly selected subset of  $R_+$  is sufficient instead of using all the pixel intensities in  $R_+$ . If we select  $N$  points from  $R_+$  in order to estimate  $\hat{p}_+$  and another  $N$  points from  $R_-$ , the computational cost using FGT per each iteration is  $\mathcal{O}(c(M + N + N) + c(N + N) + c(M + N) + c(N + N) + c(M + N))$ , where the integral in the second and third term in (15) takes  $\mathcal{O}(c(M + N))$  time by FGT. Given the size of the narrow band, a reasonable choice of  $N$  will be a linear function of  $M$ . This results in the overall complexity of  $\mathcal{O}(M)$ , i.e. linear in the size of the narrow band.

In general, FGT is also possible for estimation of multi-dimensional density functions, which will allow us to extend our framework to color and vector-valued images. For  $d$  dimensional data, the complexity of FGT is now  $\mathcal{O}(c^d(M + N))$  [17], with the same  $M$  and  $N$  as the above. The only difference in computational complexity from the case of gray level images is in the constant factor  $c^d$ . Therefore, the computational complexity is still linear in the size of the narrow band, if our method is extended to vector-valued images.

Since our energy functional involves a curve length penalty term, we have a curvature flow term in (15). If we replace the curve length penalty term by  $\oint_{\vec{C}} g(\vec{C}(s)) ds$  as in the geodesic active contours [7], the evolution equation (15) will have  $\text{div}(g(x, y) \frac{\nabla \phi}{|\nabla \phi|}) |\nabla \phi|$  replacing the curvature flow, where  $\phi$  is the corresponding level set function. In this case, the fast geodesic active contours proposed by Goldenberg et al. [14] can be combined with our framework, just as the simplified Mumford Shah model of Chan and Vese [9] has been combined with the fast geodesic active contours in Kimmel et al. [24]. For further information on the fast geodesic active contours, we refer the readers to [29].

#### *D. Discussion on the Gradient Flow*

The first term of the gradient flow expression in (15) is a log-likelihood ratio which compares the hypotheses that the observed image intensity  $G(\vec{C})$  at a given point on the active contour  $\vec{C}$  belongs to the foreground region  $R_+$  or the background region  $R_-$  based upon the current estimates of the distributions  $p_+$  and  $p_-$ . By this log-likelihood ratio term, the pixel on the boundary is merged to either the region  $R_+$  or the region  $R_-$  such that the updated regions are more homogeneous.

To understand the second and third terms in (15), let us consider the analogy to the generic flow in (14). We have the second term of (14) because the integrand  $\varepsilon(\cdot)$  in (13) depends on the curve. Similarly, we have the second and third terms in the gradient flow (15) because the integrands of the entropy estimates (11) and (12), which are logarithms of Parzen density estimates, depend on the curve.

These second and third terms reinforce and refine what the first term does. The first term alone does not take into account the fact that a deformation of the curve results in updating the data samples used for the two density estimates. It is the two additional terms that compensate for the change of density estimates.

These second and third terms, as well as the use of the nonparametric density estimates distinguish this active contour model from the region competition algorithm of Zhu and Yuille [45], which involves alternating iterations of two operations: estimating the distribution parameters inside and outside the curve; and likelihood ratio tests to evolve the curve. In that algorithm, changes in the distributions are not directly coupled with likelihood ratio tests. In contrast, the changes in the nonparametric density estimates are built directly into our curve evolution equation through these two terms.

#### IV. EXTENSION TO MULTI-PHASE SEGMENTATION

In this section, we provide an extension of the two-region version of our technique to images with more than two regions. To this end, we incorporate the multi-phase segmentation formulation of [10] into our information-theoretic, nonparametric segmentation framework. Our method uses  $m$  level set functions to segment up to  $2^m$  regions, and the resulting curve evolution equation (motion equation) turns out to be a natural generalization of nonparametric region competition.

##### A. $n$ -ary Segmentation Problem and Mutual Information

We extend the two-region image segmentation problem to an  $n$ -ary (i.e.  $n$ -region) version, where  $R_1, \dots, R_n$  denote the true unknown regions, and the image intensity at pixel  $x$ , denoted by  $G(x)$ , is drawn from the density  $p_i$  if  $x \in R_i$ , where  $p_i$ 's are unknown. Figure 2(a) illustrates this image model when  $n = 4$ .

The goal of  $n$ -ary image segmentation by curve evolution is to move a set of curves  $\{\vec{C}_1, \dots, \vec{C}_m\}$  (equivalently, a set of level set functions  $\{\phi_1, \dots, \phi_m\}$ ) such that these curves partition the image domain into the true regions  $R_1, \dots, R_n$ . Each curve  $C_i$  partitions the image domain into the

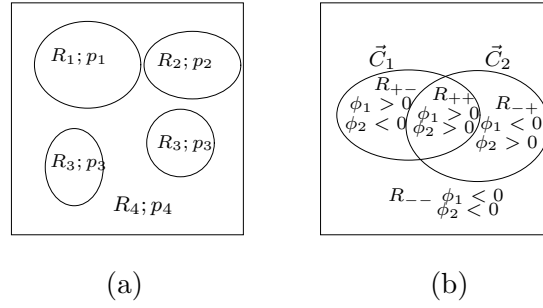


Fig. 2. Multi-phase segmentation image model. (a) Illustration of the case where  $n = 4$ : true regions  $R_1, \dots, R_4$ , with the associated distributions  $p_1, \dots, p_4$ . (b) Illustration of the two curves ( $\vec{C}_1, \vec{C}_2$ ) and the regions  $R_{++}, R_{+-}, R_{-+}, R_{--}$  partitioned by the curves .

two regions, the region inside the curve and the region outside the curve ( $\phi_i$  does the same thing by its sign). Thus the  $m$  level set functions partition the image domain into  $2^m$  regions, each of which we label by the signs of the level set functions in that region. For instance, when  $m = 2$ , we have 4 regions,  $R_{++}, R_{+-}, R_{-+}, R_{--}$  as illustrated in Figure 2(b).

Given the partitioning by the curves  $\mathbf{C} \triangleq \{\vec{C}_i\}_{i=1}^m$ , we can label each pixel  $x$  by its label  $L_{\mathbf{C}}(x)$ . For instance, if  $x \in R_{++}$ ,  $L_{\mathbf{C}}(x) = L_{++}$ . More formally, this partitioning of the image domain by the curves  $\mathbf{C}$  gives us a label

$$L_{\mathbf{C}} : \Omega \rightarrow \{L_{++++}, \dots, L_{----}\},$$

which is a mapping from the image domain  $\Omega$  to a set of  $2^m$  labeling symbols  $\{L_{++++}, \dots, L_{----}\}$  defined as follows:

$$L_{\mathbf{C}}(x) = L_{s(i)} \text{ if } x \in R_{s(i)}, 1 \leq i \leq 2^m, \quad (16)$$

where  $s(i)$  is the  $i$ th element in the set  $\{++++, \dots, ----\}$ . By a straightforward generalization of (9), we propose the following energy functional for multi-phase segmentation:

$$E(\mathbf{C}) = -|\Omega| \hat{I}(G(X); L_{\mathbf{C}}(X)) + \alpha \sum_{i=1}^m \oint_{\vec{C}_i} ds, \quad (17)$$

where the mutual information estimate is naturally extended to:

$$\hat{I}(G(X); L_{\mathbf{C}}(X)) = \hat{h}(G(X)) - \sum_{i=1}^{2^m} Pr(L_{\mathbf{C}}(X) = L_{s(i)}) \hat{h}(G(X) | L_{\mathbf{C}}(X) = L_{s(i)}) \quad (18)$$

### B. The Gradient Flows

We now compute the gradient flow to minimize  $E(\mathbf{C})$  of (17). For notational convenience, we consider the case where  $m = 2$ , but the development could easily be generalized to any  $m$ .

In (18), we have  $2^m = 4$  conditional entropies to estimate, namely,  $\hat{h}(G(X)|L_{\mathbf{C}}(X) = L_{++}), \dots, \hat{h}(G(X)|L_{\mathbf{C}}(X) = L_{--})$ . We compute these estimates in a way that is analogous to what we did for the two-region case. For example,  $\hat{h}(G(X)|L_{\mathbf{C}}(X) = L_{++})$  is given by

$$\begin{aligned} \hat{h}(G(X)|L_{\mathbf{C}}(X) = L_{++}) &= -\frac{1}{|R_{++}|} \int_{R_{++}} \log \hat{p}_{++}(G(x)) dx \\ &= -\frac{1}{|R_{++}|} \int_{R_{++}} \log \left( \frac{1}{|R_{++}|} \int_{R_{++}} K(G(x) - G(\hat{x})) d\hat{x} \right) dx, \end{aligned} \quad (19)$$

and the other entropy estimates are obtained in a similar way.

Generalizing our results from Section III, and using the multi-phase segmentation formulation of [10], we compute the first variation of the energy functional  $E(\mathbf{C})$  in (17), and obtain the following coupled motion equations:

$$\begin{aligned} \frac{\partial \vec{C}_1}{\partial t} &= \vec{N}_1 \left[ -\alpha \kappa_1 + H(\phi_2(\vec{C}_1)) \left( \log \frac{\hat{p}_{++}(G(\vec{C}_1))}{\hat{p}_{-+}(G(\vec{C}_1))} + \frac{1}{|R_{++}|} \int_{R_{++}} \frac{K(G(x) - G(\vec{C}_1))}{\hat{p}_{++}(G(x))} dx \right. \right. \\ &\quad \left. \left. - \frac{1}{|R_{-+}|} \int_{R_{-+}} \frac{K(G(x) - G(\vec{C}_1))}{\hat{p}_{-+}(G(x))} dx \right) + (1 - H(\phi_2(\vec{C}_1))) \left( \log \frac{\hat{p}_{+-}(G(\vec{C}_1))}{\hat{p}_{--}(G(\vec{C}_1))} \right. \right. \\ &\quad \left. \left. + \frac{1}{|R_{+-}|} \int_{R_{+-}} \frac{K(G(x) - G(\vec{C}_1))}{\hat{p}_{+-}(G(x))} dx - \frac{1}{|R_{--}|} \int_{R_{--}} \frac{K(G(x) - G(\vec{C}_1))}{\hat{p}_{--}(G(x))} dx \right) \right] \end{aligned} \quad (20)$$

$$\begin{aligned} \frac{\partial \vec{C}_2}{\partial t} &= \vec{N}_2 \left[ -\alpha \kappa_2 + H(\phi_1(\vec{C}_2)) \left( \log \frac{\hat{p}_{++}(G(\vec{C}_2))}{\hat{p}_{+-}(G(\vec{C}_2))} + \frac{1}{|R_{++}|} \int_{R_{++}} \frac{K(G(x) - G(\vec{C}_2))}{\hat{p}_{++}(G(x))} dx \right. \right. \\ &\quad \left. \left. - \frac{1}{|R_{+-}|} \int_{R_{+-}} \frac{K(G(x) - G(\vec{C}_2))}{\hat{p}_{+-}(G(x))} dx \right) + (1 - H(\phi_1(\vec{C}_2))) \left( \log \frac{\hat{p}_{-+}(G(\vec{C}_2))}{\hat{p}_{--}(G(\vec{C}_2))} \right. \right. \\ &\quad \left. \left. + \frac{1}{|R_{-+}|} \int_{R_{-+}} \frac{K(G(x) - G(\vec{C}_2))}{\hat{p}_{-+}(G(x))} dx - \frac{1}{|R_{--}|} \int_{R_{--}} \frac{K(G(x) - G(\vec{C}_2))}{\hat{p}_{--}(G(x))} dx \right) \right] \end{aligned} \quad (21)$$

where  $H(\cdot)$  is the Heaviside function ( $H(\phi)=1$  if  $\phi \geq 0$  and  $H(\phi) = 0$  if  $\phi < 0$ ).

Equations (20), (21) involve log-likelihood ratio tests comparing the hypotheses that the observed image intensity  $G(\vec{C}_i)$  at a given point on the active contour  $\vec{C}_i$  belongs to one region or the other.

As illustrated in Figure 2(b),  $\vec{C}_1$  delineates either the boundary between  $R_{++}$  and  $R_{-+}$ , or the boundary between  $R_{+-}$  and  $R_{--}$ , when  $\vec{C}_1$  lies inside or outside curve  $\vec{C}_2$ , respectively.

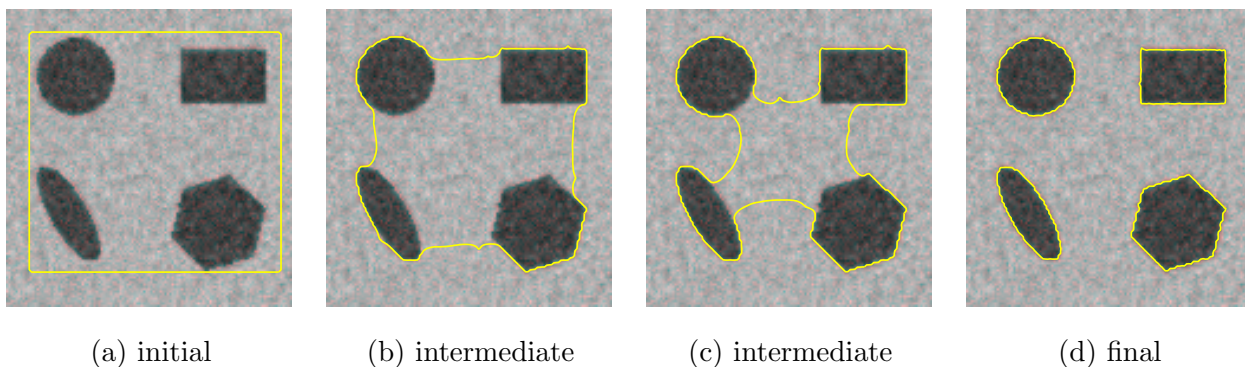


Fig. 3. Evolution of the curve on a synthetic image; the different mean case.

Equation (20) exactly reflects this situation and reveals the region competition between regions adjacent to curve  $\vec{C}_1$ . Similarly, Equation (21) expresses the region competition between regions adjacent to curve  $\vec{C}_2$ .

## V. EXPERIMENTAL RESULTS

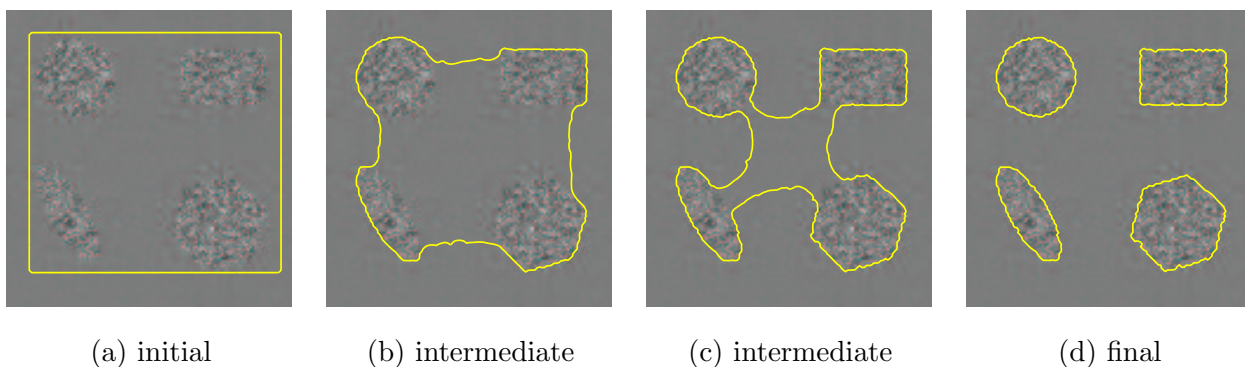


Fig. 4. Evolution of the curve on a synthetic image; the different variance case.

We present experimental results on synthetic images of geometric objects, and a number of real images. In all the examples, the regularization parameter  $\alpha$  in (9) or (17) is chosen subjectively based upon our qualitative assessment of the segmented imagery. In cases where prior information is available about the objects in the scene, it may be possible to learn an appropriate distribution of regularizers based upon the known smoothness characteristics of the object boundaries coupled with the signal-to-noise ratios of the images to be segmented.

We use synthetic images generated by several sets of distributions. Figure 3 shows the result produced by our technique for the case where the two distributions for the foreground and the background are Gaussian with different means and the same variance. Figure 4 shows the result

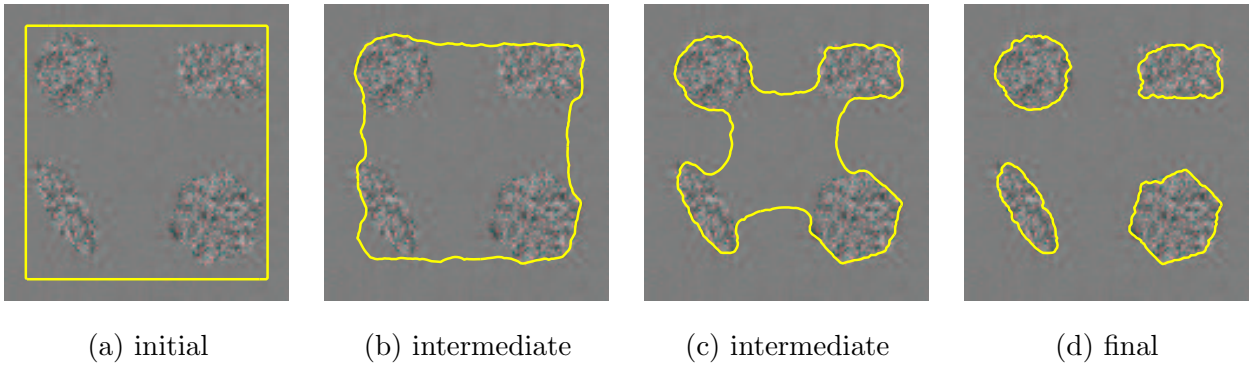


Fig. 5. Evolution of the curve on a synthetic image without the additional two terms; the different variance case.

for the case where the two distributions for the foreground and the background are Gaussian with different variances and the same mean. For these two cases, the method of Yezzi et al. [44] would require the selection of the appropriate statistic (i.e. mean and variance for the first and second cases respectively) *a priori*, whereas our method solves the segmentation problem without that information.

For the result in Figure 3, we measured the run time for both our nonparametric method and parametric counterpart in [44]. On an Intel Xeon 2.2 GHz cpu, the nonparametric method took 167 seconds (image size is 126 by 121), whereas the parametric method took 26 seconds. The parametric method is of less computational cost when the parametric method is well-matched to the problem here. However, if there is a mismatch between the image and the parametric model, there will be losses in terms of both the accuracy of the segmentation and the computational cost.

As we mentioned in Section III-D, the motion equation for the curve (15) contains three data-driven terms and a curvature term. We now provide an empirical analysis of the relative contribution of the first data-driven term (the log-likelihood ratio) versus the other two data-driven terms, to the overall curve evolution. To this end, we consider the example in Figure 3. We compute the numerical values of the log-likelihood ratio, the second term, and the third term of the gradient flow (15) at each point on the curve, for multiple snapshots during the iterative curve evolution process. In order to analyze the general behavior of these terms, we combine all those data obtained throughout the curve evolution process and show their histograms in Figure 6. Figure 6(a) and Figure 6(b) show histograms of the values taken by the second term



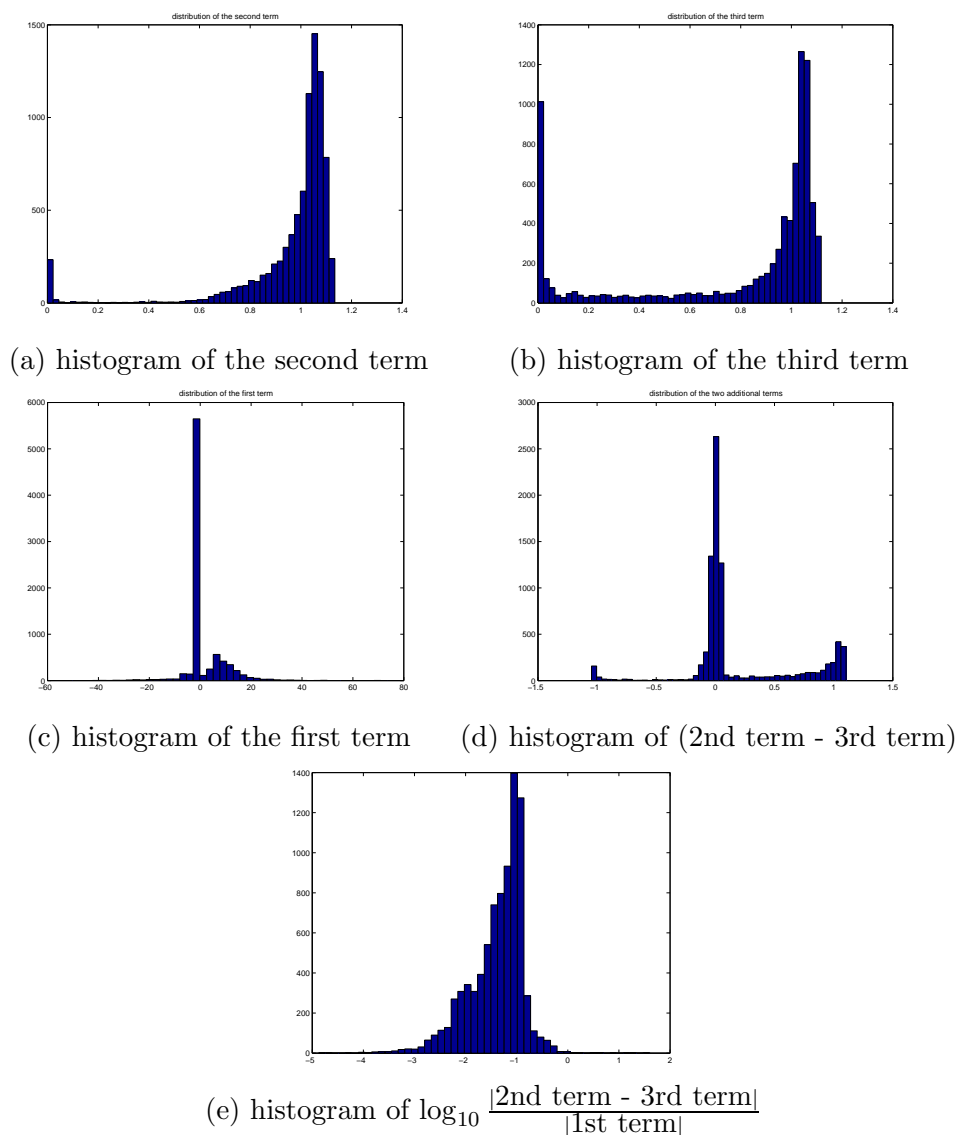


Fig. 6. Histograms of the three terms of the gradient flow for the points on the boundaries of Figure 3.

and the third term respectively. We observe that the values of both terms are often close to 1, and lie in a limited range (mostly between 0 and 1.5). We analyze this observation in more detail in Appendix-E. Figure 6(c) and Figure 6(d) show histograms of the values taken by the first term and the other two terms (i.e. the second term minus the third term). Since both the second and the third term have a limited range, their difference (which is their overall contribution to the evolution) is also in a limited range (mostly between -1.5 and 1.5), as is shown in Figure 6(d). Finally, Figure 6(e) shows a histogram of  $\log_{10} \frac{|2\text{nd term} - 3\text{rd term}|}{|1\text{st term}|}$ . We can observe that mostly the first term has a larger magnitude than the other two terms, hence it

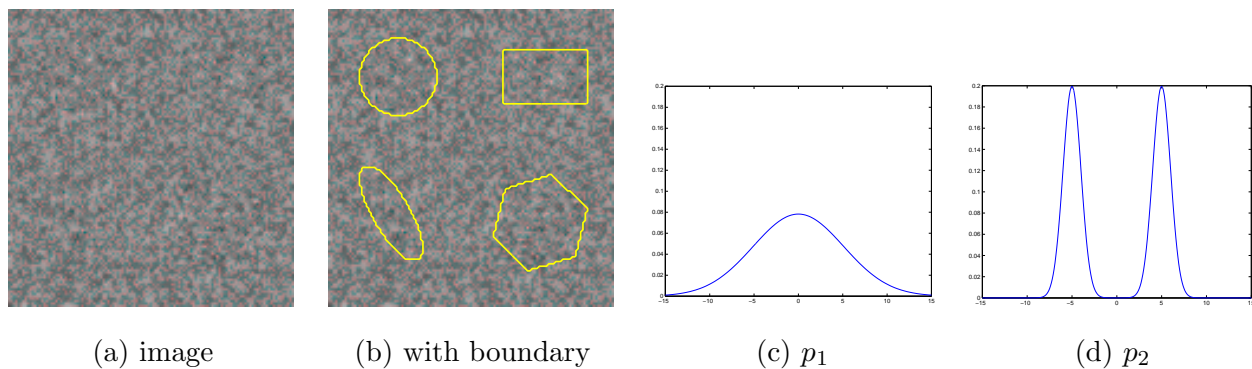


Fig. 7. Example image with two regions (boundaries marked in (b)), where the foreground has a unimodal density  $p_1$ , and the background has a bimodal density  $p_2$ . The two densities  $p_1$  and  $p_2$  have the same mean and the same variance.

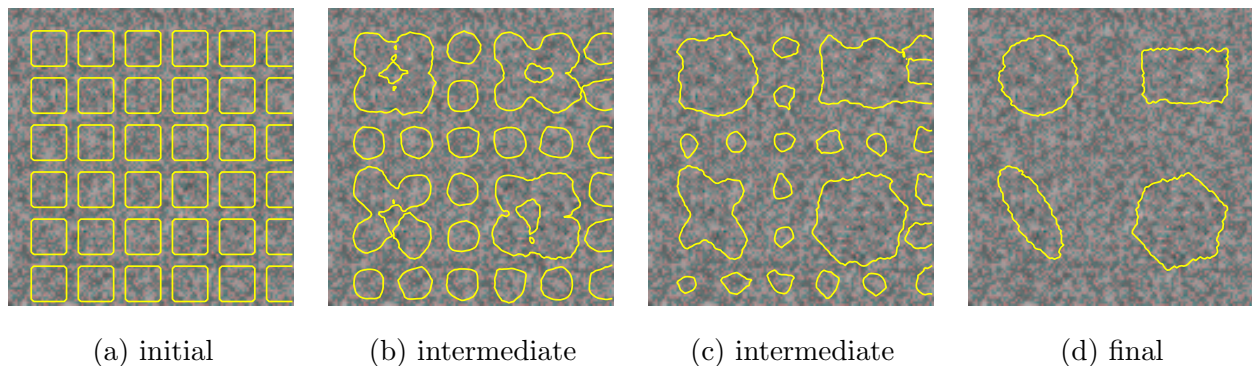
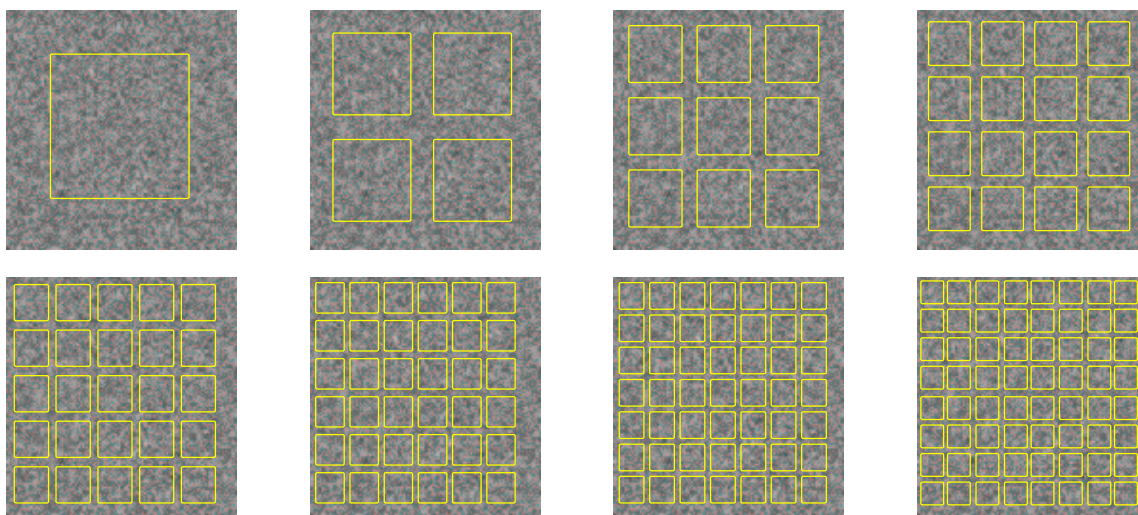


Fig. 8. Evolution of the curve on a synthetic image; unimodal versus bimodal densities.

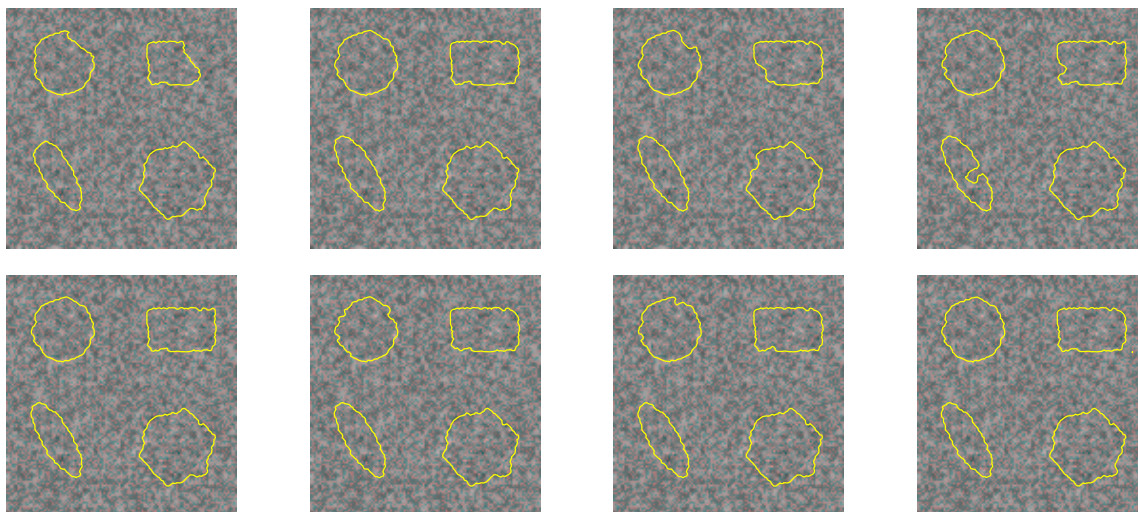
is the dominant contributor to the curve evolution. Consequently, for the experiment in Figure 3, we obtain a similar segmentation results without including the two additional terms.

However, for other types of images, the log-likelihood ratio can be small, and the other two terms can become more important affecting the performance of the segmentation. For instance, if we do not include the additional two terms for the segmentation of the image in Figure 4(a), we observe a loss in the accuracy of the segmentation as illustrated in Figure 5. We observe that the sharp corners of the rectangle are missed. A similar performance loss due to excluding these additional terms is also pointed out by Jehan-Besson [21]. Based on these empirical observations, we believe this is an issue that requires further analysis in future works.

The next synthetic example we consider involves a more challenging image shown in Figure 7(a). The underlying distributions of the foreground and the background are a unimodal Gaussian density and a bimodal density with two Gaussian components as illustrated in Figure 7(c) and



(a) initial curves



(b) corresponding segmentation results

Fig. 9. Segmentations of the image in Figure 7(a) with various initializations. (a) Eight different initializations with varying number of seeds. (b) Corresponding segmentation results.

Figure 7(d) respectively. The two distributions have the same mean and same variance, so it is hard even for a human observer to separate the foreground from the background. In order to let the readers see the foreground, we show the actual boundaries by a curve in Figure 7(b). For this kind of image, the methods based on means and variances such as that proposed by Yezzi *et al.* [44] would no longer work. Figure 8 shows our segmentation results. As shown in Figure 8(a), we have used an automatic initialization with multiple seeds. The power of the multiple-seed initialization is that it observes entire regions, and the evolution of the curve occurs globally.

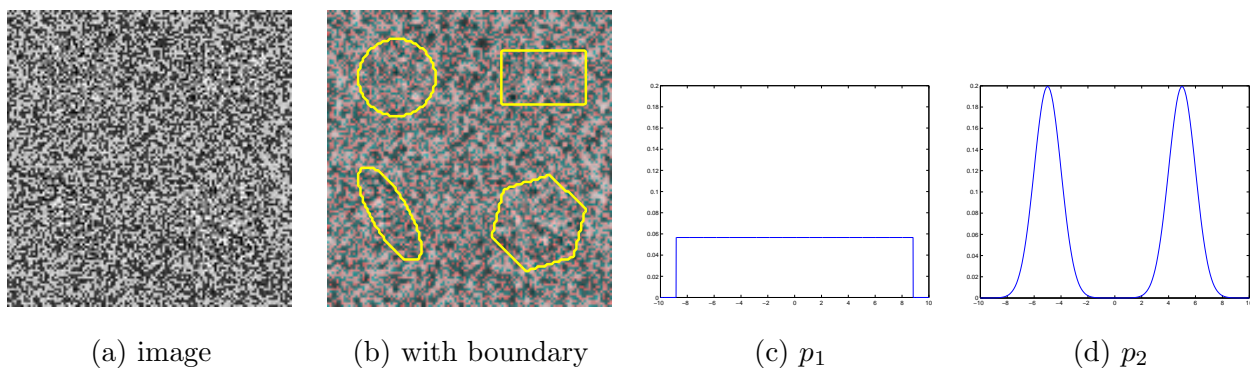


Fig. 10. Example image with two regions (boundaries marked in (b)), where the foreground has a uniform density  $p_1$ , and the background has a bimodal density  $p_2$ . The two densities  $p_1$  and  $p_2$  have the same mean and the same variance.

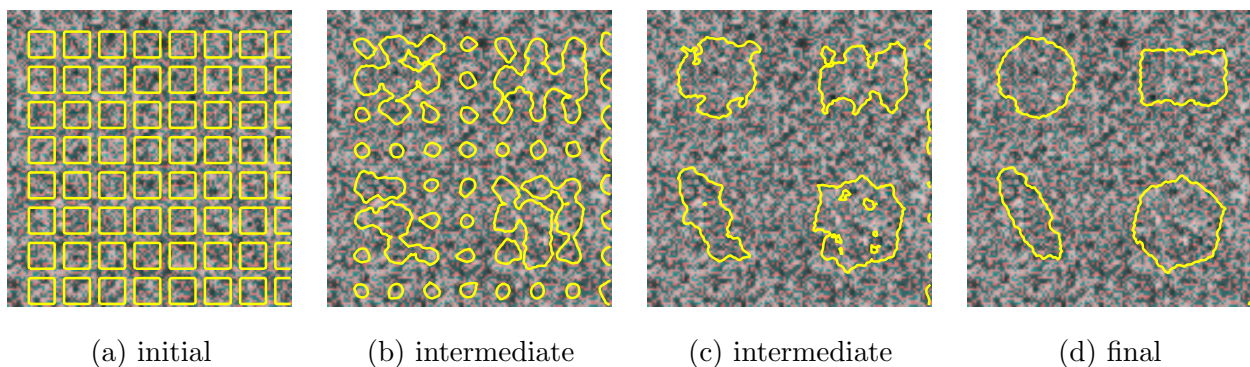


Fig. 11. Evolution of the curve on a synthetic image; uniform (foreground) versus bimodal (background) densities.

Figure 8(b) and Figure 8(c) show the intermediate stages of the evolution, where the seeds in the background region gradually shrink at each iteration, whereas those in the foreground region grow. The final result shown in Figure 8(d) appears to be an accurate segmentation. Similarly, the next synthetic example in Figure 10 involves two distributions with the same mean and the same variance, where the foreground distribution is uniform and the background one is bimodal with two Gaussian components. As shown in Figure 11, our method can detect the foreground objects without any prior knowledge about the probability densities involved.

We empirically analyze the sensitivity of our segmentation results to initialization. In Figure 9, we run our algorithm on the same image as the one generated from unimodal and bimodal densities in Figure 7 with different initializations. Figure 9(a) shows various initializations with different number of seeds, and Figure 9(b) shows the corresponding segmentation results. As the

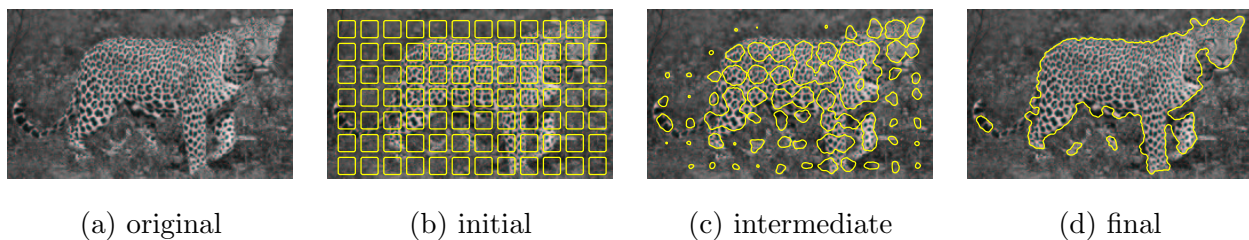


Fig. 12. Evolution of the curve on a leopard image.

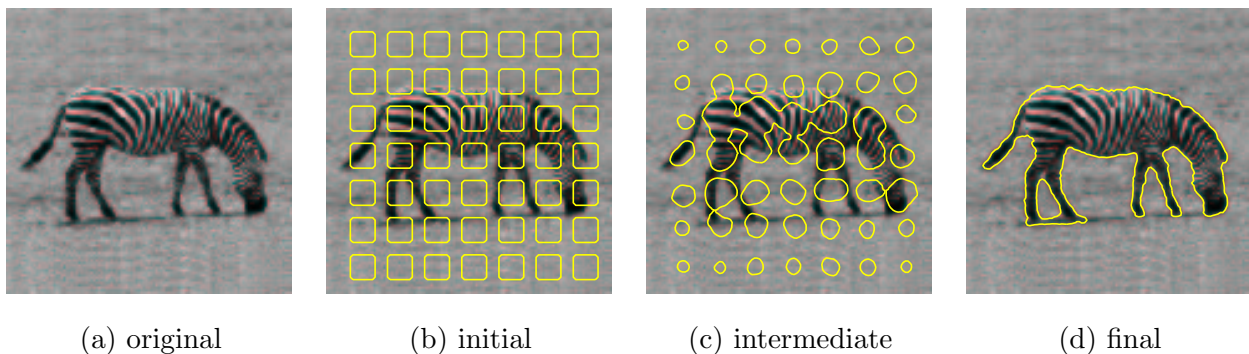


Fig. 13. Evolution of the curve on a zebra image. (Input image: courtesy of Nikos Paragios)

upper row of Figure 9(b) shows, the segmentation can be suboptimal if we have a small number of seeds indicating that the segmentations depend on the initializations. However, the lower row of Figure 9(b) shows that as long as the number of seeds is large enough, the segmentation result is stable with respect to initializations even for this challenging example. It will be a worthwhile future work to analyze the dependence of the curve evolution on the initializations. At this point we can give a rule of thumb for initializations with multiple seeds that the seeds need to cover the entire region such that they intersect with both the foreground and the background with high probability and that the number of seeds need to be large enough in order to avoid local minima.

Let us now consider the challenging examples in Figure 8 and Figure 11. If we did not have access to the underlying truth (as shown in Figure 7 and Figure 10), then based on the data and the results in Figure 8 and Figure 11, one might naturally ask the question of whether there are really two regions (i.e. foreground and background) here as the segmentations suggest, or whether there is only a single region. This raises the issue of statistical significance of a given result. We can address this issue by considering the null hypothesis  $H_0$  that there is only one region versus the alternative hypothesis  $H_1$  that there are two regions. We present the details of

this analysis in Appendix-B.2, where we observe that the key quantity involved here is again the mutual information. Specifically, the log-likelihood ratio  $\log \frac{p(\{G(x)|x \in \Omega\}|H_1)}{p(\{G(x)|x \in \Omega\}|H_0)}$  is given by the size of data times the mutual information estimate, i.e.  $|\Omega| \hat{I}(G(X); L_{\bar{C}}(X))$ , which leads to the following interpretations: First, the higher the mutual information, the more different the density estimates  $\hat{p}_+$ ,  $\hat{p}_-$  are, and thus the more confidence we have. Second, the larger the size of data, the more accurate those density estimates are. Based on these observations, we take  $\hat{I}(G(X); L_{\bar{C}}(X))$  as a statistic, and generate samples of this statistic under the null hypothesis  $H_0$  that there is a single region. The procedure for generating these samples is described in Appendix-B.3. Next we compute the sample mean  $E[\hat{I}|H_0]$  and the sample variance  $Var[\hat{I}|H_0]$  of  $\hat{I}(G(X); L_{\bar{C}}(X))$  under  $H_0$ . Finally, we evaluate whether the mutual information estimate  $\hat{I}_{seg}(G(X); L_{\bar{C}}(X))$  produced by our segmentation result is a likely outcome under the null hypothesis. For this evaluation, we simply use the Z-value,  $Z \triangleq \frac{\hat{I}_{seg} - E[\hat{I}|H_0]}{\sqrt{Var[\hat{I}|H_0]}}$ , which measures the distance between the observed value  $\hat{I}_{seg}$  and the mean under  $H_0$ , in terms of the number of standard deviations. Large values indicate that the result is significant, hence the null hypothesis can be rejected. For the result shown in Figure 8(d) and Figure 11(d) the Z-values are 4.24 and 5.63, respectively. These values are unlikely to occur under the null hypothesis, which thereby indicates that the segmentation results we have are statistically significant.

We now report the result for a leopard image and a zebra image shown in Figure 12 and Figure 13 respectively. Both of these are challenging segmentation problems, where methods based on single statistics may fail. Figure 12(d) shows the segmentation result for the leopard image. The final curve captures the main body of the leopard and some parts of its tail and legs. The parts of the tail and the legs that are missing look similar to the background, which makes a perfect segmentation difficult. Figure 13 shows the success of our method in segmenting the zebra image, which is the identical zebra image used in Paragios *et al.* [31]. Their supervised texture segmentation algorithm requires an image patch taken from the object and an image patch taken from the background in advance as an input to the algorithm. In contrast, the merit of our method is that we do not have to know or choose which feature to use and that the method non-parametrically estimates probability density functions and use that as a statistical feature. It is noticeable that our method, which is unsupervised, can segment this complex image as accurately as their supervised algorithm. Regarding the computational costs, on an Intel Xeon 2.2 GHz cpu, the nonparametric method took 211 seconds for segmenting the zebra image, whose

size is 115 by 115.

Although our method can segment textured images without a prior training, there are some classes of images where our framework breaks down. For instance, if one region has a texture with a marginal distribution  $p_1$ , and the other region has a different texture with the same marginal distribution  $p_1$ , then such an image can not be segmented without using a preprocessing such as one based on filter banks.

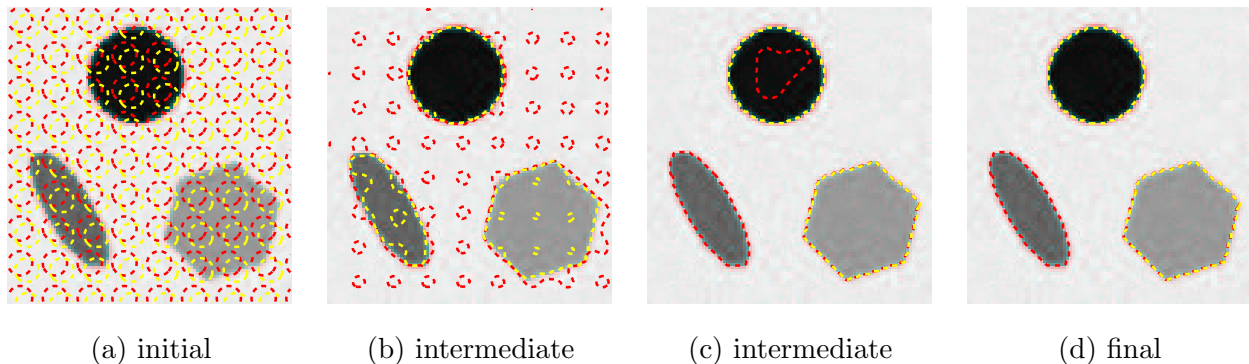


Fig. 14. Evolution of the curve on a synthetic image; three regions with different mean intensities.

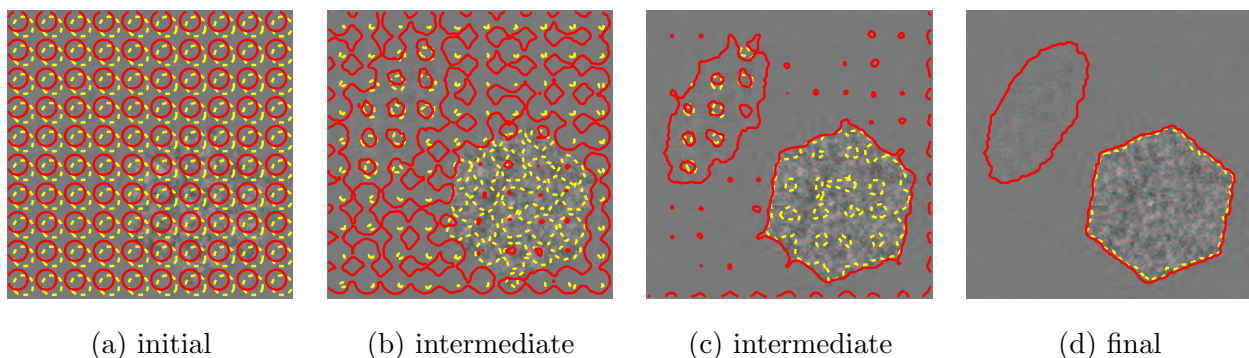


Fig. 15. Evolution of the curve on a synthetic image; three regions with different mean intensities.

Now we present the results of our information-theoretic, multi-phase segmentation method on synthetic images of geometric objects, as well as real images. The image shown in Figure 14(a) contains four regions (circle, ellipse, hexagon, and background) with Gaussian distributions with different means. Hence, in this case we have  $m = 2$ ,  $n = 4$ . The initial, intermediate, and final stages of our curve evolution algorithm are shown in Figure 14, where the four regions  $R_{++}, R_{+-}, R_{-+}, R_{--}$  determined by the two curves capture the circle, the background, the hexagon, and the ellipse, respectively. Note that, methods such as that of [44] would also work for this simple example, but would require the selection of an appropriate statistic (in this case the

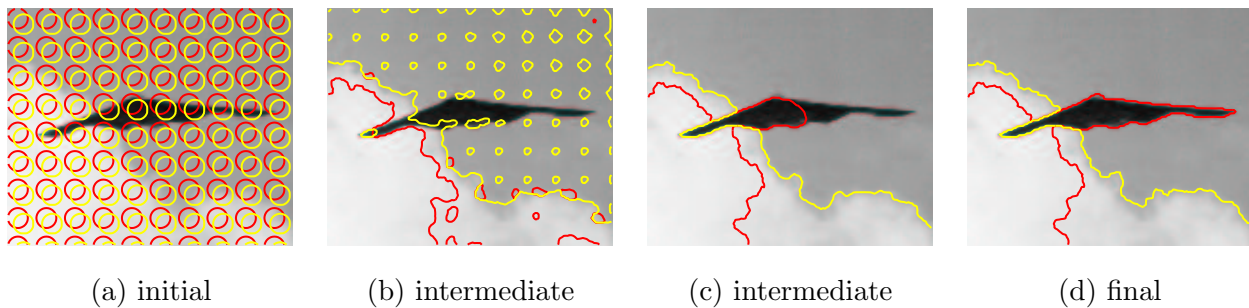


Fig. 16. Evolution of the curve on an aircraft image.

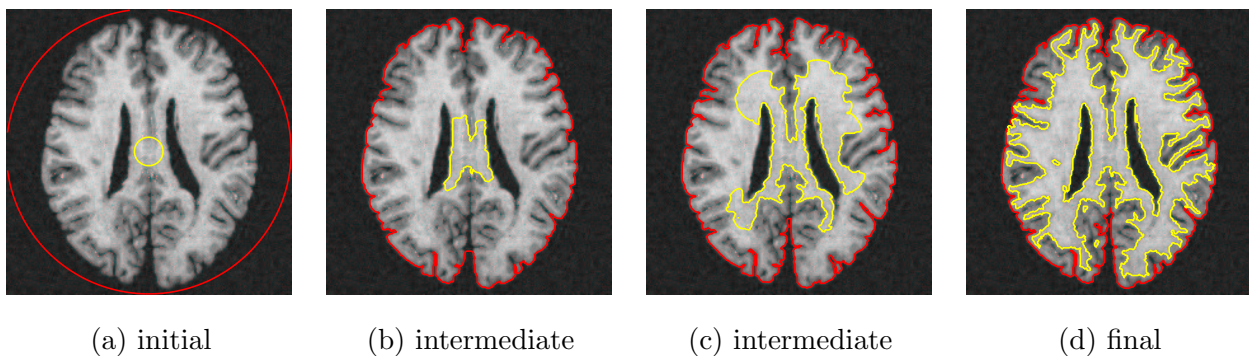


Fig. 17. Evolution of the curve on a brain image.

mean) *a priori*, whereas our method does not. The Mumford Shah-based multi-phase technique of [10], would also work in this case. Figure 15(a) contains an example with three regions having Gaussian distributions with different variances, hence  $m = 2, n = 3$ . In this case,  $R_{+-}$ ,  $R_{-+}$ , and  $R_{--}$  capture the background, the hexagon, and the ellipse, respectively, whereas  $R_{++}$  shrinks and disappears.

Figure 16(a) shows an image of an airplane. The two curves in the final segmentation in Figure 16(d) capture the four regions, the airplane, the sky, the white clouds, and the darker clouds.

Figure 17(a) shows a brain pathology image, which has three regions, the background, the white matter, and the gray matter. This kind of brain images involve not only complex boundaries but also a topological constraint that the white matter is inside the gray matter. The proposed multiphase segmentation method can handle this topology and the three regions capture the white matter, the gray matter, and the background.



## VI. CONCLUSION

We have developed a new information-theoretic image segmentation method based on non-parametric statistics and curve evolution. We have formulated the segmentation problem as one of maximizing the mutual information between the region labels and the pixel intensities, subject to curve length constraints. We have derived the curve evolution equations for the optimization problem posed in our framework. Due to the nonparametric aspect of our formulation, the proposed technique can automatically deal with a variety of segmentation problems, in which many currently available curve evolution-based techniques would either completely fail or at least require the *a priori* extraction of representative statistics for each region. We use fast techniques for the implementation of nonparametric estimation, which keep the computational complexity at a reasonable level. Our experimental results have shown the strength of the proposed technique in accurately segmenting real and synthetic images.

## APPENDIX

## A. Proof of the Fact about Mutual Information

In this appendix, we prove a statement from Section II, namely that the mutual information  $I(G(X); L_{\vec{C}}(X))$  is maximized if and only if  $\vec{C}$  is the correct segmentation, i.e. if  $R_+ = R_1$ ,  $R_- = R_2$  (or equivalently  $R_+ = R_2$ ,  $R_- = R_1$ ). We remind the readers that this analysis makes use of the knowledge of  $R_1, R_2, p_1, p_2$  so that we can compute the MI. Since  $I(G(X); X)$  is independent of the label  $L_{\vec{C}}(\cdot)$ , it is sufficient to show that

$$I(G(X); L_{\vec{C}}(X)) \leq I(G(X); X) \quad (22)$$

and that equality holds if and only if  $R_+ = R_1$ ,  $R_- = R_2$  (or equivalently  $R_+ = R_2$ ,  $R_- = R_1$ ).

*Proof:* The inequality is basically the data processing inequality [12]. We will follow the proof in [12].

By using the chain rule, we can expand the mutual information between  $G(X)$  and  $\{X, L_{\vec{C}}(X)\}$ , namely  $I(G(X); X, L_{\vec{C}}(X))$  in the following two different ways:

$$I(G(X); X, L_{\vec{C}}(X)) = I(G(X); L_{\vec{C}}(X)) + I(G(X); X | L_{\vec{C}}(X)) \quad (23)$$

$$= I(G(X); X) + I(G(X); L_{\vec{C}}(X) | X) \quad (24)$$

Note that given  $X = x$ ,  $L_{\bar{C}}(X)$  is just a constant  $L_{\bar{C}}(x)$ . Thus  $G(X)$  and  $L_{\bar{C}}(X)$  are conditionally independent given  $X$ , and we have  $I(G(X); L_{\bar{C}}(X)|X) = 0$ . Since  $I(G(X); X|L_{\bar{C}}(X)) \geq 0$ , we have

$$I(G(X); X) \geq I(G(X); L_{\bar{C}}(X)). \quad (25)$$

The equality holds if and only if  $I(G(X); X|L_{\bar{C}}(X)) = 0$ , i.e.  $G(X)$  and  $X$  are conditionally independent given  $L_{\bar{C}}(X)$ . Now it suffices to show that  $p_{G(X)|L_{\bar{C}}(X)} = p_{G(X)|X, L_{\bar{C}}(X)}$  if and only if  $R_+ = R_1$ ,  $R_- = R_2$  (or equivalently  $R_+ = R_2$ ,  $R_- = R_1$ ). The remainder of the proof is based on the fact that  $p_{G(X)|L_{\bar{C}}(X)}$  is not homogeneous, (i.e. it is a mixture of  $p_1$  and  $p_2$ ) unless  $L_{\bar{C}}(\cdot)$  gives a correct segmentation, whereas  $p_{G(X)|X, L_{\bar{C}}(X)}$  is always homogeneous.

Note that the conditional densities  $p_{G(X)|L_{\bar{C}}(X)=L_+}$  and  $p_{G(X)|L_{\bar{C}}(X)=L_-}$  are mixtures of  $p_1$  and  $p_2$  as given in (6) and (7):

$$p_{G(X)|L_{\bar{C}}(X)=L_+} = \frac{|R_+ \cap R_1|}{|R_+|} p_1 + \frac{|R_+ \cap R_2|}{|R_+|} p_2 \quad (26)$$

$$p_{G(X)|L_{\bar{C}}(X)=L_-} = \frac{|R_- \cap R_1|}{|R_-|} p_1 + \frac{|R_- \cap R_2|}{|R_-|} p_2 \quad (27)$$

On the other hand, the conditional density  $p_{G(X)|X=x, L_{\bar{C}}(X)=L_{\bar{C}}(x)}$  is  $p_1$  if  $x \in R_1$  and  $p_2$  if  $x \in R_2$ .

Suppose that  $R_+ = R_1$ ,  $R_- = R_2$ . Then (26) and (27) give us that  $p_{G(X)|L_{\bar{C}}(X)=L_+} = p_1$  and  $p_{G(X)|L_{\bar{C}}(X)=L_-} = p_2$ . Similarly, if  $R_+ = R_2$ ,  $R_- = R_1$ , then  $p_{G(X)|L_{\bar{C}}(X)=L_+} = p_2$  and  $p_{G(X)|L_{\bar{C}}(X)=L_-} = p_1$ . In either case, we have  $p_{G(X)|L_{\bar{C}}(X)} = p_{G(X)|X, L_{\bar{C}}(X)}$ .

However, unless  $R_+ = R_1$ ,  $R_- = R_2$  (or equivalently  $R_+ = R_2$ ,  $R_- = R_1$ ), at least one of  $p_{G(X)|L_{\bar{C}}(X)=L_+}$  and  $p_{G(X)|L_{\bar{C}}(X)=L_-}$  is a mixture of  $p_1$  and  $p_2$ , thus  $p_{G(X)|L_{\bar{C}}(X)} \neq p_{G(X)|X, L_{\bar{C}}(X)}$ .

Therefore,  $p_{G(X)|L_{\bar{C}}(X)} = p_{G(X)|X, L_{\bar{C}}(X)}$  if and only if  $R_+ = R_1$ ,  $R_- = R_2$  (or equivalently  $R_+ = R_2$ ,  $R_- = R_1$ ), and this completes the proof. ■

Remark: The inequality (22) is also true for the case where  $L_{\mathbf{C}}(\cdot)$  is an  $n$ -ary label, and the equality holds if and only if  $p_{G(X)|L_{\mathbf{C}}(X)} = p_{G(X)|X, L_{\mathbf{C}}(X)}$ . Consequently, the equality holds if the label  $L_{\mathbf{C}}(\cdot)$  gives a correct segmentation. Now we prove that the equality does not hold if the label gives an incorrect segmentation. Since  $p_{G(X)|X, L_{\mathbf{C}}(X)}$  is always homogeneous, the equality holds only if  $p_{G(X)|L_{\mathbf{C}}(X)}$  is homogeneous. However, if the segmentation is incorrect,  $p_{G(X)|L_{\mathbf{C}}(X)=L_{s(i)}}$  is a mixture for at least one  $L_{s(i)}$  thus  $p_{G(X)|L_{\mathbf{C}}(X)} \neq p_{G(X)|X, L_{\mathbf{C}}(X)}$ . This proves the same fact for the  $n$ -ary label case.

## B. Statistical Interpretation and Analysis

### B.1 MAP Estimation Interpretation of the Energy Functional

The curve that minimizes the energy functional is given by

$$\arg \min_{\vec{C}} E(\vec{C}) = \arg \min_{\vec{C}} |\Omega| \hat{h}(G(X)|L_{\vec{C}}(X)) + \alpha \oint_{\vec{C}} ds. \quad (28)$$

Now the conditional entropy term corresponds to the negative logarithm of the likelihood as follows:

$$\begin{aligned} |\Omega| \hat{h}(G(X)|L_{\vec{C}}(X)) &= |\Omega| Pr(L_{\vec{C}}(X) = L_+) \hat{h}(G(X)|L_{\vec{C}}(X) = L_+) \\ &\quad + |\Omega| Pr(L_{\vec{C}}(X) = L_-) \hat{h}(G(X)|L_{\vec{C}}(X) = L_-) \\ &= -|\Omega| \frac{|R_+|}{|\Omega|} \frac{1}{|R_+|} \int_{R_+} \log \hat{p}_+(G(x)) dx - |\Omega| \frac{|R_-|}{|\Omega|} \frac{1}{|R_-|} \int_{R_-} \log \hat{p}_-(G(x)) dx \\ &= - \int_{\Omega} \log \hat{p}_{G(X)|L_{\vec{C}}(X)=L_{\vec{C}}(x)}(G(x)|L_{\vec{C}}(X) = L_{\vec{C}}(x)) dx, \end{aligned} \quad (29)$$

where the last expression is the negative log-likelihood of the data  $\{G(x)|x \in \Omega\}$  in terms of the estimated density. On the other hand, the curve length term can be interpreted as the negative logarithm of prior probability for the curve,  $\oint_{\vec{C}} ds = -\log p(\vec{C})$ . Therefore, minimizing the energy functional corresponds to finding the maximum a posteriori estimate of the label.

### B.2 MI as a Confidence Measure

We express the question of whether the image has only a single region or two regions as the following hypothesis testing problem:

$$H_0 : p_1(x) = p_2(x) \quad (\text{single region}) \quad (30)$$

$$H_1 : p_1(x) \neq p_2(x) \quad (\text{two regions}) \quad (31)$$

Under the null hypothesis  $H_0$ , the data  $\{G(x)|x \in \Omega\}$  have a single unknown density  $p_1 = p_2$ , and in this case  $p_{G(X)} = p_1 = p_2$ , whose estimate is  $\hat{p}_{G(X)}$ . Thus the log-likelihood is given by

$$\log p(\{G(x)|x \in \Omega\}|H_0) = \int_{\Omega} \log \hat{p}_{G(X)}(G(x)) dx \quad (32)$$

$$= -|\Omega| \hat{h}(G(X)) \quad (33)$$

Under the alternative hypothesis, the data have two unknown densities  $p_1$  and  $p_2$ , and their estimates are  $\hat{p}_+$  and  $\hat{p}_-$ . Thus (29) gives the negative of the log-likelihood of the data under  $H_1$ .

Therefore, we have the log-likelihood ratio in terms of the data size and the mutual information estimate as follows:

$$\log \frac{p(\{G(x)|x \in \Omega\}|H_1)}{p(\{G(x)|x \in \Omega\}|H_0)} = -|\Omega|\hat{h}(G(X)|L_{\vec{C}}(X)) + |\Omega|\hat{h}(G(X)) \quad (34)$$

$$= |\Omega|\hat{I}(G(X); L_{\vec{C}}(X)) \quad (35)$$

This gives a quantitative measure of the belief that  $H_1$  is true.

### B.3 Computing the Z-value

To evaluate the significance of a segmentation result (indicating the existence of two regions in the image), we need to generate samples of the statistic under the null hypothesis that there is a single region. We obtain such samples through random permutations of the binary label. More formally, we define the permutation of the binary labels  $L_{\pi, \vec{C}}(\cdot)$  induced by a permutation of the pixels  $\pi : \Omega \rightarrow \Omega$  as follows:

$$L_{\pi, \vec{C}}(x) \triangleq L_{\vec{C}}(\pi(x)).$$

In a similar way to [15], we perform the following procedure:

- Repeat  $M$  times (with index  $m = 1$  to  $M$ ):
  - sample a random permutation  $\pi_m$  from a uniform distribution over the set of all permutations,
  - compute the MI statistic  $I_m = \hat{I}(G(X); L_{\pi_m, \vec{C}}(X))$
- compute sample mean and sample variance of  $\{I_1, \dots, I_M\}$ .

These sample mean and sample variance are used as estimates of  $E[\hat{I}|H_0]$  and  $Var[\hat{I}|H_0]$ .

### C. Gradient Flows for “Nested” Region Integrals

In Section III-B, we stated the gradient flow for a general nested region integral. In this section, we provide a derivation of the gradient flow (via the first variation) of a curve  $\vec{C}(t)$  for minimizing an energy integral  $E$ , which is a region integral over the curve’s interior  $R(\vec{C}(t))$  (we have used a shorthand notation  $R$  for  $R(\vec{C}(t))$  in the main body of the paper). In our derivation, we use the results of Delfour and Zolesio [13]. Alternative derivations for this type of region integrals can be found in [3], [21].

For a simple region integral in the form of

$$E(\vec{C}(t)) = \int_{R(\vec{C}(t))} f(x) dx, \quad (36)$$

where the integrand does not depend on the curve or time  $t$  (dependence on the curve  $\vec{C}(t)$  implies dependence on time  $t$ ), we have the following expression for the time derivative:

$$\frac{dE}{dt} = \oint_{\vec{C}(t)} \langle \vec{C}_t, f\vec{N} \rangle ds \quad (37)$$

where we use subscript  $t$  to denote partial derivative as in  $\vec{C}_t = \frac{\partial \vec{C}}{\partial t}$ . From the above derivative, the form of the gradient flow for  $\vec{C}$  (the negative of the gradient so that the region integral decreases most rapidly) is revealed to be [45]

$$\frac{\partial \vec{C}}{\partial t} = -f\vec{N}. \quad (38)$$

We now consider a general class of region-based energy functionals  $E$  where the integrand  $f$  depends upon another family of region integrals  $\varepsilon(x, t)$  over  $R(\vec{C}(t))$ . Note that the “nested” region integrals  $\varepsilon(x, t)$  depend on  $t$ , since  $R(\vec{C}(t))$  (the interior of  $\vec{C}(t)$ ) changes as the curve evolves over time. More precisely, we assume as in (13)

$$E(\vec{C}(t)) = \int_{R(\vec{C}(t))} f(\varepsilon(x, t)) dx \quad \text{where} \quad \varepsilon(x, t) = \int_{R(\vec{C}(t))} g(x, \hat{x}) d\hat{x} \quad (39)$$

If we let  $\varphi(x, t) \triangleq f(\varepsilon(x, t))$ , our energy functional can be written as

$$E(\vec{C}(t)) = \int_{R(\vec{C}(t))} \varphi(x, t) dx \quad (40)$$

where  $\varphi$  depends on time  $t$ . Delfour and Zolesio [13] have considered region integrals in the form of (40), and Theorem 4.2 in [13](page 352) provides derivatives of such region integrals w.r.t. time  $t$ . Using their results, the derivative of (40) is given by:

$$\frac{\partial E(\vec{C}(t))}{\partial t} = \int_{R(\vec{C}(t))} \varphi_t(x, t) dx + \oint_{\vec{C}(t)} \langle \varphi(x, t)\vec{N}, \frac{\partial \vec{C}}{\partial t} \rangle ds \quad (41)$$

where  $\varphi_t(x, t) = \frac{\partial}{\partial t} f(\varepsilon(x, t)) = f'(\varepsilon(x, t))\varepsilon_t(x, t)$ . Since  $\varepsilon(x, t)$  in (39) does have the form of a simple region integral (36) for each  $x$ , whose integrand  $g$  does not depend on  $\vec{C}$ . As such, we can write  $\varepsilon_t$  as follows:

$$\varepsilon_t(x, t) = \oint_{\vec{C}(t)} \langle \vec{C}_t, g(x, \vec{C})\vec{N} \rangle ds \quad (42)$$

Plugging this into the above expression for  $\frac{dE}{dt}$  yields

$$\frac{dE}{dt} = \oint_{\vec{C}(t)} \langle \vec{C}_t, (f \circ \varepsilon) \vec{N} \rangle ds + \int_{R(\vec{C}(t))} f'(\varepsilon(x, t)) \left[ \oint_{\vec{C}(t)} \langle \vec{C}_t, g(x, \vec{C}) \vec{N} \rangle ds \right] dx \quad (43)$$

$$= \oint_{\vec{C}(t)} \langle \vec{C}_t, (f \circ \varepsilon) \vec{N} \rangle ds + \int_{R(\vec{C}(t))} \oint_{\vec{C}(t)} \langle \vec{C}_t, f'(\varepsilon(x, t)) g(x, \vec{C}) \vec{N} \rangle ds dx \quad (44)$$

$$= \oint_{\vec{C}(t)} \left\langle \vec{C}_t, \left[ f \circ \varepsilon + \int_{R(\vec{C}(t))} f'(\varepsilon(x, t)) g(x, \vec{C}) dx \right] \vec{N} \right\rangle ds \quad (45)$$

$$(46)$$

revealing the following gradient flow for  $\vec{C}$  (where  $t$  is omitted as an argument for simplicity):

$$\frac{\partial \vec{C}}{\partial t} = - \left[ f(\varepsilon(\vec{C})) + \int_{R(\vec{C}(t))} f'(\varepsilon(x)) g(x, \vec{C}) dx \right] \vec{N}, \quad (47)$$

which is the result we stated in (14).

#### D. Derivation of the Curve Evolution Formula

This section presents the derivation of the curve evolution formula (15) given in Section III-C. We begin by rewriting the energy functional (9) as follows:

$$E(\vec{C}) = -|\Omega| \hat{h}(G(X)) + E_+(\vec{C}) + E_-(\vec{C}) + \alpha \oint_{\vec{C}} ds, \quad (48)$$

where the components  $E_+(\vec{C})$  and  $E_-(\vec{C})$  are given by

$$E_+(\vec{C}) = |\Omega| Pr(L_{\vec{C}}(X) = L_+) \hat{h}(G(X) | L_{\vec{C}}(X) = L_+) \quad (49)$$

$$= - \int_{R_+} \log \left( \frac{1}{|R_+|} \int_{R_+} K(G(x) - G(\hat{x})) d\hat{x} \right) dx \quad (50)$$

$$E_-(\vec{C}) = |\Omega| Pr(L_{\vec{C}}(X) = L_-) \hat{h}(G(X) | L_{\vec{C}}(X) = L_-) \quad (51)$$

$$= - \int_{R_-} \log \left( \frac{1}{|R_-|} \int_{R_-} K(G(x) - G(\hat{x})) d\hat{x} \right) dx. \quad (52)$$

We now proceed with a calculation of the gradient flow for  $E_+$  noting that the flow for  $E_-$  will have a similar form (but with an opposite sign). Since  $\frac{1}{|R_+|}$  in (50) also depends on the curve,

we start by breaking  $E_+$  into two integrals:

$$E_+ = -(E_+^1 + E_+^2) \quad (53)$$

$$E_+^1 = - \int_{R_+} \log |R_+| dx = -|R_+| \log |R_+| \quad (54)$$

$$E_+^2 = \int_{R_+} \underbrace{\overbrace{\log}^{f(\cdot)} \left( \underbrace{\int_{R_+} \overbrace{K(G(x) - G(\hat{x}))}^{g(x, \hat{x})} d\hat{x}}_{\varepsilon(x, t)} \right)}_{f(\varepsilon(x, t))} dx, \quad (55)$$

where the second integral  $E_+^2$  exhibits the structure of the general nested form given in (13) (with the integrand  $f(\cdot)$ , the nested integral  $\varepsilon(\cdot)$ , and the nested integrand  $g(\cdot)$  labeled accordingly).

Using (14), the gradient flow for  $E_+^2$ , which we denote by  $\nabla_{\vec{c}} E_+^2$ , is given by

$$\begin{aligned} \nabla_{\vec{c}} E_+^2 &= - \left[ \overbrace{\log \left( \int_{R_+} K(G(\vec{C}) - G(\hat{x})) d\hat{x} \right)}^{f(\varepsilon(\vec{C}))} \right. \\ &\quad \left. + \int_{R_+} \frac{1}{\underbrace{\left( \int_{R_+} K(G(x) - G(\hat{x})) d\hat{x} \right)}_{f'(\varepsilon(x))}} \overbrace{K(G(x) - G(\vec{C}))}^{g(x, \vec{C})} dx \right] \vec{N} \quad (56) \end{aligned}$$

$$= - \left[ \log |R_+| + \log \hat{p}_+(G(\vec{C})) + \frac{1}{|R_+|} \int_{R_+} \frac{K(G(x) - G(\vec{C}))}{\hat{p}_+(G(x))} \right] \vec{N}, \quad (57)$$

while the gradient flow for  $E_+^1$  is given by

$$\nabla_{\vec{c}} E_+^1 = -(\nabla_{\vec{c}} |R_+|) \log |R_+| - \nabla_{\vec{c}} |R_+| = (1 + \log |R_+|) \vec{N}. \quad (58)$$

Adding these gradients yields

$$\nabla_{\vec{c}} E_+ = -(\nabla_{\vec{c}} E_+^1 + \nabla_{\vec{c}} E_+^2) = \left[ -1 + \log \hat{p}_+(G(\vec{C})) + \frac{1}{|R_+|} \int_{R_+} \frac{K(G(x) - G(\vec{C}))}{\hat{p}_+(G(x))} \right] \vec{N}. \quad (59)$$

The gradient for  $E_-$  has a similar structure (but with an opposite sign since the outward normal with respect to  $R_-$  is given by  $-\vec{N}$  rather than  $\vec{N}$ )

$$\nabla_{\vec{c}} E_- = - \left[ -1 + \log \hat{p}_-(G(\vec{C})) + \frac{1}{|R_-|} \int_{R_-} \frac{K(G(x) - G(\vec{C}))}{\hat{p}_-(G(x))} \right] \vec{N}. \quad (60)$$

Finally, the overall gradient flow for  $E(\vec{C})$  of (9) is obtained as follows:

$$\begin{aligned} \frac{\partial \vec{C}}{\partial t} = & \left[ \log \frac{\hat{p}_+(G(\vec{C}))}{\hat{p}_-(G(\vec{C}))} + \frac{1}{|R_+|} \int_{R_+} \frac{K(G(x) - G(\vec{C}))}{\hat{p}_+(G(x))} dx \right. \\ & \left. - \frac{1}{|R_-|} \int_{R_-} \frac{K(G(x) - G(\vec{C}))}{\hat{p}_-(G(x))} dx \right] \vec{N} - \alpha \kappa \vec{N}. \end{aligned} \quad (61)$$

### E. Approximations of the Second and Third Terms

In Section V, we have empirically observed that the second and third terms in the curve evolution expression in (15) have a limited range. Here we show that under certain assumptions, the values of these terms approach 1. In particular, provided that  $|R_+ \cap R_1| \gg 1$  and  $|R_+ \cap R_2| \gg 1$ , we have

$$\frac{1}{|R_+|} \int_{R_+} \frac{K(G(x) - G(\vec{C}))}{\hat{p}_+(G(x))} dx \approx 1. \quad (62)$$

Similarly, provided that  $|R_- \cap R_1| \gg 1$  and  $|R_- \cap R_2| \gg 1$ , we have

$$\frac{1}{|R_-|} \int_{R_-} \frac{K(G(x) - G(\vec{C}))}{\hat{p}_-(G(x))} dx \approx 1. \quad (63)$$

### Derivation

Let  $\lambda = \frac{|R_+ \cap R_1|}{|R_+|}$ , then  $\hat{p}_+ \approx \lambda p_1 + (1 - \lambda) p_2$ .

Now the approximation is as follows:

$$\begin{aligned} & \frac{1}{|R_+|} \int_{R_+} \frac{K(G(x) - G(\vec{C}))}{\hat{p}_+(G(x))} dx \\ &= \frac{|R_+ \cap R_1|}{|R_+|} \left[ \frac{1}{|R_+ \cap R_1|} \int_{R_+ \cap R_1} \frac{K(G(x) - G(\vec{C}))}{\hat{p}_+(G(x))} dx \right] \\ & \quad + \frac{|R_+ \cap R_2|}{|R_+|} \left[ \frac{1}{|R_+ \cap R_2|} \int_{R_+ \cap R_2} \frac{K(G(x) - G(\vec{C}))}{\hat{p}_+(G(x))} dx \right] \end{aligned} \quad (64)$$

$$\approx \lambda E_{p_1} \left[ \frac{K(Y - G(\vec{C}))}{\hat{p}_+(Y)} \right] + (1 - \lambda) E_{p_2} \left[ \frac{K(Y - G(\vec{C}))}{\hat{p}_+(Y)} \right] \quad (65)$$

$$\approx \lambda \int \frac{p_1(y) K(y - G(\vec{C}))}{\lambda p_1(y) + (1 - \lambda) p_2(y)} dy + (1 - \lambda) \int \frac{p_2(y) K(y - G(\vec{C}))}{\lambda p_1(y) + (1 - \lambda) p_2(y)} dy \quad (66)$$

$$= \int K(y - G(\vec{C})) dy \quad (67)$$

$$= 1 \quad (68)$$

The derivation of (63) is similar to that of (62).



## ACKNOWLEDGMENTS

We would like to acknowledge Ramani Duraiswami for referring us to fast Gauss Transform method for calculating Parzen density estimates.

## REFERENCES

- [1] D. Adalsteinsson and J. A. Sethian. A fast level set method for propagating interfaces. *J. Comput. Phys.*, 118:269–277, 1995.
- [2] R. Adams and L. Bischof. Seeded region growing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16(6):641–647, 1994.
- [3] G. Aubert, M. Barlaud, O. Faugeras, and S. Jehan-Besson. Image segmentation using active contours: Calculus of variation or shape optimization ? *SIAM Journal on Applied Mathematics*, 63(6):2128–2154, 2003.
- [4] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. van der Meulen. Nonparametric entropy estimation: An overview. *International Journal of Math. Stat. Sci.*, 6(1):17–39, 1997.
- [5] J. F. Canny. A computational approach to edge detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.
- [6] V. Caselles, F. Catte, T. Col, and F. Dibos. A geometric model for active contours in image processing. *Numerische Mathematik*, 66:1–31, 1993.
- [7] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *International Journal on Computer Vision*, 22(1):61–79, 1997.
- [8] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic snakes. *Int. J. Computer Vision*, 1998.
- [9] T. Chan and L. Vese. Active contours without edges. *IEEE Trans. on Image Processing*, 10(2):266–277, February 2001.
- [10] T. F. Chan and L. A. Vese. An efficient variational multiphase motion for the mumford-shah segmentation model. In *Proc. Asilomar Conf. on Signals, Systems, and Computers*, pages 490–494, 2000.
- [11] D. L. Chopp. Computing minimal surfaces via level set curvature flow. *J. Comput. Phys.*, 106:77–91, 1993.
- [12] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.
- [13] M. C. Delfour and J. Zolesio. *Shapes and Geometries: Analysis, Differential Calculus, and Optimization*. SIAM, 2001.
- [14] R. Goldenberg, R. Kimmel, E. Rivlin, and M. Rudzsky. Fast geodesic active contours. *IEEE Trans. on Image Processing*, 10(10):1467–1475, 2001.
- [15] P. Golland and B. Fischl. Permutation tests for classification: Towards statistical significance in image-based studies. In *Proc. Information Processing in Medical Imaging*, volume 2732 of *LNCS*, pages 330–341, 2003.
- [16] M. Grayson. The heat equation shrinks embedded plane curves to round points. *Journal of Differential Geometry*, 26:285–314, 1987.
- [17] L. Greengard and V. Rokhlin. A new version of the fast multipole method for the Laplace equation in three dimensions. *Acta Numerica*, 6:229–269, 1997.
- [18] L. Greengard and J. Strain. The fast Gauss transform. *SIAM J. Sci. Stat. Comput.*, 12(1):79–94, 1991.

- [19] M. Heiler and C. Schnörr. Natural image statistics for natural image segmentation. In *Proc. of the ICCV*, pages 1259–1266, 2003.
- [20] J. Huang and D. Mumford. Statistics of natural images and models. In *Proc. of the ICCV(1)*, pages 541–547, 1999.
- [21] S. Jehan-Besson and M. Marlaud. Dream<sup>2</sup>s: Deformable regions driven by an eulerian accurate minimization method for image and video segmentation. *International Journal of Computer Vision*, 53:45–70, 2003.
- [22] J. Kim, J. W. Fisher, M. Cetin, A. Yezzi, Jr., and A. S. Willsky. Incorporating complex statistical information in active contour-based image segmentation. In *Proc. IEEE Conf. on Image Processing*, volume 2, pages 655–658, 2003.
- [23] J. Kim, J. W. Fisher, A. Yezzi, Jr., M. Cetin, and A. S. Willsky. Nonparametric methods for image segmentation using information theory and curve evolution. In *Proc. IEEE Conf. on Image Processing*, volume 3, pages 797–800, 2002.
- [24] R. Kimmel and A. M. Bruckstein. Regularized laplacian zero crossings as optimal edge integrators. *International Journal on Computer Vision*, 53(3):225–243, 2003.
- [25] A. Leonardis, A. Gupta, and R. Bajcsy. Segmentation of range image as the search for geometric parametric models. *International Journal of Computer Vision*, 14(3):253–277, 1995.
- [26] J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *International Journal of Computer Vision*, 43(1):7–27, 2001.
- [27] R. Malladi, J. Sethian, and B. Vemuri. Shape modeling with front propagation: a level set approach. *IEEE Trans. on Pattern Anal. Machine Intell.*, 17:158–175, 1995.
- [28] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications in Pure and Applied Mathematics*, 42(4):577–685, 1989.
- [29] S. Osher and N. Paragios, editors. *Geometric Level Set Methods in Imaging, Vision, and Graphics*, chapter Fast Edge Integration, pages 59–77. Springer Verlag, 2002.
- [30] S. Osher and J. Sethian. Fronts propagating with curvature-dependent speed: algorithms based on the Hamilton-Jacobi formulation. *Journal of Computational Physics*, 79:12–49, 1988.
- [31] N. Paragios and R. Deriche. Geodesic active regions and level set methods for supervised texture segmentation. *Int. J. Computer Vision*, 2002.
- [32] E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.
- [33] R. C. Reininger and J. D. Gibson. Distributions of the tow-dimensional DCT coefficients for images. *IEEE Trans. on Communications*, 31(6):835–839, 1983.
- [34] R. Ronfard. Region-based strategies for active contour models. *Int. J. Computer Vision*, 13:229–251, 1994.
- [35] C. Samson, L. Blanc-Feraud, G. Aubert, and J. Zerubia. A level set method for image classification. In *Int. Conf. Scale-Space Theories in Computer Vision*, pages 306–317, 1999.
- [36] J. A. Sethian. *Level Set Methods: Evolving Interfaces in Geometry, Fluid Mechanics, Computer Vision, and Material Science*. Cambridge University Press, 1996.
- [37] J. Strain. The fast Gauss transform with variable scales. *siamSSC*, 12(5):1131–1139, 1991.

- [38] M. Tang and S. Ma. General scheme of region competition based on scale space. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(12):1366–1378, 2001.
- [39] A. Tsai, A. Yezzi, Jr., and A. S. Willsky. Curve evolution implementation of the Mumford-Shah functional for image segmentation, denoising, interpolation, and magnification. *IEEE Trans. on Image Processing*, 10(8):1169–1186, 2001.
- [40] G. Unal, H. Krim, and A. Yezzi. A vertex-based representation of objects in an image. In *Proc. IEEE Conf. on Image Processing*, volume 1, pages 896–899, 2002.
- [41] G. Unal, A. Yezzi, and H. Krim. Information-theoretic active polygons for unsupervised texture segmentation. *preprint, submitted to IJCV*, 2002.
- [42] P. Viola. *Alignment by Maximization of Mutual Information*. PhD thesis, Massachusetts Institute of Technology, 1995.
- [43] A. Yezzi, Jr., S. Kichenassamy, A. Kumar, P. Olver, and A. Tannenbaum. A geometric snake model for segmentation of medical imagery. *IEEE Trans. on Medical Imaging*, 16(2):199–209, April 1997.
- [44] A. Yezzi, Jr., A. Tsai, and A. Willsky. A statistical approach to snakes for bimodal and trimodal imagery. In *Int. Conf. on Computer Vision*, pages 898–903, 1999.
- [45] S. C. Zhu and A. Yuille. Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(9):884–900, September 1996.