Int J Comput Vis (2009) 84: 97–111
DOI 10.1007/s11263-009-0234-0

Local Histogram Based Segmentation Using the Wasserstein Distance

Kangyu Ni · Xavier Bresson · Tony Chan · Selim Esedoglu

Received: 12 July 2008 / Accepted: 24 March 2009 / Published online: 4 April 2009

Abstract We propose and analyze a nonparametric region-based active contour model for segmenting cluttered scenes. The proposed model is unsupervised and assumes pixel intensity is independently identically distributed. Our proposed energy functional consists of a geometric regularization term that penalizes the length of the partition boundaries and a region-based image term that uses histograms of pixel intensity to distinguish different regions. More specifically, the region data encourages segmentation so that local histograms within each region are approximately homogeneous. An advantage of using local histograms in the data term is that histogram differentiation is not required to solve the energy minimization problem. We use *Wasserstein distance with exponent 1* to determine the dissimilarity between two histograms. The Wasserstein distance is a metric and is able to faithfully measure the distance between two histograms, compared to many pointwise distances. Moreover, it is insensitive to oscillations, and therefore our model is robust to noise. A fast global minimization method based on (Chan et al. in SIAM J. Appl. Math. 66(5):1632–1648, 2006; Bresson et al. in J. Math. Imaging Vis. 28(2):151–167, 2007) is employed to solve the proposed model. The advantages of using this method are two-fold. First, the computational time is less than that of the method by gradient descent of the associated Euler-Lagrange equation (Chan et al. in Proc. of SSVM, pp. 697–708, 2007). Second, it is able to find a global minimizer. Finally, we propose a variant of our model that is able to properly segment a cluttered scene with local illumination changes.

Keywords Image segmentation · Unsupervised · Wasserstein distance · Image processing · Computer vision · Nonparametric

1 Introduction

Image segmentation plays an important role in computer vision. It involves a process that simplifies an image by partitioning the image domain into several regions. Many existing methods segment an image according to edge information and/or region information. Examples of edge-basded methods are snake (Kass et al. 1991), balloon (Cohen 1991), and geodesic active contours based (Caselles et al. 1997; Kichenesamy et al. 1996) methods, which use edge detection functions and evolve contours towards sharp gradients of pixel intensity. This classic active contour approach is widely used in medical image processing. Although these methods are quite effective, they are usually not robust to noise because noise also has large gradients. One way to process a noisy image is to add a smoothing step prior segmentation, but doing this also smoothes image edges. Region-based active contour models incorporate region information so that image within each segmented region has a uniform characteristics, such as intensities and textures.

This research is supported by ONR grant N00014-09-1-0105 and NSF grant DMS-0610079.

K. Ni (✉) · X. Bresson · T. Chan
University of California, Los Angeles, USA
e-mail: kni66@math.ucla.edu

X. Bresson
e-mail: xbresson@math.ucla.edu

T. Chan
e-mail: chan@math.ucla.edu

S. Esedoglu
University of Michigan, Ann Arbor, USA
e-mail: esedoglu@umich.edu

These methods are therefore robust to noise and furthermore able to detect objects with either sharp or smooth edges. One of the first region-based active contour models is the Mumford-Shah model (Mumford and Shah 1989), which approximates an image by a piecewise smooth function, with a length penalizing term of the edge set of the piecewise smooth function. However, this model is difficult to solve in practice because of the unknown edge set, in addition to the unknown smooth function. Ambrosio and Tortorelli (1990) approximates the Mumford-Shah functional by approximating the edge set by smooth functions, and this is easier to solve. The active contours without edges (ACWE) model (Chan and Vese 2001) is a variant of the piecewise constant Mumford-Shah model. It approximates an image by a two-phase piecewise constant function and is based on a level-set implementation (Osher and Sethian 1988). The solution is easily obtained by the minimizing flow derived by computing the variation of the energy with respect to the level set function. Region competition (Zhu and Yuille 1996) is a statistical and variational model based on minimizing a generalized Bayes and Minimum description length criterion. This model penalizes the boundary length and the Bayes error within each region, in which appropriate probability distributions are chosen. The ACWE, region competition, and other parametric region-based active contour models, such as (Yezzi et al. 1999; Paragios and Deriche 2002), had the assumption that the probability density function (pdf) of the pixel intensity in each region is up to a few parameters. For example, often a Gaussian distribution is assumed with mean and variance the only unknowns. However, many natural images are not necessarily described by Gaussian distribution.

Nonparametric region-based active contour models, such as (Aubert et al. 2005; Herbulot et al. 2004, 2006; Kim et al. 2005; Michailovich et al. 2007), use the entire pdf, or histogram, to drive the segmentation. Therefore, they do not suffer from the above limitations. Our model is related to, yet different from, existing models. In (Aubert et al. 2005; Herbulot et al. 2006), the segmentation model is supervised, and the data descriptors directly depend on the regions, which consequently involves histogram differentiation in the evolution equations. Unsupervised segmentation models in (Herbulot et al. 2004; Kim et al. 2005) take an information-theoretic approach and their data descriptors also directly depend on the regions and therefore also requires histogram differentiation. The model in (Michailovich et al. 2007) maximizes the Bhattacharyya distance between the histogram inside the segmentation curve and the histogram outside the curve. In our work, the data descriptors do not directly depend on the regions and therefore our model does not involve histogram differentiation. This is achieved through the use of local histograms. The local histogram of a pixel is defined as the total number of each gray level on a local region of that pixel. The local region of a pixel, for instance, may be chosen to be a square patch centered at that pixel. These local histograms of intensity are used as the image feature. The proposed model finds a partition such that the local histograms in each region are similar to one another. Local statistics have also been used for segmentation in (Zhu et al. 2005), but the model is parametric, in which intensity statistics are assumed to be Gaussian distributions.

Many existing nonparametric segmentation models are quite effective for many natural images. Among the popular distances used for comparing two histograms are the χ^2 statistics, Kullback-Leibler divergence, and the Bhattacharyya distance (Georgiou et al. 2007). A common feature of these distances is that they are pointwise with respect to histogram bins. As addressed in our previous work in (Chan et al. 2007), this may not be reliable for histogram comparison even under simple circumstances. For example, a pointwise distance between two delta functions with disjoint supports is the same no matter how close or how far the supports are from each other. This is a situation that arises often in segmentation applications, since for example an image, which consists of two objects with approximately constant intensities within each region but distinct intensity means, would fall into this category. The above mentioned existing nonparametric methods commonly use the Parzen window method (Parzen 1962) to approximate and smooth histograms. The smoothing operation may alleviate the above issue with pointwise distances. However, the degree of smoothness is generally a user-selected parameter and is often crucial for segmentation. To overcome the issue with pointwise distances, we use an optimal transport distance, which extends as a metric to measure such as the delta functions and does not require histograms to be smoothed. For this reason, we believe this to be the more natural and appropriate way to compare histograms for segmentation. The optimal transport ideas has been employed in other contexts in image processing, such as image registration (Haker et al. 2004) and classification (Rubner et al. 1998).

The optimal transport, or the Monge-Kantorovich, problem is to find the most efficient plan to rearrange one probability measure into another. We will describe Kantorovich's version (Kantorovich 1942) here. Let (X, μ) and (Y, ν) be two probability measure spaces. Let π be a probability measure on the product space $X \times Y$ and $\Pi(\mu, \nu) = \{\pi \in P(X \times Y) : \pi[A \times Y] = \mu[A]$ and $\pi[X \times B] = \nu[B]$ hold for all measureable sets $A \in X$ and $B \in Y\}$ be the set of admissible transference plans. For a given cost function $c : X \times Y \to \mathbb{R}$, where $c(x, y)$ means the cost of moving from location x to location y, the total transport cost associ-

ated to plan $\pi \in \Pi(\mu, \nu)$ is

$$I[\pi] = \int_{X\times Y} c(x, y)d\pi(x, y). \tag{1}$$

The optimal transport cost between μ and ν is

$$T_c(\mu, \nu) = \inf_{\pi\in\Pi(\mu,\nu)} I[\pi]. \tag{2}$$

More detail can be found in (Rachev and Rüschendorf 1998; Villani 2003). In the case when X and Y are the real line, $\mathbb{R}$, and the cost function is $c(x, y) = |x - y|^p$, the optimal transport cost has a closed-form solution,

$$T_p(\mu, \nu) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|^p dt, \tag{3}$$

where F and G are the cumulative distribution functions of μ and ν, respectively, and F^{-1} and G^{-1} represent their respective inverse functions. The optimal transport distance, commonly called the *Wasserstein distance with exponent* p, is $W_p(\mu, \nu) = T_p(\mu, \nu)^{1/p}$ and defines a metric. Furthermore, if the cost function is the Euclidean distance $c(x, y) = |x - y|$,

$$\begin{aligned} W_1(\mu, \nu) &= \int_0^1 |F^{-1}(t) - G^{-1}(t)|dt \\ &= \int_{\mathbb{R}} |F(x) - G(x)|dx, \end{aligned} \tag{4}$$

where the last equality is obtained by Fubini-Tonelli Theorem and the proof is provided in the Appendix.

Finally, note that the proposed model shown in this paper is based on the statistics of image intensity, but can certainly be replaced by other features, such as gradient, curvature, orientation and scale. To conclude this section, we list the main contributions of this paper in the following:

1. the novelty of using the Wasserstein distance to properly compare histograms without a smoothing approximation for histograms,
2. a segmentation model that does not need to differentiate histograms to find a solution,
3. the use of the fast global minimization method (Bresson et al. 2007) to solve the proposed model, which significantly improves the previous model (Chan et al. 2007) in two ways, the computational time is less than the standard method and initialization can be arbitrary,
4. mathematical properties of the proposed model are presented.

2 Related Works

Kim et al. (2005) took an information-theoretic approach and proposed a nonparametric region-based active contour model. Given an image $I : \Omega \to [0, L]$ with two regions, in each of which pixel intensities are independently identically distributed, a curve $\vec{C}$ is evolved towards the boundary. Denote the region inside (resp. outside) the curve $\vec{C}$ by Σ (resp. Σ^c). Define the region labels associated with curve $\vec{C}$ by

$$L_{\vec{C}}(x) = \begin{cases} L_1 & \text{if } x \in \Sigma, \\ L_2 & \text{if } x \in \Sigma^c. \end{cases}$$

The proposed model maximizes the mutual information between the image pixel intensities and region labels, subject to a constraint on the total length of the region boundaries:

$$\inf_{\vec{C}} \oint_{\vec{C}} ds - \lambda|\Omega|M(I(X); L_{\vec{C}}(X)), \tag{5}$$

where λ is a positive parameter, $|\cdot|$ is the 2-dimensional Lebesgue measure, i.e. area, and M stands for mutual information, defined as:

$$M(I(X); L_{\vec{C}}(X)) = h(I(X)) - h(I(X)|L_{\vec{C}}(X)). \tag{6}$$

Since entropy of image $h(I(X))$ is constant, maximizing the mutual information between $I(X)$ and $L_{\vec{C}}(X)$ minimizes the conditional entropy $h(I(X)|L_{\vec{C}}(X))$. The curve $\vec{C}$ is evolved so that knowing which region a pixel belongs to decreases the uncertainty of the pixel intensity. The conditional entropy is

$$\begin{aligned} &h(I(X)|L_{\vec{C}}(X)) \\ &\quad = -\frac{1}{|\Omega|}\left(\int_\Sigma \log P_1(I(x))dx + \int_{\Sigma^c} \log P_2(I(x))dx\right), \end{aligned} \tag{7}$$

where the probability density functions $P_1(I(x))$ and $P_2(I(x))$ of each region are approximated using the Parzen window method (Parzen 1962),

$$P_1(I(x)) = \frac{1}{|\Sigma|}\int_\Sigma K(I(x) - I(\hat{x}))d\hat{x}, \tag{8}$$

$$P_2(I(x)) = \frac{1}{|\Sigma^c|}\int_{\Sigma^c} K(I(x) - I(\hat{x}))d\hat{x}. \tag{9}$$

The Gaussian function $K(z) = (1/\sqrt{2\pi\sigma^2})e^{-z^2/2\sigma^2}$ is used as a smoothing kernel, where σ is a scalar parameter that controls the smoothness of the approximation. The minimization problem (5) is solved by the following gradient flow:

$$\begin{aligned} \frac{\partial\vec{C}}{\partial t} = \lambda\Bigg[&\log\frac{P_1(I(\vec{C}))}{P_2(I(\vec{C}))} + \frac{1}{|\Sigma|}\int_\Sigma \frac{K(I(x) - I(\vec{C}))}{P_1(I(x))}dx \\ &- \frac{1}{|\Sigma^c|}\int_{\Sigma^c} \frac{K(I(x) - I(\vec{C}))}{P_2(I(x))}dx\Bigg]\vec{N} - \kappa\vec{N}, \end{aligned} \tag{10}$$

where $\vec{N}$ is the outward normal and κ is the curvature of $\vec{C}$. The implementation for (10) is by the level-set method with narrow band approach.

Herbulot et al. (2004) also took a nonparametric region-based active contours approach and used information entropy as competition between two regions:

$$\inf_{\vec{C}} \oint_{\vec{C}} ds + \lambda h(I(X), \Sigma) + \lambda h(I(X), \Sigma^c), \tag{11}$$

where entropy of pixel intensities in each region is

$$h(I(X), \Sigma) = -\int_{\Sigma} P_1(I(x)) \log P_1(I(x)) dx \tag{12}$$

$$h(I(X), \Sigma^c) = -\int_{\Sigma^c} P_2(I(x)) \log P_2(I(x)) dx. \tag{13}$$

The probability density functions $P_1(I(x))$ and $P_2(I(x))$ are approximated using the Parzen window method as described in (8) and (9). The minimization is solved by the following gradient flow:

$$\begin{aligned}\frac{\partial \vec{C}}{\partial t} = \lambda \Bigg[& -(P_1(\log P_1 + 1) - P_2(\log P_2 + 1)) \\ & - \frac{1}{|\Omega|} \Bigg(h(I(X), \Sigma) - h(I(X), \Sigma^c) \\ & + \int_{\Sigma} K(I(x) - I(\vec{C})) \log P_1(I(x)) dx \\ & + \int_{\Sigma^c} K(I(x) - I(\vec{C})) \log P_2(I(x)) dx \Bigg) - \kappa \Bigg] \vec{N},\end{aligned} \tag{14}$$

The curve evolution is implemented by using smoothing B-splines.

3 Proposed Model I

In this section, we discuss an unsupervised segmentation model proposed in our previous work (Chan et al. 2007) for cluttered images. Suppose the observed gray-scale image $I : \Omega \to [0, L]$ is measurable and has two regions of interests. Let $\mathcal{N}_{x,r}$ be the local region centered at x with radius r. Define the local histogram of a pixel $x \in \Omega$ by

$$P_x(y) := \frac{|\{z \in \mathcal{N}_{x,r} \cap \Omega : I(z) = y\}|}{|\mathcal{N}_{x,r} \cap \Omega|}, \tag{15}$$

for $0 \le y \le L$. Define the corresponding cumulative distribution function by

$$F_x(y) := \frac{|\{z \in \mathcal{N}_{x,r} \cap \Omega : I(z) \le y\}|}{|\mathcal{N}_{x,r} \cap \Omega|}, \tag{16}$$

for $0 \le y \le L$. These are the image data used in the following proposed segmentation model:

$$\begin{aligned}\inf_{\Sigma, P_1, P_2} \Bigg\{ & E_1(\cdot, \cdot, \cdot | I) = \mathrm{Per}(\Sigma) \\ & + \lambda \int_{\Sigma} W_1(P_1, P_x) dx + \lambda \int_{\Sigma^c} W_1(P_2, P_x) dx \Bigg\},\end{aligned} \tag{17}$$

where $\mathrm{Per}(\Sigma)$ is the perimeter of the set Σ. This minimization problem finds an optimal region $\Sigma \subseteq \Omega$ and approximates the local histograms inside Σ (resp. Σ^c) by a constant histogram P_1 (resp. P_2). Recall that W_1 is the Wasserstein distance with exponent 1, described in the introduction:

$$W_1(P_1, P_2) = \int_0^L |F_1(y) - F_2(y)| dy. \tag{18}$$

The energy functional (17) can be formulated in terms of the level set method (Osher and Sethian 1988). The boundary between Σ and Σ^c is represented by the 0-level set of a Lipschitz function $\phi : \Omega \to \mathbb{R}$.

$$\begin{aligned}\inf_{\phi, F_1, F_2} \Bigg\{ & E_1(\cdot, \cdot, \cdot | I) = \int_{\Omega} |\nabla H(\phi(x))| dx \\ & + \lambda \int_{\Omega} H(\phi(x)) \int_0^L |F_1(y) - F_x(y)| dy\, dx \\ & + \lambda \int_{\Omega} [1 - H(\phi(x))] \int_0^L |F_2(y) - F_x(y)| dy\, dx \Bigg\},\end{aligned} \tag{19}$$

where H is the Heaviside function, $\int_{\Omega} |\nabla H(\phi(x))| dx$ represents $\mathrm{Per}(\Sigma)$, and $H(\phi)$ (resp. $1 - H(\phi)$) defines Σ (resp. Σ^c).

The minimization of (19) can be achieved by a two-step scheme, which gives a local minimum. First, we fix ϕ and minimize with respect to F_1 and F_2, respectively. Variations with respect to F_1 and F_2 yield the following optimality conditions that should be held for all $0 \le y \le L$,

$$\int H(\phi(x)) \frac{F_1(y) - F_x(y)}{|F_1(y) - F_x(y)|} dx = 0 \tag{20}$$

and

$$\int [1 - H(\phi(x))] \frac{F_2(y) - F_x(y)}{|F_2(y) - F_x(y)|} dx = 0, \tag{21}$$

respectively. The solutions to (20) and (21) are

$$F_1(y) = \text{median of } F_x(y), \quad \text{over } \{x : \phi(x) \ge 0\} \tag{22}$$

and

$$F_2(y) = \text{median of } F_x(y), \quad \text{over } \{x : \phi(x) < 0\}. \tag{23}$$

To see this intuitively, for a fixed y, the quotient in (20) is equal to $+1$ if $F_1(y) > F_x(y)$ and is equal to -1 if $F_1(y) < F_x(y)$. The factor $H(\phi(x))$ in front of the quotient is equal to 1 or 0, depending on whether x is inside Σ or outside Σ, respectively. Since equation (20) requires the integral of $+1$

and -1 over all x inside Σ equals zero, the unknown $F_1(y)$ has to separate the higher half values from the lower half, and therefore is the median.

Next, with fixed F_1 and F_2, the gradient descent of Euler-Lagrange equation for ϕ gives

$$\phi_t = \delta(\phi)\Bigg[\nabla\cdot\left(\frac{\nabla\phi}{|\nabla\phi|}\right) - \lambda\int_0^L (|F_1(y) - F_x(y)| - |F_2(y) - F_x(y)|)dy\Bigg], \tag{24}$$

where δ is a regularized Dirac function and $\nabla\cdot(\frac{\nabla\phi}{|\nabla\phi|})$ is the curvature of the level sets. In implementation, δ is a non-compactly supported approximation as in (Chan and Vese 2001) and steps (22), (23), and (24) are iterated alternately, until convergence to a steady state solution. Note that for computational efficiency, F_x is treated only dependent on the pixel location x and is computed only once before minimization. Near the boundary of the regions, this is not accurate because the local region of x may across over both regions. However, in our experiment, when the size of the local region is properly chosen, the final contours seem to be quite accurate. Another issue regarding the size of the local region is that it depends on the location in the image and the size of textures and/or clutters. In this paper, a constant size for all pixel locations is given by the user.

Numerically, (24) has a serious time-step restriction, in addition to being a second-order equation. The curvature term, the first term of (24), is approximated by

$$\frac{\partial}{\partial x}\left(\frac{\phi_x}{\sqrt{\phi_x^2+\phi_y^2+\epsilon^2}}\right) + \frac{\partial}{\partial y}\left(\frac{\phi_y}{\sqrt{\phi_x^2+\phi_y^2+\epsilon^2}}\right), \tag{25}$$

where $\epsilon > 0$ so that the denominators are not zero but small enough to stay close to the solution. By the CFL condition, the time-step restriction of the explicit scheme for (24) as in (Osher and Fedkiw 2002) is $\Delta t \le c\cdot\epsilon\cdot(\Delta x)^2$, where c is a constant. The factor ϵ comes from (25) when $\phi_x^2+\phi_y^2 = 0$. This time-step restriction can be improved to $\Delta t \le c\cdot(\Delta x)^2$ with Chambolle's method (Chambolle 2004), where $c = 1/8$. The application of Chambolle's method on the proposed model is presented in Sect. 4.3.

4 Fast Global Minimization of Model I

4.1 Global Minimization of Model I

Like many variational segmentation models, model (17) suffers from being non-convex (with respect to Σ) and is therefore sensitive to initializations. The requirement of reasonable initializations conflicts the purpose of automatic segmentation. Numerically, a non-compactly supported Dirac function is used in (Chan and Vese 2001) to increase the chances of finding global minimizers of the piecewise constant segmentation model. Theoretically, based on the framework of (Bresson et al. 2007; Chan et al. 2006; Mory and Ardon 2007), we propose the following global minimization of Model I:

$$\min_{0\le u\le 1, P_1, P_2}\Bigg\{E_2(\cdot,\cdot,\cdot|I) = \int_\Omega |\nabla u(x)|dx + \lambda\int_\Omega W_1(P_1,P_x)u(x)dx + \lambda\int_\Omega W_1(P_2,P_x)(1-u(x))dx\Bigg\}. \tag{26}$$

This model is closely related to model (17) but overcomes the non-convexity. Let $\mathbf{1}_S$ denote the characteristic function of set S. Model (26) extends the original minimization over the non-convex set $\{u\in BV(\Omega): u = \mathbf{1}_\Sigma$, for some set Σ with finite perimeter$\}$ to the convex set $\{u\in BV(\Omega): 0\le u\le 1\}$. Thus, (26) is convex with respect to u and, unlike (17), does not have (non-global) local minima with respect to the geometric unknown.

The major advantage of (26) is that initializations can be arbitrary. The relation between (17) and (26) is that, for fixed F_1 and F_2, a global minimizer of (17) can be found through a global minimizer of (26). This relation is stated in the following theorem, which is based on the geometric properties of total variation.

Theorem 1 (Global Minimizers) *Suppose* $I(x)\in[0,1]$. *If* P_1, *and* P_2 *are fixed, and* $u(x)$ *is any minimizer of* $E_2(\cdot,P_1,P_2|I)$, *then for a.e.* $\rho\in[0,1]$, $\mathbf{1}_{\{x:u(x)>\rho\}}(x)$ *is a global minimizer of* $E_1(\cdot,P_1,P_2|I)$.

Proof Based on (Chan et al. 2006), by the coarea formula and setting $\Sigma(\rho) := \{x : u(x) > \rho\}$, we can write E_2 in terms of E_1

$$\begin{aligned} &E_2(u,P_1,P_2|I)\\ &= \int_0^1\Bigg\{\mathrm{Per}(\Sigma(\rho)) + \lambda\int_{\Sigma(\rho)} W_1(P_1,P_x)dx + \lambda\int_{\Omega-\Sigma(\rho)} W_1(P_2,P_x)dx\Bigg\}d\rho\\ &= \int_0^1 E_1(\Sigma(\rho),P_1,P_2|I)d\rho, \end{aligned} \tag{27}$$

Therefore, if u is a minimizer of $E_2(\cdot,P_1,P_2|I)$, then for a.e. $\rho\in[0,1]$, $\Sigma(\rho)$ is a minimizer of $E_1(\cdot,P_1,P_2|I)$. □

4.2 Existence of Global Minimization Solutions

In this section, we show the existence of minimizers for (26) and convexity of (26) with respect to each variable.

Theorem 2 (Existence of Solutions for u) *For fixed P_1 and P_2,*

$$\min_{0\le u\le 1}\Big\{E_2(\cdot,P_1,P_2|I)=\int_\Omega|\nabla u(x)|dx + \lambda\int_\Omega W_1(P_1,P_x)u(x)dx + \lambda\int_\Omega W_1(P_2,P_x)(1-u(x))dx\Big\} \tag{28}$$

has a solution $u\in BV(\Omega)$ with $0\le u\le 1$.

Proof Let $\{u_n\}\in BV(\Omega)$ with $0\le u\le 1$ be a minimizing sequence. Then, $\int_\Omega|Du_n|$ is uniformly bounded. Since every uniformly bounded sequence in $BV(\Omega)$ is relatively compact in $L^1(\Omega)$, there exists a subsequence $\{u_{n_k}\}$ converging to some $u\in BV(\Omega)$. Since $u_{n_k}\to u$ in $L^1(\Omega)$, we have $u_{n_k}\to u$ in measure, i.e. $|\{x:|u_{n_k}(x)-u(x)|\ge\epsilon\}|\to 0$ as $\epsilon\to 0$. Since we also have $0\le u_{n_k}\le 1$, u satisfies $0\le u\le 1$. Finally, one can check easily that u is indeed a minimizer by the lower semicontinuity of $BV(\Omega)$ and Fatou's lemma. □

For fixed u, the minimizer for F_1 (resp. F_2) has an explicit solution. Variations of E_2 with respect to F_1 and F_2 yield the following optimality conditions that should hold for all $0\le y\le L$:

$$\int u(x)\frac{F_1(y)-F_x(y)}{|F_1(y)-F_x(y)|}dx=0 \tag{29}$$

and

$$\int[1-u(x)]\frac{F_2(y)-F_x(y)}{|F_2(y)-F_x(y)|}dx=0, \tag{30}$$

respectively. Therefore,

$$F_1(y)=\text{weighted (by } u(x)\text{) median of } F_x(y), \tag{31}$$

and

$$F_2(y)=\text{weighted (by } 1-u(x)\text{) median of } F_x(y), \tag{32}$$

We will next show that $E_2[u,P_1,P_2|I]$ is convex with respect to each variable. First, E_2 is convex with respect to u because $\int_\Omega|Du(x)|dx$ is convex in u and the set $\{u\in BV(\Omega):0\le u\le 1\}$ is convex. Second,

Theorem 3 *The minimization problem*

$$\min_{P_1\in P(\Omega)}E_2[u,\cdot,P_2|I]$$

is convex, where $P(\Omega)$ denotes the set of Borel probability measures on Ω.

Proof $E_2[u,\cdot,P_2|I]$ is convex in P_1 because the Wasserstein distance is a metric and in particular satisfies the triangle inequality. Since $P(\Omega)$ is a convex set, minimization with fixed u and P_2 is a convex problem. □

Similarly, the minimization $\min_{P_2\in P(\Omega)}E_2[u,P_1,\cdot|I]$ is convex. Therefore, $E_2[u,P_1,P_2|I]$ is convex with respect to each variable.

4.3 Fast Minimization Scheme

Minimizing the proposed energy E_2 in (26) with respect to u can be efficiently solved by applying methods in (Aujol et al. 2006; Bresson et al. 2007). The regularization and data terms in (26) can be decoupled by using a new variable v to replace u in the data term and adding a convex term that forces v and u to be close to each other:

$$\min_{u,0\le v\le 1}\int_\Omega|\nabla u(x)|dx+\frac{1}{2\theta}\int_\Omega(u(x)-v(x))^2dx + \lambda\int_\Omega r(x,F_1,F_2)v(x)dx, \tag{33}$$

where

$$r(x,F_1,F_2)=\int_0^L|F_1(y)-F_x(y)|-|F_2(y)-F_x(y)|dy,$$

and $\theta>0$ is a small parameter. Minimizing the convex variational model (33) can be approached by alternately solving the following coupled problems:

$$\min_u\int_\Omega|\nabla u(x)|+\frac{1}{2\theta}(u(x)-v(x))^2dx \tag{34}$$

and

$$\min_{0\le v\le 1}\int_\Omega\frac{1}{2\theta}(u(x)-v(x))^2+\lambda r(x,F_1,F_2)v(x)dx. \tag{35}$$

The minimization problem in (34) can be efficiently achieved by Chambolle's method (Chambolle 2004), based on the dual formulation of the total variation norm. The derived solution is

$$u(x)=v(x)-\theta\,\mathrm{div}\,p(x), \tag{36}$$

where $p=(p^1,p^2)$ solves $\nabla(\theta\,\mathrm{div}\,p-v)-|\nabla(\theta\,\mathrm{div}\,p-v)|p=0$ and is solved by a fixed point method,

$$p^{n+1}=\frac{p^n+\delta t\nabla(\mathrm{div}\,p^n-v/\theta)}{1+\delta t|(\mathrm{div}\,p^n-v/\theta)|}. \tag{37}$$

Table 1 Properties of the proposed model and Kim et al. (2005) and Herbulot et al. (2004) models

	Our model	Kim et al. (2005)	Herbulot et al. (2004)
existence of solution	✓	✓	✓
global minimum/convexity	✓	x	x
fast minimization	✓	x	✓
insensibility to noise	✓	–	–
no need to smooth histograms (noiseless case)	✓	x	x
local change of lighting	✓	x	x
complexity for one iteration	O(Lmn)	O(M)	O(LM)
time-step restriction	$\frac{1}{8}\cdot(\Delta x)^2$	$c\cdot\epsilon\cdot(\Delta x)^2$	$c\cdot\epsilon\cdot(\Delta x)^2$
computational time	1 mins	3 mins	4 mins
handle topological changes	✓	✓	x

The solution of (35) is found as in (Bresson et al. 2007):

$$v(x) = \max\{\min\{u(x) - \theta\lambda r(x, F_1, F_2), 1\}, 0\}. \tag{38}$$

The proposed fast minimization scheme is to iterate (31), (32), (37), (36), and (38) alternately, until convergence.

5 Properties of Proposed Models and Comparison with Other Models

The proposed model has several desired mathematical properties as shown in Table 1. In Sect. 4.2, we show the existence of solution and the convexity of the model in each variable. Based on Chambolle's dual method regarding the length-penalizing term, the solution converges after a small number of iterations, compared to directly solving the associated Euler-Lagrange equation. Moreover, since the Wasserstein distance is insensitive to oscillations, our model is intrinsically robust to noise. On the other hand, it does not require histograms to be smoothed, which has to be done for many segmentation models even for noiseless images. For instance, the Wasserstein distance is able to distinguish the distance between any pair of delta functions with disjoint supports. Many distances do not tell apart the distance between two disjointly supported histograms unless the histograms are smoothed. The complexity of computing one iteration is O(Lmn). The time-step restriction is $\Delta t \leq \frac{1}{8}\cdot(\Delta x)^2$, as discussed in Sect. 3. For a 144×144 image as in Fig. 1, the computational time for a solution to converge is approximately one minute. Since the partition is implicitly embedded in function u, the model is able to handle topological changes.

Kim et al.'s model (Kim et al. 2005) also has existence of solution and minimizes over a non-convex set $\{u \in BV(\Omega) : u = 1_\Sigma,$ for some set Σ with finite perimeter $\}$, thus does not guarantee to get a global minimizer. The gradient flow (10) has a curvature term and the convergence can be slow, due to the CFL condition discussed in Sect. 3. The probability density functions are estimated by the Parzen window method. This enables their model to handle noise but introduces a user-selected parameter, i.e. kernel width. They use the fast Gauss transform to compute probability densities, which reduces the complexity of computing one iteration to $O(M)$, where M is the size of the narrow band. The time-step restriction is $\Delta t \leq c\cdot\epsilon\cdot(\Delta x)^2$, for some small ϵ and a constant c. Typically, ϵ is taken to be about 0.01. The level-set method is used for curve evolution and thus allows topological changes.

Herbulot et al. (2004) use smoothing B-splines to implement their derived evolution equation instead of the usual level-set method to avoid extensive computational time. The complexity of each iteration is O(LM), where L is the number of gray levels and M is the size of the narrow band. The time-step restriction is $\Delta t \leq c\cdot\epsilon\cdot(\Delta x)^2$. The parametric method using B-splines does not handle topological changes of the contours. They further use smoothing B-splines in order to be more robust to noise. The tradeoff between the smoothness and interpolation error is controlled by a parameter that has to be chosen by the user. Their model also minimizes over a non-convex set, thus does not guarantee to get a global minimizer.

6 Description of Model II

We propose a variant of Model I that properly handles segmentation when the captured scene is under uneven lighting exposure, due to reasons such as the location of the light source and camera. The original model considers the data term globally, i.e. compares all the local histograms within each region. Therefore, when the local lighting changes significantly, local histograms of the same feature may have similar shapes but are far apart by a translation in the intensity axis. As a result, the Wasserstein distance between them is large and thus the original model is not designed to deal with uneven lighting. To model this variation, we introduce

a function $a(x)$, representing the translation in the intensity axis, and propose a new model:

$$\inf_{\Sigma,a,F_1,F_2}\Big\{E_3(\Sigma,a,F_1,F_2|I)=\mathrm{Per}(\Sigma)+\frac{\alpha}{2}\int|\nabla a(x)|^2dx+\lambda\int_\Sigma\int_0^L|F_1(y)-F_x(y-a(x))|dy\,dx+\lambda\int_{\Sigma^c}\int_0^L|F_2(y)-F_x(y-a(x))|dy\,dx\Big\}. \quad (39)$$

This model allows local histograms to translate on the intensity axis in order to find a best fit among one another within each region. A regularity constraint $\int|\nabla a(x)|^2dx$ is imposed to ensure smoothness of a.

To solve the minimization, we have the following three-step scheme. The evolution equations for F_1, F_2 and ϕ can be derived similarly as in Sect. 3:

$$F_1(y)=\text{median of } F_x(y-a(x)),\quad \text{over } \{\phi\geq 0\}, \quad (40)$$

$$F_2(y)=\text{median of } F_x(y-a(x)),\quad \text{over } \{\phi<0\}, \quad (41)$$

$$\phi_t=\delta(\phi)\Big[\nabla\cdot\Big(\frac{\nabla\phi}{|\nabla\phi|}\Big)-\lambda\int_0^L\Big(|F_1(y)-F_x(y-a(x))|-|F_2(y)-F_x(y-a(x))|\Big)dy\Big]. \quad (42)$$

The minimization with respect to $a(x)$ is to solve:

$$\inf_a E_3(\Sigma,\cdot,F_1,F_2|I)=\frac{\alpha}{2}\int|\nabla a(x)|^2dx+\lambda\int_\Sigma\int_0^L|F_1(y)-F_x(y-a(x))|dy\,dx+\lambda\int_{\Sigma^c}\int_0^L|F_2(y)-F_x(y-a(x))|dy\,dx. \quad (43)$$

Without the first term, $a(x)$ can be solved explicitly by

$$a_0(x)=\begin{cases}F_1^{-1}(0.5)-F_x^{-1}(0.5) & \text{if } \phi(x)>0,\\ F_2^{-1}(0.5)-F_x^{-1}(0.5) & \text{if } \phi(x)\leq 0.\end{cases}$$

Therefore, the problem of (43) can be transformed into the following:

$$\inf_a\frac{1}{2}\int|a(x)-a_0(x)|^2dx+\frac{\alpha}{2}\int|\nabla a(x)|^2dx. \quad (44)$$

The solution to (44) is $a(x)-\alpha\triangle a(x)=a_0(x)$, which can be easily solved, for example, by the fast Fourier transform. We may also employ the fast global minimization technique for Model II, instead using (42).

7 Experimental Results

7.1 Comparison with Other Methods

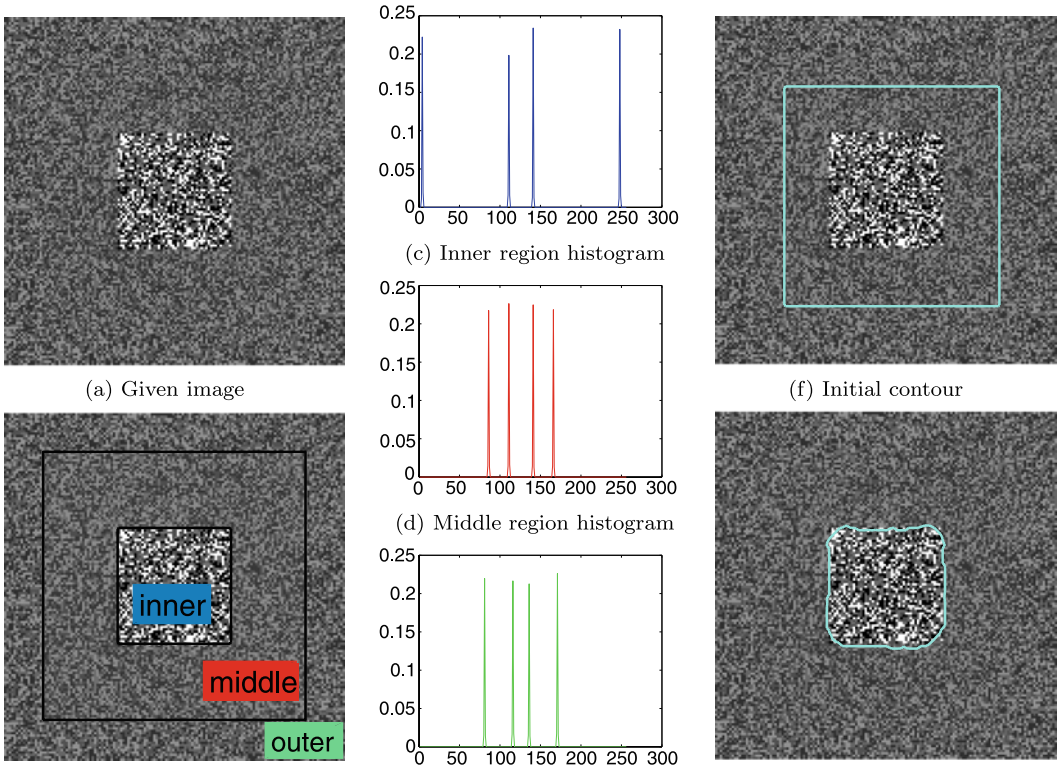
As explained in Sect. 5, our model does not require histograms to be smoothed for proper segmentation. In contrast, previous methods use Parzen window method (Parzen 1962) to estimate pdfs, which requires a smoothness parameter selection. If the bandwidth of the kernel is too small, point-wise metrics cannot detect similar intensities. Figure 1(a) is a synthetic image with three regions, in each of which the pixel intensity is independently identically distributed (b). The pixels in the inner region take intensities 3, 110, 140, and 247, with probability about 0.25 each. The pixels in the middle region take intensities 85, 110, 140, and 165, with probability about 0.25 each. The pixels in the outer region take intensities 80, 115, 135, and 170, with probability about 0.25 each. The middle and outer regions are perceptually similar and so are their corresponding intensity histograms, (d) and (e), respectively. A desired partition is to distinguish the inner region from the rest. The initial contour is shown in (f). Our model does not have the smoothing parameter and correctly segments the inner region from the rest because of the use of the Wasserstein distance.

On the other hand, Kim et al.'s model (Kim et al. 2005) needs a careful selection of the smoothness parameter σ (variance of the Gaussian kernel) in order to segment correctly. Figure 2(a) is the final contour with $\sigma=5$, which incorrectly groups the inner and middle regions together. This is because the histograms of the inner and middle regions overlap 50% but the histograms of the middle and outer region do not overlap. In (b), the segmentation with $\sigma=10$ is correct because the intensity pdf is greatly smoothed and thus mutual information is able distinguish the inner region from the rest. When $\sigma=50$, the final contours (c) incorrectly separate pixels with intensity $=3,247$ from pixels with intensity $=110,115,135,140$.

We emphasize here that nonparametric models are able to deal with a greater variety of images than parametric models. In this experiment, the object and background have the same intensity mean and variance. In Fig. 3(a), we show the boundaries of the objects in red curves and the corresponding histograms in each region. Figure 3(c) and (b) are the final contours of our proposed model and ACWE, respectively. The proposed model is able to distinguish the objects from the background. On the other hand the ACWE model cannot handle this case due to its parametric nature.

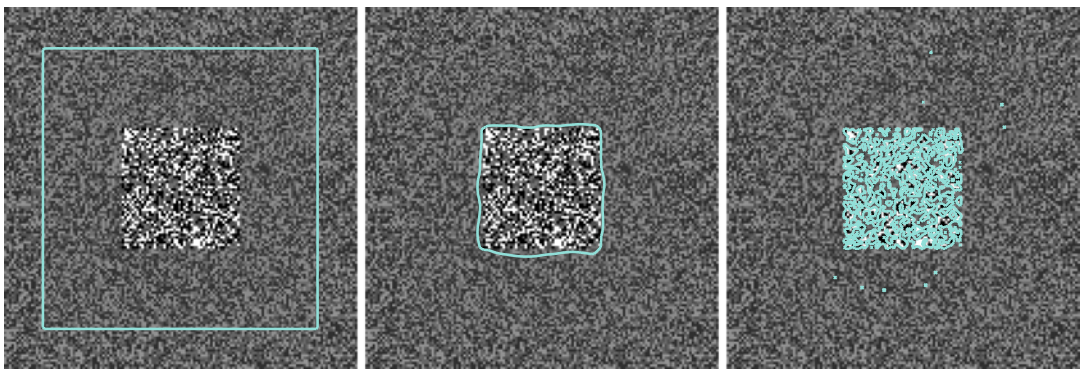
7.2 Comparison between Original Model and Fast Global Minimization

The proposed fast global minimization in Sect. 4 improves the original minimization in (Chan et al. 2007) described in



(a) Given image (b) Intensity is i.i.d in each region (c) Inner region histogram (d) Middle region histogram (e) Outer region histogram (f) Initial contour (g) Model I

Fig. 1 The given image (**a**) has three regions (**b**), in each of which pixel intensity is independently identically distributed. (**c**), (**d**), and (**e**) are the intensity histograms of the pixels in the inner, middle, and outer regions, respectively. The pixels in the inner region take intensities 3, 110, 140, and 247, with probability about 0.25 each; the pixels in the middle region take intensities 85, 110, 140, and 165; and the pixels in the outer region take intensities 80, 115, 135, and 170. The middle and outer regions look similar, as well as their corresponding histograms. Wasserstein distance does not require histograms to be smoothed in order to compare histograms in a reasonable manner. The final contour of proposed model I, in (**g**), correctly distinguish the inner region from the rest



(a) Kim et al. with $\sigma = 5$ (b) Kim et al. with $\sigma = 10$ (c) Kim et al. with $\sigma = 50$

Fig. 2 Kim et al.'s model (Kim et al. 2005) needs a proper selection of the smoothness parameter σ in order to segment correctly. (**a**) is the final contour with $\sigma = 5$, which incorrectly groups the inner and middle regions (see Fig. 1(b)). The segmentation with $\sigma = 10$ (**b**) is correct because the intensity pdf is greatly smoothed and thus mutual information is able distinguish the inner region from the rest. When $\sigma = 50$ (**c**), the final contours separate pixels with intensity 3 and 247 from pixels with intensity 110, 115, 135, and 140

Sect. 3 of model I. Figure 4 is a downsized 175×135 image of cheetah. In Sects. 4.1 and 4.2, we explain that the global minimization model is convex and therefore all local minima are global minima. A gradient descent method is guaranteed to find a global minimizer. We experiment with several images with different and arbitrary initializations and

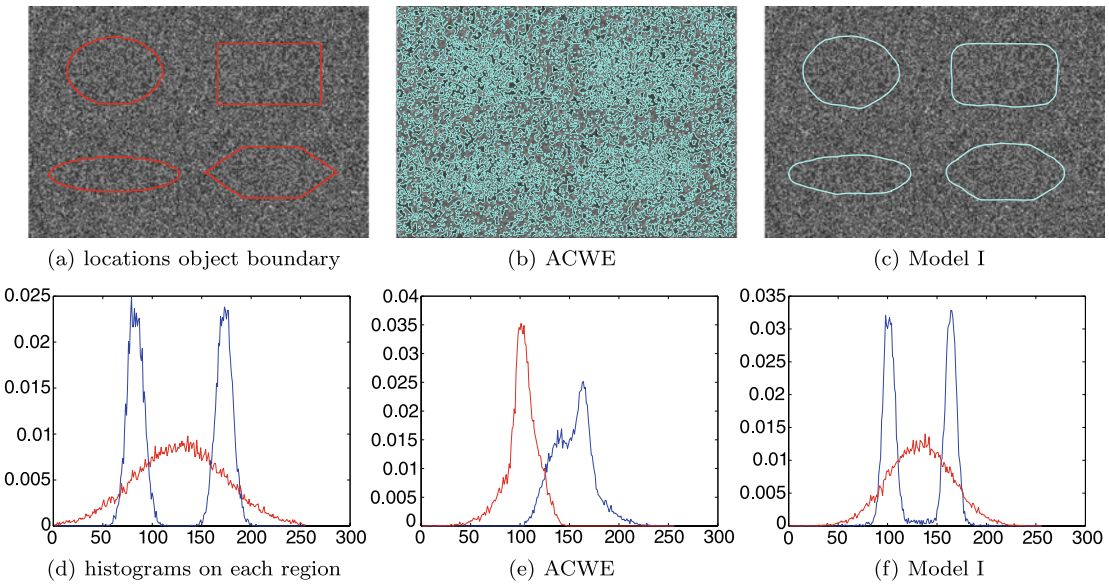


Fig. 3 Objects and background regions have the same intensity mean and variance. (**a**) shows the location of object boundary. (**b**) shows the final contours of ACWE model. (**c**) shows the final contours of proposed model I. (**d**), (**e**), and (**f**) are the histograms of each region for the contours in (**a**), (**b**), and (**c**), respectively. One can see that a non-parametric segmentation model is needed for this image in order to distinguish different regions. This is because the histograms are distinct but have the same parameters, i.e. mean and variance

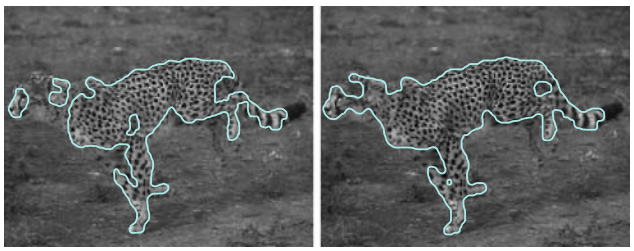


Fig. 4 Down-sized cheetah image of Berkeley Segmentation Dataset. (**a**) shows the final contours of the method in our previous work in (Chan et al. 2007) and (**b**) shows the final contours of the model in (26). Global minimization scheme improves segmentation result as well as the computational time, from about two hours to two minutes

all arrive at similar results for each image. This is a nice consequence of the proposed model being convex with respect to each variable. On the other hand, the original minimization is non-convex and thus requires initializations to be reasonably close to the final contours. Moreover, the fast global minimization improves the speed from two hours to two minutes.

7.3 Robustness to Noise/More Results of Model I and Model II

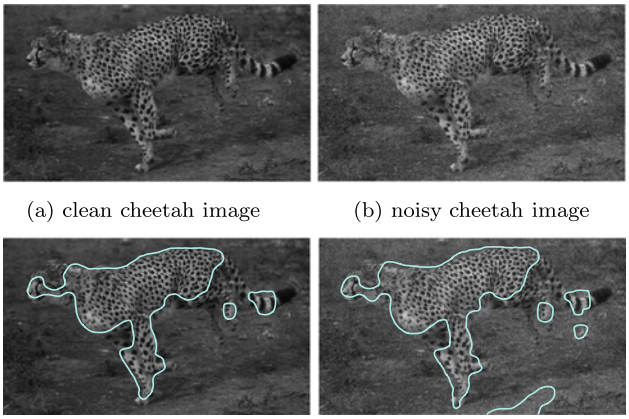
Figure 5(a) is a clean image of cheetah and (b) is with noise. The final contour shown in (d) by the global minimization of Model I is able to segment the cheetah patterns and is nearly as good as the result in (c) of the clean image. In this experiment, the radius of the local region is 11.

Figure 6 shows other experiments of Model I. The first experiment is a 285×281 image consisting of two Brodatz textures. The final contours are shown in (a) and the corresponding histograms on each region are plotted in (c). Model I is able distinguish these two Brodatz textures, even though their intensity distributions are highly discontinuous. The second is a 481×321 image of tiger; (b) shows the final contours by Model I and (d) shows the histograms in each region. The final contour successfully selects the tiger patterns.

Figure 7 shows that Model II improves Model I when there are local lighting changes in the image. The first experiment is a 384×223 image of cheetah. In (a), Model I is able to capture some of the cheetah patterns but not near the back legs, due to the local lighting difference. Final contours of Model II, in (b), are more accurate. Another experiment is a 282×218 image of fish. The final contours by Model I, in (d), do not select the fish patterns accurately, because the local illumination is significantly uneven. Model II, on the other hand, is able to overcome this difficulty, as shown in (e) the final contours separates the fish patterns from the background.

7.4 Implementation Issues

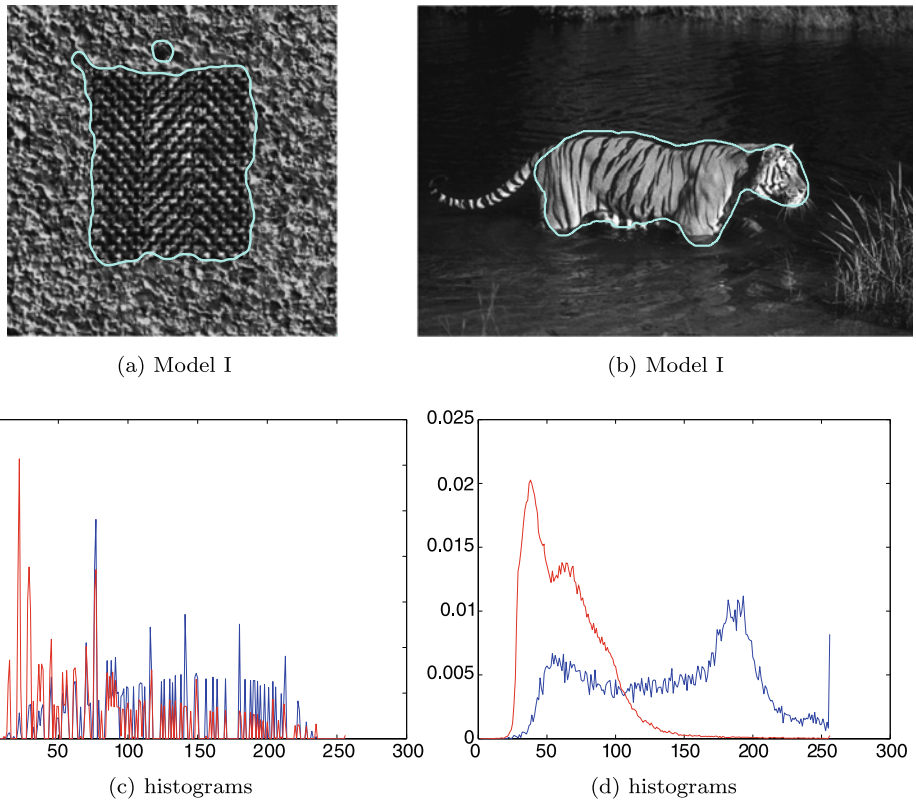
We show a method to solve the weighted median for $F_1(y)$ in (31) in the discrete case.



(a) clean cheetah image (b) noisy cheetah image

(c) final contour of clean image (d) final contour of noisy image

Fig. 5 Experimental results of Model I. (**a**) is the original clean image of cheetah. (**b**) is the image of cheetah with added noise. The final contours of the noisy image, in (**d**), is nearly as good as the final contours of the clean image, in (**c**). Model I is robust to noise because the Wasserstein distance is insensitive to oscillations



(a) Model I (b) Model I

(c) histograms (d) histograms

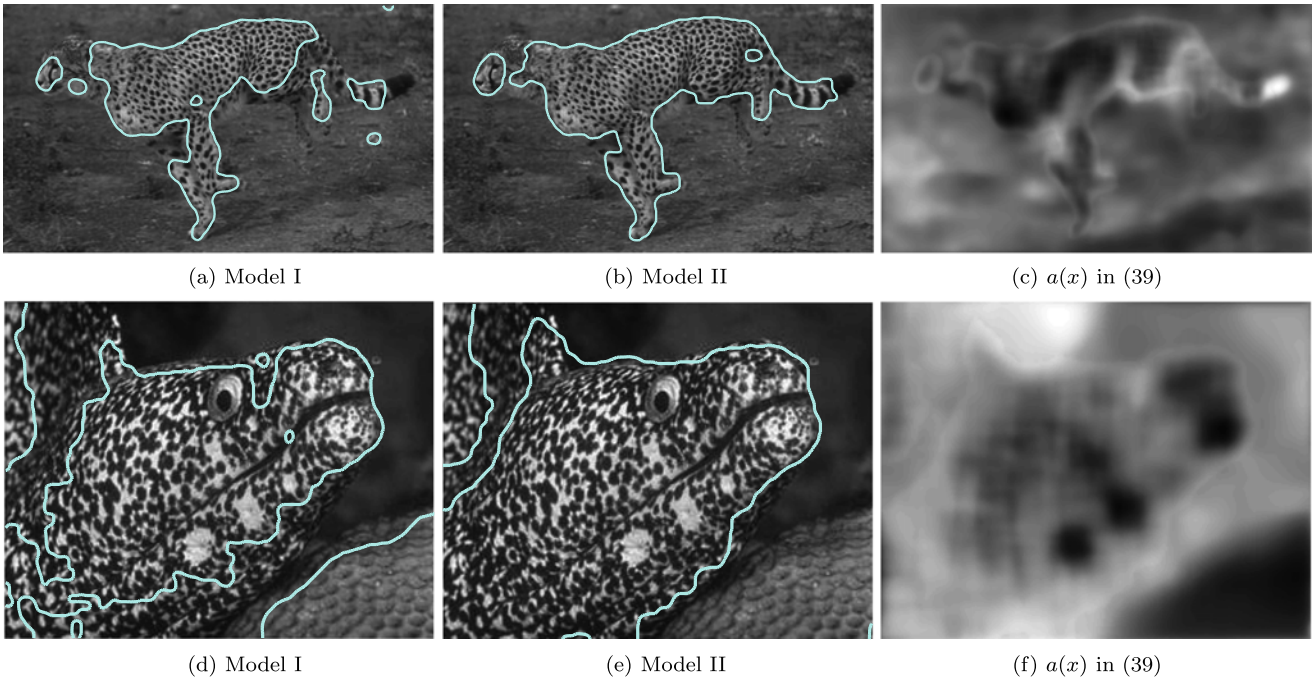
Fig. 6 Experimental results of Model I. (**a**) shows the final contours of a synthetic image consisting of two Brodatz textures. (**c**) shows the corresponding histograms of each region. Notice that they are highly discontinuous and the supports of histograms overlap greatly. (**b**) shows the final contours of an image of tiger from Berkeley Segmentation Dataset. (**d**) shows the corresponding histograms of each region

For each $y = 0, 1, \ldots, L$,

1. Compute the weighted histogram, H_y, of value $F_x(y)$ with weight $u(x)$. More precisely, for all pixels $x \in \Omega$, each value $F_x(y)$ is counted $u(x)$ times. Then, normalize the weighted histogram, H_y, by dividing by the total count, $\sum_{x\in\Omega} u(x)$.
2. For each weighted histogram H_y, compute the cumulative distribution C_y.
3. The weighted median is then $F_1(y) = C_y^{-1}(0.5)$.

The calculation of $F_2(y)$ is similar and with weight $1-u(x)$.

We empirically demonstrate the segmentation results are not sensitive to the size of the local histogram region, within a reasonable range. The experiment is on a 384×223 image of cheetah, shown in Fig. 5(a). Figure 8 shows final contours by global minimization of Model I with different local region sizes, radius ranging from 1 to 25. If the size is smaller than the clutter features, the final contour partitions clutter features into smaller regions, an undesired result. If the size is large enough, our results show the cheetah patterns are segmented correctly.



(a) Model I (b) Model II (c) $a(x)$ in (39)

(d) Model I (e) Model II (f) $a(x)$ in (39)

Fig. 7 (**a**) shows the final contours by Model I on the image of cheetah in Fig. 5(a). (**b**) shows the final contours by Model II. (**c**) is the final function $a(x)$ in (39). This smoothness component allows local illumination changes and captures more of the cheetah pattern. The second row is the experiment of an image of fish from Berkeley Segmentation Dataset. (**d**) shows the final contours by Model I and (**e**) is by Model II. (**f**) is the final function $a(x)$ in (39). Model II is able to capture more of the fish pattern for this image

7.5 Limitations and Extensions

Our segmentation model is formulated for gray-scale images but can be extended to color images. The data term can be generalized because the Wasserstein distance is defined on any space of probability measures. However, the implementation would be much more complicated because there is no closed form for the Wasserstein distance between two probability measures on Euclidean spaces with dimensions larger than one. The Earth Movers Distance between signatures is equivalent to the Wasserstein distance when signatures have the same total mass (or normalized discrete pdfs) and the optimization has been investigated in (Rubner et al. 1998). This can be a possible direction to extend our segmentation model. Works in (Chartrand et al. 2005; Haker et al. 2004) numerically solve the optimal maps of the optimal transport problem on $\mathbb{R}^2$ and may also be applied to our extension. Another limitation is that our model assumes the given image has two regions of clutters. Many natural images have more than two regions and requires a multi-phase segmentation model. This limitation can be easily overcome, since our model has a natural extension to multi-phase segmentation as in (Vese and Chan 2002). Moreover, since our model only uses the intensity probability density, it does not take into account higher-order characteristics, such as gradient, scale, and orientation. For example, if two textures have the same intensity probability density, our model is not able to distinguish them. However, histograms of suitable descriptors can be used instead of or combined with intensity. On the other hand, our segmentation model can contribute to segmentation algorithms, such as (Sapiro and Caselles 1997; Tu and Zhu 2002) that incorporate many image characteristics, including clutter.

8 Conclusions

In this paper, we proposed a fast global minimization of a local histogram based model using the Wasserstein distance with exponent 1 to segment cluttered scenes. Our model is different from previous nonparametric region-based active contour models in three ways. The first is the use the Wasserstein distance, which is able to properly compare both continuous and discontinuous histograms. We are not claiming the Wasserstein distance is better than other distances used for nonparametric segmentation in the literature but rather raising the fundamental limitations with pointwise distances. Second, the proposed model does not need to differentiate histograms to find the solutions. Many existing models require histograms to be differentiated and thus rely on a smoothing step, usually the Parzen window method. The third is the application of the global minimization method for a nonparametric model. Consequently, the segmentation results are not sensitive to initializations. The

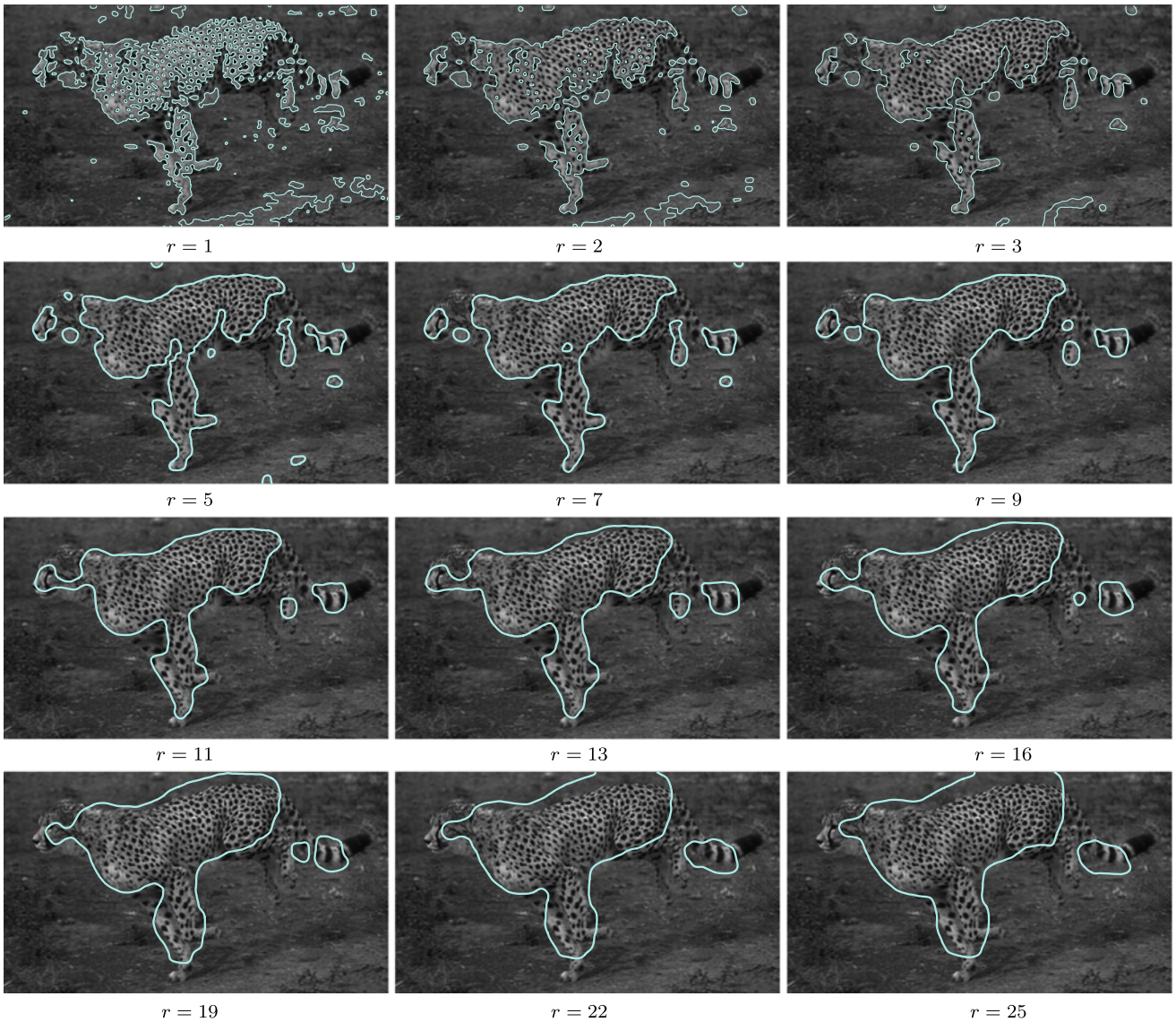


Fig. 8 The size r of the local region in model I, described by (26), needs to be equal or bigger than the smallest features of interest in the given image. The images shown here are the final contours with different sizes r, ranging from 1 to 25. The segmentation results are not too sensitive to the size of the local region, but are more accurate when the size is closer to that of the smallest image features of interest

second and third were made possible by the use of local histograms. We have proved a number of desired mathematical properties of the model and provided experimental verifications. In the future, we will generalize our model to color images and multi-phase segmentation. The former can be achieved by using the fast minimization of vectorial total variation in (Bresson and Chan 2007) and adapting the numerical scheme for computing the optimal transport distance in (Rubner et al. 1998; Chartrand et al. 2005; Haker et al. 2004). The later can be approached by applying methods such as the multi-phase level set framework (Vese and Chan 2002).

Appendix

Theorem *Let μ and ν be two probability measures on $\mathbb{R}$. Let $F : \mathbb{R} \to [0, 1]$ and $G : \mathbb{R} \to [0, 1]$ be the corresponding cumulative distribution functions. Then,*

$$\int_{\mathbb{R}} |F(x) - G(x)|dx = \int_0^1 |F^{-1}(t) - G^{-1}(t)|dt.$$

Proof Without loss of generality, suppose both F and G are supported on $[0, L]$. First, we will show that

$$\int_0^L F(x)dm_1(x) = L - \int_0^1 F^{-1}(t)dm_2(t), \tag{45}$$

where m_1 denotes Lebesgue measure restricted on $[0, L]$ and m_2 denotes Lebesgue measure restricted on $[0, 1]$.

Let $S_F = \{(x, t) \in [0, L] \times [0, 1] : t \leq F(x)\}$. It is easy to check that S_F is $B_{[0,L]} \times B_{[0,1]}$-measurable, where $B_{[0,L]}$ and $B_{[0,1]}$ denote the Borel σ-algebra restricted on $[0, L]$ and $[0, 1]$, respectively. Then,

$$\begin{aligned} m_1 \times m_2(S_F) &= \int \mathbf{1}_{S_F}(x, t)d(m_1 \times m_2) \\ &= \int_0^1 \left(\int_0^L \mathbf{1}_{S_F}(x, t)dm_1(x) \right) dm_2(t) \qquad (46) \\ &= \int_0^L \left(\int_0^1 \mathbf{1}_{S_F}(x, t)dm_2(t) \right) dm_1(x), \qquad (47) \end{aligned}$$

where the last two equalities are by Fubini-Tonelli Theorem since everything is σ-finite. Next,

$$\begin{aligned} \text{Eq. (46)} &= \int_0^1 \left(\int_0^L \mathbf{1}_{S_F}(x, t)dm_1(x) \right) dm_2(t) \\ &= \int_0^1 \left(\int_0^L \mathbf{1}_{\{x:F^{-1}(t)\leq x\leq L\}}(x)dm_1(x) \right) dm_2(t) \\ &= \int_0^1 \left(\int_0^L \mathbf{1}_{[0,L]}(x) - \mathbf{1}_{[0,F^{-1}(t)]}(x)dm_1(x) \right) dm_2(t) \\ &= L - \int_0^1 F^{-1}(t)dm_2(t) \qquad (48) \end{aligned}$$

and

$$\begin{aligned} \text{Eq. (47)} &= \int_0^L \left(\int_0^1 \mathbf{1}_{S_F}(x, t)dm_2(t) \right) dm_1(x) \\ &= \int_0^L \left(\int_0^1 \mathbf{1}_{\{t:0\leq t\leq F(x)\}}(t)dm_2(t) \right) dm_1(x) \\ &= \int_0^L \left(\int_0^1 \mathbf{1}_{[0,F(x)]}(t)dm_2(t) \right) dm_1(x) \\ &= \int_0^L F(x)dm_1(x). \qquad (49) \end{aligned}$$

By (48) and (49), we have proved (45). Similarly,

$$\int_0^L G(x)dm_1(x) = L - \int_0^1 G^{-1}(t)dm_2(t). \tag{50}$$

Without loss of generality, we may assume $F(x) - G(x) \geq 0$, because we can partition $[0, L]$ into a finite subintervals so that $F(x) - G(x)$ is monotone in each subinterval. Then,

$$\begin{aligned} &\int_0^L |F(x) - G(x)|dx \\ &= \int_0^L F(x) - G(x)dm_1(x) \\ &= L - \int_0^1 F^{-1}(t)dm_2(t) - L + \int_0^1 G^{-1}(t)dm_2(t) \\ &= \int_0^1 G^{-1}(t)dm_2(t) - \int_0^1 F^{-1}(t)dm_2(t) \\ &= \int_0^1 |F^{-1}(t) - G^{-1}(t)|dt. \qquad (51) \end{aligned}$$

□

References

Ambrosio, L., & Tortorelli, V. M. (1990). Approximation of functionals depending on jumps by elliptic functionals via Gamma convergence. *Communications on Pure and Applied Mathematics*, *43*, 999–1036.

Aubert, G., Barlaud, M., Faugeras, O., & Jehan-Besson, S. (2005). Image segmentation using active contours: calculus of variations or shape gradients? *SIAM Journal on Applied Mathematics*, *1*(2), 2128–2145.

Aujol, J. F., Gilboa, G., Chan, T., & Osher, S. (2006). Structure-texture image decomposition—modeling, algorithms, and parameter selection. *International Journal of Computer Vision*, 67(1).

Bresson, X., & Chan, T. (2007). *Fast minimization of the vectorial total variation norm and applications to color image processing*. UCLA CAM report 07-25.

Bresson, X., Esedoglu, S., Vandergheynst, P., Thiran, J. P., & Osher, S. (2007). Fast global minimization of the active contour/snake model. *Journal of Mathematical Imaging and Vision*, *28*(2), 151–167.

Caselles, V., Kimmel, R., & Sapiro, G. (1997). Geodesic active contours. *International Journal of Computer Vision*, *22*(1), 61–79.

Chambolle, A. (2004). An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, *20*(1–2), 89–97.

Chan, T. F., & Vese, L. A. (2001). Active contours without edges. *IEEE Transactions on Image Processing*, *10*(2), 266–277.

Chan, T., Esedoglu, S., & Nikolova, M. (2006). Algorithms for finding global minimizers of image segmentation and denoising models. *SIAM Journal on Applied Mathematics*, *66*(5), 1632–1648.

Chan, T., Esedoglu, S., & Ni, K. (2007). Histogram based segmentation using Wasserstein distances. In *Proceedings of SSVM* (pp. 697–708).

Chartrand, R., Vixie, K., Wohlber, B., & Bollt, E. (2005). A gradient descent solution to the Monge-Kantorovich problem. Preprint: LA-UR-04-6305.

Cohen, L. (1991). On active contour models and balloons. *Computer Vision, Graphics, and Image Processing*, *53*, 211–218.

Georgiou, T., Michailovich, O., Rathi, Y., Malcolm, J., & Tannenbaum, A. (2007). Distribution metrics and image segmentation. *Linear Algebra and its Applications*, *405*, 663–672.

Haker, S., Zhu, L., & Tannenbaum, A. (2004). Optimal mass transport for registration and warping. *International Journal of Computer Vision*, *60*(3), 225–240.

Herbulot, A., Jehan-Besson, S., Barlaud, M., & Aubert, G. (2004). Shape gradient for image segmentation using information theory. In *ICASSP* (Vol. 3, pp. 21–24).

Herbulot, A., Jehan-Besson, S., Duffner, S., Barlaud, M., & Aubert, G. (2006). Segmentation of vectorial image features using shape gradients and information measures. *Journal of Mathematical Imaging and Vision*, *25*(3), 365–386.

Kantorovich, L. V. (1942). On the translocation of masses. *Doklady Akademii Nauk SSSR*, *37*, 199–201.

Kass, M., Witkin, A., & Terzopoulos, D. (1991). Snakes: active contours model. *International Journal of Computer Vision*, *1*, 1167–1186.

Kichenesamy, S., Kumar, A., Olver, P., Tannenbaum, A., & Yezzi, A. (1996). Conformal curvature flows: from phase transitions to active vision. *Archive for Rational Mechanics and Analysis*, *134*, 275–301.

Kim, J., Fisher, J. W., Yezzi, A., Cetin, M., & Willsky, A. S. (2005). A nonparametric statistical method for image segmentation using information theory and curve evolution. *IEEE Transactions on Image Processing*, *14*, 1486–1502.

Michailovich, O., Rathi, Y., & Tannenbaum, A. (2007). Image segmentation using active contours driven by the Bhattacharya gradient flow. *IEEE Transactions on Image Processing*, *16*(11), 2787–2801.

Mory, B., & Ardon, R. (2007). Fuzzy region competition: a convex two-phase segmentation framework. In *Proceedings of SSVM* (pp. 214–226).

Mumford, D., & Shah, J. (1989). Optimal approximation by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics*, *42*, 577–685.

Osher, S., & Fedkiw, R. (2002). *Applied mathematical sciences: Vol. 153. Level set methods and dynamic implicit surfaces*. New York: Springer.

Osher, S., & Sethian, J. A. (1988). Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulation. *Journal of Computational Physics*, *79*, 12–49.

Paragios, N., & Deriche, R. (2002). Geodesic active regions: a new paradigm to deal with frame partition problems in computer vision. *Journal of Visual Communication and Image Representation*, *13*(1–2), 249–268.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, *33*, 1065–1076.

Rachev, S., & Rüschendorf, L. (1998). *Mass transportation problems. Vol. I: Theory, Vol. II: Applications. Probability and its applications*. New York: Springer.

Rubner, Y., Tomasi, C., & Guibas, L. J. (1998). A metric for distributions with applications to image databases. In: *IEEE international conference on computer vision* (pp. 59–66).

Sapiro, G., & Caselles, V. (1997). Histogram modification via differential equations. *Journal of Differential Equations*, *135*(2), 238–268.

Tu, Z., & Zhu, S. (2002). Image segmentation by data-driven Markov chain Monte Carlo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(5).

Vese, L. A., & Chan, T. F. (2002). A multiphase level set framework for image segmentation using the Mumford and Shah model. *International Journal of Computer Vision*, *50*(3), 271–293.

Villani, C. (2003). *Graduate studies in mathematics: Vol. 58. Topics in optimal transportation*. Providence: American Mathematical Society.

Yezzi, A. Jr., Tsai, A., & Willsky, A. (1999). A statistical approach to snakes for bimodal and trimodal imagery. In *International conference on computer vision* (pp. 898–903).

Zhu, S. C., & Yuille, A. (1996). Region competition: unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*(9), 884–900.

Zhu, W., Jiang, T., & Li, X. (2005). Local region based medical image segmentation using J-divergence measures. In *Proceedings of EMBC*.