WASSERSTEIN ACTIVE CONTOURS

Gabriel Peyré,

CNRS and CEREMADE,
Université Paris-Dauphine, France
gabriel.peyre@ceremade.dauphine.fr

Jalal Fadili and Julien Rabin

GREYC, CNRS-ENSICAEN
Université de Caen, France,
jalal.fadili@greyc.ensicaen.fr, *julien.rabin@unicaen.fr*

ABSTRACT

In this paper, we propose a novel and rigorous framework for region-based active contours that combines the Wasserstein distance between statistical distributions in arbitrary dimension and shape derivative tools. To speed-up the computation and be able to handle high-dimensional features and large-scale data, we introduce an approximation of the differential of the Wasserstein distance between histograms. The framework is flexible enough to allow either minimization of the Wasserstein distance to prior distributions, or maximization of the distance between the distributions of the regions to be segmented (*i.e. region competition*). Numerical results reported demonstrate the advantages of the proposed optimal transport distance with respect to point-wise metrics.

Index Terms— Image segmentation, Optimal transport

1. INTRODUCTION

1.1. Overview of the Literature

Contour-based vs. region-based segmentation methods. Active contours for image segmentation methods can be broadly classified as being either edge-based or region-based. Starting from the seminal work on the snakes model [1], contour-based actives contours are driven towards image edges through the minimization of an energy usually derived from the image gradient magnitude.

In Region-Based Active Contours (RBAC) approaches, the evolution equation is generally deduced from a general criterion that includes both region integrals and boundary integrals; see *e.g.* [2, 3]. The main issue when dealing with RBAC models is the computation of the velocity vector in the evolution equation from the energy functional, especially when the descriptors are region-dependent, as will be the case in this work. To circumvent this difficulty, we here take benefit from the shape derivation principles, see [4, 5].

Statistical segmentation A number of authors have proposed RBAC energy functionals involving statistical region-based terms. These are typically functions of the distribution of some image attributes within the region. The distribution can be either parametric or non-parametric, see for instance[6, 7]. In the non-parametric approach, the energy functional usually involves a point-wise distance between non-parametric kernel estimates (*e.g.* Parzen kernel) of the underlying densities. Such an approach requires a proper choice of the smoothing kernel bandwidth.

Optimal transport for imaging. To avoid the drawbacks faced with traditional statistical distances, the authors in [8] propose to use the Wasserstein distance.

The Wasserstein distance originates from the theory of optimal transport [9]. It defines a natural metric between probability distributions. This Wasserstein distance has found a wide range of applications for imaging problems such as the comparison of histogram features for image retrieval, shape recognition, histogram specification and color transfer [10].

1.2. Relation to Previous Work and Contributions

To the best of our knowledge, the work of [8] is the first, and so far the only one, to clearly address the statistical segmentation problem using a Wasserstein distance. Their work clearly emphasizes the usefulness of optimal transport methods to deal with statistically localized features.

Our work, however, departs significantly from theirs in many important ways. First, unlike their work which focused on scalar features with the L^1 Wasserstein distance, we consider in §3 a general setting in *arbitrary dimension*. Secondly, contrary to [8] which does not take into account the explicit dependence of the Wasserstein distance to the region, the distance in our energy functional is explicitly *region-dependent*. Using shape derivative tools, we also provide the exact derivative of the Wasserstein distance with respect to the domain boundary and then deduce the active contour velocity field. In contrast, the work [8] use a patch-based local histogram estimation to avoid taking into account the dependence of the statistics to the region. While this solution appears appealing by its simplicity, and provides good results on synthetic and natural images, it faces an important dilemma when choosing the patch size, which is governed by a tradeoff between accuracy and robustness.

We eventually propose in section 3.2 an approximation for the computation of the Wasserstein distance between dis-

This work has been supported by the European Research Council (ERC project SIGMA-Vision) and the French National Research Agency (project NatImages).

tributions in high dimension and then illustrate its practical interest with numerical experiments (§ 4).

2. STATISTICAL SEGMENTATION

Notations In the sequel, we consider a feature map $I : u \in \Sigma \to \mathbb{R}^d$, where $u \in \Sigma$ indexes the pixel location, and d is the dimension of the feature of interest (for instance $d = 3$ for a color image). We also consider a histogram binning grid $\Omega \subset \mathbb{R}^d$.

In the following, we consider Σ as a continuous domain (equipped with the Lebesgue measure) and Ω as a finite discrete grid (equipped with the counting measure). We thus define the Hilbert spaces $\mathrm{L}^2(\Omega)$ and $\mathrm{L}^2(\Sigma)$ endowed with the inner products $\langle A,\, B\rangle_\Omega = \sum_{x\in\Omega} A(x)B(x)$ and $\langle f,\, g\rangle_\Sigma = \int_\Sigma f(u)g(u)\mathrm{d}u$. The set of statistical distributions on Ω is $\mathcal{D}(\Omega) = \{A \backslash A(x) \geqslant 0 \quad \text{and} \quad \langle A,\, \mathbb{1}\rangle_\Omega = 1\} \subset \mathrm{L}^2(\Omega)$, where $\mathbb{1}(x) = 1, \forall x \in \Omega$, such that $\langle A,\, \mathbb{1}\rangle_\Omega = \sum_{x\in\Omega} A(x)$.

2.1. Kernel Density Estimator

Given a fixed feature map I, and a non-negative weight function $w \in \mathrm{L}^2(\Sigma)$, the kernel density estimator of the distribution underlying I is given by the mapping $\mathrm{L}^2(\Sigma) \mapsto: \mathcal{D}(\Omega)$

$$\forall\, x \in \Omega,\ P(w) = \frac{\Psi(w)}{\langle \Psi^*(\mathbb{1}),\, w\rangle_\Sigma} \tag{1}$$

where the mapping $\Psi : \mathrm{L}^2(\Sigma) \to \mathrm{L}^2(\Omega)$ and its adjoint Ψ^* are defined as $\Psi(w) : x \mapsto \displaystyle\int_\Sigma w(u)\psi_s(x - I(u))\mathrm{d}u$, and $\Psi^*(f) : u \mapsto \displaystyle\sum_{x\in\Omega} f(x)\psi_s(x - I(u))$, using a non-negative symmetric smooth localized window ψ_s which is referred to as the kernel, and is parametrized by its bandwidth s. Observe that the normalization map $\Psi^*(\mathbb{1})$ corresponds to $[\Psi^*(\mathbb{1})](u) = \sum_{x\in\Omega} \psi_s(x - I(u))\ \forall\, u \in \Sigma$.

A common kernel function is a Gaussian kernel, and the corresponding estimator is termed the *Parzen estimator*. The choice of s is even more crucial and results from a traditional bias-variance tradeoff, and should be adapted to the discretization grid Ω and smoothness of the underlying density.

2.2. Statistical Distance-based Segmentation

Let's consider the problem of variational segmentation of the image domain in two regions $\Sigma = \Gamma \cup \Gamma^c$, where Γ is a regular bounded open set. Γ and its complement Γ^c share the same boundary $\partial\Gamma$ (denoted also C for short), with normals pointing in opposite directions. The goal is to find a (local) minimizer of an energy including both region (Wasserstein fidelity) and boundary (regularity) functionals. The key principle is to construct a PDE from the energy criterion that changes the shape of the current boundary curve according to some velocity field which can be thought of as a direction of descent of the energy criterion.

Shape derivatives of statistical distances. Let's define the region functional

$$E(\Gamma, B) = W(P(\chi_\Gamma), B) \tag{2}$$

for any fixed distribution $B \in \mathcal{D}(\Omega)$, where $\chi_\Gamma(u)$ is the characteristic function of $\Gamma \subset \Sigma$, i.e. $\chi_\Gamma(u) = 1$ if $u \in \Gamma$, and 0 otherwise, and where W is a well chosen distance (see § 3).

Introducing the time τ for the evolution, and considering $m \in [0, 1] \mapsto \mathsf{C}(m, \tau)$ to be a parametric representation of the boundary $\partial\Gamma$ at time τ, a gradient flow of this boundary may be defined from the so-called *shape gradient* $\mathbf{v}_\Gamma$ as

$$\frac{\partial\, \mathsf{C}(m,\tau)}{\partial\tau} = \mathbf{v}_\Gamma(\mathsf{C}(m,\tau)) \quad \text{and} \quad \mathsf{C}(\cdot, 0) = \mathsf{C}_0\,. \tag{3}$$

Proposition 1. *The shape gradient $\mathbf{v}_\Gamma$ ensuring that (3) converges to a stationary point of $E(\Gamma, B)$ is*

$$\forall\, u \in \partial\Gamma \quad \mathbf{v}_\Gamma(u) = G_{\Gamma,B}(u)\, \mathbf{N}_u$$
$$\textit{where} \quad G_{\Gamma,B} = [DP(\chi_\Gamma)^*]\,(\nabla_1 W(P(\chi_\Gamma), B))\,, \tag{4}$$

where $\mathbf{N}_u$ is the unit inward normal to $\partial\Gamma$ at u, and $\nabla_1 W$ is the (sub)gradient of W with respect to its first variable. The adjoint Gâteaux derivative DP^* of the kernel density estimator in direction μ is given by

$$DP(w)^* : \mu \in \mathrm{L}^2(\Omega) \mapsto \frac{\Psi^*(\mu) - \Psi^*(\mathbb{1})\langle P(w),\, \mu\rangle_\Omega}{\langle \Psi^*(\mathbb{1}),\, w\rangle_\Sigma} \in \mathrm{L}^2(\Sigma).$$

Proof. See for instance[5, Theorem 6.1][11, Theorem 2]. □

Level set implementation. The minimization of (2) with respect to the domain Γ may be obtained by introducing an auxiliary function $\varphi : \Sigma \to \mathbb{R}$, which is often chosen to be the signed distance to $\partial\Gamma$. Thus Γ is represented as

$$\Gamma = \{u \in \Sigma \setminus \varphi(u) > 0\} \text{ and } \partial\Gamma = \{u \in \Sigma \setminus \varphi(u) = 0\}\,.$$

The energy (2) is rewritten as $W(P(H(\varphi)), B)$ where $H = \chi_{[0,+\infty)}$ is the Heaviside function. Introducing an artificial time variable τ, the evolution equation (3) associated to the energy (2) then becomes

$$\frac{\partial\varphi(u,\tau)}{\partial\tau} = -|\nabla\varphi(u,\tau)|G_{\Gamma,B}(u)\,. \tag{5}$$

Note that the velocity function $G_{\Gamma,B}$ is computed now for the whole image domain Σ. However, the signed distance function φ is not a solution of the PDE (5), and in practice, it must be periodically re-initialized so that it remains a distance function. This is important to ensure numerical stability of the method.

2.3. Statistical Segmentation by Region Competition

In the same vein as [11, 6], we restrict our attention in this paper to a non-parametric variational segmentation method that seeks the maximization of the distance between the respective

distributions in Γ and Γ^c, i.e. region competition. Of course, our approach can be applied to other energy functionals just as well, *e.g.* those with terms that favor region homogeneity. The energy functional to be minimized reads

$$\min_{\Gamma} \mathcal{E}(\Gamma) = -W(P(\chi_\Gamma), P(\chi_{\Gamma^c})) + \lambda r(\mathsf{C}) \,, \qquad (6)$$

where $r(\mathsf{C})$ is a boundary regularity term, *e.g.* the curve length. Written using the level set formalism, this corresponds to the solution of

$$\min_{\varphi} \; -W(P(H(\varphi)), P(H(-\varphi))) + \lambda R(\varphi) \,, \qquad (7)$$

where $R(\varphi)$ is a suitable regularization associated to $r(\mathsf{C})$. For instance, if $r(\mathsf{C})$ is the length, then $R(\varphi)$ is the TV regularization $R(\varphi) = \int |\nabla\varphi(u)| \mathrm{d}u$.

The equivalent level set evolution PDE (5) that drives an initial contour to a stationary point (hopefully a local minimizer) of (6) is

$$\frac{\partial\varphi(u,\tau)}{\partial\tau} = -|\nabla\varphi(u,\tau)| \left(-G_{\Gamma,B_{\Gamma^c}}(u) + G_{\Gamma^c,B_\Gamma}(u) - \lambda\kappa\right) \,,$$

where $\Gamma = \{u \setminus \varphi(u,\tau) > 0\}$ is the domain at time τ (we have dropped the dependency on τ for the sake of clarity), where the histogram inside and outside Γ are

$$B_\Gamma = P(H(\varphi(\cdot,\tau)) \quad \text{and} \quad B_{\Gamma^c} = P(H(-\varphi(\cdot,\tau)),$$

where G_{Γ,B_Γ} is the velocity defined in (4), and $\kappa = \mathrm{div}\left(\frac{\nabla\varphi}{|\nabla\varphi|}\right)$ is the mean curvature of the boundary. In practice, this PDE is discretized with a sufficiently small time step (for instance computed using a line search).

3. THE WASSERSTEIN DISTANCE AND THE SLICED WASSERSTEIN APPROXIMATION

3.1. Wasserstein Distance on a Non-ordered 1-D Grid

We assume here that A and $B \in \mathcal{D}(\Omega)$ are two discrete distributions defined over the 1-D grid points $\Omega = \{x_i \in \mathbb{R}\}_{i=1}^N$ that are sorted in increasing order, *s.t.* $x_i \leqslant x_{i+1}$

L^p Wasserstein Distance. The L^p Wasserstein distance on the real line for any $p \geqslant 1$ may then be written as follows

$$W(A,B) = \int_0^1 |R_A^{-1}(t) - R_B^{-1}(t)|^p \mathrm{d}t \,,$$

where $R_\mu(s) = \int_{-\infty}^s \mu(x)\mathrm{d}x$ is the cumulative distribution function (CDF) of μ and $R_\mu^{-1}(t) = \inf\{s \setminus R_\mu(s) \geqslant t\}$ its pseudo-inverse. The latter is well defined as the CDF is non-decreasing. When $\mu \in \mathcal{D}(\Omega)$ is discrete, the CDF is equal to $R_\mu(s) = \sum_{x_i \leqslant s} \mu(x_i)$.

1-D Wasserstein distance sub-differential. The following proposition[1] gives the *sub-gradients set* of the Wasserstein distance for 1-D discrete distributions. Note that when the subdifferential is not a singleton, one can take any subgradient of $W(A,B)$ in lieu of $\nabla_1 W(A,B)$ in Eq. (4).

Proposition 2. *Let $A, B \in \mathcal{D}(\Omega)$. For $p \geqslant 1$, the set of sub-gradients of $A \mapsto W(A,B)$ are written as*

$$\{\, \nabla_1 W(A,B) \,\} : x_i \mapsto \sum_{j \geqslant i} |x_j - \tilde{x}_j|^p - |x_{j+1} - \tilde{x}_j|^p \quad (8)$$

$$\textit{where} \begin{cases} \tilde{x}_j = x_k & \textit{if} \quad R_B(x_{k-1}) < R_A(x_j) < R_B(x_k), \\ \tilde{x}_j \in [x_k, x_{k+1}] & \textit{if} \quad R_A(x_j) = R_B(x_k). \end{cases}$$

3.2. Sliced Wasserstein Distance for multivariate features

For large-scale and high-dimensional histograms, the computation of $W(A,B)$ and $\nabla_1 W(A,B)$ for the Wasserstein distance is too demanding, both in time and memory, with a complexity of $O(|\Omega|^3)$. To speed-up the computation, we follow [10] and consider an alternative distance that mimics the Wasserstein distance, but is faster to compute.

Sliced Distance Approximation We define the sliced distance as

$$SW(A,B) = \sum_{\theta \in \Theta} W(A_\theta, B_\theta) \qquad (9)$$

where Θ is a finite subset of the unit sphere in $\mathbb{R}^d$, and $A_\theta \in \mathcal{D}(\Omega_\theta)$ is the projected distribution in the direction θ, defined on 1-D grid points $\Omega_\theta = \{x_\theta = \langle x, \theta\rangle\}_{x\in\Omega}$, that has the same values as A, i.e. $\forall\, x \in \Omega, A_\theta(x_\theta) = A(x)$. The sliced Wasserstein distance is thus a sum of 1-D Wasserstein distances between the projected distributions. Note that the grid Ω_θ that supports these distributions is non-uniform, thus requiring the sorting of bins after projection (see § 3.1). This step, which is achieved in $O(|\Omega|\log(|\Omega|))$ operations, may be precomputed for a fixed set of orientations Θ, or parallelized.

Sliced Wasserstein sub-differential The following proposition[1] enables us to compute efficiently the gradient of the considered distance using Formula (8).

Proposition 3. *The function $A \mapsto SW(A,B)$ is closed and convex and its subdifferential at A is such that*

$$\forall\, x \in \Omega, \quad \partial_1 SW(A,B)(x) = \sum_{\theta\in\Theta} \partial_1 W(A_\theta, B_\theta)(x_\theta)\theta.$$

4. NUMERICAL RESULTS AND CONCLUSION

Wasserstein distance vs. KL divergence We first illustrate the difficulty of point-wise statistical metrics for image segmentation, as already reported in [8] with a different method, in comparison with the Wasserstein distance. We consider here a toy-example with 1-D features and the Kullback-Leibler symmetrized divergence. Figure 1, left, shows an example of a gray-scale image with three regions delimited by two circles of increasing radii $r_0 < r_1$. The distribution

[1] The proof, given in [12], is omitted here due to the lack of space.

of the intensity values within each region is a Gaussian with a mean value in $\{0.05, 0.8, 0.95\}$ and small variances, so that the resulting image is a mixture of three localized and barely overlapping Gaussians. As can be seen from the histograms in Figure 1 top-right, the correct segmentation should group together the two outer regions which have close means.

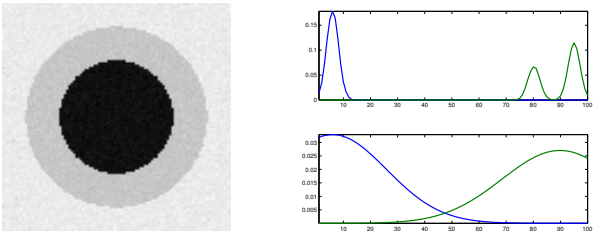

Fig. 1. Left: *gray-scale image example with 3 regions delimited by 2 circles with radii* $r_0 < r_1$. Right: *estimated densities of the dark region* Γ_{r_0} *(in blue) and its complement* $\Gamma^c_{r_0}$ *(in green) using* $P(\chi_\Gamma)$ *from Eq.* (1) *and bandwidth* $s = 10^{-2}$ *(top) and* $s = 0.2$ *(bottom).*

Figure 2, shows the energy landscape for circular regions Γ_r of radius r. The leftmost figure shows this energy for the Wasserstein distance, which provides the proper segmentation. Indeed, circle of radius r_0 is the only local, and hence global, maximum of the L^2 ($p = 2$) Wasserstein distance between the inside and outside regions. The situation is radically different with the KL divergence (rightmost figure), where a spurious local minimum for $r = r_1$ persists, unless an extremely large kernel bandwidth is used. The weakness of point-wise statistical metrics is thus apparent when localized feature histograms come into play, and the smoothing impacts significantly the spatial localization.

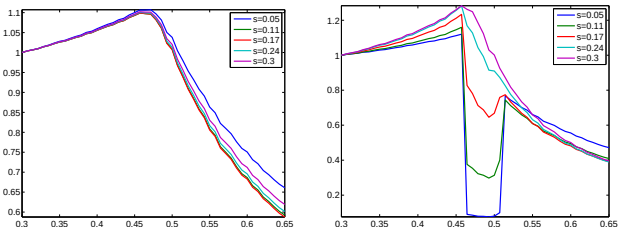

Fig. 2. *Energy* $W(P(\chi_{\Gamma_r}), P(\chi_{\Gamma_r^c}))$ *for a centered circle* Γ_r *of radius* r *as a function of* r. *Each curve corresponds to a different smoothing bandwidth* s *in the Parzen kernel estimator.*

Natural image segmentation We illustrate now the interest of our approach for multi-dimensional features, here using color distributions[2]. The results are displayed in Figure 3 where the initial contour C_0 is drawn in red and the final one $C(;\infty)$ in blue. For these examples we use $|\Theta| = 12$ random directions in 3-D for the computation of the sliced L^2 Wasserstein distance (9). Experiments show that the color distribution is split into two distinctive parts as expected.

[2]A more in-depth study is proposed in [12].

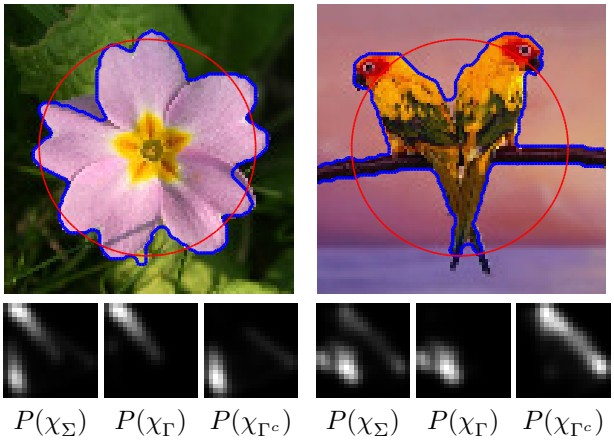

Fig. 3. Top: *color image segmentation results.* Bottom: *color distributions of respectively the whole image, the inside and the outside of the region delimited by the final (blue) contour (only the two first PCA components are displayed here).*

Conclusion We have proposed a mathematically grounded way to handle the statistical segmentation problem in arbitrary dimension. Our framework combines wisely the Wasserstein statistical distance and shape derivative tools.

5. REFERENCES

[1] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *IJCV*, vol. 1, no. 4, pp. 321–331, Jan. 1988.

[2] R. Ronfard, "Region-based strategies for active contour models," *IJCV*, vol. 13, no. 2, pp. 229–251, 1994.

[3] T. Chan and L. Vese, "Active contours without edges"," *IEEE Trans. Image Proc.*, vol. 10, no. 2, pp. 266–277, 2001.

[4] M. C. Delfour and J.-P. Zolésio, *Shapes and geometries: analysis, differential calculus, and optimization*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001.

[5] G. Aubert, M. Barlaud, O. Faugeras, and S. Jehan-Besson, "Image segmentation using active contours: Calculus of variations or shape gradients ?," *SIAM Applied Mathematics*, vol. 63, no. 6, pp. 2128–2154, 2003.

[6] A. Herbulot, S. Jehan-Besson, S. Duffner, M. Barlaud, and G. Aubert, "Segmentation of vectorial image features using shape gradients and information measures," *JMIV*, vol. 25, no. 3, pp. 365–386, Oct. 2006.

[7] G. Unal, A. J. Yezzi, and H. Krim, "Information-theoretic active polygons for unsupervised texture segmentation," *IJCV*, vol. 62, no. 3, pp. 199–220, 2005.

[8] K. Ni, X. Bresson, T. Chan, and S. Esedoglu, "Local histogram based segmentation using the wasserstein distance," *IJCV*, vol. 84, pp. 97–111, August 2009.

[9] C. Villani, *Topics in Optimal Transportation*, American Mathematical Society, 2003.

[10] J. Rabin, G. Peyré, J. Delon, and M. Bernot, "Wasserstein barycenter and its application to texture mixing," *Proc. SSVM'11*, 2011.

[11] S. Jehan-Besson, M. Barlaud, G. Aubert, and O. Faugeras, "Shape gradients for histogram segmentation using active contours," in *Proc. ICCV*, 2003.

[12] J. Fadili, G. Peyré, and J. Rabin, "Wasserstein active contours," Tech. Rep., 2011, http://hal.archives-ouvertes.fr/hal-00593424/.