Report for CSE 5339 2018 — (OTMLSA)
Optimal Transport in Machine Learning and Shape Analysis

Convex Color Image Segmentation with Optimal Transport

Siyang Zhang

April 24, 2018

1 Introduction

Some previous works in image segmentation use mean gray level value to segment image into homogeneous regions. In textured image segmentation, local histograms are used to extend the mean value image segmentation model. Statistical based image segmentation uses parametric models (mean, variance) or empirical distributions combining the Kullback-Leibler divergence. Later, optimal transport is investigated to compare local 1-dimensional histograms. Then the Wasserstein distance is proposed to compare global multi-dimensional histograms. The drawback of these non-convex active contours methods is the sensitiveness to the initial contour, and to reduce such dependence on initialization choice, convex formulations are designed to compute the global distance between histograms.

Yildizoglu(2013) provides a fast algorithm using ℓ_1 norm between cumulative histograms. His work focuses on 1-dimensional global histogram-based segmentation of grayscale images. Swoboda(2013) proposes a convex formulation to deal with low dimensional histograms, using sub-iterations to compute the proximity operator of the Wasserstein distance. Cuturi(2013) provides the entropic regularization of optimal transport distances for handling accurate discretization of histograms. Papadakis and Rabin's work in this paper is related to the work of Cuturi(2016) that using the Legendre-Fenchel transform of regularized transport cost for image segmentation.

2 Histogram-based Segmentation

Some notations are as following: denote $\langle .,. \rangle$ as the Euclidean inner product and the ℓ_2 norm $||.|| = \sqrt{\langle .,. \rangle}$. Denote A^* as the conjugate operator of A and satisfies $\langle Ax, y \rangle = \langle x, A^* y \rangle$. $\mathbf{1}_n$ and $\mathbf{0}_n \in \mathbb{R}^n$ are denoted as the n-dimensional vectors of ones and zeros. ℓ_p norm is referred as $||x||_p = (\sum_i |x_i|^p)^{\frac{1}{p}}$, and the norm of a linear operator A is $||A|| = sup_{||x||=1} ||Ax||$. Id stands for the identity operator and the identity matrix $Id_n = diag(\mathbf{1}_n)$, while the operator $diag(x)$ stands for a square matrix whose diagonal is x. A histogram with n bins is a vector $h \in \mathbb{R}^n_+$ with non-negative entries. The set $S_{m,n} := \{x \in \mathbb{R}^n_+, \langle x, \mathbf{1}_n \rangle = m\}$ is the simplex of histogram vectors of total mass m, therefore $S_{1,n}$ is the n-dimensional discrete probability simplex of $\mathbb{R}^n$. The Kronecker δ symbol is $\delta_{i,j} = 1$ if $i = j$, and $\delta_{i,j} = 0$ otherwise. The operators Prox and Proj are denoted as the Euclidean proximity and projection operators, such that $Prox_f(x) = argmin_y \frac{1}{2}||y - x||^2 + f(x)$ and $Proj_S(x) = argmin_{y \in S} ||y - x|| = Prox_{X_S}(x)$. The indicator and characteristic functions of a set S are: $\mathbb{1}_S(x) = \begin{cases} 1 & ifx \in S \\ 0 & otherwise \end{cases}$, $\iota_S(x) = \begin{cases} 0 & ifx \in S \\ \infty & otherwise \end{cases}$.

First lets consider the global histogram-based binary segmentation between two parts of a greyscale image. Let Ω be the N-pixel image domain, note N as the size of Ω (N $= |\Omega|$). Let $I : \Omega \mapsto \Lambda \subset \mathbb{R}^d$ be the image. h^0 and h^1 are two given reference histograms with $\sum_{\lambda\in\Lambda} h^i(\lambda) = 1, i = 0, 1$. the author defines the binary segmentation represented by $u : \Omega \mapsto \{0, 1\}$, saying that the histogram on $\Omega_0 := \{x \in \Omega, u(x) = 0\}$ is close to h^0 and the histogram on $\Omega_1 := \{x \in \Omega, u(x) = 1\}$ is close to h^1. Then compute the histogram on the region Ω_1 by:

$$h_u(\lambda) = \frac{1}{|\Omega_1|} \sum_{x\in\Omega} u(x) \nVdash_{I=\lambda}(x) = \frac{1}{\sum_{x\in\Omega} u(x)} \sum_{x\in\Omega} u(x) \nVdash_{I=\lambda}(x) \tag{1}$$

A metric between histograms and a norm $||.||$ on $\mathbb{R}^\Lambda$ is needed to solve the segmentation problem. The total variation regularization is considered to determine the interface length between two partitions. Therefore, the image is segmented by minimizing the following non-convex energy over the set $\{0, 1\}^N$:

$$J(u) = TV(u) + ||(h_u - h^1)_{\lambda\in\Lambda}|| + ||(h_{1-u} - h^0)_{\lambda\in\Lambda}|| \tag{2}$$

where $TV(u)$ is the total variation norm of the binary image u, relating to the perimeter of $\Omega_1 := \{x \in \Omega, u(x) = 1\}$. To handle the problem of energy minimization, some relaxations and reformulations are needed. The first step of relaxation is using a weight function (probability map) as a segmentation variable $u : \Omega \mapsto [0, 1]$. Therefore, a threshold can be applied to segment the image by $\Omega_t(u) := \{x \in \Omega | u(x) \geqslant t\}$.

In order to define a convex model, the data term is reformulated to compare histograms:

$$\begin{aligned} ||h_u - h^1|| &= \left|\left| \left(\frac{1}{\sum_\Omega u(x)} \sum_\Omega u(x) \nVdash_{I=\lambda}(x) - h^1(\lambda) \right)_{\lambda\in\Lambda} \right|\right| \\ &= \left|\left| \frac{1}{\sum_\Omega u(x)} \left(\sum_\Omega u(x) \nVdash_{I=\lambda}(x) - \left(\sum_\Omega u(x) \right) h^1(\lambda) \right)_{\lambda\in\Lambda} \right|\right| \end{aligned} \tag{3}$$

If assume that $|\Omega_1| = \sum_\Omega u(x)$ is known, then the distance is obtained:

$$\left|\left| \left(\sum_\Omega u(x)((1)_{I=\lambda}(x) - h^1(\lambda)) \right)_{\lambda\in\Lambda} \right|\right| \tag{4}$$

Notice that the distance is convex in u. A weighting factor $\beta \in [0, 1]$ is needed to balance the data term of two partitions after normalizations, such that the ratio $\beta = \frac{\sum_\Omega u(x)}{|\Omega|} = \frac{|\Omega_1|}{|\Omega|}$.

Denoting $g^1_\lambda(x) := \nVdash_{I=\lambda}(x) - h^1(\lambda)$, and $g^0_\lambda(x) := \nVdash_{I=\lambda}(x) - h^0(\lambda)$, observe the final convex model as:

$$J(u) = TV(u) + \frac{1}{\beta} ||(\langle u, g^1_\lambda \rangle_\Omega)_{\lambda\in\Lambda}|| + \frac{1}{1-\lambda} ||(\langle 1-u, g^0_\lambda \rangle_\Omega)_{\lambda\in\Lambda}|| \tag{5}$$

The rest is choosing the optimal transport distance to compare histograms.

Now lets consider $I : x \in \Omega \mapsto I(x) \in \mathbb{R}^d$ as a color image, then a feature image $FI(x) \in \mathbb{R}^n$ is introduced to rewrite the equation (1):

$$h(u) : y \in \mathbb{R}^n \mapsto \frac{1}{\sum_{x\in\Omega} u(x)} \sum_{x\in\Omega} u(x) \delta_{FI(x)}(y) \tag{6}$$

where $h(u)$ is the empirical discrete probability distribution of features FI using the binary map u, and F is the feature-transform of n-dimensional descriptors. By introducing the feature image, denote $H_X(u)$ as the quantized, non-normalized, and weighted feature histogram, with the relaxed variable $u : \Omega \mapsto [0,1]$ and the feature set $X = \{X_i \in \mathbb{R}^n\}_{1 \leqslant i \leqslant M_X}$, $H_X(u)$ write as:

$$(H_X(u))_i = \sum_{x \in \Omega} u(x) \mathbb{1}_{C_X(i)}(FI(x)), \ \forall i \in \{1, ...M_X\} \tag{7}$$

denoting that i as a bin index of M_X bins, X_i as the centroid of the corresponding bin, and $C_X(i) \subset \mathbb{R}^n$ as the corresponding set of features. So the H_X can be rewrite as a linear operator:

$$H_X : u \in \mathbb{R}^N \mapsto \mathbb{1}_X u \in \mathbb{R}^{M_X}, with \mathbb{1}_X(i,j) := 1 \ if \ FI(j) \in C_X(i), 0 \ otherwise \tag{8}$$

Notice that $\mathbb{1}_X \in \mathbb{R}^{M_X \times N}$ indicates which pixels of FI contribute to bin index i of the histogram H_X. Therefore, $\langle H_X(u), \mathbf{1}_X \rangle = \sum_{x \in \Omega} u(x) = \langle u, \mathbf{1}_N \rangle$, so that $H_X(u) \in S_{M_X, \langle u, 1 \rangle}$. Now rewrite equation (5) to find the minimum of segmentation energy using ℓ_1 distance:

$$J(u) = \rho TV(u) + \frac{1}{\beta} ||a\langle u, \mathbf{1}_N \rangle - H_A u||_1 + \frac{1}{N - \gamma} ||b\langle \mathbf{1}_N - u, \mathbf{1}_N \rangle - H_B(\mathbf{1}_N - u)||_1 \tag{9}$$

Considering the discrete probability segmentation map, the problem can be constrained as:

$$min_{u \in [0,1]^N} J(u) = min_{u \in \mathbb{R}^N} \left\{ J(u) := J(u) + \iota_{[0,1]^N}(u) \right\} \tag{10}$$

3 Monge-Kantorovitch distance

Let a, b be a pair of histograms such that $a \in S_{M_a,k}$ and $b \in S_{M_b,k}$, consider the Monge-Kantorovitch optimal transport problem as the discrete formulation between a and b. Note $C_{A,B} \in \mathbb{R}^{M_a \times M_b}$ as a fixed assignment cost matrix between the corresponding histogram centroids $A = \{A_i\}_{1 \leqslant i \leqslant M_a}$ and $B = \{B_j\}_{1 \leqslant j \leqslant M_b}$, defining the sets of admissible histogram and transport matrices as:

$$S := \left\{ a \in \mathbb{R}^{M_a}, b \in \mathbb{R}^{M_b} | a > 0, b > 0 \ and \ \langle a, \mathbf{1}_{M_a} \rangle = \langle b, \mathbf{1}_{M_b} \rangle \right\} \tag{11}$$

$$P(a,b) := \left\{ P \in \mathbb{R}_+^{M_a \times M_b}, P\mathbf{1}_{M_b} = a \ and \ P^T \mathbf{1}_{M_a} = b \right\} \tag{12}$$

Now the optimal transport plan is obtained to minimize the global transport cost, note as:

$$\forall (a,b) \in S, \ \mathbf{MK}(a,b) := min_{P \in P(a,b)} \left\{ \langle P, C \rangle = \sum_{i=1}^{M_a} \sum_{j=1}^{M_b} P_{i,j} C_{i,j} \right\} \tag{13}$$

For readability and the use of duality, it can be reformulated to:

$$\forall a, b \ , \mathbf{MK}(a,b) = min_{P \in P(a,b)} \langle P, C \rangle + \iota_S(a,b) \tag{14}$$

4 Sinkhorn distance

The definition of Sinkhorn distance is $d_{M,a}(r,c) := min_{P \in U_a(r,c)} \langle P, M \rangle$. By consider an entropic constraint in optimal transport, Sinkhorn distance provides computational method and restrict the low cost joint probabilities. Consider the entropy-regularized optimal transport problem on set S and rewrite the equation (14) as:

$$\mathbf{MK}_\lambda(a,b) := min_{P \in P(a,b)} \left\{ \langle P, C \rangle - \frac{1}{\lambda} h(P) \right\} + \iota_S(a,b) \tag{15}$$

It can be read as:

$$\mathbf{MK}_\lambda(\alpha) := min_{P \in \mathbb{R}^{M_a M_b}_{s.t. p \geqslant 0, L^T p = \alpha}} \langle p, c + \frac{1}{\lambda} log(p) \rangle + \iota_S(\alpha) \tag{16}$$

As Cuturi(2013) writes the Lagrangian of such problem with $\beta = \begin{bmatrix} u \\ v \end{bmatrix}$ to the constraint $L^T p = \alpha$, now the respective solution p^*_λ and P^*_λ from equation (15) and (16) can be write as:

$$\log p^*_\lambda = \lambda(L\beta - c) - 1 \Leftrightarrow (\log P^*_\lambda)_{i,j} = \lambda(u_i + v_i - C_{i,j}) - 1 \tag{17}$$

Sinkhorn proves the alternate normalization of rows and columns of any positive matrix M converges to a unique bistochastic matrix $P = diag(x)Mdiag(y)$. Therefore, the solution P^*_λ can be found by a fixed-point iteration algorithm with setting $M_\lambda = e^{-\lambda C}$:

$$P^*_\lambda = diag(x^\infty)M_\lambda diag(y^\infty) \; where \; x^{k+1} = \frac{a}{M_\lambda y^k} \; and \; y^{k+1} = \frac{b}{M^T_\lambda x^k} \tag{18}$$

where a and b are the matrix marginals as desired. Hence, the Sinkhorn distance or the derivatives can be used to design algorithms to compute the regularized optimal transportation.

Considering the set S does not prescribe histogram sums as admissible histograms, the histograms' total mass can be bounded above by N ($N = |\Omega|$) by alternative setting:

$$S_{\leqslant N} := \left\{ a \in \mathbb{R}^{M_a}, b \in \mathbb{R}^{M_b} | a > 0, b > 0, \langle a, \mathbf{1}_{M_a} \rangle = \langle b, \mathbf{1}_{M_b} \rangle \leqslant N \right\} \tag{19}$$

A normalized variant of the entropic regularization is proposed as the transport matrix P^*_λ is not normalized:

$$h(p) := Nh(\frac{p}{N}) = -N\mathbf{KL}(\frac{p}{N}||\mathbf{1}) = -\langle p, \log p \rangle + \langle p, \mathbf{1} \rangle \log N \tag{20}$$

5 Co-segmentation

The framework in this work can also be extended to unsupervised co-segmentation in multiple images. Now considering two images I^1 and I^2 with respectively domains Ω_1 and Ω_2. The image segmentation problem is converted to jointly segment a common object among all the images without priority. To solve this problem, the goal is to find the largest regions with similar feature distributions. The model is as following denoting $u = (u^1; u^2)$:

$$J(u) := ||H_1 u^1 - H_2 u^2||_1 + \sum_{k=1}^{2} \rho TV(u^k) - \delta ||u^k||_1 \tag{21}$$

6 Proposition

Proposition 1 (Cuturi-Doucet). *The convex conjugate of* $\mathbf{MK}_\lambda(\alpha)$ *reads*

$$\mathbf{MK}^*_\lambda(\beta) = \frac{1}{\lambda}\langle Q_\lambda(\beta), \mathbf{1} \rangle \; with \; Q_\lambda(\beta) := e^{\lambda(L\beta - c) - \mathbf{1}} \tag{22}$$

Corollary 1. *The convex conjugate of the normalized Sinkhorn distance*

$$\mathbf{MK}_{\lambda, \leqslant N}(\alpha) := min_{p \in \mathbb{R}^{M_a M_b}_{s.t. p \geqslant 0, l^T p = \alpha}} \left\{ \langle p, c + \frac{1}{\lambda} \log p - \frac{\log N}{\lambda} \mathbf{1} \rangle \right\} + \iota_{S \leqslant N}(\alpha) \tag{23}$$

reads, using the matrix-valued function $Q_\lambda(.) \mapsto e^{\lambda(L-c)-1}$ *defined in (19)*

$$\mathbf{MK}^*_{\lambda, \leqslant N}(\beta) = \begin{cases} \frac{N}{\lambda}\langle Q_\lambda(\beta), \mathbf{1} \rangle & if \langle Q_\lambda(\beta), \mathbf{1} \rangle \leqslant 1 \\ \frac{N}{\lambda} \log\langle Q_\lambda(\beta), \mathbf{1} \rangle + \frac{N}{\lambda} & if \langle Q_\lambda(\beta), \mathbf{1} \rangle \geqslant 1 \end{cases} \tag{24}$$

References

[1] N. Papadakis et al. "Convex Histogram-Based Joint Image Segmentation with Regularized Optimal Transport Cost." https://arxiv.org/abs/1610.01400.

[2] J. Rabin et al. "Convex Color Image Segmentation with Optimal Transport Distances."https://arxiv.org/abs/1503.01986.

[3] R. Yildizoglu et al. "A convex formulation for global histogram based binary segmentation." https://hal.archives-ouvertes.fr/hal-00834068/document.

[4] P. Swoboda et al. "Variational Image Segmentation and Cosegmentation with the Wasserstein Distance." https://ipa.math.uni-heidelberg.de/dokuwiki/Papers/WassersteinCoSegmentation13.pdf

[5] M. Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport." https://arxiv.org/abs/1306.0895.

[6] M. Cuturi et al. "Fast Computation of Wasserstein Barycenters." https://arxiv.org/abs/1310.4375

[7] M. Cuturi et al. "A Smoothed Dual Approach for Variational Wasserstein Problems." https://arxiv.org/abs/1503.02533

[8] Marco Cuturi's webpage. http://marcocuturi.net.

[9] http://marcocuturi.net/dagstuhl.pdf.