HAL
open science

Convex Color Image Segmentation with Optimal Transport Distances

Julien Rabin, Nicolas Papadakis

To cite this version:

Julien Rabin, Nicolas Papadakis. Convex Color Image Segmentation with Optimal Transport Distances. International Conference on Scale Space and Variational Methods in Computer Vision (SSVM'15), May 2015, Lège Cap Ferret, France. pp.256-269. hal-01133447

HAL Id: hal-01133447

https://hal.science/hal-01133447

Submitted on 19 Mar 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Convex Color Image Segmentation with Optimal Transport Distances

Julien Rabin[1] and Nicolas Papadakis[2]

[1] GREYC, Université de Caen, CNRS UMR 6072, France
julien.rabin@unicaen.fr
[2] CNRS, IMB, UMR 5251, Université de Bordeaux, France
nicolas.papadakis@math.u-bordeaux.fr

Abstract. This work is about the use of regularized optimal-transport distances for convex, histogram-based image segmentation. In the considered framework, fixed exemplar histograms define a prior on the statistical features of the two regions in competition. In this paper, we investigate the use of various transport-based cost functions as discrepancy measures and rely on a primal-dual algorithm to solve the obtained convex optimization problem.

Keywords: Optimal transport, Wasserstein distance, Sinkhorn distance, convex optimization, image segmentation

1 Introduction

Optimal transport Optimal transport theory has received a lot of attention during the last decade as it provides a powerful framework to address problems which embed statistical constraints. Its successful application in various image processing tasks has demonstrated its practical interest (see *e.g.* [7,8,11,5]). Some limitations have been also shown and partially addressed, such as time complexity, regularity and relaxation [1,4].

Segmentation Statistical based image segmentation has been thoroughly studied in the literature, first using parametric models (such as the mean and variance), and then empirical distributions combined with adapted statistical distances, such as the Kullback-Leibler divergence. In this work, we are interested in the use of the optimal transport framework for Image segmentation. This has been first investigated in [7] for 1D features, then extended to multi-dimensional features using approximations of the optimal transport cost [5,9], and adapted to region-based active contour in [9], relying on a non-convex formulation. In [12], a convex formulation is proposed, making use of sub-iterations to compute the proximity operator of the Wasserstein distance, which use is restricted to low dimensions.

In this paper, we extend the convex formulation for two-phase image segmentation of [14] for non-regularized as well as regularized [1,2] optimal transport distances. This work shares some common features with the recent work of [3] in which the authors investigate the use of the Legendre-Fenchel transform of regularized transport cost for imaging problems.

2 Convex histogram-based image segmentation

2.1 Notation

We consider here vector spaces equipped with the scalar product $\langle .\,,\,. \rangle$ and the norm $\|.\| = \sqrt{\langle .\,,\,. \rangle}$. The conjugate operator of A is denoted by A^* and satisfies $\langle Ax, y \rangle = \langle x, A^* y \rangle$. We denote as $\mathbf{1}_n$ and $\mathbf{0}_n \in \mathbb{R}^n$ the n-dimensional vectors full of ones and zeros respectively, x^T the transpose of x, and ∇ the discrete gradient operator, while Id stands for the identity operator. The operator $\mathrm{diag}(x)$ defines a square matrix whose diagonal is x. Functions ι_S and $\mathbb{1}_S$ are respectively the characteristic and indicator functions of a set S. *Proj* and *Prox* stands respectively for the Euclidean projection and proximity operator. The set $\mathcal{S}_{k,n} := \{x \in \mathbb{R}^n_+, \langle x,\, \mathbf{1}_n \rangle = k\}$ is the simplex of histogram vectors ($\mathcal{S}_{1,n}$ being therefore the discrete probability simplex of $\mathbb{R}^n$).

2.2 General formulation of distribution-based image segmentation

Let $I : x \in \Omega \mapsto I(x) \in \mathbb{R}^d$ be a color image, defined over the N-pixel domain Ω ($N = |\Omega|$), and $\mathcal{F}$ a feature-transform of n-dimensional descriptors $\mathcal{F}I(x) \in \mathbb{R}^n$. We would like to define a binary segmentation $u : \Omega \mapsto \{0, 1\}$ of the whole image domain, using two fixed probability distributions of features a and b. Following the variational model introduced in [14], we consider the energy

$$J(u) = \rho\, TV(u) + D(a, h(u)) + D(b, h(\mathbf{1} - u)) \tag{1}$$

where $\rho \geq 0$ is the regularization parameter,

- the fidelity terms are defined using $D(.,.)$, a dissimilarity measure between features;
- $h(u)$ is the empirical discrete probability distribution of features $\mathcal{F}I$ using the binary map u, which is written as a sum of Dirac masses

$$h(u) : y \in \mathbb{R}^n \mapsto \frac{1}{\sum_{x \in \Omega} u(x)} \sum_{x \in \Omega} u(x) \delta_{\mathcal{F}I(x)}(y)\ ;$$

- $TV(u)$ is the total variation norm of the binary image u, which is related to the perimeter of the region $R_1(u) := \{x \in \Omega \,|\, u(x) = 1\}$ (co-area formula).

Observe that this energy is highly non-convex since h is a non linear operator, and that we would like to find a minimum over the non-convex set $\{0, 1\}^N$.

2.3 Convex relaxation of histogram-based segmentation energy

The authors of [14] propose some relaxations and a reformulation in order to handle the minimization of energy (1) using convex optimization tools.

Probability map The first relaxation consists in using a segmentation variable $u : \Omega \mapsto [0, 1]$ which is a weight function (probability map). A threshold is therefore required to get a binary segmentation of the image into two regions $R_t(u) := \{x \in \Omega \,|\, u(x) \geq t\}$ and its complement $R_t(u)^c$.

Feature histogram The feature histogram of the probability map is denoted $H_{\mathcal{X}}(u)$ and defined as the **quantized, non-normalized, and weighted histogram** of the feature image $\mathcal{F}I$ using the relaxed variable $u : \Omega \mapsto [0,1]$ and a feature set $\mathcal{X} = \{X_i \in \mathbb{R}^n\}_{1 \leq i \leq M_X}$ composed of M_X bins

$$(H_{\mathcal{X}}(u))_i = \sum_{x \in \Omega} u(x) \mathbb{1}_{\mathcal{C}_{\mathcal{X}}(i)}(\mathcal{F}I(x)), \quad \forall\, i \in \{1, \dots M_X\}$$

where i a bin index, X_i is the centroid of the corresponding bin, and $\mathcal{C}_{\mathcal{X}}(i) \subset \mathbb{R}^n$ is the corresponding set of features (*e.g.* the Voronoï cell obtained from *hard assignment* method). Therefore, we can write $H_{\mathcal{X}}$ as a linear operator

$$H_{\mathcal{X}} : u \in \mathbb{R}^N \mapsto \mathbb{1}_{\mathcal{X}} \cdot u \in \mathbb{R}^{M_X}, \quad \text{with } \mathbb{1}_{\mathcal{X}}(i,j) := 1 \text{ if } \mathcal{F}I(j) \in \mathcal{C}_{\mathcal{X}}(i), 0 \text{ otherwise.}$$

Note that $\mathbb{1}_{\mathcal{X}} \in \mathbb{R}^{M_X \times N}$ is a fixed assignment matrix that indicates which pixels of $\mathcal{F}I$ contribute to each bin i of the histogram. As a consequence, $\langle H_{\mathcal{X}}(u),\, \mathbf{1}_X \rangle = \sum_{x \in \Omega} u(x) = \langle u,\, \mathbf{1}_N \rangle$, so that $H_{\mathcal{X}}(u) \in \mathcal{S}_{M_X, \langle u,\, \mathbf{1} \rangle}$.

Exemplar histograms The segmentation is driven from two fixed histograms $a \in \mathcal{S}_{M_a,1}$ and $b \in \mathcal{S}_{M_b,1}$, which are normalized (*i.e.* sum to 1), have respective dimension M_a and M_b, and are obtained using the respective sets of features $\mathcal{A}$ and $\mathcal{B}$. In order to measure the similarity between the non-normalized histogram $H_{\mathcal{A}} u$ and the normalized histogram a, while obtaining a convex formulation, we follow [14] and consider the fidelity term $D\left(a\langle u,\, \mathbf{1}_N \rangle, H_{\mathcal{A}} u\right)$, where the constant vector a has been scaled to $H_{\mathcal{A}} u \in \mathcal{S}_{M_a, \langle u,\, \mathbf{1} \rangle}$.

Segmentation energy Observe that the problem can now be written as finding the minimum of the following energy

$$\tilde{E}(u) = \rho\, TV(u) + \tfrac{1}{\gamma} D\left(a\langle u,\, \mathbf{1}_N \rangle, H_{\mathcal{A}} u\right) + \tfrac{1}{N-\gamma} D\left(b\langle \mathbf{1}_N - u,\, \mathbf{1}_N \rangle, H_{\mathcal{B}}(\mathbf{1}_N - u)\right).$$

The constant $\gamma \in (0, N)$ is meant to compensate for the fact that the binary regions $R_t(u)$ and $R_t(u)^c$ may have different size. More precisely, as we are interested in a discrete probability segmentation map, we consider the following constrained problem:

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

Simplification From now on, and without loss of generality, we will assume that all histograms are computed using the same set of features, *namely* $\mathcal{A} = \mathcal{B}$. We will also omit unnecessary subscripts in order to simplify notation. Moreover, we also omit the parameter γ since its value seems not to be critical in practice, as demonstrated in [14]. Finally, introducing linear operators

$$A := a\mathbf{1}_N^T \in \mathbb{R}^{M \cdot N} \quad \text{and} \quad B := b\mathbf{1}_N^T \in \mathbb{R}^{M \cdot N} \tag{2}$$

such that $Au = (a\mathbf{1}_N^T)u = a\langle u, \mathbf{1}_N \rangle$, we have the following minimization problem:

$$\min_u \ \rho \|\nabla u\| + D(Au, Hu) + D(B(1-u), H(1-u)) + \iota_{[0,1]^N}(u). \tag{3}$$

Notice that matrix $H \in \mathbb{R}^{M \cdot N}$ is sparse (with N non zero values) and A and B are of rank 1, so that storing or manipulating these matrices is not an issue.

In [14], the distance function D was defined as the L_1 norm. In the following sections, we investigate the use of similarity measure based on optimal transport, which is known to be more robust and appropriate for histogram comparison. The next paragraph details the optimization framework used in this work.

2.4 Optimization

In order to solve (3), we consider the following dualization of the problem using the Legendre-Fenchel transforms of the L^2 norm and the function D

$$\min_{u \in \mathbb{R}^N} \max_{\substack{p_A, q_A, p_B, q_B \in \mathbb{R}^M \\ p_C \in \mathbb{R}^{2N}}} \langle Hu, p_A \rangle + \langle Au, q_A \rangle + \langle H(\mathbf{1} - u), p_B \rangle + \langle B(\mathbf{1} - u), q_B \rangle + \langle \nabla u, p_C \rangle \\ + \iota_{[0,1]^N}(u) - D^*(p_A, q_A) - D^*(p_B, q_B) - \iota_{\|\cdot\| \leqslant \rho}(p_C), \tag{4}$$

where $\iota_{\|\cdot\| \leqslant \rho}$ is the characteristic function of the convex ℓ_2 ball of radius ρ, while D^* is the dual of the function D. In order to accommodate the different models studied in this paper, we assume here that D^* is a sum of two convex functions $D^* = D_1^* + D_2^*$, where D_1^* is non-smooth and D_2^* is differentiable and has a Lipschitz continuous gradient.

We recover a general primal-dual problem of the form

$$\min_u \max_p \langle Ku, p \rangle + \iota_{[0,1]^N}(u) + H(u) - F^*(p) - G^*(p), \tag{5}$$

with primal variable $u \in \mathbb{R}^N$ and dual vector $p = [p_A^T, q_A^T, p_B^T, q_B^T, p_C^T]^T \in \mathbb{R}^{4M+2N}$, where

- $K = [H^T, A^T, -H^T, -B^T, \nabla^T]^T \in \mathbb{R}^{(4M+2N) \times N}$ is a sparse, linear operator;
- H is convex and smooth ($H(u) = 0$ in the setting of problem (5)) with Lipschitz continuous gradient ∇H with constant L_H;
- $\iota_{[0,1]^N}(u)$ is convex and non-smooth;
- $F^*(p) = D_1^*(p_A, q_A) + D_1^*(p_B, q_B) + \iota_{\|\cdot\| \leqslant \rho}(p_C)$ is convex and non-smooth;
- $G^*(p) = D_2^*(p_A, q_A) + D_2^*(p_B, q_B) - \langle H\mathbf{1}_N, p_B \rangle - \langle B\mathbf{1}_N, q_B \rangle$ is convex and differentiable with Lipschitz constant L_{G^*}.

To solve this problem, we consider the precondionned primal dual algorithm of [6]

$$\begin{cases} u^{k+1} = \mathrm{Proj}_{[0,1]^N}\left(u^k - \tau(K^T p^k + \nabla H(u^k))\right) \\ p^{k+1} = \mathrm{Prox}_{\sigma F^*}\left(p^k + \sigma(K(2u^{k+1} - u^k) - \nabla G^*(p^k))\right) \end{cases} \tag{6}$$

that converges to a saddle point of (5) as soon as (see for instance [6])

$$\left(\tfrac{1}{\tau} - L_H\right)\left(\tfrac{1}{\sigma} - L_{G^*}\right) \geqslant \|K\|^2. \tag{7}$$

3 Monge-Kantorovitch distance for image segmentation

3.1 Wasserstein Distance and Optimal Transport problem

Optimal Transport problem We consider in this work the discrete formulation of the Monge-Kantorovitch optimal mass transportation problem (see *e.g.* [13]) between a pair of histograms $a \in \mathcal{S}_{M_a,k}$ and $b \in \mathcal{S}_{M_b,k}$. Given a fixed assignment cost matrix $C_{\mathcal{A},\mathcal{B}} \in \mathbb{R}^{M_a \times M_b}$ between the corresponding histogram centroids $\mathcal{A} = \{A_i\}_{1 \le i \le M_a}$ and $\mathcal{B} = \{B_j\}_{1 \le j \le M_b}$, an optimal transport plan minimizes the global transport cost, defined as a weighted sum of assignments

$$\forall\, (a,b) \in \mathcal{S}, \quad \mathbf{MK}(a,b) := \min_{P \in \mathcal{P}(a,b)} \left\{ \langle P,\, C \rangle = \sum_{i=1}^{M_a} \sum_{j=1}^{M_b} P_{i,j} C_{i,j} \right\}. \tag{8}$$

The sets of admissible histogram and transport matrices are respectively

$$\mathcal{S} := \{a \in \mathbb{R}^{M_a}, b \in \mathbb{R}^{M_b} \mid a \ge 0, b \ge 0 \text{ and } \langle a,\, \mathbf{1}_{M_a} \rangle = \langle b,\, \mathbf{1}_{M_b} \rangle\}, \tag{9}$$

$$\mathcal{P}(a,b) := \{P \in \mathbb{R}_+^{M_a \times M_b}, P\mathbf{1}_{M_b} = a \text{ and } P^T \mathbf{1}_{M_a} = b\}. \tag{10}$$

Observe that the norm of histograms is not prescribed in $\mathcal{S}$, and that we only consider histograms with positive entries since null entries do not play any role.

Wasserstein distance When using $C_{i,j} = \|A_i - B_j\|^p$, then $\mathbf{W}_p(a,b) = \mathbf{MK}(a,b)^{1/p}$ is a metric between normalized histograms. In the general case where C does not verify such a condition, by a slight abuse of terminology we will refer to the **MK** transport cost function as the Monge-Kantorovich *distance*.

Monge-Kantorovich distance In the following, due to the use of duality, it would be more convenient to introduce the following reformulation:

$$\forall\, a, b \qquad \mathbf{MK}(a,b) = \min_{P \in \mathcal{P}(a,b)} \langle P,\, C \rangle + \iota_{\mathcal{S}}(a,b). \tag{11}$$

LP formulation We can rewrite the optimal transport problem as a linear program (LP) with vector variables. The primal and dual problems write

$$\mathbf{MK}(\alpha) = \min_{\substack{p \in \mathbb{R}^{M_a \cdot M_b} \\ \text{s.t. } p \ge 0,\ L^T p = \alpha}} \langle c,\, p \rangle + \iota_{\mathcal{S}}(\alpha) \;=\; \max_{\substack{\beta \in \mathbb{R}^{M_a + M_b} \\ \text{s.t. } L\beta \le c}} \langle \alpha,\, \beta \rangle. \tag{12}$$

where α is the concatenation of histograms: $\alpha^T = [a^T, b^T]$ and the unknown vector $p \in \mathbb{R}^{M_a \cdot M_b}$ corresponds to the bi-stochastic matrix P being read column-wise (*i.e.* $P_{i,j} = p_{i+(j-1)\cdot M_a}$). The $M_a + M_b$ linear marginal constraints on p are defined by the matrix $L^T \in \mathbb{R}^{(M_a+M_b)\times(M_a M_b)}$ through equation $L^T p = \alpha$, where

$$L^T = \begin{bmatrix} \mathbf{1}_{M_b} e_1^T & \mathbf{1}_{M_b} e_2^T & \cdots & \mathbf{1}_{M_b} e_{M_a}^T \\ \mathrm{Id}_{M_b} & \mathrm{Id}_{M_b} & \cdots & \mathrm{Id}_{M_b} \end{bmatrix} \quad \text{with} \quad e_i(j) = \delta_{i-j} \;\; \forall\, j \le M_b.$$

Note that we have the following property: $(L\alpha)_{i,j} = \left(L \left[\begin{smallmatrix} a \\ b \end{smallmatrix}\right]\right)_{i,j} = a_i + b_j$.

The dual formulation shows that the function $\mathbf{MK}(\alpha)$ is not strictly convex in α. We draw the reader's attention to the fact that the indicator of set $\mathcal{S}$ is not required anymore with the dual formulation, which will later come in handy.

Dual distance From Eq. (12), we have that the Legendre–Fenchel conjugate of **MK** writes simply as the characteristic function of the set $\mathcal{L}_c := \{\beta \,|\, L\beta - c \leq 0\}$

$$\forall \beta \in \mathbb{R}^{M_a+M_b}, \quad \mathbf{MK}^*(\beta) = \iota_{L\beta \leqslant c}(\beta). \tag{13}$$

3.2 Integration in the segmentation framework

We propose to substitute in problem (3) the dissimilarity functions by the convex Monge-Kantorovich optimal transport cost (11).

In order to apply our minimization scheme described in (6), we should be able to compute the proximity operator of $\mathbf{MK}^*$, which is the projection onto the convex set $\mathcal{L}_c$. However, because the linear operator L is not invertible, we cannot compute this projector in a closed form and an optimization problem should be solved at each iteration of the process (6) as in [12].

Bidualization To circumvent this problem, we resort to a bidualization to rewrite the **MK** distance as a primal-dual problem. First, we have that $\mathbf{MK}^*(\beta) = f^*(L\beta)$ with $f^*(r) = \iota_{r \leq c}(r)$, so that $f(r) = \langle r,\, c\rangle + \iota_{r\geq 0}(r)$. Then,

$$\begin{aligned}
\mathbf{MK}^*(\beta) &= f^*(L\beta) = \max_r \langle r,\, L\beta\rangle - f(r) = \max_r \langle r,\, L\beta - c\rangle - \iota_{\cdot\geq 0}(r) \\
\mathbf{MK}(\alpha) &= \max_\beta \langle \alpha,\, \beta\rangle - f^*(L\beta) = \max_\beta \langle \alpha,\, \beta\rangle + \min_r \langle r,\, c - L\beta\rangle + \iota_{\cdot\geq 0}(r) \\
&= \min_r \max_\beta \langle r,\, c\rangle + \iota_{\cdot\geq 0}(r) + \langle \alpha - L^T r,\, \beta\rangle.
\end{aligned} \tag{14}$$

Segmentation problem Plugging the previous expression into Eq. (4) enables us to solve it using algorithm (6). Indeed, introducing new primal variables $r_A, r_B \in \mathbb{R}^{M^2}$ related to transport mapping, we recover the following primal dual problem

$$\begin{aligned}
\min_{\substack{u\in\mathbb{R}^N \\ r_A, r_B \in \mathbb{R}^{M^2}}} \max_{\substack{p_A, q_A, p_B, q_B \in \mathbb{R}^M \\ p_C \in \mathbb{R}^{2N}}} & \langle Hu, p_A\rangle + \langle Au, q_A\rangle + \langle H(\mathbf{1}-u), p_B\rangle + \langle B(\mathbf{1}-u), q_B\rangle \\
& \langle r_A,\, c - L\begin{bmatrix} p_A \\ q_A \end{bmatrix}\rangle + \langle r_B,\, c - L\begin{bmatrix} p_B \\ q_B \end{bmatrix}\rangle + \langle \nabla u, p_C\rangle \\
& + \iota_{[0,1]^N}(u) + \iota_{\cdot\geq 0}(r_A) + \iota_{\cdot\geq 0}(r_B) - \iota_{\|\cdot\|\leqslant\rho}(p_C).
\end{aligned} \tag{15}$$

Observe that now we have a linear term $H(u, r_A, r_B) = \langle r_A + r_B,\, c\rangle$ whose gradient has a Lipschitz constant $L_H = 0$. We have also gained extra non smooth characteristic functions $\iota_{\cdot\geq 0}$, whose proximity operators are trivial (projection onto the positive quadrant $\mathbb{R}_+^{M^2}$: $\mathrm{prox}_{\iota_{\geq 0}}(x) = \max\{\mathbf{0}, x\}$).

Advantages and drawback The main advantage of this new segmentation framework is that it makes use of optimal transport to compare histograms of features, without sub-iterative routines such as solving optimal transport problems to compute sub-gradients or proximity operators (see for instance [1,12]), or without making use of approximation (such as the Sliced-Wasserstein distance [9],

generalized cumulative histograms [8] or entropy-based regularization [2]). Last, the proposed framework is not restricted to Wasserstein distances, since it enables the use of any cost matrix, and does not depend on features dimensionality.

However, a major drawback of this method is that it requires two additional primal variables r_A and r_B whose dimension is M^2 in our simplified setting, M being the dimension of histograms involved in the model. As soon as $M^2 \gg N$, the number of pixels, the proposed method could be significantly slower than when using L^1 as in [14] due to time complexity and memory limitation. This is more likely to happen when considering high dimensional features, such as patches or computer vision descriptors, as M increases with feature dimension n.

4 Regularized MK distance for image segmentation

As already mentioned in the last section, the previous approach based on optimal transport may be very slow for large histograms. In such a case, we propose to use instead the entropy smoothing of optimal transport recently proposed and investigated in [1,2,3], that may offer increased robustness to outliers [1]. While it has been initially studied for probability simplex $\mathcal{S}_1$, we here investigate its use for our framework with unnormalized histograms on $\mathcal{S}$.

4.1 Sinkhorn distances $\mathbf{MK}_\lambda$

The entropy-regularized optimal transport problem (11) on set $\mathcal{S}$ (Eq. (9)) is

$$\mathbf{MK}_\lambda(a,b) := \min_{P\in\mathcal{P}(a,b)} \left\{\langle P,\, C\rangle - \tfrac{1}{\lambda}h(P)\right\} + \iota_{\mathcal{S}}(a,b), \tag{16}$$

where the entropy of the matrix P is defined as $h(P) := -\langle P,\, \log P\rangle$. Thanks to the negative entropy term which is strictly convex, the regularized optimal transport problem has a unique minimizer, denoted $P^\star_\lambda$, which can be recovered using a fixed point algorithm studied by Sinkhorn (see *e.g.* [1]). The regularized transport cost $\mathbf{MK}_\lambda(a,b)$ is thus referred to as the *Sinkhorn distance*.

Interpretation Another way to express the negative entropic term is:

$$-h(p) : p \in \mathbb{R}^k_+ \mapsto \mathbf{KL}(p\|\mathbf{1}_k) \in \mathbb{R}, \qquad \text{with } k = M_a \cdot M_b$$

that is the Kullback-Leibler divergence between transport map p and the uniform mapping. This shows that, as λ decreases, the model encourages smooth, uniform transport so that the mass is spread everywhere. This also explains why this distance shows better robustness to outliers, as reported in [1]. To conclude, one thus would like to use in practice large values of λ to be close to the original Monge-Kantorovich distance, but low enough to deal with feature perturbation.

Structure of the solution First, the Sinkhorn distance (16) reads as

$$\mathbf{MK}_\lambda(\alpha) := \min_{\substack{p\in\mathbb{R}^{M_a\cdot M_b}\\ \text{s.t. } p\geq 0,\, L^T p=\alpha}} \langle p,\, c + \tfrac{1}{\lambda}\log p\rangle + \iota_{\mathcal{S}}(\alpha). \tag{17}$$

As demonstrated in [1], when writing the Lagrangian of this problem with a multiplier β to take into account the constraint $L^T p = \alpha$, we can show that the respective solutions $p^\star_\lambda$ and $P^\star_\lambda$ of problem (16) and (17) write

$$\log p^\star_\lambda = \lambda(L\beta - c) - \mathbf{1} \Leftrightarrow (\log P^\star_\lambda)_{i,j} = \lambda(u_i + v_i - C_{i,j}) - 1 \text{ with } \beta = \begin{bmatrix} u \\ v \end{bmatrix}.$$

Remark 1. The constant -1 is due to the fact that we use the unnormalized KL divergence $\mathbf{KL}(p\|\mathbf{1}_k)$, instead of $\mathbf{KL}(p\|\frac{1}{k}\mathbf{1}_k)$ for instance.

Sinkhorn algorithm Sinkhorn showed that the alternate normalization of rows and columns of any positive matrix M converges to a unique bistochastic matrix $P = \operatorname{diag}(x) M \operatorname{diag}(y)$. The following fixed-point iteration algorithm can thus be used to find the solution $P^\star_\lambda$: setting $M_\lambda = e^{-\lambda C}$, one has

$$P^\star_\lambda = \operatorname{diag}(x^\infty) M_\lambda \operatorname{diag}(y^\infty) \quad \text{where } x^{k+1} = \frac{a}{M_\lambda\, y^k} \text{ and } y^{k+1} = \frac{b}{M_\lambda^T\, x^k},$$

where a and b are the desired marginals of the matrix. This result enables us to design fast algorithms to compute the regularized optimal transport plan, and the the Sinkhorn distance or its derivative, as demonstrated in [1,2].

4.2 Legendre–Fenchel transformation of Sinkhorn distance MK_λ

Now, in order to use the Sinkhorn distance in algorithm (6), we need to compute its Legendre-Fenchel transform, which has been expressed in [2].

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

$$\boldsymbol{MK}^*_\lambda(\beta) = \tfrac{1}{\lambda}\langle Q_\lambda(\beta), \mathbf{1}\rangle \qquad \textit{with } Q_\lambda(\beta) := e^{\lambda(L\beta - c) - \mathbf{1}}. \tag{18}$$

We obtain a simple expression of the Legendre–Fenchel transform which is C^∞, but unfortunately, its gradient is not Lipschitz continuous.

To overcome this problem, we propose two solutions in the next paragraphs: either we use a new normalized Sinkhorn distance (§ 4.3), whose gradient is Lipschitz continuous (§ 4.4), or we rely on the use of proximity operator (§ 4.6).

4.3 Normalized Sinkhorn distance $\mathrm{MK}_{\lambda,\leq N}$ on $\mathcal{S}_{\leq N}$

As the set $\mathcal{S}$ of admissible histograms does not prescribe the sum of histograms, we consider here a different setting in which the histograms' total mass are bounded above by N, the number of pixels of the image domain Ω

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

Moreover, as the transport matrix $P^\star_\lambda$ is not normalized (*i.e.* $\langle P^\star_\lambda,\, \mathbf{1}\rangle \leq N$), we also propose to use a slightly normalized variant of the entropic regularization:

$$\tilde{h}(p) := N h\left(\tfrac{p}{N}\right) = -N\, \mathbf{KL}\left(\tfrac{p}{N}\|\mathbf{1}\right) = -\langle p,\, \log p\rangle + \langle p,\, \mathbf{1}\rangle \log N. \tag{20}$$

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

$$\boldsymbol{MK}_{\lambda,\leq N}(\alpha) := \min_{\substack{p\in\mathbb{R}^{M_a\cdot M_b}\\ s.t.\ p\geq 0,\ L^T p=\alpha}} \left\{\langle p,\ c+\tfrac{1}{\lambda}\log p - \tfrac{\log N}{\lambda}\mathbf{1}\rangle\right\} + \iota_{\mathcal{S}_{\leq N}}(\alpha) \tag{21}$$

reads, using the matrix-valued function $Q_\lambda(.) \mapsto e^{\lambda(L.-c)-\mathbf{1}}$ *defined in* (18)

$$\boldsymbol{MK}^*_{\lambda,\leq N}(\beta) = \begin{cases} \frac{N}{\lambda}\langle Q_\lambda(\beta),\mathbf{1}\rangle & \text{if } \langle Q_\lambda(\beta),\mathbf{1}\rangle \leqslant 1 \\ \frac{N}{\lambda}\log\langle Q_\lambda(\beta),\mathbf{1}\rangle + \frac{N}{\lambda} & \text{if } \langle Q_\lambda(\beta),\mathbf{1}\rangle \geqslant 1 \end{cases} \tag{22}$$

Proof. The proof [10, A.1] is omitted here for the sake of shortness.

Observe that the dual function $\mathbf{MK}^*_{\lambda,\leq N}(\beta)$ is continuous for $\langle Q_\lambda(\beta^\star),\ \mathbf{1}\rangle = 1$. Note also that the optimal matrix now is written $P^\star_\lambda = NQ_\lambda(\beta^\star)$ if $\langle Q_\lambda(\beta^\star),\ \mathbf{1}\rangle \leq 1$, and $P^\star_\lambda = N\frac{Q_\lambda(\beta^\star)}{\langle Q_\lambda(\beta^\star),\mathbf{1}\rangle}$ otherwise.

4.4 Gradient of $\mathbf{MK}^*_{\lambda,\leq N}$

From Corollary 1, we can express the gradient of $\mathbf{MK}^*_{\lambda,\leq N}$ which is continuous (writing Q in place of $Q_\lambda(\beta)$ to simplify expression)

$$\nabla\mathbf{MK}^*_{\lambda,\leq N}(\beta) = \begin{cases} N & \left(Q\mathbf{1}_{M_b}, \mathbf{1}^T_{M_a}Q\right) & \text{if } \langle Q,\mathbf{1}\rangle \leqslant 1 \\ \frac{N}{\langle Q,\mathbf{1}\rangle} & \left(Q\mathbf{1}_{M_b}, \mathbf{1}^T_{M_a}Q\right) & \text{if } \langle Q,\mathbf{1}\rangle \geqslant 1 \end{cases}. \tag{23}$$

We emphasis here that we retrieve a similar expression than the one originatively demonstrated in [3], where the authors consider the Sinkhorn distance on the probability simplex $\mathcal{S}_1$ (*i.e.* the special case where $N = 1$ and $\langle Q,\mathbf{1}\rangle = 1$).

Proposition 2. *The gradient* $\nabla\boldsymbol{MK}^*_{\lambda,\leq N}$ *is a Lipschitz continuous function of constant* $L_{\boldsymbol{MK}^*}$ *bounded by* $2\lambda N$.

Proof. The proof [10, A.2] is omitted here for the sake of shortness.

4.5 Optimization using $\nabla\mathbf{MK}^*_{\lambda,\leq N}$

The general final problem we want to solve can be expressed as:

$$\min_u \rho TV(u) + \mathbf{MK}_{\lambda,\leq N}(H_a u, Au) + \mathbf{MK}_{\lambda,\leq N}(H_b(\mathbf{1}-u), B(\mathbf{1}-u)) + \iota_{[0,1]^N}(u). \tag{24}$$

Using the Legendre–Fenchel transform, the problem (24) can be reformulated as:

$$\min_u \max_{\substack{p_A,q_A\\ p_B,q_B,p_C}} \langle H_a u, p_A\rangle + \langle Au, q_A\rangle + \langle H_b(\mathbf{1}-u), p_B\rangle + \langle B(\mathbf{1}-u), q_B\rangle + \langle \nabla u, p_C\rangle$$
$$+ \iota_{[0,1]^N}(u) - \mathbf{MK}^*_{\lambda,\leqslant N}(p_A,q_A) - \mathbf{MK}^*_{\lambda,\leqslant N}(p_B,q_B) - \iota_{\|.\|\leqslant\rho}(p_C),$$

and can be optimized with the algorithm (6). Using proposition 2, ∇G^* is a Lipschitz continuous function with constant L_{G^*} checking $L_{G^*} = 2L_{\mathbf{MK}^*} + \|H_b\| + \|B\| = 2\lambda N + \|H_b\| + \|B\|$, where N is the number of pixels. It will be large for high resolution images and huge for good approximations of the **MK** cost (*i.e.* $\lambda \gg 1$). Such a scheme may thus involve a very slow explicit gradient ascent in the dual update (6). In such a case, we can resort to the alternative scheme proposed in the next subsection.

4.6 Optimization using proximity operator of $\mathbf{MK}^*_{\boldsymbol{\lambda}^*}$

An alternative optimization of (24) consists in using the proximity operator of $\mathbf{MK}^*_\lambda$. Since we cannot compute the proximity operator of $\mathbf{MK}^*_\lambda$ in a closed form, we resort instead to a bidualization, as previously done in Section 3.2.

Considering now the normalized function $\mathbf{MK}_\lambda(\alpha)$ using entropy normalization (20) on set $\mathcal{S}$, we thus have $\mathbf{MK}^*_\lambda(\beta) = \frac{N}{\lambda}\langle Q_\lambda(\beta), \mathbf{1}\rangle = g^*_\lambda(L\beta)$.

Proposition 3. *The proximity operators of* $g^*_\lambda(q) = \frac{N}{\lambda}\langle e^{\lambda(q-c)-\mathbf{1}}, \mathbf{1}\rangle$ *is*

$$prox_{\tau g^*_\lambda}(p) = p - \frac{1}{\lambda} W\left(\lambda\tau N e^{\lambda(p-c)-\mathbf{1}}\right). \tag{25}$$

where W *is the* Lambert function*, such that* $w = W(z)$ *is solution of* $we^w = z$*. The solution is unique as* $z = \lambda\tau N e^{\lambda(p-c)-\mathbf{1}} \geqslant 0$*.*

Proof. The proof [10, A.3] is omitted here for the sake of shortness.

Remark 2. Note that the Lambert function can be evaluated very fast.

5 Experiments

Experimental setting In this experimental section, exemplar regions are defined by the user with scribbles (see Figures 1 to 5). These regions are only used to built prior histograms, so erroneous labeling is tolerated. Histograms a and b are built using hard-assignment on $M = 8^n$ clusters, which are obtained with the K-means algorithm. We use either RGB color ($\mathcal{F} = \mathrm{Id}$ and $n = d = 3$) or the gradient color norm ($\mathcal{F} = \|\nabla.\|$ and again $n = d = 3$) features. The cost matrix is defined from the Euclidean metric $\|\cdot\|$ in $\mathbb{R}^n$ space, combined with the concave function $1 - e^{-\gamma\|\cdot\|}$, which is known to be more robust to outliers. Region $R_t(u)$ is obtained with threshold $t = \frac{1}{2}$, as illustrated in Figure 1. Approximately 1 minute is required to run 500 iterations and segment a 1 Megapixel color image.

Results Figure 1 shows the influence of the threshold t used to get a binary segmentation. A small comparison with the model of [14] is then given in Figure 2. This underlines the robustness of optimal transport distance with respect to bin-to-bin L^1 distance. Contrary to optimal transport, when a color is not present in the reference histograms, the L^1 distance does not take into account the color distance between bins which can lead to incorrect segmentation. The robustness is further illustrated in Figure 3. It is indeed possible to use a prior histogram from a different image, even with a different clustering of the feature space. Note that it is not possible with a bin-to-bin metric, which requires the same clustering. Figure 4 shows comparisons between the non-regularized model, quite fast but high dimensional model, with the regularized model, using a low dimensional formulation. One can see that setting a large value of λ gives interesting results. On the other hand, using a very small value of λ always yields poor segmentation results. Some other results are proposed in the supplementary material (Section B).

Some last examples on texture segmentation are presented in Figure 5 where the proposed method is perfectly able to recover the textured areas.

6 Conclusion and future work

Several formulations have been proposed in this work to incorporate transport-based distances in convex variational model for image processing, using either regularization of the optimal-transport or not.

Different perspectives have yet to be investigated, such as the final thresholding operation, the use of capacity transport constraint relaxation [4], of other statistical features, of pre-conditionned optimization algorithms, and the extension to region-based segmentation and to multi-phase segmentation problem.

References

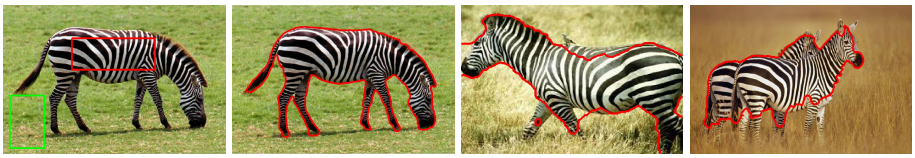
1. M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Neural Information Processing Systems (NIPS'13)*, pages 2292–2300, 2013.
2. M. Cuturi and A. Doucet. Fast computation of wasserstein barycenters. In *International Conference on Machine Learning (ICML'14)*, pages 685–693, 2014.
3. M. Cuturi, G. Peyré, and A. Rolet. A smoothed dual approach for variational wasserstein problems. *Preprint arXiv:1503.02533*, Mar. 2015.
4. S. Ferradans, N. Papadakis, G. Peyré, and J.-F. Aujol. Regularized discrete optimal transport. *SIAM J. Imaging Sciences*, (1):212–238, 2014.
5. M. Jung, G. Peyré, and L. D. Cohen. Texture segmentation via non-local non-parametric active contours. EMMCVPR'11, pages 74–88, Berlin, Heidelberg, 2011. Springer-Verlag.
6. D. Lorenz and T. Pock. An inertial forward-backward algorithm for monotone inclusions. *Journal of Mathematical Imaging and Vision*, 51(2):311–325, 2015.
7. K. Ni, X. Bresson, T. Chan, and S. Esedoglu. Local histogram based segmentation using the wasserstein distance. *Int. J. of Computer Vision*, 84(1):97–111, 2009.
8. N. Papadakis, E. Provenzi, and V. Caselles. A variational model for histogram transfer of color images. *IEEE Trans. on Image Processing*, 20(6):1682–1695, 2011.
9. G. Peyré, J. Fadili, and J. Rabin. Wasserstein active contours. In *IEEE International Conference on Image Processing (ICIP'12)*, 2012.
10. J. Rabin and N. Papadakis. Convex Color Image Segmentation with Optimal Transport Distances. *ArXiv e-prints 1503.01986*, Mar. 2015.
11. J. Rabin and G. Peyré. Wasserstein regularization of imaging problem. In *IEEE International Conderence on Image Processing (ICIP'11)*, pages 1541–1544, 2011.
12. P. Swoboda and C. Schnörr. Variational image segmentation and cosegmentation with the wasserstein distance. In *EMMCVPR*, pages 321–334, 2013.
13. C. Villani. *Topics in Optimal Transportation*. AMS, 2003.
14. R. Yıldızoglu, J.-F. Aujol, and N. Papadakis. A convex formulation for global histogram based binary segmentation. In *EMMCVPR*, pages 335–349, 2013.

Input $t = 0.9$ $t = 0.5$ $t = 0.1$

Fig. 1. Influence of the threshold on OT segmentation ($\lambda = \infty$).

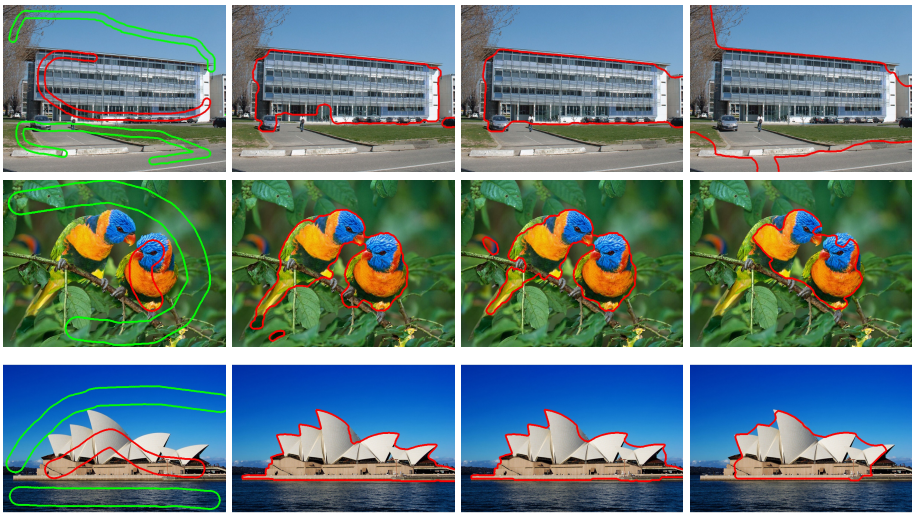
Input L_1 OT ($\lambda = \infty$)

Fig. 2. Robustness of OT with respect to L_1: the blue colors that are not in the reference histograms are considered as background with OT distance and as foreground with the L_1 model, since no color transport is taken into account.



Input histograms Segmentation 1 Segmentation 2 Segmentation 3

Fig. 3. Illustration of the interest of optimal transport for comparison of histograms. Prior histograms taken from image 1 are used to segment images 2 and 3.



Input $\lambda = \infty$ $\lambda = 100$ $\lambda = 10$

Fig. 4. Comparison of segmentations obtained from the proposed models. The input areas are used to compute the reference color distributions a and b. The non-regularized model corresponds to $\lambda = +\infty$, increasing regularization effects are then shown.

Input $\lambda = \infty$ Input $\lambda = \infty$

Fig. 5. Texture segmentation using histograms of gradient norms.