J Math Imaging Vis
DOI 10.1007/s10851-017-0725-5

Convex Histogram-Based Joint Image Segmentation with Regularized Optimal Transport Cost

Nicolas Papadakis[1] · **Julien Rabin**[2]

Received: 23 September 2016 / Accepted: 7 March 2017

Abstract We investigate in this work a versatile convex framework for multiple image segmentation, relying on the regularized optimal mass transport theory. In this setting, several transport cost functions are considered and used to match statistical distributions of features. In practice, global multidimensional histograms are estimated from the segmented image regions and are compared to reference models that are either fixed histograms given a priori, or directly inferred in the non-supervised case. The different convex problems studied are solved efficiently using primal–dual algorithms. The proposed approach is generic and enables multiphase segmentation as well as co-segmentation of multiple images.

Keywords Optimal transport · Wasserstein distance · Sinkhorn distance · Image segmentation · Convex optimization

1 Introduction

Optimal Transport in Imaging 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. In contrast to most distances from information theory (e.g., the Kullback–Leibler divergence), optimal transport takes into account the spatial location of the density mode and defines robust distances between empirical distributions. The geometric nature of optimal transport, as well as the ability to compute optimal displacements between densities through the corresponding transport map, makes this theory progressively mainstream in several applied fields. In image processing, the warping provided by the optimal transport has been used for video restoration [26], color transfer [52], texture synthesis [36], optical nanoscopy [11] and medical imaging registration [32]. It has also been applied to interpolation in computer graphics [7,70] and surface reconstruction in computational geometry [27].

The optimal transport distance has also been successfully used in various image processing and machine learning tasks, image retrieval [50,65], image segmentation [45], image decomposition [39] or texture synthesis [58].

Some limitations have been also shown and partially addressed, such as time complexity [6,23,66], regularization and relaxation of the transport map [28] for imaging purposes.

Image Segmentation Image segmentation has been the subject of active research for more than 20 years (see, e.g., [2,22] and references therein). For instance, we can refer to the seminal work of Mumford and Shah [44], or to its very popular approximation with level sets developed by Chan and Vese in [17]. This last work provides a very flexible algorithm to segment an image into two homogeneous regions, each one being characterized by its mean gray level value.

In the case of textured images, a lot of extensions of [17] have been proposed to enhance the mean value image segmentation model by considering other kinds of local information. For instance, local histograms are used in [45,82], Gabor filters in [72], wavelet packets in [3] and textures are characterized thanks to the structure tensor in [9,63].

✉ Nicolas Papadakis
nicolas.papadakis@math.u-bordeaux.fr

✉ Julien Rabin
julien.rabin@unicaen.fr

[1] IMB, UMR 5251, Université de Bordeaux, 33400 Talence, France

[2] Normandie Univ, ENSICAEN, CNRS, GREYC, 14000 Caen, France

Published online: 22 March 2017

Advanced statistical-based image segmentation models using first parametric models (such as the mean and variance), and then empirical distributions combined with adapted statistical distances such as the Kullback–Leibler divergence, have been thoroughly studied in the literature. One can, for instance, refer to the works in [1,10,33,37] that consider the global histograms of the regions to segment and are also based on the Chan and Vese model [17]. It is important to notice that this class of approaches involves complex shape gradient computations for the level set evolution equation. Moreover, as these methods all rely on the evolution of a level set function [47], it leads to non-convex methods that are sensitive to the initialization choice and only a local minimizer of the associated energy is computed.

Recently, convexification methods have been proposed to tackle this problem, as in [5,8,12,14,15,46,54,78]. The original Chan and Vese model [17] can indeed be convexified, and a global solution can be efficiently computed, for instance with a primal–dual algorithm. By means of the coarea formula, a simple thresholding of this global solution provides a global minimizer of the original non-convex problem. The multiphase segmentation model based on level sets [72] can also be treated with convexification methods. However, the thresholding of the estimated global minima does not anymore ensure the recovery of a global optimal multiphase segmentation [53]. Notice that such approaches have not been developed yet for global histogram segmentation with length boundary regularization.

Other models as in [4,30,62,73,80] use graph-based methods and max-flow formulations [55] in order to obtain good minima without level set representation. Nevertheless, these approaches are restricted to bin-to-bin distances (for instance, ℓ_1 [62], Bhattacharyya [80], ℓ_2 [59] or χ^2 [75]) between features' histograms that are not robust enough to deal with non-uniform quantification or data outliers.

Optimal Transport and Image Segmentation The use of Optimal transport for image segmentation has been first investigated in [45] for comparing local $1D$ histograms. In [42,51], active contours approaches using the Wasserstein distance for comparing global multidimensional histograms of the region of interest have been proposed. Again, these non-convex active contours methods are sensitive to the initial contour. Moreover, their computational cost is very large, even if they include some approximations of the Wasserstein distance as in [51].

In order to deal with global distance between histograms while being independent of any initialization choice, convex formulations have been designed [71,79]. In [79], a ℓ_1 norm between cumulative histograms is considered and gives rise to a fast algorithm. This is related to optimal transport only for 1D histograms of grayscale images. The authors of [71] proposed to rely on the Wasserstein distance. In order to be able to optimize the corresponding functional, it requires to make use of sub-iterations to compute the proximity operator of the Wasserstein distance, the use of which is restricted to low-dimensional histograms. Hence, we considered in [57] a fast and convex approach involving regularization of the optimal transport distance through the entropic regularization of [23]. In this paper, we investigate in detail this regularized model and look at its extension to multiphase segmentation.

Co-segmentation As already proposed in [71], the studied convex framework can be extended to deal with the unsupervised co-segmentation of two images. The problem of co-segmentation [74] consists in segmenting simultaneously multiple images that contain the same object of interest without any prior information. When the proportion between the size of the object and the size of the image is the same in all images, the model of [71] can be applied. It aims at finding regions in different images having similar color distributions. However, this model is not suited for cases where the scale of the object varies. In the literature, state-of-the art approaches rely on graph representation. They are able to deal with small-scale changes [64] or large ones by considering global information on image subregions [34] pre-computed with dedicated algorithms. Notice that convex optimization algorithms involving partial duality of the objective functional have been used for co-segmentation based on local features [35]. Such an approach is able to deal with scale change of objects, but it relies on high-dimensional problems scale with $O(N^2)$, where N is the total number of pixels. The use of robust optimal transport distances within a low-dimensional formulation for the global co-segmentation of objects of different scales is thus an open problem that is addressed in this paper.

Contributions The global segmentation models presented in this paper are based on the convex formulation for two-phase image segmentation of [79] involving ℓ_1 distances between histograms. Following [57,71], we consider the use of Wasserstein distance for global segmentation purposes. As in [57], we rely on the entropic regularization [23,24] of optimal transport distances in order to deal with accurate quantization of histograms. Hence, this paper shares some common features with the recent work of [25] in which the authors investigate the use of the Legendre–Fenchel transform of regularized transport cost for imaging problems.

With respect to the preliminary version of this work presented in a conference [57], the contributions of this paper are the following:

- we give detailed proofs of the computation of the functions and operators involved by the entropic regularization of optimal transport between non-normalized histograms.

– we generalize the framework to the case of multiphase segmentation in order to find a partition of the images with respect to several priors;
– we provide numerous experiments exhibiting the properties of our framework;
– we extend our model to the co-segmentation of multiple images. Two convex models are proposed. The first one is able to co-segment an object with constant size in two images for general ground costs. The second one can deal with different scales of a common object contained in different images for a specific ground cost.

This paper is also closely related to the framework proposed in [71]. With respect to this method, our contributions are:

– the use of regularized optimal transport distances for dealing with high-dimensional histograms;
– the generalization of the framework to multiphase segmentation;
– the definition of co-segmentation model for more than two images dealing with scale changes of objects.

2 Convex Histogram-Based Image Segmentation

2.1 Notation and Definitions

We consider here vector spaces equipped with the Euclidean inner product $\langle .\,,\,. \rangle$ and the ℓ_2 norm $\|.\| = \sqrt{\langle .\,,\,. \rangle}$. The adjoint linear operator of A is denoted by A^* and satisfies $\langle A.\,,\,. \rangle = \langle .\,,\,A^*. \rangle$. We denote as $\mathbf{1}_n$ and $\mathbf{0}_n \in \mathbb{R}^n$ the n-dimensional vectors filled with ones and zeros, respectively, x^T the transpose of x, while Id stands for the identity operator. The concatenation of the vectors x and y into a vector is denoted $(x; y)$. Operations and functions on vectors and matrices are meant componentwise, such as inequalities:

$$X \leq Y \quad \Leftrightarrow \quad X_{ij} \leq Y_{ij} \quad \forall\, i, j$$

or exponential and logarithm functions:

$$(\exp X)\,(i, j) = \exp X_{i,j} \qquad \log X = \left(\log X_{i,j}\right)_{i,j}.$$

We refer to ℓ_p norm as $\|x\|_p = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}$. The norm of a linear operator A is $\|A\| = \sup_{\|x\|=1} \|Ax\|$.

The operator $\mathrm{diag}(x)$ defines a square matrix whose diagonal is the vector x. The identity matrix is $\mathrm{Id}_n = \mathrm{diag}(\mathbf{1}_n)$. The functions $\mathbb{1}_S$ and χ_S are, respectively, the indicator and characteristic functions of a set S

$$\mathbb{1}_S(x) = \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{otherwise} \end{cases}, \quad \chi_S(x) = \begin{cases} 0 & \text{if } x \in S \\ \infty & \text{otherwise} \end{cases}.$$

The Kronecker δ symbol is $\delta_{i,j} = 1$ if $i = j$, and $\delta_{i,j} = 0$ otherwise.

A histogram with n bins is a vector $h \in \mathbb{R}^n_+$ with nonnegative entries. The set

$$\mathcal{S}_{m,n} := \{x \in \mathbb{R}^n_+, \langle x,\, \mathbf{1}_n \rangle = m\} \tag{1}$$

is the simplex of histogram vectors of total mass m ($\mathcal{S}_{1,n}$ being the n-dimensional probability simplex).

The operators Prox and Proj stand, respectively, for the Euclidean proximity and projection operators:

$$\mathrm{Prox}_f(x) = \mathrm{argmin}_y \tfrac{1}{2}\|y - x\|^2 + f(x)$$
$$\mathrm{Proj}_S(x) = \mathrm{argmin}_{y \in S} \|y - x\| = \mathrm{Prox}_{\chi_S}(x)\,.$$

Functions f for which the proximity operator is known in closed form, or at least can be evaluated at a given point explicitly, are usually referred to as *simple*.

The Legendre–Fenchel conjugate f^* of a proper, lower semicontinuous and convex function f is

$$f^*(y) = \sup_x \langle x,\, y \rangle - f(x),$$

and satisfies the equality: $f^{**} = f$.

2.2 General Formulation of Distribution-Based Image Segmentation

For sake of simplicity, we first describe the binary segmentation problem. The following framework can be extended to multiphase segmentation, as latter shown in Sect. 2.4.

Let $I : x \in \Omega \mapsto I(x) \in \mathbb{R}^d$ be a multidimensional image, defined over the N-pixel domain Ω ($N = |\Omega|$), and $\mathcal{F}$ a feature-transform into n-dimensional descriptors: $\mathcal{F}I(x) \in \mathbb{R}^n$. The border of the domain is denoted $\partial\Omega$. 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 [79], we consider the energy

$$E(u) = \rho \mathrm{TV}(u) + S(a, h(u)) + S(b, h(\mathbf{1} - u)) \tag{2}$$

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

- the fidelity terms are defined using $S(.,.)$, a dissimilarity measure between distributions of 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

$$h(u) : y \in \mathbb{R}^n \mapsto \frac{1}{\sum_{x \in \Omega} u(x)} \sum_{x \in \Omega} u(x) \delta_{\mathcal{F}I(x)}(y)\,; \tag{3}$$

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

Observe that this energy is highly non-convex, h being a nonlinear operator, and that we would like to find a minimum $u^\star \in \{0,1\}^N$ over a non-convex set.

2.3 Convex Relaxation of Histogram-Based Segmentation Energy

The authors of [79] propose some relaxations and a reformulation in order to handle the minimization of energy (2) using convex optimization tools.

2.3.1 Probability Map

First, it consists in considering solutions from the convex envelope of the binary set, i.e., using a segmentation variable $u : \Omega \mapsto [0,1]$ which can be interpreted as a weight function (probability map). A threshold is therefore required to obtain a binary segmentation of the image into the region corresponding to the prior distribution a

$$R_t(u) := \{x \in \Omega \,|\, u(x) \geq t\}, \tag{4}$$

its complement $R_t(u)^c$ corresponding to prior distribution b. Other post-processing partition techniques may be considered and are discussed later.

It is worth mentioning that for the specific TV-ℓ_1 approach of [46], where the dissimilarity measure $S(u, u_0) = \|u - u_0\|_1$ is the ℓ_1 distance between the segmentation variable u and a given *prior binary* segmentation variable u_0, such a relaxation still guarantees to find a global solution for the non-convex problem. However, there is no such a property in our general setting.

2.3.2 Feature Histogram

Considering the continuous domain of the feature space, as done, for instance, in [51], may become numerically intractable for high dimensional descriptors. We consider instead histograms, as already proposed in [45,79].

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 indexed by $i \in \{1, \ldots M_X\}$

$$(H_{\mathcal{X}}(u))_i = \sum_{x \in \Omega} u(x) \mathbb{1}_{\mathcal{C}_{\mathcal{X}}(i)}(\mathcal{F}I(x)), \tag{5}$$

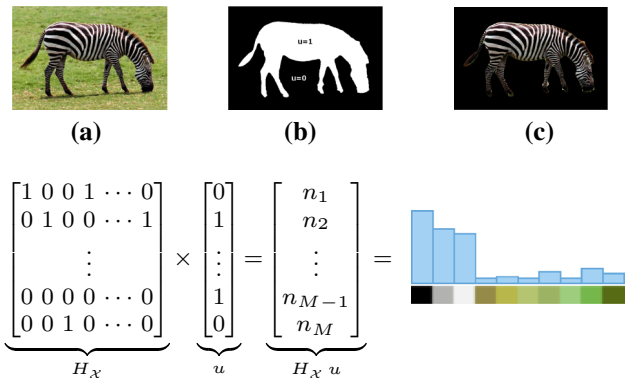


Fig. 1 Illustration of the color histogram computed from a binary region. **a** Image I. **b** Binary segmentation map u. **c** Corresponding region $I \odot u(x) = I(x).u(x)$. The hard assignment linear operator $H_{\mathcal{X}}$ encodes the position of each pixel in the clustered color space. The histogram value n_i represents here the number of pixels $I(x)$ of the region characterized by $u(x) = 1$ that belong to the feature cluster $C_{\mathcal{X}}(i)$

where X_i is the centroid of the corresponding bin i, and $\mathcal{C}_{\mathcal{X}}(i) \subset \mathbb{R}^n$ is the corresponding set of features (e.g., the Voronoi cell obtained from nearest-neighbor assignment). We can write $H_{\mathcal{X}}$ as a linear operator

$$H_{\mathcal{X}} : u \in \mathbb{R}^N \mapsto H_{\mathcal{X}} \cdot u \in \mathbb{R}^{M_X}, \tag{6}$$

with matrix notation $H(i,x) := 1$ if $\mathcal{F}I(x) \in \mathcal{C}_{\mathcal{X}}(i)$ and 0 otherwise. Note that $H_{\mathcal{X}} \in \mathbb{R}^{M_X \times N}$ is a fixed *hard assignment* matrix that indicates which pixels of $\mathcal{F}I$ contribute to each bin of the histogram. As a consequence, we have the property

$$\langle H_{\mathcal{X}}\, u,\, \mathbf{1}_{M_{\mathcal{X}}} \rangle = \sum_{x \in \Omega} u(x) = \langle u,\, \mathbf{1}_N \rangle, \tag{7}$$

so that $H_{\mathcal{X}}(u) \in \mathcal{S}_{\langle u,\, \mathbf{1} \rangle, M_X}$. This linear operator computing the histogram of a particular region of the image is illustrated in Fig. 1 for RGB color feature.

2.3.3 Exemplar Histograms

The segmentation is driven by two fixed histograms $a \in \mathcal{S}_{1,M_a}$ and $b \in \mathcal{S}_{1,M_b}$, 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 [79] and consider $S\left(a\langle u,\, \mathbf{1}_N \rangle,\, H_{\mathcal{A}}(u)\right)$ as fidelity term, where the constant vector a has been scaled to $H_{\mathcal{A}}(u) \in \mathcal{S}_{\langle u,\, \mathbf{1} \rangle, M_a}$.

Note that this approach, based on the comparison of unnormalized histogram pairs as a data fidelity term, is also used in [43,61] for co-segmentation. We will further discuss the

consequence of such a choice for this problem in the dedicated Sect. 6. 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 and consider $H_{\mathcal{A}} = H_{\mathcal{B}}$ and $M_a = M_b = M$ in order to simplify the notation.

We introduce the linear operators

$$A := a\,\mathbf{1}_N^\mathsf{T} \in \mathbb{R}^{M\cdot N} \quad \text{and} \quad B := b\,\mathbf{1}_N^\mathsf{T} \in \mathbb{R}^{M\cdot N} \tag{8}$$

such that $Au = (a\mathbf{1}^\mathsf{T})u = a\langle u, \mathbf{1}\rangle$.

2.3.4 Spatial Regularization

As previously mentioned, we rely on the total variation of the segmentation variable in order to control its spatial regularity. We thus denote as $D : \mathbb{R}^N \mapsto \mathbb{R}^{2N}$, the finite difference operator on the bi-dimensional cartesian grid Ω. The gradient at a pixel coordinate $x = (i, j) \in \Omega$ is $Du(x) = v(x) = (v_1(x); v_2(x))$ where one has $\forall\, x \in \Omega\backslash\partial\Omega$ (i.e., excluding of the domain's border):

$$\begin{aligned} v_1(i, j) &= u(i, j) - u(i-1, j), \\ v_2(i, j) &= u(i, j) - u(i, j-1). \end{aligned}$$

On $\partial\Omega$, we use homogeneous Dirichlet conditions:

$$v_1(0, j) = u(0, j), \quad v_2(i, 0) = u(i, 0), \quad \forall i, j$$

meaning that a pixel x outside Ω is considered as background ($u(x) = 0$).

The usual discrete definition of the isotropic total variation used in the problem (2) is

$$\mathrm{TV}(u) := \|Du\|_{1,2} = \sum_{x\in\Omega} \|Du(x)\|_2\ , \tag{9}$$

where the $\ell_{1,2}$ norm for a gradient field $v = (v_1; v_2)$ corresponds to

$$\|v\|_{1,2} := \sum_{x\in\Omega} \sqrt{v_1(x)^2 + v_2(x)^2}.$$

2.3.5 Segmentation Energy

Gathering all previous ingredients, the convex segmentation problem (2) can now be discretized as the following constrained problem:

$$\begin{aligned} \min_{u\in[0,1]^N} \ & \rho\ \|Du\|_{1,2} + \tfrac{1}{\gamma}\, S(Au, Hu) \\ & + \tfrac{1}{N-\gamma}\, S(B(\mathbf{1}-u), H(\mathbf{1}-u)). \end{aligned} \tag{10}$$

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. We nevertheless omit this parameter (by taking $\gamma = N/2$) since its value seems not to be critical in practice, as demonstrated in [79]. This model now compares the histograms of the regions with the rescaled reference histograms, instead of normalized distributions defined in Eq. (3). Notice that the matrix $H \in \mathbb{R}^{M\cdot N}$ is sparse (with only N nonzero values) and A and B are of rank 1, so that storing or manipulating these matrices is not an issue.

In [79], the distance function S was defined as the ℓ_1 norm. In Sects. 3 and 4, we investigate the use of similarity measures based on optimal transport cost, which is known to be more robust and relevant for histogram comparison [56]. In the next paragraphs, we first investigate some extensions of the previous framework, and then, we describe the optimization method used to solve the proposed variational problems.

2.4 Convex Multiphase Formulation

Let $a_1, \ldots a_K$ be $K \geq 2$ input histograms. The previous framework can be extended to estimate a partition of the domain Ω of an input image with respect to these histograms.

Multiple Probability Map A simple way to extend the binary model defined in Formula (2) is to describe the partition of the image into K regions for each pixel $x \in \Omega$ by a binary variable $u(x) \in \{0, 1\}^K$:

$$u(x) = (u_1(x); \ldots u_K(x)) \ \text{ s.t. } \ \langle u(x),\, \mathbf{1}_K\rangle = 1$$

where $u_k(x)$ states whether the pixel at x belongs to the region indexed by k or not.

The extension of the convex optimization problem (10) is then obtained by the relaxation of u into a probability vector map, as done, for instance, in [49,81]: $u(x) \in \mathcal{S}_{1,K}$ so that $u_k(x)$ defines the probability the pixel at x belongs to the region indexed by k

$$\min_{\substack{u=(u_k\in\mathbb{R}^N)_{1\leq k\leq K} \\ \text{s.t. } u(x)\in\mathcal{S}_{1,K}}} \ \sum_{k=1}^{K} \|Du_k\|_{1,2} + S(A_k u_k, H_k u_k)\ , \tag{11}$$

where $H_k = H$ for all k in the simplified setting, and where A_k indicates the linear operator that multiplies histogram a_k by the total sum of the entries of u_k, as previously defined in Eq. (8).

Notice that other convex relaxations for multiphase segmentation with non-ordered labels and total variation regularization have been proposed in the literature [38,53]. The one proposed in [38] is nevertheless less tight than [81]. On the other hand, the convexification of [53] is even tighter but harder to optimize, while giving very close results, even on

pathological cases after thresholding (see [53] for a detailed comparison).

2.5 Other Model Variations

In addition to the multiple labeling extension, some other variations of the previous framework are discussed in this section.

Soft Assignment Histogram For simplicity, we have assumed previously that each operator H_k is an *hard assignment* operators [see definition (6)]. In the proposed framework, these histogram operators could be instead defined from soft assignment, which might reduce quantization noise. In this case, special care should be taken regarding the conditioning of the matrix H, as some rows of H could be arbitrarily close to zero.

Supervised Soft Labeling In our framework, prior histograms $\{a_k\}_K$ may be given a priori but can also be defined from *scribbles* drawn on the input image by the user. In the experiments, we will consider binary scribbles $s_k : \Omega \mapsto \{0, 1\}$ so that prior histograms are defined as [assuming that condition (7) is fulfilled]

$$a_k = \frac{H_k\, s_k}{\langle s_k,\, \mathbf{1}\rangle}\ .$$

This approach makes it possible for the model (11) to correct user mislabeling, as the segmentation variables u_k do not need to agree with the user labeling s_k. Alternatively, it may be possible to use hard labeling constraints instead without increasing the model complexity.

Multi-image Segmentation The framework enables to segment multiple images with the same prior histograms that can be defined by scribbles from different images. Without adding interaction terms to measure the similarity between the segmentation variables of each image, the corresponding optimization problems can be solved separately for each image.

2.6 Optimization

All convex segmentation problems studied in this work are addressed using primal–dual forward-backward optimization schemes. Depending on the properties of the convex function S chosen to measure similarity between histograms, several algorithms can be considered.

In particular, when S is a Lipschitz-differentiable function (using, for instance, a quadratic ℓ_2, Huber loss or χ^2 distance), even simpler forward-backward algorithm can be used. However, such a choice of function is known to be not very well suited for histogram comparison (see, for instance, [56]) and more robust distances are therefore preferred, such as the ℓ_1 norm in [78].

As a consequence, and without loss of generality, we do not address this specific case in the following and consider the most general setting, without any assumptions on S (or S^*, its Legendre–Fenchel transform) aside from being convex and lower semicontinuous.

Two-Phase Segmentation Model In order to reformulate (10) as a primal–dual optimization problem, we resort to variable splitting, using the Legendre–Fenchel transforms of the discrete TV norm and the function S to obtain

$$\begin{aligned}\min_{u}\max_{\substack{v\\ p_A,q_A\\ p_B,q_B}}\ & \langle Du,\, v\rangle + \langle Hu,\, p_A\rangle + \langle H(\mathbf{1}-u),\, p_B\rangle \\ & + \langle Au,\, q_A\rangle + \langle B(\mathbf{1}-u),\, q_B\rangle \\ & - S^*(p_A, q_A) - S^*(p_B, q_B) \\ & + \chi_{[0,1]^N}(u) - \chi_{\|.\|_{\infty,2}\le\rho}(v)\end{aligned} \tag{12}$$

where the primal variable is $u = (u(x))_{x\in\Omega} \in \mathbb{R}^N$ (corresponding to the segmentation map), and dual variables are $v = (v(x))_{x\in\Omega} \in \mathbb{R}^{2N}$ (related to the gradient field) and $p_A, p_B, q_A, q_B \in \mathbb{R}^M$ (related to the histograms). Notice that S^* is the convex conjugate of the function S. In this new problem formulation, $\chi_{\|.\|_{\infty,2}\le\rho}$ is the characteristic function of the convex $\ell_{\infty,2}$ ball of radius ρ, as we have for the discrete isotropic TV norm

$$\begin{aligned}\rho\mathrm{TV}(u) &= \rho\sum_{x\in\Omega}\sup_{\|v(x)\|\le 1}\langle v(x),\, Du(x)\rangle \\ &= \sup_{v}\langle v,\, Du\rangle - \sum_{x\in\Omega}\chi_{\|.\|\le\rho}(v(x)) \\ &= \sup_{v}\langle v,\, Du\rangle - \chi_{\|.\|_{\infty,2}\le\rho}(v).\end{aligned}$$

In order to accommodate the different models studied in this paper, we assume here that S^* is a sum of two convex functions $S^* = S_1^* + S_2^*$, where S_1^* is non-smooth and S_2^* is differentiable with Lipschitz continuous gradient.

We recover a general primal–dual problem of the form

$$\min_{u}\max_{p}\ \langle \mathcal{K}u,\, p\rangle + R(u) + T(u) - F^*(p) - G^*(p) \tag{13}$$

with primal variable $u \in \mathbb{R}^N$ and dual variable $p = (p_A; q_A; p_B; q_B; v) \in \mathbb{R}^{4M+2N}$, where

- $\mathcal{K} = [H^\mathsf{T}, A^\mathsf{T}, -H^\mathsf{T}, -B^\mathsf{T}, D^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^{(4M+2N)\times N}$ is a sparse matrix;

- T is convex and smooth, with Lipschitz continuous gradient ∇T with constant L_T. For now, we have $T(u) = 0$ and $L_T = 0$ in the setting of problem (12).
- R is convex and non-smooth. In problem (12), we have $R = \chi_{\mathcal{C}}$ the indicator function of the convex domain $\mathcal{C} = [0, 1]^N$;
- $F^*(p) = S_1^*(p_A, q_A) + S_1^*(p_B, q_B) + \chi_{\|.\|_{\infty,2} \leq \rho}(v)$ is convex and non-smooth;
- $G^*(p) = S_2^*(p_A, q_A) + S_2^*(p_B, q_B) - \langle H\mathbf{1}_N,\ p_B\rangle - \langle B\mathbf{1}_N,\ q_B\rangle$ is convex and differentiable, with Lipschitz continuous gradient with constant L_{G^*}. From definition of H and B, one has $B\mathbf{1}_N = Nb$ and $H\mathbf{1}_N = Nh_I$ where h_I is the normalized histogram of features of the image I.

To solve this problem, we consider the primal dual (PD) algorithm of [16,77]

$$\begin{cases} u^{(t+1)} = \mathrm{Prox}_{\tau R}\big(u^{(t)} - \tau(\mathcal{K}^\mathsf{T} p^{(t)} + \nabla T(u^{(t)}))\big) \\ p^{(t+1)} = \mathrm{Prox}_{\sigma F^*}\big(p^{(t)} + \sigma\mathcal{K}(2u^{(t+1)} - u^{(t)}) \\ \qquad\qquad -\sigma\nabla G^*(p^{(t)})\big) \end{cases} \tag{PD}$$

where the notation $u^{(t)}$ indicates the variable at discrete time indexed by t. For problem (12), one has $\mathrm{Prox}_{\tau R} = \mathrm{Proj}_{[0,1]^N}$. The application $\mathrm{Prox}_{\sigma F^*}$ depends on the non-smooth part of similarity function S and can be written, due to separability, as

$$\mathrm{Prox}_{\sigma F^*}(p) = \Big(\mathrm{Prox}_{\sigma S_1^*}(p_A, q_A);\ \mathrm{Prox}_{\sigma S_1^*}(p_B, q_B);\ \mathrm{Proj}_{\|.\|_{\infty,2}\leq\rho}(v)\Big)$$

where

$$\mathrm{Proj}_{\|.\|_{\infty,2}\leq\rho}(v)(x) = \frac{v(x)}{\max\big\{\|v(x)\|/\rho,\, 1\big\}}. \tag{14}$$

The algorithm (PD) is guaranteed to converge from any initialization of $(u^{(0)}, p^{(0)})$ to a saddle point of (13) as soon as the step parameters σ and τ satisfy (see, for instance, [16, Eq. 21])

$$\left(\tfrac{1}{\tau} - L_T\right)\left(\tfrac{1}{\sigma} - L_{G^*}\right) \geq \|\mathcal{K}\|^2. \tag{15}$$

The worst-case estimate for this norm is

$$\|\mathcal{K}\| = 4\sqrt{N} + \sqrt{8}.$$

Proof See "Appendix 1".

Preconditioning As a consequence of the large value of $\|\mathcal{K}\|^2$ scaling with the primal variable dimension, the gradient step parameters (τ, σ) may be very small to satisfy Eq. (15), which results in a slow convergence.

Fortunately, this algorithm can benefit from the recent framework proposed in [19,41], using preconditioning. The idea is to change the metric by using—fixed or variable—matrices $\mathbf{T}$ and $\mathbf{\Sigma}$ in lieu of scalar parameters τ and σ in algorithm (PD).

Following the guideline proposed in [41] to design diagonal and constant conditioning matrices, we define

$$\mathbf{T} := \mathrm{diag}\,(\boldsymbol{\tau}) \quad \text{and}$$
$$\mathbf{\Sigma} := \mathrm{diag}\,(\boldsymbol{\sigma}) = \mathrm{diag}\,(\boldsymbol{\sigma}_H, \boldsymbol{\sigma}_a, \boldsymbol{\sigma}_H, \boldsymbol{\sigma}_b, \boldsymbol{\sigma}_D)$$

where

$$\begin{aligned} \frac{1}{\boldsymbol{\tau}(x)} &= \frac{L_T}{\gamma} + r\sum_{i=1}^{4M+2N} |\mathcal{K}_{i,x}|\,, \\ \frac{1}{\boldsymbol{\sigma}(i)} &= \frac{L_{G^*}}{\delta} + \frac{1}{r}\sum_{x=1}^{N} |\mathcal{K}_{i,x}|\,. \end{aligned} \tag{16}$$

For the setting of problem (12), considering an hard assignment matrix H and writing the operator D in matrix form, we have

$$\begin{aligned} \frac{1}{\boldsymbol{\tau}(x)} &= 4r + r\sum_{y=1}^{2N} |D_{y,x}| \leq 8r \\ \frac{1}{\boldsymbol{\sigma}_H} &= \frac{L_{G^*}}{\delta}\mathbf{1}_M + \frac{N}{r}h_I \quad \text{with } h_I = \tfrac{1}{N}H\mathbf{1}_N \\ \frac{1}{\boldsymbol{\sigma}_h} &= \frac{L_{G^*}}{\delta}\mathbf{1}_M + \frac{N}{r}h \quad \text{for histogram } h = a \text{ and } b \\ \frac{1}{\boldsymbol{\sigma}_D(y)} &= \frac{1}{r}\sum_{x=1}^{N} |D_{y,x}| \leq \frac{2}{r}. \end{aligned}$$

The scaling parameters $r > 0$ and $\delta \in (0, 2)$ enable to balance the update between the primal and the dual variables. We observed that the preconditioning allows for the use of very unbalanced histograms (that is far from being uniform) that otherwise could make the convergence arbitrarily slow.

Other acceleration methods, such as variable metric [19] and inertial update [41], may be considered.

Multiphase Optimization The algorithm (PD) can still be used to minimize problem (11). The only two differences are the size of the variables and the convex constraint set $\mathcal{C}$. First, we consider now multidimensional primal and dual variables, i.e., respectively $u : x \in \Omega \mapsto (u_k(x))_{k=1}^K$ and $p = (p_k)_{k=1}^K$ with $p_k = (p_A^k; q_A^k; p_B^k; q_B^k; v^k)$. Furthermore, the constraint set $\mathcal{C}$ for the primal variable u is defined for each pixel $u(x)$ as the simplex $\mathcal{S}_{1,K}$ [defined in Eq. (1)], so that:

$$R(u) = \sum_{x\in\Omega} \chi_{\mathcal{S}_{1,K}}(u(x))\,. \tag{17}$$

In this setting, the definition of the diagonal preconditioners *for each phase k* is the same as in (16).

Eventually, the primal variable $u^\star = u^{(\infty)}$ provided by the algorithm (PD) only solves the relaxed segmentation problem and has to be post-processed to obtain a partition of the image, as discussed in the next paragraph.

2.7 Binarization of the Relaxed Solution

The solution $u^\star$ of the relaxed segmentation problems studied before is a probability map, i.e., $u^\star(x) \in [0, 1]$. Although in practice we have observed (see the experimental Sect. 5), as already reported in [46,81] for other models, that the solution is often close to be binary, i.e., $u^\star(x) \approx 0$ or 1, some thresholding is still required to obtain a proper labeling of the image.

Following for instance [81], we simply select for every pixel x the most likely label based on probability maps solutions $\{u_k^\star\}_{1\leq k\leq K}$, that is

$$x \mapsto \text{argmin}_k \ \{u_k^\star(x)\}_{1\leq k\leq K} \ . \tag{18}$$

Recall that in general, there is no correspondence between this thresholded solution and the global minimizer of the non-relaxed problem over binary variables.

In the specific case of the $K = 2$ phase segmentation problem, the previous processing boils down to using a threshold $t = \frac{1}{2}$ to define $u_t(x) = \mathbb{1}_{u^\star(x)>t}$. A better strategy would be to optimize the global threshold t such that the objective functional $J(u_t)$ is minimized. However, due to the complexity of the measures S considered in this work, this method is not considered here.

3 Monge–Kantorovich Distance for Image Segmentation

We investigate in this section the use of optimal transport costs as a distance function S in the previous framework.

3.1 Optimal Mass Transportation Problem and the Wasserstein Distance

Optimal Transport Problem Following [57], we consider in this work the discrete formulation of the Monge–Kantorovich optimal mass transportation problem (see, e.g., [76]) between a pair of normalized histograms a and b. Given a fixed assignment cost matrix $C_{\mathcal{A},\mathcal{B}} \in \mathbb{R}^{M\times M}$ between the corresponding histogram centroids $\mathcal{A} = \{A_i\}_{1\leq i\leq M}$ and $\mathcal{B} = \{B_j\}_{1\leq j\leq M}$, an optimal transport plan minimizes the global transport cost, defined as a weighted sum of assignments $\forall\,(a, b) \in \Delta$:

$$\mathbf{MK}(a, b) := \min_{P\in\mathcal{P}(a,b)} \left\{ \langle P,\, C\rangle = \sum_{i,j=1}^{M} P_{i,j}C_{i,j} \right\}. \tag{19}$$

The set of admissible histograms is

$$\Delta := \{a, b \in \mathbb{R}_+^M \,,\ \langle a,\, \mathbf{1}_M\rangle = \langle b,\, \mathbf{1}_M\rangle\}, \tag{20}$$

and the polytope of admissible transport matrices reads

$$\mathcal{P}(a, b) := \{P \in \mathbb{R}_+^{M\times M},\, P\mathbf{1}_M = a \text{ and } P^\mathsf{T}\mathbf{1}_M = b\}. \tag{21}$$

Observe that the norm of histograms is not prescribed in Δ 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 we recover the Wasserstein distance

$$\mathbf{W}_p(a, b) = \mathbf{MK}(a, b)^{1/p}, \tag{22}$$

which 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 refer to the **MK** transport cost function as the Monge–Kantorovich *distance*.

ℓ_1 distance As previously mentioned, the ℓ_1 norm is a popular metric in statistics and signal processing, in particular for image segmentation. When penalized by a factor $\frac{1}{2}$, it is also known as the *total variational distance* or the *statistical distance* between discrete probability distributions. As a matter of fact, such a distance can also be seen as a special instance of optimal transport when considering the cost function $C_{i,j} = 2(1 - \delta_{ij})$ and the same set of features $\mathcal{A} = \mathcal{B}$. See "Appendix 2" for more details.

This relation illustrates the versatility and the advantages of optimal transport for histogram comparison as it allows to adapt the distance between histogram features and to use different features for each histogram, as opposed to bin-to-bin metric.

Monge–Kantorovich Distance In the following, due to the use of duality, it is more convenient to introduce the following reformulation for general cost matrix C and $\forall\, a, b \in \mathbb{R}^M$

$$\begin{aligned}\mathbf{MK}(a, b) &= \inf_{P\in\mathcal{P}(a,b)} \langle P,\, C\rangle \\ &= \begin{cases} \min_{P\in\mathcal{P}(a,b)} \langle P,\, C\rangle & \text{if } (a, b) \in \Delta \\ +\infty & \text{otherwise.}\end{cases}\end{aligned} \tag{23}$$

Notice that the optimal transport matrix P is not necessarily unique and does not exist for $(a, b) \notin \Delta$.

Linear Programming Formulation We can rewrite the optimal transport problem as a linear program with vector

variables. The associated primal and dual problems can be, respectively, written as

$$\begin{aligned}\mathbf{MK}(\alpha) &= \inf_{p\in\mathbb{R}^{M^2} \text{ s.t. } p\geq\mathbf{0},\ L^\mathsf{T}p=\alpha} \langle c,\ p\rangle \\ &= \max_{\beta\in\mathbb{R}^{2M} \text{ s.t. } L\beta\leq c} \langle \alpha,\ \beta\rangle,\end{aligned} \tag{24}$$

where $\alpha = (a; b) \in \mathbb{R}^{2M}$ is the concatenation of the two histograms and the unknown vector $p \in \mathbb{R}^{M^2}$ corresponds to the bistochastic matrix P being read columnwise (i.e., $P_{i,j} = p_k$ with 1D index $k(i, j) = i + M(j - 1)$). The $2M$ linear marginal constraints on p are defined by the matrix $L^\mathsf{T} \in \mathbb{R}^{2M\times M^2}$ through equation $L^\mathsf{T}p = \alpha$, where

$$L^\mathsf{T} = \begin{bmatrix} \mathrm{Id}_M & \mathrm{Id}_M & \cdots & \mathrm{Id}_M \\ e_1\mathbf{1}_M^\mathsf{T} & e_2\mathbf{1}_M^\mathsf{T} & \cdots & e_M\mathbf{1}_M^\mathsf{T}\end{bmatrix} = \begin{bmatrix} \mathbf{1}_M^\mathsf{T}\otimes\mathrm{Id}_M \\ \mathrm{Id}_M\otimes\mathbf{1}_M^\mathsf{T}\end{bmatrix},$$

with $e_i(j) = \delta_{ij}\ \ \forall\ j \leq M$. As a consequence,

$$(L\alpha)_{k(i,j)} = \left(L\begin{bmatrix}a\\b\end{bmatrix}\right)_{k(i,j)} = \left(a\mathbf{1}^\mathsf{T} + \mathbf{1}b^\mathsf{T}\right)_{i,j} = a_i + b_j.$$

From the dual formulation (24) that contains a linear objective with inequality constraints, one can observe that the function $\mathbf{MK}(\alpha)$ is not strictly convex in α and not differentiable everywhere. We also draw the reader's attention to the fact that the constraint $\alpha \in \Delta$ is not required anymore with the dual formulation, which will later come in handy.

Conjugate Monge–Kantorovich Distance From Eq. (24), we have that the Legendre–Fenchel conjugate of **MK** can be simply written as the characteristic function of the set $\{\beta \in \mathbb{R}^{2M}\ ,\ L\beta - c \leq \mathbf{0}\}$

$$\mathbf{MK}^*(\beta) = \chi_{L\beta\leq c}(\beta) \qquad \forall\,\beta\in\mathbb{R}^{2M}, \tag{25}$$

where c denotes the vector representation of the cost matrix C (i.e., $C_{i,j} = c_{i+M(j-1)}$).

3.2 Integration in the Segmentation Framework

We propose to substitute in problem (10) the similarity function S with the convex Monge–Kantorovich optimal transport cost (23).

3.2.1 Proximity Operator

In order to apply the minimization scheme described in the algorithm (PD), as $\mathbf{MK}^*$ is not differentiable, we should be able to compute the proximity operator of $\mathbf{MK}^*$. Following (25), it boils down to the projection onto the convex set $\{\beta\ ,\ L\beta\ \leq\ c\}$. However, because the linear operator L is not invertible, this projector cannot be computed in a closed form and the corresponding optimization problem should be solved at each iteration of the primal–dual process.

A similar strategy is employed in [71] with the quadratic Wasserstein distance [defined in (22), using $p = 2$], where the proximity operator of $\mathrm{Prox}_{\mathbf{W}_2^2(.,a)}(Hu)$ with respect to the primal variable u is computed using quadratic programming algorithm. To reduce the resulting time complexity, a reformulation is proposed which does not depends on the size N of the pixel grid, but rather on the number of bins M, as in our framework with the computation of $\mathrm{Prox}_{\mathbf{MK}^*}$.

3.2.2 Biconjugation

To circumvent this problem, we resort to biconjugation to rewrite the **MK** transport cost as a primal–dual problem itself. First, we can write $\mathbf{MK}^*(\beta) = f^*(L\beta)$ with $f^* = \chi_{.\leq c}$, so that $f(r) = \langle r,\ c\rangle + \chi_{.\geq\mathbf{0}}(r)$. Then, using variable splitting

$$\begin{aligned}\mathbf{MK}^*(\beta) = f^*(L\beta) &= \max_r\ \langle r,\ L\beta\rangle - f(r) \\ &= \max_r\ \langle r,\ L\beta - c\rangle - \chi_{.\geq\mathbf{0}}(r)\end{aligned} \tag{26}$$

and

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

where min and max are swapped by virtue of the minimax theorem (the characteristic function being lower semicontinuous for variable r). With this *augmented* representation of the transportation problem, it is no longer necessary to compute the proximity operator of $\mathbf{MK}^*$.

3.2.3 Segmentation Problem

Plugging the previous expression into Eq. (12) enables us to solve it using algorithm (PD). Indeed, introducing new primal variables $r_A, r_B \in \mathbb{R}^{M^2}$ related to transport mappings for the binary segmentation problem, we recover the following primal dual formulation (extension for multiphase segmentation is straightforward using Sect. 2.6)

$$\begin{aligned}\min_{\substack{u\\ r_A,r_B}}\ \max_{\substack{v\\ p_A,q_A\\ p_B,q_B}}\quad & \langle Du,\ v\rangle - \chi_{\|.\|_{\infty,2}\leq\rho}(v) \\ &+ \langle Au,\ q_A\rangle + \langle B(\mathbf{1}-u),\ q_B\rangle \\ &+ \langle Hu,\ p_A\rangle + \langle H(\mathbf{1}-u),\ p_B\rangle \\ &+ \langle r_A,\ c - L\begin{bmatrix}p_A\\q_A\end{bmatrix}\rangle + \langle r_B,\ c - L\begin{bmatrix}p_B\\q_B\end{bmatrix}\rangle \\ &+ \chi_{[0,1]^N}(u) + \chi_{.\geq\mathbf{0}}(r_A) + \chi_{.\geq\mathbf{0}}(r_B).\end{aligned} \tag{27}$$

Using the canonic formulation (13), we consider now

$$\mathcal{K} = \begin{bmatrix} H & -L^\top & \mathbf{0} \\ A & & \\ -H & \mathbf{0} & -L^\top \\ -B & & \\ D & \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{28}$$

In addition, observe that there is now an additional linear term $T(u, r_A, r_B) = \langle r_A + r_B, c\rangle$ whose gradient $\nabla T = (\mathbf{0}_N; c; c)$ has a Lipschitz constant $L_T = 0$. As in problem (12), we still have $R = \chi_{\mathcal{C}}$ which can here be written as

$$\chi_{\mathcal{C}}(u, r_A, r_B) = \chi_{[0,1]^N}(u) + \chi_{\cdot\geq\mathbf{0}}(r_A) + \chi_{\cdot\geq\mathbf{0}}(r_B)\ .$$

The proximity operator of the characteristic function $\chi_{\cdot\geq\mathbf{0}}$ boils down to the projection onto the nonnegative orthant $\mathbb{R}_+^{M^2}$:

$$\mathrm{Proj}_{\cdot\geq\mathbf{0}}(r) = \max\{\mathbf{0}, r\}. \tag{29}$$

The preconditioners for the problem (27) are computed using the definition (16) for the operator $\mathcal{K}$ defined in Formula (28).

3.2.4 Advantages and Drawbacks

The main advantage of this 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, [51,71]), or without making use of approximation (such as the Sliced-Wasserstein distance [51], generalized cumulative histograms [48] or entropy-based regularization [24]). Last, the proposed framework is not restricted to Wasserstein distances, since it enables the use of any cost matrix and does not depend on the feature dimension.

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 [79] 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 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 [23–25]. This strategy is also used by the *soft assign* algorithm [60] to solve linear and quadratic assignment problems.

While offering good properties for optimization, it is also reported [23] to give a good approximation of optimal transportation and increased robustness to outliers. While it has been initially studied for a pair of vectors on the probability simplex $\mathcal{S}_{1,M}$, we follow our preliminary work [57] and investigate in details its use for our framework with unnormalized histograms on Δ.

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

The entropy-regularized optimal transport problem is

$$\mathbf{MK}_\lambda(a, b) = \inf_{P\in\mathcal{P}(a,b)} \langle P,\ C\rangle - \tfrac{1}{\lambda}h(P) \tag{30}$$

where the entropy of the matrix P is defined as $h(P) := -\langle P, \log P\rangle$ (with the convention that $h(\mathbf{0}) = 0$). Thanks to the strictly convex negative entropy term, the regularized optimal transport problem has a unique minimizer, denoted $P_\lambda^\star$. It can be recovered using a fixed-point algorithm as demonstrated by Sinkhorn (see, e.g., [60,69]). The regularized transport cost $\mathbf{MK}_\lambda(a, b)$ is thus referred to as the *Sinkhorn distance*.

4.1.1 Interpretation

Another way to express the negative entropic term is:

$$-h : p \in \mathbb{R}_+^{M^2} \mapsto \mathbf{KL}(p\|\mathbf{1}_{M^2}) \in \mathbb{R},$$

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 [23].

Hence, one would like to consider large values of λ to be close to the original Monge–Kantorovich distance, but low enough to deal with feature intrinsic variability and noise. As detailed after, the estimation of this regularized distance involves terms of the form $\exp(-\lambda C)$. For numerical reasons, the process is limited to low values of λ in practice, so that the Sinkhorn distances are rough approximations of the Monge–Kantorovich distances.

4.1.2 Structure of the Solution

First, using the same vectorial notation as in Eq. (24), the Sinkhorn distance (30) reads as

$$\mathbf{MK}_\lambda(\alpha) := \inf_{\substack{p\in\mathbb{R}^{M^2} \\ \text{s.t. } p\geq\mathbf{0},\ L^\mathsf{T}p=\alpha}} \langle p,\ c+\tfrac{1}{\lambda}\log p\rangle. \tag{31}$$

As demonstrated in [23], when writing the Lagrangian of this problem with a multiplier β to take into account the constraint $L^\mathsf{T}p=\alpha$, we can show that the respective solutions $P_\lambda^\star$ and $p_\lambda^\star$ of problem (30) and (31) can be written as

$$\log p_\lambda^\star = \lambda(L\beta - c) - \mathbf{1} \text{ with } \beta = \begin{bmatrix} x \\ y \end{bmatrix}$$
$$\Leftrightarrow (\log P_\lambda^\star)_{i,j} = \lambda(x_i + y_j - C_{i,j}) - 1.$$

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.

4.1.3 Sinkhorn Algorithm

Sinkhorn showed [69] that the alternating normalization of rows and columns of any positive matrix M converges to a unique bistochastic matrix

$$P = \operatorname{diag}(x) M \operatorname{diag}(y)$$

with the desired marginals. The corresponding fixed-point iteration algorithm can be used to find the solution $P_\lambda^\star$: setting $M_\lambda = e^{-\lambda C}$, one has

$$P_\lambda^\star = \operatorname{diag}(x^{(\infty)}) M_\lambda \operatorname{diag}(y^{(\infty)})$$
$$\text{with } x^{(t+1)} = \frac{a}{M_\lambda\, y^{(t)}} \text{ and } y^{(t+1)} = \frac{b}{M_\lambda^\mathsf{T}\, x^{(t+1)}},$$

where a and b are the desired marginals of the matrix. With this result, one can design a fast algorithm to compute the regularized optimal transport plan, the Sinkhorn distance or its derivative, as shown in [23,24].

4.2 Conjugate Sinkhorn Distance $\mathbf{MK}_\lambda^*$

Now, in order to use the Sinkhorn distance in the algorithm (PD), we need to compute its Legendre–Fenchel transform, which expression has been studied in [24].

Proposition 1 (Cuturi–Doucet) *The convex conjugate of $\boldsymbol{MK}_\lambda(\alpha)$ defined in (31) reads*

$$\boldsymbol{MK}_\lambda^*(\beta) = \tfrac{1}{\lambda}\langle q_\lambda(\beta), \mathbf{1}\rangle$$
$$\textit{with } q_\lambda(\beta) := e^{\lambda(L\beta - c)-\mathbf{1}}. \tag{32}$$

With matrix notations, writing $\beta = (\beta_1; \beta_2)$, we have equivalently $\mathbf{MK}_\lambda^*(\beta) = \frac{1}{\lambda}\langle Q_\lambda(\beta),\ \mathbf{1}\rangle$

$$\text{with } Q_\lambda(\beta_1,\beta_2) := e^{\lambda(\beta_1\mathbf{1}^\top + \mathbf{1}\beta_2^\top - C)-\mathbf{1}}. \tag{33}$$

This simple expression of the Legendre–Fenchel transform is C^∞, but unfortunately, its gradient is not Lipschitz continuous. We propose two solutions in order to recover the setting of the general primal dual problem (13) and be able to minimize the segmentation energy involving Sinkhorn distances. We either define a new normalized Sinkhorn distance $\mathbf{MK}_{\lambda,\leq N}$ (Sect. 4.3), whose gradient is Lipschitz continuous (Sect. 4.4), or we rely on the use of the proximity operator of $\mathbf{MK}_\lambda$ (Sect. 4.6). A discussion follows to compare the two approaches.

4.3 Normalized Sinkhorn Distance $\mathbf{MK}_{\lambda,\leq N}$

As the set Δ 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

$$\Delta_{\leq N} := \left\{ a, b \in \mathbb{R}_+^M \;\middle|\; \langle a,\ \mathbf{1}_M\rangle = \langle b,\ \mathbf{1}_M\rangle \leq N \right\}. \tag{34}$$

Next, as the total mass transported by an admissible matrix $P_\lambda^\star$ from a to b is not known in our segmentation problem (i.e., $\langle P_\lambda^\star,\ \mathbf{1}\rangle \leq N$), we use a slight variant of the entropic regularization:

$$\begin{aligned} \tilde{h}(p) &:= Nh\left(\tfrac{p}{N}\right) = -N\,\mathbf{KL}\left(\tfrac{p}{N}\|\mathbf{1}\right) \\ &= -\langle p,\ \log p\rangle + \langle p,\ \mathbf{1}\rangle \log N \ \geq 0. \end{aligned} \tag{35}$$

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

$$\begin{aligned} &\boldsymbol{MK}_{\lambda,\leq N}(\alpha) \\ &:= \inf_{\substack{p\in\mathbb{R}^{M^2} \\ \text{s.t. } p\geq\mathbf{0},\ L^\mathsf{T}p=\alpha}} \langle p,\ c+\tfrac{1}{\lambda}\log p - \tfrac{1}{\lambda}\log N\ \mathbf{1}\rangle + \chi_{\Delta_{\leq N}}(\alpha) \end{aligned} \tag{36}$$

reads

$$\boldsymbol{MK}_{\lambda,\leq N}^*(\beta) = \begin{cases} \frac{N}{\lambda}\langle q_\lambda(\beta),\mathbf{1}\rangle & \text{if } \langle q_\lambda(\beta),\mathbf{1}\rangle \leq 1, \\ \frac{N}{\lambda}\log\langle q_\lambda(\beta),\mathbf{1}\rangle + \frac{N}{\lambda} & \text{if } \langle q_\lambda(\beta),\mathbf{1}\rangle \geq 1, \end{cases} \tag{37}$$

using the vector-valued function $q_\lambda(.) \mapsto e^{\lambda(L.-c)-\mathbf{1}}$ defined in (32).

Proof See "Appendix 3".

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

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

By the virtue of Corollary 1, we can express the gradient of $\mathbf{MK}^*_{\lambda,\leq N}$ which is continuous

$$\nabla \mathbf{MK}^*_{\lambda,\leq N}(\beta) = N \, \frac{\big(Q_\lambda(\beta)\,\mathbf{1};\, Q_\lambda(\beta)^\top \mathbf{1}\big)}{\max\{1, \langle Q_\lambda(\beta), \mathbf{1}\rangle\}}, \tag{38}$$

using expression of $Q_\lambda(\beta)$ from Eq. (33). In vectorial notation, we have a simpler expression using matrix L:

$$\nabla \mathbf{MK}^*_{\lambda,\leq N}(\beta) = N \, \frac{L^\top q_\lambda(\beta)}{\max\{1, \langle q_\lambda(\beta), \mathbf{1}\rangle\}}. \tag{39}$$

We emphasize here that, when restricting the Sinkhorn distance to histograms on the probability simplex $\mathcal{S}_{1,M}$ (i.e., the special case where $N = 1$ and $\langle Q_\lambda(\beta), \mathbf{1}\rangle = 1$), or more generally on $\Delta_{\leq 1}$, we retrieve a similar expression than the one originally derived in [25].

Finally, the normalized Sinkhorn transport cost can be used in the generic optimization scheme due to the following property.

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

Proof See "Appendix 4".

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

The binary segmentation problem (10) with normalized Sinkhorn transport cost can be expressed as:

$$\begin{aligned} \min_u \quad & \chi_{[0,1]^N}(u) + \rho \mathrm{TV}(u) + \mathbf{MK}_{\lambda,\leq N}(Hu, Au) \\ & + \mathbf{MK}_{\lambda,\leq N}(H(\mathbf{1}-u), B(\mathbf{1}-u)). \end{aligned} \tag{40}$$

Using the Fenchel transform, the problem (40) can be reformulated as:

$$\begin{aligned} \min_u \max_{\substack{v \\ p_A, q_A \\ p_B, q_B}} \quad & \langle Du, v\rangle + \chi_{[0,1]^N}(u) - \chi_{\|\cdot\|_{\infty,2}\leq\rho}(v) \\ & + \langle Hu, p_A\rangle + \langle H(\mathbf{1}-u), p_B\rangle \\ & + \langle Au, q_A\rangle + \langle B(\mathbf{1}-u), q_B\rangle \\ & - \mathbf{MK}^*_{\lambda,\leq N}(p_A, q_A) - \mathbf{MK}^*_{\lambda,\leq N}(p_B, q_B), \end{aligned}$$

and can be optimized with the algorithm (PD), setting $S_1^* = 0$ and $S_2^* = \mathbf{MK}^*_{\lambda,\leq N}$. Using proposition 2, ∇G^* is a Lipschitz continuous function with constant $L_{G^*} = 2L_{\mathbf{MK}^*}$. The definition of the diagonal preconditioners is the same as in problem (12), using Formula (16). The extension to multiphase segmentation is also analogous to problem (12) (see the last paragraph of Sect. 2.6).

Advantages and Drawbacks It has been shown in [25] that, aside from an increased robustness to outliers, the smoothing of the optimal transport cost offers significant numerical stability. However, the optimization scheme may be slow due to the use of the unnormalized simplex $\Delta_{\leq N}$. In practice, the Lipschitz constant L_{G^*} will be large for high-resolution images (i.e., large values of N) and for tight approximations of the $\mathbf{MK}$ cost (i.e., $\lambda \gg 1$). It will lead to low values of time step parameters in (16) and involve a slow explicit gradient ascent in the dual update of the algorithm (PD). In such a case, we can resort to the alternative scheme proposed hereafter.

4.6 Primal–dual Formulation of $\mathbf{MK}_\lambda$

An alternative optimization of (40) consists in using the proximity map of G^*. Since we cannot compute such an operator for $\mathbf{MK}^*_\lambda$ in a closed form, or in an effective way, we resort instead to a biconjugation, as previously done in Sect. 3.2.2.

Biconjugation For consistency with the previous section, we consider again the normalized entropy (35) to define the regularized cost function $\mathbf{MK}_{\lambda,N}$ on the set Δ in order to obtain the factor N:

$$\mathbf{MK}_{\lambda,N}(\alpha) := \inf_{\substack{p\in\mathbb{R}^{M^2} \\ \text{s.t. } p\geq\mathbf{0},\ L^\mathsf{T}p=\alpha}} \langle p,\ c + \tfrac{1}{\lambda}\log(p/N)\rangle. \tag{41}$$

Simple calculations show that the dual conjugate in Eq. (32) becomes

$$\mathbf{MK}^*_{\lambda,N}(\beta) = \tfrac{N}{\lambda}\langle q_\lambda(\beta), \mathbf{1}_N\rangle. \tag{42}$$

Introducing the dual conjugate function

$$g^*_\lambda(q) := \tfrac{N}{\lambda}\langle e^{\lambda(q-c)-\mathbf{1}},\ \mathbf{1}\rangle \tag{43}$$

that is convex and continuous, we have

$$\mathbf{MK}^*_{\lambda,N}(\beta) = g^*_\lambda(L\beta) = \max_r\ \langle r,\ L\beta\rangle - g_\lambda(r) \tag{44}$$

and

$$\begin{aligned} \mathbf{MK}_{\lambda,N}(\alpha) &= \max_\beta\ \langle \alpha,\ \beta\rangle - g^*_\lambda(L\beta) \\ &= \min_r \max_\beta\ \langle \alpha - L^\mathsf{T}r,\ \beta\rangle + g_\lambda(r). \end{aligned}$$

This reformulation, combined with the following expression of the proximity function of g_λ, enables to solve efficiently the segmentation problem with $\mathbf{MK}_{\lambda,N}$.

Proposition 3 *The proximity operator of the function g_λ, the conjugate of g_λ^* defined in Eq. (43), is*

$$Prox_{\tau g_\lambda}(r) = \frac{\tau}{\lambda} W\left(\frac{\lambda}{\tau} N e^{\lambda(\frac{r}{\tau}-c)-\mathbf{1}}\right) \tag{45}$$

where W is the Lambert function*, such that $w = W(z)$ is solution of $we^w = z$. The solution is unique as $z \geq 0$.*

Proof See "Appendix 5".

Remark 2 Note that the Lambert function can be evaluated very fast, using an efficient parallel algorithm that requires a few iterations [21].

Segmentation Problem Plugging the Formula (43) into Eq. (12) provides the following primal dual problem

$$\begin{aligned}\min_{\substack{u\\ r_A,r_B}} \max_{\substack{v\\ p_A,q_A\\ p_B,q_B}} \quad & \langle Du,\, v\rangle - \chi_{\|.\|_{\infty,2}\leq\rho}(v) \\ & + \langle Au,\, q_A\rangle + \langle B(\mathbf{1}-u),\, q_B\rangle \\ & + \langle Hu,\, p_A\rangle + \langle H(\mathbf{1}-u),\, p_B\rangle \\ & - \langle r_A,\, L\begin{bmatrix} p_A \\ q_A\end{bmatrix}\rangle - \langle r_B,\, L\begin{bmatrix} p_B \\ q_B\end{bmatrix}\rangle \\ & + \chi_{[0,1]^N}(u) + g_\lambda(r_A) + g_\lambda(r_B)\end{aligned} \tag{46}$$

Again, we can use a variant of the primal–dual algorithm described in (PD), augmented by primal variables r_A and r_B. The operator $\mathcal{K}$ is the same than in Formula (28). The proximity function $\text{Prox}_{\tau R}$ corresponds to

$$\begin{aligned}&\text{Prox}_{\chi_{[0,1]^N}(u)+\tau g_\lambda(r_A)+\tau g_\lambda(r_B)}(u, r_A, r_B) \\ &\quad = \left(\text{Proj}_{[0,1]^N}(u); \text{Prox}_{\tau g_\lambda}(r_A); \text{Prox}_{\tau g_\lambda}(r_B)\right).\end{aligned}$$

4.7 Comparison of the Two Approches

In the previous sections, two variants of the entropic regularized transportation problem have been introduced: $\mathbf{MK}_{\lambda,\leq N}$ in (36) and $\mathbf{MK}_{\lambda,N}$ in (41). We underline the fact that, while having different definitions, these two metrics provide the same numerical result for any of the segmentation problems investigated in this paper, as the corresponding primal–dual optimal solutions satisfy the same property (i.e., the mass of histograms in Δ cannot exceed the total number of pixels N) for which the metrics behave identically.

5 Segmentation Experiments

The following table gives an overview of the optimal transport distances S studied previously and its respective properties.

S	**MK**	$\mathbf{MK}_{\lambda,N}$	$\mathbf{MK}_{\lambda,\leq N}$
Section	Section 3	Section 4.1	Section 4.3
Definition	(23)	(41)	(36)
Smooth	No	Yes	Yes
Dual conjugate	(32)	(42)	(37)
Bi-conjugate (BC)	(26)	(44)	–
Dual gradient (G)	–	–	(38)
Lipschitz gradient	No	No	Yes
PD formulation	(27)	(46)	(12)

Experimental Setting In this experimental section, exemplar regions are defined by the user with scribbles (see, for instance, Fig. 2) or bounding boxes (Fig. 8). These regions are only used to build prior histograms, so erroneous labeling is tolerated. The 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 features ($\mathcal{F} = \text{Id}$ and $n = d = 3$) or the gradient norm of color features ($\mathcal{F} = \|D.\|$ computed on each color channel, so that $n = 3$). 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 [65]. Approximately 1 minute is required to run 500 iterations and segment a 1 megapixel color image.

To account for the case where a region boundary coincides with the image border $\partial\Omega$, we enlarge the size of the domain Ω by 1 pixel and we force variable u to be null on the border. That way, the model does not favor regions that lie across the boundary.

Throughout the experiments, the diagonal preconditioning is defined using Formula (16) with $r = \delta = 1$. We have observed an impressive convergence acceleration (approximately 3 orders of magnitude) due to preconditioning.

Projection Onto the Simplex For the multiphase segmentation described in paragraph 2.4, the projector onto the discrete probability set $\mathcal{S}_{1,K}$ of the K variables u_k [see also relation (17)] can be computed pointwise in linear time complexity, see, for instance, [20].

Thresholding As previously stated, the segmentation map u^* obtained by minimizing the functional (10) is not binary in general. The final segmentation is obtained by thresholding the global minima $u^\star$ with $t = \frac{1}{2}$ (see Sect. 2.7). This leads to the best compromise in our experiments, as illustrated in Fig. 3 that shows the influence of the threshold t used to get a binary segmentation.

5.1 Regularized Versus Non-regularized MK Distances

As previously discussed in Sect. 4.7, the solutions when using the gradient of $\mathbf{MK}_{\lambda,\leq N}$ or the proximity operator based on

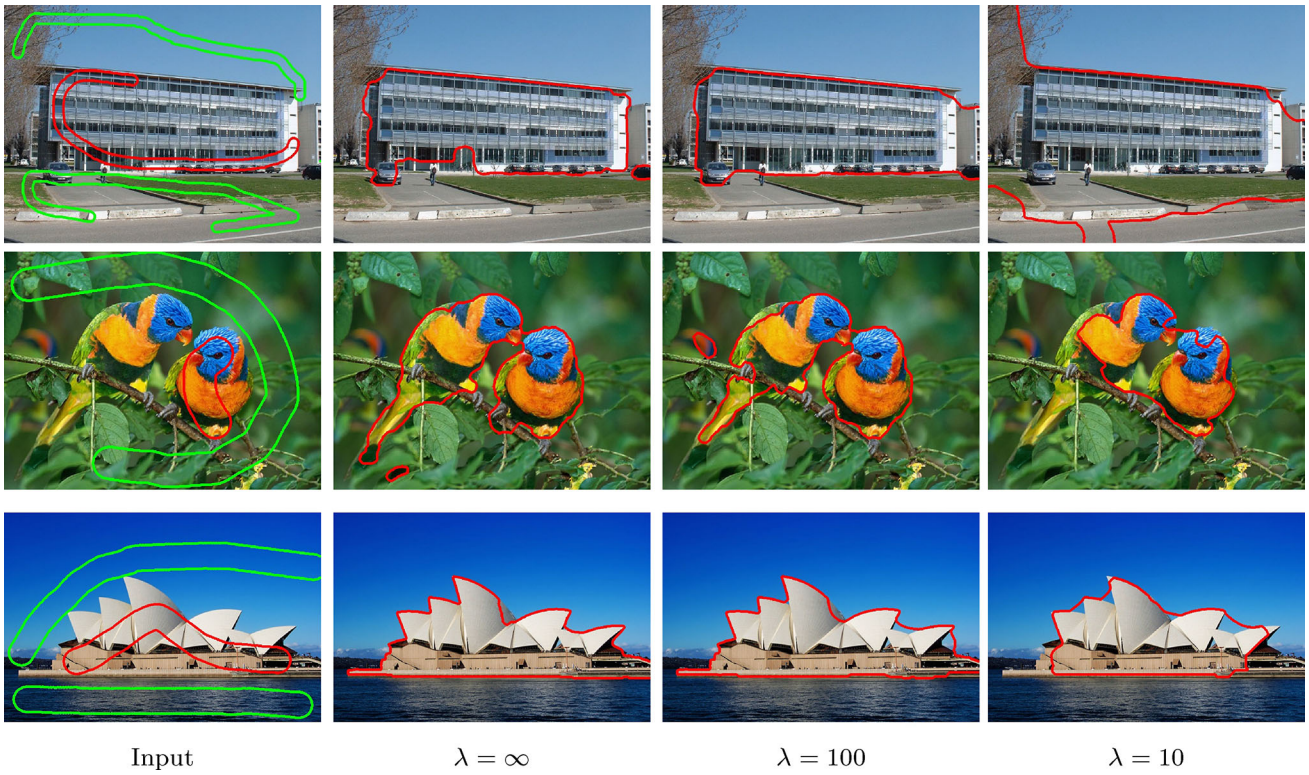


Fig. 2 Comparison of segmentations obtained from the proposed models with $\mathbf{MK}_\lambda$ cost function. The input images are partially labeled by the user, and the corresponding areas are used to compute the reference color distributions a and b. The segmented regions, obtained from the thresholded solution, are contoured in *red*. Different values of the regularization parameter λ of the transport cost are used, $\lambda = \infty$ corresponding to the non-regularized model (Color figure online)

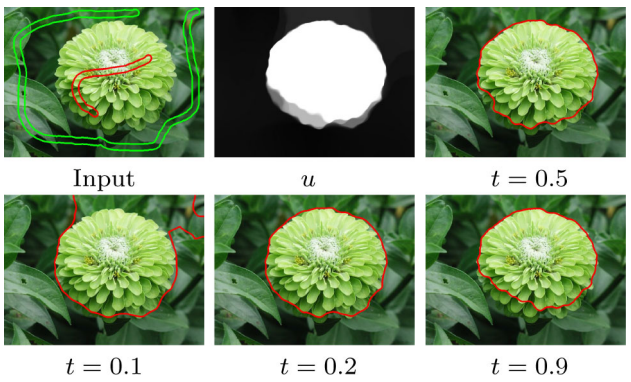


Fig. 3 Illustration of the image segmentation method with **MK** transport cost and the influence of the final threshold parameter t on the segmentation result. The user defines scribbles which indicate the object to be segmented (here in *red*) and the background to be discarded (in *green*). The output image u is a regularized weight map that gives the probability of a pixel to belong to the object. This probability map u is finally thresholded with a parameter t to segment the input image into a region $R_t(u)$, which contour is displayed in *red* (Color figure online)

$\mathbf{MK}_{\lambda,N}$ are the same when $N = |\Omega|$, even if the respective optimization schemes are different. As a consequence, we simply denote by $\mathbf{MK}_\lambda$ when referring to these methods. We also indicate **MK** or $\lambda = \infty$ when not using any regularization.

We first illustrate the influence of the parameter λ in the regularized distance $\mathbf{MK}_\lambda$. Figure 2 gives a comparison between the non-regularized model ($\lambda = \infty$) with the regularized model. The non-regularized model gives excellent segmentation results, but it requires high-dimensional representation through biconjugation (27). For regularized distances, we considered the smooth low-dimensional formulation (40), but the high- dimensional representation (46) gives similar results. One can see that setting a large value of λ gives good results. On the other hand, using a very small value of λ always yields poor segmentation results. In practice, if not specified otherwise, we consider $\lambda = 100$ in our experiments, as higher values may lead to numerical issues (floating point precision).

5.2 Comparisons with Other Segmentation Models Including Wasserstein Distance

We first exhibit the advantage of considering global data terms over histograms, such as in Eq. (2). We present a comparison with the convex model proposed in [45] that includes a local data term over color distributions:

(a) I, a, b **(b)** Local [45] **(c)** Global

Fig. 4 Comparison with a local model. **a** Input image and regions where reference distributions a and b are estimated. The segmentation fails for the local histogram model (**b**) as it classifies the *orange* areas in the first class and the darker ones in the second class. The global histograms on the segmented zones are not close to the given ones, contrary to the global model (**c**) (Color figure online)

$$\tilde{E}(u) = \rho \mathrm{TV}(u) + \sum_{x\in\Omega} \mathbf{MK}(a, h_{V(x)})u(x) + \mathbf{MK}(b, h_{V(x)})(1-u(x))$$

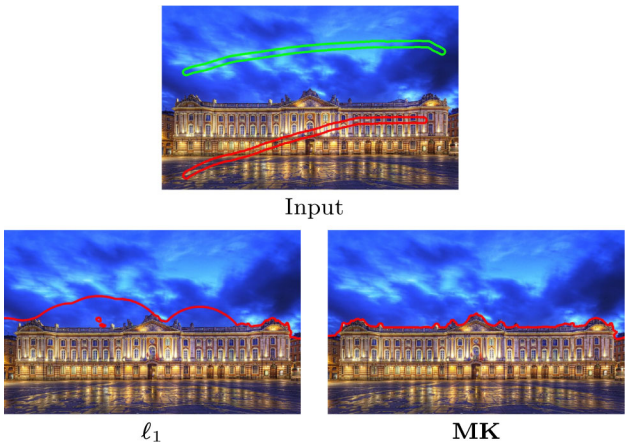
where $h_{V(x)}$ is the color distribution over the neighborhood $V(x)$ of pixel x. This model, that can be optimized globally [78], measures the local color distribution of the image with respect to the reference foreground and background distributions a and b. As illustrated in Fig. 4, such a local model is not able to perform a global segmentation. Here the orange colors are more probable in the region related to the butterfly, so in small neighborhoods the flowers are classified as the butterfly, and the darker regions are segmented as being in the background. This example illustrates the importance of considering global histogram comparisons to get a global segmentation of an image. Indeed, the global distance between histograms (c) is able to recover the butterfly, whereas the local approach (b) completely fails. Local approaches are therefore only relevant when the local histograms correctly approximate the global ones.

Next, we illustrate the advantage of having a convex model that does not depend on the initialization. We compare our results with the ones obtained with the Wasserstein active contour method proposed in [51]. Such an approach consists in deforming a level set function in order to minimize globally the Wasserstein distance between the reference histograms and the one induced by the segmentation. To evolve the level set, this formulation requires complex shape gradient computations. As illustrated in Fig. 5, even if this model can give good segmentations that are close to the ones we obtained in Fig. 4c, its initialization may be a critical step as really different segmentations are obtained with very similar initializations.

We show comparisons with the global model of [79] that includes ℓ_1 distances between histograms. 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 similarity between different bins, which can lead to incorrect segmentation. This is illustrated with the blue colors

(a) Initialization **(b)** Result

Fig. 5 Comparison with a non-convex model. The Wasserstein active contours method [51], initialized in two different ways in (**a**), provides the segmentations presented in (**b**), illustrating the non-convexity of the model. When carefully initialized, it leads to a segmentation close to the one obtained with our global approach (see Fig. 4c)



Input

ℓ_1 **MK**

Fig. 6 Robustness of **MK** compared to ℓ_1. The *blue colors* that are not in the reference histograms are considered correctly as background with **MK** distance, but as foreground with the ℓ_1 model where no color comparison is performed (Color figure online)

in Fig. 6 where the ℓ_1 distance leads to an incorrect segmentation by associating some blue tones to the building area.

Figure 7 finally illustrates the robustness of the optimal transport distance compared to bin-to-bin ℓ_1 distance. A small variation of the reference histograms may involve a large change in the final segmentation with ℓ_1 distance, whereas segmentations obtained with **MK** or regularized $\mathbf{MK}_\lambda$ are stable.

The robustness is further illustrated in Figure 8. It is indeed possible to use a prior histogram from a different image, even with a different clustering of the feature space. This is not

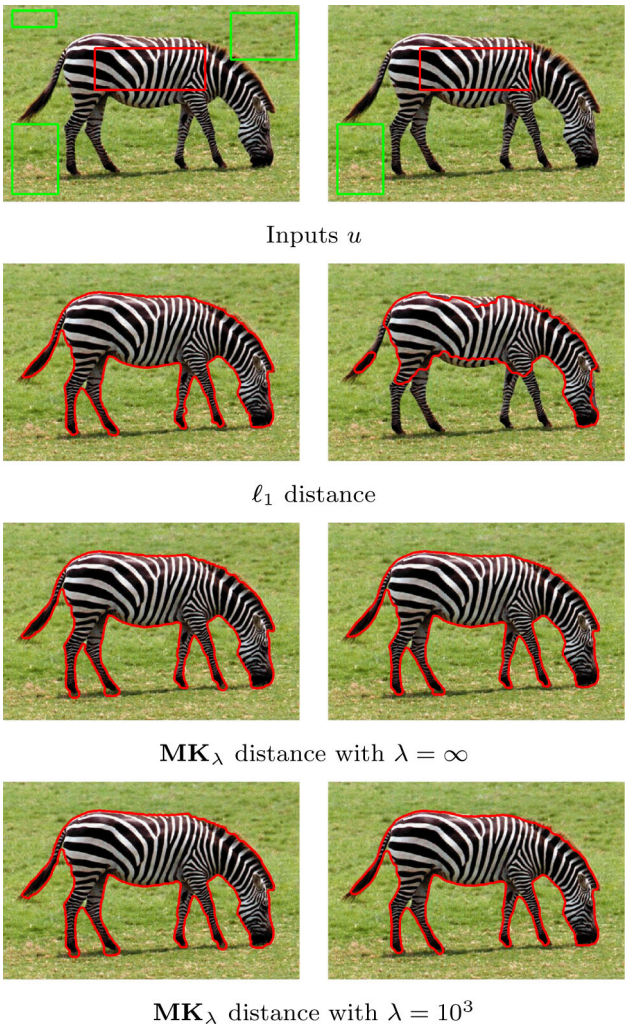


Fig. 7 Comparison of the results obtained from the proposed segmentation models (using $\mathbf{MK}_\lambda$ distances) together with the ℓ_1 distance used in [79], for different initialization. The same regularization parameter ρ is used for every segmentation. Note that the optimal transport similarity measure is a more robust statistical metric between histograms than ℓ_1

possible with a bin-to-bin metric, such as ℓ_1, which requires the same clustering.

5.3 Convergence Behavior of the Different Models

We now compare the behavior of the different metrics S and algorithms with respect to the evolution of their corresponding energy functions and computational cost. The experiments have been performed for two different histogram quantifications: $M_\mathcal{X} = 8^3$ and $M_\mathcal{X} = 16^3$ with the zebra image of Fig. 7 of dimension 730×487 pixels.

As illustrated in Fig. 9, the energy of the ℓ_1 model of [79] converges very fast in practice. This model relies on the formulation (12) where the proximity operator of the dual of the ℓ_1 metric can be computed in a closed form. Con-

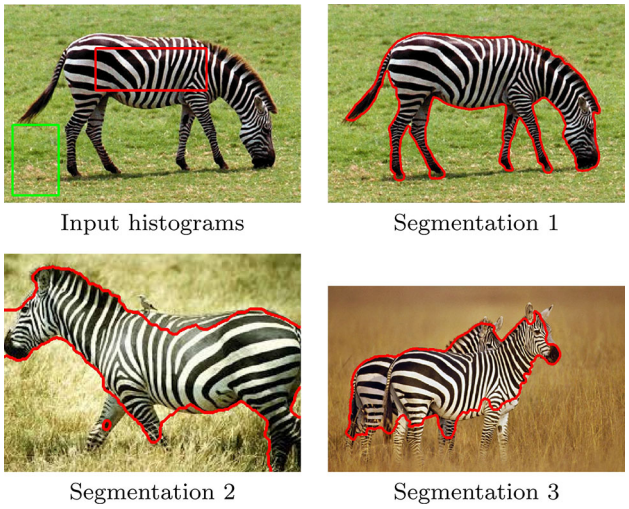


Fig. 8 Illustration of the interest of optimal transport cost for the comparison of histograms. Its robustness makes it possible to use prior histograms from different images (in this example, histograms are estimated from image 1 and used to segment all images)

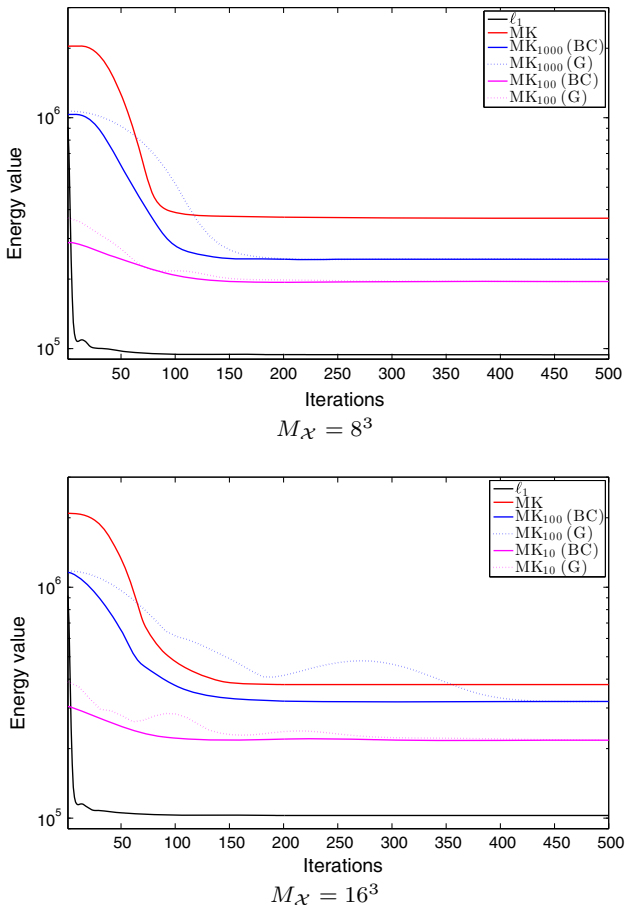


Fig. 9 Evolution in semi logarithmic scale of the functional values for the different metrics S: ℓ_1, **MK** and $\mathbf{MK}_\lambda$ with $\lambda = 100$ and 1000. The optimization for metric $\mathbf{MK}_\lambda$ is done with the two proposed formulations: gradient descent (G) on the dual (12) and biconjugation (BC) with model (46)

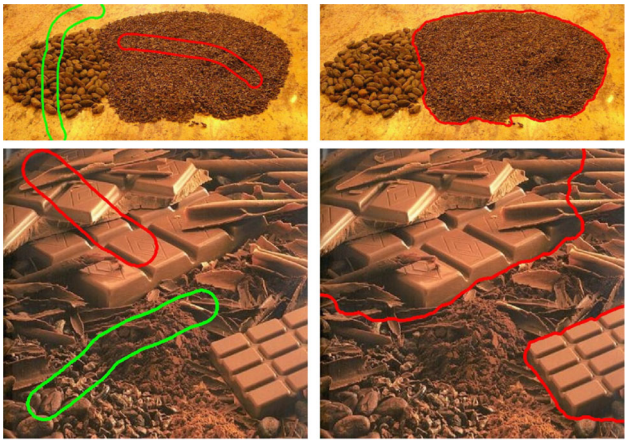


Fig. 10 Texture segmentation using joint histograms of color gradient norms. In this example, only gradient information is taken into account, illustrating the versatility of the optimal transport framework

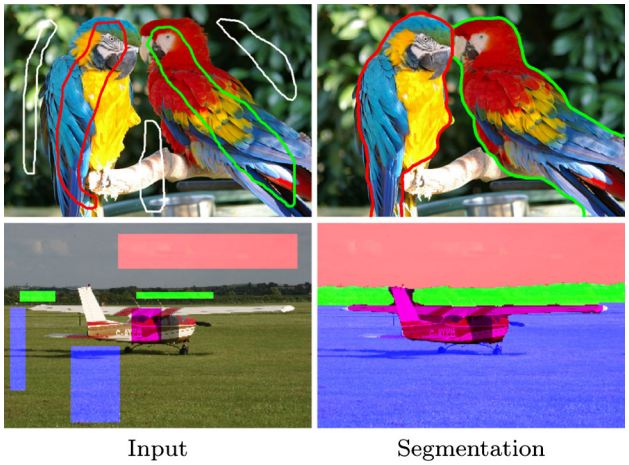


Fig. 11 Multiphase segmentation with three regions (*first line*) and four regions (*second line*)

sidering biconjugation (BC) through formulations (27) and (46) also involves an almost monotonous convergence of the energy function for both the **MK** metric and the entropy-regularized ones $\mathbf{MK}_\lambda$. Finally, for the entropy-regularized metric $\mathbf{MK}_\lambda$, we can also rely on the formulation (12), by realizing a gradient (G) descent on the dual of $\mathbf{MK}^*_\lambda$. In this case, the convergence is slower compared to biconjugation. As expected, it can also be observed that for larger values of λ (and thus large value of the Lipschitz constant of $\nabla\mathbf{MK}_\lambda$ from Proposition 2), the convergence of the gradient descent is additionally slowed down by small time steps.

From the computational point of view, the implementation of (G) and (BC) using, respectively, (46) and (12) yields approximately the same computation time for one iteration (7 s for 200 iterations and $M_\mathcal{X} = 16^3$ with MATLAB implementation). Indeed, it is mainly due to the manipulation of the cost matrix $M_\lambda = e^{-\lambda C}$ that has to be stored in memory. Note, however, that the computation of the gradient of $\nabla\mathbf{MK}^*_{\lambda,\leq N}$ relies on the matrix Q_λ (Eq. (33)) that can be computed efficiently using convolution. More specifically, when using the cost matrix $C_{ij} = ||X_i - X_j||^2$ on a cartesian grid, we could rely on separable 1D gaussian convolutions as demonstrated in [70] so that the gradient descent approach (G) would became significantly faster that the (BC) approach. This solution has not been investigated here because the cost matrix in our setting is a concave function of the distance, as detailed in the beginning of the section.

5.4 Generalizations

Texture segmentation examples are presented in Fig. 10 where the proposed method is perfectly able to recover the textured areas. We considered here the joint histogram of gradient norms on the three color channels. The complexity of the algorithm is the same as for color features, as long as we use the same number of clusters to quantize the feature space.

We finally present experiments involving more than two partitions in Fig. 11. In the first line, three regions are considered for the background and the two parrots. Even if the two parrots share similar colors, the model is able to find a global segmentation of the image into three regions. In the second line of Fig. 11, we considered four regions for the sky, the grass, the forest and the plane. The approach is able to deal with the color variations inside each class in order to perform a correct segmentation.

6 Unsupervised Co-segmentation

In this section, we extend our framework to the unsupervised co-segmentation of multiple images. We invite the reader to see the following reference [74] for a complete review.

6.1 Co-segmentation of Two Images

We first consider two images I^1 and I^2 the domain of which are, respectively, Ω_1 and Ω_2 composed of N_1 and N_2 pixels. Assuming that the images contain a common object, the goal is now to jointly segment them without any additional prior.

Model for Two Images To that end, following the model used in [71,74], we aim at finding the *largest* regions that have similar feature distributions. To define the segmentation maps u^1 and u^2 related to each image, we consider the following model first investigated in [43,61], denoting $u = (u^1; u^2)$:

$$J'(u) := S\left(H_1u^1, H_2u^2\right) + \sum_{k=1}^{2} \rho\mathrm{TV}(u^k) - \delta\|u^k\|_1 \qquad (47)$$

where for a nonnegative variable u^k we have a total mass $\|u^k\|_1 = \langle u^k, \mathbf{1}_{N_k}\rangle$. When $u^k \in \{0,1\}^{N_k}$, this term corresponds to the area of the region segmented in image I^k. Such a ballooning term encourages the segmentation of large regions. Without this term, a global optimum would be given by $u^k = \mathbf{0}$.

Following definition (5), the operator $H_k(i,x)$ is 1 if pixel $I^k(x)$ belongs to the cluster $\mathcal{C}_{\mathcal{X}_k}(i)$ and 0 otherwise. As before, the value of the segmentation variables u^k is relaxed into the convex intervals $[0,1]^{N_k}$.

In [74], several cost functions S are benchmarked for the model defined in Eq. (47), such as ℓ_1 and ℓ_2. It is demonstrated that ℓ_1 performs the best. In [71], the Wasserstein distance is used again to measure the similarity of the two histograms. In the following, we investigate the use of these two metrics in our setting.

Property of the Segmented Regions To begin with, note that when considering optimal transport cost to define S, one has to constrain the histograms to have the same mass, i.e., $(H_1u^1, H_2u^2) \in \Delta$. When using assignment operators such as in (5), this boils down to constrain the segmentation variables to have the same mass, i.e., $\langle u^1, \mathbf{1}_{N_1}\rangle = \langle u^2, \mathbf{1}_{N_2}\rangle$.

When looking for a binary solution, this condition implies that the two regions corresponding to the segmentation of each image have the exact same number of pixels. This means that the model is not robust to small-scale changes in appearance with optimal cost transport, while this is the case when using the ℓ_1 metric, as demonstrated in [74].

One simple way to remove this restriction from the model is to use the same formulation introduced in Sect. 2.3.3 for segmentation. Unfortunately, this boils down to defining the similarity measure with

$$S(H_1u_1\langle u_2, \mathbf{1}\rangle, H_2u_2\langle u_1, \mathbf{1}\rangle)$$

which is obviously non-convex and does not fit the optimization framework used in this paper.

It is not the first time that the conservation of mass in the optimal transport framework is reported to limit its practical interest in imaging problem, and several variations have been proposed to circumvent it. Without entering into details, a common idea is to discard the conservation of mass when the two histograms are *unbalanced* and to define alternative transport maps that may create or annihilate mass. As an example, a solution might be to transport the minimum amount of mass between the unnormalized histograms and penalize the remaining, as done by the distance introduced in [50] and similarly in [31]. Other models have recently been investigated, such as in [40], and [18,29]. However, the application of such metric for our setting is far from being straightforward and need careful analysis that is left for future work.

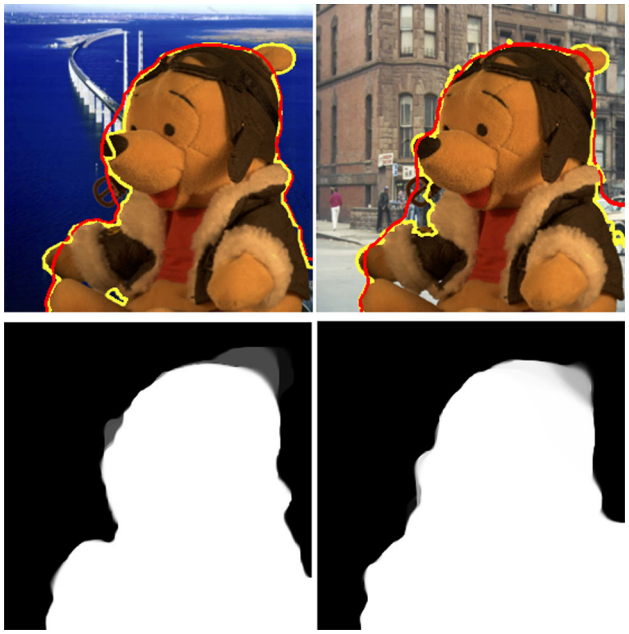


Fig. 12 Co-segmentation and optimal transport with or without entropic regularization. The results obtained with the model (47) with the entropic regularization (in *red*), that approximate the method of [71] (in *yellow*, image courtesy of [71]). The estimated segmentation maps u^k are binary almost everywhere. The threshold $t = \frac{1}{2}$ is used to obtain the final co-segmentation regions (Color figure online)

Optimization To solve the relaxed problem

$$\min_{u\in[0,1]^{N_1+N_2}} J'(u)$$

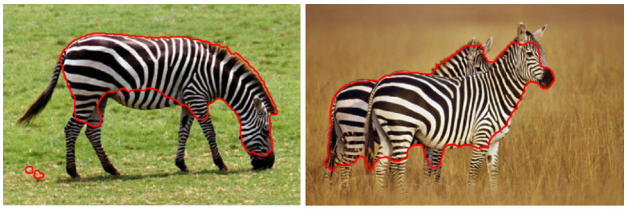
using either ℓ_1, **MK** or $\mathbf{MK}_\lambda$ as a cost function S, we rely again on the primal–dual formulation (13) of the problem and the algorithm (PD). Notice that the minor difference with previous segmentation problems is the presence of the linear ballooning term and that there is only one dissimilarity term.

Experiments We now illustrate the behavior of this model. Again, we underline that the convex cosegmentation model (47) is not new, as our approach only differ algorithmically from [71,74] when using ℓ_1 or **MK** as a cost function. Therefore, we only focus on results obtained using optimal transport with entropic regularization (setting $\lambda = 100$).

In the synthetic experiment of Fig. 12 containing exactly the same object with different backgrounds, we compare our approach with the one of [71] that does not include entropic regularization.[1] Both methods give similar co-segmentations.

When considering images where the common object has a similar scale in both images, Fig. 13 shows that the condition

[1] Another main difference is that [71] makes use of super-pixel representation to reduce the complexity, whereas we use a pixel representation.



u^1 u^2

Fig. 13 Co-segmentation of two zebras with the model (47). The convex constraint $\|u^1\|_1 = \|u^2\|_1$ enforces the segmented regions to have the same area. As the obtained result is not binary, the areas may be different after the thresholding

$\|u^1\|_1 = \|u^2\|_1$ is not restrictive and our method still gives acceptable co-segmentations.

Nevertheless, in a more general setting, we cannot expect the common objects to have the same scale in all images. We leave the study of alternative optimal transport-based distance such as [18,29] for future work.

6.2 Co-segmentation of P Images

We consider now the generalization of the previous co-segmentation model to an arbitrary number of $P \geq 2$ images.

Complexity A natural extension of (47) for more than two images would be to penalize the average dissimilarity between all image pairs, writing, for instance,

$$J''(u) = \sum_{k=1}^{P-1} \left(\sum_{l=k+1}^{P} S(H_k u^k, H_l u^l) \right) + \rho \mathrm{TV}(u^k) - \delta \|u^k\|_1 + \chi_{[0,1]^{N_k}}(u^k) \tag{48}$$

which would require to compute $\binom{P}{2}$ similarity terms $S(H_k u^k, H_l u^l)$. However, the complexity of such a model scales quadratically with the number of images P which is not desirable.

To that end, we consider instead the following barycentric formulation which scales linearly with P

$$J'''(u, b) = \sum_{k=1}^{P} S(H_k u^k, b) + \rho \mathrm{TV}(u^k) - \delta \|u^k\|_1 + \chi_{[0,1]^{N_k}}(u^k) + \chi_{\geq \mathbf{0}}(b) \tag{49}$$

where b is the estimated barycentric distribution between the histograms of the segmented regions in all images. Note that in the *unsupervised* case studied in this section, the barycenter b has to be estimated jointly with segmentation variables.

Model Properties with Optimal Transport Costs Combining the $O(P^2)$ model (48) or the linear barycentric formulation (49) with optimal transport-based cost functions **MK** and $\mathbf{MK}_\lambda$ results in the scaling problem previously reported with model (47), as definitions of **MK** (23) and $\mathbf{MK}_\lambda$ (30) constraint each pair of histograms to have the same mass. Non-convex formulation or unbalanced transport costs [18,29] should again be considered as a solution, but do not fit in the proposed optimization framework.

Model properties with ℓ_1 In order to circumvent this issue, we consider the ℓ_1 case instead, i.e., using

$$S(H_k u^k, b) = \|H_k u^k - b\|_1.$$

As stated before in paragraph 3.1, the ℓ_1 distance between normalized histograms can be seen as the total variation distance, a specific instance of the **MK** distance that naturally extends to unnormalized histograms. In this setting, recall that histograms must have the same number of bins M and the exact same feature clusters $\mathcal{C}_\mathcal{X}$ (see Sect. 2.3.2).

Optimization The minimization of the functional (49) for fixed histograms $H_k u^k$ boils down to the smooth Wasserstein barycenter problem studied in [25]. The authors investigate the dual formulation of this primal problem and show that it can be solved by a projected gradient descent on the dual problem. They resort to a splitting strategy, defining P primal histogram variables $(b_k \in \mathbb{R}^M)_{k=1\ldots P}$ with the linear constraint $b_1 = \cdots = b_P$. Using a similar approach, one obtain the following primal–dual formulation

$$\min_{u,b} \max_{h,v} \sum_{k=1}^{P} \langle H_k u^k - b,\, h_k \rangle - \delta \langle u^k,\, \mathbf{1}_{N_k} \rangle + \langle D_k u^k,\, v_k \rangle - \chi_{\|h_k\|_\infty \leq 1} - \chi_{\|v_k\|_{\infty,2} \leq \rho} + \chi_{u^k \in [0,1]^{N_k}} + \chi_{b \geq \mathbf{0}}$$

which fits the canonic form of Problem (13). In the above equation, D_k refers to the finite difference operator in the grid Ω_k of image I^k.

The algorithm (PD) requires to compute the Euclidean projector onto the nonnegative orthant (29) and on the ℓ_∞ unit ball [similarly to Eq. (14)]

$$\mathrm{Proj}_{\|.\|_\infty \leq 1}(h)(i) = \frac{h(i)}{\max\{|h(i)|, 1\}}.$$

Experiments To illustrate the validity of the proposed ℓ_1 model, we repeat the toy experiment of Fig. 12 in Fig. 14, where the same object is shown in two images with different backgrounds. While there is no more constraint on the size of the objects to segment as in Eq. (47), the model (49) is still able to get a good co-segmentation of the data.

In Figs. 15 and 16, we illustrate how this new model is able to segment objects of different scales in $P = 5$ images and to learn its representative color distribution. The area of the zebra can be very different in each image.

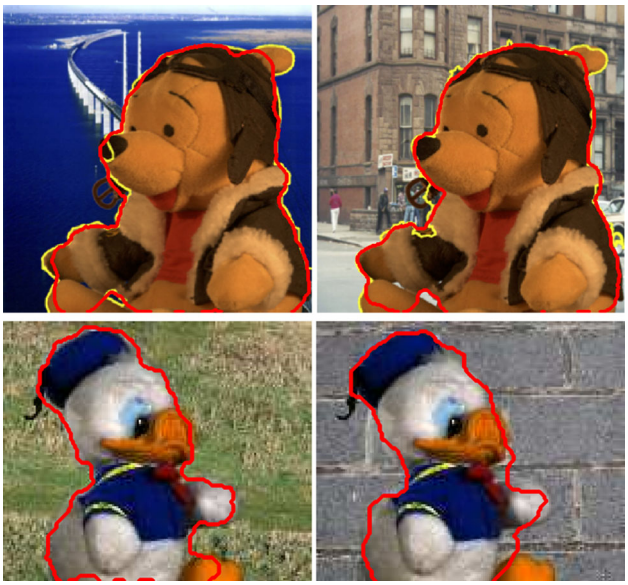


Fig. 14 Examples of co-segmentation of $P = 2$ images with model (49). The objects to segment have the same scale. Comparing with the first line with Fig. 12, results similar to the model (47) and the method [71] (in *yellow*), whereas no constraint is considered here on the size of the regions (Color figure online)

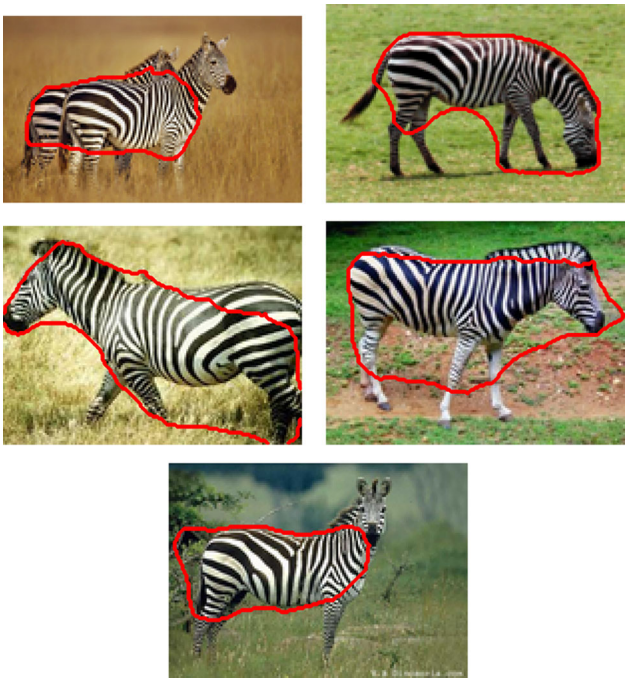


Fig. 15 Co-segmentation of $P = 5$ images with model (49)

Note that for simplicity the same regularization parameter ρ and ballooning parameter δ are used for all images in the model (49), whereas one should tune separately these parameters according to each image in order to obtain more accurate co-segmentations. Moreover, it seems necessary to add information on the background for improving these

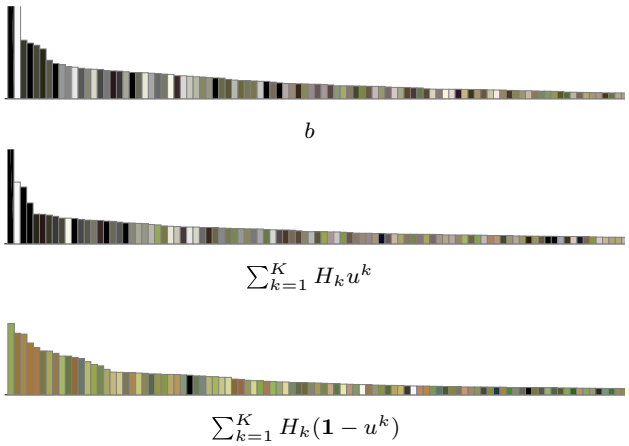


Fig. 16 Estimated color histograms from the experiment in Fig. 15. For visualization, each histogram is sorted and displayed for the 100 most frequent colors. The first color palette is the learnt barycenter histogram b in the model (49). It mainly contains shades of *black* and *white* colors corresponding to the zebras. For a comparison, the histograms for the segmentation variables u^k and $\mathbf{1} - u^k$ are displayed, as illustrated in Fig. 1. As expected, the second color palette corresponding to the segmented object is very similar to the histogram b, while the histogram of the background is completely different (Color figure online)

results. In the previous examples of Figs. 14 and 15, the backgrounds were sufficiently different in the different images to be discarded by the model. As soon as the backgrounds of the co-segmented images contain very similar information (for instance, gray regions outside the gnome in images of Fig. 17), the ballooning term in Eq. (49) forces the model to include these areas in the co-segmentations.

7 Conclusion and Future Work

In this work, several formulations have been proposed to incorporate transport-based cost functions in convex variational models for supervised segmentation and unsupervised co-segmentation. The proposed framework includes entropic regularization of the optimal transport distance and deals with multiple objects as well as collections of images.

As already demonstrated for the image segmentation problem, optimal transport yields significant advantages when comparing feature distributions from segmented regions (robustness with respect to histogram quantization, the possibility to incorporate prior information on features into the transport cost, definition of a metric between different types of features, etc.). When considering entropic regularization, the algorithmic scheme is yet very similar to the one obtained for the ℓ_1 norm, at the only expense of requiring more memory. We observed, as reported in [25], that such regularization offers practical acceleration for small histograms but also improves robustness to outliers (such as noise or rare fea-

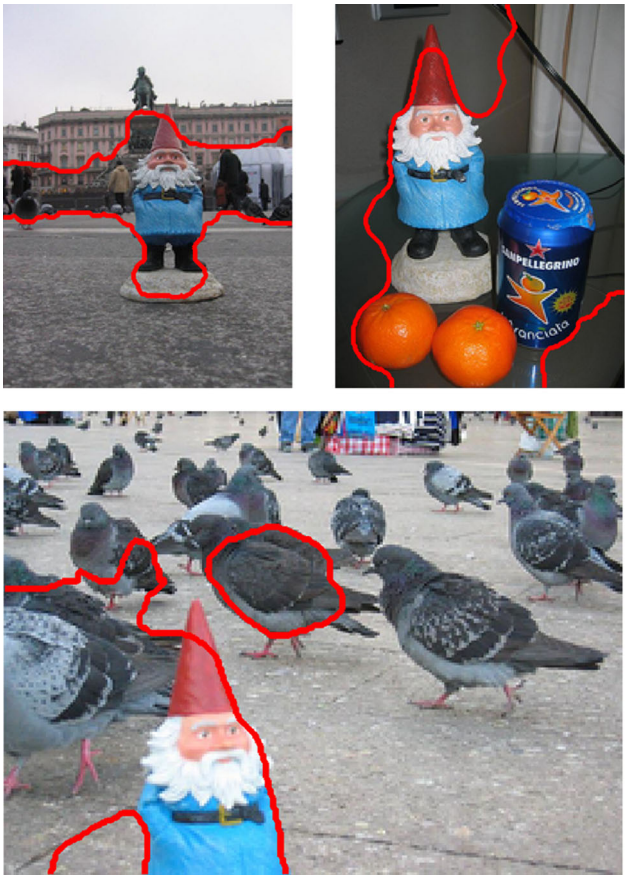


Fig. 17 Example of a failure case of unsupervised co-segmentation with a dramatic change of the object of interest (illumination and scale), where the backgrounds have similar *color* distribution (Color figure online)

tures). However, we also emphasized that large regularization significantly decreases the quality of the results.

The main limitation highlighted and discussed in this work is the lack of scale invariance for the unsupervised co-segmentation problem due to the convex formulation. In comparison, non-convex formulations of optimal transport with probability distribution such as [51] yield such invariance, while usual cost functions such as ℓ_1 offer some robustness [74]. A promising perspective to overcome this restriction is the use of the unbalanced optimal transport framework recently studied in [18,29].

In the future, other potential improvements of the model will be investigated, such as the optimization of the final thresholding operation, some variable metrics during optimization, the use of a capacity transport constraint relaxation [28], the incorporation of other statistical features and the integration of additional priors such as the shape of the objects [67,68].

Acknowledgements The authors acknowledge support from the CNRS in the context of the "Défi Imag'In" project CAlcul des VAriations pour L'Imagerie, l'Edition et la Recherche d'Images (CAVALIERI). This study has been carried out with financial support from the French State, managed by the French National Research Agency (ANR) in the frame of the Investments for the future Programme IdEx Bordeaux (ANR-10-IDEX-03-02). The authors would like to thank Gabriel Peyré and Marco Cuturi for sharing their preliminary work and Jalal Fadili for fruitful discussions on convex optimization.

Appendix 1: Norm of $\mathcal{K}$

Proof We recall that $\mathcal{K} = [H^\mathsf{T}, A^\mathsf{T}, -H^\mathsf{T}, -B^\mathsf{T}, D^\mathsf{T}]^\mathsf{T}$ so that

$$\|\mathcal{K}\| \leq 2\|H\| + \|A\| + \|B\| + \|D\|.$$

For rank one operator A and B, we can write

$$\|B\| = \max_{\|x\|=1} \|Bx\| = \max_{\|x\|=1} |\langle x, \mathbf{1}_N \rangle| \|b\| = \|\mathbf{1}_N\| \|b\| = \sqrt{N}\|b\|$$

and for histogram b subject to $b \geq 0$ and $\langle b, \mathbf{1} \rangle = 1$, we have $\|b\|_2 \leq \|b\|_1 = 1$ and the same for histogram a. For the hard assignment operator

$$\|H\| \leq \|H\|_F = \sqrt{\sum_{x\in\Omega} \sum_{i=1}^{M} H_{i,x}^2} = \sqrt{N}$$

where the equality holds when assignment matrix is $H = \mathbf{1}_N^\mathsf{T}$. The finite difference operator D verifies (see, for instance, [13])

$$\|D\| \leq \sqrt{8}.$$

Finally, we obtain

$$\|\mathcal{K}\| \leq 4\sqrt{N} + \sqrt{8}.$$

Addendix 2: Proof of the Special ℓ_1 case

Proof We consider here that two histograms $(a, b) \in \Delta$ and the cost matrix C such that $C_{i,j} = 2(1 - \delta_{ij})$. This cost is only null when not moving mass (that is $j = i$) and constant otherwise, so that an optimal matrix $P \in \mathcal{P}(a, b)$ must satisfy $P_{i,i} = \min(a_i, b_i)$ to minimize the transport cost $\langle P, C \rangle$. Therefore, we have that

$$\begin{aligned}
\mathbf{MK}(a, b) &= 2 \sum_i \sum_{j\neq i} P_{i,j} = 2 \sum_i (a_i - \min(a_i, b_i)) \\
&= 2 \sum_i (a_i - b_i) \mathbb{1}_{a_i > b_j} \\
&= 2 \sum_j (b_j - a_j) \mathbb{1}_{a_i < b_j}.
\end{aligned}$$

The last equality is obtained by symmetry. Then, adding the two last equalities, we obtain the desired result

$$\mathbf{MK}(a,b) = \|a-b\|_1.$$

Addendix 3: Proof of Corollary 1

Proof For sake of simplicity, the notation $\mathbf{1}$ without subindex refers to either the vector $\mathbf{1}_M$ or a matrix $\mathbf{1}_{M\times M}$ depending on the context. Let us consider the problem (36) using Lagrangian multipliers:

$$\begin{aligned}
&\mathbf{MK}_{\lambda,\leq N}(a,b)\\
&= \inf_{P\in\mathcal{P}(a,b)} \langle P, C\rangle + \tfrac{1}{\lambda}\langle P, \log(P/N)\rangle + \chi_{\Delta_{\leq N}}(a,b)\\
&= \min_{P\geq \mathbf{0}} \max_{\substack{u,v\\ w\leq 0}} \left\{ \begin{array}{l} \langle P, \frac{1}{\lambda}\log P/N + C\rangle + \langle u, a - P\mathbf{1}\rangle \\ +\langle v, b - P^\mathsf{T}\mathbf{1}\rangle + w(N - \mathbf{1}^\mathsf{T}P\mathbf{1}) \end{array} \right\}\\
&= \max_{\substack{u,v\\ w\leq 0}} \left\{ \begin{array}{l} \langle u, a\rangle + \langle v, b\rangle + Nw \\ +\min_{P\geq \mathbf{0}} \langle P, \frac{1}{\lambda}\log P/N + C - u\mathbf{1}^\mathsf{T} - \mathbf{1}v^\mathsf{T} - w\,\mathbf{1}\rangle \end{array} \right\}
\end{aligned}$$

using the fact that the (normalized) negative entropy is continuous and convex. The corresponding Lagrangian is

$$\begin{aligned}
&\mathcal{L}(P,u,v,w) = \langle u, a\rangle + \langle v, b\rangle + Nw\\
&\quad \frac{1}{\lambda}\langle P, \log P - \log N + \lambda(C - u\mathbf{1}^\mathsf{T} - \mathbf{1}v^\mathsf{T} - w\mathbf{1})\rangle.
\end{aligned}$$

The first-order optimality condition $\partial_P \mathcal{L}(P^\star, u, v, w) = \mathbf{0}$ gives:

$$\log P^\star - \log N + \lambda(C - u\mathbf{1}^\mathsf{T} - \mathbf{1}v^\mathsf{T} - w\mathbf{1}) + \mathbf{1} = \mathbf{0},$$

that is

$$P^\star_{i,j} = Ne^{-1+\lambda w}\, e^{-\lambda(C_{i,j}-u_i-v_j)} \geq 0. \tag{50}$$

Using this expression in $\mathcal{L}(P^\star, u, v, w)$

$$\begin{aligned}
&\mathbf{MK}_{\lambda,\leq N}(a,b)\\
&= \max_{u,v,w\leq 0} \langle u, a\rangle + \langle v, b\rangle + Nw - \frac{1}{\lambda}\langle P^\star, \mathbf{1}\rangle\\
&= \max_{u,v} \left\{ \begin{array}{l} \langle u, a\rangle + \langle v, b\rangle \\ +\max_{w\leq 0} Nw - \frac{N}{\lambda}e^{-1+\lambda w}\sum_{i,j} e^{-\lambda(C_{i,j}-u_i-v_j)} \end{array} \right\}
\end{aligned} \tag{51}$$

Observe that the expression of $P^\star(u,v,w)$ in (50), that becomes $P^\star = Ne^{\lambda w}Q_\lambda(u,v)$ using definition (33), is scaled by the Lagrangian variable w which corresponds to the constraint $\langle P^\star, \mathbf{1}\rangle \leq N$. We consider now whether or not this equality holds.

Case 1 $\langle P^\star, \mathbf{1}\rangle = N$. Let us first consider the case where the constraint is saturated, that is when $w < 0$ due to the complementary slackness property. The maximum of function $f(w) = Nw - \frac{N}{\lambda}e^{\lambda w}\langle Q_\lambda, \mathbf{1}\rangle$ which is concave ($\partial^2_w f(w) < 0$), is obtained for $w^\star$ subject to

$$e^{\lambda w^\star} = \frac{1}{\langle Q_\lambda, \mathbf{1}\rangle} = \left(\sum_{i,j} e^{-\lambda(C_{i,j}-u_i-v_j)-1}\right)^{-1} \leq 1.$$

One can check that the equality $\sum_{i,j} P^\star_{i,j} = N$ is indeed satisfied. In addition, the maximum of f verifies

$$\begin{aligned}
f(w^\star) &= Nw^\star - \frac{N}{\lambda} = \frac{N}{\lambda}(\lambda w^\star - 1) = \frac{N}{\lambda}\log e^{\lambda w^\star - 1}\\
&= -\frac{N}{\lambda}\log\langle Q_\lambda, \mathbf{1}\rangle - \frac{N}{\lambda}
\end{aligned}.$$

The problem (51) becomes

$$\begin{aligned}
&\mathbf{MK}_{\lambda,\leq N}(a,b)\\
&\quad = \max_{u,v}\ \langle u, a\rangle + \langle v, b\rangle - \frac{N}{\lambda}\log\sum_{i,j} e^{-\lambda(C_{i,j}-u_i-v_j)}.
\end{aligned} \tag{52}$$

From the definition of the Legendre–Fenchel transformation, this implies that

$$\mathbf{MK}_{\lambda,\leq N}(u,v) = \left(\frac{N}{\lambda}\log\left(\sum_{i,j} e^{-\lambda(C_{i,j}-u_i-v_j)}\right)\right)^*.$$

As these functions are convex, proper and lower semi-continuous, we have that $\mathbf{MK}^{**}_{\lambda,N} = \mathbf{MK}_{\lambda,N}$ which concludes the proof for the case $\langle Q_\lambda(u,v), \mathbf{1}\rangle \geq 1$.

Case 2 $\langle P^\star, \mathbf{1}\rangle < N$. Now we consider the case where the constraint is not saturated, i.e., $w = 0$. The expression of $P^\star(u,v,w)$ in (50) becomes $P^\star = NQ_\lambda(u,v)$. Going back to relation (51), we have directly

$$\mathbf{MK}_{\lambda,\leq N}(u,v) = \max_{u,v}\ \langle u, a\rangle + \langle v, b\rangle - \frac{N}{\lambda}\langle Q_\lambda(u,v), \mathbf{1}\rangle$$

which concludes the proof for the case $\langle Q_\lambda(u,v), \mathbf{1}\rangle \leq 1$.

Addendix 4: Proof of Proposition 2

Proof The derivative $\nabla\mathbf{MK}^*_{\lambda,\leq N}(X)$ with $X = (u; v)$ is Lipschitz continuous *iff* there exists $L_{\mathbf{MK}^*} > 0$ such that

$$\|\nabla\mathbf{MK}^*_{\lambda,\leq N}(X) - \nabla\mathbf{MK}^*_{\lambda,\leq N}(X')\| \leq L_{\mathbf{MK}^*}\|X - X'\|.$$

We denote as $\mathcal{U}$ the set of vectors $X = (u; v) \in \mathbb{R}^{2M}$ such that $\langle Q_\lambda(u, v), \mathbf{1}\rangle > 1$ [where Q_λ is defined in Eq. (33)]. We denote $\mathcal{V} = U^c$ the complement of $\mathcal{U}$ in $\mathbb{R}^{2M}$. Observe that the set $\mathcal{V}$ is convex, as it corresponds to a sublevel set of the convex function $\mathbf{MK}^*_{\lambda,\leq N}$.

Due to the expression of the gradient in (38) that is different on sets $\mathcal{U}$ and $\mathcal{V}$, we will consider the following three cases.

Case 1 Let $X, X' \in \mathcal{U}$. As $\nabla\mathbf{MK}^*_{\lambda,\leq N}$ is differentiable in the set $\mathcal{U}$, it is a Lipschitz function *iff* the norm of the Hessian matrix $\mathcal{H}$ of $\mathbf{MK}^*_{\lambda,\leq N}$ is bounded. Denoting $\{\mu_i\}_{i=1}^{2M}$ the eigenvalues of $\mathcal{H}$, its spectral norm is defined as $\|\mathcal{H}\|_2 = \max_i |\mu_i|$. Moreover, as $\mathbf{MK}^*_{\lambda,\leq N}$ is convex, all eigenvalues are nonnegative. Thus, we have that the norm of $\mathcal{H}$ is bounded by its trace: $\|\mathcal{H}\| \leq \mathrm{Tr}(\mathcal{H}) = \sum_i \mu_i = \sum_i \mathcal{H}_{ii}$.

The Hessian matrix $\mathcal{H}$ of $\mathbf{MK}^*_{\lambda,\leq N}$ is defined as:

$$\mathcal{H} = \begin{bmatrix} \mathcal{H}^{11} & \mathcal{H}^{12} \\ \mathcal{H}^{21} & \mathcal{H}^{22} \end{bmatrix}, \quad \text{with} \quad \mathcal{H}^{mn} = \nabla_m \nabla_n^\mathsf{T} \mathbf{MK}^*_{\lambda,\leq N}.$$

Combining Eqs. (38) and (33), we have

$$\begin{aligned}(\nabla_1 \mathbf{MK}^*_{\lambda,\leq N}(u, v))_i &= \partial_{u_i} \mathbf{MK}^*_{\lambda,\leq N}(u, v) \\ &= N \left(\frac{Q_\lambda(u, v)\mathbf{1}}{\langle Q_\lambda(u, v), \mathbf{1}\rangle} \right)_i = N \frac{\sum_l e^{-\lambda(C_{i,l} - u_i - v_l)}}{\sum_{k,l} e^{-\lambda(C_{k,l} - u_k - v_l)}}.\end{aligned}$$

Hence, the diagonal elements of the matrix read

$$\begin{aligned}\mathcal{H}^{11}_{ii} &= \partial^2_{u_i} \mathbf{MK}^*_{\lambda,\leq N}(u, v) \\ &= \lambda N \frac{\sum_l e^{-\lambda(C_{i,l} - u_i - v_l)} \sum_{k \neq i,l} e^{-\lambda(C_{k,l} - u_k - v_l)}}{\left(\sum_{k,l} e^{-\lambda(C_{k,l} - u_k - v_l)}\right)^2}, \\ \mathcal{H}^{22}_{jj} &= \partial^2_{v_j} \mathbf{MK}^*_{\lambda,\leq N}(u, v) \\ &= \lambda N \frac{\sum_k e^{-\lambda(C_{k,j} - u_k - v_j)} \sum_{k,l \neq j} e^{-\lambda(C_{k,l} - u_k - v_l)}}{\left(\sum_{k,l} e^{-\lambda(C_{k,l} - u_k - v_l)}\right)^2}.\end{aligned}$$

Computing the trace of the matrix $\mathcal{H}$, we obtain

$$\|\mathcal{H}\| \leq \mathrm{Tr}(\mathcal{H}) = \sum_i \mathcal{H}^{11}_{ii} + \sum_j \mathcal{H}^{22}_{jj} \leq 2\lambda N.$$

Case 2 We now consider $X, X' \in \mathcal{V}$. In this case, we have for $X = (u, v)$:

$$\mathbf{MK}^*_{\lambda,\leq N}(u, v) = N \langle Q_\lambda(u, v), \mathbf{1}\rangle = \frac{N}{\lambda} \sum_{i,j} e^{-1-\lambda(C_{i,j} - u_i - v_j)}.$$

As the second partial derivative with respect to u_i reads $\partial^2_{u_i} \mathbf{MK}^*_{\lambda,\leq N}(u, v) = N\lambda \sum_j e^{-1-\lambda(C_{i,j} - u_i - v_j)}$, the trace of the Hessian matrix is:

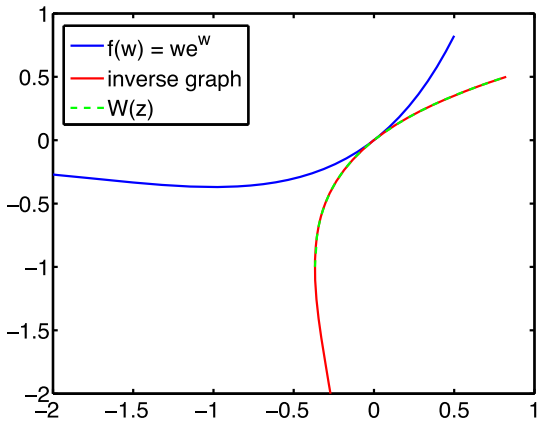


Fig. 18 Graph of the Lambert function $W(z)$

$$\begin{aligned}\mathrm{Tr}(\mathcal{H}) &= N\lambda \left(\sum_{i,j} e^{-1-\lambda(C_{i,j} - u_i - v_j)} + \sum_{j,i} e^{-1-\lambda(C_{i,j} - u_i - v_j)} \right) \\ &= 2\lambda N \langle Q_\lambda(u, v), \mathbf{1}\rangle \leq 2\lambda N,\end{aligned}$$

since $(u, v) \in \mathcal{V}$.

Case 3 We consider $X \in \mathcal{U}$ and $X' \in \mathcal{V}$. As $\mathcal{V}$ is a convex set and $\mathcal{U}$ its complement, we denote as Y the vector that lies in the segment $[X; X']$ and belongs to ∂V, the boundary of $\mathcal{V}$, that is satisfying $\langle Q_\lambda(Y), \mathbf{1}\rangle = 1$. We thus have $\|X - Y\| + \|X' - Y\| = \|X - X'\|$ so that

$$\begin{aligned}&\|\nabla\mathbf{MK}^*_{\lambda,\leq N}(X) - \nabla\mathbf{MK}^*_{\lambda,\leq N}(X')\| \\ &\quad \leq \|\nabla\mathbf{MK}^*_{\lambda,\leq N}(X) - \nabla\mathbf{MK}^*_{\lambda,\leq N}(Y)\| \\ &\qquad + \|\nabla\mathbf{MK}^*_{\lambda,\leq N}(X') - \nabla\mathbf{MK}^*_{\lambda,\leq N}(Y)\| \\ &\quad \leq 2\lambda N(\|X - Y\| + \|X' - Y\|) = 2\lambda N \|X - X'\|\end{aligned} \tag{53}$$

which concludes the proof.

Addendix 5: Proof of Proposition 3

Proof We are interested in the proximity operator of g, which convex conjugate is $g^*(q) = \frac{N}{\lambda}\langle e^{\lambda(q-c)-\mathbf{1}}, \mathbf{1}\rangle$. First, notice that the proximity operator of g can be computed easily from the proximity operator of g^* through Moreau's identity:

$$\mathrm{Prox}_{\tau g}(p) + \tau \mathrm{Prox}_{g^*/\tau}(p/\tau) = p \quad \forall\, \tau > 0, \forall\, p.$$

We now recall that the Lambert function W is defined as:

$$z = we^w \Leftrightarrow w = W(z)$$

where w can take two real values for $z \in]-\frac{1}{e}, 0]$, and only one on $]0, \infty[$, as illustrated in Fig. 18. As z will always be positive in the following, we do not consider complex values.

The proximity operator of g^* at point p reads (as g^* is convex, the Prox operator is univalued):

$$\begin{aligned}\mathrm{Prox}_{\tau g^*}(p) &= q^\star \in \operatorname*{argmin}_q \tfrac{1}{2\tau}\|q-p\|^2 + g^*(q)\\ &= \operatorname*{argmin}_q \sum_k \tfrac{1}{2\tau}(q_k-p_k)^2 + \tfrac{N}{\lambda}e^{\lambda(q_k-c_k)-1}.\end{aligned}$$

This problem is separable and can be solved independently $\forall k$. Taking the derivative of the previous relation with respect to q_k, the first-order optimality condition gives:

$$\begin{aligned}&q_k^\star - p_k + \tau N e^{\lambda(q_k^\star-c_k)-1} = 0\\ \Leftrightarrow &(p_k-q_k^*)e^{-\lambda q_k^\star} = \tau N e^{-\lambda c_k - 1}\\ \Leftrightarrow &\lambda(p_k-q_k^*)e^{\lambda(p_k-q_k^*)} = \lambda\tau N e^{\lambda(p_k-c_k)-1}.\end{aligned}$$

Using the Lambert function, we get:

$$\begin{aligned}&\lambda(p_k-q_k^\star) = W(\lambda\tau N e^{\lambda(p_k-c_k)-1})\\ \Leftrightarrow &q_k^* = p_k - \frac{1}{\lambda}W(\lambda\tau N e^{\lambda(p_k-c_k)-1}).\end{aligned}$$

The proximity operator of g^*/τ thus reads

$$\mathrm{Prox}_{g^*/\tau}(p) = p - \frac{1}{\lambda}W\left(\tfrac{\lambda}{\tau}N e^{\lambda(p-c)-1}\right),$$

hence

$$\mathrm{Prox}_{\tau g}(p) = \frac{\tau}{\lambda}W\left(\tfrac{\lambda}{\tau}N e^{\lambda(\frac{p}{\tau}-c)-1}\right).$$

References

1. Aubert, G., Barlaud, M., Faugeras, O., Jehan-Besson, S.: Image segmentation using active contours: calculus of variations or shape gradients? SIAM Appl. Math. **63**(6), 2128–2154 (2003)
2. Aubert, G., Kornprobst, P.: Mathematical Problems in Image Processing, volume 147 of Applied Mathematical Sciences. Springer, New York (2002)
3. Aujol, J.-F., Aubert, G., Blanc-Féraud, L.: Wavelet-based level set evolution for classification of textured images. IEEE Trans. Image Process. **12**(12), 1634–1641 (2003)
4. Ayed, I., Chen, H., Punithakumar, K., Ross, I., Li, S.: Graph cut segmentation with a global constraint: recovering region distribution via a bound of the bhattacharyya measure. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'10), pp. 3288–3295 (2010)
5. Berkels, B.: An unconstrained multiphase thresholding approach for image segmentation. In: Scale Space and Variational Methods in Computer Vision (SSVM'09), vol. 5567, pp. 26–37 (2009)
6. Bonneel, N., Rabin, J., Peyré, G., Pfister, H.: Sliced and radon Wasserstein barycenters of measures. J. Math. Imaging Vis. **51**(1), 22–45 (2015)
7. Bonneel, N., van de Panne, M., Paris, S., Heidrich, W.: Displacement interpolation using Lagrangian mass transport. In: ACM Transactions on Graphics (SIGGRAPH Asia'11), vol. 30(6) (2011)
8. Brown, E., Chan, T.F., Bresson, X.: Completely convex formulation of the Chan–Vese image segmentation model. Int. J. Comput. Vis. **98**(1), 103–121 (2012)
9. Brox, T., Rousson, M., Deriche, R., Weickert, J.: Unsupervised segmentation incorporating colour, texture, and motion. Comput. Anal. Images Patterns **2756**, 353–360 (2003)
10. Brox, T., Weickert, J.: Level set segmentation with multiple regions. IEEE Trans. Image Process. **15**(10), 3213–3218 (2006)
11. Brune, C.: 4D imaging in tomography and optical nanoscopy. PhD thesis, University of Münster, Germany (2010)
12. Chambolle, A.: An algorithm for mean curvature motion. Interfaces Free Boundaries Math. Model. Anal. Computat. **6**(2), 195–218 (2004)
13. Chambolle, A.: An algorithm for total variation minimization and applications. J. Math. Imaging Vis. **20**(1), 89–97 (2004)
14. Chambolle, A., Darbon, J.: On total variation minimization and surface evolution using parametric maximum flows. Int. J. Comput. Vis. **84**(3), 288–307 (2009)
15. Chambolle, A., Pock, T.: A first-order primal–dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**, 120–145 (2011)
16. Chambolle, A., Pock, T.: On the ergodic convergence rates of a first-order primal–dual algorithm. J. Math. Program. **159**(1–2), 253–287 (2016)
17. Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Trans. Image Process. **10**(2), 266–277 (2001)
18. Chizat, L., Schmitzer, B., Peyré, G., Vialard, F.-X.: Unbalanced optimal transport: geometry and Kantorovich formulation. arXiv:1508.05216 (2015)
19. Combettes, P., Condat, L., Pesquet, J.-C., Vu, B.: A forward–backward view of some primal–dual optimization methods in image recovery. In: IEEE International Conference on Image Processing (ICIP'14), pp. 4141–4145 (2014)
20. Condat, L.: Fast projection onto the simplex and the l1 ball. Math. Program. **158**(1), 575–585 (2016)
21. Corless, R., Gonnet, G., Hare, D., Jeffrey, D., Knuth, D.: On the lambert W function. Adv. Comput. Math. **5**(1), 329–359 (1996)
22. Cremers, D., Rousson, M., Deriche, R.: A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape. Int. J. Comput. Vis. **72**(2), 215 (2007)
23. Cuturi, M.: Sinkhorn distances: lightspeed computation of optimal transport. In: Neural Information Processing Systems (NIPS'13), pp. 2292–2300 (2013)
24. Cuturi, M., Doucet, A.: Fast computation of Wasserstein barycenters. In: International Conference on Machine Learning (ICML'14), pp. 685–693 (2014)
25. Cuturi, M., Peyré, G.: A smoothed dual approach for variational Wasserstein problems. SIAM J. Imaging Sci. **9**(1), 320–343 (2016)
26. Delon, J.: Movie and video scale-time equalization application to flicker reduction. IEEE Trans. Image Process. **15**(1), 241–248 (2006)
27. Digne, J., Cohen-Steiner, D., Alliez, P., de Goes, F., Desbrun, M.: Feature-preserving surface reconstruction and simplification from defect-laden point sets. J. Math. Imaging Vis. **48**(2), 369–382 (2014)
28. Ferradans, S., Papadakis, N., Peyré, G., Aujol, J.-F.: Regularized discrete optimal transport. SIAM J. Imaging Sci. **7**(3), 1853–1882 (2014)
29. Frogner, C., Zhang, C., Mobahi, H., Araya-Polo, M., Poggio, T.: Learning with a wasserstein loss. arXiv: 1506.05439 (2015)
30. Gorelick, L., Schmidt, F.R., Boykov, Y., Delong, A., Ward, A.: Segmentation with non-linear regional constraints via line-search cuts. In: European Conference on Computer Vision (ECCV'12), pp. 583–597 (2012)

31. Gramfort, A., Peyré, G., Cuturi, M.: Fast optimal transport averaging of neuroimaging data. In: Information Processing in Medical Imaging (IPMI'15) (2015)
32. Haker, S., Zhu, L., Tannenbaum, A., Angenent, S.: Optimal mass transport for registration and warping. Int. J. Comput. Vis. **60**(3), 225–240 (2004)
33. Herbulot, A., Jehan-Besson, S., Duffner, S., Barlaud, M., Aubert, G.: Segmentation of vectorial image features using shape gradients and information measures. J. Math. Imaging Vis. **25**(3), 365–386 (2006)
34. Hochbaum, D., Singh, V.: An efficient algorithm for co-segmentation. In: International Conference on Computer Vision (ICCV'09), pp. 269–276 (2009)
35. Joulin, A., Bach, F., Ponce, J.: Discriminative clustering for image co-segmentation. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'10), pp. 1943–1950 (2010)
36. Jung, M., Peyré, G., Cohen, L.D.: Texture segmentation via non-local non-parametric active contours. In: Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR'11), pp. 74–88 (2011)
37. Kim, J., Fisher, J.W., Yezzi, A., Cetin, M., Willsky, A.S.: A nonparametric statistical method for image segmentation using information theory and curve evolution. IEEE Trans. Image Process. **14**(10), 1486–1502 (2005)
38. Lellmann, J., Kappes, J., Yuan, J., Becker, F., Schnörr, C.: Convex multi-class image labeling by simplex-constrained total variation. In: Scale Space and Variational Methods in Computer Vision (SSVM'09), pp. 150–162 (2009)
39. Lellmann, J., Lorenz, D.A., Schönlieb, C., Valkonen, T.: Imaging with Kantorovich–Rubinstein discrepancy. SIAM J. Imaging Sci. **7**(4), 2833–2859 (2014)
40. Lombardi, D., Maitre, E.: Eulerian models and algorithms for unbalanced optimal transport. ESAIM Math. Model. Numer. Anal. **49**(3), 1717–1744 (2015)
41. Lorenz, D., Pock, T.: An inertial forward–backward algorithm for monotone inclusions. J. Math. Imaging Vis. **51**(2), 311–325 (2015)
42. Mendoza, C., Perez-Carrasco, J.-A., Saez, A., Acha, B., Serrano, C.: Linearized multidimensional earth-mover's-distance gradient flows. IEEE Trans. Image Process. **22**(12), 5322–5335 (2013)
43. Mukherjee, L., Singh, V., Dyer, C.R.: Half-integrality based algorithms for cosegmentation of images. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'09), pp. 2028–2035 (2009)
44. Mumford, D., Shah, J.: Optimal approximation by piecewise smooth functions and associated variational problems. Commun. Pure Appl. Math. **42**, 577–685 (1989)
45. Ni, K., Bresson, X., Chan, T., Esedoglu, S.: Local histogram based segmentation using the Wasserstein distance. Int. J. Comput. Vis. **84**(1), 97–111 (2009)
46. Nikolova, M., Esedoglu, S., Chan, T.F.: Algorithms for finding global minimizers of image segmentation and denoising models. SIAM J. Appl. Math. **66**(5), 1632–1648 (2006)
47. Osher, S., Sethian, J.A.: Fronts propagating with curvature-dependent speed: algorithms based on Hamilton–Jacobi formulations. J. Comput. Phys. **79**(1), 12–49 (1988)
48. Papadakis, N., Provenzi, E., Caselles, V.: A variational model for histogram transfer of color images. IEEE Trans. Image Process. **20**(6), 1682–1695 (2011)
49. Papadakis, N., Yıldızoglu, R., Aujol, J.-F., Caselles, V.: High-dimension multi-label problems: convex or non convex relaxation? SIAM J. Imaging Sci. **6**(4), 2603–2639 (2013)
50. Pele, O., Werman, M.: Fast and robust earth mover's distances. In: IEEE International Conference on Computer Vision (ICCV'09), pp. 460–467 (2009)
51. Peyré, G., Fadili, J., Rabin, J.: Wasserstein active contours. In: IEEE International Conference on Image Processing (ICIP'12) (2012)
52. Pitié, F., Kokaram, A.C., Dahyot, R.: Automated colour grading using colour distribution transfer. Comput. Vis. Image Underst. **107**, 123–137 (2007)
53. Pock, T., Chambolle, A., Cremers, D., Bischof, H.: A convex relaxation approach for computing minimal partitions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09), pp. 810–817 (2009)
54. Pock, T., Cremers, D., Bischof, H., Chambolle, A.: Global solutions of variational models with convex regularization. SIAM J. Imaging Sci. **3**(4), 1122–1145 (2010)
55. Punithakumar, K., Yuan, J., Ayed, I.B., Li, S., Boykov, Y.: A convex max-flow approach to distribution-based figure-ground separation. SIAM J. Imaging Sci. **5**(4), 1333–1354 (2012)
56. Rabin, J., Delon, J., Gousseau, Y.: Transportation distances on the circle. J. Math. Imaging Vis. **41**(1–2), 147–167 (2011)
57. Rabin, J., Papadakis, N.: Convex color image segmentation with optimal transport distances. In: Scale Space and Variational Methods in Computer Vision (SSVM'15), pp. 241–252 (2015)
58. Rabin, J., Peyré, G., Delon, J., Bernot, M.: Wasserstein barycenter and its application to texture mixing. In: Scale Space and Variational Methods in Computer Vision (SSVM'11), pp. 435–446 (2012)
59. Ranchin, F., Chambolle, A., Dibos, F.: Total variation minimization and graph cuts for moving objects segmentation. In: Scale Space and Variational Methods in Computer Vision (SSVM'07), pp. 743–753 (2007)
60. Rangarajan, A., Yuille, A., Mjolsness, E.: Convergence properties of the softassign quadratic assignment algorithm. Neural Comput. **11**(6), 1455–1474 (1999)
61. Rother, C., Minka, T., Blake, A., Kolmogorov, V.: Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 1, pp. 993–1000 (2006)
62. Rother, C., Minka, T., Blake, A., Kolmogorov, V.: Cosegmentation of image pairs by histogram matching-incorporating a global constraint into MRFs. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'06), pp. 993–1000 (2006)
63. Rousson, M., Brox, T., Deriche, R.: Active unsupervised texture segmentation on a diffusion based feature space. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'03), pp. 699–704 (2003)
64. Rubio, J., Serrat, J., Lopez, A., Paragios, N.: Unsupervised co-segmentation through region matching. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'12), pp. 749–756 (2012)
65. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. Int. J. Comput. Vis. **40**(2), 99–121 (2000)
66. Schmitzer, B.: A sparse multiscale algorithm for dense optimal transport. J. Math. Imaging Vis. **56**(2), 238–259 (2016)
67. Schmitzer, B., Schnörr, C.: Modelling convex shape priors and matching based on the Gromov–Wasserstein distance. J. Math. Imaging Vis. **46**(1), 143–159 (2013)
68. Schmitzer, B., Schnörr, C.: Globally optimal joint image segmentation and shape matching based on Wasserstein modes. J. Math. Imaging Vis. **52**(3), 436–458 (2015)
69. Sinkhorn, R., Knopp, P.: Concerning nonnegative matrices and doubly stochastic matrices. Pac. J. Math. **21**(2), 343–348 (1967)
70. Solomon, J., de Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., Guibas, L.: Convolutional wasserstein distances: efficient optimal transportation on geometric domains. In: ACM Transactions on Graphics (SIGGRAPH'15) (2015)
71. Swoboda, P., Schnörr, C.: Variational image segmentation and cosegmentation with the wasserstein distance. In: Energy Mini-

mization Methods in Computer Vision and Pattern Recognition (EMMCVPR'13), pp. 321–334 (2013)
72. Vese, L.A., Chan, T.F.: A multiphase level set framework for image segmentation using the Mumford and Shah model. Int. J. Comput. Vis. **50**(3), 271–293 (2002)
73. Vicente, S., Kolmogorov, V., Rother, C.: Joint optimization of segmentation and appearance models. In: IEEE International Conference on Computer Vision (ICCV'09), pp. 755–762 (2009)
74. Vicente, S., Kolmogorov, V., Rother, C.: Cosegmentation revisited: models and optimization. In: European Conference on Computer Visio (ECCV'10), pp. 465–479 (2010)
75. Vicente, S., Rother, C., Kolmogorov, V.: Object cosegmentation. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'11), pp. 2217–2224 (2011)
76. Villani, C.: Topics in Optimal Transportation. AMS, Providence (2003)
77. Vu, B.C.: A splitting algorithm for dual monotone inclusions involving cocoercive operators. Adv. Comput. Math. **38**(3), 667–681 (2013)
78. Yıldızoğlu, R., Aujol, J.-F., Papadakis, N.: Active contours without level sets. In: IEEE International Conference on Image Processing (ICIP'12), pp. 2549–2552 (2012)
79. Yıldızoglu, R., Aujol, J.-F., Papadakis, N.: A convex formulation for global histogram based binary segmentation. In: Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR'13), pp. 335–349 (2013)
80. Yuan, Y., Ukwatta, E., Tai, X., Fenster, A., Schnörr, C.: A fast global optimization-based approach to evolving contours with generic shape prior. UCLA Technical Report CAM, pp. 12–38 (2012)
81. Zach, C., Gallup, D., Frahm, J., Niethammer, M.: Fast global labeling for real-time stereo using multiple plane sweeps. In: International Workshop on Vision, Modeling, and Visualization (VMV'08), pp. 243–252 (2008)
82. Zhu, S., Lee, T., Yuille, A.: Region competition: unifying snakes, region growing, energy/bayes/mdl for multi-band image segmentation. In: IEEE International Conference on Computer Vision (ICCV'95), pp. 416–423 (1995)

Nicolas Papadakis graduated from INSA ROUEN, France, in 2004. He received his Ph.D. in Applied mathematics from Université de Rennes 1 in 2007 and his Habitation thesis from Université de Bordeaux in 2015. Nicolas is currently a CNRS confirmed researcher in image processing at Institut de Mathématiques de Bordeaux.

Julien Rabin graduated from École Normale Supérieure de Cachan, France, in 2006. He received his Ph.D. in image processing and computer vision from Télécom ParisTech in 2009, under the supervision of Julie Delon and Yann Gousseau. Since 2011, Julien is Assistant Professor at GREYC, an image processing laboratory at ENSICAEN and Normandie University.