Variational Image Segmentation and Cosegmentation with the Wasserstein Distance

Paul Swoboda and Christoph Schnörr

Image and Pattern Analysis Group & HCI
Dept. of Mathematics and Computer Science, University of Heidelberg

Abstract. We present novel variational approaches for segmenting and cosegmenting images. Our supervised segmentation approach extends the classical Continuous Cut approach by a global appearance-based data term enforcing closeness of aggregated appearance statistics to a given prior model. This novel data term considers non-spatial, deformation-invariant statistics with the help of the Wasserstein distance in a single global model. The unsupervised cosegmentation model also employs the Wasserstein distance for finding the common object in two images. We introduce tight convex relaxations for both presented models together with efficient algorithmic schemes for computing global minimizers. Numerical experiments demonstrate the effectiveness of our models and the convex relaxations.

Keywords: Wasserstein distance, (co)segmentation, convex relaxation.

1 Introduction

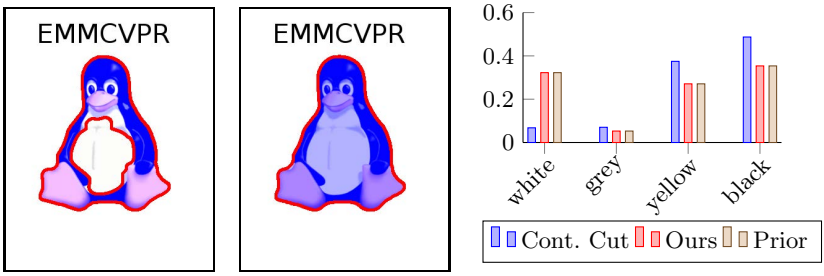
The segmentation problem for k classes consists of finding a partition $(\Omega_1, \dots, \Omega_k)$ of a domain Ω, which means $\Omega_1, \dots, \Omega_k \subset \Omega$, $\Omega_i \cap \Omega_j = \varnothing$ for $i \neq j$ and $\bigcup_{i=1}^k \Omega_i = \Omega$, such that an energy $E(\Omega_1, \dots, \Omega_k)$ is minimized. A commonly used energy functional comes from the minimal partition problem:

$$E(\Omega_1, \dots, \Omega_k) = \frac{1}{2} \sum_{i=1}^{k} \mathrm{Per}(\Omega_i; \Omega) + \sum_{i=1}^{k} \int_{\Omega_i} d^i(x) dx \,, \tag{1}$$

where $\mathrm{Per}(\Omega_i, \Omega)$ is the perimeter of the set Ω_i in Ω and $d^i \in L^1(\Omega), i \in \{1, \dots, k\}$. By minimizing the above functional, k sets are found such that their boundaries are short and the areas they cover are dictated by which potential function d^i has the lowest value. See [5, 12, 16] for treatments of this problem, including relaxations, discretizations and extensions of the minimization problem (1). In the case of two classes this is the well-known Continuous Cut segmentation model, see [8]. This model can be exactly solved by variational methods, see [9].

Often the potential functions $d^i(x) = -\log(p^i(I(x)))$ are chosen as the negative log-likelihood of some probability density p^i modelling the data. Using such potentials d^i poses in general the following problems:

A. Heyden et al. (Eds.): EMMCVPR 2013, LNCS 8081, pp. 321–334, 2013.



(a) Cont. Cut (b) Ours (c) Foreground histograms

Fig. 1. Inadequacy of local costs for segmentation. Figure (a) shows the result of the Continuous Cut segmentation, Figure (b) the result of our approach and Figure (c) the resulting and prior foreground color histograms. The blue areas in Figures (a) and (b) denote the areas determined to be foreground by the respective algorithms. The ground truth foreground is the penguin, while the background is the white area behind it as well as the "EMMCVPR" inscription. We set $d^i(x) = -\log(p^i(I(x)))$ in the Continuous Cut model with accurate distributions p^i for the two classes. White and black color can be found in fore- and background, hence local potentials d^i for both classes are not discriminative or may lead to wrong segmentations. Although the local potentials d^i used in the Continuous Cut model indicate that the "EMMCVPR" inscription should be foreground, it is labelled correctly as background, because the regularization strength is set high. However the white belly of the penguin is labelled wrong, because white is more probable to be background and the regularizer is not able to fill in the correct information. In contrast, our approach correctly determines fore- and background, because it works on the appearance histograms of the *whole* segmentation and enforces them to be close to the prespecified ones as can be seen in Figure (c).

1. For some probability densities p^i the resulting potential functions d^i may not be discriminative or even misleading for some $x \in \Omega$. See Figure 1 for an illustration.
2. For individual components of the resulting partition, the corresponding appearance measures may not match well the model distributions p^i.
3. In unsupervised settings like cosegmentation, which is the task of finding the same object in two different images, we have no knowledge of the probability distribution coming from the object we wish to cosegment. Consequently, no probability models p^i or potential functions d^i are available and must be inferred as part of the optimization problem.

These problems more or less persist, even if we use more elaborate potential functions. We resolve this issue by making our data term *dependent* on the whole segmentation.

We propose to solve the first and second of the stated problems by introducing a global term which directly works on global appearance measures. By using such

a term, we force each of the subsets Ω_i of the partition $(\Omega_1, \ldots, \Omega_k)$ to have an appearance measure which is near a prespecified one. To approach the third problem, we introduce a closely related global term, which depends on both appearance measures of the common object in the two images and ensures that they are similar.

1.1 Related Work

Segmentation. Foreground/background segmentation with the Wasserstein distance was already proposed in the two papers [15] and [7].

Peyré et al. introduce in [15] a data term based on the Wasserstein distance and an approximation thereof for reasons of efficiency. The model proposed there is not convex, so it may get stuck in local minima. By contrast, we derive a fully convex model and work directly with the Wasserstein distance.

The work of Chan et al. in [7] boils down to the Continuous Cut model. The novelty is the computation of the local costs d^1 and d^2 from (1). They are computed by comparing patches around pixels to a foreground and a background histogram with the Wasserstein distance. The model remains convex, as it amounts to solving a Continuous Cut, so global minimizers can be computed very efficiently with existing methods. Our approach differs in that we use the Wasserstein distance (i) on arbitrary images opposed to grayvalue images and (ii) as a truly global data term that *depends* on the segmentation. We point out however that the limitation to grayvalue images in [7] is only made for computational reasons as the one dimensional Wasserstein distance is very fast to compute and is not an inherent limitation of the algorithm in [7].

Cosegmentation. Rother et al. introduce in [18] the cosegmentation task into the literature. To solve the problem, they propose to find a MAP configuration of an MRF with pairwise potentials for spatial coherency and a global constraint to actually cosegment two images. The resulting MRF is not easy to optimize however, and the authors employ a trust region algorithm, which they call trust region graph cut. The algorithm they employ is not guaranteed to find a global optimum, may get stuck in local optima and is dependent upon initialization. In comparison, we solve a convex relaxation that is not dependent upon initialization and gives a reasonably tight global optimum of the relaxed problem.

Vicente et al. give in [19] an overview over several models for cosegmentation. They all have in common that they seek the object to be cosegmented to have similar appearance histograms. The approaches considered in [19] fall into two categories: (i) the histogram matching term may not be very general or (ii) may be difficult to optimize. Approaches falling into category (ii) are solved with EM-type algorithms which alternatingly compute appearance models and then match according to them. Our approach can match appearance measures very flexibly and leads to a *single convex model*, hence solving both of the problems of the approaches encountered in the paper [19].

Another approach to cosegmentation is presented in [20], where object proposals for the objects to be cosegmented are computed and taken as labels in a graphical model. This approach is different from ours, as it relies heavily on object proposals, which are computed with sophisticated but mathematically less explicit methods from the realm of computer vision. For these proposals a big array of complex features is computed. These features are used to compare objects in different images and find the matching ones. Our model does not need object proposals to be computed but finds the cosegmented objects in a mathematically more explicit variational manner by minimizing one single convex energy function. Still, sophisticated features can be introduced in our model as well, however this is not the focus of this paper.

1.2 Contribution

We present

- A new variational model for supervised segmentation with global appearance-based data-terms, see Section 2,
- a new variational model for unsupervised cosegmention of two images based on the similarity of the appearance measures of the respective cosegmentations, see Section 3,
- convex relaxations for both models together with efficient numerical schemes to minimize them, see Section 4,
- experimental validation of the proposed approach, see Section 5.

1.3 Notation

For vectors or vector valued functions $u = \left(u^1, \dots, u^k\right)^\top$ we will denote its i-th entry by u^i. Throughout the paper let $\Omega \subset \mathbb{R}^l$ be the image domain, typically $\Omega = [0,1]^2$. We will denote images by $I, I_1, I_2 : \Omega \to \mathcal{M}$. Images will take values in a measurable space $(\mathcal{M}, \Sigma)$. $\mathcal{M}$ denotes the values an image can take, while $\Sigma \subset 2^{\mathcal{M}}$ is a σ-algebra over $\mathcal{M}$. We also assume we are given a measurable similarity function $c : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$. An example is the k-dimensional euclidean space with the Borel σ-algebra: $(\mathcal{M}, \Sigma) = (\mathbb{R}^l, \mathcal{B}(\mathbb{R}^l))$, $c(v_1, v_2) = \|v_1 - v_2\|^p$. For gray-value images we have $l = 1$ and for color images $l = 3$.

For $v \in \mathcal{M}$ consider the dirac measure $\delta_v(A) = \begin{cases} 0, v \notin A \in \Sigma \\ 1, v \in A \in \Sigma \end{cases}$.

Given a measurable subset $\Theta \subset \Omega$ of the image domain and an image $I : \Omega \to \mathcal{M}$, consider the measure $\mu^I_\Theta : \Sigma \to \mathbb{R}_+$ which records the values which I takes on the subset Θ:

$$\mu^I_\Theta = \int_\Theta \delta_{I(x)} dx \,. \tag{2}$$

Please note that the right hand side of (2) is a measure-valued integral, hence again a measure. It follows that for a measurable set $A \in \Sigma$, we have $\mu^I_\Theta(A) = \int_\Theta \mathbb{1}_{\{I(x) \in A\}} dx$, which is the area in $\Theta \subset \Omega$ where I takes values in $A \subset \mathcal{M}$.

Therefore, the measure μ^I_Θ captures the appearance of the image region $\Theta \subset I$. See Figure 2 for an illustration.

In the discrete case, i.e. $\Omega = \{1, \dots, n\}$, $\mathcal{M} = \{1, \dots, m\}$, the appearance measure μ^I_Θ is the histogram of the image values on the subset Θ: $\mu^I_\Theta(A) = \#\{x \in \Theta : I(x) \in A\}$. The general setup however allows to state the model in a continuous setting and makes notation easier.

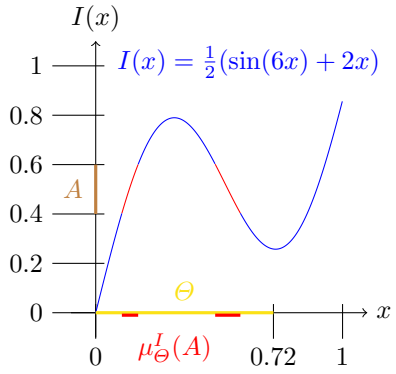


Fig. 2. Illustration of the construction of the appearance measure $\mu^I_\Theta(A)$ for a subregion $\Theta = [0, 0.72] \subset \Omega = [0, 1]$ and a subset of values $A = [0.4, 0.6] \subset \mathcal{M} = [0, 1]$. The blue parts of the curve $I(x) = \frac{1}{2}(\sin(6x) + 2x)$ do not contribute to $\mu^I_\Theta(A)$ while the red ones do. Note that Definition (2) applies also to vector-valued and more generally to $\mathcal{M}$-valued images.

For a convex formulation of our models we introduce the space of functions

$$\mathcal{E}_k = \left\{ u \in BV(\Omega)^k \, : u(x) \in \{e_1, \dots, e_k\} \text{ a.e. } x \in \Omega \right\}, \tag{3}$$

where e_i are the unit vectors in $\mathbb{R}^k$ and $BV(\Omega)$ is the space of functions of bounded variations, see [2] for an introduction to this topic.

$\mathcal{E}_k$ is not a convex set and therefore it is not amenable for use in minimization problems in practice. Hence we consider the convex hull of $\mathcal{E}_k$:

$$\Delta_k = \left\{ u \in BV(\Omega)^k \, : u(x) \in \operatorname{conv}\{e_1, \dots, e_k\} \text{ a.e. } x \in \Omega \right\}, \tag{4}$$

which is the space of functions having values in the k-dimensional unit simplex.

1.4 Wasserstein Distance

Given two measures $\mu_1, \mu_2 : \Sigma \to \mathbb{R}_+$ with $\mu_1(\mathcal{M}) = \mu_2(\mathcal{M})$, the Wasserstein distance $W(\mu_1, \mu_2) \in \mathbb{R}$ of these two measures is computed by evaluating the cost of an optimal rearrangement of μ_1 onto μ_2 with regard to a similarity function c on $\mathcal{M}$. Specifically, consider the space of all *rearrangements* of μ_1 onto μ_2, that is all measures on $\mathcal{M} \times \mathcal{M}$ with marginals μ_1 and μ_2:

$$\Pi(\mu_1, \mu_2) = \{\pi \text{ a measure on } \mathcal{M} \times \mathcal{M} \, : \begin{array}{l} \pi(A \times \mathcal{M}) = \mu_1(A) \\ \pi(\mathcal{M} \times B) = \mu_2(B) \end{array} \quad \forall A, B \in \Sigma\}. \tag{5}$$

Measures in Π are also known as *coupling measures* or *transport plans* in the literature. We will stick to the name *coupling measures*. The Wasserstein distance

is defined as the infimum over all possible rearrangements with regard to the cost c:

$$W(\mu_1, \mu_2) = \inf_{\pi \in \Pi(\mu_1, \mu_2)} \int_{\mathcal{M} \times \mathcal{M}} c \, d\pi \,, \tag{6}$$

It can be shown that under mild assumptions on c the infimum is attained and the distance is finite, see [21] for an in-depth treatise of the Wasserstein distance. The Wasserstein distance is a metric on the space of probability measures for c a metric on $\mathcal{M}$, hence it gives a reasonable distance for measures for c properly chosen.

The minimization problem (6) has linear objective and constraints and is therefore a linear optimization problem, which means it is globally solvable. Moreover it is jointly convex in both of its arguments under mild conditions as well, so it is naturally usable in a convex variational setting, see Theorem 4.8 in [21]

Finally, the Wasserstein distance offers much flexibility in modelling similarity and dissimilarity of measures by choosing an appropriate cost function c in (6).

2 Variational Model for Supervised Segmentation

We will combine into a single variational problem the spatial regularization from the minimal partition problem (1), appearance measures from subsets of the image domain constructed by (2) and the Wasserstein distance (6) for comparing the resulting measures to obtain a new model for segmenting images.

We assume in this setting that one image $I : \Omega \to \mathcal{M}$ and k probability measures μ^i over $\mathcal{M}$ are given. For a partition $(\Omega_1, \ldots, \Omega_k)$ of Ω we enforce the measures $\mu^I_{\Omega_i}$ to be similar to the prespecified measures μ^i by using the Wasserstein distance (6).

Replacing the data term with the potential functions d^i in the minimal partition problem (1) by the Wasserstein distance yields

$$E_{seg}(\Omega_1, \ldots, \Omega_k) = \frac{1}{2} \sum_{i=1}^{k} \mathrm{Per}(\Omega_i, \Omega) + \sum_{i=1}^{k} W\left(\mu^I_{\Omega_i}, |\Omega_i| \cdot \mu^i\right). \tag{7}$$

The additional multiplicative factor $|\Omega_i|$ in the second argument of the Wasserstein distance above is needed to ensure that measures of equal mass are compared, as otherwise the Wasserstein distance is ∞. This is due to the fact that the space (5) of coupling measures Π is empty for measures of differing masses.

Minimizing (7) over all partitions $(\Omega_1, \ldots, \Omega_k)$ of Ω results in partitions, which have regular boundaries due to the perimeter term, and the appearance measures of the partition $\mu^I_{\Omega_i}$ being similar to the given appearance measures μ^i. Note that the measures $\mu^I_{\Omega_i}$ depend on the partition through Ω_i.

As for the minimal partition problem in [5, 9, 12, 16], we replace the sets Ω_i by indicator functions $u^i = \mathbb{1}_{\Omega_i}$ and minimize over them.

Proposition 1. *Let* $u^i = \mathbb{1}_{\Omega_i}$. *Then* (7) *is equal to*

$$J_{seg}(u) = \tfrac{1}{2}\sum_{i=1}^k \int_\Omega |Du^i|\,dx + \sum_{i=1}^k W\Big(\int_\Omega u^i(x)\delta_{I(x)}dx, \int_\Omega u^i(x)dx \cdot \mu^i\Big)\,, \tag{8}$$

where the Total Variation $\int_\Omega |Du^i|\,dx$ *is to be understood as*

$$\int_\Omega |Du^i|\,dx \coloneqq \sup\left\{\int_\Omega u^i \cdot \operatorname{div}(g)dx \;:\; g \in C_c^1(\Omega), \|g\|_\infty \le 1\right\}. \tag{9}$$

Minimizing (7) *over all partitions* $(\Omega_1, \dots, \Omega_k)$ *such that each* Ω_i *has a finite perimeter is equivalent to minimizing* (8) *over* $u \in \mathcal{E}_k$ *given by* (3).

Proof. A partition $(\Omega_1, \dots, \Omega_k)$ corresponds to a vector-valued function $u \in \mathcal{E}_k$ bijectively by $\Omega_i \Leftrightarrow u^i = \mathbb{1}_{\Omega_i}$. By the Coarea formula $\operatorname{Per}(\Omega_i, \Omega) = \int_\Omega |Du^i|\,dx$ holds, see [2]. The Wasserstein term is equal, since

$$\mu^I_{\Omega_i} = \int_{\Omega_i} \delta_{I(x)}dx = \int_\Omega u^i(x)\delta_{I(x)}dx \quad \text{and} \quad |\Omega_i| = \int_\Omega u^i(x)dx\,. \tag{10}$$

Thus, $J_{seg}(u) = E_{seg}(\Omega_1, \dots, \Omega_k)$, which proves the first claim.

The equivalence of both minimization problems stems from the fact, that sets of finite perimeter correspond bijectively to indicator functions of finite variation, see again [2], and partitions correspond bijectively to vector-valued functions such that $\sum_{i=1}^k u^i = \mathbb{1}$ and $u \in \{0,1\}^k$, hence

$$\inf_{u \in \mathcal{E}_k} J_{seg}(u) = \inf_{(\Omega_1,\dots,\Omega_k) \text{ is a partition}} E_{seg}(\Omega_1, \dots, \Omega_k)\,, \tag{11}$$

which proves the second claim.

The functional $J_{seg}(\cdot)$ from (8) is convex, as the Total Variation term is convex and the Wasserstein term is so as well by Theorem 4.8 in [21]. However $\mathcal{E}_k$ is a nonconvex set, so taken together minimizing $\min_{u\in\mathcal{E}_k} J_{seg}(u)$ is not a convex problem. Thus, for practically finding a minimizer of (8), we have to relax the domain over which we optimize. The following problem is convex, as Δ_k is the convex hull of $\mathcal{E}_k$:

$$\inf_{u \in \Delta_k} J_{seg}(u)\,. \tag{12}$$

Remark 1. It is possible to introduce additional local costs $d^i : \Omega \to \mathbb{R}$ without compromising convexity of (8), i.e. to minimize

$$\inf_{u \in \Delta_k} J_{seg}(u) + \sum_{i=1}^k \int_\Omega d^i(x)u^i(x)dx\,. \tag{13}$$

Numerically it comes at a marginal cost to do so. However we chose not to use local costs to demonstrate most directly the power of the global Wasserstein cost.

Remark 2. (8) is the Continuous Cut model when we choose $k = 2$, two points $v_1, v_2 \in \mathcal{M}$ and $\mu^1 = \delta_{v_1}$ and $\mu^2 = \delta_{v_2}$, as then we can replace the Wasserstein distance by multiplication with a local data term. The resulting model is the minimal partition problem (1) for two classes. [9] shows that a global minimizer of the non-relaxed problem can be obtained by thresholding.

3 Variational Model for Unsupervised Cosegmentation

Let two images $I_1, I_2 : \Omega \to \mathcal{M}$ be given and let $\mathcal{M}$ and c be as above. Suppose an object is present in both images, but we have no information about the appearance, location or size of it, Thus, we consider the fully unsupervised setting. The task is to search for two sets $\Omega_1, \Omega_2 \subset \Omega$ such that Ω_1 and Ω_2 are the areas occupied in I_1 resp. I_2 by the common object. Let $\mu^{I_1}_{\Omega_1}$ and $\mu^{I_2}_{\Omega_2}$ be the appearance measures of the common object in images I_1 and I_2 respectively. We know that both appearance measures should be very similar. Therefore we will use the Wasserstein distance $W\left(\mu^{I_1}_{\Omega_1}, \mu^{I_2}_{\Omega_2}\right)$ as a penalization term for enforcing similarity of the appearance measures $\mu^{I_1}_{\Omega_1}$ and $\mu^{I_2}_{\Omega_2}$.

Consider the energy

$$E_{coseg}(\Omega_1, \Omega_2) = \sum_{i=1}^{2} \mathrm{Per}(\Omega_i, \Omega) + W\left(\mu^{I_1}_{\Omega_1}, \mu^{I_2}_{\Omega_2}\right) + \sum_{i=1}^{2} P \cdot |\Omega \backslash \Omega_i| \tag{14}$$

where $P > 0$ and $P \cdot |\Omega \backslash \Omega_i|$ penalizes not selecting an area as the common object. This latter term is called the ballooning term in [19] and is needed to avoid the empty cosegmentation. Minimizing (14) results in two sets Ω_1 and Ω_2 which have a short boundary due to the perimeter term and such that the appearance measures $\mu^{I_1}_{\Omega_1}$ and $\mu^{I_2}_{\Omega_2}$ are similar. Note that neither $\mu^{I_1}_{\Omega_1}$ nor $\mu^{I_2}_{\Omega_2}$ are known but completely depend on the segmentation.

The main difference between the segmentation model (7) and the cosegmentation model (14) is that in the segmentation model the second argument in the Wasserstein distance is fixed while we allow it to vary in the cosegmentation model.

By the same arguments as in Section 2 and Proposition 1, we can establish a similar correspondence between (14) and a suitable convex formulation in the space of indicator functions.

Proposition 2. *Let* $u^i = \mathbb{1}_{\Omega_i}$. *Then* (14) *is equal to*

$$J_{coseg}(u^1, u^2) = \begin{array}{l} \sum_{i=1}^{2} \int_\Omega |Du^i| \, dx + W\left(\int_\Omega u^1(x)\delta_{I_1(x)} dx, \int_\Omega u^2(x)\delta_{I_2(x)} dx\right) \\ + \sum_{i=1}^{2} P \cdot \int_\Omega (1 - u^i(x)) \, dx \end{array}. \tag{15}$$

Minimizing $E_{coseg}(\Omega_1, \Omega_2)$ (14) *over all sets* $\Omega_1, \Omega_2 \subset \Omega$ *with finite perimeter is equivalent to minimizing* $J_{coseg}(u^1, u^2)$ *over all* $\{0, 1\}$*-valued functions of finite variation.*

As in Section 2, J_{coseg} is convex, whereas the space of $\{0,1\}$-valued functions is not. Relaxing to functions $u^i : \Omega \to [0,1]$ yields a convex relaxation.

Note that due to aggregating the appearance in the two measures $\mu^{I_1}_{\Omega_1}$ and $\mu^{I_2}_{\Omega_2}$ in a *translation-, rotation- and deformation-invariant* way, the resulting cosegmentation energy also exhibits these properties.

Remark 3. (14) implicitly defines the size constraint $|\Omega_1| = |\Omega_2|$, since the Wasserstein distance requires both measures to have equal mass. Weakening this constraint is beyond the scope of this paper.

4 Numerical Implementation

It is common to solve convex large-scale non-smooth problems with first order algorithms like [3, 6, 10]. To efficiently solve our models with such schemes, it is necessary to split our energies into suitable convex funtions, such that the proximity operators for each function can be computed efficiently. Our splitting results in $2+k$ convex non-smooth functions for the segmentation functional (8) and 3 such functions with an additional linear term for the cosegmentation functional (15). We use the Generalized Forward-Backward Splitting Algorithm [17], which can handle an arbitrary number of convex functions in a flexible way.

In practice our image domain is discrete. Here we assume $\Omega = \{1, \dots, n\}^2$. The gradient operator will be approximated by forward differences.

We can rewrite the energy function (8) for the segmentation problem as follows by splitting variables for the gradient operator:

$$J_{seg}(u,g) = \chi_{\{\nabla u = g\}} + \chi_{\{u \in \Delta_k\}} + \|g\| + \sum_{i=1}^{k} W^i_{seg}(u^i)\,, \tag{16}$$

where $W^i_{seg}(u) = W\left(\sum_{x\in\Omega} u(x)\delta_{I(x)}, \left(\sum_{x\in\Omega} u(x)\right)\mu^i\right)$ are the Wasserstein terms in (8) and $\chi_{\text{True}} = 0$, $\chi_{\text{False}} = +\infty$ stands for the indicator function. The energy (15) for the cosegmentation problem can be split as follows:

$$J_{coseg}(u,g) = \sum_{i=1}^{2} \left\{\chi_{\{\nabla u^i = g^i\}} + \|g^i\|\right\} + \langle d, u\rangle + \chi_{\{u\in[0,1]^{|\Omega|}\}} + W_{coseg}(u^1, u^2)\,, \tag{17}$$

where $W_{coseg}(u_1,u_2) = W\left(\sum_{x\in\Omega} u^1(x)\delta_{I_1(x)}, \sum_{x\in\Omega} u^2(x)\delta_{I_2(x)}\right)$ is the Wasserstein term in (15) and $\langle d, u\rangle$ takes care of the balloning term.

Solving (8) and (15) with the Generalized Forward-Backward Splitting algorithm from [17] requires solving efficiently the proximity operators for each convex function in (16) and (17). The proximity operator for a function G at point u^0 is defined by

$$\text{prox}_G(u^0) = \text{argmin}_u \frac{1}{2}\|u - u^0\|^2 + G(u)\,. \tag{18}$$

Proximity operators for all the convex functions in (16) and (17) except for the Wasserstein term can be computed very efficiently by standard methods:

- $\text{prox}_{\delta_{\{\nabla u = g\}}}(u^0, g^0)$ is the projection onto the set $\{\nabla u = g\}$ and can be computed with Fourier transforms.
- $\text{prox}_{\Delta_k}(u^0)$ is the projection onto the simplex and can be computed in a small finite number of steps with the algorithm from [14].
- $\text{prox}_{\|g\|}(g^0)$ amounts to computing the shrinkage operator.

See again [17] concerning how these proximity operators are combined.

The Wasserstein proximity operator can be computed efficiently with the technique detailed below.

4.1 Dimensionality Reduction for the Proximity Operator of the Wasserstein Distance

In general, computing the proximity operator of the Wasserstein distance can be expensive and requires solving a quadratic program with $|\Omega| + |\mathcal{M}|^2$ variables. However due to symmetry we can significantly reduce the size of the quadratic program to $|\mathcal{M}|^2$ variables, such that the Wasserstein proximation step is *independent of the size of the image*.

In practice we will solve the problem on an image grid $\Omega = \{1, \dots, n\}^2$ and the number of values a pixel can take is usually significantly smaller than the number of pixels (e.g. 256 values for gray-value images and for color pictures we may cluster the colors to reduce the number of distinct values as well, while the number of pixels $|\Omega| = n^2$ can be huge). Hence, we may assume $|\Omega| \gg |\mathcal{M}|$.

In the following we only discuss the segmentation case due to space constraints.

Due to the representation of the Wasserstein distance (6), the proximity operator $\text{prox}_{W^i_{seg}}(u^0) = \text{argmin}_u \|u - u^0\|^2 + W^i_{seg}(u)$ of the Wasserstein distance in the segmentation problem (16) can be written equivalently as

$$\begin{array}{rl} \text{argmin}_{\{u,\pi\}} & \sum_{x\in\Omega}(u(x) - u^0(x))^2 + \int_{\mathcal{M}\times\mathcal{M}} c(v_1, v_2)\, d\pi(v_1, v_2) \\ \text{s.t.} & \pi(\mathcal{M} \times A) = \sum_{\{x \in I^{-1}(A)\}} u(x) \quad \forall A \subset \mathcal{M} \\ & \pi(B \times \mathcal{M}) = \left(\sum_{x\in\Omega} u(x)\right) \mu^i(B) \quad \forall B \subset \mathcal{M} \\ & \pi \geq 0 \end{array} \tag{19}$$

Note that the Wasserstein distance term above is invariant to permutations of values inside each set $\{I^{-1}(v)\}\ \forall v \in \mathcal{M}$. The quadratic term $\sum_{x\in\Omega}(u(x) - u^0(x))^2\, dx$ also possesses similar symmetries. This enables us to reduce the number of variables as follows:

Let $n_v = \#\{I^{-1}(v)\}$ be the number of pixels which take the value $v \in \mathcal{M}$ and let $\mu^0 = \sum_{x\in\Omega} u^0(x)\delta_{I(x)}$. Consider the problem

$$\begin{array}{rl} \text{argmin}_{\pi\in\mathcal{P}(\mathcal{M}\times\mathcal{M})} & \int_{\mathcal{M}} n_v \cdot \left(\pi(\mathcal{M} \times \{v\}) - \mu^0(\{v\})\right)^2 dv + \int_{\mathcal{M}\times\mathcal{M}} c(v_1, v_2)\, d\pi(v_1, v_2) \\ \text{s.t.} & \pi(B \times \mathcal{M}) = \pi(\mathcal{M} \times \mathcal{M}) \cdot \mu^1(B) \quad \forall B \subset \mathcal{M} \\ & \pi \geq 0 \end{array} \tag{20}$$

The relation between the two minimization problems (19) and (20) is:

Lemma 1. *The minimization problems* (19) *and* (20) *are equivalent in the following sense: For* $I(x) = v \in \mathcal{M}$ *the optimal solutions* $\hat{u}$ *of* (19) *and* $\hat{\pi}$ *of* (20) *correspond to each other via the relation*

$$\hat{u}(x) = u^0(x) + \frac{\hat{\pi}(\mathcal{M} \times \{v\}) - \mu^0(\{v\})}{n_v}. \tag{21}$$

Lemma 1 allows for efficiently solving (19) via (20) and (21).

5 Experiments

To show the performance of our method we have restricted ourselves to only consider colors as features. Hence the features alone are not very distinctive, but the whole energy function makes our approach work. Our label space $\mathcal{M}$ is the CIE 1931 color space and our cost function c will be derived from the euclidean distance on the above color space. *More sophisticated features can be used in our variational models with no additional computational cost in the minimization procedure. Choosing such features however goes beyond the scope of this paper, that is purely devoted to the novel variational approach, rather than to specific application scenarios.* Also, more sophisticated regularizers can be employed as well, e.g. one could vary weights in the total variation term or use nonlocal versions of it, see [11] for the latter.

5.1 Segmentation

In our experimental setting we assume that we have probability measures μ^1, μ^2 at hand for the foreground and background classes, which we employ in the global Wasserstein data-term. We could in addition determine potential functions to enhance segmentation results and solve model (13), e.g. by $d^i(x) = -\log(p^i(I(x)))$, where p^i is the density of μ^i. We chose to not use the latter to show the strength of the global Wasserstein term alone and the tightness of our relaxation. See [5, 9, 12, 16] for numerical examples of segmentation results with potential functions alone.

For the foreground and background appearance measures we chose a part of the foreground and background of the image respectively and constructed prior appearance measures μ^1, μ^2 from them. In a preprocessing step, we clustered the color values of the image by the k-means method [13]. The number of prototypes was set to 50. The quadratic problem in the prox-step (20) of the Wasserstein distance is thus a 50×50 convex quadratic problem and efficiently solvable. We conducted four experiments with textured objects, for which it is not always easy to find discriminative prototypical vectors, but where the color histogram catches much information about the objects' appearance, see figure 3. Note for example that the cheetah's fur has the same color as the sand in the image, but the distribution of the black dots and the color of the rest of the fur is still distinctive. The fish has black regions, exactly as in the background, but the white and black pattern is distinctive again, so a reasonable segmentation can be obtained.

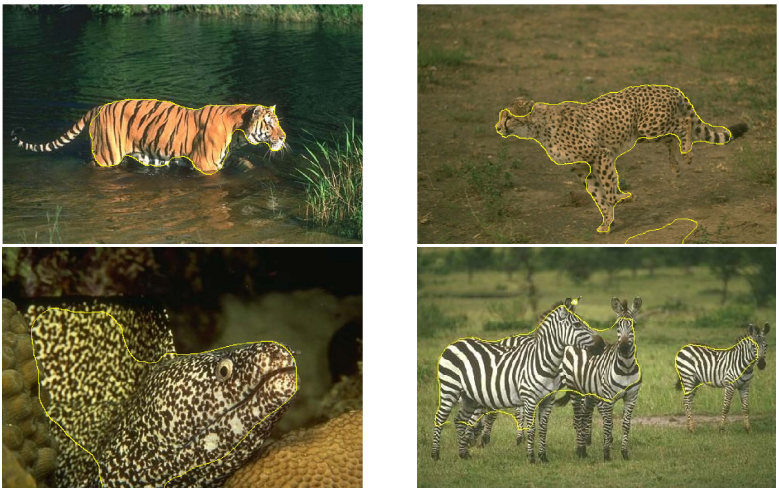


Fig. 3. Supervised segmentation experiments with global segmentation-dependent data term using the Wasserstein distance. Note that because the results correspond to *global* optima of a single convex functional, undesired parts of the partition are solely due to the – in our case: simple color – features and the corresponding prior appearance measures.

5.2 Cosegmentation

For cosegmentation we first subdivide the image into superpixels with SLIC [1]. Then we modify the cost function c as follows: For each superpixel in image 1 we consider k nearest superpixels in image 2 and vice versa. For these pairs we let c be the euclidean distance. For all other pairs of superpixels we set c to ∞. Obviously, the optimal transport plan will be zero where the distance c is ∞, hence we may disregard such variables. By this procedure we reduce the problem size and computational complexity substantially while not reducing the quality of the solution. The prox-step $\text{prox}_{W_{coseg}}(u^1, u^2)$ can be further reduced with a technique similar to the one presented in Section 4.1.

Four experiments can be seen in figure 4. The foreground objects were taken from the dataset [4]. We rotated these objects, translated them and added different backgrounds. As the Wasserstein term does not depend upon location and spatial arrangement of the pixels contributing to the cosegmentation, we could find the common objects independently of where and in which orientation they were located in the images without explicitly enumerating over all different possible such configurations, but by *solving a single convex optimization problem to its global optimum*. Note that in this unsupervised setting, no prior knowledge about the objects is used.

In both experimental settings our method produced functions u^i which were nearly indicator functions except on some parts of the boundaries. Empirically, our relaxation seems to be quite tight.

Fig. 4. Unsupervised cosegmentation: foreground regions in two images are separated at arbitrary locations where the Wasserstein distance between the corresponding histograms is small. This distance depends on the unknown segmentation, and both are consistently determined by a single convex variational problem. No prior knowwledge at all was used in these unsupervised experiments.

6 Conclusion

We presented new variational models for segmentation and cosegmentation. Both utilize the Wasserstein distance as a global term for enforcing closeness between suitable appearance measures. We also derived convex relaxations of the models and presented efficient numerical methods for minimizing them. Both models can be easily augmented by using different regularizers or additional data terms and any features known from the literature.

Acknowledgements. The authors would like to thank Marco Esquinazi for helpful discussions.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: SLIC Superpixels. Technical report, EPFL (June 2010)
2. Ambrosio, L., Fusco, N., Pallara, D.: Functions of Bounded Variation and Free Discontinuity Problems (Oxford Mathematical Monographs). Oxford University Press, USA (2000)
3. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. Found. Trends Mach. Learning 3(1), 1–122 (2010)
4. Wang, J., Gelautz, M., Kohli, P., Rott, P., Rhemann, C., Rother, C.: Alpha matting evaluation website

5. Chambolle, A., Cremers, D., Pock, T.: A Convex Approach to Minimal Partitions. SIAM J. Imag. Sci. 5(4), 1113–1158 (2012)
6. Chambolle, A., Pock, T.: A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. Journal of Mathematical Imaging and Vision 40(1), 120–145 (2011)
7. Chan, T., Esedoglu, S., Ni, K.: Histogram Based Segmentation Using Wasserstein Distances. In: Sgallari, F., Murli, A., Paragios, N. (eds.) SSVM 2007. LNCS, vol. 4485, pp. 697–708. Springer, Heidelberg (2007)
8. Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Trans. Imag. Proc. 10(2), 266–277 (2001)
9. Chan, T.F., Esedoglu, S., Nikolova, M.: Algorithms for Finding Global Minimizers of Image Segmentation and Denoising Models. SIAM J. Appl. Math. 66(5), 1632–1648 (2006)
10. Eckstein, J., Bertsekas, D.P.: On the Douglas—Rachford splitting method and the proximal point algorithm for maximal monotone operators. Mathematical Programming 55, 293–318 (1992)
11. Gilboa, G., Osher, S.: Nonlocal Operators with Applications to Image Processing. Multiscale Modeling & Simulation 7(3), 1005–1028 (2008)
12. Lellmann, J., Schnörr, C.: Continuous Multiclass Labeling Approaches and Algorithms. SIAM J. Imag. Sci. 4(4), 1049–1096 (2011)
13. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/1966, vol. 1, pp. 281–297 (1967)
14. Michelot, C.: A finite algorithm for finding the projection of a point onto the canonical simplex of rn. J. Optim. Theory Appl. 50(1), 195–200 (1986)
15. Peyré, G., Fadili, J., Rabin, J.: Wasserstein Active Contours. Technical report, Preprint Hal-00593424 (2011)
16. Pock, T., Schoenemann, T., Graber, G., Bischof, H., Cremers, D.: A Convex Formulation of Continuous Multi-label Problems. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 792–805. Springer, Heidelberg (2008)
17. Raguet, H., Fadili, J., Peyré, G.: Generalized Forward-Backward Splitting. Technical report, Preprint Hal-00613637 (2011)
18. Rother, C., Minka, T., Blake, A., Kolmogorov, V.: Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In: CVPR, pp. 993–1000. IEEE, Washington, DC (2006)
19. Vicente, S., Kolmogorov, V., Rother, C.: Cosegmentation revisited: Models and optimization. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 465–479. Springer, Heidelberg (2010)
20. Vicente, S., Rother, C., Kolmogorov, V.: Object cosegmentation. In: CVPR, pp. 2217–2224. IEEE (2011)
21. Villani, C.: Optimal Transport: Old and New, 1st edn. Grundlehren der mathematischen Wissenschaften. Springer (November 2008)