Object Segmentation by Shape Matching with Wasserstein Modes

Bernhard Schmitzer and Christoph Schnörr

Image and Pattern Analysis Group, Heidelberg University

Abstract. We gradually develop a novel functional for joint variational object segmentation and shape matching. The formulation, based on the Wasserstein distance, allows modelling of local object appearance, statistical shape variations and geometric invariance in a uniform way. For learning of class typical shape variations we adopt a recently presented approach and extend it to support inference of deformations *during* segmentation of new query images. The resulting way of describing and fitting trained shape variations is in style reminiscent of contour-based variational shape priors, but does not require an intricate conversion between the contour and the region representation of shapes. A well-founded hierarchical branch-and-bound scheme, based on local adaptive convex relaxation, is presented, that provably finds the global minimum of the functional.

Keywords: Wasserstein distance, non-convex optimization, convex relaxation

1 Introduction

Object segmentation and shape matching are fundamental problems in image processing and computer vision that underlie many high-level approaches to understanding the content of an image. They are intimately related: segmentation of the foreground object is a prerequisite for shape matching in a sequential analysis of an image. On the other hand, if performed simultaneously, matching can guide the segmentation process by supplying additional information about the object shape in a noisy environment where unsupervised segmentation would fail. Naturally, joint application is a more complicated problem.

We propose a new functional for simultaneous segmentation and statistical model-based shape matching within a single variational approach. The mathematical framework allows to *combine* key concepts - appearance modelling, modelling and description of deformable regions or contours, geometric invariance - in a uniform way. We rely on convex relaxation and a hierarchical branch-and-bound scheme for global optimization.

1.1 Related Literature

Wasserstein Distance and Image Registration. Optimal transport is a popular tool for object matching and image registration [3, 10] due to its favourable properties: Choosing the cost function to be the squared Euclidean distance between pixels gives access to the rich theory of the Monge formulation of optimal transport. Also, thanks to the linear programming relaxation due to Kantorovich such problems usually involve convex functionals. On the other hand, directly converting grey values to mass densities, as often done [3], is not robust to noise. A naïve extension to noise handling will only work if the query and the reference image are aligned properly (see e.g. [7] for an attempt to alleviate these restrictions). Additionally, when measuring the similarity between two objects via plain optimal transport, their distance will exclusively be determined by the resulting optimal transport cost. *During* the registration process there is no way to benefit from prior knowledge to distinguish common and uncommon types of deformations. However, *after* having computed the registrations, there are promising ideas how to extract and analyze information on the deformations from the optimal registrations [10]. The observed deformation fields are viewed as elements of the tangent space of a reference shape where then standard machine learning techniques (e.g. PCA) can be applied.

Variational Image Segmentation and Contour Based Shape Priors. For object segmentation variational approaches with shape priors, based on contour spaces have received a lot of attention [1, 2]. The manifold of shapes, described by closed contours has been studied extensively [8, 5]. Again, here working in the tangent space of a reference shape enables application of machine learning tools to learn object typical deformations from training data.

However, the map between the contour representation of a shape and the region representation by its indicator function is mathematically complex. Thus, when combined with region based variational segmentation functionals, the contour based shape priors tend to yield highly non-convex functionals that rely on a good initialization to give useful, only locally optimal, results (e.g. [2]). There are approaches to model shape statistics directly on the set of indicator functions, yielding overall convex functionals [4, 6]. But due to required convexity, these shape representations are rather simplistic and lack important features such as geometric invariance.

1.2 Contribution

In this paper we propose a new functional for noise robust joint object segmentation and shape matching based on the Wasserstein distance. We start with a functional for variational segmentation where we regularize the segmentation with the Wasserstein distance to a reference measure. This functional has several limitations, hinted at above and further discussed in Section 3.1. To overcome these, we enhance the functional by additional degrees of freedom, obtaining an advantageous new approach:

(i) The optimal segmentation & matching become invariant under translations and approximately invariant w.r.t. rotations and scale transformations.
(ii) Prior knowledge on object-typical deformations can be learned from training samples and exploited *during* the registration process. Although the

mathematical representation is different, the scope of our approach is similar to that of contour-based shape priors.

(iii) The overall functional is non-convex. Yet, instead of heuristic local optimization we propose a hierarchical branch-and-bound scheme to obtain global optimizers. We show how bounds can be obtained by adaptive convex relaxation that becomes tighter as the hierarchy-scale becomes finer and how successive refinement of the bound computations converges towards the global optimum (Propositions 1 and 2).

Organization. After a brief review of the mathematical background in the next section, we will gradually motivate and develop the full functional in Sec. 3. Global optimization of the non-convex functional is discussed in Sec. 4, key properties of the approach are illustrated with experiments in Sec. 5, before we reach a short conclusion at the end.

2 Mathematical Background: Wasserstein Distance

For any measurable space A denote by $\mathcal{P}(A)$ the set of non-negative measures thereon. For two measurable spaces A, B, a measure μ on A and a measurable map $f : A \to B$, the push-forward $f_\sharp\, \mu$ of μ onto B via f is defined by $f_\sharp\, \mu(\sigma) = \mu(f^{-1}(\sigma))$ for all measurable $\sigma \subset B$.

Let X be a measurable space with measures $\mu, \nu \in \mathcal{P}(X)$ of the same total mass. Then the set of *couplings* $\Pi(\mu, \nu)$ between μ and ν is given by

$$\Pi(\mu,\nu) = \{\pi \in \mathcal{P}(X \times X) :\ \pi(\sigma \times X) = \mu(\sigma),\, \pi(X \times \sigma) = \nu(\sigma) \text{ for all measurable } \sigma \subset X\}\,. \tag{1}$$

For a metric $d_X : X \times X \to \mathbb{R} \cup \{\infty\}$ the Wasserstein distance is determined by

$$D(\mu,\nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \left\{ \int_{X \times X} d_X^2(x,y)\, d\pi(x,y) :\ \pi \in \Pi(\mu,\nu) \right\} \right)^{1/2} . \tag{2}$$

The minimizing coupling π, determining the Wasserstein distance D is called *optimal transport* in the literature [9].

For absolutely continuous measures on $X = \mathbb{R}^n$, metrized with the Wasserstein distance w.r.t. the Euclidean metric, the optimal $\pi \in \Pi(\mu, \nu)$ is induced by a unique map $\varphi : X \to X$, i.e. $\pi = (\mathrm{id}, \varphi)_\sharp\, \mu$. That is, at any point x, the mass of μ is transported to the unique location $\varphi(x)$. Further, these measures constitute a Riemannian manifold. When φ is the optimal transport map between μ and ν then the vector field $t(x) = \varphi(x) - x$ corresponds to a vector in the tangent space of μ. For two vectors t_1, t_2 in the tangent space the inner product is given by $\langle t_1, t_2 \rangle_\mu = \int_X \langle t_1(x), t_2(x) \rangle_{\mathbb{R}^n}\, d\mu(x)$.

We adopt the idea from [10] to use PCA on the tangent space to learn typical object deformations. However, in our approach we will be able to benefit from this learned knowledge during segmentation/matching of new images.

3 Variational Approach

3.1 Problem Setup and Naïve Approach.

We will now, step by step, describe the problem to be solved, set out the notation and develop the final form of our proposed functional.

Let Y be the image domain which we want to separate into fore- and background. We can describe the separation by an indicator function $u : Y \to \{0, 1\}$. To obtain feasible optimization problems, one typically relaxes the constraint that u must be binary to the interval $[0, 1]$. In this paper we use optimal transport as a regularizer. Therefore we interpret the relaxed function u as the density of a measure ν. For simplicity we will define our functionals directly over the set of measures, drop u and translate the $[0, 1]$-constraint appropriately.

To find the optimal segmentation ν of Y we want to combine local information with prior knowledge on the shape of the sought-after object. This information is given by a reference measure μ on a template space X. Let both X and Y be embedded into $\mathbb{R}^2$, i.e. $X, Y \subset \mathbb{R}^2$. Considering the literature, one might be tempted to optimize ν w.r.t. a local data term and regularize by its Wasserstein distance to μ in $\mathbb{R}^2$. The optimal coupling between μ and the optimal ν can then be interpreted as a registration between the template and its counterpart in the image. A corresponding functional could look like this:

$$E_0(\nu) = \frac{1}{2} D(\mu, \nu)^2 + G(\nu) = \frac{1}{2} \inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} \|x - y\|^2 \, d\pi(x, y) + G(\nu) \tag{3}$$

where G is the function that encodes local appearance information. An illustration of the functional is given in Fig. 1a. For the remainder of the paper we choose G to be linear in ν:

$$G(\nu) = \int_Y g(y) \, d\nu(y) \tag{4}$$

Here g is a function of local weights, $g(y) < 0$ ($g(y) > 0$) indicating foreground (background) affinity of point $y \in Y$. For the applicability of the framework presented in this paper, G can be any 1-homogeneous convex function.

An optimal segmentation is then described by an optimizer of E_0 w.r.t. the following feasible set:

$$\mathcal{S}(M) = \{\nu \in \mathcal{P}(Y) : \; \nu(Y) = M, \, \nu(\sigma) \leq \mathcal{L}(\sigma) \text{ for all measurable } \sigma \subset Y\} \, . \tag{5}$$

Here $\mathcal{L}$ denotes the Lebesgue measure on Y. This constraint is equivalent to the density of ν being a relaxed indicator function. $M = \mu(X)$ is the total mass of μ to ensure that the Wasserstein distance $D(\mu, \nu)$ is well defined for all feasible ν.

Limitations. In addition to the mass constraint, as discussed in the introduction, functional (3) has two major shortcomings. The first is the dependence of the optimal ν on the relative embedding of X and Y into the $\mathbb{R}^2$ plane. Assume both μ and ν were fixed. Then any optimal coupling $\pi \in \Pi(\mu, \nu)$ would be still

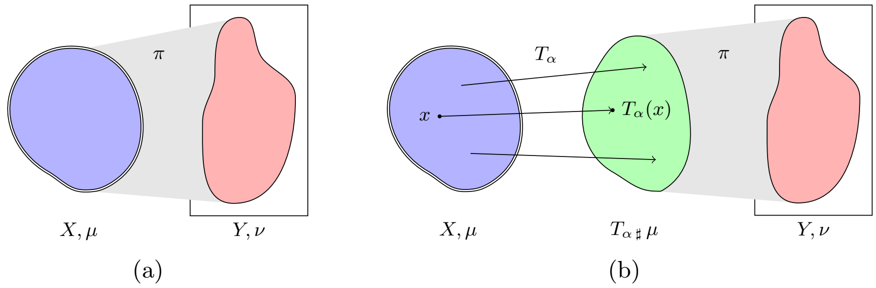


Fig. 1: Illustration of functionals $E_0(\nu)$, eq. (3), and $E_1(\alpha, \nu)$, eq. (8): (a) The segmentation in Y is described by measure ν which is regularized by the Wasserstein distance to a template measure μ, living on X. This simple approach introduces strong bias, depending on the relative location of X and Y, and lacks the ability to explicitly model typical object deformations. (b) In the enhanced functional the template measure μ is deformed by the map T_α, resulting in the push-forward $T_{\alpha\,\sharp}\,\mu$. The segmentation ν is then regularized by its Wasserstein distance to $T_{\alpha\,\sharp}\,\mu$. The corresponding optimal coupling π gives a registration between the foreground part of the image and the deformed template.

be optimal after relative translation of X and Y (of course taking into account the coordinate transformation caused by the translation). However, since we do not consider ν to be fixed, as in other approaches, but optimize over ν, we cannot exploit this quasi-invariance. Any fixed embedding of X and Y will always introduce a bias, encouraging ν to have its mass close to μ, breaking translation invariance, which is clearly not what we want.

The second problem is that any deformation between μ and ν is uniformly penalized by its transportation distance. No information on more or less common deformations can be encoded.

To overcome these restrictions we propose to additionally optimize over the embedding of X into $\mathbb{R}^2$.

3.2 Wasserstein Modes

Let $T_\alpha : X \to \mathbb{R}^2$ be a family of functions, parametrized by some vector $\alpha \in \mathbb{R}^n$, used to adjust the position of X to obtain better matches between template and query image. We choose:

$$T_\alpha(x) = x + \sum_{i=1}^{n} \alpha_i \cdot t_i(x) \tag{6}$$

This linear decomposition will give us enough flexibility to deform X while keeping the resulting functionals easy to handle. We refer to the functions t_i as *modes*. They can be used to make the approach invariant w.r.t. translation, approximately invariant under rotation and scale and introduce prior knowledge on learned object deformations into the functional. The enhanced version of (3) that we consider in this paper is:

$$\begin{aligned} E_1(\alpha,\nu) &= \frac{1}{2} D(T_{\alpha\,\sharp}\,\mu,\nu)^2 + F(\alpha) + G(\nu) \\ &= \frac{1}{2} \inf_{\pi \in \Pi(T_{\alpha\,\sharp}\,\mu,\nu)} \int_{X\times Y} \|x-y\|^2\, d\pi(x,y) + F(\alpha) + G(\nu) \end{aligned} \tag{7}$$

Note that by a standard argument from measure theory we can rewrite this as

$$E_1(\alpha,\nu) = \frac{1}{2} \inf_{\pi \in \Pi(\mu,\nu)} \int_{X\times Y} \|T_\alpha(x)-y\|^2\, d\pi(x,y) + F(\alpha) + G(\nu)\,. \tag{8}$$

F is a function that can introduce statistical knowledge on the distribution of the coefficients α. The functional E_1 is illustrated in Fig. 1b. For simplicity, in the course of this paper we choose

$$F(\alpha) = \frac{1}{2} \alpha^\top \Sigma^{-1}\, \alpha\,, \tag{9}$$

for some symmetric, positive semi-definite Σ^{-1}. We choose a basis in which $\Sigma^{-1} = \mathrm{diag}(\{\Sigma_i^{-1}\}_{i=1}^n)$ is diagonal. So coefficients with $\Sigma_i^{-1} = 0$ can move freely and and coefficients $\Sigma_i^{-1} > 0$ model i.i.d. Gaussian distributions $\alpha_i \sim \mathcal{N}(0,\Sigma_i^2)$.

The functional E_1 has an intuitive and transparent interpretation: With the coefficients α we describe a finite dimensional submanifold of known shapes, F modelling their plausibility. ν is the segmentation-measure, its local plausibility measured by G. $D(T_{\alpha\,\sharp}\,\mu,\nu)$ allows the optimal segmentation to be more flexible than the finite-dimensional submanifold given by the modes would allow, by measuring the distance of ν from the most plausible point on the manifold, and actually carrying out the corresponding assignment.

We will now discuss the choices for the modes t_i to model different types of variation in position and shape of X.

3.3 Geometric Invariance and Statistical Shape Deformation

Euclidean Isometries. If one chooses

$$t_{\mathrm{t1}}(x) = (1,0)^\top, \quad t_{\mathrm{t2}}(x) = (0,1)^\top \tag{10}$$

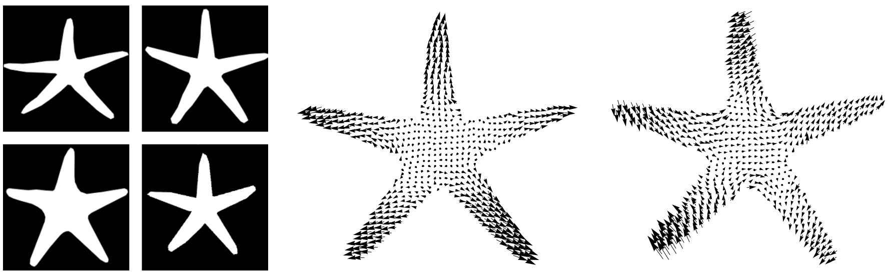
one can use the corresponding coefficients $\alpha_{\mathrm{t1,t2}}$ to translate the template X and thus reintroduce translation invariance, that the simple approach (3) lacks. Further, let $R(\phi)$ be the 2-dimensional rotation matrix by angle ϕ. Then choose

$$t_{\mathrm{r}}(x) = \left.\frac{d}{d\phi} R(\phi)\right|_{\phi=0} x = (-x_2, x_1)^\top \tag{11}$$

to approximately model rotation of X up to angles of about $\pm 30°$. For explicit invariance under these transformations one chooses $\Sigma_{\mathrm{t1}}^{-1} = \Sigma_{\mathrm{t2}}^{-1} = \Sigma_{\mathrm{r}}^{-1} = 0$.

Learning Class Typical Deformations. In this section we describe how modes $t_{\mathrm{d}i}$ can be learned that model class-typical shape variations from a set of training samples. These modes can then be used to allow X to be deformed in the learned way, to prefer known deformations over unknown deformations during the segmentation process.

Let $\{\mu_i\}_{i=1}^m$ be a set of training segmentations, given as indicator-measures: the support of μ_i marks the foreground of the corresponding segmentation. Assume that all μ_i have the same mass. We arbitrarily choose μ_1 to be the reference segmentation and compute the optimal transport couplings $\{\pi_{1,i}\}_{i=1}^k$ between the reference and the other segmentations, *optimized over rotation.* As discussed earlier, for fixed measures (μ_1, μ_i) the optimal coupling does not depend on the relative translation. It is easy to see that the relative translation with smallest cost is the one where the centers of mass coincide [10]. Note that the optimal coupling $\pi_{1,1}$ simply transports mass from all pixels onto themselves. The relative transportation maps that underlie the optimal couplings $\pi_{1,i}$ are then elements of the tangent space of the manifold of measures at μ_1. As in [10], we can then perform a classical principal component analysis (PCA) on the set of tangent vectors to obtain the *mean deformation* t_{m}, a set of *principal deformation modes* $\{t_{\mathrm{d}i}\}_i$ and the corresponding parameters Σ_i^{-1} for the statistical term. Together the pairs $(t_{\mathrm{d}i}, \Sigma_i^{-1})$ act like the well-known contour based shape priors. *However, in our approach no difficult conversion between different mathematical shape representations is necessary.*



(a) (b) (c)

Fig. 2: Learning class-typical deformation modes for starfish: (a) four ground-truth segmentations for learning (b) first principal component: modelling elongation of arms (c) second principal component: modelling angles between arms.

The choice of a reference template is arbitrary and rather heuristic. In our application it is however not possible to take the average of all given samples as reference (as done in [10]) since we work with indicator measures and the mean would no longer be an indicator measure. For a proof of concept however, we consider this choice sufficient and it should be noted that the arbitrary choice of the reference measure is somewhat alleviated by the PCA where the mean of all observed transportation maps is determined. During segmentation the template μ will then be generated by the reference template μ_i, shifted by this mean.

Scale Transformation. The presented framework of deformation modes can, with some slight extensions, also be used to approximately model scale variations of the template.

Let us first have a look at how the map T_α transforms the measure μ via push-forward: μ has uniform mass density 1 within its support since it is an indicator measure (i.e. its density is an indicator function). At $T_\alpha(x)$ the density of the measure $T_{\alpha\,\sharp}\mu$ will depend on the Jacobian determinant $J_{T_\alpha}(x)$. For the decomposition (6) and small coefficients α we find:

$$J_{T_\alpha}(x) = 1 + \sum_{i=1}^{n} \alpha_i \cdot \operatorname{div} t_i(x) + \mathcal{O}(\alpha^2)$$

The rotation mode has zero divergence and since the deformation modes learned from the training samples map indicator measures onto similar indicator measures their divergence should also be small. Since the translation modes are constant, they have no influence on the Jacobian. Hence, the presented method to deform the reference template, is in fact a reasonable approximate description of a set of 'allowed' indicator measures.

Now for rescaling, the corresponding mode is $t_\mathrm{s}(x) = x$. The Jacobian determinant of $x \mapsto x + \alpha_\mathrm{s} \cdot t_\mathrm{s}(x)$ is $(1+\alpha_\mathrm{s})^2$, hence to keep $T_{\alpha\,\sharp}\mu$ an indicator measure, we must multiply it by $(1+\alpha_\mathrm{s})^2$. To make the resulting functional scale invariant, we choose $\Sigma_\mathrm{s}^{-1} = 0$ and to rescale the functional by $(1+\alpha_\mathrm{s})^2$. One then gets:

$$E_{1,\mathrm{s}}(\alpha,\nu) = (1+\alpha_\mathrm{s})^{-2}\left(\frac{1}{2}D((1+\alpha_\mathrm{s})^2 \cdot T_{\alpha\,\sharp}\,\mu,\nu)^2 + F(\alpha) + G(\nu)\right) \tag{12}$$

This functional will have to be optimized subject to the modified condition $\nu \in \mathcal{S}\big(\mu(X)\cdot(1+\alpha_\mathrm{s})^2\big)$, with the mass of ν adjusted according to α_s. Obviously the terms depending on the mass of the measures should be rescaled. But it is reasonable to also rescale the term F since on a larger scale also the deformation modes need to have higher coefficients to obtain the same relative deformation. Since we chose F to be quadratic, the chosen renormalization is exactly the one to cancel that effect.

The scale invariance is only approximate near $\alpha_\mathrm{s} = 0$ because of the rasterization applied for practical implementation. When the grid sizes of X and Y become too different, one will expect numerical artifacts.

4 Optimization

Eliminating ν. The functional E_1 is convex in ν for fixed α and vice versa. But instead of a heuristic alternating optimization scheme we propose a hierarchical branch and bound approach that yields global optimizers and also applies to the scale-invariant version $E_{1,\mathrm{s}}$.

For some fixed α let

$$E_2(\alpha) = \min_{\nu \in \mathcal{S}(\mu(X))} E_1(\alpha, \nu)\,, \qquad E_{2,\mathrm{s}}(\alpha) = \min_{\nu \in \mathcal{S}((1+\alpha_\mathrm{s})^2 \cdot \mu(X))} E_{1,\mathrm{s}}(\alpha, \nu)\,. \tag{13}$$

These can be computed by solving the convex optimization problem in ν for fixed α. The remaining problem is to optimize E_2 w.r.t. α. Let now $\mathcal{A}$ be a set of α-parameters and define a functional over such sets:

$$E_3(\mathcal{A}) = \inf_{\nu \in \mathcal{S}(\mu(X))} \frac{1}{2} \inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} \left(\min_{\alpha \in \mathcal{A}} \|T_\alpha(x) - y\|^2 \right) d\pi(x,y) + \min_{\alpha \in \mathcal{A}} F(\alpha) + G(\nu)\,. \tag{14}$$

This is an adaptive convex relaxation of minimizing E_2 over a given interval. The relaxation becomes tighter as the interval becomes smaller.

For our global optimization scheme the following proposition is required:

Proposition 1. *The functional E_3 has the following properties:*

$$\text{(i)}\; E_3(\mathcal{A}) \leq \min_{\alpha \in \mathcal{A}} E_2(\alpha), \qquad \text{(ii)}\; \lim_{\mathcal{A} \to \{\alpha_0\}} E_3(\mathcal{A}) = E_2(\alpha_0),$$
$$\text{(iii)}\; \mathcal{A}_1 \subset \mathcal{A}_2 \Rightarrow E_3(\mathcal{A}_1) \geq E_3(\mathcal{A}_2)\,. \tag{15}$$

Proof. For the lower bound property (i) note that for any feasible ν and $\pi \in \Pi(\mu,\nu)$:

$$\int_{X \times Y} \left(\min_{\alpha \in \mathcal{A}} \|T_\alpha(x) - y\|^2 \right) d\pi(x,y) \leq \min_{\alpha \in \mathcal{A}} \int_{X \times Y} \|T_\alpha(x) - y\|^2 \, d\pi(x,y)$$

This inequality holds also for the infimum of $\pi \in \Pi(\mu,\nu)$ and $\nu \in \mathcal{S}\big(\mu(X)\big)$. So the first term in E_3 is bounded by $\min_{\alpha \in \mathcal{A}} 1/2\, D(T_{\alpha\,\sharp}\,\mu, \nu)^2$. The separate minimization of the D and F term is obviously smaller than the joint minimization, so the bound property holds.

For the limit property (ii) note that $\|T_\alpha(x) - y\|^2$ and $F(\alpha)$ are continuous functions in α. Hence, when $\mathcal{A} \to \{\alpha_0\}$ all involved minimizations will converge towards the respective function value at α_0 and E_3 converges as desired.

For the hierarchical bound property (iii) just note that for fixed π and ν minimization over the larger set will never yield the larger result for all occurrences of α. This relation will then also hold after minimization. □

With slight modifications E_3 can be extended to the case $E_{3,\mathrm{s}}$ with a scaling-mode involved: the additional rescaling factor will also be independently optimized over and the feasible sets $\mathcal{S}(\mu(X)), \Pi(\mu,\nu)$ in the initial optimization must be replaced by $\mathcal{S}(\mu(X), (1+\alpha_{\mathrm{s,l}})^2, (1+\alpha_{\mathrm{s,u}})^2), \Pi(\mu, \nu, (1+\alpha_{\mathrm{s,l}})^2, (1+\alpha_{\mathrm{s,u}})^2)$

where $\alpha_{\mathrm{s,l}}$ and $\alpha_{\mathrm{s,u}}$ are lower and upper bound for the scale coefficient for all $\alpha \in \mathcal{A}$. The modified feasible sets are defined by

$$\mathcal{S}(M, s_1, s_2) = \bigcup_{s_1 \leq s \leq s_2} \mathcal{S}(s \cdot M) \tag{16}$$

$$\begin{aligned} \Pi(\mu, \nu, s_1, s_2) =& \{\pi \in \mathcal{P}(X \times X) :\ s_1 \cdot \mu(\sigma) \leq \pi(\sigma \times Y) \leq s_2 \cdot \mu(\sigma), \\ & \pi(X \times \sigma) = \nu(\sigma) \text{ for all measurable } \sigma \subset X\}\,. \end{aligned} \tag{17}$$

Since from any π the measure ν can be obtained by marginalization, the nested optimization over ν and π can actually be performed at once, by only minimizing π and transferring the constraints of ν onto π.

So E_3, $E_{3,\mathrm{s}}$ can be computed by solving linear programs, which can also be rewritten as optimal transport problems with suitable dummy nodes, to use specialized, more efficient solvers.

Branch and Bound in α. Let $L = \{(\mathcal{A}_i, b_i)\}_i$ be a list of α-parameter intervals $\mathcal{A}_i$ and lower bounds b_i on E_2 on these respective intervals. For such a list consider the following refinement procedure:

`refine`**(L):**
find the element $(\mathcal{A}_{\hat{i}}, b_{\hat{i}}) \in L$ with the smallest lower bound $b_{\hat{i}}$
let $\texttt{subdiv}(\mathcal{A}) = \{\mathcal{A}_{\hat{i},j}\}_j$ be a subdivision of the interval $\mathcal{A}_{\hat{i}}$ into smaller intervals
compute $b_{\hat{i},j} = E_3(\mathcal{A}_{\hat{i},j})$ for all $\mathcal{A}_{\hat{i},j} \in \texttt{subdiv}(\mathcal{A})$
remove $(\mathcal{A}_{\hat{i}}, b_{\hat{i}})$ from L and add $\{(\mathcal{A}_{\hat{i},j}, b_{\hat{i},j})\}_j$ for $\mathcal{A}_{\hat{i},j} \in \texttt{subdiv}(\mathcal{A})$

This allows the following statement:

Proposition 2. *Let L be a list of finite length. Let the subdivision in* `refine` *be such that any interval will be split into a finite number of smaller intervals, and that any two points will eventually be separated by successive subdivision.* $\texttt{subdiv}(\{\alpha_0\}) = \{\{\alpha_0\}\}$. *Then repeated application of* `refine` *to the list L will generate an adaptive piecewise underestimator of E_2 throughout the union of the intervals $\mathcal{A}$ appearing in L. The sequence of smallest lower bounds will converge to the global minimum of E_2.*

Proof. Obviously the sequence of smallest lower bounds is non-decreasing (see Proposition 1 (iii)) and never greater than the minimum of E_2 throughout the considered region. So it must converge to a value which is at most this minimum. Assume that $\{\mathcal{A}_i\}_i$ is a sequence with $\mathcal{A}_{i+1} \in \texttt{subdiv}(\mathcal{A}_i)$ such that $E_3(\mathcal{A}_i)$ is a subsequence of the smallest lowest bounds of L (there must be such a sequence since L is finite). Since `subdiv` will eventually separate any two distinct points, this sequence must converge to a singleton $\{\alpha_0\}$ and the corresponding subsequence of smallest lowest bounds converges to $E_3(\{\alpha_0\}) = E_2(\alpha_0)$. Since the sequence of smallest lowest bounds converges, and the limit is at most the minimum of E_2, $E_2(\alpha_0)$ must be the minimum. □

In practice we start with a coarse grid of hypercubes covering the space of reasonable α-parameters (translation throughout the image, rotation and scale within bounds where the approximation is valid and the deformation-coefficients in ranges adjusted to the corresponding Gaussian covariances) and the respective E_3-bounds. Any hypercube with the smallest bound will then be subdivided into equally sized smaller hypercubes, leading to an adaptive 2^n-tree cover on the considered parameter range.

We stop the refinement, when the interval with the lowest bound has edge-lengths that correspond to an uncertainty in $T_\alpha(x)$ which is in the range of the pixel discretization of X and Y. Further refinement would only reveal structure determined by rasterization effects.

5 Experiments

Implementation Details. Computation of the E_3-lower bound requires local optimization w.r.t. α for the cost function entries of the optimal transport term. Given the linear decomposition of T_α these are low-dimensional constrained quadratic programs that can quickly be solved. For a given α-interval $\mathcal{A}$ the locally minimized cost function $\min_{\alpha \in \mathcal{A}} \|T_\alpha(x) - y\|^2$ has low values where α values in $\mathcal{A}$ allow $T_\alpha(x)$ to be close to y and rises quickly elsewhere. Exploiting this, we only consider a sparse subset of the full product space $X \times Y$ to speed up computation. To ensure that we still obtain the global optimum, we add overflow variables. As long as no mass is transported onto these dummy variables, the global optimum is attained. Otherwise, more coupling combinations need to be added.

Setup and Numerical Results. For learning of the object class 'starfish' we used about 20 ground truth segmentations. We took the first four principal components as modes, capturing about 70% of the variance in the training set. Together with translation and rotation this amounts to seven modes to be optimized over.

To the test images we applied a simple local color model, based on seeds, to obtain affinity coefficients g, eq. (4). Note that we specifically chose test images with inhomogeneously colored foreground objects and insufficient seeds for color model training, to obtain coefficients on which a purely local segmentation would fail and the benefit of shape-modelling can be demonstrated.

Fig. 3 illustrates the approach for a typical example. Position and pose of the sought-after object are correctly estimated by the modes, *independent of the position of X, i.e. without requiring a good initial guess.* Figure 4 gives original image, affinity coefficients g and the resulting segmentation for the example in Fig. 3 in column 1 and for further examples. The segmentations in Fig. 4 sometimes exhibit small holes or fluctuations along the boundary, even though the underlying object position and pose are very accurately determined (see Fig. 3) and the computed matching is smooth. These irregularities on the pixel level are induced by noise in g and could be removed by local regularization of the boundary of ν (e.g. total variation). As long as such an extension yields a convex functional G, it is still compatible with our approach. To make the acting

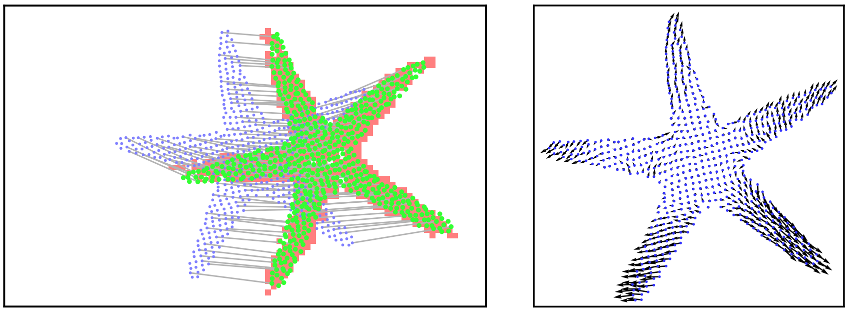


Fig. 3: Left: Illustration of the approach on example data, analogous to Fig. 1b. Small blue dots indicate the arbitrary position of X relative to the image Y (bounding box). Large green dots give the position of $T_\alpha(X)$, the map is indicated by grey lines. The optimal segmentation ν is given by the red shaded region in the image. As intended, the *modes* model the Euclidean isometries (the true object position is not known beforehand and is not relevant for the result), and the major deformations. The Wasserstein distance handles the remaining degrees of freedom, guided by the local data term. Right: The deformation of X by the non-Euclidean modes. Length and relative orientation of the arms are adjusted.

of the presented Wasserstein-regularization as transparent as possible, however, we chose to omit such fine-tuning.

Scale Invariance and Representation Flexibility. In this section we demonstrate two further important properties of our approach: scale invariance and flexibility in application. Due to the general formulation of optimal transport, adaption to superpixels is straightforward, which facilitates application to large images. In a discrete implementation Y need not be a regular grid (pixel-level) in $\mathbb{R}^2$, but can be any set of points.

We illustrate both aspects in Fig. 5. Our approach, equipped with a prior trained on fish, is run on an image with a large and a small fish. We demonstrate scale invariance by actually artificially breaking it: by modelling the scale-coefficient α_{s} to be Gaussian, through the choice of the mean scale $\alpha_{\mathrm{s,m}}$ we can trigger which of the two fish is segmented, while the wrong sized one is ignored. Except for the mean scale, no modifications in the approach were made.

6 Conclusion

We developed a novel variational approach for joint image segmentation and shape matching. The formulation, based on the Wasserstein distance, allows to combine modelling of appearance, statistical shape deformation and geometric invariance in a uniform way, by allowing the reference template to be moved and deformed. We extended previous work on analyzing observed deformation fields for object classification to be applicable already during matching of new query

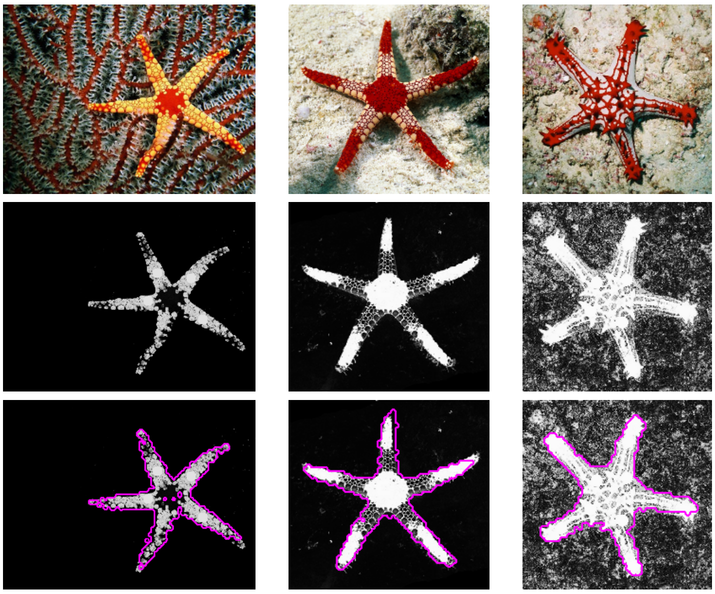


Fig. 4: Segmentation Results with Starfish-Prior. First row: original images. Second row: affinity coefficients g, based on a primitive local color model. There is false-positive clutter, foreground parts are poorly detected or missing. Third row: optimal segmentations, based on joint segmentation and matching. The objects are correctly located, clutter is ignored, missing parts are 'filled in'. Different variants are segmented with the same prior, due to statistical deformation modelling with modes.
On a small scale fluctuations may appear, although the underlying matching is smooth (cp. Fig. 3). These could be handled by enhancing the functional G to locally regularize the boundary of the segmentation.

images. The resulting way of describing and fitting trained shape variations is in style reminiscent of contour-based variational shape priors, but does not require an intricate conversion between the contour and the region representation of shapes. A well-founded hierarchical branch-and-bound scheme, based on local adaptive convex relaxation, is presented, that provably finds the global minimum.

At some points, development of the presented approach is not yet complete (e.g. modelling rotation beyond linear approximation, a more satisfying way to find a reference template, adding local boundary regularization of ν to suppress fluctuations). Yet, as experiments demonstrated, the functional is able to perform robust segmentation and matching in a noisy environment, which, due to geometric invariance, does not depend on a proper initialization. Additionally, the scale invariance property and the natural portability onto superpixel images have been illustrated.

Acknowledgement. This work was supported by the DFG grant GRK 1653.

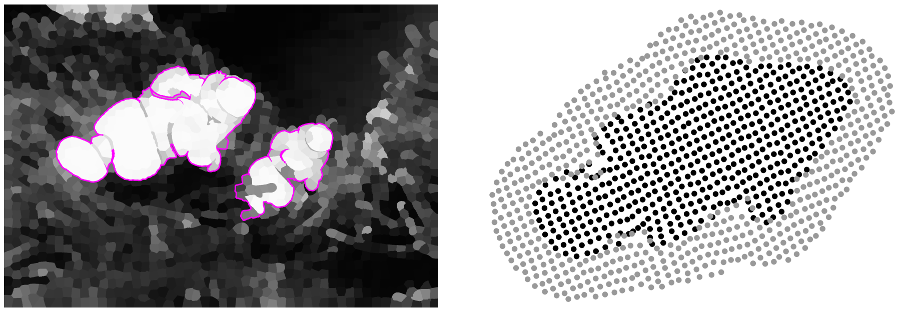


Fig. 5: Scale Invariance and Superpixels. Left: Foreground-affinities g, eq. (4), of a superpixel over-segmentation of an image with two fish of different size, both of which induce strong local minima of our approach. By artificially breaking scale invariance through modelling the scale coefficient α_{s} to be Gaussian, one can choose which one shall be segmented by setting the mean scale. Other than that, the setup was absolutely identical. One obtains similar segmentation results with a free-moving scale coefficient, if one, in turn, erases one of the fish from the g-coefficients. Right: Template X for fish-experiment. To prevent that the small fish is simply immersed in the big one, one must explicitly model a region of background around the fish, by reversing the affinity coefficients for this region of X. Black (grey) dots indicate fore-(back-)ground.

References

1. Charpiat, G., Faugeras, O., Keriven, R.: Shape statistics for image segmentation with prior. In: Computer Vision and Pattern Recognition (CVPR 2007). pp. 1 –6 (2007)
2. Cremers, D., Kohlberger, T., Schnörr, C.: Shape statistics in kernel space for variational image segmentation. Patt. Recognition 36(9), 1929–1943 (2003)
3. Haker, S., Zhu, L., Tannenbaum, A., Angenent, S.: Optimal mass transport for registration and warping. Int. J. Comput. Vision 60, 225–240 (December 2004)
4. Klodt, M., Cremers, D.: A convex framework for image segmentation with moment constraints. In: ICCV (2011)
5. Michor, P., Mumford, D.: Riemannian geometries on spaces of plane curves. Journal of the European Mathematical Society 8(1), 1–48 (2006)
6. Schmitzer, B., Schnörr, C.: Convex coupling continuous cuts and shape priors. In: Scale Space and Variational Methods (SSVM 2011). pp. 423–434 (2012)
7. Schmitzer, B., Schnörr, C.: Modelling convex shape priors and matching based on the Gromov-Wasserstein distance. Journal of Mathematical Imaging and Vision 46(1), 143–159 (2013)
8. Sundaramoorthi, G., Mennucci, A., Soatto, S., Yezzi, A.: A new geometric metric in the space of curves, and applications to tracking deforming objects by prediction and filtering. SIAM Journal on Imaging Sciences 4(1), 109–145 (2011)
9. Villani, C.: Optimal Transport: Old and New. Springer (2009)
10. Wang, W., Slepčev, D., Basu, S., Ozolek, J.A., Rohde, G.K.: A linear optimal transportation framework for quantifying and visualizing variations in sets of images. International Journal of Computer Vision 101, 254–269 (2012)