Sequential Convex Relaxation for Mutual Information-Based Unsupervised Figure-Ground Segmentation*

Youngwook Kee[†,‡] Mohamed Souiai[‡] Daniel Cremers[‡] Junmo Kim[†]

†KAIST, South Korea
‡TU Munich, Germany

Abstract

We propose an optimization algorithm for mutual-information-based unsupervised figure-ground separation. The algorithm jointly estimates the color distributions of the foreground and background, and separates them based on their mutual information with geometric regularity. To this end, we revisit the notion of mutual information and reformulate it in terms of the photometric variable and the indicator function; and propose a sequential convex optimization strategy for solving the nonconvex optimization problem that arises. By minimizing a sequence of convex sub-problems for the mutual-information-based nonconvex energy, we efficiently attain high quality solutions for challenging unsupervised figure-ground segmentation problems. We demonstrate the capacity of our approach in numerous experiments that show convincing fully unsupervised figure-ground separation, in terms of both segmentation quality and robustness to initialization.

1. Introduction

1.1. Unsupervised Figure-Ground Separation

The unsupervised segmentation of figure and ground is inherently a chicken-and-egg problem: *Where is the object and what distinguishes it from the background?* A common assumption in image segmentation is that the object and background have different color distributions. Yet, if these are completely unknown and if they show considerable overlapping, then the joint estimation of color distributions and segmentation becomes a major algorithmic challenge.

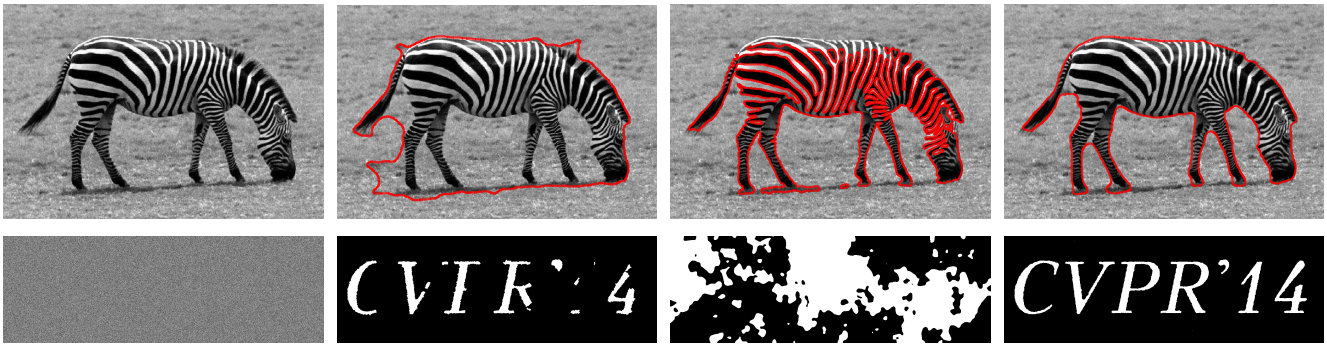
Figure 1 shows two segmentation problems: The top row is a zebra image where humans easily distinguish the zebra from the background; and the bottom row shows a synthetic image where intensity distributions of each region are significantly overlapped—see Figure 6 for details. While existing PDE-based techniques—resorting to the level-set method [21]—often provide somewhat reasonable segmentation results if good initializations are given, these methods tend to fail on the more challenging synthetic image. This is not just because we hardly know which initializations are appropriate for a given task, but the level-set based segmentation techniques are inherently bound to get stuck in local minima. On the other hand, a convex formulation [7] of the Chan-Vese model [9], which is independent of any initialization, guarantees near-optimality of the solutions. Unfortunately, the Chan-Vese model does not take into account any higher-order statistics of color distributions, which means it not only fails on the synthetic image but is also barely able to separate meaningful regions in natural images since, in practice, they exhibit complex coloring.

Therefore, it seems to be fairly obvious that a convex formulation of a functional that takes all possible collections of statistical moments into account would outperform those based on the level-set method or inspired by the Chan-Vese model.

1.2. Related Work

Local Search on Nonparametric Methods Since the color distributions from which pixel values of the object and background are drawn would not be properly assumed as parametric models, the segmentation algorithms that make use of non-parametric statistical methods are superior to those resorting to the parametric methods. In conjunction with the philosophy of unsupervised image segmentation, the nonparametric methods have been incorporated with local search algorithms. Kim *et al*. [17] and Herbulot *et al*. [15] proposed curve evolution equations driven by information-theoretic energy functionals—mutual information and conditional entropy, respectively—using the Parzen-Rosenblatt window method [25, 22]. Similarly, Michailovich *et al*. [18] derived the gradient flow of the Bhattacharyya coefficient for curve evolution using non-

*This work was supported by an Erasmus Mundus BEAM fellowship (No. L031000107).



Input | Kim *et al*. [17] | Brown *et al*. [7] | Our approach

Figure 1. We propose a sequential convex programming approach for mutual-information-based unsupervised image segmentation. In contrast to curve evolution techniques, *e.g*. Kim *et al*. [17], our approach reasonably separates photometric distributions of figure and ground and is robust to initialization, therefore it allows to achieve high quality solutions. Obviously, even a convex formulation of the Chan-Vese model [9], *e.g*. Brown *et al*. [7], fails entirely when applied on above examples due to the complex coloring of the images.

parametric methods. Note that all the curve evolution equations mentioned above were implemented using the level-set method [21], which is a gradient-based local search algorithm.

Convex Relaxation Methods During the era of classical variational and graph-theoretic approaches, the considerable efforts taken to find globally optimal solutions of energy models for various types of computer vision problems have recently opened up the realm of convex relaxation techniques [11]. For example, the seminal work of Nikolova *et al*. [20] tackles the two-phase instance of the Mumford-Shah functional and devised a convex formulation in terms of the indicator function of the object. Based on their work and the calibration approach of Alberti *et al*. [2], Pock *et al*. [23] proposed an efficient algorithm to find near-optimal solutions of the piecewise smooth Mumford-Shah functional. Similarly, Brown *et al*. [7] completely convexified the Chan-Vese model [9] with respect to region-based variables and the geometric unknown.

Convex Relaxation Methods for Statistical Distances The convex relaxation techniques have been gradually expanding to find globally (or near) optimal solutions for energy models inspired by various types of statistical distance measures. A notable relaxation (among others such as those making use of user inputs, *e.g*. scribbles or bounding boxes) is the work of Punithakumar *et al*. [24]. The authors proposed a sequential bound optimization technique for the Bhattacharyya coefficient between a given (*a priori* known) distribution and distributions of the candidate figures to be segmented. However, to the best of our knowledge, little attention has been paid to convex relaxation methods for unsupervised image segmentation, except for the work of Ni *et al*. [19]. In their work, they proposed an energy functional which consists of the Wasserstein distance between (local) intensity histograms from disjoint regions, and the total variation regularization. The energy is only convex with respect to the geometric unknown, and is minimized by an alternating scheme.

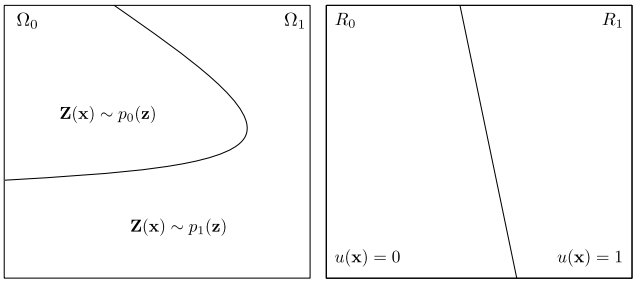
1.3. Contribution

In this paper, we consider the fully unsupervised figure-ground segmentation problem, namely the chicken-and-egg problem of jointly computing a segmentation and respective color distributions. This problem combines maximal color separation with spatial regularity and can be formulated as the following functional:

$$\min_{F \subset \Omega} \mathrm{Per}(F;\Omega) - \lambda \mathcal{D}(P_F, P_G), \tag{1}$$

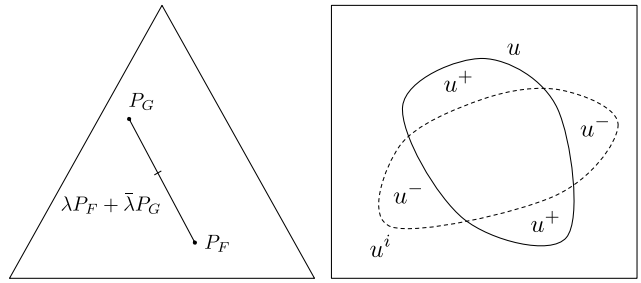
where Ω and F denote an image domain and the foreground region, respectively; and the perimeter of F is denoted as $\mathrm{Per}(F;\Omega)$. P_F and P_G denote the color distributions of the foreground and background, respectively; and $\mathcal{D}(\cdot,\cdot)$ is some metric on the space of probability distributions (weighted by a parameter $\lambda > 0$). Our main contribution is to establish a family of convex upper bounds of the mutual information—which can be rewritten as a convex combination of Kullback-Leibler divergences—in such a way that it can be efficiently minimized by sequential convex relaxation. A generalization for vector-valued images turns out to be straightforward; on the other hand, there is no closed form representation for vector-valued images in the case of the Wasserstein distance [19].

The remainder of the manuscript is organized as follows: In Section 2, we introduce, step by step, a functional which combines the negative mutual information of color distributions with a weighted total variation regularization encoding the boundary length. In Section 3, we show how this formulation can be minimized by constructing a sequence of convex upper bounds, each of which can in turn be efficiently



(a) Ground truth model (b) Binary labels

Figure 2. **Probabilistic setting for figure-ground separation.** (**a**) Photometric distributions of Ω_0 and Ω_1, where each region is associated with $p_0(\mathbf{z})$ and $p_1(\mathbf{z})$, respectively. (**b**) An indicator function $u(\mathbf{x})$ decides foreground R_1 and background R_0.



(a) Mutual information (b) Labeling configuration

Figure 3. **(a)** On the space of probability distributions, the mutual information measures similarity between two distributions P_F and P_G (by means of a convex combination of Kullback-Leibler divergences between P_F and $\lambda P_F + \bar{\lambda} P_G$, and between P_G and $\lambda P_F + \bar{\lambda} P_G$). **(b)** We use the same labeling configuration as proposed in [24].

minimized by convex relaxation techniques. In Section 4, we experimentally validate the capacity of our approach by showing convincing, fully unsupervised figure-ground separation; then we end with a conclusion.

1.4. Notation

Throughout the paper, we mainly deal with information-theoretic quantities from a variational point of view. We clarify a system of notation to be used as follows: capital and lower case letters are used for denoting random and deterministic variables, respectively; and we use bold face letters for vectors.

2. Information-Theoretic Energy Functional

We briefly summarize the motivation of an information-theoretic energy for unsupervised image segmentation proposed by Kim *et al*. [17], then reformulate the energy in terms of an *indicator function* and its *total variation*, thereby convexifying the curve length penalization term.

2.1. Image Modelling

Let $I : \Omega \subset \mathbb{R}^2 \to \mathcal{Z} \subset \mathbb{R}^n$ be an observed image function that maps a bounded open set Ω with Lipschitz boundary to a photometric space $\mathcal{Z}$, *e.g*. intensity or color space. As in Figure 2 (a), we model collections of photometric variables as independent and identically distributed spatial random processes as follows:

$$\begin{aligned} \{\mathbf{Z}(\mathbf{x}) \in \mathcal{Z} \mid \mathbf{x} \in \Omega_1\} &\overset{\text{iid}}{\sim} p_1(\mathbf{z}), \\ \{\mathbf{Z}(\mathbf{x}) \in \mathcal{Z} \mid \mathbf{x} \in \Omega_0\} &\overset{\text{iid}}{\sim} p_0(\mathbf{z}). \end{aligned} \tag{2}$$

where Ω_1 and Ω_0 are the ground truth object and background, respectively. Note that $I(\cdot)$ is a realization of $\mathbf{Z}(\cdot)$; and both $p_1(\mathbf{z})$ and $p_0(\mathbf{z})$ are joint probability density functions when $n \geq 2$. For a given indicator function $u \in BV(\Omega; \{0, 1\})$, the *space of functions of bounded variation* [3]—we use $BV(\Omega)$ as a solution space in this paper—the figure R_1 and the ground R_0 which form a partition of Ω is defined by

$$\begin{aligned} R_1 &= \{\mathbf{x} \in \Omega \mid u(\mathbf{x}) = 1\}, \\ R_0 &= \{\mathbf{x} \in \Omega \mid u(\mathbf{x}) = 0\}, \end{aligned} \tag{3}$$

which are shown in Figure 2 (b).

Having established the above statistical assumption and correspondences, we introduce the notion of mutual information between the photometric variable and the indicator function.

2.2. Mutual Information between the Photometric Variable and the Indicator Function

Let us randomly extract a pixel $\mathbf{X}$ from Ω for a given image; its photometric variable $\mathbf{Z}(\mathbf{X})$ is then probabilistically realized by the following distribution

$$\begin{aligned} p_{\mathbf{Z}(\mathbf{X})}(\mathbf{z}) &= \sum_{i=\{0,1\}} \Pr(\mathbf{X} \in \Omega_i)\, p_{\mathbf{Z}(\mathbf{X})|\mathbf{X}\in\Omega_i}(\mathbf{z}), & (4) \\ &= |\Omega_0|/|\Omega|\, p_0(\mathbf{z}) + |\Omega_1|/|\Omega|\, p_1(\mathbf{z}), & (5) \end{aligned}$$

where $|\cdot|$ denotes the 2-dimensional Lebesgue measure, *i.e*., area. The joint distribution of the photometric variable consists of two sources of uncertainty: first, the uncertainty of pixel location, or the region where a given pixel belongs, which is modeled by $\Pr(\mathbf{X} \in \Omega_i)$; the second source of uncertainty is photometric distributions in each region, *i.e*., $p_{\mathbf{Z}(\mathbf{X})|\mathbf{X}\in\Omega_i}(\mathbf{z})$. The indicator function $u(\mathbf{X})$, on the other hand, is given by

$$u(\mathbf{X}) = \begin{cases} 1 & \text{with probability } |R_1|/|\Omega|, \\ 0 & \text{with probability } |R_0|/|\Omega|. \end{cases} \tag{6}$$

It should be noted that the function $u(\cdot)$ is deterministic itself, but the probability nature of $u(\cdot)$ in (6) derives merely from the uncertainty of pixel location.

We now consider the conditional entropy (or uncertainty) $h(\mathbf{Z}(\mathbf{X})|u(\mathbf{X}))$. Since the differential entropy is concave, *i.e.*, $h(\alpha p_0 + (1-\alpha)p_1) \geq \alpha h(p_0) + (1-\alpha)h(p_1)$, $\alpha \in (0,1)$, for every p_0, $p_1 \in \mathcal{P}$ (the space of probability density functions which is a convex set), the idea is to minimize the conditional entropy by flipping pixel labels to make their photometric distributions (*i.e.*, $p_{\mathbf{Z}(\mathbf{X})|u(\mathbf{X})=0}(\mathbf{z})$ or $p_{\mathbf{Z}(\mathbf{X})|u(\mathbf{X})=1}(\mathbf{z})$, which are mixtures of p_0 and p_1) as homogeneous as possible. Thereby, we achieve an optimal labeling configuration where the photometric distributions are similar to the model distributions, namely p_0 and p_1. The unsupervised figure-ground segmentation problem, therefore, can be captured by maximizing the mutual information between the photometric variable and the indicator function, where the mutual information is defined as

$$\begin{aligned}
&\mathcal{I}(\mathbf{Z}(\mathbf{X}); u(\mathbf{X})) \\
&= h(\mathbf{Z}(\mathbf{X})) - h(\mathbf{Z}(\mathbf{X})|u(\mathbf{X})) \qquad (7)\\
&= h(\mathbf{Z}(\mathbf{X})) - \Pr(u(\mathbf{X})=1)h(\mathbf{Z}(\mathbf{X})|u(\mathbf{X})=1) \\
&\quad - \Pr(u(\mathbf{X})=0)h(\mathbf{Z}(\mathbf{X})|u(\mathbf{X})=0), \qquad (8)
\end{aligned}$$

where $\Pr(u(\mathbf{X})=1) = \Pr(\mathbf{X} \in R_1) = |R_1|/|\Omega|$ and $\Pr(u(\mathbf{X})=0) = \Pr(\mathbf{X} \in R_0) = |R_0|/|\Omega|$. Indeed, the mutual information is maximized if, and only if, $\Omega_0 = R_0$, $\Omega_1 = R_1$ (or equivalently $\Omega_0 = R_1$, $\Omega_1 = R_0$)—readers can find the proof in [17]—which comes basically from the data processing inequality [10]. Interestingly, the mutual information proposed can be interpreted as a disparity measure between $p_{\mathbf{Z}(\mathbf{X})|u(\mathbf{X})=1}(\mathbf{z})$ and $p_{\mathbf{Z}(\mathbf{X})|u(\mathbf{X})=0}(\mathbf{z})$ using a convex combination of Kullback-Leibler (KL) divergences—often called λ-divergence—as follows

$$\begin{aligned}
&\mathcal{I}(\mathbf{Z}(\mathbf{X}); u(\mathbf{X})) \qquad (9)\\
&= \lambda \mathcal{D}_{\mathrm{KL}}(P_F || \lambda P_F + \bar{\lambda} P_G) + \bar{\lambda} \mathcal{D}_{\mathrm{KL}}(P_G || \lambda P_F + \bar{\lambda} P_G)
\end{aligned}$$

where $\lambda = |\Omega_1|/|\Omega|$, and P_F and P_G are the photometric distributions from which densities $p_{\mathbf{Z}(\mathbf{X})|u(\mathbf{X})=1}(\mathbf{z})$ and $p_{\mathbf{Z}(\mathbf{X})|u(\mathbf{X})=0}(\mathbf{z})$ are derived. Although the KL divergence is not a distance metric on the space of probability distributions, it as a premetric measures discrepancy between P_F and P_G via $\lambda P_F + \bar{\lambda} P_G$. In Figure 3 (a), we illustrate how the mutual information measures dissimilarity between P_F and P_G on the space of probability distributions.

2.3. Energy Functional

Nevertheless, the mutual information $\mathcal{I}(\mathbf{Z}(\mathbf{X}); u(\mathbf{X}))$ cannot be computed in practice since the distributions involved are unknown to us; otherwise, the maximization of the mutual information would be straightforward by the likelihood ratio test, *i.e.*, $\frac{p_1(\mathbf{z}(\mathbf{x}))}{p_0(\mathbf{z}(\mathbf{x}))} \gtrless \gamma$. Therefore, we estimate the mutual information using the *integral estimate*[1] of entropy proposed by Dmitriev *et al.* [12]; they showed that the estimate converges *almost surely*, that is $\Pr\left(\lim_{n\to\infty} \hat{h}(\mathbf{Z}(\mathbf{X})) = h(\mathbf{Z}(\mathbf{X}))\right) = 1$, where the entropy estimate $\hat{h}(\mathbf{Z}(\mathbf{X})) = -\int_{\mathcal{Z}} \hat{p}_{\mathbf{Z}(\mathbf{X})}(\mathbf{z}) \log \hat{p}_{\mathbf{Z}(\mathbf{X})}(\mathbf{z})\, \mathrm{d}\mathbf{z}$; the kernel density estimate $\hat{p}_{\mathbf{Z}(\mathbf{X})}(\mathbf{z})$ of $p_{\mathbf{Z}(\mathbf{X})}(\mathbf{z})$ is defined as $\frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{z} - I(\mathbf{x}_i))$, where $\mathbf{H}$ is the $n \times n$ bandwidth matrix. Note that we use $\frac{1}{|\Omega|}\int_{\Omega} K_{\mathbf{H}}(\mathbf{z} - I(\mathbf{x}))\, \mathrm{d}\mathbf{x}$ for $\hat{p}_{\mathbf{Z}(\mathbf{X})}(\mathbf{z})$ in the continuous setting. The conditional entropy estimates $\hat{h}(\mathbf{Z}(\mathbf{X})|u(\mathbf{X}) = 1)$ and $\hat{h}(\mathbf{Z}(\mathbf{X})|u(\mathbf{X}) = 0)$ are similarly defined by using $\hat{p}_{\mathbf{Z}(\mathbf{X})|u(\mathbf{X})=1}(\mathbf{z}) = \frac{\int_{\Omega} K_{\mathbf{H}}(\mathbf{z}-I(\mathbf{x}))u(\mathbf{x})\, \mathrm{d}\mathbf{x}}{\int_{\Omega} u(\mathbf{x})\, \mathrm{d}\mathbf{x}}$, and $\hat{p}_{\mathbf{Z}(\mathbf{X})|u(\mathbf{X})=0}(\mathbf{z}) = \frac{\int_{\Omega} K_{\mathbf{H}}(\mathbf{z}-I(\mathbf{x}))(1-u(\mathbf{x}))\, \mathrm{d}\mathbf{x}}{\int_{\Omega} 1-u(\mathbf{x})\, \mathrm{d}\mathbf{x}}$. Here, a number of possible definitions for the kernel function $K_{\mathbf{H}}(\mathbf{z})$ and $\mathbf{H}$ can be found in [26].

We now combine the mutual information estimate with a (weighted) total variation (TV) regularization encoding boundary length. Furthermore, we multiply the area of the image domain to the mutual information estimate in order to incorporate the total amount of information between the indicator function and the entire image—see [17] for details. Then, the overall energy model is given by:

$$\inf_{u \in BV(\Omega;\{0,1\})} \underbrace{-\lambda|\Omega|\hat{\mathcal{I}}(\mathbf{Z}(\mathbf{X}); u(\mathbf{X})) + \int_{\Omega} g|Du|}_{=:E(u)}, \qquad (10)$$

where $\lambda > 0$ is a parameter. The second term is the weighted TV-norm of u [3]—geometrically equivalent to $\mathrm{Per}_g(F;\Omega)$—which is the exact counterpart of that in the geodesic active contours [6], where $g(\mathbf{x})$ is an edge indicator function. Let us claim that there exists a minimizer of (10).

Proposition 1. *There exists a* $u(\cdot)$ *such that* $\inf\{E(u) \,|\, u \in BV(\Omega;\{0,1\})\}$ *is achieved.*

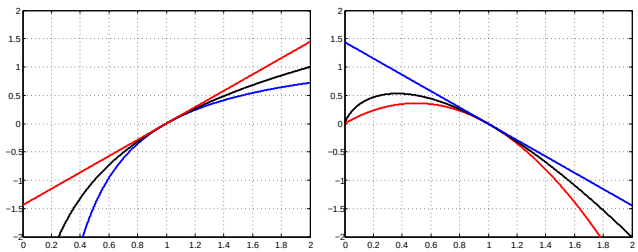
Proof. See the Supplementary Material. □

However, in spite of the existence of a solution, the overall energy functional would not be able to be efficiently minimized because the convexity of $E(u)$ is not guaranteed even when the domain $\{0,1\}$ is relaxed to its convex hull $[0,1]$. Note that the proof of Proposition 1 obviously holds in the relaxed domain $[0,1]$.

Proposition 2. $-\lambda|\Omega|\hat{\mathcal{I}}(\mathbf{Z}(\mathbf{X}); u(\mathbf{X}))$ *is concave in* $u \in BV(\Omega;[0,1])$.

Proof. See the Supplementary Material. □

[1] Note that entropy terms in [17], on the other hand, are estimated by using the *resubstitution estimate* proposed by Ahmed *et al.* [1]—which is essentially the law of large numbers. They showed that the estimate is mean square consistent. Readers can refer to an overview of the nonparametric entropy estimation in [5].



(a) (b)

Figure 4. **Fundamental inequality** when $b = 2$. **(a)** Red, black, and blue curves correspond to $(x-1)\log_b e$, $\log_b x$, and $\left(1-\frac{1}{x}\right)\log_b e$, respectively. **(b)** Blue, black, and red curves correspond to $-x\left(1-\frac{1}{x}\right)\log_b e$, $-x\log_b x$, and $-x(x-1)\log_b e$, respectively.

Let $u^\star_{\text{bin}}$ be a thresholded version of $u^\star_{\text{rel}}$ which is a global minimizer of $E(u)$ over the relaxed domain $[0,1]$, that is $u^\star_{\text{bin}} = \mathbf{1}_{\{u^\star_{\text{rel}} \geq \theta\}}$ for any threshold value $\theta \in (0,1)$; and let u_{opt} be a globally optimal binary solution of $E(u)$. Then, under the condition that the coarea formula [14] holds in $E(u)$, the global optimality of $u^\star_{\text{bin}}$, or $E(u^\star_{\text{bin}}) = E(u_{\text{opt}})$, is typically guaranteed by the thresholding theorem [20]. Unfortunately, one cannot obtain an optimal binary solution by simply thresholding any relaxed solutions, because the coarea formula does not hold in the case of $E(u)$. Instead, we can give a per instance energy bound for the optimal solution: $|E(u^\star_{\text{bin}}) - E(u_{\text{opt}})| \leq |E(u^\star_{\text{bin}}) - E(u^\star_{\text{rel}})|$. That is, by evaluating $E(u)$ at $u^\star_{\text{rel}}$ and $u^\star_{\text{bin}}$, one can check how close $u^\star_{\text{bin}}$ is to u_{opt} *energetically*. However, efficient minimization of $E(u)$ over $BV(\Omega;[0,1])$ appears hopeless, as it is nonconvex. Yet, $E(u)$ is a form of the difference of convex functionals where the Majorization-Minimization principle [16] could kick in.

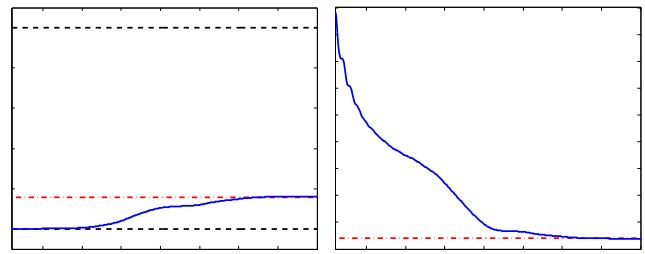
Indeed, in what follows, we tackle the challenging nonconvex optimization problem by constructing a *convex surrogate*, thereby minimizing the original nonconvex energy by means of sequential convex relaxation.

3. Sequential Convex Programming

To begin with, we summarize the Majorization-Minimization (MM) principle [16] as follows.

MM Philosophy Let $F(u|u^i)$ denote a function of u whose form depends on u^i. In the MM principle, one constructs a surrogate function $F(u|u^i)$ which majorizes the objective $E(u)$, meaning for all u, $F(u|u^i) \geq E(u)$ and $F(u^i|u^i) = E(u^i)$, and minimizes it instead. Indeed, if u^{i+1} is a minimizer of $F(u|u^i)$ for all $i \in \mathbb{Z}_+$, the iterative minimization of $F(u|u^i)$ gives a sequence of energies $(E(u^i))_{i\geq 1}$ which is monotonically decreasing. Furthermore, $(E(u^i))_{i\geq 1}$ is convergent if $E(u)$ is bounded below.

Therefore, the task is to construct an energy upper bound $F(u|u^i)$ for $E(u)$ in (10). Note that the iterative bound optimization framework proposed by Punithakumar *et al.* [24] is basically the MM principle.



(a) (b)

Figure 5. **(a)** Mutual information estimate $\hat{\mathcal{I}}(\mathbf{Z}(\mathbf{X}); u^i_{\text{bin}}(\mathbf{X}))$ (**blue line**) is bounded by 0 and 1 (**black lines**) and converges to the ground truth (**red line**). **(b)** Energy $E(u^i_{\text{bin}})$ (**blue line**) is monotonically decreasing to the ground truth (**red line**).

3.1. Energy Upper Bounds and Convex Relaxation

Definition 1. *Let $u^i \in BV(\Omega;\{0,1\})$ be a previous label and $u \in BV(\Omega;\{0,1\})$ be a current label. We denote the area increase and decrease as respectively,*

$$u^+(\mathbf{x}) = \begin{cases} u(\mathbf{x}) & \textit{where } u^i(\mathbf{x}) = 0, \\ 0 & \textit{otherwise;} \end{cases} \tag{11}$$

$$u^-(\mathbf{x}) = \begin{cases} 1 - u(\mathbf{x}) & \textit{where } u^i(\mathbf{x}) = 1, \\ 0 & \textit{otherwise.} \end{cases} \tag{12}$$

The above labeling configuration proposed in [24]—see also Figure 3 (b)—is one of the building blocks to construct an energy upper bound of $E(u)$ with the following inequality.

Lemma 1 (Fundamental inequality)**.** *For any $b > 0$, and $x > 0$,*

$$\left(1 - \frac{1}{x}\right)\log_b e \leq \log_b x \leq (x-1)\log_b e, \tag{13}$$

where equalities on both sides are satisfied if, and only if $x = 1$.

Proof. See the proof in [27]. □

We illustrate the inequality (13) in Figure 4 to give an intuition for the bounds which are going to be presented. Indeed, it turns out these two building blocks play a significant role to derive the following theoretical results.

Proposition 3. *Given a binary label u^i, for any labeling function $u \in BV(\Omega;\{0,1\})$, we have the following energy upper bound:*

$$E(u) \leq \underbrace{\lambda\left(-|\Omega|\hat{\mathcal{I}}(\mathbf{Z}(\mathbf{X}); u^i(\mathbf{X})) + J(u,u^i)\right) + \int_\Omega g|Du|}_{=:F(u|u^i)}, \tag{14}$$

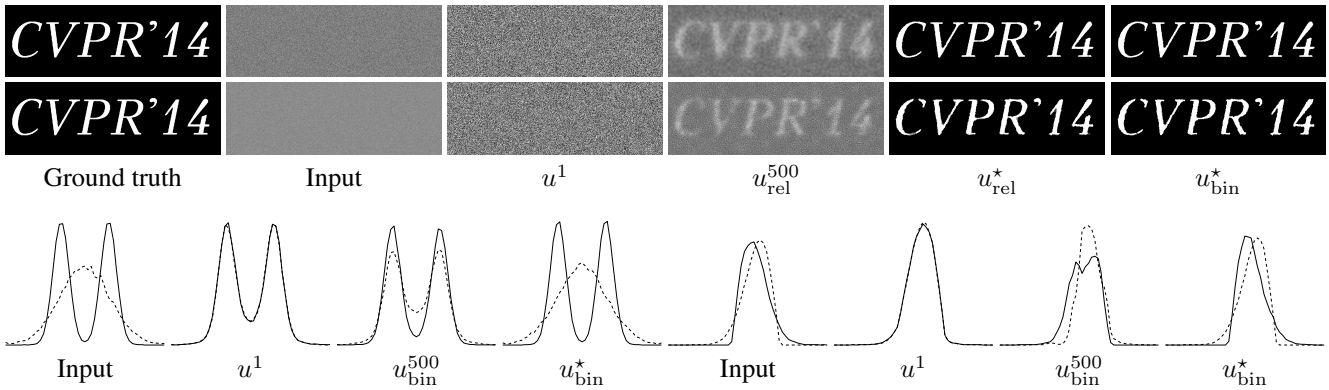


Figure 6. **Synthetic image experiments for visualization of the intensity distribution separation. Case 1 (top row and four PDFs from the left most) :** Object (**white**) and the background (**black**) of the "CVPR'14" image are associated with an unimodal and bimodal distribution, respectively, of the same $\mu = 0.5$ and $\sigma = 0.16$. **Case 2 (second row and the remaining PDFs) :** By setting $Z_G = 2\mathbb{E}[Z_F] - Z_F$, where $Z_F \sim \text{Rayleigh}(0.15\sqrt{2/\pi}) + 0.4$ [18], we obtain two Rayleigh distributions from which all the even central moments drawn are the same.

where

$$J(u, u^i) = \int_\Omega C_1^i (u - u^i)\,\mathrm{d}\mathbf{x} + \int_\Omega (u - u^i)\,\mathrm{d}\mathbf{x} \cdot \left[C_2^i \int_\Omega C_4 u\,\mathrm{d}\mathbf{x} - C_3^i \int_\Omega C_4 (1-u)\,\mathrm{d}\mathbf{x} \right], \tag{15}$$

and

$$C_1^i = \int_{\mathcal{Z}} K_{\mathbf{H}}(\mathbf{z} - I(\mathbf{x})) \cdot \log \left(\frac{\hat{p}_{\mathbf{Z}(\mathbf{X})|u^i(\mathbf{X})=0}(\mathbf{z})}{\hat{p}_{\mathbf{Z}(\mathbf{X})|u^i(\mathbf{X})=1}(\mathbf{z})} \right) \mathrm{d}\mathbf{z}, \tag{16}$$

$$C_2^i = \left(\int_\Omega u^i\,\mathrm{d}\mathbf{x} \right)^{-1}, \tag{17}$$

$$C_3^i = \left(\int_\Omega (1 - u^i)\,\mathrm{d}\mathbf{x} \right)^{-1}, \tag{18}$$

$$C_4 = \int_{\mathcal{Z}} K_{\mathbf{H}}(\mathbf{z} - I(\mathbf{x}))\,\mathrm{d}\mathbf{z} \cdot \log \mathrm{e}; \tag{19}$$

furthermore[2]*, the equality holds if, and only if,* $u(\mathbf{x}) = u^i(\mathbf{x})$*, for all* $\mathbf{x} \in \Omega$*.*

Proof. See the Supplementary Material. □

Corollary 1. $(E(u^i))_{i\geq 1}$ *is convergent.*

Proof. $(E(u^i))_{i\geq 1}$ is bounded below, since $0 \leq \hat{\mathcal{I}}(\mathbf{Z}(\mathbf{X}); u(\mathbf{X})) \leq 1$ bit, and $\int_\Omega g|Du| \geq 0$. □

Corollary 2. *The upper bound of* $E(u)$ *in* (14) *is convex in* $u \in BV(\Omega; [0,1])$*.*

Proof. For every $u_1, u_2 \in BV(\Omega; [0,1])$ and $0 \leq t \leq 1$, $\int_\Omega g|D(tu_1 + (1-t)u_2)| \leq \int_\Omega g|Dtu_1| + \int_\Omega g|D(1-t)u_2| = t\int_\Omega g|Du_1| + (1-t)\int_\Omega g|Du_2|$; and $J''(u, u^i) > 0$. □

[2]Note that C_1^i and C_4 are functions of $\mathbf{x}$; on the other hand, C_2^i, C_3^i are constants.

Consequently, the original nonconvex optimization problem (10) can be efficiently solved by minimizing the convex upper bound in Proposition 3. In addition, the mutual information estimate $\hat{\mathcal{I}}(\mathbf{Z}(\mathbf{X}); u^i(\mathbf{X}))$ in (14) is constant for a given u^i, we clearly have the following sequential convex programming problem:

$$\inf_{u \in BV(\Omega;[0,1])} \lambda J(u, u^i) + \int_\Omega g|Du|. \tag{20}$$

Here, one can easily check that each of the convex subproblems admits the existence of a minimizer as shown in the proof of Proposition 1.

For each instance again, let u_{opt} be a globally optimal binary solution of $F(u|u^i)$, and let $u^\star_{\text{bin}}$ be a thresholded version of $u^\star_{\text{rel}}$ which is a global minimizer of $F(u|u^i)$ over the relaxed domain $[0,1]$. Unfortunately, the coarea formula [14] does not hold for the proposed upper bounds. In other words, we simply cannot make use of the thresholding theorem [20] to show that $F(u^\star_{\text{bin}}|u^i) = F(u_{\text{opt}}|u^i)$. Instead, we have the following *a posteriori* optimality bound:

$$\left| F(u^\star_{\text{bin}}|u^i) - F(u_{\text{opt}}|u^i) \right| \leq \left| F(u^\star_{\text{bin}}|u^i) - F(u^\star_{\text{rel}}|u^i) \right|. \tag{21}$$

In practice, the relative bounds are typically less than 1%—see Section 4 for actual numerical values we obtained.

3.2. Implementation

3.2.1 Primal-Dual Optimization

The convex energy upper bounds in (20) are not differentiable because the Euler-Lagrange equation of the (weighted) TV term exhibits a singularity at $u = 0$. In order to remedy that, we reformulate the bounds to equivalent

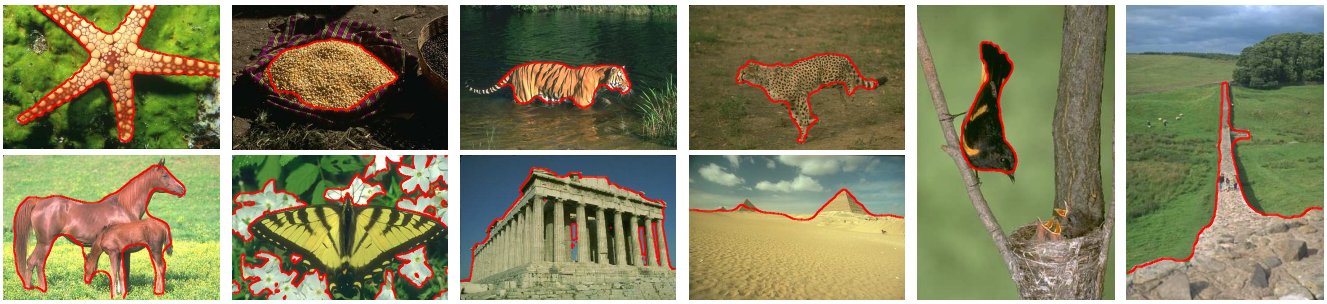


Figure 7. **Unsupervised figure-ground segmentation for vector-valued images.** Interestingly, the results obtained by minimizing the mutual-information-based functional do not always match human expectations: For example, rather than separating the butterfly from the background, the algorithm separated the white flowers, which are more likely distinguishable from the remaining colors.

saddle-point problems [13] as follows:

$$\min_{u \in BV(\Omega;[0,1])} \sup_{\xi \in \mathcal{K}_g} \lambda J(u, u^i) - \int_\Omega u \operatorname{div} \xi \, d\mathbf{x}, \tag{22}$$

with the convex set $\mathcal{K}_g = \left\{ \xi \in C_c^1(\Omega, \mathbb{R}^2) : |\xi| \leq g \right\}$, where $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^2$ and g is an edge indicator function. We solve (22) by means of a primal-dual algorithm of Arrow-Hurwicz type [4], which essentially performs a projected gradient ascent in the dual variable followed by a projected gradient descent in the primal variable. Additionally, we perform a subsequent over-relaxation step for the primal variables as in [8], in order to improve the convergence of the algorithm. Overall, the algorithm point-wise iterates the following update steps:

$$\xi^{k+1}(\mathbf{x}) = \Pi_{\mathcal{K}_g} \left(\xi^k(\mathbf{x}) + \tau \nabla \bar{u}^k(\mathbf{x}) \right) \tag{23}$$

$$u^{k+1}(\mathbf{x}) = \Pi_{[0,1]} \left(u^k(\mathbf{x}) - \sigma \left(\partial_u J - \operatorname{div} \xi^k(\mathbf{x}) \right) \right) \tag{24}$$

$$\bar{u}^{k+1}(\mathbf{x}) = 2u^{k+1}(\mathbf{x}) - u^k(\mathbf{x}), \tag{25}$$

where (23) and (24) are the update steps for the dual and primal variables ($\Pi_{\mathcal{K}_g}$ and $\Pi_{[0,1]}$ are orthogonal projection operators onto the sets $\mathcal{K}_g$ and $[0,1]$), respectively; (25) is an over-relaxation step. For the step sizes (τ, σ), we chose sufficiently small values.

3.2.2 Kernel Density Estimation

We use a Gaussian function with a diagonal matrix $\mathbf{H}$ for $K_{\mathbf{H}}(\mathbf{z})$. For the sake of efficient implementation, we channel-wise estimate the color distributions of vector-valued images as follows,

$$K_{\mathbf{H}}(\mathbf{z} - I(\mathbf{x})) = \prod_{j=1}^{n} \frac{1}{2\pi\sigma_j^2} \exp\left(-\frac{|\mathbf{z}_j - I_j(\mathbf{x})|^2}{2\sigma_j^2} \right), \tag{26}$$

where σ_j is the j-th diagonal element of $\mathbf{H}$ and $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^n$. This estimation derives from the assumption that the color channels are mutually independent which is not necessarily the case in natural images. However, computing the log-likelihood ratio at each step is computationally demanding during optimization and would render our approach impractical. By decoupling the color channels, we are able to achieve drastic improvements of the computation time with reasonably meaningful figure-ground separation. The runtime of the algorithm is approximately 5 minutes with unoptimized Matlab code on an 400×1200 image. Our algorithm runs on a Intel Core I7 CPU computer with 2.67 GHz and 4 GB of memory.

4. Experimental Results

The segmentation results obtained by our approach are mostly well matched to human expectations, compared to those from algorithms based on the Chan-Vese model [9] and its convex formulations. The reason is that the mutual information takes into account all the higher-order statistics, whereas the algorithms for the Chan-Vese model including its convex formulations consider the 1st central moment only. We present two compelling examples in Figure 6 where such algorithms completely fail to separate the hidden distributions. Here, the figure and ground basically cannot be distinguished up to the 2nd central moment. Even in the challenging Rayleigh distribution example where all the even central moments are the same [18], the proposed method is capable of separating the foreground and background. We also demonstrate Corollary 1 in Figure 5 by plotting the evolution of the mutual information and the overall energy with respect to the number of iterations for the unimodal/bimodal distribution case. Indeed, in many cases, the proposed algorithm converges to the global optimum of the original problem, and it is robust to initialization compared to [17] and [18].

The gradient based curve evolution of Kim *et al.* [17] fails to segment the zebra image in Figure 1 where even a bounding box was used for an initialization to enclose the object to be segmented. Even with a good initialization—multiple contours are typically used in the curve evolution to cover the entire image—it may not well separate the foreground and background in the challenging synthetic image

in the bottom row of Figure 1.

As we mentioned in Section 1.3, extending the proposed method to vector-valued images is straightforward. Some results from the the Berkeley Segmentation Dataset are shown in Figure 7. In the case of the segmented flowers in the butterfly image, we did not expect such an interesting separation. This example could provide evidence why unsupervised figure-ground separation is not trivial, even to humans. As for the relative optimality bounds, we obtain in average values between 0.1% to 0.5%.

5. Conclusion

We have proposed an optimization algorithm for unsupervised figure-ground separation. Through sequential convex relaxation of a sequence of convex upper bounds, we have efficiently minimized the original nonconvex mutual-information-based energy functional. In contrast to the approach where convex relaxation techniques are combined with the Wasserstein distance, our approach is straightforwardly generalizable to vector-valued images. Experimental results are remarkable in that they are in fairly good agreement with those from humans; yet, some results are quite surprising as the proposed approach is able to separate distributions where humans can hardly see the differences.

Lastly, we should mention that it does not seem possible to guarantee optimality for solutions of the original nonconvex problem due to the nature of sequential optimization.

References

[1] I. Ahmad and P.-E. Lin. A nonparametric estimation of the entropy for absolutely continuous distributions (corresp.). *IEEE Trans. Inf. Theory*, 22(3):372–375, 1976. 4

[2] G. Alberti, G. Bouchitté, and G. Dal Maso. The calibration method for the Mumford–Shah functional and free-discontinuity problems. *Calc. Var. Partial Differential Equations*, 16(3):299–333, 2003. 2

[3] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford: Clarendon Press, 2000. 3, 4

[4] K. J. Arrow, L. Hurwicz, and H. Uzawa. *Studies in Linear and Non-linear Programming*. Stanford University Press, 1958. 7

[5] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. van der Meulen. Nonparametric entropy estimation: An overview. *Int. J. Math. Stat. Sci.*, 6:17–40, 1997. 4

[6] X. Bresson, S. Esedoglu, P. Vandergheynst, J.-P. Thiran, and S. Osher. Fast global minimization of the active contour/snake model. *J. Math. Imaging Vis.*, 28(2):151–167, 2007. 4

[7] E. S. Brown, T. F. Chan, and X. Bresson. Completely convex formulation of the chan-vese image segmentation model. *Int. J. Comput. Vision*, 98(1):103–121, 2012. 1, 2

[8] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40(1):120–145, 2011. 7

[9] T. F. Chan and L. A. Vese. Active contours without edges. *IEEE Trans. Image Process.*, 10(2):266–277, 2001. 1, 2, 7

[10] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006. 4

[11] D. Cremers, T. Pock, K. Kolev, and A. Chambolle. Convex relaxation techniques for segmentation, stereo and multiview reconstruction. In *Markov Random Fields for Vision and Image Processing*. MIT Press, 2011. 2

[12] Y. G. Dmitriev and F. P. Tarasenko. On the estimation of functionals of the probability density and its derivatives. *Theory Probab. Appl.*, 18(3):628–633, 1974. 4

[13] I. Ekeland and R. Témam. *Convex Analysis and Variational Problems*. Classics in Applied Mathematics. SIAM, 1999. 7

[14] H. Federer. *Geometric Measure Theory*. Springer, 1969. 5, 6

[15] A. Herbulot, S. Jehan-Besson, S. Duffner, M. Barlaud, and G. Aubert. Segmentation of vectorial image features using shape gradients and information measures. *J. Math. Imaging Vis.*, 25(3):365–386, 2006. 1

[16] D. R. Hunter and K. Lange. A tutorial on MM algorithms. *Amer. Statist.*, 58(1):30–37, 2004. 5

[17] J. Kim, J. W. Fisher III, A. Yezzi, M. Çetin, and A. S. Willsky. A nonparametric statistical method for image segmentation using information theory and curve evolution. *IEEE Trans. Image Process.*, 14(10):1486–1502, 2005. 1, 2, 3, 4, 7

[18] O. Michailovich, Y. Rathi, and A. Tannenbaum. Image segmentation using active contours driven by the Bhattacharyya gradient flow. *IEEE Trans. Image Process.*, 16(11):2787–2801, 2007. 1, 6, 7

[19] K. Ni, X. Bresson, T. F. Chan, and S. Esedoglu. Local histogram based segmentation using the wasserstein distance. *Int. J. Comput. Vision*, 84(1):97–111, 2009. 2

[20] M. Nikolova, S. Esedoglu, and T. F. Chan. Algorithms for finding global minimizers of image segmentation and denoising models. *SIAM J. Appl. Math.*, 66(5):1632–1648, 2006. 2, 5, 6

[21] S. Osher and J. A. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton–Jacobi formulations. *J. Comput. Phys.*, 79(1):12–49, 1988. 1, 2

[22] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962. 1

[23] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. An algorithm for minimizing the piecewise smooth Mumford–Shah functional. In *ICCV*, 2009. 2

[24] K. Punithakumar, J. Yuan, I. B. Ayed, S. Li, and Y. Boykov. A convex max-flow approach to distribution based figure-ground separation. *SIAM J. Imaging Sci.*, 2012. 2, 3, 5

[25] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956. 1

[26] J. S. Simonoff. *Smoothing Methods in Statistics*. Springer, 1996. 4

[27] R. W. Yeung. *Information Theory and Network Coding*. Springer, 2008. 5