Inferring Segmentation Label and Color Distribution in a Unified Framework using Global Constraints

Viet-Quoc Pham†, Keita Takahashi††, Takeshi Naemura†††

Abstract In this paper, we propose a unified framework for inferring the segmentation label and color distribution of an image region of interest. Recent studies have shown that segmentation with global consistency measures outperforms conventional techniques based on pixel-wise measures. However, such global approaches require a precise input distribution to obtain the correct extraction. To overcome this strict assumption, we propose a new approach in which the given reference distribution plays a guiding role in inferring the latent distribution and its consistent region. The inference is based on an assumption that the latent distribution resembles the distribution of the consistent region but is distinct from the distribution of the complement region. We state the problem as the minimization of an energy function consisting of global similarities and implement an iterative scheme for jointly optimizing distribution and segmentation. Rich experimental results demonstrate the advantages of using our approach with various segmentation problems.

Key words: image segmentation, graph cut, global constraint energy, background model estimation

1. Introduction

Foreground-background segmentation is a technique for extracting a foreground region in a picture or photograph from its background. The segmentation result can be used for many applications, such as object recognition and image editing. Because segmentation is a strictly ambiguous problem, most approaches rely on additional user guidance or some prior information regarding the region of interest. Such input may be given in the form of hard constraints, such as marked pixels in seeded segmentation\(^1\)\textendash\(4\), or soft constraints, such as color cues of foreground and/or background regions\(^5\)\textendash\(8\).

This paper addresses the problem of foreground/background segmentation with a reference distribution. The given distribution may come from the same image or from a different image, depending on the purpose of the segmentation problem. Most approaches to this problem treat it as a binary labeling problem of minimizing a cost function. This cost function often consists of two constraint terms related to the label \(L\): appearance consistency and spatial smoothness, because the desired segmentation result is a smoothed region consistent with the distribution. Depending on the formulation of the appearance consistency term, we divided these approaches into two classes: pixel-wise measure based and global measure based.

In standard techniques based on pixel-wise consistency measures\(^1\)\textendash\(3\), the appearance consistency term evaluates each pixel individually for how well it fits the distribution. Therefore, this term, often called the likelihood term, can be expressed by a sum of unary terms. The cost function then consists of only unary and pairwise (spatial smoothness) terms.

\[
\mathcal{E}(L) = \sum_{\text{pixel } i} f_i(L_i) + \sum_{\text{neighbor } (i,j)} g_{(i,j)}(L_i, L_j) \quad (1)
\]

Minimization of such a function can be performed effectively by using GraphCut optimization\(^9\)\textendash\(10\). In this optimization method, minimizing the cost function is equivalent to computing a max-flow of a corresponding graph, and it can thus guarantee global optima and computational effectiveness. In\(^1\)\textendash\(3\), the distribution is fixed, and the optimal solution of the segmentation label can be obtained from one GraphCut computation. However, in situations where the given distribution is not reliable, such as the interactive segmentation interface from a user-defined bounding box\(^6\), only one GraphCut calculation cannot guarantee a good solution. Rother et al. proposed an iterated process called GrabCut that jointly optimizes segmentation and distributions\(^6\). However, in each iteration, the appearance consistency is still based on a pixel-wise model. Because the optimization does not consider the global consistency of the segmentation data, the typical GraphCut methods are subject to several disadvantages, such as the boundary-length...
bias (shrinking bias), which makes segmentation prone to shortcutting, as shown in \(4\).

Recent studies\(5,7\) have shown that segmentation with global similarity measures outperforms conventional techniques based on pixel-wise consistency measures. Their energy functions have a common form as follows.

\[
E(L) = f(1,2,...,n)(L_1, L_2, ..., L_n) + \sum_{\text{neighbor } (i,j)} g(i,j)(L_i, L_j)
\]  

(2)

The presence of the global term \(f(1,2,...,n)(\cdot)\), which evaluates global similarities, makes the minimization problem quite challenging. The energy function cannot be expressed purely in the form of unary and pairwise terms and thus cannot be solved directly by GraphCut. Rother et al. formulated the appearance consistency by using the \(L_1\) norm of the difference between histograms to find an image region consistent with a reference histogram\(7\). The optimization is based on a method called Trust Region Graph Cut (TRGC) that computes a sequence of parametric linear approximations. As stated in a study by I. Ben Ayed et al.\(5\), TRGC suffers from the fundamental disadvantage of dependence on the region size. In a different approach, this team used a global constraint based on the Bhattacharyya measure\(5\). They introduced an auxiliary labeling method to estimate the upper bound of the constraint term and performed optimization of an auxiliary function with GraphCut. When the reference distribution was learned from the ground truth, they demonstrated that their segmentation results have extremely high qualities that outperform TRGC\(5\) and most active contour based methods\(1,11\)-\(13\). We tested this method on a sample image, the results of which are shown in Fig. 1. The reference distribution was computed from the white-masked region. In Fig. 1(a), we show the result with the distribution learned from the true region. In Fig. 1(b), we show the result with the inexact distribution learned from the region outside the true background. While we agree that the segmentation is successful when the reference distribution is precise (see Fig. 1(a)), requiring precise distribution is a strict assumption and not practical. For example, in some interactive segmentation interfaces\(6\), the distribution is learned from a user-defined region and is generally not precise. In such situations, directly applying global distribution matching will result in a poor performance, as shown in Fig. 1(b).

In this paper, we propose an approach for image segmentation from an inexact reference distribution in which the given reference distribution plays a guiding role to infer the latent distribution and its consistent region in a unified framework. The inference is based on the assumption that the latent distribution resembles the distribution of the consistent region but is distinct from the distribution of the complement region. We state the problem as the minimization of an energy function consisting of global similarities based on the Bhattacharyya distance. We first infer the latent distribution by joint estimation with the segmentation label, based on a process of *iterated distribution matching*. When a well-estimated distribution is obtained, we optimize the segmentation label with a final refinement. We evaluated our algorithm on the GrabCut dataset, and demonstrate the advantages of using our approach for various segmentation problems, including interactive segmentation, background subtraction, co-segmentation, and multiple image segmentation.

A previous conference version of this paper has been published in \(14\). In this paper, we provide a complete discussion on our inference model, a new application and experimental results. We present our algorithm in Section 2. Experimental results are shown in Section 3. We introduce some important applications of our model in Section 4. Finally, Section 5 concludes our paper.

2. Algorithm

2.1 Inference of Latent Distribution

Let \(M\) be the given reference distribution. The purpose of our algorithm is to compute the latent distribution \(H\) and its consistent region, guided by \(M\). In this section, we address the problem of segmentation from a user-defined bounding box\(6\) to illustrate our algorithm. In this case, \(M\) is the (known) color distribution of the region outside the box. Our algorithm is performed on the image region inside the box, and its outputs are both the distribution \(H\) of the background color inside this domain, and the consistent background region \(H^L_0\) defined by the segmentation label \(L\). Our key idea is that the two unknowns \(H\) and \(L\) should balance the fol-
lowing conditions.

- [Cond.1] $\mathcal{H}$ is close to the reference distribution $\mathcal{M}$.
- The segmentation label $L$ separates the domain (the region inside the box) into two partially smoothed parts, $R_L^0 = \{p \mid L_p = 1\}$ and $R_L^1 = \{p \mid L_p = 0\}$, that satisfies
  - [Cond.2] $\mathcal{H}$ resembles $R_L^0$ distribution,
  - [Cond.3] $\mathcal{H}$ is distinct from $R_L^1$ distribution.

The third condition [Cond.3] may be not understandable at first glance. In fact, without this condition, the optimized solution of $\mathcal{H}$ turns out to be $\mathcal{M}$, and the problem then is to find an image region $R_L^0$ consistent with the reference distribution $\mathcal{M}$. Consequently, for this case, we obtain the same model as $^5$, which does not estimate the latent distribution, neither consider the distribution of the complement region. Its resulting performance is poor, as in Fig. 1(b).

We state the problem as the minimization of an energy function unifying the above conditions.

\[
\mathcal{E} (\mathcal{H}, L) = - B(\mathcal{M}, \mathcal{H}) - B(\mathcal{P}_0(L), \mathcal{H}) + B(\mathcal{P}_1(L), \mathcal{H}) + \lambda S(L)
\]  

(3)

where $\mathcal{P}_i(L)$ is the distribution of the region $R_L^i$ and $B(f, g)$ denotes the similarity of the two distributions. Inspired by previous successful usages of the Bhattacharyya coefficient $^5$, we also use this distance to formulate $B$:

\[
B(f, g) = \sum_{z \in Z} \sqrt{f(z)g(z)}
\]

(4)

$S(L)$ is a pairwise term that assigns a smoothness penalty to $L$. The presence of $S$ is to prevent $L$ from spatial dispersal, which may cause $\mathcal{H}$ to be over-fitted. However, the proportion of $S$ in (3) should be small so as not to affect the main conditions related to $\mathcal{H}$. I. Ben Ayed et al.’s model $^5$ considered only the second and fourth terms in (3). As compared to theirs, our model is more complex, including both positive and negative terms of Bhattacharyya distance $B$. Our optimization scheme requires both the lower and upper bounds of $B$, while only the lower bound was considered in $^5$.

We minimize $\mathcal{E} (\mathcal{H}, L)$ by jointly optimizing $\mathcal{H}$ and $L$, i.e. alternatively fixing one unknown and optimizing the other.

\[
\begin{align*}
L^{(t)} &= \arg \min_L \mathcal{E}(\mathcal{H}, \mathcal{L}^{(t-1)}) \\
\mathcal{H}^{(t)} &= \arg \min_{\mathcal{H}} \mathcal{E}(L^{(t)}, \mathcal{H})
\end{align*}
\]

(5)

The problem is shown in Lemma 2.2. An important advantage of the above two properties is that we can generate an non-increasing sequence \{\mathcal{E} (L^{(t)})\}, as proved in Lemma 2.3. Therefore, a local minimum of $\mathcal{E}_1$ can be obtained by iteratively optimizing the auxiliary function, as illustrated in Fig. 2.

\[
\mathcal{E}_1(L, L^*) \geq \mathcal{E}_1(L)
\]

(6)

\[
\mathcal{E}_1(L, L) = \mathcal{E}_1(L)
\]

(7)

(8)

(9)

Preparation Before getting into our formulation, we first introduce some notations. Let $I$ be the target image region for segmentation. In the case of bounding-box based segmentation, $I$ is the region inside the bounding box. $I_p$ denotes the color value of pixel $p$. $\mathcal{P}_i$ is the kernel density estimate (KDE) of the distribution $\mathcal{P}_i(L)$ ($i = 0, 1$):

\[
\mathcal{P}_i(z) = \frac{\sum_{p \in R^i} K_z(I_p)}{A(R^i)} \quad \forall z \in Z
\]

(10)
where $A(R)$ denotes the number of pixels within region $R$. In this paper, we use histograms for density estimation:

$$K_z(I_p) = \begin{cases} 
1 & \text{if } I_p \in z \\
0 & \text{if } I_p \notin z 
\end{cases} \quad (11)$$

Here, $z$ is the domain of one histogram bin (e.g., if we use a $128 \times 128 \times 128$ histogram, each $z$ has a cube size of $2 \times 2 \times 2$). $Z$ is the set of all $z$s.

**Lemma 2.1.** Given a fixed label $L^*$, for any label $L$ satisfying $R^*_1 \supset R^*_1$, i.e., the foreground region defined by $L$ contains the foreground region defined by $L^*$, and $\forall \alpha \geq 0$, we have an upper bound of $\mathcal{F}(L) = \mathcal{F}_1(L) + \mathcal{F}_2(L)$ as

$$Q(L, L^*, \alpha) = \sum_{p \in R^*_1} m_p(1) + \left(\text{sign}(\mathcal{F}(L^*))\alpha + 1\right) \sum_{p \in R^*_0} m_p(0) \quad (12)$$

where $m_p(1)$ and $m_p(0)$ are unary terms given for each $p$:

$$m_p(1) = -\frac{\delta_{L^*_0}}{2A(R^*_1)} \sum_{z \in z} K_z(I_p) \sqrt{\mathcal{H}(z) - \mathcal{P}_{L^*_1}(z)}$$

$$+ \frac{\delta_{L^*_0}}{A(R^*_0)} \left(\sum_{z \in z} K_z(I_p) \sqrt{\mathcal{H}(z) - \mathcal{P}_{L^*_0}(z)} + \mathcal{F}(L^*)\right)$$

$$m_p(0) = \frac{\mathcal{F}(L^*)}{A(R^*_0)}$$

Here, $\delta$ is the Kronecker delta function. *

**Proof.** Finding the upper bounds of $\mathcal{F}_1$ and $\mathcal{F}_2$ requires both an upper bound and a lower bound of the Bhattacharyya distance $B$. A lower bound of $B$ was considered in a previous study 5. In this paper, we propose an upper bound of $B$, which has not been considered before. Details of the proof are explained in Appendix A.

**Lemma 2.2.** For $\alpha = 0$, the following auxiliary function satisfies the two conditions (8) and (9)

$$\hat{\mathcal{E}}_1(L, L^*, \alpha) = Q(L, L^*, \alpha) + \lambda S(L) \quad (14)$$

**Proof.** (8) is proved directly from Lemma 2.1. When $L^* = L$, we have $\forall p \in R^*_1$: $\delta_{L^*_0} = 0$. Therefore, $\sum_{p \in R^*_1} m_p(1) = 0$, and

$$\hat{\mathcal{E}}_1(L, L, 0) = \sum_{p \in R^*_0} \frac{\mathcal{F}(L)}{A(R^*_0)} + \lambda S(L) = \mathcal{E}_1(L) \quad (15)$$

This equality proves (9).

**Lemma 2.3.** $\mathcal{E}_1$ is non-increasing under the update

$$L^{(\tau+1)} := \arg\min_{L: R^*_1 \supset R^*_1} \hat{\mathcal{E}}_1(L, L^{(\tau)}) \quad (16)$$

* When $R^*_1 = \emptyset$, we consider $A(R^*_1)$ and $\mathcal{P}_{L^*_0}(z)$ as 1.

**Proof.** $\mathcal{E}_1(L^{(\tau+1)}) \leq \hat{\mathcal{E}}_1(L^{(\tau+1)}, L^{(\tau)}) \leq \hat{\mathcal{E}}_1(L^{(\tau)}, L^{(\tau)}) = \hat{\mathcal{E}}_1(L^{(\tau)})$ \hfill \Box

(2) Optimizing $\mathcal{H}$

Because $L$ is constant, minimizing $\mathcal{E}(L^{(t)}, \mathcal{H})$ in (6) is equivalent to minimizing the following function.

$$\mathcal{E}_2(\mathcal{H}) = -B(\mathcal{M}, \mathcal{H}) - B(\mathcal{P}_0(L), \mathcal{H}) + B(\mathcal{P}_1(L), \mathcal{H}) \quad (17)$$

**Lemma 2.4.** The optimized solution of $\mathcal{E}_2(\mathcal{H})$ is

$$\mathcal{H}(z) = \left\{ \begin{array}{ll}
0 & (\forall z \in Z^-) \\
\frac{a^2}{\sum_z \mathcal{H}(z)} & (\forall z \in Z^+) 
\end{array} \right. \quad (18)$$

where $a_z$ is constant with fixed $L$.

$$a_z = \sqrt{\mathcal{M}(z) + \mathcal{P}_{L^*_1}(z) - \mathcal{P}_{L^*_0}(z)} \quad (19)$$

and $Z^+ = \{z | a_z > 0\}$, $Z^- = \{z | a_z \leq 0\}$.

**Proof.** From (17) and (19), we have

$-\mathcal{E}_2(\mathcal{H}) = \sum_z a_z \sqrt{\mathcal{H}(z)} \leq \sum_z a_z \sqrt{\mathcal{H}(z)} \leq \sqrt{\left(\sum_z a^2\right) \left(\sum_z \mathcal{H}(z)\right)} \leq \sqrt{\left(\sum_z a^2\right) \left(\sum_z \mathcal{H}(z)\right)} \quad (22)$

(22) is given by Cauchy-Schwarz inequality, and (23) is from the fact that $\sum_{z \in Z} \mathcal{H}(z) = 1$. Equality holds if and only if (18) is satisfied. \hfill \Box

(3) Procedure Overview

The overview of the procedure to estimate the latent distribution $\mathcal{H}$ is as follows.

**Procedure 1:** Iterated distribution matching

1: Initialize $\mathcal{H}$: $\mathcal{H}^{(0)} = \mathcal{M}$
2: repeat
3: Update $L^{(t)} = \arg\min \mathcal{E}(L, \mathcal{H}^{(t-1)})$ using Procedure 2
4: Update $\mathcal{H}^{(t)} = \arg\min \mathcal{E}(L^{(t)}, \mathcal{H})$ by applying (18) \hfill \forall t
5: until convergence
6: return $\mathcal{H} = \mathcal{H}^{(t)}$

The updating process of $\mathcal{H}$ in line 4 comes directly from Lemma 2.4. Updating $L$ in line 3 by minimizing $\hat{\mathcal{E}}_1(L)$ is instructed by Lemma 2.1, 2.2 and 2.3. We use the same procedure as I. Ben Ayed et al. 5 to perform this process.

**Procedure 2:** Auxiliary labeling

1: Initialize $\mathcal{H}$: $\mathcal{H}^{(0)} = \mathcal{M}$
2: repeat
3: Update $L^{(t)} = \arg\min \mathcal{E}(L, \mathcal{H}^{(t-1)})$ using Procedure 2
4: Update $\mathcal{H}^{(t)} = \arg\min \mathcal{E}(L^{(t)}, \mathcal{H})$ by applying (18) \hfill \forall t
5: until convergence
6: return $\mathcal{H} = \mathcal{H}^{(t)}$
1: Initialize $L^*: L^*_p = 0 \ \forall p \in I$
2: Initialize $\alpha: \alpha = \alpha_0$ (small positive value)
3: repeat
4: Optimize auxiliary function: $L^{(r)} = \arg\min_{L: R^t \supset R^t_p} \tilde{E}_1(L, L^*, \alpha)$
5: Update $L^*: L^* = L^{(r)}$
6: Decrease $\alpha: \alpha = \alpha^q$ with $\rho > 1$
7: until convergence

This process converges when $\alpha$ approaches 0. The presence of $\alpha$ is necessary to put an additive penalty to each $p$ in $R^t_0$, with the purpose of biasing the resulting label towards enlarging the foreground region. The auxiliary function $\tilde{E}_1$ in line 4 contains only unary and pairwise terms and can thus be optimized effectively by applying GraphCut optimization\(^9\). The condition $R^t_k \supset R^t_p$ can be satisfied easily by adding a hard constraint to the label, as was used in\(^1\).

In our implementation, to prevent noise from propagating, we erode $R^t_p$ by some small pixels (typically two) before setting it as the foreground hard constraint.

2.2 Final Label Refinement

Let $\tilde{H}$ be the latent distribution obtained from Procedure 1. As shown in Fig. 3, $\tilde{H}$ improves the reference distribution $\mathcal{M}$ significantly by its close approach to the ground truth distribution. In this section, we discuss how to use this inferred distribution to extract the consistent image region.

A straight-forward approach is to directly apply the resulting label $L$ in Procedure 1 by optimizing the cost function $\mathcal{E}(L, \tilde{H})$ over $L$ (see (3)). We employ this optimization approach, however, with a small modification. It is important to note that the objective of the cost function $\mathcal{E}(L, \tilde{H})$ is to balance the three main conditions related to $\tilde{H}$, as stated in the beginning of Section 2.1. However, at this point, when the distribution $\tilde{H}$ is well-estimated, we do not need to take the term $-B(\mathcal{M}, \tilde{H})$ into account. Instead, we should increase the influence of the consistency constraint $-B(P_0(L), \tilde{H})$ and the smoothness prior $S(L)$, because they are the two main factors that define the region consistent with the estimated distribution. Therefore, we modify the cost function as follows.

$$\tilde{E}(L) = \gamma B(P_1(L), \tilde{H}) - B(P_0(L), \tilde{H}) + \lambda' S(L)$$  \hspace{1cm} (24)

where $\gamma < 1$ and $\lambda' > \lambda$. When $\gamma = 0$, we obtain the same cost function as\(^5\). Finally, the segmentation result is obtained by optimizing this function $\tilde{L} = \arg\min \tilde{E}(L)$. To minimize $\tilde{E}(L)$, we perform an additive procedure that is almost identical to Procedure 2. The only difference is that the weighting factors of $\mathcal{F}_1(L)$ and $S(L)$ are changed.

3. Experiment

In this section, we present experimental results using the proposed method for the problem of bounding-box based interactive segmentation\(^6\).

We first give some implementation details. The smoothness prior term $S(L)$ is given as

$$S(L) = \sum_{(p, q) \in N} \delta_{L_p + L_q} \left( \frac{1}{1 + ||I_p - I_q||^2} + \frac{\epsilon}{||p - q||} \right)$$  \hspace{1cm} (25)

where $\epsilon = 10^{-3}$ and $N$ is the set of all 8-connected neighbor pixel pairs. In parenthesis, the first term contributes to region continuity and the second term minimizes the length of partition boundary. The weighting factor of $S(L)$ in (3) and (24) is $\lambda = 2 \times 10^{-3}$ and $\lambda' = 5 \times 10^{-3}$. The parameters in Procedure 2 are fixed at $\alpha_0 = 0.85$ and $\rho = 1.1$. We use the RGB color model, and a 3-dimensional histogram with $256 \times 256 \times 256$ bins was used as a density estimate.

In all the experiments described in this paper, we found that Procedure 1 converges after no more than three iterations, and Procedure 2 converges after, at most, five iterations, so we use these values to fix the number of iterations. For (24), we assign the weighting factor $\gamma$ of $B(P_1(L), \tilde{H})$ to 0. With this setting, the procedure to find the final segmentation result, by minimizing $\tilde{E}(L)$, converges faster and just three iterations can assure the convergence.

Inference of latent distribution: We investigate how the iterated distribution matching process improves the reference distribution. Fig. 3 shows the inferred distribution $\mathcal{H}^{(t)}$ along with label $L^{(t)}$, corresponding to the number of iterations $t$ of Procedure 1. The reference distribution $\mathcal{M}$ is computed from the outside region of the box. Our objective is to extract a background region inside the bounding box, guided by the reference distribution. When $t = 0$, i.e. in the case when the reference distribution $\mathcal{M}$ was used directly to extract the consistent region, as same as the method of\(^5\), the segmentation quality is poor due to the large difference between the reference and the ground truth distributions. By applying the iterated distribution matching process, the latent distribution $\mathcal{H}^{(t)}$, guided by $\mathcal{M}$, closely approaches the ground truth distribution, thus significantly improving the segmentation result.

Quantitative and qualitative evaluation in the case of interactive image segmentation: We report the results for the GrabCut database\(^6\) of 49 images with associated user-defined bounding box \(^*\). The color distribution of the outside

\(^*\) Like other research works, we exclude the image "cross.jpg" since the bounding box covers the whole image.
region of the box is used as the reference distribution. We perform iterated distribution matching (Procedure 1) to infer the latent background distribution $\tilde{H}$ (Section 2.1). This inferred distribution is then used to extract the consistent background region (Section 2.2). We employ error rate as a measurement to evaluate the results

\[
error\ rate = \frac{\text{No. of misclassified pixels}}{\text{No. pixels in inference region}} \quad (26)
\]

Here, the inference region is the region inside the box. We compute the average error rate over all sample images and compare our results with two related works: GrabCut\textsuperscript{6)} and Dual Decomposition (DD)\textsuperscript{8)}. As shown in Table 1, we obtained the best result among the three methods. We also tested\textsuperscript{5)} for the same problem, but the result was much poorer. The method in\textsuperscript{5)} assumed that the provided reference distribution is precise, which is not suitable for the bounding-box based segmentation task.

We also provide some segmentation samples in Fig. 4 and compare them to those obtained with the GrabCut method\textsuperscript{6)}. In most cases, our method provides visually better results than GrabCut. In contrast with GrabCut, which is based on pixel-wise fitting to estimate the appearance model, our approach uses global similarity measures in both the inference process of distribution and the optimization process of segmentation. As a result, our method is less prone to shortcutting and more robust to thin objects. However, there still remain some limitations to be resolved. Our method is more prone to failure where foreground and background regions have a considerable overlapping of color distribution, like the example of the tiger in Fig. 4. The weighting factors of $F_1(L)$ and $F_2(L)$ (the importance rates of condition 2 and 3 in Section 2.1) should be tuned manually or by learning for a better result. Besides, the values of $\alpha_0$ and $\rho$ should be selected carefully because they will affect the convergence rate and the final results. Experiment results in Fig. 5 show that too large $\rho$ values or too small $\alpha_0$ values will give worse results.

Regarding bounding-box based segmentation, it is worth mentioning the algorithm of Lempitsky et al.\textsuperscript{15)}. They employed a prior called “tightness”, stating that the segmented region should be sufficiently close to each side of the bounding box. Their algorithm is very slow due to its mathematical
complexity. Because this topological prior is out of the configuration of our problem, we do not consider this method in our performance comparison. However, we will investigate the possibility of combining this valuable prior with our own method in our future work.

**Calculation cost:** The running time of our method, in comparison with other methods, is shown in Table 1. All experiments were performed on a laptop PC with Intel Core 2 Duo, 3.0-GHz CPU and 4-GB RAM. Because the optimization process of our method consists of a small number of min-cut/ max-flow calculations\(^9\), we can obtain as fast a running speed as GrabCut.

### 4. Applications

In this section, we present three segmentation problems that can be solved using our method: background subtraction, co-segmentation and multiple image segmentation. While the first two applications can use the proposed model directly, the third application requires an extension of our model to deal with multiple images.

#### 4.1 Background Subtraction

The first application is background subtraction. Our process is very simple. We first calculate the reference distribution from the given background image and then apply our method for the segmentation task. We should notice that our method still works when the reference background comes from another image. An experimental result is shown in Fig. 6. Although we use the term ‘background subtraction’ here, our method is different from standard background subtraction methods like\(^{16}\), which require a stable background to perform segmentation. In Fig. 6, our method still succeeds even though the backgrounds of the target image and
the given background image are not identical due to camera movement.

4.2 Co-segmentation

The second problem is background co-segmentation. In this problem, we aim to simultaneously segment the common background parts of an image pair. The idea of co-segmentation was first introduced by Rother et al.7) and has elicited great interest recently71718). Our method can be applied to solve this problem: from the given image pair, alternatively choose one image and compute the reference background distribution from its entire region, then perform segmentation on the other image. A successful co-segmentation result from a difficult sample is shown in Fig. 7, in which the color distributions of the foreground and background are very similar in both images.

4.3 Multiple Image Segmentation: Inferring Background Model behind The Scene

In the third application, we discuss an important extension of our method to deal with multiple images. Different from the problem of “co-segmentation” above, here, we consider multiple images coming from the same scene, such as a video sequence. Our first motivation is to simultaneously perform segmentation on multiple images using a reference distribution as the only input. The second motivation is to infer the color distribution of the occluded background from the images. The result has many advantages besides the segmentation purpose. For example, we can use the resulting background distribution for the scene recognition task by matching it with the database.

(1) Inference Model

A straight-forward approach to this problem is to apply the proposed model in Section 2 on each image separately. However, as shown in Fig. 8, even if the segmentation tasks are successful, we only know the color distribution of each occluded background region in each image, which is often different from the true background of the scene. This explains the necessity of a new inference model, which takes the common background distribution and segmentation labels of all images as latent variables. We state the problem as the minimization of an energy function, which is an extension of (3)

\[
E(H, L_1, L_2, ..., L_n) = -B(M, H) + \sum_{1 \leq i \leq n} \{ -B(P_0(L_i^0), H) + B(P_1(L_i), H) + \lambda S(L_i) \}
\]  

(27)
where $\mathcal{H}$ is the latent distribution of the common background, $\mathbf{L}_i$ is the segmentation label of the $i^{th}$ image from the given $n$ images. We again apply the joint optimization, by alternatively fixing one unknown and optimizing the others.

\[
\begin{align*}
\arg\min_{\mathbf{L}_i} & -B(P_0(\mathbf{L}_i), \mathcal{H}) + B(P_1(\mathbf{L}_i), \mathcal{H}) + \lambda S(\mathbf{L}_i) \\
\text{subject to } & (\forall i = 1, n)
\end{align*}
(28)
\]

\[
\arg\min_{\mathcal{H}} -B(\mathcal{M}, \mathcal{H}) + \sum_{1 \leq i \leq n} (-B(P_0(\mathbf{L}_i), \mathcal{H}) + B(P_1(\mathbf{L}_i), \mathcal{H}))
(29)
\]

$\mathcal{H}$ is initialized by the reference distribution $\mathcal{H}^{(0)} = \mathcal{M}$. Optimization in (28) is performed by using the same method in Section 2 on each image $i$ separately. Optimization in (29) has the same solution as (18) with a small variation on the definition of constant $a_z$

\[
a_z = \sqrt{\mathcal{M}(z)} + \sum_{1 \leq i \leq n} (\sqrt{P_{0i}(z)} - \sqrt{P_{1i}(z)})
(30)
\]

As stated in Section 2, optimization related to $\mathbf{L}_i$ requires several iterations of GraphCut optimizations, thus its performance has a bigger calculation cost than optimization of $\mathcal{H}$ in (29). Therefore, the total calculation cost of our inference model in (27) increases near linearly in the number of images. This is another advantage of our model.

(2) Evaluation

To evaluate the performance of our model, we performed an experiment on four images in the first row of Fig. 8. These images were extracted from a video sequence, which was generated by superimposing a moving object on a known background. The object in the four images occludes different parts from the same background. The only manual input is a bounding-box around the object in the first image. We use the region outside this box to learn the reference distribution $\mathcal{M}$. Our objective is to infer the segmentation labels and the common background distribution of the four images by minimizing the energy function in (27). The iteration of joint optimization in (28) and (29) converged after 3 iterations. We show the segmentation labels in each iteration on row 2 - 4 in Fig. 8. Some segmentation errors appeared in the first iteration were corrected at the convergence point.

To evaluate the inferred background distribution $\mathcal{H}$, we calculated the Bhattacharyya distance between $\mathcal{H}$ and the ground-truth background distribution. The results are shown in Table. 2. We can conclude that the inferred distribution closely approaches the ground truth distribution from this result.
Table 2 Bhattacharyya distance to ground-truth distribution.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bhattacharyya measure</td>
<td>0.966</td>
<td>0.985</td>
<td>0.986</td>
</tr>
</tbody>
</table>

![image](app.Fig.1) Given a fixed label $L^*$, the variable label $L$, which separates the image into two regions $R_0^L$ and $R_1^L$, satisfies $R_1^L \supset R_0^L$.

5. Conclusion

In this paper, we propose an approach for image segmentation from an inexact reference distribution in which the given reference distribution plays a guiding role to infer the latent distribution and its consistent region label in a unified framework. The inference is based on the assumption that the latent distribution resembles the distribution of the consistent region but is distinct from the distribution of the complement region. We state the problem as the minimization of an energy function consisting of global similarities based on the Bhattacharyya distance and then implement a novel iterated distribution matching process for jointly optimizing distribution and segmentation. Experiments on many samples including the GrabCut dataset demonstrate that our approach outperforms the state-of-the-art approaches and can deal with various segmentation problems, including interactive segmentation, background subtraction, co-segmentation and multiple image segmentation.

Appendix

A. Proof of Lemma 2.1

First, we find an upper bound of $\mathcal{F}_1(L)$. This estimation is first proposed in this paper. From the given condition $R_1^L \supset R_0^L$, we have (see app. Fig. 1)

$$R_1^L = R_1^L \cup (R_1^L \cap R_0^L)$$ (app.1)

The kernel density estimate in (10) can be rewritten as

$$p_L^L(z) = \frac{\sum_{p \in R_1^L} K_z(I_p) + \sum_{p \in R_1^L \cap R_0^L} K_z(I_p)}{A(R_1^L)} + \frac{\sum_{p \in R_1^L \cap R_0^L} K_z(I_p)}{A(R_1^L \cap R_0^L)}$$

$$\leq \frac{\sum_{p \in R_1^L} K_z(I_p) + \sum_{p \in R_1^L \cap R_0^L} K_z(I_p)}{A(R_1^L)}$$ (app.2)

This inequality leads to an upper bound of $\mathcal{F}_1(L)$ (we omit $z$ as arguments to simplify the equations)

$$\mathcal{F}_1(L) = \sum_{z \in Z} p_L^L \mathcal{H}$$

$$\leq \sum_{z \in Z} \sqrt{\left(\frac{\sum_{p \in R_1^L \cap R_0^L} K_z(I_p)}{A(R_1^L)}\right) \mathcal{H}}$$

$$= \sum_{z \in Z} \sqrt{1 + \frac{\sum_{p \in R_1^L \cap R_0^L} K_z(I_p)}{\sum_{p \in R_1^L} K_z(I_p)}} \mathcal{H}$$ (app.3)

We apply the following inequality to bound the first term in the above sum

$$\sqrt{1 + \frac{x}{2}} \leq 1 + \frac{x}{2}, \quad \forall x \geq 0$$ (app.4)

to obtain an upper bound of $\mathcal{F}_1(L)$, which contains only unary and constant terms

$$\mathcal{F}_1(L) \leq \sum_{z \in Z} \left(1 + \frac{1}{2} \frac{\sum_{p \in R_1^L \cap R_0^L} K_z(I_p)}{\sum_{p \in R_1^L} K_z(I_p)}\right) \sqrt{p_L^L} \mathcal{H}$$

$$= \mathcal{F}_1(L^*) + \sum_{z \in Z} \left(1 + \frac{1}{2} \frac{\sum_{p \in R_1^L \cap R_0^L} K_z(I_p)}{\sum_{p \in R_1^L} K_z(I_p)}\right) \sqrt{p_L^L} \mathcal{H}$$

$$= \mathcal{F}_1(L^*) + \sum_{p \in R_1^L} \frac{\delta_{L_p^L=0}}{\sqrt{2A(R_1^L)}} \frac{\mathcal{H}(z)}{p_L^L(z)}$$ (app.5)

Next, we estimate an upper bound of $\mathcal{F}_2(L)$ using a method similar to 5. From the given condition $R_1^L \supset R_0^L$:

$$R_0^L = R_0^L \setminus (R_1^L \cap R_0^L)$$ (app.6)

The kernel density estimate in (10) can be rewritten as

$$p_L^L(z) = \frac{\sum_{p \in R_1^L} K_z(I_p) - \sum_{p \in R_1^L \cap R_0^L} K_z(I_p)}{A(R_1^L)} + \frac{\sum_{p \in R_1^L \cap R_0^L} K_z(I_p)}{A(R_1^L \cap R_0^L)}$$

$$\geq \frac{\sum_{p \in R_0^L} K_z(I_p) - \sum_{p \in R_1^L \cap R_0^L} K_z(I_p)}{A(R_0^L)}$$ (app.7)

This inequality leads to an estimation of $\mathcal{F}_2(L)$

$$\mathcal{F}_2(L) = -\sum_{z \in Z} \sqrt{p_0^L} \mathcal{H}$$

$$\leq -\sum_{z \in Z} \sqrt{\frac{p_L^L - \sum_{p \in R_1^L \cap R_0^L} K_z(I_p)}{A(R_1^L)}} \mathcal{H}$$

$$= -\sum_{z \in Z} \sqrt{1 - \frac{\sum_{p \in R_0^L} K_z(I_p)}{\sum_{p \in R_1^L \cap R_0^L} K_z(I_p)}} \sqrt{p_L^L} \mathcal{H}$$ (app.8)

The following inequality is then applied

$$\sqrt{1 - x} \geq 1 - x, \quad \forall x : 0 \leq x \leq 1$$ (app.9)

to obtain an upper bound of $\mathcal{F}_2(L)$
Finally, combining (app.5) and (app.10) gives an upper bound of $F(L) = \tilde{F}_1(L) + \tilde{F}_2(L)$

$F(L) \leq F(L^*) + \sum_{p \in R_1^*} \delta_{L^*} = 0 \sum_{z \in Z} K_z(I_p) \sqrt{\mathcal{H}(z) / P_{b_k}^L(z)}$

\[ = \tilde{F}_2(L^*) + \sum_{p \in R_1^*} \frac{\delta_{L^*} = 0}{A(R^*_k)} \sum_{z \in Z} K_z(I_p) \sqrt{\mathcal{H}(z) / P_{b_k}^L(z)} \quad \text{(app.11)} \]

The constant term $F(L^*)$ is then expressed by unary terms

$F(L^*) = \frac{\tilde{F}_2(L^*) R(R^*_k)}{A(R^*_k)} + \frac{F(L^*) A(R^*_k)}{A(R^*_k)}$

\[ \leq \sum_{p \in R_1^*} \frac{\delta_{L^*} = 0}{A(R^*_k)} \frac{F(L^*)}{A(R^*_k)} + \text{sign}(F(L^*)) \alpha + 1 \sum_{p \in R_1^*} \frac{F(L^*)}{A(R^*_k)} \quad \text{(app.12)} \]

Notice that $\text{sign}(F(L^*))$ is the sign of $F(L^*)$ and $\alpha \geq 0$.

Combining (app.11) and (app.12) proves Lemma 2.1.

References