Expert Systems With Applications 79 (2017) 90–100

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Robust interactive image segmentation using structure-aware labeling

Changjae Oh, Bumsub Ham, Kwanghoon Sohn*

The School of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea

ARTICLE INFO

Article history:
Received 14 September 2016
Revised 3 February 2017
Accepted 19 February 2017
Available online 27 February 2017

Keywords:
Interactive image segmentation
Co-occurrence probability
Graph-based optimization

ABSTRACT

Interactive image segmentation has remained an active research topic in image processing and graphics, since the user intention can be incorporated to enhance the performance. It can be employed to mobile devices which now allow user interaction as an input, enabling various applications. Most interactive segmentation methods assume that the initial labels are correctly and carefully assigned to some parts of regions to segment. Inaccurate labels, such as foreground labels in background regions for example, lead to incorrect segments, even by a small number of inaccurate labels, which is not appropriate for practical usage such as mobile application. In this paper, we present an interactive segmentation method that is robust to inaccurate initial labels (scribbles). To address this problem, we propose a structure-aware labeling method using occurrence and co-occurrence probability (OCP) of color values for each initial label in a unified framework. Occurrence probability captures a global distribution of all color values within each label, while co-occurrence one encodes a local distribution of color values around the label. We show that nonlocal regularization together with the OCP enables robust image segmentation to inaccurately assigned labels and alleviates a small-cut problem. We analyze theoretic relations of our approach to other segmentation methods. Intensive experiments with synthetic and manual labels show that our approach outperforms the state of the art.

1. Introduction

Image segmentation aims to split an image into discriminative regions that provide high-level information (e.g., objects and object parts). Unsupervised image segmentation (e.g., (Shi & Malik, 2000)) has been used as an off-the-shelf processing for a great variety of applications such as object proposal (Arbeláez, Pont-Tuset, Barron, Marques, & Malik, 2014) and object tracking (Ren & Malik, 2007). This segmentation approach does not need training data, and thus it does not reflect a user's intention.

Recent works on interactive image segmentation address this problem using *human prior*, e.g., bounded box (Rother, Kolmogorov, & Blake, 2004), snap (Mortensen & Barrett, 1995), and scribble (Bai & Wu, 2014; Boykov & Jolly, 2001; Casaca, Nonato, & Taubin, 2014; Grady, 2006; Jung & Kim, 2012; Kim, Lee, & Lee, 2008; Shen, Du, & Li, 2014; Subr, Paris, Soler, & Kautz, 2013; Unger, Pock, Trobin, Cremers, & Bischof, 2008). For example, the user manually labels some parts of regions to segment, and all the label information is then propagated into other parts of regions. This approach enhances segmentation accuracy with providing wider applications, e.g. image manipulation (Chen, Li, & Tang, 2013; Levin, Lischinski, & Weiss, 2004; Sheng, Sun, Magnor, & Li, 2014), 2D-3D conversion (Zhang, Zhou, Wang, & Gao, 2013).

Most interactive segmentation methods assume that the initial labels are reliable and correctly assigned to the regions the user wants to extract, and have focused on improving segmentation accuracy (Boykov & Jolly, 2001; Casaca et al., 2014; Grady, 2006; Jung & Kim, 2012; Kim et al., 2008; Shen et al., 2014; Unger et al., 2008). Such approaches thus require a lot of attention to label assignments, and give correct segments only with accurate initial labels. In addition, they are easily distracted by a small number of inaccurately assigned labels, which is not appropriate for practical use, e.g., in mobile applications.

To address these issues, we introduce a structure-aware labeling method that is robust to inaccurate initial labels. The proposed model considers both global and local color distribution around initial labels. In other words, we leverage occurrence and co-occurrence probability (OCP) of color values for each initial label. The OCP captures a global distribution of all color values *within* each label and a local distribution of color values *around* the label simultaneously. That is, the OCP considers color distributions of initial labels and the spatial relationship of color values, which enables estimating the reliability of each label effectively. We show that nonlocal regularization together with the OCP alleviates the influence of inaccurate labels on the segmentation result, and addresses a small-cut problem.

* Corresponding author.
E-mail addresses: ocj1211@yonsei.ac.kr (C. Oh), mimo@yonsei.ac.kr (B. Ham), khsohn@yonsei.ac.kr (K. Sohn).

http://dx.doi.org/10.1016/j.eswa.2017.02.031
0957-4174/

Thanks to the structure-aware model with nonlocal regularization, our work shows several interesting features comparing to existing approaches: First, the proposed method outperforms existing methods when inaccurate labels are included. Second, the proposed method still shows competitive results when all initial labels are correctly located.

The remainder of this paper is organized as follows. In Section 2, we review representative works related to ours. Our approach to robust image segmentation is then described in Section 3. In Section 4, we present experimental results and comparisons between our approach and the state of the art. Finally, we conclude the paper in Section 5.

2. Related works

In this section, we review current interactive segmentation methods related to our work.

2.1. Interactive image segmentation

Various graph-based optimization techniques have been adapted to interactive image segmentation in such a way that image pixels and their relationships are incorporated into corresponding graph models (Boykov & Jolly, 2001; Casaca et al., 2014; Grady, 2006; Kim et al., 2008). An approach using graph-cut (GC) (Boykov & Jolly, 2001) is the representative work. This method minimizes color variation within foreground objects and background by minimizing total edge weights in the cut, but this causes a small-cut problem. A single region can be split into many sub-regions. This problem can be addressed by using a random walk (RW) model (Grady, 2006). This model regards each pixel as a random walker, and compute the first arrival probability that the random walker reaches to each label. Although the RW-based approach is free from the small-cut problem, it does not handle weak boundaries and highly textured-regions. The first arrival probability does not consider the relation between the starting location of each random walker and the locations inside the initial labels. The random walk with restart (RWR) model (Kim et al., 2008) considers this relation, and shows an excellent performance for an image that has weak boundaries and textures. In Casaca et al. (2014), the Laplacian coordinate (LC) has been introduced a higher-order graph model of the RW based image segmentation. The higher-order interaction between pixels further regularizes segmentation results by propagating the label information to longer distances. More recently, a subMarkov RW has been proposed, which unifies existing RW-based models by regarding the random walker with Markov transition probability (Dong, Shen, Shao, & Van Gool, 2016).

Despite differences in graph construction and optimization, these methods share the same limitation (Boykov & Jolly, 2001; Casaca et al., 2014; Couprie, Grady, Najman, & Talbot, 2011; Dong et al., 2016; Grady, 2006; Kim et al., 2008): they focus on improving segmentation accuracy, and do not consider the properties of the labels (e.g., density and location). All the initial labels should thus be inliers, i.e., they are correctly assigned to the regions to segment, and carefully set to the optimal location for the best performance each method can achieve. Otherwise, the performance decreases drastically even by a small number of inaccurate labels.

2.2. Robust interactive image segmentation

There have been many attempts to handle the inaccurate initial labels. Unlike the approaches in the previous section, they regard the initial labels as soft constraints to label propagation (Bai & Wu, 2014; Jung & Kim, 2012; Rother et al., 2004; Subr et al., 2013; Unger et al., 2008). In Unger et al. (2008), the initial foreground and background labels are classified into three types of labels (foreground, background, and ambiguous ones) by thresholding, where the threshold value is determined using the distributions of all color values in the initial labels. Similar approach was adapted to the dense conditional random field (CRF) model, where the probability of becoming foreground and background regions is assigned to the initial labels (Subr et al., 2013). Bai and Wu used a labeling heuristics that the labels along the boundary of initial labels are more likely to be incorrect (Bai & Wu, 2014). These approaches can discriminate accurate and inaccurate labels, but need fine-tuned parameters (e.g., the threshold value in Subr et al. (2013); Unger et al. (2008)) for each image to achieve the best performance. Using a bounding box annotation to assign initial labels is the representative approach in interactive image segmentation (Rother et al., 2004). This assumes that the region inside box contains a mixture of foreground and background labels, while the entire region outside the box is background (Rother et al., 2004). Starting from the initial labels, the color distribution of each label is estimated iteratively, and the foreground object is then segmented out of the box. This approach gives satisfactory segmentation results even with the simple bounding box annotation, but the bounding box should contain an entire object to segment for estimating the initial color distribution of background labels correctly. The object-level information such as salient objects can be used to alleviate this problem (Jung & Kim, 2012). Recently, Zemene and Pelillo have presented a graph-based clustering on superpixels, which addresses robust segmentation by neglecting inaccurate foreground labels by considering background labels (Zemene & Pelillo, 2016). Most of current robust segmentation methods use the global color distributions of initial labels only. This is problematic when foreground and background have similar color distributions. In contrast, our approach considers both global and local distributions of color values for the labels in a unified framework, addressing this problem.

3. Proposed approach

3.1. Background and motivation

We review the framework of interactive image segmentation. Let us consider an undirected graph $G = (V, E)$ with a set of nodes V and edges E. Each node $v_i \in V$ identifies an image pixel i, and its neighboring nodes v_j are connected to an edge $e_{ij} \in E$. With an assumption that neighboring pixels tend to have a similar label when their color and spatial distance are small, an edge weight is computed using the Gaussian function as follows:

$$w_{ij} = \exp\left(-\frac{{c_{ij}}^2}{2\sigma_c^2} - \frac{{s_{ij}}^2}{2\sigma_s^2}\right), \tag{1}$$

where $c_{ij} = \left\|\mathbf{c}_i - \mathbf{c}_j\right\|$ and $s_{ij} = \left\|\mathbf{s}_i - \mathbf{s}_j\right\|$ are the distances between color values ($\mathbf{c}_i$ and $\mathbf{c}_j$) and spatial locations ($\mathbf{s}_i$ and $\mathbf{s}_j$), respectively. The effects of c_{ij} and s_{ij} are controlled by σ_c and σ_s, respectively.

Let us suppose k initial labels, i.e., l_k, are given. Interactive image segmentation can then be formulated as a classification problem, where all pixels in the input image are classified to one of given labels l_k. That is, for each initial label l_k, a classifying function x^k is estimated by minimizing the following objective function (Couprie et al., 2011; Sinop & Grady, 2007):

$$\mathcal{E}(x^k) = \lambda \sum_i z_i^k \left(x_i^k - g_i^k\right)^2 + \sum_i \sum_{j \in \mathcal{N}_i} w_{ij} \left(x_i^k - x_j^k\right)^2. \tag{2}$$

It consists of fidelity and regularization terms, balanced by the regularization parameter λ. $\mathcal{N}_i$ is the local neighborhood of the

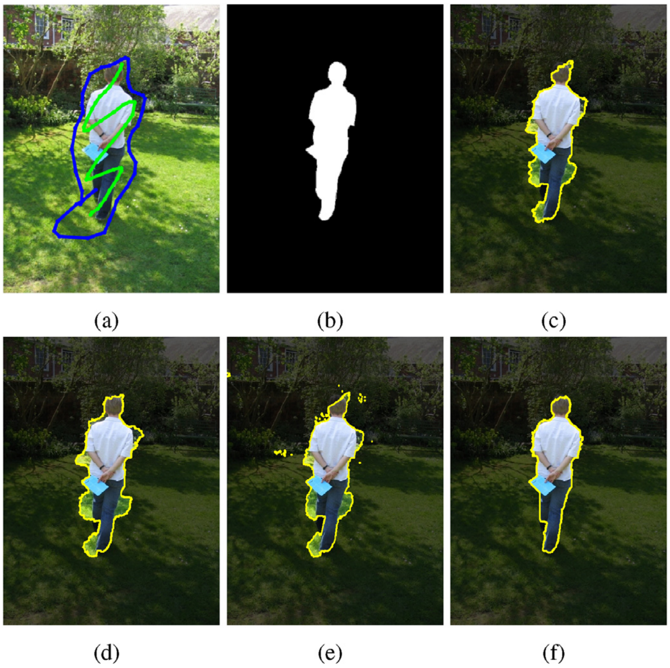


Fig. 1. Segmentation examples for inaccurate initial labels. (a) Initial labels, (b) ground truth, (c) RW (Grady, 2006), (d) RWR (Kim et al., 2008), (e) LC (Casaca et al., 2014), and (f) ours. Current segmentation methods regard the initial labels as correct ones, and thus they are easily distracted even by a small number of inaccurately assigned labels. Our approach addresses this problem, and gives a correct segmentation result.

pixel i. g_i^k and z_i^k denote an existence of the label l_k and its reliability, respectively, at the pixel i. They are set to 1 if i is assigned to the label l_k, and otherwise, 0. Using the classifying function x^k, the label S_i is assigned to the pixel i by

$$S_i = \arg\max_{l_k} x_i^k. \tag{3}$$

Conventionally, interactive segmentation methods have assumed that the initial labels are correct, i.e., both g_i^k and z_i^k are assigned to 1. This assumption often leads to serious problems if the initial labels include inaccurate ones, as shown in Fig. 1. Note that the methods in Fig. 1(c–e) use similar objective functions to ours in (2). This figure shows that current segmentation methods are easily distracted even by a small number of inaccurately assigned labels. In other words, they does not handle the inaccurate labels. To address this problem, we present a structure-aware labeling method that penalizes inaccurate initial labels (Fig. 1(f)).

3.2. Model

The overall process of the proposed method is shown in Fig. 2. Given initial labels (Fig. 2(a)), the OCP is estimated by computing both the global distribution of all color values within each label (occurrence statistics) and local distribution of color values around the label (co-occurrence statistics) (Fig. 2(b)) (Chang & Krumm, 1999; Lu, Tan, & Lim, 2014). Note that the OCP has been widely used in computer vision, since it gives the global probability distribution of the color value which is accompanied by contextual cues, e.g., spatial relations, to capture image structure effectively (Chang & Krumm, 1999; Lu et al., 2014). Our approach inherits this property, which enables estimating the reliability of each label effectively, and adaptively discriminating between accurate and inaccurate labels (Fig. 2(c)). Furthermore, we adopt a nonlocal regularization framework, which captures the pairwise connectivity more effectively than local regularization (Fig. 2(d)).

By comprising the OCP with a nonlocal pairwise connection, we minimize an objective function of the form:

$$\mathcal{E}(x^k) = \lambda \mathcal{D}\left(x^k\right) + \mathcal{S}\left(x^k\right), \tag{4}$$

where $\mathcal{D}\left(x^k\right)$ and $\mathcal{S}\left(x^k\right)$ are structure-aware fidelity and regularization terms, respectively, which will be described in the following sections.

3.2.1. Structure-aware fidelity

Without loss of generality, we assume that the number of inaccurately assigned labels is smaller than that of the accurate ones. Let $p_k(m_1, m_2)$ denote the OCP for initial labels of l_k, given an intensity range of [0, 255]. The OCP is defined as follows:

$$p_k(m_1, m_2) = \frac{h_k(m_1, m_2)}{\sum_{\bar{m}_1} \sum_{\bar{m}_2} h_k(\bar{m}_1, \bar{m}_2)}, \tag{5}$$

where $h_k(m_1, m_2)$ is an image co-occurrence histogram for kth initial labels defined as:

$$h_k(m_1, m_2) = \sum_i z_i^k \left(\delta(I_i - m_1) \cdot \sum_{j \in N_i} \delta(I_j - m_2) \right). \tag{6}$$

I_i denotes an intensity value at the pixel i, and $\delta(\cdot)$ represents a Dirac delta function. The image co-occurrence histogram $h_k(m_1, m_2)$ counts the number of pixels when the intensities of a reference pixel i and its neighboring pixels $j \in N_i$ are m_1 and m_2, respectively. This means that the diagonal part of the OCP $p_k(m_1, m_2)$ is the global color distribution of the kth initial label l_k, and each of its row captures the local color distribution around that label. Thus, the OCP encodes both global color distribution of initial labels and their local color distribution, simultaneously.

We define the reliability of the label l_k as the aggregation of the OCPs in a local neighborhood B_i, as follows[1]:

$$r_i^k = \sum_{j \in B_i} p_k\left(I_i, I_j\right). \tag{7}$$

The reliability r_i^k effectively captures structural information around the label l_k. In other words, inaccurate labels have lower reliabilities than accurate ones, which suppresses the inaccurate labels and alleviates propagation errors. The OCP is computed around the initial labels.

Finally, we propose the structure-aware data cost for each label l_k:

$$\mathcal{D}\left(x^k\right) = \sum_i z_i^k \left(x_i^k - r_i^k g_i^k\right)^2. \tag{8}$$

It captures the confidence of the labels using both global and local color distributions of them.

3.2.2. Nonlocal regularization

To capture the relationship between neighbors effectively, we adopted a nonlocal pairwise connection by computing k-nearest neighbors (k-NN) in the feature space (Chen et al., 2013; Muja & Lowe, 2009). Given a reference pixel i, a five-dimensional feature vector $\mathbf{f}_i = \left(I_i^L, I_i^a, I_i^b, \eta s_i^x, \eta s_i^y\right)$ is used. The parameter η controls the influence between *Lab* color components $\mathbf{c}_i = \left(I_i^L, I_i^a, I_i^b\right)$ and spatial coordinates $\mathbf{s}_i = \left(s_i^x, s_i^y\right)$ in searching k-NN, i.e., spatially closer neighbors are clustered as η increases. The nonlocal regularization term is then constructed as follows:

$$\mathcal{S}(x) = \sum_i \sum_{j \in \mathcal{K}_i} w_{ij} \left(x_i^k - x_j^k\right)^2, \tag{9}$$

[1] For a color image, a *Lab* color space is used where each channel is normalized to [0, 255], and the reliability r_i^k for each color channel is averaged.

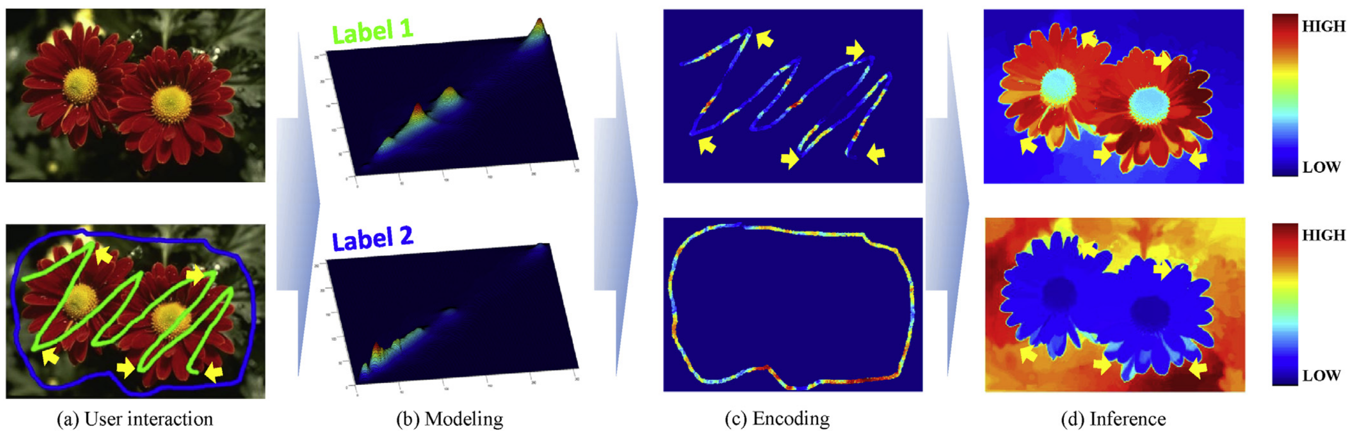


Fig. 2. Overview of the proposed method. (a) For an input image, initial labels are allocated. (b) The OCP is computed for each label to capture the global distribution of color values within each label and the local distribution of color values around the label, that captures the spatial relationship of the color values. (c) The reliability of each label is encoded using the OCP. (d) Finally, we optimize each classifying function while suppressing the effect of inaccurate initial labels. As highlighted with yellow arrows in (c) and (d), inaccurate initial labels are penalized, suppressing the propagation of inaccurate label information during optimization. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

where $\mathcal{K}_i$ represents a nonlocal neighborhood for the pixel i. The weight between $\mathbf{f}_i$ and $\mathbf{f}_j$, w_{ij}, is computed similar to (1). While the local smoothness constraint in (2) simply considers a spatially similar neighborhood, the nonlocal neighborhood in (9) connects all pixels that might have similar values to x_i^k. The k-NN connectivity thus propagates initial labels effectively to their neighbors having similar appearance.

3.3. *Solver*

The objective function in (4) can be represented as:

$$E(X_k) = (X_k - R_k G_k)^T \Lambda (X_k - R_k G_k) + X_k{}^T L X_k, \tag{10}$$

where X_k is a vector formed by concatenating all x_i^k. Λ and R_k are diagonal matrices where $[\Lambda]_{ii} = \lambda$ and $[R_k]_{ii} = r_i^k$. G_k is a vector concatenating the initial labels g_i^k:

$$[G_k]_i = \begin{cases} g_i^k & z_i^k = 1 \\ 0 & \text{otherwise} \end{cases}. \tag{11}$$

The similarity between nonlocal neighbors is represented by the Laplacian matrix, $L = D - W$. D is a diagonal degree matrix where each diagonal entry $d_{ii} = \sum_j w_{ij}$, and W is a weight matrix defined as:

$$W_{ij} = \begin{cases} w_{ij} & e_{ij} \in E \\ 0 & \text{otherwise} \end{cases}. \tag{12}$$

Since (10) is quadratic w.r.t to X_k, a closed-form solution can be obtained by solving the following linear equation:

$$X_k = (L + \Lambda)^{-1} \Lambda R_k G_k. \tag{13}$$

Note that nonlocal regularization uses only a few pixels for the weight computation, and thus $(L + \Lambda)$ is sparse. This enables solving our linear equation efficiently by sparse solvers (Krishnan, Fattal, & Szeliski, 2013). Finally, each label is assigned to all pixels as in (3). We summarize the proposed method in Algorithm 1.

3.4. *Comparison to other methods*

Here, we show the relationship between our model and conventional ones: RW (Grady, 2006), RWR (Ham, Min, & Sohn, 2013; Kim et al., 2008), LC (Casaca et al., 2014), accurate binary selection (ABS) (Subr et al., 2013), and ratio cut (RC) (Bai & Wu, 2014). Our model formulates interactive segmentation as a continuous optimization problem, which is closely related to RW, RWR, and LC.

Algorithm 1 Interactive image segmentation using OCP.

Require: Original image I and initial labels l_k
1: Compute the OCP $p_k(m_1, m_2)$ according to (5).
2: Encode $p_k(m_1, m_2)$ to the reliability of initial labels in (7).
3: Construct the weight matrix W_{ij} according to (12).
4: Compute X_k as in (13).
5: Select the segmentation label of each pixel S_i by (3).

They solve similar linear equations to (13). Unlike these methods, ABS and RC solve the segmentation problem in a discrete domain. Table 1 summarizes the properties of current segmentation methods.

3.4.1. *Continuous optimization*

The energy functions of RW ($\mathcal{E}^{RW}$), RWR ($\mathcal{E}^{RWR}$) and LC ($\mathcal{E}^{LC}$) can be represented as follows:

$$\mathcal{E}_k^{RW}\left(x_i^k\right) = \sum_i z_i^k \left(x_i^k - g_i^k\right)^2 + \lambda \sum_i \sum_{j \in \mathcal{N}_i} w_{ij} \left(x_i^k - x_j^k\right)^2, \tag{14}$$

$$\mathcal{E}_k^{RWR}\left(x_i^k\right) = \sum_i z_i^k \left(x_i^k - g_i^k\right)^2 + \lambda \sum_i \sum_{j \in \mathcal{N}_i} \bar{w}_{ij} \left(x_i^k - x_j^k\right)^2, \tag{15}$$

$$\mathcal{E}_k^{LC}\left(x_i^k\right) = \sum_i z_i^k \left(x_i^k - g_i^k\right)^2 + \lambda \sum_i \left(\sum_{j \in \mathcal{N}_i} w_{ij} \left(x_i^k - x_j^k\right) \right)^2, \tag{16}$$

where $\bar{w}_{ij} = w_{ij} / \sum_{j \in \mathcal{N}_i} w_{ij}$. As shown in (14)–(16), RW, RWR, and LC are associated with each other. For example, for unconstrained pixels, i.e. $z_i^k = 0$, label information is propagated to neighboring pixels $\mathcal{N}_i$ as follows:

$$x_i^k = \sum\nolimits_{j \in \mathcal{N}_i} \bar{w}_{ij} x_j^k. \tag{17}$$

That is, the initial labels in these methods are propagated in the same manner. The RW optimizes the energy function using initial labels as hard constraints, i.e., $z_i^k = \infty$ for initial labels. The segmentation result thus depends heavily on the location of initial labels. In contrast to RW, RWR and LC alleviate this problem by setting $z_i = 1$, but still regard all the initial labels as correct ones. Accordingly, these methods do not allow inaccurate initial labels for segmentation. On the contrary, our model considers the reliability of initial labels, effectively suppressing inaccurate labels and giving robust segmentation results.

Table 1
Comparison of current interactive segmentation methods and their properties.

Method	Energy function	Optimization	Robustness	Multi-label
RW (Grady, 2006)	$\lambda \sum_i z_i^k \left(x_i^k - g_i^k\right)^2 + \sum_i \sum_{j\in\mathcal{N}_i} w_{ij}\left(x_i^k - x_j^k\right)^2$	Numerical solver	No	Yes
RWR (Kim et al., 2008)	$\lambda \sum_i z_i^k \left(x_i^k - g_i^k\right)^2 + \sum_i \sum_{j\in\mathcal{N}_i} \bar{w}_{ij}\left(x_i^k - x_j^k\right)^2$	Numerical solver	No	Yes
LC (Casaca et al., 2014)	$\lambda \sum_i z_i^k \left(x_i^k - g_i^k\right)^2 + \sum_i \left(\sum_{j\in\mathcal{N}_i} w_{ij}\left(x_i^k - x_j^k\right)\right)^2$	Numerical solver	No	Yes
ABS (Subr et al., 2013)	$\lambda \sum_i \psi_{ABS}(\theta_i) + \sum_{i,j} \left(\mu\left(\theta_i, \theta_j\right) \sum_m k^{(m)}\left(\mathbf{f}_i, \mathbf{f}_j\right)\right)$	Discrete optimization	Yes	No
RC (Bai & Wu, 2014)	$\lambda \sum_i \left(\psi_{GMM}(\theta_i) - \alpha U(\theta_i)\right) + \sum_i \sum_{j\in\mathcal{N}_i} w_{ij}\left\|\theta_i - \theta_j\right\|$	Discrete optimization	Yes	No
Ours	$\lambda \sum_i z_i^k \left(x_i^k - r_i^k g_i^k\right)^2 + \sum_i \sum_{j\in\mathcal{K}_i} w_{ij}\left(x_i^k - x_j^k\right)^2$	Numerical solver	Yes	Yes

3.4.2. Discrete optimization

The ABS and RC formulate image segmentation as a discrete optimization problem. For discrete variables $\theta_i \in \theta$ where binary labels (foreground, background) are available, the ABS optimizes the following energy function:

$$\mathcal{E}^{ABS}(\theta) = \sum_i \psi_{ABS}(\theta_i) + \lambda \sum_{i,j} \left(\mu\left(\theta_i, \theta_j\right) \sum_m k^{(m)}\left(\mathbf{f}_i, \mathbf{f}_j\right)\right), \quad (18)$$

where $\psi_{ABS}(\theta_i)$ is the data term employed in Subr et al. (2013). In $\psi_{ABS}(\theta_i)$, the ABS heuristically penalizes initial probabilities for foreground/background labels according to the initial labels. $\mu(\theta_i, \theta_j)$ and $k^{(m)}(\mathbf{f}_i, \mathbf{f}_j)$ denote a simple Potts model and a kernel similarity function, respectively. Details are described in Subr et al. (2013).

In RC, the solution is estimated by optimizing θ and the ratio α iteratively as follows:

$$\mathcal{E}^{RC}(\theta) = \sum_i \left(\psi_{GMM}(\theta_i) - \alpha U(\theta_i)\right) + \lambda \sum_i \sum_{j\in\mathcal{N}_i} w_{ij}\left|\theta_i - \theta_j\right|, \quad (19)$$

where ψ_{GMM} is the data term using a Gaussian mixture model (GMM) (Rother et al., 2004), and $U(\theta)$ is a utility function that penalizes inaccurate labels by assuming that centerline of labels are more likely to be correct than the outer ones.

As shown in the above, current robust segmentation methods (ABS and RC) consider the global color distributions of initial labels only, or exploit a heuristic strategy to handle inaccurate labels. In contrast, the proposed method captures structural information underlying the initial labels.

4. Experiments

4.1. Experimental details

The experiments were performed on a PC with Intel Core i7 Quad processor (3.40GHz). We used the Grabcut dataset (Rother et al., 2004) with 50 images from Berkeley dataset (Martin, Fowlkes, Tal, & Malik, 2001) for binary segmentation, and a subset of 50 images from the Graz dataset (Santner, Pock, & Bischof, 2010) for multi-label segmentation. We compared our approach to the state of the art: RW (Grady, 2006), RWR (Kim et al., 2008), LC (Casaca et al., 2014), ABS (Subr et al., 2013), and RC (Bai & Wu, 2014). All the experiments were performed by the public codes provided by the authors with default parameter settings. In the proposed method, we set the bandwidth parameters, σ_r and σ_s, to 0.2. The regularization parameter λ and k-NN parameter η are fixed to 0.0005 and 1.8, respectively. We empirically set 7×7 size of square windows for the neighborhoods N_i and B_i. The size of the neighborhood $\mathcal{K}_i$ is set to 10 and 20 in the Grabcut and Graz datasets, respectively. For a quantitative evaluation, we measure the Dice score (Dice, 1945; Santner et al., 2010) between segmentation results S^k and ground truths GT^k for each label l_k as follows:

$$Dice(S^k, GT^k) = \frac{2\left|S^k \cap GT^k\right|}{\left|S^k\right| + \left|GT^k\right|}, \quad (20)$$

where $|S^k \cap GT^k|$ indicates the overlapping area between the regions S^k and GT^k. We report an average Dice score over all labels.

4.1.1. Initial labeling

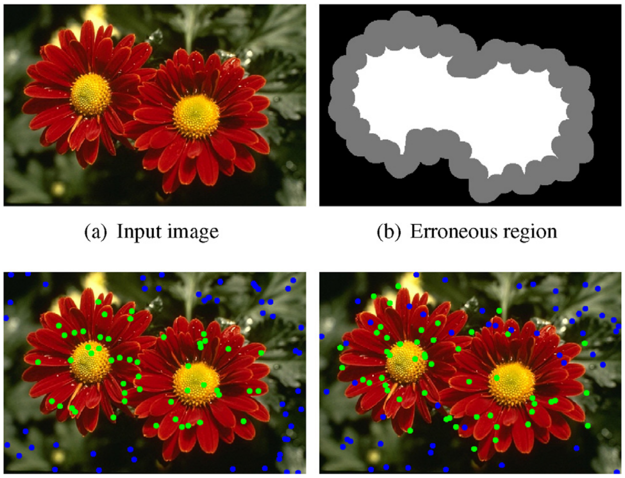
We generate both manual and synthetic labels as initial labels for the experiments. For the manual labeling, a user roughly draw initial labels to the Grabcut dataset (Rother et al., 2004) and the Graz dataset (Santner et al., 2010). Each image is sparsely labeled into two or three regions colored with red, blue, and green.

For the synthetic labeling, following the experimental protocol in Subr et al. (2013) and (Bai & Wu, 2014), we synthesize initial labels with the ground truth, and evaluate segmentation accuracy by varying the amount of incorrect initial labels: Using the ground-truth labels, we split an input image to foreground, background, and erroneous regions (Fig. 3(a) and (b))). The erroneous regions are generated near object boundaries by dilation and erosion with the radius of 15 pixels. Except for the experiment with varying label density (Section 4.2.2), we randomly select 50 points in each of foreground and background, dilate these points by the radius of 5 pixels, and use them as initial labels. We control the number of incorrect initial labels in foreground and background from 0% to 40% of total number of initial labels, respectively, in such a way that some of foreground (background) labels are set to the background (foreground) ones in the erroneous region (Fig. 3(c) and (d)). Here, we use the Grabcut dataset (Rother et al., 2004).

4.2. Performance analysis

4.2.1. Effectiveness of the OCP

We compare the segmentation results obtained by our model, but with different data terms (with and without the reliability term) in Fig. 4. The reliability of initial labels is estimated using the OCP that encodes global and local color distributions underlying the labels, and this figure verifies the effectiveness of using the OCP in interactive image segmentation. Since the OCP penalizes inaccurate initial labels, the inferred probability of foreground is highly discriminative compared to the one estimated without using the OCP (Fig. 4(c) and (e)). Accordingly, the OCP enables obtaining more robust segmentation results to the outliers (Fig. 4(d)



(a) Input image (b) Erroneous region

(c) Initial labels without error (d) Initial labels with 30% error

Fig. 3. Synthetic label generation: Using ground-truth labels, we split (a) an input image into (b) foreground (white), background (black), and erroneous (gray) regions. We randomly generate same amount of pixels in each of foreground and background, and use them as initial labels. (c-d) The error level of initial labels is controlled by switching some of foreground labels in the erroneous regions to the background ones, and *vice versa*. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Average and standard deviation (in brackets) of Dice scores on the Grabcut dataset (Rother et al., 2004) with different data terms and changing error rate. Higher Dice scores are better.

	Error rate(%)				
	0	10	20	30	40
Ours w/o OCP	0.9205 (0.141)	0.9043 (0.147)	0.8292 (0.122)	0.8237 (0.129)	0.8181 (0.135)
Ours w/ OCP	0.9308 (0.070)	0.9256 (0.084)	0.9213 (0.087)	0.9152 (0.105)	0.8990 (0.090)

and (f)). This also alleviates a small-cut problem that usually occurs to nonlocal regularization. Table 2 compares the average Dice scores on the Grabcut dataset (Rother et al., 2004) by changing the error rate in initial labels while fixing the total number of initial labels. As expected, using the OCP gives a better quantitative result, and the results without the OCP drastically decrease when the inaccurate labels are included.

For further analysis, we present a precision-recall (PR) curve of each case. Here, the PR curve reflects the performance of foreground appearance in precision and recall by thresholding the final foreground appearance map with different values (ranging from 0 to 255). As shown in Fig. 5, the results with the OCP commonly perform well with higher precision in the case of fixed recall than the ones without the OCP. Specifically, noting that high recall is important to the performance of image segmentation, the precision without the OCP becomes lower than the one with the OCP when the recall is high.

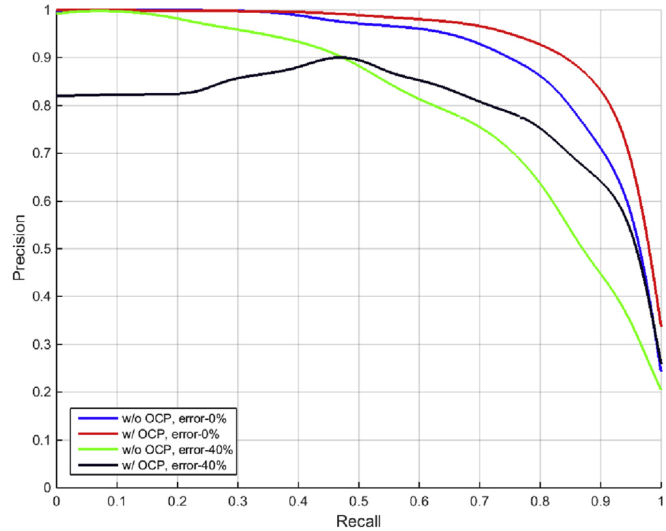


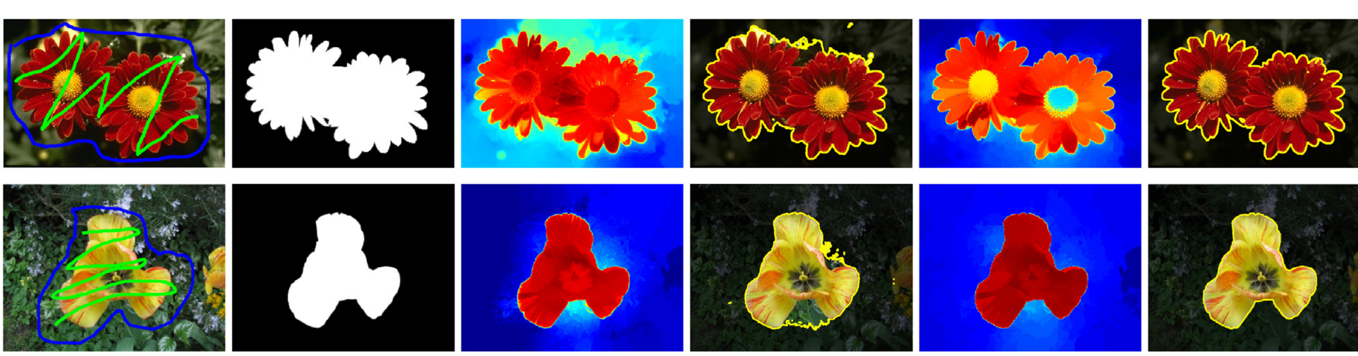
Fig. 5. PR curves of the foreground map with and without the OCP. Overall, the results with the OCP outperform the ones without the OCP.

4.2.2. Experiments with varying label density

We evaluate the performance difference of the proposed method by changing the density of initial labels. In other words, we changed the total number of random points to 20, 40, 60, 80, and 100, containing 40% of inaccurate labels. The PR curves and the Dice scores of the experiments are presented in Fig. 6 and Table 3, respectively. It shows that the proposed method generally performs well since the initial labels are propagated to longer distance with non-local neighborhood, and inaccurate initial labels are suppressed thanks to the OCP.

4.2.3. Experiments with varying neighborhood size

Fig. 7 shows a qualitative comparison for different sizes of neighborhood $\mathcal{K}_i$. This shows that better segmentation results can



(a) (b) (c) (d) (e) (f)

Fig. 4. Influence of the reliability term using the OCP on segmentation results. (a) Initial labels, (b) ground truths, (c) and (e) the inferred probabilities of foreground (red: high, blue: low) without and with the reliability term, respectively, and (d) and (f) corresponding segmentation results. The probability of foreground in (e) is highly discriminative compared to the one in (c), and this gives more accurate segmentation results. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

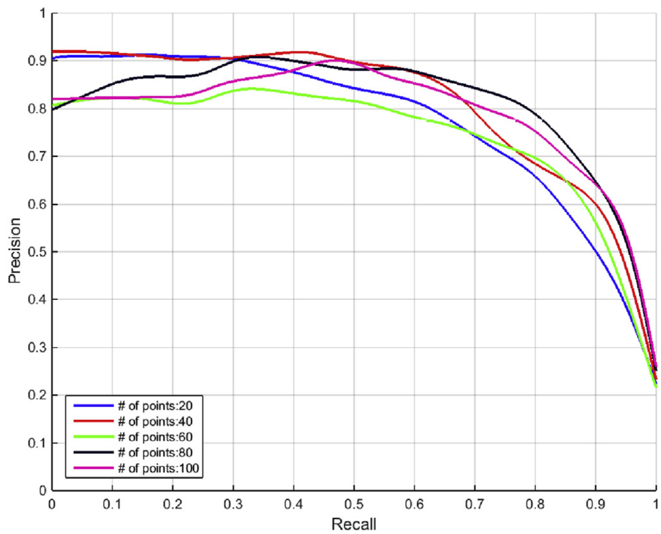


Fig. 6. PR curves of the foreground map by varying the density of initial labels. All cases show competitive results, demonstrating robustness to label density.

Table 3
Average and standard deviation (in brackets) of Dice scores on the Grabcut dataset (Rother et al., 2004) by changing the number of initial labels containing 40% errors. Higher Dice scores are better.

# Points	20	40	60	80	100
Dice Score	0.8858 (0.110)	0.8897 (0.112)	0.8931 (0.103)	0.8903 (0.090)	0.8990 (0.085)

generally be obtained with a large neighborhood size. Regularization using a small size of neighborhood does not capture image structure well, and causes a small-cut problem.

4.2.4. Computational time

We measured the runtime of the proposed method in Algorithm 1. Note that the source code is implemented in MATLAB and tested with a single CPU without hardware acceleration. For an input image (481 × 321), the main overheads of computation can be divided into four categories: modeling (Algorithm 1: line 1), encoding (Algorithm 1: line 2), weight computation (Algorithm 1: line 3), and optimization (Algorithm 1: line 4). Table 4 summarizes the computation time (in seconds) of the proposed method with varying dominant parameters. We ran the algorithm fifty times and measured the average and standard deviation values of runtime. It shows that the OCP estimation depends on the density of initial labels, and the weight construction and the optimization are affected by the size of $\mathcal{K}_i$. Note that, for practical implementation, the proposed method can be further accelerated by existing strategies, such as histogram quantization (Cheng, Mitra, Huang, Torr, & Hu, 2015), label sampling (Bie, Huang, & Wang, 2011), and fast solver (Min et al., 2014).

Table 4
Average and standard deviation (in brackets) of running time (in seconds) of the proposed method.

(# Points, $\lvert\mathcal{K}_i\rvert$)	Modeling	Encoding	Weight	Optimization	Total
(20, 10)	21.0 (0.108)	2.49 (0.273)	2.10 (0.051)	22.4 (0.301)	49.5 (0.737)
(20, 15)	21.1 (0.112)	2.53 (0.271)	3.32 (0.081)	27.9 (0.375)	57.5 (0.841)
(100, 10)	20.5 (0.111)	9.87 (0.475)	2.11 (0.055)	22.7 (0.305)	58.8 (0.944)
(100, 15)	20.1 (0.113)	9.85 (0.471)	3.01 (0.086)	28.8 (0.365)	64.3 (1.030)

4.3. Comparison with existing methods via synthetic labels

A quantitative comparison on the Grabcut dataset (Rother et al., 2004) with the state of the art is shown in Fig. 8. Most methods except ABS (Subr et al., 2013) give similar Dice scores with accurate initial labels, but the performance decreases as the number of incorrect labels increases. On the contrary, our approach is robust to the inaccurate labels, and maintains a nearly similar performance over all error levels. Fig. 9 shows examples of segmentation results generated with 20% of inaccurate labels in (a). The RW (Grady, 2006), RWR (Kim et al., 2008), and LC (Casaca et al., 2014) assume that all initial labels are correct, and thus they give incorrect segmentation results even with a small portion of inaccurate labels (Fig. 9(c–e)). The LC (Casaca et al., 2014) using nonlocal regularization propagates label information to longer distance than using local regularization as in RW (Grady, 2006) and RWR (Kim et al., 2008). This may help to improve segmentation accuracy, but with inaccurate labels the accuracy rather decreases. Although ABS (Subr et al., 2013) discriminates between accurate and inaccurate labels, a fully-connected CRF model in Subr et al. (2013) causes a small-cut problem (Fig. 9(f)). The RC (Bai & Wu, 2014) shows a

Fig. 7. (From left to right) initial labels and segmentation results for different sizes of neighborhood $\mathcal{K}_i$. The size of neighborhood is set to 2, 4, 7, 10, and 15, respectively. As the neighborhood size increases, a better segmentation accuracy can be achieved without a small-cut problem.

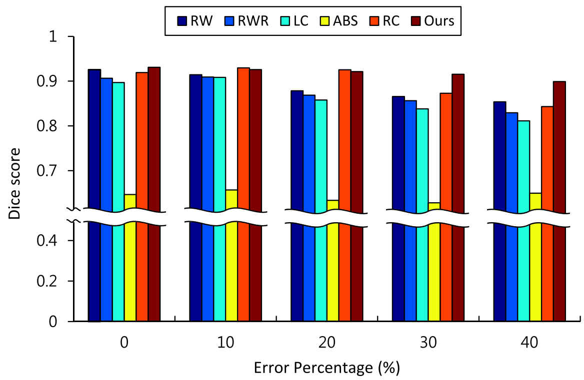
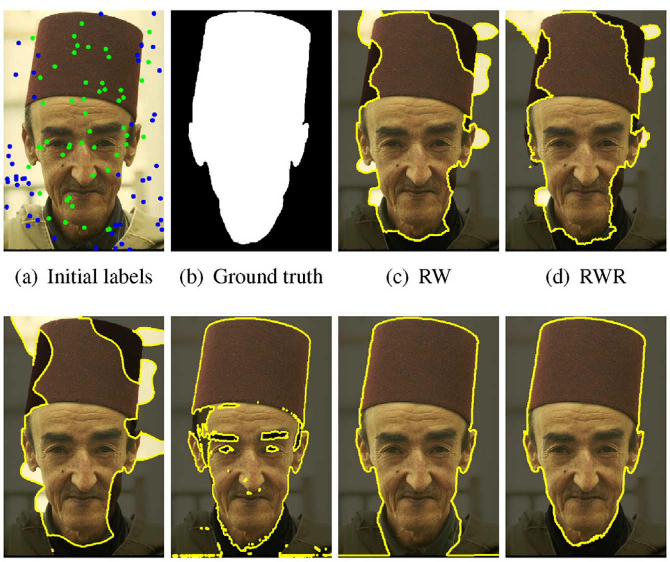


Fig. 8. Comparison of average Dice scores on the Grabcut dataset (Rother et al., 2004) with varying the error level of initial labels. Our approach shows robustness to inaccurate initial labels, and outperforms the state of the art, especially for the high error level of initial labels.



(a) Initial labels (b) Ground truth (c) RW (d) RWR

(e) LC (f) ABS (g) RC (h) Ours

Fig. 9. Segmentation results for synthetic labels. (a) initial labels with 20% error, (b) ground truth, (c) RW (Grady, 2006), (d) RWR (Kim et al., 2008), (e) LC (Casaca et al., 2014), (f) ABS (Subr et al., 2013), (g) RC (Bai & Wu, 2014), and (h) ours. Our approach substantially outperforms other methods.

Table 5
Average and standard deviation (in brackets) of Dice scores for binary segmentation on the Grabcut dataset (Rother et al., 2004).

Method	Dice score
RW	0.9015 (0.053)
RWR	0.9057 (0.055)
LC	0.8839 (0.050)
ABS	0.7155 (0.210)
RC	0.9094 (0.090)
Ours	**0.9207 (0.081)**

Table 6
Average and standard deviation (in brackets) of Dice scores for multi-label segmentation on the Graz dataset (Santner et al., 2010).

Method	Dice score
RW	0.8720 (0.080)
RWR	0.8851 (0.083)
LC	0.8650 (0.079)
ABS	N.A.
RC	N.A.
Ours	**0.9011 (0.098)**

comparable performance to ours with a small number of inaccurate labels (Fig. 9(g)). The energy function in this method penalizes the boundaries of initial labels, and rejects many outliers iteratively. This is problematic when a whole part of the labels is assigned inaccurately, and unexpected segmentation results may be obtained if the solution in the previous step is incorrect. Unlike these methods, our approach using the OCP considers global and local color distribution underlying the initial labels, and effectively suppresses inaccurate labels, giving robust segmentation results (Fig. 9(h)).

4.4. Comparison with existing methods via manual labels

In this section, we show segmentation results obtained from manually assigned initial labels. Table 5 compares the average Dice scores for binary segmentation on the Grabcut dataset (Rother et al., 2004). This verifies once more that, similar to the results in Fig. 8, RW (Grady, 2006), RWR (Kim et al., 2008), and RC (Bai & Wu, 2014) show a better performance than LC (Casaca et al., 2014), and ABS (Subr et al., 2013) gives the worst results. This table also shows that our approach outperforms all methods. We compare the average Dice scores for multi-label segmentation on the Grabcut dataset (Rother et al., 2004) in Table 6. Note that ABS and RC cannot deal with a multi-label segmentation problem.

Figs. 10 and 11 show qualitative comparisons for binary and multi-label segmentations, respectively. We manually assign initial labels in such a way that a part of regions to segment contains inaccurate labels. The RW (Grady, 2006) leverages the first arrival probability that a random walker starting at each pixel arrives at the initial labels. This means that the random walker starting at the initial labels does not move. Namely, this method uses the initial labels as a hard constraint on segments, and assume that these labels are correctly assigned. The segmentation results are thus not correct especially around the inaccurate labels (Figs. 10(b) and 11(b)). The RWR (Kim et al., 2008) alleviates this problem using a global color distribution of the initial labels, but this does not differentiate the inaccurate labels from the initial ones explicitly (Figs. 10(c) and 11(c)). The LC (Casaca et al., 2014) is more sensitive to the outliers than RW and RWR (Figs. 10(d) and 11(d)). Unlike these methods, the ABS (Subr et al., 2013) is robust to outliers, but suffers from a small-cut problem (Fig. 10(e)). The RC (Bai & Wu, 2014) gives robust segmentation results. However, this method can handle binary labels only, and may give unexpected results (Fig. 10(f)). Our approach penalizes the inaccurate labels successfully, and gives robust segmentation results to the outliers (Figs. 10(g) and 11(e)). In addition, the proposed method outperforms in dividing fine boundaries. Thanks to the nonlocal pairwise connection, the underlying image structure is captured effectively in optimizing the problem.

From the intensive experiments, when inaccurate initial labels are provided, the proposed method shows outstanding results both qualitatively and quantitatively. In addition, it is worth noting that the proposed method also performs competitive results even correct initial labels are provided, which indicates that the structure-aware model provides more than a robustness to initial labels.

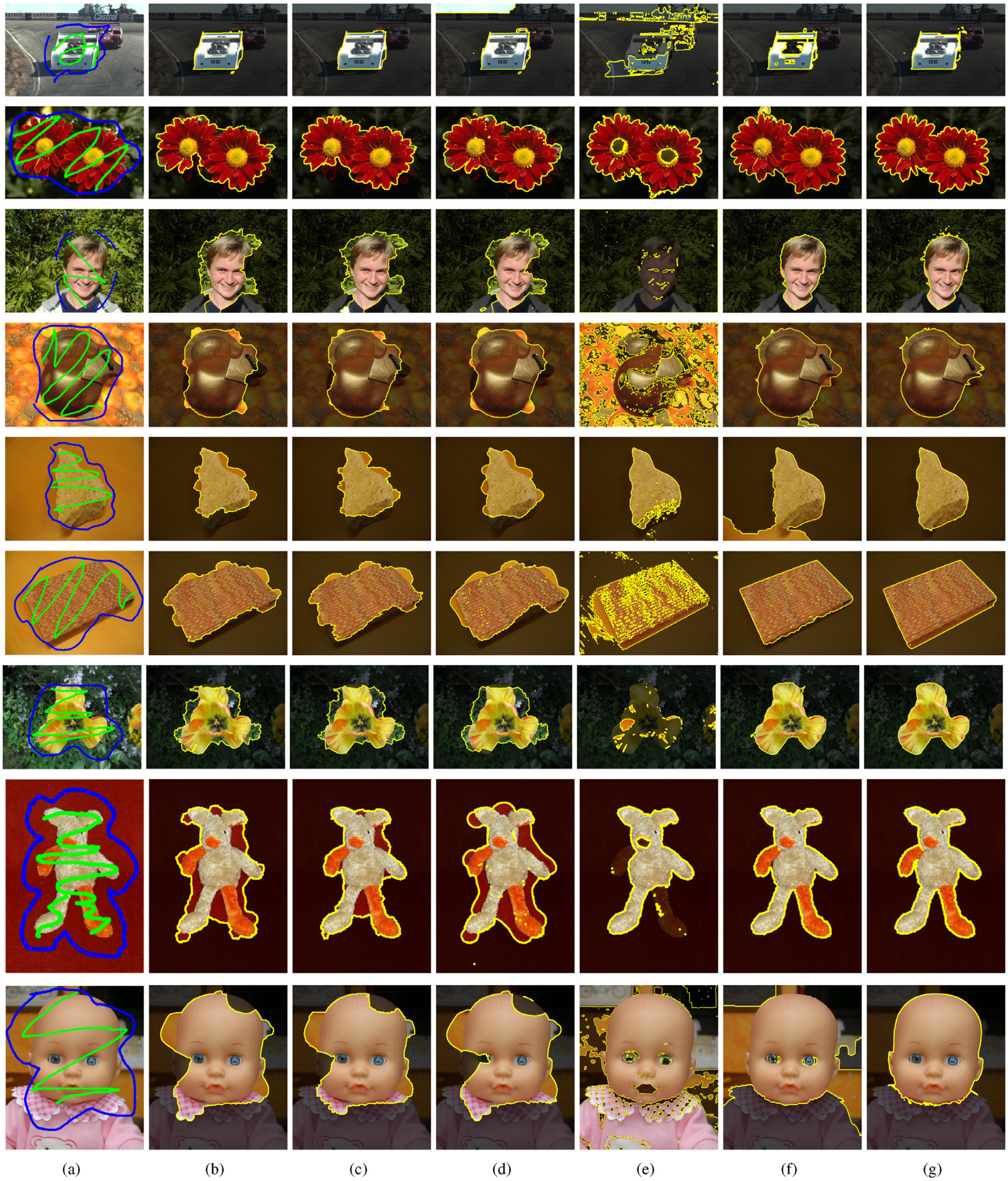


Fig. 10. Binary segmentation with manual labeling on the Grabcut dataset (Rother et al., 2004). (a) Input images with initial labels, (b) RW (Grady, 2006), (c) RWR (Kim et al., 2008), (d) LC (Casaca et al., 2014), (e) ABS (Subr et al., 2013), (f) RC (Bai & Wu, 2014), and (g) ours. Our method gives robust segmentation results to the inaccurate initial labels with alleviating a small-cut problem.

4.5. Limitations

Our approach shares the same limitation as current segmentation methods: Since most of the existing methods use the color distribution of initial labels, they cannot handle the case where all the initial labels have similar color distributions (e.g., Fig. 12(a)). In addition, our method penalizes inaccurate labels using the OCP of color values for each initial label. Thus, when a small number of correct labels are assigned in distinct regions (e.g., left-hand side in Fig. 12(b)), the OCP considers these labels as outliers, and sup-

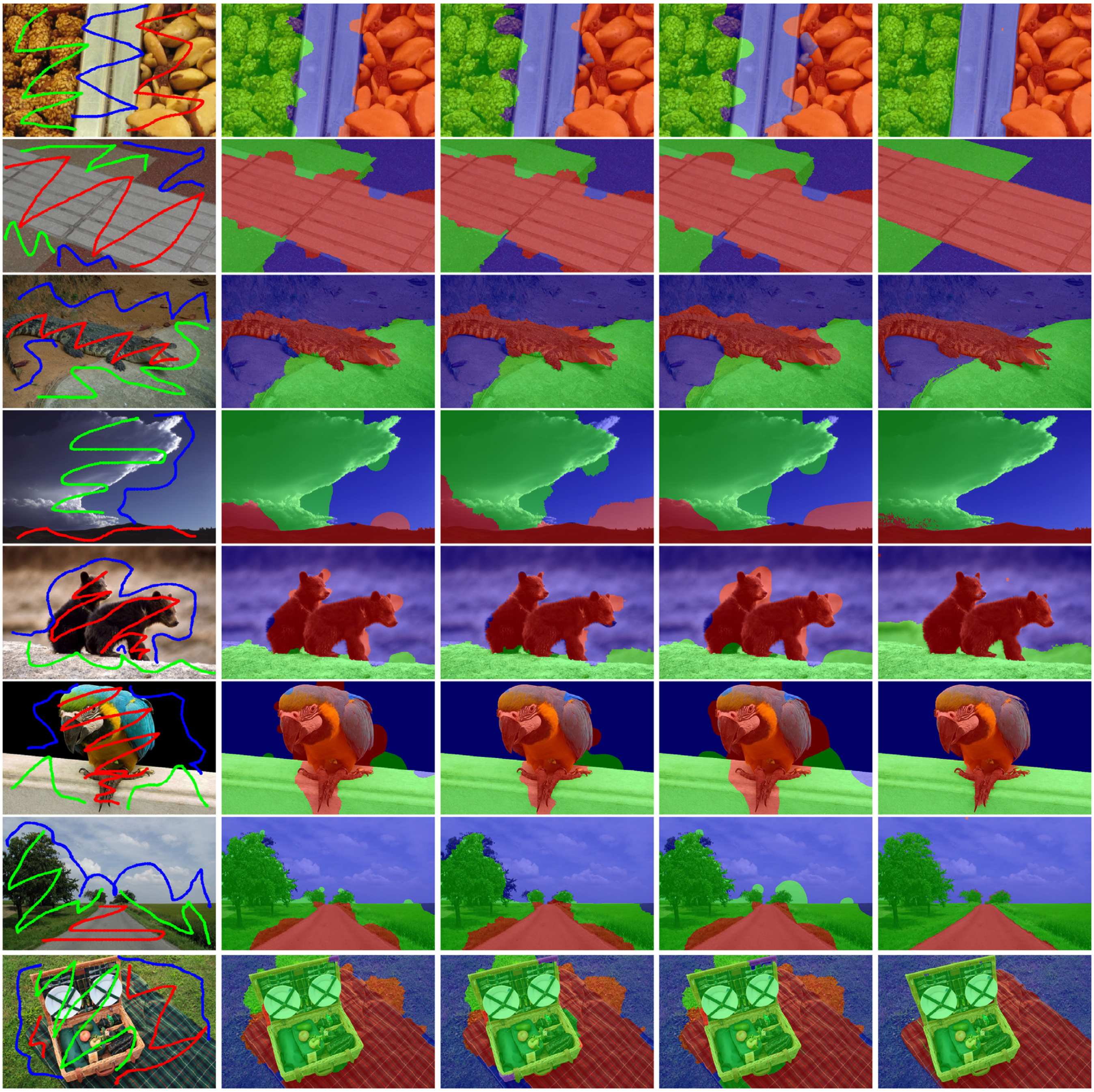


Fig. 11. Multi-label segmentation with manual labeling on the Graz dataset (Santner et al., 2010). (a) Input images with initial labels, (b) RW (Grady, 2006), (c) RWR (Kim et al., 2008), (d) LC (Casaca et al., 2014), and (e) ours. Our method is much more robust to the inaccurate initial labels than the state of the art.

presses them. These limitations can be addressed by employing other features as a description of initial labels, e.g., texture and object prior.

5. Conclusion and future work

We have presented a structure-aware labeling method for interactive segmentation. Interactive segmentation is formulated as a convex optimization problem. The OCP in our model considers global and local distributions of color values in initial labels. Contrary to the existing methods, our model is robust to inaccurate initial labels since the local color distribution captures the spatial relationship of these color value. The experimental results have demonstrated that our approach alleviates the influence of inaccurate initial labels effectively, and outperforms the state of the art for both binary and multi-label segmentation problems.

For future work, we can extend our model to address the existing limitations by integrating the proposed method in the framework of deep learning. It enables using highly reliable object prior for interactive image segmentation. In addition, the underlying theme of the proposed method can be applied to solve other semi-supervised computer vision and graphics problems that need to handle inaccurate initial supervision, such as interactive video segmentation, image manipulation, and 2D-to-3D conver-

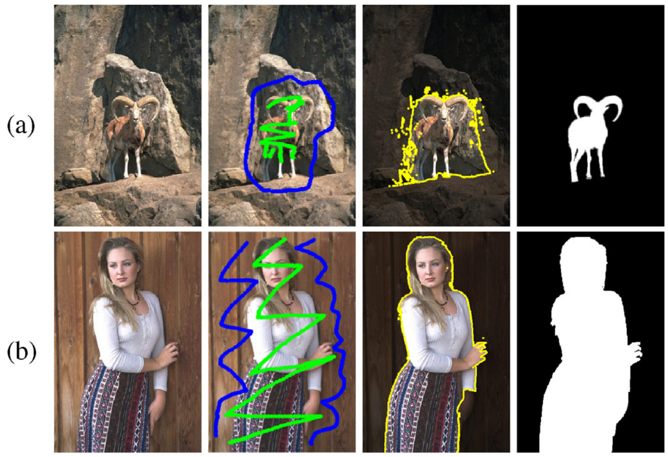


Fig. 12. Difficult examples. (From left to right) inputs, initial labels, segmentation results, and ground truths. Our method gives incorrect segmentations when (a) initial labels have similar color distributions and (b) a small number of correct labels are assigned in a distinct region (e.g., left-hand side).

sion. Similarly, it can be applied as a post-processing to the salient object detection which commonly generates coarse outputs especially in boundary region. Finally, learning features in convolutional neural networks (CNN) can be a promising future work (Bearman, Russakovsky, Ferrari, & Fei-Fei, 2016; Lin, Dai, Jia, He, & Sun, 2016). Unlike existing works use dense labels which require labor-intensive tasks to generate, we can train CNN with spare user labels.

Acknowledgment

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. R0115-16-1007).

References

Arbeláez, P., Pont-Tuset, J., Barron, J. T., Marques, F., & Malik, J. (2014). Multiscale combinatorial grouping. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 328–335).

Bai, J., & Wu, X. (2014). Error-tolerant scribbles based interactive image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 392–399).

Bearman, A., Russakovsky, O., Ferrari, V., & Fei-Fei, L. (2016). Whats the point: Semantic segmentation with point supervision. In *European conference on computer vision* (pp. 549–565).

Bie, X., Huang, H., & Wang, W. (2011). Real time edit propagation by efficient sampling. In *Computer graphics forum: 30* (pp. 2041–2048).

Boykov, Y. Y., & Jolly, M.-P. (2001). Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *Proceedings of the IEEE international conference on computer vision: 1* (pp. 105–112).

Casaca, W., Nonato, L. G., & Taubin, G. (2014). Laplacian coordinates for seeded image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 384–391).

Chang, P., & Krumm, J. (1999). Object recognition with color cooccurrence histograms. In *Proceedings of the IEEE conference on computer vision and pattern recognition: 2*.

Chen, Q., Li, D., & Tang, C.-K. (2013). Knn matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(9), 2175–2188.

Cheng, M.-M., Mitra, N. J., Huang, X., Torr, P. H., & Hu, S.-M. (2015). Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(3), 569–582.

Couprie, C., Grady, L., Najman, L., & Talbot, H. (2011). Power watershed: A unifying graph-based optimization framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(7), 1384–1399.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology, 26*(3), 297–302.

Dong, X., Shen, J., Shao, L., & Van Gool, L. (2016). Sub-markov random walk for image segmentation. *IEEE Transactions on Image Processing, 25*(2), 516–527.

Grady, L. (2006). Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(11), 1768–1783.

Ham, B., Min, D., & Sohn, K. (2013). A generalized random walk with restart and its application in depth up-sampling and interactive segmentation. *IEEE Transactions on Image Processing, 22*(7), 2574–2588.

Jung, C., & Kim, C. (2012). A unified spectral-domain approach for saliency detection and its application to automatic object segmentation. *IEEE Transactions on Image Processing, 21*(3), 1272–1283.

Kim, T. H., Lee, K. M., & Lee, S. U. (2008). Generative image segmentation using random walks with restart. In *Proceedings of the European conference on computer vision* (pp. 264–275).

Krishnan, D., Fattal, R., & Szeliski, R. (2013). Efficient preconditioning of laplacian matrices for computer graphics. *ACM Transactions on Graphics (TOG), 32*(4), 142.

Levin, A., Lischinski, D., & Weiss, Y. (2004). Colorization using optimization. In *ACM transactions on graphics (tog): 23* (pp. 689–694).

Lin, D., Dai, J., Jia, J., He, K., & Sun, J. (2016). Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3159–3167).

Lu, S., Tan, C., & Lim, J.-H. (2014). Robust and efficient saliency modeling from image co-occurrence histograms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*(1), 195–201.

Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the IEEE international conference on computer vision: 2* (pp. 416–423).

Min, D., Choi, S., Lu, J., Ham, B., Sohn, K., & Do, M. N. (2014). Fast global image smoothing based on weighted least squares. *IEEE Transactions on Image Processing, 23*(12), 5638–5653.

Mortensen, E. N., & Barrett, W. A. (1995). Intelligent scissors for image composition. In *Proceedings of the 22nd annual conference on computer graphics and interactive techniques* (pp. 191–198).

Muja, M., & Lowe, D. G. (2009). Fast approximate nearest neighbors with automatic algorithm configuration. In *In VISAPP international conference on computer vision theory and applications* (pp. 331–340).

Ren, X., & Malik, J. (2007). Tracking as repeated figure/ground segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–8).

Rother, C., Kolmogorov, V., & Blake, A. (2004). Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG): 23* (pp. 309–314).

Santner, J., Pock, T., & Bischof, H. (2010). Interactive multi-label segmentation. In *Proceedings of the Asian conference on computer vision* (pp. 397–410).

Shen, J., Du, Y., & Li, X. (2014). Interactive segmentation using constrained laplacian optimization. *IEEE Transactions on Circuits and Systems for Video Technology, 24*(7), 1088–1100.

Sheng, B., Sun, H., Magnor, M., & Li, P. (2014). Video colorization using parallel optimization in feature space. *IEEE Transactions on Circuits and Systems for Video Technology, 24*(3), 407–417.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(8). 888–805

Sinop, A. K., & Grady, L. (2007). A seeded image segmentation framework unifying graph cuts and random walker which yields a new algorithm. In *Proceedings of the IEEE international conference on computer vision* (pp. 1–8). IEEE.

Subr, K., Paris, S., Soler, C., & Kautz, J. (2013). Accurate binary image selection from inaccurate user input. In *Proceedings of the computer graphics forum: 32* (pp. 41–50). Wiley Online Library.

Unger, M., Pock, T., Trobin, W., Cremers, D., & Bischof, H. (2008). Tvseg-interactive total variation based image segmentation. In *Proceedings of the British machine vision conference: 31* (pp. 44–46).

Zemene, E., & Pelillo, M. (2016). Interactive image segmentation using constrained dominant sets. In *European conference on computer vision* (pp. 278–294).

Zhang, Z., Zhou, C., Wang, Y., & Gao, W. (2013). Interactive stereoscopic video conversion. *IEEE Transactions on Circuits and Systems for Video Technology, 23*(10), 1795–1808.