Pattern Recognition 79 (2018) 440–451

Contents lists available at ScienceDirect

Pattern Recognition

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

Diffusive likelihood for interactive image segmentation

Tao Wang[a], Zexuan Ji[a], Quansen Sun[a,*], Qiang Chen[a], Qi Ge[b], Jian Yang[a,*]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, No. 200, Street Xiao Ling Wei, Nanjing 210094 China
[b] School of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, China

ARTICLE INFO

Article history:
Received 4 September 2017
Revised 22 January 2018
Accepted 18 February 2018
Available online 19 February 2018

Keywords:
Interactive image segmentation
Likelihood diffusion
Perceptual learning
Graph cuts

ABSTRACT

The performance of conventional interactive image segmentation methods is strongly affected by seed quantity and position, and it is difficult for them to maintain global data coherence due to the bias that is caused by limited interactions. Furthermore, the pixel-level relationships in these methods are too local to capture long-range connectivity cues, which often causes them to obtain under-segmented results. To solve these problems, this paper proposes an interactive segmentation method that is based on likelihood diffusion and perceptual learning. The diffusive likelihood strategy is proposed for accurately estimating the prior label probability from limited user inputs. Superpixel-level grouping cues are utilized to enforce continuity during the segmentation process. The geometrical adjacency and long-range grouping cues are fused in the proposed framework to ensure that the segmentation results maintain proximity and continuity. The final results can be obtained by applying a joint optimization technique to solve a pair of sub-module functions. Experiments on the Berkeley segmentation data set and the Microsoft GrabCut database demonstrate that the proposed method outperforms state-of-the-art methods.

1. Introduction

Image segmentation can be described as partitioning an image into several homogeneous connected regions based on a similarity criterion by utilizing low-level visual feature, and extracting one or more objects that are of interest to the user from the complex background environment [1]. Segmented semantic regions or contours that are associated with real-word entities or scenes are the basis for further advanced image processing [2,3]. Therefore, image segmentation is a key step from image processing to image analysis and a fundamental problem in computer vision.

Many image segmentation methods have been proposed in the literature. Segmentation schemes can be classified into unsupervised, semi-supervised and fully supervised approaches. Unsupervised schemes, such as the minimum spanning tree (MST)-based method [4] and the mean shift method [5], can automatically segment the image without any prior information. In [4], image edge maps are utilized to build MSTs instead of the original images, which improve the stability of the original MST algorithm. However, because no prior knowledge is provided in these approaches, it is not possible for them to determine the target in which the user is interested. Therefore, such methods are only applied to specific segmentation tasks or certain types of images. Fully supervised schemes, such as the fully convolutional network approach [6], collect complete labelled images for model training. After the training phase, the classification process is executed based on the constructed model and segmentation results that are consistent with the training samples can be produced. However, for the same image, different users may not be interested in the same target. The model needs to be retrained once the user-desired object changes and the training phase is generally time consuming. Semi-supervised schemes, which are also referred to interactive approaches, such as the graph cut (GC) approach [7], allow the user to provide simple interactions to represent the label information during the segmentation. Compared with the other two types of segmentation schemes, the advantage of semi-supervised schemes is the ability to specify the users' intentions to obtain results that meet their demands. Furthermore, the prior label information that is provided by the user also helps to improve the segmentation performance.

The interactive segmentation algorithms can be generally classified into boundary-based approaches and region-based approaches. In boundary-based approaches, the user is asked to provide an initial area or contour that is close to the desired boundary and the algorithms evolve the initial area or contour to the desired boundary [8,9]. In general, a high level of accuracy is required from the user in specifying the initial contour to obtain a satisfactory segmentation result. In region-based approaches, the user is asked to provide an initial labeling of some pixels as seeds that belong to the foreground or the background, after which the algorithm

* Corresponding author.
E-mail addresses: wangtaoatnjust@163.com (T. Wang), sunquansen@njust.edu.cn (Q. Sun), csjyang@njust.edu.cn (J. Yang).

https://doi.org/10.1016/j.patcog.2018.02.023
0031-3203/

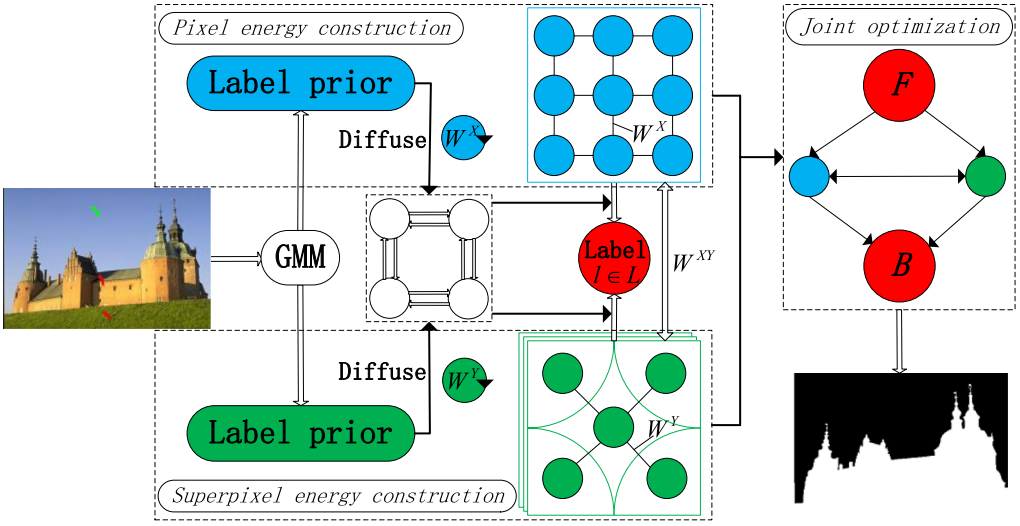

Fig. 1. Overview of the segmentation framework.

completes the labeling for all pixels [10–12]. Compared with boundary-based approaches, the main advantage of region-based approaches is that the cost of interactive effort is very modest since the user only needs to sketch the foreground and background in a few locations [13].

Typical examples of region-based approaches include GC [7], Random Walk (RW) [11] and the Shortest Path (SP) [14]. In these approaches, unary and pairwise potentials are generally constructed for the segmentation, which correspond to the region and boundary information, respectively. A unary potential measures the similarity of a pixel to labels, while the pairwise potential quantifies the similarity between pairs of pixels. These relationships can be represented via a graph and the energy function is constructed based on this graph. Since these energy functions are generally convex, it is possible to find the global optimum by the graph-theory-based optimization algorithms. As shown in [10], the quality of these interactive methods is strongly affected by seed quantity and position, and they are likely to fail to maintain global data coherence due to the bias that is caused by limited interactions. Furthermore, under-segmented results are always produced because the relationships of pixels are too local to capture long-range connectivity cues.

To address the above problems, in this work, we propose an interactive image segmentation method that is based on likelihood diffusion and perceptual learning. Fig. 1 shows a block diagram of the proposed algorithm. The main contributions of the proposed approach are as follows:

First, a likelihood diffusive method is proposed for obtaining an accurate estimate of prior label probability from limited seeds information. The global affinity can be utilized to overcome the sensitivity to seeds. Furthermore, an equivalence relationship between likelihood diffusion and likelihood learning is established to further verify the validity of the diffusive strategy. **Second**, superpixel-based grouping cues are introduced to enforce continuity for the object extraction. The relationships between pixels and superpixels are utilized to transfer communications between low-order and high-order cliques. **Last**, the segmentation model, which combines proximity and continuity, is constructed for utilizing the geometrical adjacency and long-range grouping cues. A joint optimization technique is utilized to solve a pair of sub-module functions based on the max-flow/min-cut algorithm [15].

2. Related work

It is shown in [1] that GC, RW and SP minimize a similar energy function under different L^q-norms ($q = 1, 2, \infty$). These methods have been widely used and many extended algorithms have been proposed in the literature.

Graph Cut: In the work of Boykov and Jolly [7], the interactive graph cut method was first proposed for segmenting grayscale medical images. Specific pixels are labelled by the user as foreground or background to provide the hard constraints, which helps to make the resulting segmentation adhere with the user's intention. Then, statistical histograms are utilized to estimate intensity distributions of the foreground and background from the seeded pixels. The segmentation problem is formulated through a cost function that consists of unary and pairwise potentials for utilizing the region and boundary properties of the segment. Minimizing the binary cost function is equivalent to finding a minimum cut on a specific graph, in which two virtual terminals are added to represent the foreground and background labels, connections between pixels and terminals are utilized to represent the region information, and connections between neighbouring pixels are utilized to represent the boundary information.

Many extensions of GC have been proposed in the literature. Lazy snapping (LS) [16] constructed the graph based on superpixels instead of pixels to improve efficiency. A coarse-to-fine user interface is also designed to provide instant visual feedback. GrabCut [17] extended the graph cut approach to colour images by utilizing a Gaussian mixture model (GMM) to model the F and B regions. Incomplete trimaps are also provided to simplify the user interaction through an iterative optimization process. Texture-aware model (TAM) [13,18] combined the colour and texture information to overcome the difficulties in handling images with textures. For objects with specific shapes, shape priors are introduced into the graph cut framework [19] to restrict segmentation results to a particular class of shapes, which helps to improve the accuracy for objects that lack salient edges. ACP-cut [10] combined GC and a semi-supervised kernel matrix model to better preserve the details near object boundaries. Moreover, the seed information is propagated to achieve discriminative structure learning and reduce the computational complexity. However, most of these methods require many seeds for learning the foreground and background models, which makes them sensitive to the number of seeds. Furthermore, they have difficulty maintaining global coherence in their results because of the bias that is caused by limited seeds information.

Random Walk: In the RW model [11], the probability that a random walker that starts at a pixel first reaches the foreground or background seeds is computed for each unseeded pixel. Then, each pixel is classified into the corresponding group according to the maximal probability. However, as shown in [20], RW is prone to producing "flatter" results since its methodology does not exhibit anisotropic behaviour. Random Walk with Restart [21] introduced the steady-state probability between unseeded pixels and seeded pixels for solving the weak boundary and texture problems of RW. In contrast to methods that are mentioned above that formally minimize the "distance" between pairwise pixels, Laplacian Coordinates (LC) [20] minimized the average distance while better controlling anisotropic propagation of labels. Sub-Markov Random Walk (SMRW) [22] estimated the global label prior based on the seeds to improve the segmentation accuracy of thin and elongated objects.

Shortest Path: In the SP model [14], the distances of paths from unseeded pixels to seeds are computed. The, a pixel is assigned the foreground label if there is a shorter path to a foreground seed than to any background seed. It intrinsically belongs to the L^{∞}-norm approaches [1]. Although SP is attractive in terms of speed, it is strongly influenced by the position of the seeds.

Higher-order Model: To improve the robustness to user inputs, many perceptual grouping methods have been proposed by utilizing superpixels to capture long-range grouping cues. In these approaches [23–25], pixels that are within the same superpixel are assigned the same label. However, superpixels tend not to emphasize sufficiently the proximity, thereby generating isolated regions. To obtain reliable results, the relationships among pixels and superpixels are fused to enforce proximity and continuity by the GC framework [26,27]. However, only the relationships from superpixels to pixels are explored in these methods, regardless of the inherent impacts between pixels and superpixels. To further improve the segmentation performance, the multi-layer relationships between pixels and multiple superpixels are combined to segment pixels and superpixels together by the RW framework [28,29]. Although better performance can be achieved with these methods, the optimization of their matrix equations involves inverting a large matrix, which results in high computational complexity.

3. Image segmentation by likelihood diffusion and perceptual learning

This paper proposes an interactive image segmentation method with likelihood diffusion and perceptual learning. For an input image, the user needs to first provide some seeds for the foreground and background. Then, the GMM is utilized to estimate the initial probabilities of pixels and superpixels from the limited seed information. A likelihood diffusion strategy is further proposed for obtaining more accurate label prior probabilities by considering the global similarity relationships. A segmentation model is constructed based on the combination of pixel-based energy and superpixel-based energy. The output segmentation result can be produced by a joint optimization method that is based on graph cuts (shown in Fig. 1).

3.1. Construction of a graph

An image can be represented by a graph $G = (\mathrm{X} \cup \mathrm{Y}, W)$, where $\mathrm{X} = \{x_i\}_{i=1}^{\mathrm{N_X}}$ is a collection of pixels, $\mathrm{Y} = \{y_i\}_{i=1}^{\mathrm{N_Y}}$ is a collection of superpixels, and $\mathrm{N_X}$ and $\mathrm{N_Y}$ are the numbers of pixels and superpixels, respectively. Multiple superpixels can be produced by a single unsupervised segmentation algorithm, such as mean shift [5], with different parameters. $W = \begin{bmatrix} W^{\mathrm{X}} & W^{\mathrm{XY}} \\ W^{\mathrm{YX}} & W^{\mathrm{Y}} \end{bmatrix}$ is a similarity matrix, where $W^{\mathrm{X}} = [W_{ij}^{\mathrm{X}}]_{\mathrm{N_X}\times\mathrm{N_X}}$ represents the relationships of pairwise pixels, $W^{\mathrm{Y}} = [W_{ij}^{\mathrm{Y}}]_{\mathrm{N_Y}\times\mathrm{N_Y}}$ represents the relationships of pairwise superpixels, and $W^{\mathrm{XY}} = [W_{ij}^{\mathrm{XY}}]_{\mathrm{N_X}\times\mathrm{N_Y}}$ $(W^{\mathrm{YX}} = (W^{\mathrm{XY}})^{\mathrm{T}})$ represents the relationships between pixels and superpixels.

W^{X} is defined as a typical Gaussian function:

$$W_{ij}^{\mathrm{X}} = \begin{cases} \exp(-\beta \left\| c_i - c_j \right\|_2) & if\ x_j \in \aleph_i \\ 0 & otherwise \end{cases} \tag{1}$$

where c_i denotes the intensity feature at pixel x_i, β is a constant that controls the strength of the weight, and $\aleph_i$ represents the neighbourhood of x_i. If two neighbouring pixels have similar features, their weight is large, and vice versa.

W^{Y} is defined based on pairwise superpixels:

$$W_{ij}^{\mathrm{Y}} = \begin{cases} \exp(-\beta \left\| c_i - c_j \right\|_2) & if\ y_j \in \aleph_i \\ 0 & otherwise \end{cases} \tag{2}$$

where the intensity feature c_i of superpixel y_i is defined as the mean intensity feature of all pixels in superpixel y_i.

W^{XY} is defined based on the relationships between pixels and superpixels:

$$W_{ij}^{\mathrm{XY}} = \begin{cases} \exp(-\beta \left\| c_i - c_j \right\|_2) & if\ x_i \in y_j \\ 0 & otherwise \end{cases} \tag{3}$$

3.2. Estimation of prior label probability

User inputs represent the prior label information. We utilize the GMM to estimate the initial prior label probability based on the seed information that is provided by the user. After GMM, the Gaussian parameters set can be produced as: $\theta^l = \{\theta_1^l, ..., \theta_K^l\}$ with $\theta_k^l = \{\nu_k^l, \mu_k^l, \sum_k^l\}$ $(l \in L)$, where L represents the set of labels, K represents the number of Gaussian components, and ν_k^l, μ_k^l, and $\sum_k^l$ denote the mixture coefficient, the mean, and the covariance of the kth Gaussian component, respectively. The initial probability $\bar{p}(x_i/y_i|l)$ of pixel/superpixel x_i/y_i with label l can be obtained as follows:

$$\bar{p}(x_i/y_i|l) = \max_{k\in\{1,...,K\}} \nu_k^l \frac{\exp(-\frac{1}{2}[c_i - \mu_k^l]^T (\sum_k^l)^{-1} [c_i - \mu_k^l])}{\sqrt{(2\pi)^{\dim} |\sum_k^l|}} \tag{4}$$

where dim is the dimension of c_i. Then, the value of $\bar{p}(x_i/y_i|l)$ is normalized with the constraint $\sum_l \bar{p}(x_i/y_i|l) = 1$.

It is difficult for the GMM to accurately estimate the prior probability when the number of seeds is limited. Inspired by the label propagation method [30] which learns the local and global consistency based on a label diffusive process, in this paper, we propose a likelihood diffusion method for propagating the initial probability for more accurate estimation of the prior label information. The process of likelihood diffusion is defined as:

$$(P_l^X)^{(t)} = \alpha_p Q^X (P_l^X)^{(t-1)} + (1 - \alpha_p)\bar{P}_l^X \tag{5}$$

where $P_l^X = [p(x_i|l)]_{\mathrm{N_X}\times 1}$ is the diffusive likelihood probability of the pixel; Q^X is the row-normalized matrix of W^X: $Q^X = (D^X)^{-1} \times W^X$, where $D^X = diag([d_1^X, ..., d_{\mathrm{N_X}}^X])$ with $d_i^X = \sum_{j=1}^{\mathrm{N_X}} W_{ij}^X$; $\bar{P}_l^X = [\bar{p}(x_i|l)]_{\mathrm{N_X}\times 1}$ is the initial likelihood probability, which is estimated by the GMM; $0 < \alpha_p < 1$ is the controlling parameter; and t represents the iterative step. Benefitting from the diffusive strategy, the global similarity affinity can be utilized to obtain an accurate probability estimate, even with limited seeds.

The closed form of the diffusive matrix P_l^X at step t can be written as:

$$(P_l^X)^{(t)} = (\alpha_p Q^X)^{t-1} \bar{P}_l^X + (1 - \alpha_p) \sum_{i=0}^{t-1} (\alpha_p Q^X)^i \bar{P}_l^X \tag{6}$$

Because $0 < \alpha_p < 1$, we can derive:

$$\lim_{t\to\infty}(\alpha_p Q^X)^{t-1}\bar{P}_l^X = 0 \text{ and } \lim_{t\to\infty}\sum_{i=0}^{t-1}(\alpha_p Q^X)^i = (I-\alpha_p Q^X)^{-1} \quad (7)$$

where I is the identity matrix.

Hence, after self-normalization, the diffusive probability converges to

$$\lim_{t\to\infty}(P_l^X)^{(t)} = (1-\alpha_p)(D^X - \alpha_p W^X)^{-1}\bar{P}_l^X \quad (8)$$

The multiplication of the inversion of a matrix by a single vector can be efficiently performed by the MATLAB division operator '\'. Then, the value of $p(x_i|l)$ is normalized with the constraint $\sum_l p(x_i|l) = 1$.

The above likelihood diffusive strategy can be interpreted as the minimization of the following cost function:

$$E(P_l^X) = \sum_{i,j=1}^{N_X} W_{ij}^X \cdot \left|p(x_i|l) - p(x_j|l)\right|^2 + \eta\sum_{i=1}^{N_X} d_i^X\Big[\tilde{p}(x_i|l)\cdot\left|p(x_i|l) - \frac{1}{d_i^X}\right|^2 + \sum_{l'\in L/l}\tilde{p}(x_i|l')\cdot|p(x_i|l)|^2\Big] \quad (9)$$

where $\eta = (1-\alpha_p)/\alpha_p$. The first term constrains the neighbouring pixels with high similarities to having similar likelihood probabilities. The second term constrains the likelihood estimate to maintain consistency with the initial prior probabilities. A pixel x_i should be assigned a high $p(x_i|l)$ if its initial prior probability $\tilde{p}(x_i|l)$ is high, and vice versa. The equivalence relationship between the likelihood diffusion in Eq. (5) and the likelihood learning in Eq. (9) can be used to verify the effectiveness of both.

The initial probabilities of superpixels can also be diffused and we can obtain the diffusive likelihood of superpixels as follows:

$$\lim_{t\to\infty}(P_l^Y)^{(t)} = (1-\alpha_s)(D^Y - \alpha_s W^Y)^{-1}\bar{P}_l^Y \quad (10)$$

where $P_l^Y = [p(y_i|l)]_{N_Y\times 1}$ is the diffusive likelihood probability of the superpixel, $D^Y = diag([d_1^Y, ..., d_{N_Y}^Y])$ with $d_i^Y = \sum_{j=1}^{N_Y} W_{ij}^Y$, $\bar{P}_l^Y = [\tilde{p}(y_i|l)]_{N_Y\times 1}$ is the initial likelihood probability, which is estimated by the GMM, and $0 < \alpha_s < 1$ is a parameter that controls the diffusion of superpixels.

3.3. Construction of the segmentation model

We design the segmentation model based on perceptual grouping laws, where the relationships among pixels and superpixels are fused to enforce proximity and continuity. Both geometrical adjacency and long-range cues are critical for the segmentation. The segmentation energy function with respect to the pixel is defined as follows:

$$E(f^X) = \sum_{i,j=1}^{N_X} W_{ij}^X \cdot \delta(f_i^X \neq f_j^X) + \varepsilon_p\sum_{i=1}^{N_X}(1 - p(x_i|f_i^X)) + \mu\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y} W_{ij}^{XY}\cdot\delta(f_i^X \neq f_j^Y) \quad (11)$$

where f^X denotes the labeling of pixels, $\delta(f_i^X \neq f_j^X) = 1$ if $f_i^X \neq f_j^X$ and equals 0 otherwise, and ε_p and μ are two controlling parameters. The first energy term is utilized to measure the extent to which f^X (two neighbourhood pixels) is not piecewise smooth. With a higher weight W_{ij}^X, the penalty for assigning different labels to x_i and x_j is larger. The second energy term is utilized to measure the disagreement between labeling f^X and the observed data. The third energy term is utilized to impose a higher-order constraint between pixels and superpixels. The pixel likely belongs to the same label as its corresponding superpixel if their weight W^{XY} is large.

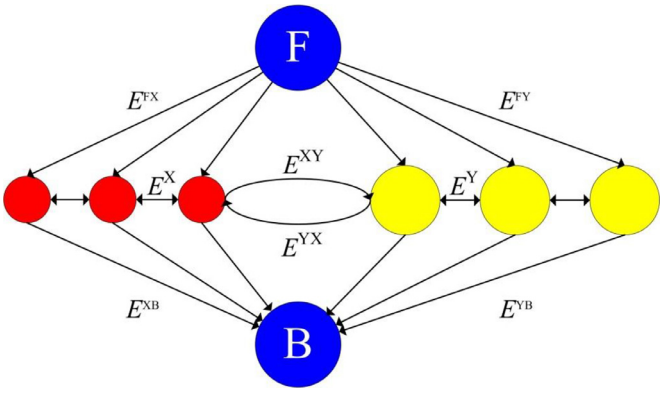

Fig. 2. Construction of the graph. Blue, red and yellow nodes represent terminals, pixels and superpixels, respectively. E^{FX} and E^{XB} (E^{FY} and E^{YB}) represent the connections between terminals and pixels (superpixels). E^X (E^Y) represents the connections between neighbouring pixels (superpixels). E^{XY} (E^{YX}) represents the connections from pixels (superpixels) to superpixels (pixels). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The segmentation energy function with respect to the superpixel is defined as:

$$E(f^Y) = \sum_{i,j=1}^{N_Y} W_{ij}^Y\cdot\delta(f_i^Y \neq f_j^Y) + \varepsilon_s\sum_{i=1}^{N_Y}(1 - p(y_i|f_i^Y)) + \mu\sum_{i=1}^{N_Y}\sum_{j=1}^{N_X} W_{ij}^{YX}\cdot\delta(f_i^Y \neq f_j^X) \quad (12)$$

where f^Y denotes the labeling of superpixels and ε_s and μ are two controlling parameters. The first energy term is utilized to measure the extent to which f^Y (two connected regions) is not piecewise smooth. The second energy term is utilized to measure the disagreement between labeling f^Y and the observed data. The third energy term considers interactions of superpixels and pixels. Each superpixel consists of multiple pixels and its label depends on the overall relationships of the pixels that it contains.

In contrast to the hard constraint relationships in these works [23-25], this paper utilizes the soft relationships between pixels and multiple superpixels to reduce the influence of inaccurate superpixels. In contrast to the methods in [26,27], which consider the hidden constraint relationships from superpixels to pixels, the mutual impacts between pixels and superpixels are considered in the proposed method. In this way, not only can higher-order cliques impose the superpixel consistency on pixels, but additionally, the pixels can provide feedback and obtain more accurate relationships. Compared with pixel-superpixel combination approaches [28,29], optimization of the sub-module functions in our method is more efficient.

3.4. Joint optimization technique

The functions in Eqs. (11)–(12) are supplementary to each other and we jointly optimize them based on graph cut. Fig. 2 illustrates the construction of the graph, in which blue nodes represent terminals, red nodes represent pixels, and yellow nodes represent superpixels. E^{FX} and E^{FY} consist of edges from F to all pixel-nodes and superpixel-nodes, respectively. E^{XB} and E^{YB} consist of edges from all pixel-nodes and superpixel-nodes to B, respectively. E^X and E^Y consist of edges of neighbouring pairwise pixels and superpixels, respectively. E^{XY} consists of edges from pixels to their corresponding superpixels and E^{YX} consists of edges from super-

Table 1
Definition of weight for each edge.

Edge	Weight
E^{FX}	$\varepsilon_p p(x\|F)$ *for* $x \in X$
E^{FY}	$\varepsilon_s p(y\|F)$ *for* $y \in Y$
E^{XB}	$\varepsilon_p p(x\|B)$ *for* $x \in X$
E^{YB}	$\varepsilon_s p(y\|B)$ *for* $y \in Y$
E^{X}	W_{ij}^{X} *for* $x_i \in X$, $x_j \in X$
E^{Y}	W_{ij}^{Y} *for* $y_i \in Y$, $y_j \in Y$
E^{XY}	μW_{ij}^{XY} *for* $x_i \in X$, $y_j \in Y$
E^{YX}	μW_{ij}^{YX} *for* $y_i \in Y$, $x_j \in X$

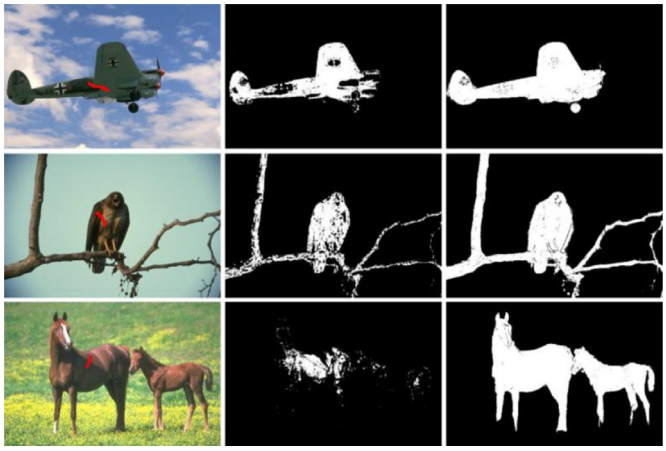

Fig. 3. Comparison of the likelihood probability map with (right) and without (middle) the likelihood diffusion.

pixels to their contained pixels. The weights of edges are defined in Table 1. The max-flow/min-cut [15] algorithm is utilized on this graph to obtain the globally optimal solution and the labeling f^X corresponds to the final segmentation result. The proposed algorithm can be easily extended for multi-label segmentation and the move-based algorithms [31] can be utilized to optimize the multi-label energy functions.

4. Experimental results

The proposed method was experimentally verified by comparing it with state-of-the-art approaches, namely, GrabCut [17], RW [11], LC [20], SMRW [22], the nonparametric higher-order method (NHO) [28], and the multi-layer graph constraint method (MGC) [29], on the Berkeley segmentation data set[1] and Microsoft GrabCut database[2] all of which have ground-truth annotations. The parameters that are involved in the proposed scheme are set as follows: (*hs, hr*) in mean shift [5] is set as {(10, 7), (10, 10), (10, 15)} to obtain three over-segmentations as our superpixels; constants β and K are fixed as 60 and 3, respectively; a 4-neighbourhood relationship is utilized for the pixel; and coefficients α_p, α_s, ε_p, ε_s, and μ are set to 0.95, 0.50, 0.01, 0.1, and 0.1, respectively.

4.1. Validity analysis

Fig. 3 shows the estimates of the likelihood probability map with/without the likelihood diffusion. The left column shows the test images, in which the red scribbles represent the seeds. The middle column shows the results that are estimated by the GMM. The GMM cannot accurately estimate the likelihoods when the number of seeds is limited. The right column shows the results that are based on the proposed likelihood diffusion. The diffusive strategy can help obtain accurate likelihoods even with limited seeds. Fig. 4 shows the comparison of the segmentation results with (right) and without (middle) the higher-order learning. Better performance can be achieved by imposing the superpixel consistency constraint.

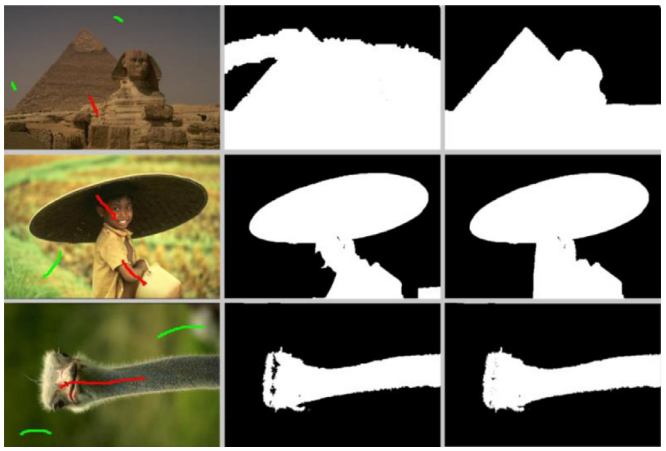

Fig. 4. Comparison of the segmentation results with (right) and without (middle) the higher-order learning.

Table 2
Mean ± standard deviation (Std) and the average rank (Ar) of PRI and VoI for the compared methods on the Berkeley segmentation data set.

Method	PRI		VoI	
	Mean ± Std	Ar	Mean ± Std	Ar
GrabCut [17]	0.68 ± 0.14	3.5	1.50 ± 0.42	4.9
RW [11]	0.59 ± 0.11	2.1	1.80 ± 0.58	5.5
LC [20]	0.62 ± 0.14	3.4	1.75 ± 0.61	5.0
SMRW [22]	0.72 ± 0.07	4.5	1.47 ± 0.34	4.2
NHO [28]	0.73 ± 0.08	4.4	1.32 ± 0.28	3.1
MGC [29]	0.74 ± 0.07	4.3	1.32 ± 0.25	3.2
Ours	**0.76 ± 0.06**	**5.8**	**1.22 ± 0.25**	**2.1**

4.2. Results on the Berkeley segmentation data set

Fig. 5 shows the comparison of state-of-the-art interactive segmentation methods. Fig. 5(a) shows the test images from the Berkeley segmentation data set with a few scribbles. Fig. 5(b)–(h) show the segmentation results of GrabCut [17], RW [11], LC [20], SMRW [22], NHO [28], MGC [29], and the proposed method, respectively. GrabCut and RW cannot obtain satisfactory results with limited seeds. The RW extension methods, namely, LC and SMRW, can obtain better results than RW. However, they are also sensitive to the seeds. Compared with the pixel-based methods, the superpixel-based methods are more robust to the seeds due to the higher-order constraint. According to the results for NHO and MGC, thin objects are not well-segmented. The proposed method produces the best results among the compared methods.

The probabilistic rand index (PRI) [32] and variation of information (VoI) [33] are utilized to quantitatively evaluate the segmentation performance on the Berkeley segmentation data set. The value of PRI ranges from 0 to 1, with a higher value representing a more accurate result. The value of VoI ranges within $[0, \infty)$, with a smaller value representing a more accurate result. Table 2 lists the mean ± standard deviation and the average rank according to the Friedman statistical test [34,35] (with a significance level of 0.05) of PRI and VoI for each compared method. The proposed method

[1] http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/S.
[2] http://research.microsoft.com/enus/um/cambridge/projects/visionimagevideoediting/segmentation/grabcut.htm.

Fig. 5. Comparison of the proposed approach with state-of-the-art approaches with a few scribbles. (a) Test image from the Berkeley segmentation data set with scribbles, (b)–(h) segmentation results of GrabCut [17], RW [11], LC [20], SMRW [22], NHO [28], MGC [29], and the proposed method.

outperforms the other methods with the largest PRI value and the smallest VoI value. The Friedman test determines the chi-square (χ^2) value as 17.06 (19.63) and the p-value as 9.1e-03 (3.2e-03) for PRI (VoI). According to the χ^2 distribution table, the critical value for $(7-1)=6$ degrees of freedom at the 0.05 significance level is 12.59. Since the χ^2 value is larger than the critical value, H_0 is rejected and H_1 is accepted, which substantiates the significant difference in behaviour among the compared methods. These quantitative and qualitative comparisons confirm the validity of the proposed method on the Berkeley segmentation data set.

4.3. Results on the Microsoft GrabCut database

We demonstrate the performance of the proposed method on the Microsoft GrabCut database. The error rate is utilized to evaluate the segmentation accuracy, which is defined as the ratio of the number of wrongly labelled pixels to the total number of unlabelled pixels. Fig. 6 illustrates example segmentations that were obtained using trimaps. Fig. 6(a) shows the trimap input in the Microsoft GrabCut database (white: foreground, dark grey: background, and light grey: the region to be classified). Fig. 6(b)–(h) show the segmentation results with error rates of GrabCut [17], RW [11], LC [20], SMRW [22], NHO [28], MGC [29], and the proposed method, respectively. Due to the influence of low contrast, the conventional methods cannot obtain accurate object boundaries. Furthermore, they are very sensitive to thin and slender objects and cannot obtain complete contours. According to a comparison of all the methods, the proposed method achieves high-quality segmentation results. Fig. 7 shows all the segmentation results and error rates of the proposed method. Table 3 summarizes the mean, standard deviation and average rank with the Friedman statistical test (with a significance level of 0.05) of error rates that are achieved by various methods. Compared with state-of-the-art methods, the proposed method achieves the lowest error rate. The Friedman test determines the χ^2 value as 108.12 and the p-value as 5.02e-21, which substantiates the significant differences in behaviour among the compared methods. These quantitative and qualitative experiments confirm the validity of the proposed method on the Microsoft GrabCut database.

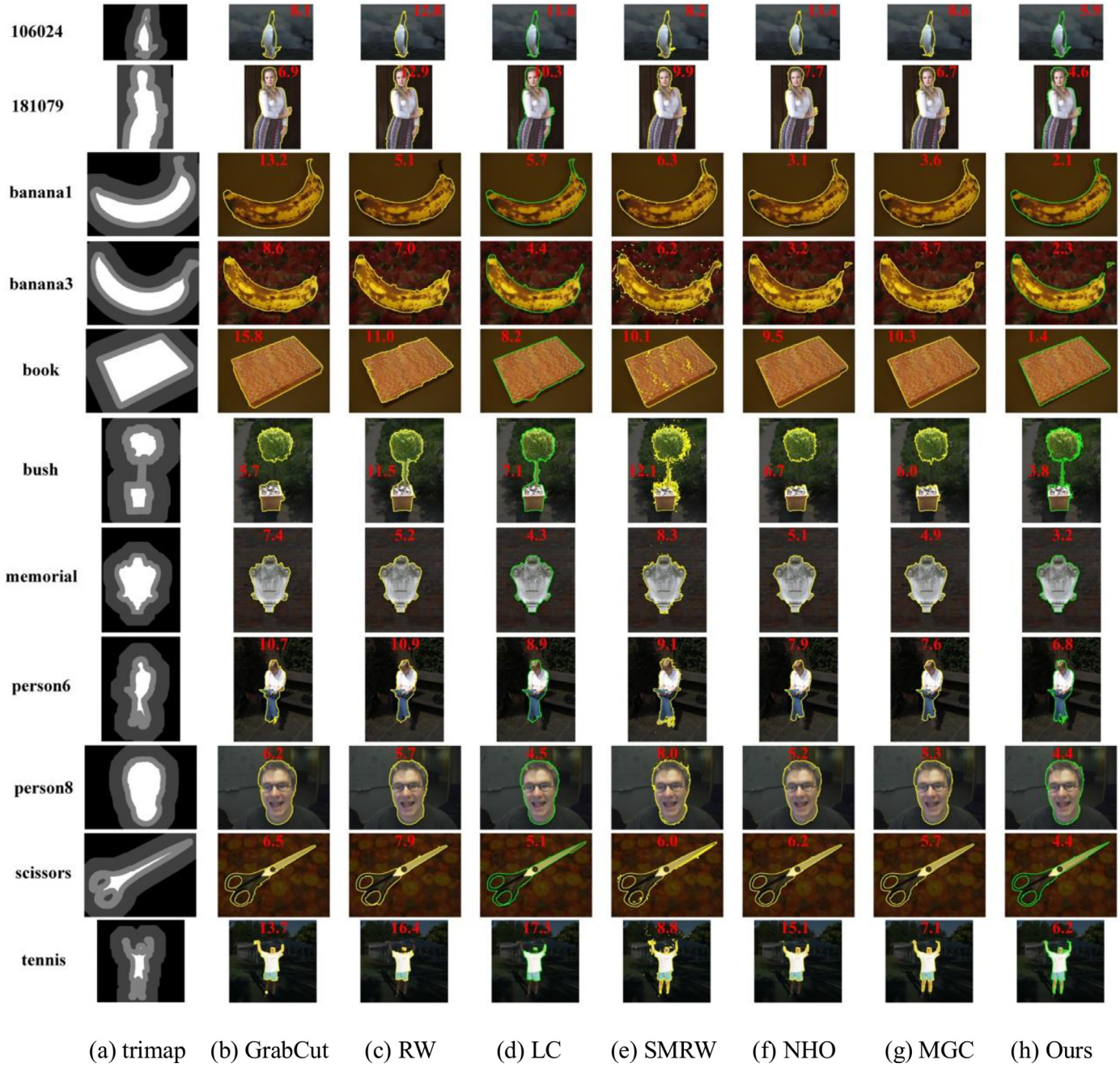

Fig. 6. Example segmentations using trimaps. (a) trimap input from the Microsoft GrabCut database (white: foreground, dark grey: background, and light grey: the region to be classified), (b)–(h) segmentation results of GrabCut [17], RW [11], LC [20], SMRW [22], NHO [28], MGC [29], and the proposed method.

4.4. Sensitivity analysis

Similar to the evaluation in [28], we analyse the sensitivity of the proposed method with respect to seed quantity and placement. The standard segmentations are obtained from the initial trimaps that are provided by the Microsoft GrabCut database. Then, the initial seeds are randomly selected from 50% to 1% of the total seed quantity. The perturbed segmentation results are recomputed from these selected seeds and compared with the standard segmentations. The normalized overlap $ao = |F1 \cap F2/F1 \cup F2|$ is used to measure the similarity of two segmentations [28], where $F1$ and $F2$ indicate the sets of pixels in two segmentations. Fig. 8 shows the comparison of example segmentations from the Microsoft GrabCut database with 50%, 30%, 10% and 1% seeds. Almost the same results are produced and even with 1% seeds, the proposed method can still obtain satisfactory segmentation results. Table 4 shows a comparison of the average normalized overlap in the Microsoft GrabCut database when varying the seed quantity among 50%, 30%, 10% and 1% of total seed quantity. When the seed quantity is large, these compared methods obtain similar results. When the seed quantity is small, the proposed method achieves the best performance. Fig. 9 shows the normalized overlap curves of each test image of our method when the seed quantity is varied. These quantitative and qualitative results show that the proposed method has strong robustness to seed quantity and placement.

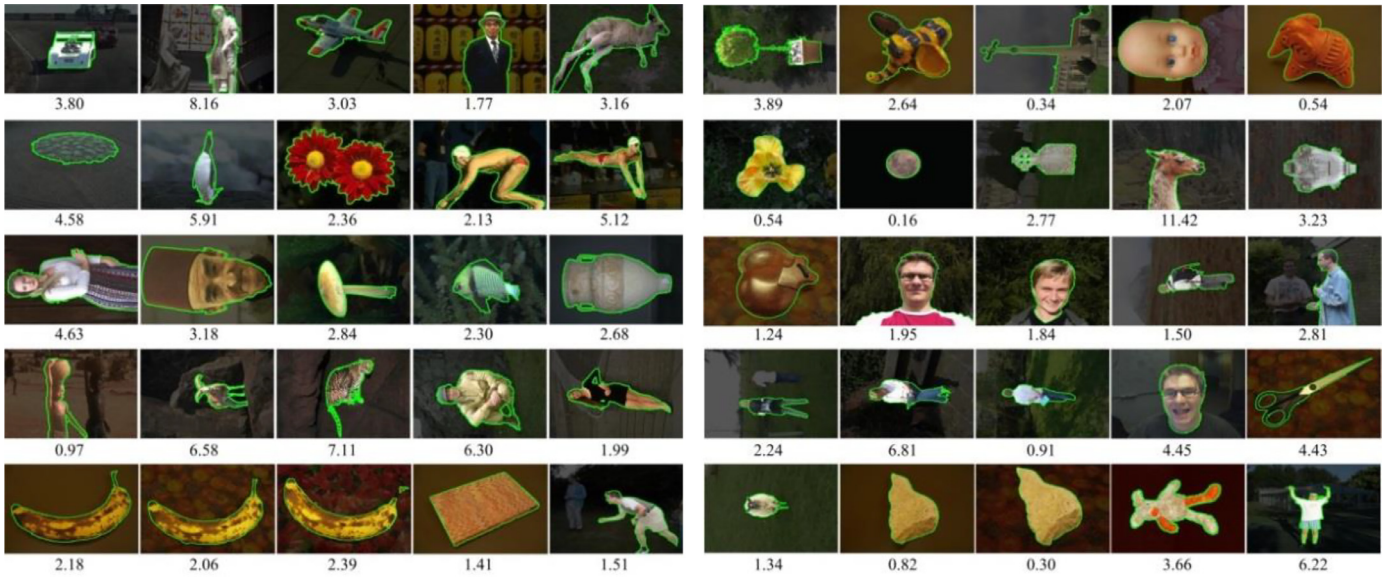

Fig. 7. Segmentation results and error rates (%) on the Microsoft GrabCut database for our method.

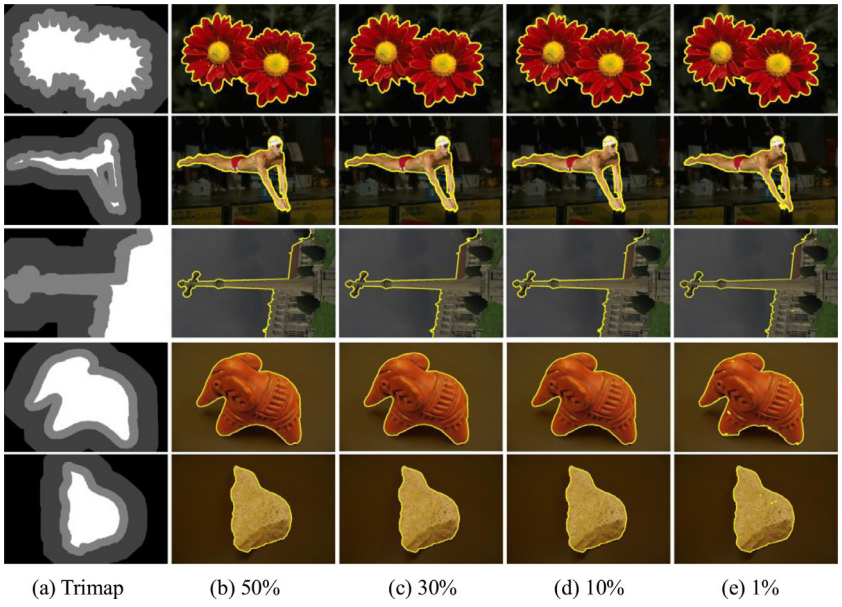

Fig. 8. Sensitivity analysis of the proposed method with respect to seed quantity and placement. (a) Trimap input from the Microsoft GrabCut database, (b)–(e) perturbed segmentations with 50%, 30%, 10% and 1% seeds.

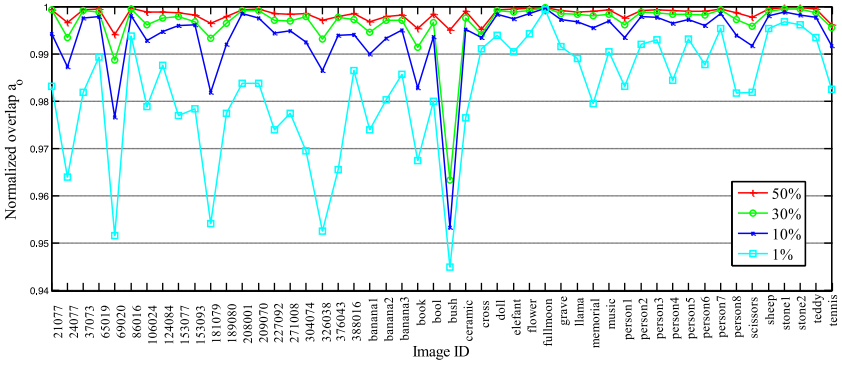

Fig. 9. Normalized overlap *ao* of each image from the Microsoft GrabCut database when varying the seed quantity among 50%, 30%, 10% and 1% of the total seed quantity.

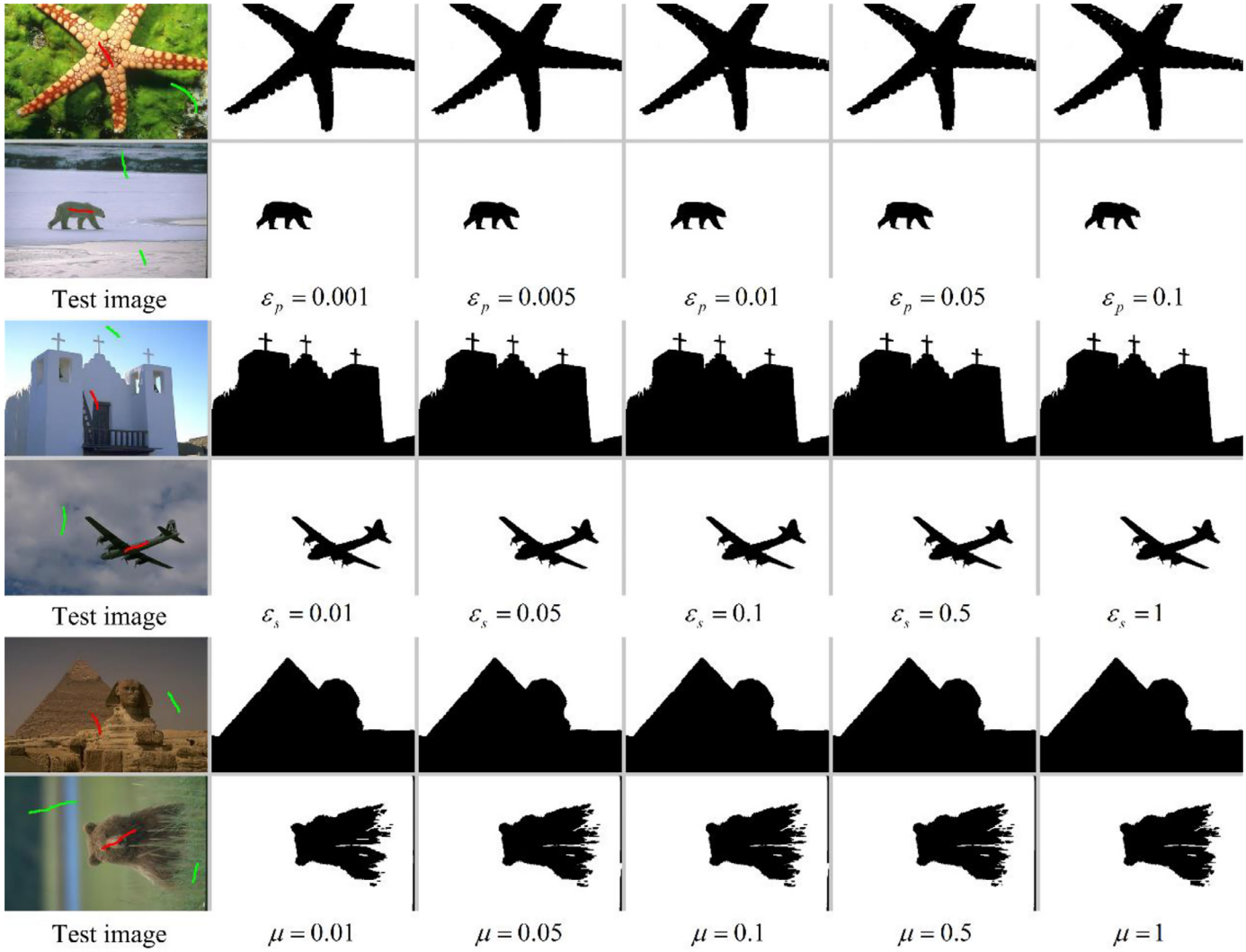

Fig. 10. Example segmentations with respect to the variation of ε_p, ε_s, and μ. Column 1 shows test images from the Berkeley segmentation data set. Columns 2–6 contain the resulting segmentations according to ε_p, ε_s, and μ, where ε_p ranges from 0.001 to 0.1, ε_s ranges from 0.01 to 1, and μ ranges from 0.01 to 1.

Table 3
Mean ± standard deviation (Std) and the average rank of error rates for the compared methods on the Microsoft GrabCut database.

Method	Error rate	
	Mean ± Std (%)	Average rank
GrabCut [17]	5.46 ± 4.2	4.7
RW [11]	6.45 ± 4.8	6.0
LC [20]	5.04 ± 3.8	4.5
SMRW [22]	4.61 ± 3.2	4.4
NHO [28]	4.25 ± 3.7	3.4
MGC [29]	3.44 ± 2.9	2.6
Ours	**3.08 ± 2.2**	**2.3**
GC [7]		6.60 (reported in [28])
GMMRF [36]		7.90 (reported in [37])
Robust P^n [26]		6.08 (reported in [28])
LS [16]		6.65 (reported in [13])
CRW [37]		4.08 (reported in [37])
TAM [13]		3.64 (reported in [13])
RMG [38]		3.79 (reported in [38])
PLL [39]		3.49 (reported in [39])

Table 4
Comparison of the average normalized overlap *ao*(%) with 50%, 30%, 10% and 1% seeds.

Percent seeds	50%	30%	10%	1%
RW [11]	**99.8**	99.1	98.0	89.5
GrabCut [17]	**99.8**	**99.7**	98.9	90.7
NHO [28]	**99.8**	**99.7**	**99.5**	96.9
Ours	**99.8**	**99.7**	99.4	**98.2**

4.5. Parameter settings

The parameters ε_p, ε_s, and μ are utilized to control the influences of the likelihood probability and the higher-order constraint on the segmentation. Fig. 10 shows examples of the segmentations with respect to the variation in the three parameters. Column 1 shows the test images from the Berkeley segmentation data set, in which the red and green scribbles represent the seeds for different labels. Columns 2–6 contain the resulting segmentations according to different values of ε_p, ε_s, and μ, where ε_p varies from 0.001 to 0.1, ε_s varies from 0.01 to 1, and μ varies from 0.01 to 1. Similar results are produced with the variation of the parameters. Table 5 shows the quantitative comparisons of the average error rates with different values of ε_p, ε_s, and μ over all 50 im-

Table 5
Average error rates (%) on the Microsoft GrabCut database with different values of parameters ε_p, ε_s, and μ.

Test Values	0.001	0.01	0.1	1
ε_p	4.16	**3.08**	3.32	3.37
ε_s	3.48	3.11	**3.08**	3.42
μ	3.75	3.15	**3.08**	3.41

Table 6
Average error rates (%) on the Microsoft GrabCut database with different values of parameters α_p and α_s.

Test Values	0.1	0.5	0.9	0.95	0.99
α_p	3.80	3.72	3.25	**3.08**	3.71
α_s	3.34	**3.08**	3.13	3.20	3.22

Table 7
Average running times of GrabCut [17], RW [11], LC [20], SMRW [22], NHO [28], MGC [29], and the proposed method on all 20 images of size 321×481 in the Microsoft GrabCut database.

Method	GrabCut	RW	LC	SMRW	NHO	MGC	Ours
Time (s)	**0.7**	0.8	3.2	5.1	11.0	5.4	3.4

ages in the Microsoft GrabCut database. The best result is obtained when $\varepsilon_p = 0.01$, $\varepsilon_s = 0.1$, and $\mu = 0.1$. Furthermore, these quantitative and qualitative experiments show that the segmentation results are not sensitive to these three parameters.

Parameters α_p and α_s are utilized to control the influence of the likelihood diffusion. Table 6 shows quantitative comparisons of the average error rates with different values of α_p and α_s over all 50 images in the Microsoft GrabCut database. The segmentation results are not sensitive to these two parameters, and the best result is obtained when $\alpha_p = 0.95$ and $\alpha_s = 0.5$. Therefore, in this paper, we experimentally set α_p and α_s to 0.95 and 0.5.

4.6. Runtimes

Table 7 shows a comparison of the average running times on all 20 test images of size 321×481 in the Microsoft GrabCut database on an Intel i7 CPU that is running at 4.20 GHz in MATLAB R2017a. The time costs of NHO [28], MGC [29], and our method do not include the over-segmentation step which takes approximately 6.0 s for the mean shift algorithm to generate three over-segmentations. Higher-order-based methods have more time costs than pixel-level-based methods. The proposed method takes approximately 3.4 s to segment an image of size 321×481. Consequently, the proposed method outperforms other higher-order approaches both in segmentation accuracy and algorithmic complexity. The likelihood diffusion and perceptual learning steps make our method more complex than pixel-level approaches. Considering the significant improvement of algorithm performance both in segmentation accuracy and the robustness to seeds, the proposed method is still competitive in practical applications. The algorithm complexity of the proposed method mainly focuses on the generation of multiple superpixels. It is worth mentioning that we can further reduce the algorithm complexity by selecting more efficient over-segmentation algorithms.

5. Conclusions

In this paper, an interactive image segmentation method that is based on diffusive likelihood and perceptual learning was proposed. The likelihood diffusion can help obtain an accurate estimate of the prior label probability from limited seed information, and the superpixel-level grouping cues can help enforce continuity for the object segmentation. A pair of energy functions that correspond to the pixel and superpixel were designed for fusing the geometrical adjacency and long-range grouping cues. To solve the proposed energy functions, a joint optimization technique was utilized to obtain the globally optimal solution. Comparison experiments on the Berkeley segmentation data set and the Microsoft GrabCut database demonstrated the effectiveness of the proposed method. Our future work will focus on refining the relationships between pixels and superpixels to further improve the efficiency of the algorithm.

Acknowledgments

This work was supported in part by the National Postdoctoral Program for Innovative Talents of China under Grant BX201700121, in part by China Postdoctoral Science Foundation under Grant 2017M621750, in part by the National Science Foundation of China under Grants 61673220, 61401209 and 61502244, in part by the Natural Science Foundation of Jiangsu Province, China under Grants BK20150859 and BK20140790.

References

[1] A.K. Sinop, L. Grady, A seeded image segmentation framework unifying graph cuts and random walker which yields a new algorithm, in: Proceedings of IEEE International Conference on Computer Vision, 2007, pp. 1–8.
[2] T. Wang, Z. Ji, Q. Sun, Q. Chen, S. Yu, W. Fan, S. Yuan, Q. Liu, Label propagation and higher-order constraint-based segmentation of fluid-associated regions in retinal SD-OCT images, Inf. Sci. 358 (C) (2016) 92–111.
[3] T. Wang, Z. Ji, Q. Sun, Q. Chen, S. Han, Image segmentation based on weighting boundary information via graph cut, J. Vis. Commun. Image Represent. 33 (2015) 10–19.
[4] Y. Liu, L. Huang, S. Wang, X. Liu, B. Lang, Efficient segmentation for region-based image retrieval using edge integrated minimum spanning tree, in: Proceedings of IEEE International Conference on Pattern Recognition, 2016, pp. 1929–1934.
[5] D. Comaniciu, P. Meer, Mean shift: A robust approach toward feature space analysis, IEEE Trans. Pattern Anal. Mach. Intell. 24 (5) (2002) 603–619.
[6] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 79 (10) (2014) 1337–1342.
[7] Y. Boykov, M. Jolly, Interactive graph cuts for optimal boundary & region segmentation of objects in ND images, in: Proceedings of IEEE International Conference on Computer Vision, 2001, pp. 105–112.
[8] E. Mortensen, W. Barrett, Interactive segmentation with intelligent scissors, Graph. Models Image Process. 60 (5) (1998) 349–384.
[9] L. Wang, C. Li, Q. Sun, Active contours driven by local and global intensity fitting energy with application to brain MR image segmentation, Comput. Med. Imaging Graph. 33 (7) (2009) 520–531.
[10] M. Jian, C. Jung, Interactive Image Segmentation using adaptive constraint propagation, IEEE Trans. Image Process. 25 (3) (2016) 1301–1311.
[11] L. Grady, Random walks for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 28 (11) (2006) 1768–1783.
[12] J. Shen, Y. Du, X. Li, Interactive segmentation using constrained Laplacian optimization, IEEE Trans. Circuits Syst. Video Technol. 24 (7) (2014) 1088–1100.
[13] H. Zhou, J. Zheng, L. Wei, Texture aware image segmentation using graph cuts and active contours, Pattern Recognit. 46 (6) (2012) 1719–1733.
[14] X. Bai, G. Sapiro, A geodesic framework for fast interactive image and video segmentation and matting, in: Proceedings of IEEE International Conference on Computer Vision, 2007, pp. 1–8.
[15] Y. Boykov, V. Kolmogorov, An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision, IEEE Trans. Pattern Anal. Mach. Intell. 26 (9) (2004) 1124–1137.
[16] Y. Li, J. Sun, C. Tang, H. Shum, Lazy snapping, ACM Trans. Graph. 23 (3) (2004) 303–308.
[17] C. Rother, V. Kolmogorov, A. Blake, Grabcut: interactive foreground extraction using iterated graph cuts, in: Proceedings of the ACM SIGGRAPH Conference, 2004, pp. 309–314.
[18] T. Wang, Z. Ji, Q. Sun, S. Han, Combining pixel-level and patch-level information for segmentation, Neurocomputing 158C (2015) 13–25.
[19] D. Freedman, T. Zhang, Interactive graph cut based segmentation with shape priors, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 755–762.
[20] W. Casaca, L.G. Nonato, G. Taubin, Laplacian coordinates for seeded image segmentation, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 384–391.
[21] T. Kim, K. Lee, S. Lee, Generative image segmentation using random walks with restart, in: Proceedings of European Conference on Computer Vision, 2008, pp. 264–275.

[22] X. Dong, J. Shen, L. Shao, L. Gool, Sub-Markov random walk for image segmentation, IEEE Trans. Image Process. 25 (2) (2016) 516–527.
[23] J. Shen, Y. Du, W. Wang, X. Li, Lazy random walks for superpixel segmentation, IEEE Trans. Image Process. 23 (4) (2014) 1451–1462.
[24] W. Wang, J. Shen, Higher-order image co-segmentation, IEEE Trans. Multimed. 18 (6) (2016) 1011–1021.
[25] P. Kohli, M. Kumar, P. Torr, P3 & beyond: solving energies with higher order cliques, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[26] P. Kohli, P. Torr, Robust higher order potentials for enforcing label consistency, Int. J. Comput. Vis. 82 (3) (2009) 302–324.
[27] T. Wang, J. Collomosse, Probabilistic motion diffusion of labeling priors for coherent video segmentation, IEEE Trans. Multimed. 14 (2) (2012) 389–400.
[28] T. Kim, K. Lee, S. Lee, Nonparametric higher-order learning for interactive segmentation, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3201–3208.
[29] T. Wang, Q. Sun, Z. Ji, Q. Chen, P. Fu, Multi-layer graph constraints for interactive image segmentation via game theory, Pattern Recognit. vol.55 (2016) 28–44.
[30] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, B. Scholkopf, Learning with local and global consistency, Adv.Neural Inf. Process. Syst. 16 (4) (2004) 321–328.
[31] Y. Boykov, O. Veksler, R. Zabih, Fast approximate energy minimization via graph cuts, IEEE Trans. Pattern Anal. Mach. Intell. 23 (11) (2001) 1222–1239.
[32] R. Unnikrishnan, C. Pantofaru, M. Hebert, Toward objective evaluation of image segmentation algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 929–944.
[33] M. Meilă, Comparing clusterings: an axiomatic view, in: Proceedings of International conference on Machine learning, 2005, pp. 577–584.
[34] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, J. Am. Stat. Assoc. 32 (1937) 674–701.
[35] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Ann. Math. Stat. 11 (1) (1940) 86–92.
[36] A. Blake, C. Rother, M. Brown, P. Perez, P. Torr, Interactive image segmentation using an adaptive GMMRF model, in: Proceedings of European Conference on Computer Vision, 2004, pp. 428–441.
[37] W. Yang, J. Cai, J. Zheng, J. Luo, User-friendly interactive image segmentation through unified combinatorial user inputs, IEEE Trans. Image Process. 19 (9) (2010) 2470–2479.
[38] T. Wang, Z. Ji, Q. Sun, Q. Chen, X. Jing, Interactive multi-label image segmentation via robust multi-layer graph constraints, IEEE Trans. Multimed. 18 (12) (2016) 2358–2371.
[39] T. Wang, Q. Sun, Q. Ge, Z. Ji, Q. Chen, G. Xia, Interactive image segmentation via pairwise likelihood learning, in: Proceedings of International Joint Conference on Artificial Intelligence, 2017, pp. 2957–2963.

Tao Wang received the B.E. degree in Computer Science and Technology, and the Ph.D. degree in pattern recognition and intelligence system from Nanjing University of Science and Technology (NUST), China, in 2012 and 2017, respectively. Currently he is doing postdoctoral research in pattern recognition and intelligence system from NUST. His research interests include image segmentation, pattern recognition, and video processing.

Zexuan Ji received the B.E. degree in Computer Science and Technology, and the Ph.D. degree in pattern recognition and intelligence system from Nanjing University of Science and Technology (NUST), China, in 2007 and 2012, respectively. Currently, he is an associate professor with the School of Computer Science and Engineering at the Nanjing University of Science and Technology. He visited the Shenzhen Institutes of Advanced Technology from eight months since Oct. 2009, and the School of Information Technologies, University of Sydney for one year since Nov. 2010. His current interests include medical imaging, image processing and pattern recognition.

Quansen Sun received the Ph.D. degree in pattern recognition and intelligence system from Nanjing University of Science and Technology (NUST), China, in 2006. He is a professor in the Department of Computer Science at NUST. His current interests include pattern recognition, image processing, remote sensing information system, and medical image analysis.

Qiang Chen received B.E. degree in computer science and Ph.D. degree in Pattern Recognition and Intelligence System from Nanjing University of Science and Technology, China, in 2002 and 2007, respectively. Currently, he is a professor in the School of Computer Science and Technology at the Nanjing University of Science and Technology. His main research topics are image segmentation, object tracking, image denoising, and image restoration.

Qi Ge received the B.Sc. degree in College of Math & Physics, Nanjing University of Information Science & Technology, Nanjing, China, in 2006, M.Sc. degree in Applied Mathematics from College of Math & Physics, Nanjing University of Information and Technology, Nanjing, China, in 2009, and Ph.D. degree in Pattern Recognition and Intelligent System in Nanjing University of Science and Technology, Nanjing, China, in 2013. Currently, she is an associate professor with the school of Telecommunications and Information Engineering at Nanjing University of Posts and Telecommunications. Her research interests include pattern recognition, image processing, and image segmentation.

Jian Yang received the B.S. degree in mathematics from the Xuzhou Normal University in 1995. He received the MS degree in applied mathematics from the Changsha Railway University in 1998 and the PhD degree from the Nanjing University of Science and Technology (NUST), on the subject of pattern recognition and intelligence systems in 2002. In 2003, he was a postdoctoral researcher at the University of Zaragoza. From 2004 to 2006, he was a Postdoctoral Fellow at Biometrics Centre of Hong Kong Polytechnic University. From 2006 to 2007, he was a Postdoctoral Fellow at Department of Computer Science of New Jersey Institute of Technology. Now, he is a professor in the School of Computer Science and Technology of NUST. He is the author of more than 80 scientific papers in pattern recognition and computer vision. His journal papers have been cited more than 2000 times in the ISI Web of Science, and 4000 times in the Web of Scholar Google. His "2DPCA" paper published in TPAMI 2004 has been cited more than 2000 in Scholar Google. His research interests include pattern recognition, computer vision and machine learning. Currently, he is an associate editor of Pattern Recognition Letters and IEEE Trans. Neural Networks and Learning Systems, respectively.