Information Sciences 538 (2020) 384–395

Contents lists available at ScienceDirect

Information Sciences

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

Error-tolerant label prior for interactive image segmentation

Tao Wang [a,*], Shengzhe Qi [a], Zexuan Ji [a], Quansen Sun [a,*], Peng Fu [a], Qi Ge [b]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[b] School of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

ARTICLE INFO

Article history:
Received 6 January 2020
Received in revised form 4 April 2020
Accepted 29 May 2020
Available online 13 June 2020

Keywords:
Interactive image segmentation
Error tolerance
Reliability learning
Graph cut

ABSTRACT

User interactions are generally utilized to impose the hard constraint and estimate the label prior probability in the interactive image segmentation methods. The conventional interactive approaches cannot work well when the user inputs contain incorrect marks. The existing error tolerance methods mainly focus on the interaction itself and try to eliminate the influence of incorrect hard constraint, which however ignore the label prior estimation. This paper mainly focuses on solving the label prior estimation problem when error marks appear. The prior probability is generally defined as the matching degree between pixels and the cluster centers of marked pixels. Incorrect interaction leads to the formation of incorrect clusters. Therefore, a reliability learning model is constructed in this paper by 1) assigning smaller weighting factors to incorrect clusters, 2) assigning larger weighting factors to correct clusters with higher matching degree. Accurate label prior probability can be obtained by the weighted averaging. Then referring to the existing methods, an error-tolerant segmentation model is designed as a ratio energy function, which can both overcome the hard constraint and the label prior limitation with error marks. Experiments on the challenging Berkeley segmentation data set and Microsoft GrabCut database demonstrate the effectiveness of the proposed method.

1. Introduction

As one of the most widely applied image segmentation method, interactive segmentation has attracted wide attention of researchers [2,4,8,9,25–26], in which user desired target can be quickly extracted with a small amount of human interaction, such as scribbles [4], object contours [14,24] and bounding boxes [19]. Interactive image segmentation plays a key role in many computer vision tasks, such as image editing, video caption and object tracking.

Many interactive image segmentation methods have been proposed, including the popular Graph Cut (GC) [4], Random Walk (RW) [8] and the Shortest Path (SP) [2]. The graph-theory-based framework is constructed for them to solve the segmentation problem. Image pixels are represented by nodes and pairwise relationships between pixels are represented by edges in the graph. Then the segmentation problem is converted to finding a minimum cut or an optimal path on the specific graph. For two-label foreground/background segmentation, the global optimal solution can be generally produced with the graph optimization. For images with strong discriminability, only a small amount of user interaction is needed for them to obtain accurate results. However, for images with similar foreground/background appearances or complex textures, more

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

https://doi.org/10.1016/j.ins.2020.05.122
0020-0255/

user interactions are required to produce satisfactory segmentations. The segmentation performance is generally sensitive to the quantity and location of the user interaction.

Benefited from the development of convolutional neural networks (CNN), many deep interactive image segmentation methods [11,15,16,18,20,23,28] have been developed recently. In these approaches, user interaction habits can be effectively learned via the CNNs. Potential image objects can be semantically perceived from limited user interaction by the deep learning techniques. The performance of these deep-based interactive methods is robust to the user interaction. Actually, only a few "clicks" are required (sometimes one "click" on the foreground) to obtain accurate segmentation results.

Though the interactive segmentation performance has been significantly improved by the deep learning-based techniques, the excellent performance of all the above methods was obtained satisfying a hard condition: all user marks must be correct. However, it is difficult to completely avoid mark errors when the user sketches scribbles or clicks on small and slender objects or around object boundary regions. Fig. 1 displays the change of segmentation results for GC [4] and the latest deep interactive method Back-propagating Refinement Scheme (BRS) [11] when mark errors appear. As shown in Fig. 1(b), the satisfactory result can be produced by GC with correct user scribbles (shown in Fig. 1(a)). However, the poor result (shown in Fig. 1(d)) is obtained for GC with error marks. As shown in Fig. 1(e), BRS obtains a satisfactory result with a few "clicks". However, the "flower" boundaries are not smooth enough. When editing boundary regions, error "clicks" easily happen (shown in Fig. 1(f)). Even for deep-based approaches, accurate segmentation results cannot be obtained with error marks. More clicks are required in order to eliminate the influence of incorrect clicks (shown in Fig. 1(g)). Fig. 1(h) shows the segmentation result of the proposed algorithm based on the user interaction in Fig. 1(c) with error marks. The motivation of this paper is to correct the possible label errors for interactive image segmentation task.

Many methods have been proposed to alleviate the problem of label errors in the interactive image segmentation [3,17,21,22]. These approaches generally focus on the user interaction itself, such as trying to find and remove the possible wrong marks. In [3], the location clue is also considered based on the assumption pixels near scribble boundary are more likely to be wrong marks. However, due to the randomness of user inputs, it is hard to directly judge the inaccurate user marks.

Error marks lead to incorrect hard constraint and incorrect prior probability. The existing error-tolerant approaches mainly focus on the hard constraint and ignore the label prior estimation, which causes them hard to obtain accurate results with incorrect prior probability. This paper focuses on solving the label prior estimation problem when error marks appear. Marked pixels are generally clustered to form multiple cluster centers and the closest distances between pixels and clusters are selected to measure how well pixels match this label. Incorrect user interaction leads to the formation of incorrect clusters, which in turn leads to incorrect estimation of the label prior probability. Based on the principles of continuity and distinguishability, we construct a believability learning model to measure the reliability factor between each pixel and each cluster by satisfying: (1) incorrect clusters are assigned small weighting factors; (2) in correct clusters, higher matching clusters are assigned larger weighting factors. The novel label prior probability can be obtained by the weighted average with all clusters, which can effectively eliminate the influence of mark errors. Then the segmentation model is designed as a ratio energy function by integrating the label prior information and the user inputs information together, which makes the segmentation error-tolerant to both the incorrect hard constraint and the incorrect prior estimation.

2. Related work

2.1. Traditional interactive methods

Typical interactive segmentation approaches include the popular GC [4], RW [8] and SP [2] algorithms. These methods minimize a similar energy function with different L-norms. In GC, the segmentation problem is formulated as finding a min-

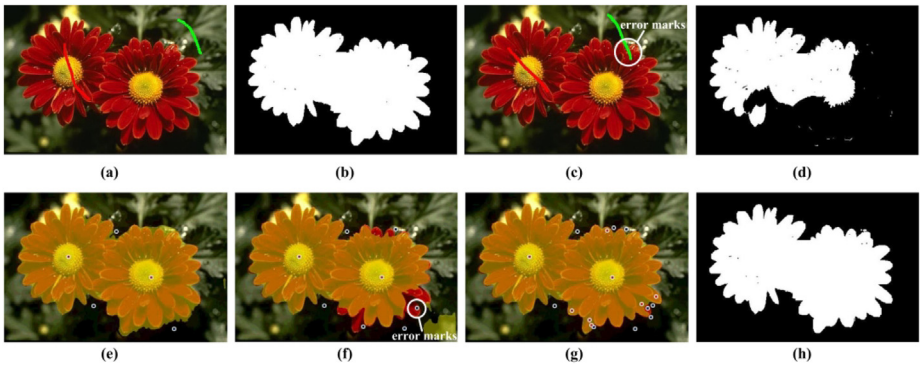

Fig. 1. Comparison of segmentation performance w/o label errors. (a)/(c) shows the test image without/with error marks, and (b)/(d) shows the corresponding segmentation result of GC [4]. (e)/(f) shows the segmentation result of the latest deep interactive approach Back-propagating Refinement Scheme (BRS) [11] with correct/incorrect "clicks". Both GC and BRS cannot obtain satisfactory results with error marks. As shown in (g), more interactions are required for BRS to overcome the influence of error marks. (h) shows the result of the proposed algorithm based on the incorrect interactions in (c).

imum cut on a specific s/t graph. Boykov and Jolly [4] first developed the interactive graph cut algorithm to segment the organ tissues in the medical images. Li *et al.* [13] replaced pixels with superpixels to improve the efficiency. Rother *et al.* [19] replaced statistical histogram with Gaussian mixture model to learn the prior appearances. Zhou *et al.* [29] fused the color and texture features to segment texture images. ACP-cut [12] propagated seed information to achieve discriminative structure clues by utilizing kernel learning. Instead of using max-flow/min-cut optimization, Heimowitz and Keller [10] formulated the segmentation as an inference problem via a probabilistic graph matching scheme. Grady *et al.* [8] first developed the interactive RW algorithm for medical image segmentation. Dong *et al.* [7] established a unified sub-Markov RW (SMRW) framework by combining the seed information vector and the prior probability vector. Casaca *et al.* [6] designed the Laplacian coordinates algorithm by measuring the average distances. Bampis *et al.* [1] extended RW to normalized RW (NRW) by weighing the contribution of every neighbor to the underlying diffusion.

2.2. Deep interactive methods

Recently, many deep learning-based approaches have been proposed for the interactive image segmentation task. Xu *et al.* [28] reduced user interactions to just a few clicks by introducing deep learning technique, which has much better understanding of objectness. User-provided clicks are transformed into distance maps which are then concatenated with the original image to compose training samples. The output probability from a trained fully convolutional network is integrated with GC optimization to obtain the segmentation result. Lin *et al.* [16] developed an algorithm to train convolutional networks for semantic segmentation supervised by scribbles. A graphical model is constructed by jointly propagating information from scribbles to unlabeled pixels and learning network parameters. Li *et al.* [15] presented an end-to-end learning approach to interactive segmentation by coupling two convolutional networks. One is trained to produce a diverse set of segmentations corresponding to user inputs, and the other is trained to select among these. Maninis *et al.* [18] developed an interactive segmentation method that requires human annotations on tight object boundaries. Song *et al.* [23] located foreground and background seeds to multiply annotations automatically. Jang and Kim [11] designed a back-propagating refinement scheme (BRS) to correct mis-segmented seeds, which shows superior performance compared with other deep interactive approaches.

Though high-quality segmentation results can be produced by the above methods, they cannot work well when incorrect scribbles or clicks appear. More additional interactive efforts are required in order to eliminate negative effects of error marks.

2.3. Existing error-tolerant methods

Liu *et al.* [17] specified new scribbles that partially overlap with the incorrect old scribbles. The hard constraints are enforced based on the new scribbles, and the old scribbles are utilized as the soft constraints for the segmentation. However, the performance still relies on the accuracy of the new scribbles. Sener *et al.* [21] proposed an error-tolerant interactive segmentation method by utilizing dynamic and iterated GCs. They tried to remove incorrect labeled pixels with some heuristics in the preprocessing step. Subr *et al.* [22] designed a dense conditional random field (CRF) model, which contains a unary term and a fully connected CRF with all pixel pairs, to produce the segmentation from possible incorrect scribbles. Bai and Wu *et al.* [3] designed a ratio energy function to tolerate mark errors in user inputs while encouraging maximum use of the label prior information. However, they only considered the role of the user interaction itself, such as the hard seed constraint, and ignored the estimation of label prior probability from inaccurate user inputs. Therefore, in their segmentation results, marked pixels and pixels near them can obtain correct labels, but other pixels are not guaranteed to obtain correct results.

3. Error-tolerant interactive image segmentation

3.1. Model analysis

Before solving the mislabeling problem in the interactive image segmentation, we analyze the reason why the conventional methods cannot work well with error marks first. Based on the roles of user interaction in the interactive image segmentation: 1) enforced as hard constraint that marked pixels keep the same labels during the segmentation, 2) utilized to estimate the label prior probability, when error marks appear we can conclude that: ***first***, the hard constraint causes the incorrect marked pixels and pixels near them to be mis-segmented (shown in Fig. 2(h): Region 1); ***second***, incorrect label prior estimation causes a large number of pixels to be mis-segmented (shown in Fig. 2(h): Region 2). The existing error-tolerant methods mainly focus on the first problem, ignoring the second problem. By comparison, we can find incorrect label prior estimation has a greater impact on the segmentation result than the hard constraint.

This paper focuses on solving the second problem: the incorrect label prior estimation. In the existing interactive methods, the clustering algorithms, such as the Gaussian mixture model, are generally utilized to cluster marked pixels for the foreground and background, respectively. Then the label prior is defined as the distance with the closest cluster center (or the maximum probability). Fig. 2 shows an example of probability maps with each cluster based on correct (Row 1st) and incorrect (Row 2nd) scribbles. For simplicity, the K-means algorithm is selected to obtain clusters C_1^F, C_2^F and C_3^F for

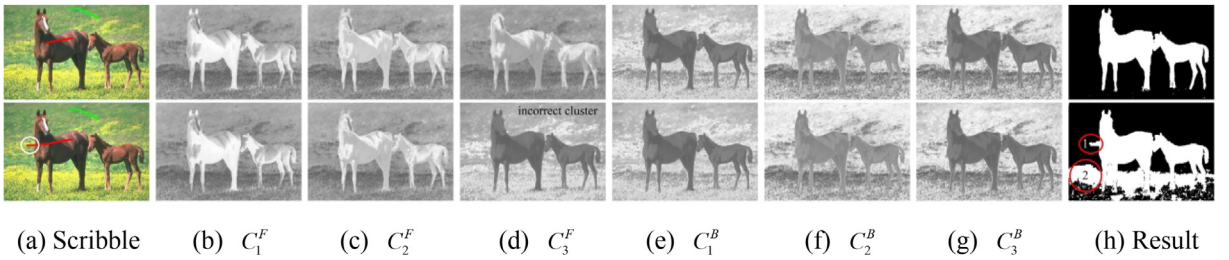

Fig. 2. Probability maps with each cluster based on correct (Row 1st) and incorrect (Row 2nd) scribbles. For each label, the number of clusters is set to 3. (b-d) shows the foreground maps with clusters C_1^F, C_2^F and C_3^F, respectively. (e-g) shows the background maps with clusters C_1^B, C_2^B and C_3^B, respectively. (h) shows the final segmentation result of GC.

the foreground, and C_1^B, C_2^B and C_3^B for the background (the number of clusters is set to 3). From the first row, all the clusters accurately represent the label information based on the correct user inputs. Therefore, the satisfactory result is obtained in Fig. 2(h). From the second row, scribbles contain mislabeled pixels, resulting in incorrect cluster C_3^F. As shown in Fig. 2(d), many background pixels have high probabilities with the foreground, resulting in an inaccurate segmentation result (shown in Fig. 2(h)). The segmentation errors in Region 1 are also partly due to the hard constraint.

It is unavoidable that incorrect user interaction can lead to erroneous clustering. Then how to accurately estimate the label prior probability based on the possible incorrect clusters. Fig. 3 shows four general strategies by taking three clusters as an example. The conventional interactive segmentation methods generally utilize the closest distance (shown in Fig. 3(a)) to estimate the label prior probability. If there is no incorrect cluster it is the most effective strategy. However, as shown in Fig. 2, it cannot work well when incorrect clusters appear. In the average strategy shown in Fig. 3(b), though the negative effects of incorrect clusters can be reduced, the positive effects of correct clusters are also weakened. Fig. 3(c) shows the optimal strategy, where the incorrect clusters are removed first and then the closest distance with the remaining clusters is computed. However, it is hard to directly determine whether the cluster is correct. In this paper, we propose a weighted average model to estimate the label prior probability based on the possible incorrect clusters (shown in Fig. 3(d)). A learned weighting factor is assigned for each pixel with each cluster, where correct/incorrect clusters should have large/small weights. The label prior probability can be defined as the weighted average distance with all the clusters.

3.2. Reliability learning

After clustering the marked pixels, the sets of clusters $C^F = \{C_k^F\}_{k=1}^{K^F}$ and $C^B = \{C_k^B\}_{k=1}^{K^B}$ can be obtained, where F/B represents the foreground/background and K^F/K^B (both set to 3) represents the number of clusters in F/B. For each label $l \in \{F, B\}$, the matching degree M_{ik}^l between each pixel $x_i \in \{x_i\}_{i=1}^N$ (N is the number of pixels) and each cluster $C_k^l \in C^l$ is defined as:

$$M_{ik}^l = exp(-||c_i - C_k^l||_2) \tag{1}$$

where c_i is the intensity feature of pixel x_i. A larger value of M_{ik}^l indicates a higher degree of matching between x_i and C_k^l.

A reasonable weighted average model should satisfy the following two conditions: ***first***, incorrect clusters should be assigned small weighting factors; ***second***, in correct clusters, higher matching clusters should be assigned larger weighting factors. Based on the above principles, take the foreground cluster C_k^F ($k \in \{1, ..., K^F\}$) as an example, the weighting factors

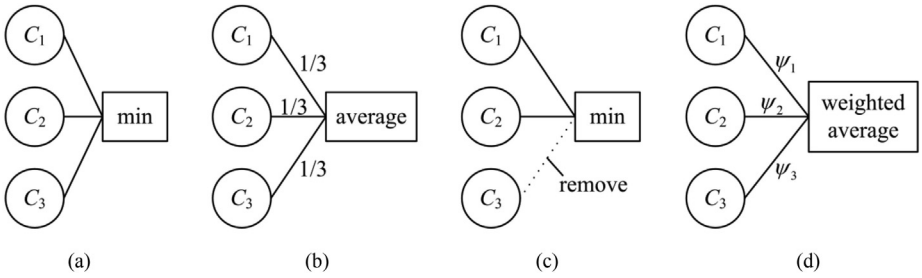

Fig. 3. Display of several label prior estimation strategies. Take three clusters C_1, C_2 and C_3 as an example, where in (a), the closest distance between pixels and clusters is selected, in (b), the average distance with all the clusters is computed, in (c), the possible incorrect cluster is removed first and then the closest distance with the remaining clusters is selected and in (d), the weighted average distance with coefficients ψ_1, ψ_2 and ψ_3 is adopted.

$\Psi_k = [\psi_{ik}]_{N\times 1}$ of each pixel with C_k^F can be learned by minimizing the following cost function:

$$E(\Psi_k) = E_{con}(\Psi_k) + \lambda_1 E_{dis}(\Psi_k) + \lambda_2 E_{aux}(\Psi_k) \tag{2}$$

where the energy term E_{con} is designed based on the continuity, utilized to impose neighboring constraints, the energy term E_{dis} is constructed based on the distinguishability, utilized to ensure the separability of the foreground and background, and the auxiliary energy term E_{aux} is built based on the unreliability, utilized to further weaken the influence of the incorrect cluster. λ_1 and λ_2 (both set to 1) are two controlling parameters utilized to balance the role of each energy term.

The continuity term E_{con} is defined as follows:

$$E_{con} = \sum_{i,j=1}^{N} w_{ij}(\psi_{ik} - \psi_{jk})^2 \tag{3}$$

where w_{ij} represents the similarity of neighboring pixels x_i and x_j, generally defined as a typical Gaussian function:

$$w_{ij} = \begin{cases} exp(-\beta \| c_i - c_j \|_2) & if\ x_j \in \aleph_i \\ 0 & otherwise \end{cases} \tag{4}$$

where c_i and c_j denote the intensity features of pixels x_i and x_j, respectively, β (set to 60) is a constant that controls the strength of the similarity, and $\aleph_i$ represents the neighbourhood of x_i. If two neighbouring pixels have similar features, the value of w_{ij} is large, and vice versa.

It can be noticed that E_{con} imposes the constraint that neighboring pixels x_i and x_j with a high similarity w_{ij} should have similar weighting factors ψ_{ik} and ψ_{jk}.

The distinguishability term E_{dis} is defined as follows:

$$E_{dis} = \sum_{i=1}^{N} d_i(\psi_{ik} - \bar{\psi}_{ik})^2 \tag{5}$$

where $d_i = \sum_{j=1}^{N} w_{ij}$ is utilized to balance the first energy term. $\bar{\psi}_{ik}$ can be understood as the possible value of the real weight if there is no error marks, which is defined based on the distinction between the foreground and the background:

$$\bar{\psi}_{ik} = \min_{\kappa\in\{1,\ldots,K^B\}} (M_{ik}^F - M_{i\kappa}^B)^2 \tag{6}$$

The value of $\bar{\psi}_{ik}$ is then normalized by satisfying the constraint $\sum_{k=1}^{K^F} \bar{\psi}_{ik} = 1$. $\bar{\psi}_{ik}$ satisfies the above two principles: ***first***, the value of $\bar{\psi}_{ik}$ is small if C_k^F is incorrect since there are similar clusters in the background; ***second***, if C_k^F is correct, a higher value of $\bar{\psi}_{ik}$ is obtained with a larger M_{ik}^F. In this case, the highest matching cluster will play the biggest role.

It can be noticed that E_{dis} imposes the constraint that the learned weighting factor ψ_{ik} should be consistent with $\bar{\psi}_{ik}$.

The auxiliary term E_{aux} is defined as follows:

$$E_{aux} = \sum_{i=1}^{N} R_{ik}\psi_{ik}^2 \tag{7}$$

where R_{ik} represents the unreliability of C_k^F, defined as:

$$R_{ik} = M_{ik}^F \times \max_{\kappa\in\{1,\ldots,K^B\}} M_{i\kappa}^B \tag{8}$$

A larger value of R_{ik} indicates C_k^F is more likely to be incorrect. R_{ik} will be large only when both M_{ik}^F and $\max_{\kappa\in\{1,\ldots,K^B\}} M_{i\kappa}^B$ are large, which corresponds to the fact that pixels affected by incorrect clusters are highly matched to both the foreground and the background (referring to pixels in Region 2 in Fig. 2).

It can be noticed that E_{aux} imposes the constraint that the weighting factor ψ_{ik} is small if the unreliability R_{ik} is large.

3.3. Optimization

The energy function in Eq. (2) can be rewritten into the following matrix form:

$$E(\Psi_k) = \Psi_k{}^T L \Psi_k + \lambda_1 (\Psi_k - \bar{\Psi}_k)^T D (\Psi_k - \bar{\Psi}_k) + \lambda_2 \Psi_k{}^T \Gamma \Psi_k \tag{9}$$

where $\Psi_k = [\psi_{ik}]_{N\times 1}$ is the learned vector of weighting factors, $L = D - W$ is the Laplacian matrix with $D = diag([d_1, \ldots, d_N])$ and $W = [w_{ij}]_{N\times N}$, $\bar{\Psi}_k = [\bar{\psi}_{ik}]_{N\times 1}$, and $\Gamma = diag([R_{ik}]_{N\times 1})$.

Differentiating $E(\Psi_k)$ with respect to Ψ_k, it has:

$$\frac{\partial E(\Psi_k)}{\partial \Psi_k} = L\Psi_k + \lambda_1 D(\Psi_k - \bar{\Psi}_k) + \lambda_2 \Gamma \Psi_k \tag{10}$$

By setting Eq. (10) to zero, we have:

$$\Psi_k = \lambda_1 (L + \lambda_1 D + \lambda_2 \Gamma)^{-1} D \bar{\Psi}_k \tag{11}$$

The value of ψ_{ik} is then normalized under the constraint $\sum_{k=1}^{K^F} \psi_{ik} = 1$. Since the matrix $L + \lambda_1 D + \lambda_2 \Gamma$ is large but very sparse, the multiplication of its inversion with a single vector typically has an efficient solution. The linear system solver implemented by the MATLAB division operator '\' is very efficient at finding the solution of Eq. (11).

3.4. Segmentation

In addition to eliminating the influence of incorrect clusters, the proposed error-tolerant model also has the characteristic that it has no negative effects on correct clusters since the highest matching cluster can still be assigned the largest weighting factor. However, the proposed method has a limitation that whether error marks appear in the foreground or the background needs to be determined before it works. An effective solution is to perform segmentation process based on the assumption that the error marks appear in the foreground or the background, respectively, and then select the optimal result between them. The extreme situation that both the foreground and the background contain error marks is an unsolvable problem since you cannot understand what the user wants to segment at all. The only way is to filter all user marks, such as removing suspicious marks based on the location and quantity information.

For the background, users can sketch scribbles far away the targets. For the foreground, it is hard to completely avoid mislabeling for small and slender objects. Therefore, error marks are more likely to appear in the foreground. This paper mainly focuses on solving the foreground errors in the experimental display. As described above, there are two reasons for incorrect segmentation caused by incorrect marks: ***first***, the hard constraint causes the incorrect marked pixels and pixels near them to be mis-segmented; ***second***, incorrect label prior estimation causes a large number of pixels to be mis-segmented. The proposed error-tolerant model focuses on solving the second problem. In order to solve these problems at the same time, we borrow ideas from this paper [3], and take the possible error marks appear in the foreground as an example, the segmentation energy function can be defined as:

$$E(f) = \frac{E_{GC}(f)}{M + U(f)} \tag{12}$$

where f represents the labeling of all pixels. $E_{GC}(f)$ is a GC-based energy, defined as:

$$E_{GC}(f) = \sum_{i,j=1}^{N} w_{ij} \cdot \delta(f_i \neq f_j) + \alpha \sum_{i=1}^{N} (1 - \Pr(x_i|f_i)) \tag{13}$$

where $f_i \in \{F, B\}$ represents the label of pixel x_i, the judgment function $\delta(f_i \neq f_j)$ equals 1 if $f_i \neq f_j$ and equals 0 otherwise, $\Pr(x_i|f_i)$ represents the prior probability of pixel x_i belonging to label f_i, and α is a controlling parameter utilized to balance these two energy terms. The first energy term, also called the boundary energy, is utilized to measure the extent to which f is not piecewise smooth. With a higher value of w_{ij}, the penalty of assigning different labels to x_i and x_j is larger. The second energy term, also called the region energy, is utilized to measure the disagreement between labeling f and the observed data.

The foreground error-tolerant prior probability is defined as follows:

$$\Pr(x_i|f_i) = \begin{cases} \sum\limits_{k \in \{1,\ldots,K^F\}} \psi_{ik} M_{ik}^F & \text{if } f_i = F \\ \max\limits_{k \in \{1,\ldots,K^B\}} M_{ik}^B & \text{if } f_i = B \end{cases} \tag{14}$$

The value of $\Pr(x_i|f_i)$ is then normalized by satisfying the constraint $\sum_{f_i \in \{F,B\}} \Pr(x_i|f_i) = 1$. The weighted average of $\psi_{ik} M_{ik}^F$ can help to obtain accurate foreground prior probability regardless of the influence of error marks.

Suggested by [3], the nonnegative utility function $U(f)$ is described as:

$$U(f) = \sum_{i=1}^{N} U(f_i) \tag{15}$$

$$U(f_i) = \begin{cases} 1 & \text{if } x_i \in S_F \text{ and } f_i = F \\ 1 & \text{if } x_i \in S_B \text{ and } f_i = B \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

where S_F and S_B represent the sets of marked pixels for the foreground and background, respectively. $U(f_i)$ counts the number of marked pixels that are respected in the final segmentation, which increases when more user input information is respected in the result.

The constant M is generally set as [3]:

$$M = \sum_{i=1}^{N} [U(f_i = F) + U(f_i = B)] \tag{17}$$

The larger M is, the more marked pixels are likely to be allowed foreground-background swap in the segmentation [3].

It tries to minimize the energy $E_{GC}(f)$ while maximizing the energy $M + U(f)$ when we minimize the energy function $E(f)$ in Eq. (12). The error-tolerance characteristics are reflected in the following two aspects: ***first***, the label prior estimation is error-tolerant benefited from our definition of the GC energy; ***second***, the optimization process of the ratio energy is to determine the optimal set of marked pixels such that the exchange of their labels promotes the maximum decrease of the GC energy $E_{GC}(f)$, thus promising the error-tolerance of the hard constraint.

Suggested by [3], the optimization of the above ratio energy function can be transformed into solving the following λ−function after convergence:

$$E^{\lambda}(f) = P(f) - \lambda Q(f) \tag{18}$$

where $P(f) = E_{GC}(f)$ and $Q(f) = M + U(f)$. More details about the ratio function optimization can refer to [3]. The detailed algorithm steps are described as:

Input: User scribbles and the parameter α

Output: Optimal labeling f^*
Compute the matching degree for F and B with Eq. (1)
For each cluster in F, compute the weighting factor with Eq. (11)
Obtain the prior probability with Eq. (14)
Set $k \leftarrow 0, f^k \leftarrow [0]_{N\times 1}$ and $\lambda^k \leftarrow P(f^k)/Q(f^k)$
While $\lambda^k \neq \lambda^{k-1}$ **do**
Compute $f^{k+1} = \arg\min_f E^{\lambda^k}(f)$ by the max-flow algorithm [5]
$\lambda^{k+1} \leftarrow P(f^{k+1})/Q(f^{k+1})$
$k \leftarrow k + 1$
end
return $f^* = f^k$

4. Experiments

To simplify parameter settings, the controlling parameters λ_1 and λ_2 in Eq. (2) are both set to 1, the number of clusters K^F and K^B are both set to 3 and the constant β in Eq. (4) is set to 60. Only the parameter α in Eq. (13) needs to be adjusted. In the following comparative experiments, the value of α is fixed as 0.01.

The subject of this paper is to solve the mislabeling problem, thus we do not rely on large scale datasets to verify the effectiveness of the proposed algorithm. The popular Berkeley segmentation data set[1] and Microsoft GrabCut database[2] are utilized in this paper, and the compared algorithms include GC [4], the latest RW-based approaches NRW [1], SMRW [7] and the probabilistic diffusion (PD) [27], the latest error-tolerant approach (ET) [3] and the latest deep interactive approach BRS [11].

4.1. Qualitative comparison results

Fig. 4 shows the qualitative comparison to state-of-the-art interactive segmentation approaches, where the odd rows and the even rows display the results with correct and incorrect user inputs, respectively. Fig. 4(a) shows the test images from the Berkeley segmentation data set. Fig. 4(b)-(h) show the segmentation results of GC [4], NRW [1], SMRW [7], PD [27], ET [3], BRS [11] and the proposed method, respectively. From the odd rows, the compared methods obtain satisfactory results in most test images based on correct user inputs. Accurate segmentations are also produced by the proposed method with correct scribbles (shown in Fig. 4(h)), which proves that the proposed error tolerant model does not play negative effects on correct clusters. From Fig. 4(g), the deep interactive method BRS can accurately extract the targets with only a few clicks,

[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. 4. Comparison to state-of-the-art interactive image segmentation methods with correct (odd row) and incorrect (even row) user inputs. (a) Test image from the Berkeley segmentation data set, (b)-(h) segmentation results of GC [4], NRW [1], SMRW [7], PD [27], ET [3], BRS [11] and the proposed method.

where the red and blue clicks represent the foreground and background seeds, respectively. Compared with the conventional interactive approaches, deep features have a stronger semantic understanding of the target. However, for uncommon objects or objects not contained in their training set, these deep-based methods still require a lot of clicks to obtain better results (shown in the first and the last images).

From the first two test images, it is hard for the user to accurately sketch scribbles or clicks on small and slender objects, such as the legs in the first image or the person in the second image. Therefore, it is hard to completely avoid error marks during the user interaction process. The even rows show the segmentation results of the compared methods when error marks appear. Compared with the odd rows, the worse results are obtained for all the compared methods with incorrect user inputs. The conventional approaches GC, NRW, SMRW and PD rely on the user inputs to obtain the prior label information. However, the incorrect hard constraint and inaccurate label prior probabilities cause them to produce poor results. ET is only error-tolerant to incorrect hard constraint, though part pixels around marked pixels might avoid being mis-segmented, the

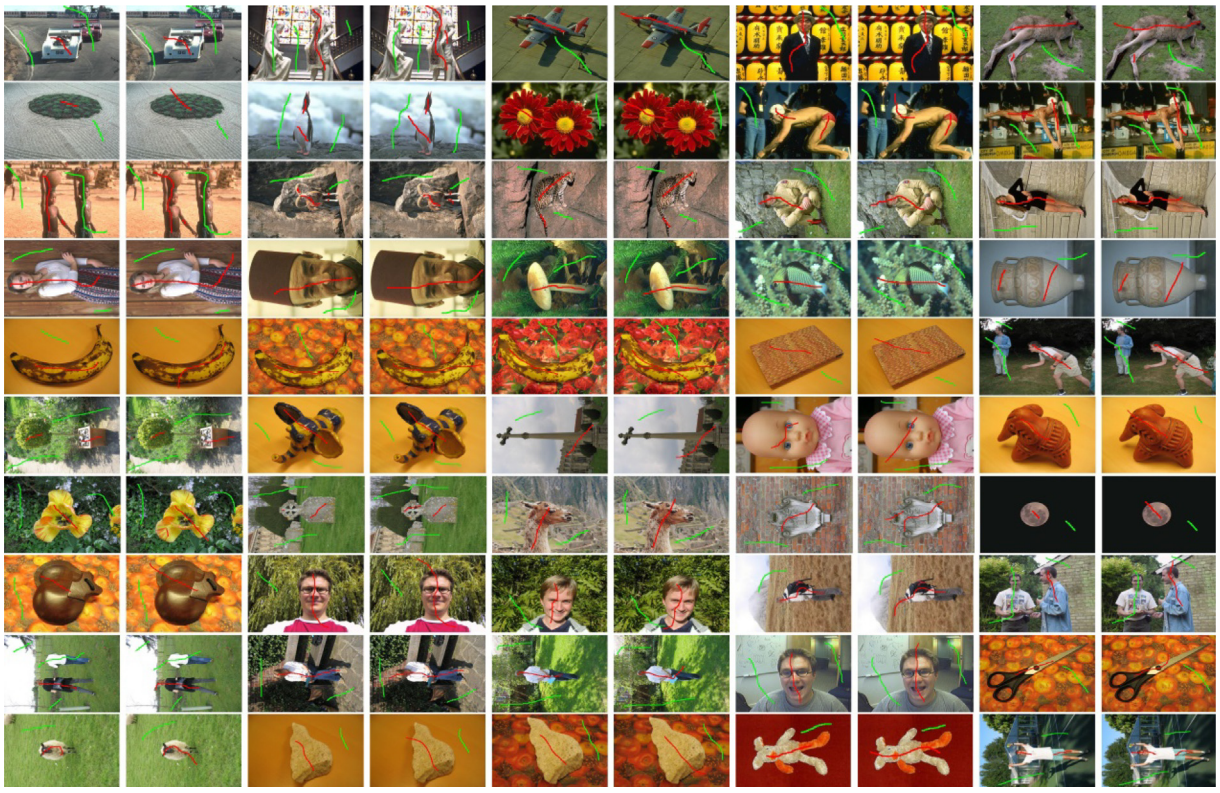

Fig. 5. Display of the test images in the Microsoft GrabCut database with correct and incorrect scribbles.

incorrect label prior probabilities prevent it from obtaining satisfactory results. From Fig. 4(g), even for the deep interactive approach BRS, the accurate results cannot be produced when error clicks appear. From Fig. 4(h), the proposed method still obtains accurate segmentation results regardless of the influence of incorrect user inputs. Compared with the odd rows, almost the same results are produced in the even rows by our method, which verifies the effectiveness of the proposed error-tolerant model. In our method, users can interact loosely, especially for those small and slender objects.

4.2. Quantitative comparison results

Microsoft GrabCut database has been widely used for the comparison of the interactive image foreground/background segmentation approaches, which consists of fifty test images with the corresponding ground-truth annotations. We quantitatively test the effectiveness of the proposed model based on this database. Since no public user scribbles are provided in this database, we sketch the correct and incorrect scribbles loosely by ourselves. Fig. 5 shows the detailed user interaction information for each test image, where in each pair of images, the first image are with correct scribbles and the second image are with incorrect scribbles.

The Jaccard index defined as the intersection-over-union (IoU) of the estimated segmentation and the ground-truth mask is employed to quantitatively evaluate the segmentation results. The IoU index has been widely adopted as it provides intuitive, scale-invariant information on the number of mislabeled pixels. Given an output segmentation O and the corresponding ground-truth mask G, the IoU score is defined as

$$IoU = \frac{|O \cap G|}{|O \cup G|} \tag{19}$$

Table 1 summarizes the average IoU scores (↑) obtained by GC [4], NRW [1], SMRW [7], PD [27], ET [3] and the proposed method with correct and incorrect scribbles on the Microsoft GrabCut database. It can be seen that GC obtains the highest IoU score 0.83 when the user inputs are correct. Error-tolerance approaches ET and the proposed algorithm also obtain good results (slightly lower than GC) when no error marks appear. However, the IoU score decreases from 0.83 to 0.68 for GC with incorrect scribbles. All the compared approaches are sensitive to the error marks. Even for the error-tolerance approach ET, its IoU score has also dropped significantly. By comparison, the proposed method obtains the highest IoU score 0.79 even with the incorrect scribbles, which shows the proposed model can effectively overcome the mislabeling limitation. Fig. 6 also

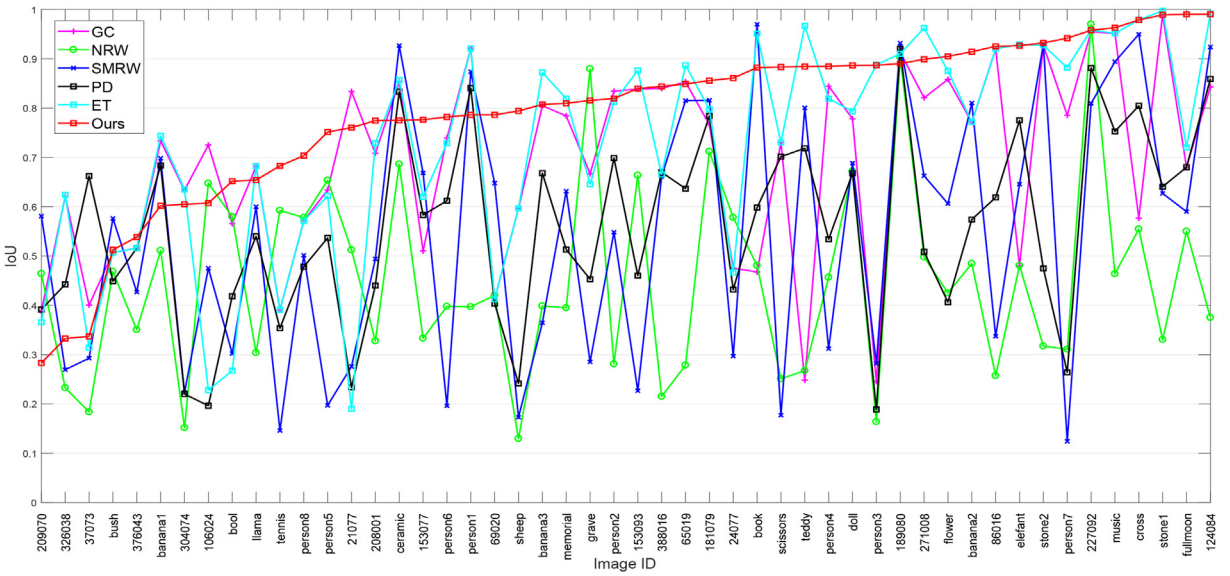

Fig. 6. IoU of each image with incorrect scribbles in the Microsoft GrabCut database by applying GC [4], NRW [1], SMRW [7], PD [27], ET [3] and the proposed method.

Table 1
The average IoU scores (↑) of GC [4], NRW [1], SMRW [7], PD [27], ET [3] and the proposed method on the Microsoft GrabCut database with correct and incorrect scribbles.

Method	GC	NRW	SMRW	PD	ET	Ours
Correct	0.83	0.59	0.59	0.79	0.81	0.81
Incorrect	0.68	0.45	0.54	0.56	0.72	0.79

shows the IoU curves of each test image with incorrect scribbles in the Microsoft GrabCut database by applying the compared methods. The images are sorted in ascending order based on the values of the proposed method. It can be clearly observed that the proposed method obtains the best performance in most cases.

4.3. Parameter settings

The controlling parameter α is used to balance the influence of the region energy and the boundary energy in Eq. (13). With a larger α, regional term plays a more important part and the details in objects can be better preserved. Comparatively smoother boundaries can be produced with a smaller α. Fig. 7 shows the quantitative evaluation of the IoU scores on the

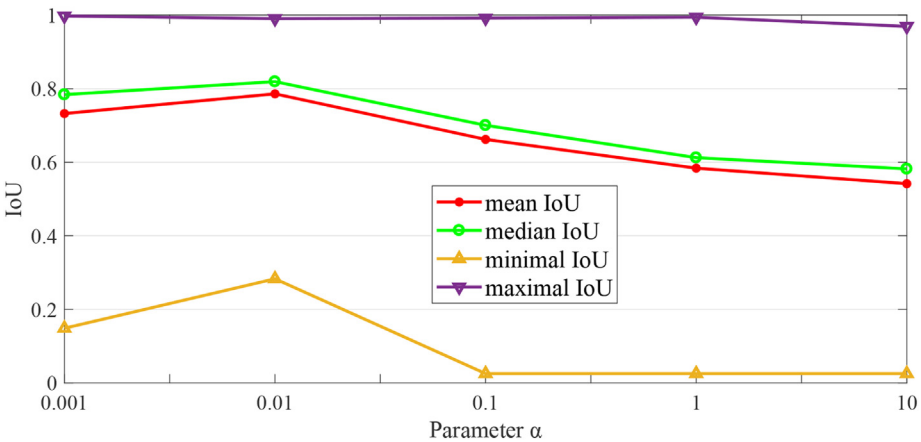

Fig. 7. IoU scores on the Microsoft GrabCut database with incorrect scribbles by varying the parameter α.

Table 2
Average running times of GC [4], NRW [1], SMRW [7], PD [27], ET [3] and the proposed method on all 20 images with size 321×481 in the Microsoft GrabCut database.

Method	GC	NRW	SMRW	PD	ET	Ours
Time (s)	0.6	6.8	5.2	1.3	1.4	1.5

Microsoft GrabCut database with incorrect scribbles by varying the values of α. The highest IoU score is obtained when $\alpha = 0.01$.

4.4. Runtimes

The algorithm cost of the proposed method mainly focuses on computing the weighting factor in Eq. (11) and solving the ratio energy function in Eq. (12). As described above, the MATLAB division operator '\' can efficiently find the solution of (11), and the ratio function in Eq. (12) can be transformed into the linear sub-module function in Eq. (18). Furthermore, the optimization process always converged in a few iterations (less than 5 iterations). Therefore, compared with the conventional approaches, our algorithm efficiency is also competitive. Table 2 lists the average running times of GC [4], NRW [1], SMRW [7], PD [27], ET [3] and the proposed method on all 20 test images with size 321×481 in the Microsoft GrabCut database on an Intel Xeon CPU running at 2.0 GHz in MATLAB. The average running time of the proposed method is 1.5 s to segment an image with size 321×481.

5. Conclusion

In this paper, an error-tolerant interactive segmentation model is constructed to eliminate the negative influence of incorrect user inputs. In order to accurately estimate the label prior probability from the possible incorrect user interaction, a reliability learning model is constructed by assigning small weights to incorrect clusters and assigning larger weights to correct clusters with higher matching degree. Accurate label prior probability can be produced by the weighted averaging with all the clusters. Satisfactory segmentation results can be obtained by solving a ratio energy function, which makes the proposed method error-tolerant to both the hard constraint and the label prior estimation. Comparison experiments confirm the effectiveness of the proposed method both in accuracy and efficiency when error marks appear.

Funding

This work was supported in part by the Natural Science Foundation of Jiangsu Province, China, under Grant BK20180458, in part by the National Science Foundation of China under Grants 61802188, 61673220, 61801222 and 61972213, and in part by the Fundamental Research Funds for the Central Universities under Grant 30919011230.

CRediT authorship contribution statement

Tao Wang: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration. **Shenzhe Qi:** Investigation. **Zexuan Ji:** Validation. **Quansen Sun:** Supervision. **Peng Fu:** Data curation. **Qi Ge:** Investigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to thank the editor and the anonymous reviewers for their critical and constructive comments and suggestions. This work was supported in part by the National Science Foundation of China under Grant 61802188, in part by the Natural Science Foundation of Jiangsu Province, China, under Grant BK20180458, in part by the National Science Foundation of China under Grants 61673220, 61801222 and 61972213, in part by the Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and in part by the Jiangsu Key Lab of Image and Video Understanding for Social Security.

References

[1] C.G. Bampis, P. Maragos, A.C. Bovik, Graph-driven diffusion and random walk schemes for image segmentation, IEEE Trans. Image Process. 26 (1) (2016) 35–50.
[2] 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.
[3] J. Bai, X. Wu, Error-tolerant scribbles based interactive image segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 392–399.
[4] 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.
[5] 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.
[6] 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.
[7] X. Dong, J. Shen, L. Shao, L. Gool, Sub-Markov random walk for image segmentation, IEEE Trans. Image Process. 25 (2) (2016) 516–527.
[8] L. Grady, Random walks for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 28 (11) (2006) 1768–1783.
[9] S. Han, W. Tao, D. Wang, X. Tai, X. Wu, Image segmentation based on GrabCut framework integrating multiscale nonlinear structure tensor, IEEE Trans. Image Process. 18 (10) (2009) 2289–2302.
[10] A. Heimowitz, Y. Keller, Image Segmentation via Probabilistic Graph Matching, IEEE Trans. Image Process. 25 (10) (2016) 4743–4752.
[11] W.D. Jang, C.S. Kim, Interactive image segmentation via backpropagating refinement scheme, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5297–5306.
[12] M. Jian, C. Jung, Interactive image segmentation using adaptive constraint propagation, IEEE Trans. Image Process. 25 (3) (2016) 1301–1311.
[13] Y. Li, J. Sun, C. Tang, H. Shum, Lazy snapping, ACM Trans. Graphics 23 (3) (2004) 303–308.
[14] Y. Li, G. Cao, T. Wang, Q. Cui, B.S. Wang, A novel local region-based active contour model for image segmentation using Bayes theorem, Inf. Sci. 506 (2020) 443–456.
[15] Z. Li, Q. Chen, V. Koltun, Interactive image segmentation with latent diversity, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 577–585.
[16] D. Lin, J. Dai, J. Jia, K. He, J. Sun, Scribblesup: Scribble-supervised convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3159–3167.
[17] J. Liu, J. Sun, H.Y. Shum, Paint selection, ACM Trans. Graphics 28 (3) (2009) 69.
[18] K.K. Maninis, S. Caelles, J. Pont-Tuset, L.V. Gool, Deep extreme cut: From extreme points to object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 616–625.
[19] 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.
[20] L. Shao, Z. Cai, L. Liu, K. Lu, Performance evaluation of deep feature learning for RGB-D image/video classification, Inf. Sci. 385 (2017) 266–283.
[21] O. Sener, K. Ugur, A. A. Alatan. Error-tolerant interactive image segmentation using dynamic and iterated graph-cuts. In: proceedings of the ACM international workshop on Interactive multimedia on mobile and portable devices, 2012, pp.9-16.
[22] K. Subr, S. Paris, C. Soler, J. Kautz, Accurate binary image selection from inaccurate user input, Comput. Graphics Forum 32 (2013) 41–50.
[23] G. Song, H. Myeong, K. L. Mu. Seednet: Automatic seed generation with deep reinforcement learning for robust interactive segmentation. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, pp.1760-1768.
[24] L. Wang, Y. Chang, H. Wang, Z. Wu, J. Pu, X. Yang, An active contour model based on local fitted images for image segmentation, Inf. Sci. 418 (2017) 61–73.
[25] T. Wang, Z. Ji, Q. Sun, Q. Chen, X. Jing, Interactive multi-label image segmentation via robust multi-layer graph constraints, IEEE Trans. Multimedia 18 (12) (2016) 2358–2371.
[26] 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.
[27] T. Wang, J. Yang, Z. Ji, Q. Sun, Probabilistic diffusion for interactive image segmentation, IEEE Trans. Image Process. 28 (1) (2019) 330–342.
[28] N. Xu, B. Price, S. Cohen, J. Yang, and T.S. Huang. Deep interactive object selection. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp.373-381.
[29] H. Zhou, J. Zheng, L. Wei, Texture aware image segmentation using graph cuts and active contours, Pattern Recogn. 46 (6) (2012) 1719–1733.