IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 30, 2019, accepted July 31, 2019, date of publication August 8, 2019, date of current version August 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2933876

Region-Based Nonparametric Model for Interactive Image Segmentation

DAN WANG, GUOQING HU, QIANBO LIU, CHENGZHI LYU, AND MD MOJAHIDUL ISLAM
School of Mechanical & Automotive Engineering, South China University of Technology, Guangzhou 510641, China

Corresponding author: Guoqing Hu (gqhu@scut.edu.cn)

This work was supported by the Nature Science Foundation of Guangdong, under Grant 2016A030313520.

ABSTRACT In this paper, we present a novel framework for interactive segmentation problems. It integrates a nonparametric model into the Conditional Random Field (CRF) framework, which can effectively combine high-level features with low-level features to represent image information. In the nonparametric model, multiple region layers are used to estimate data likelihood terms to overcome the bad regions generated by unsupervised methods. The likelihood values of each layer are calculated separately to reduce the computational cost. In addition, we analyze that the pixel layer has little effect on data estimation, so we remove it to further reduce the complexity of the algorithm. We employ the label consistency between pixels and their corresponding regions in smooth term estimation, which can be regard as a higher order potential for pixels. The data term and the smooth term are then performed together in Conditional Random Fields (CRFs) as a fine-tuning of the results. Experimental results show that the proposed method can segment images efficiently and accurately with fewer user inputs.

INDEX TERMS Higher order CRFs, interactive segmentation, region-based model, nonparametric model.

I. INTRODUCTION

Image segmentation is a fundamental task in computer vision and has obtained much attention in recent decades. It can be viewed as a labeling problem that assigning each pixel in the image a predefined label based on its color, intensity and other features. The challenge in segmentation problems is that the boundaries and appearance of natural images are often complex. In these years, various interactive segmentation algorithms have had significant success due to their ability to improve the performance by incorporating user interactions to the segmentation models, such as graph cut [1]–[3], lazy snapping [4], [5], random walker [6]–[8], geodesic segmentation [9], shortest path [10], [11], and deep convolutional networks [12], [13]. In interactive methods, users are allowed to provide a few strokes [1] or a bounding box [14] as initial labels which can help to improve the segmentation results. The primary goal is to extract object boundaries or regions with as little user interaction as possible [15].

CRF framework has been the most widely used framework for segmentation problems. It explicitly models the relationship between pixels and minimizes the energy function defined based on pixels intensities. The basic form of the energy function in CRF framework is to combine data terms with smooth terms. The data term is typically the unary potential obtained from a classifier. Thus, the segmentation performance is heavily influenced by the classifier. Parametric models are widely used as classifiers for data term estimation. Model parameters are treated as variables and optimized together with the energy function [15]. However, parametric models require a large numbers of initial inputs to get accurate segmentation results, most of which are strongly sensitive to the seed quantity and placement. The commonly used smooth term is the pairwise potential which encourages the nearby pixels to have the same label if their intensities are similar. Although pairwise models can be inferred efficiently, they are very limited to express complicated energy formulations as they can only propagate information to the pixels in the local neighborhoods [16]. They are unable to model high-level features which have been shown to be significant powerful for segmentations.

To overcome this problem, [16] proposed higher order potentials to capture high-level structural information of the images. Their algorithms were built on the image regions which had been generated by unsupervised methods. They encouraged the pixels in a region to have the same label. [17] combined a semi-supervised learning technique and region-based models together. They first

The associate editor coordinating the review of this manuscript and approving it for publication was Naveed Akhtar.

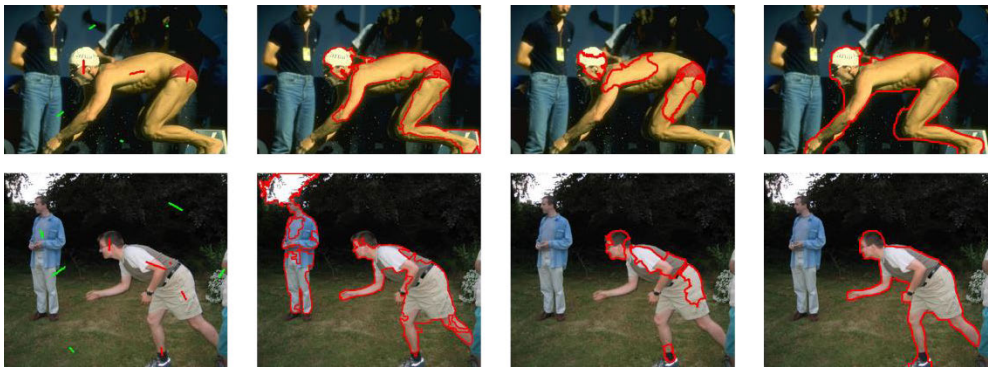


FIGURE 1. Segmentation examples with a few strokes. (a) shows the input images with strokes. (b)-(d) show segmentation results obtained by OneCut, GBMR and our method, respectively.

over-segmented regions and integrated user inputs with regions. Then a k-regular sparse graph was constructed to segment the objects. Therefore, the performance of these algorithms is heavily affected by the regions produced by unsupervised methods.

In this paper we propose a novel interactive segmentation algorithm that can effectively estimate the binary labels based on nonparametric models combining region-level information with pixels-level information. The main contributions of this research are demonstrated as follows:

1. We take advantage of the power of both nonparametric models and higher order CRF models in this paper. Compared with appearance-based models, the proposed method can reduce the computation of parameters and is robust to the quantity and placement of the initial seeds, as shown in Fig. 1(d).

2. Our method reduces the computational complexity of the problem by simple graph construction and can segment images effectively.

3. To overcome bad segmentations and label inconsistency, we employ the label consistency between pixels and their corresponding regions in the smooth term estimation, which can be regard as higher order potentials for pixels.

4. The experiments show that our method is able to produce high quality segmentations in almost real time with only a few user inputs.

The rest of this paper is organized as follows. The related works are introduced in Section II. In Section III, we explain the formulation of energy function and the proposed method in detail. We also discuss how to estimate the data likelihoods and construct our graph. Section IV shows the experimental results compared with the state-of-the-art methods. Finally, the conclusion of this research is stated in section V.

II. RELATE WORKS

The commonly used interactive segmentation approach integrates appearance models and pairwise consistency constraints into energy functions. The two most popular appearance models are the histogram model and the Gaussian Mixture Model (GMM). However, pixel-level appearance models are sensitive to user inputs as they need initial seeds to model foreground and background appearance representations. Some algorithms consider appearance model parameters as extra variables in the optimizations which makes the problem NP-hard [15]. Nonparametric models have proved powerful in segmentation problems because they do not require any basic model to describe the appearance of the image [19]. Therefore, they can accept any unknown data distribution. Márquez-Neila *et al.* [20] presented a nonparametric model for image labelling problems. They designed a patch-based representation for higher order potentials and then convert it to a pairwise form.

Many methods based on higher order potentials have been proposed to model expressive and high-level features of the images. Robust P^n [21] extended the pairwise CRF model to a higher order CRF model by incorporating higher order potentials and reformulating the energy function of GrabCut. They optimized the appearance model and the labelling problem together. OneCut [22] presented L_1 distance in the energy function and used appearance entropy to calculate the likelihood term for segmentation problems. They added auxiliary nodes into the graph to optimize higher order potentials. Regular shaped patches was utilized in [2] as higher order potentials. They considered all the pixels in a patch equally and their results were more robust to overcome noise. Krähenbühl and Koltun *et al.* [23] used densely connected CRFs to represent the remote connections between pixels, which improved the segmentation results. However, as the density of edge connections in the graph increases, so does the computational cost of the algorithm.

The optimization of higher order CRF is computational complex. It is difficult to capture boundary information especially fine structures and weak boundaries. To solve this problem, algorithms combining region-level information with pixel-level information have been developed these years. NHO [24] extracted more expressive information by combining region-level features with pixel-level features. They obtained regions by the unsupervised segmentation method [25] and constructed pairwise relationships between regions and their corresponding pixels. Tao *et al.* [26] also

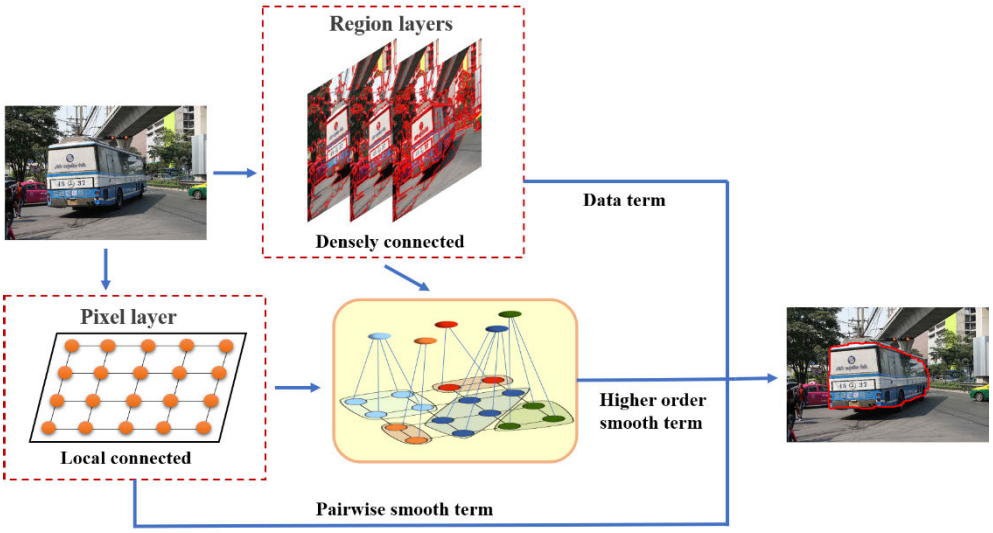


FIGURE 2. The framework of the proposed method.

added label-layer to the multilayer model to improve segmentation results. They employed the parallel partial optimality strategy to optimize the energy function. Although these approaches improve the segmentation accuracy, they are still restricted by the performance of the unsupervised segmentation method used to obtain regions. It is hard to produce satisfied results if regions do not share boundaries with objects. Besides that, the regions they got may not preserve the boundary details which can results in bad performance. [18], [26] used multiple region layers of the image in case some regions were bad. Chen *et al*. [27] proposed a full feature coverage sampling method for image matting and segmentation. They used the edges as clues to search all possible samples of the whole image area. In [28], a local similarity factor which depends on spatial distance and intensity difference was utilized to improve the results. Yu *et al.* [29] presented a novel region-based model for image segmentation. They integrated local patch similarity measure into their model. However, these methods are limited to the computational cost due to the large number of relations between layers. Our algorithm introduces a new approach incorporating region-based likelihood into the CRF models and using global connection between regions which can model long connections within the image. It can reduce the computational cost and get satisfactory performance with less user inputs compared with the state-of-the-art methods.

III. PROPOSED METHOD

We propose a novel method combining the nonparametric data term and the higher order smooth term together into CRFs in this paper. The CRFs are defined on the discrete random variables $x = \{x_i | i \in \{1, 2, \ldots, n\}\}$, where $x_i \in \mathcal{L}$ represents the label of pixel i, $\mathcal{L} = \{0, 1\}$ is the label set with 1 for foreground and 0 for background, n is the number of pixels in the image. The widely used energy function in CRF is defined as follows:

$$\mathrm{E}(\mathrm{x}) = \sum\nolimits_{i\in\Omega} D_i(x_i) + \sum\nolimits_{\{i,j\}\in\mathbb{N}} \psi_{ij}(x_i, x_j) + \sum\nolimits_{\mathrm{c}\in\mathrm{C}} \psi_{\mathrm{c}}(\mathrm{x_c}) \tag{1}$$

where D_i is the data term, ψ_{ij} is the pairwise potential, and ψ_{c} is the higher order potential defined over the clique c. Ω denotes the set of all pixels, $\mathbb{N}$ is the neighborhood defined over the image which is widely chosen to be a 4 neighborhood for local CRFs, and C refers to the set of cliques.

The data term D_i in (1) is known as the penalty of allocating label x_i to the pixel i which is typically formulated by the negative log likelihood as follows:

$$D_i(x_i) = -\log \Pr(x_i) \tag{2}$$

We introduce a nonparametric model to calculate the likelihood term, then use negative log of the probability to represent the data term in (2). The nonparametric model is based on image regions obtained by unsupervised methods. In contrast to other multilayer models [24], [26] which use complex relationships between different layers to improve accuracy, we calculate the probabilities of different layers separately to reduce computational complexity. Dense connections are used in each region layers to model large area information. This is feasible since the number of regions is limited. To overcome the problem of inconsistency between regions and object boundaries, we utilize the label consistency between pixels and their corresponding regions in the smooth term estimation, which can be regard as higher order potentials for pixels. The framework of the proposed method is shown in Fig. 2.

A. REGION LAYERS CONSTRUCTION

Firstly, we generate region layers by cluster algorithms, such as mean shift with different parameters. The set of pixels within the image I is partitioned to $R^t = \{R_k^t | k \in$

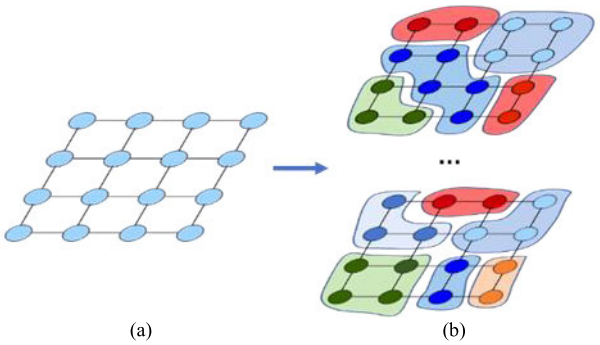


FIGURE 3. Graph for pixel layer and region layer. (a) pixel layer, (b) region layers.

$\{1, 2, \ldots, |R^t|\}$ regions where $t \in \{1, 2, \ldots, m\}$ is the layer index, m is the number of layers we choose. For layer t, R_k^t denotes region k, $|R^t|$ denotes the number of regions, which is related to the parameters we set. Let y_k^t refers to the label of region k, $k \in \{1, 2, \ldots, |R^t|\}$, Ω_k^t is the set of all pixels in region k, so $\Omega = \bigcup_k \Omega_k^t$. The graph construction for region layers is illustrated in Fig. 3.

In the graph of region layers, each node represents a region. In layer t, a region's color is defined as the average color values of its corresponding pixels as follows:

$$\overline{v_k^t} = \frac{1}{|R_k^t|} \sum\nolimits_{i \in \Omega_k^t} v_i^t \tag{3}$$

where $\overline{v_k^t}$ is the average color values of the pixels in this region.v_i^t is the color of pixel i, and $|R_k^t|$ is the pixel number inside region k.

Densely connected CRF is employed in region layers to construct long connection information within the image. It is tractable since the nodes in region layers are quite small. The local CRF with global color models [15] in segmentation problems can be replaced by densely connected CRF [30], so the image features can be obtained without parametric prior models such as histograms or GMM. We use both color information and position information to define the edge weight between region k and region q in layer t. Let β control the relative strength between φ_1 and φ_2, thus the edge weight is given by:

$$\omega_{kq}^{\mathrm{t}} = \beta\varphi_1(k, q) + (1 - \beta)\varphi_2(k, q) \quad k < q \tag{4}$$

where φ_1 is the color-dependent consistency term. It encourages neaby nodes with similar colors to be assigned the same label and is formulated as below:

$$\varphi_1(k, q) = \exp\left(-\frac{\left|\overline{v_k^t} - \overline{v_q^t}\right|^2}{\theta_1} - \frac{|p(k, q)|^2}{\theta_2}\right) \tag{5}$$

φ_2 is the global color model which takes the forms as:

$$\varphi_2(k, q) = \exp\left(-\frac{\left|\overline{v_k^t} - \overline{v_q^t}\right|^2}{\theta_3}\right) \tag{6}$$

Here $p(k, q)$ is the distance between region k and region q. It is calculated by Euclidean distance. For any pixel in region k with location (k_x, k_y) and any pixel in region q with location (q_x, q_y),

$$p(k, q) = \min_{\forall I(k_x, k_y) \in R_k^t,\ I(q_x, q_y) \in R_q^t} \sqrt{(k_x - q_x)^2 + (k_y - q_y)^2}$$

$\theta_1, \theta_2, \theta_3$ are constants that control the relative importance between color similarity and nearness.

B. LIKELIHOOD TERM ESTIMATION

To estimate the likelihood term, a nonparametric model is employed in this paper inspired by [24]. The running cost of multilayer-based methods is normally large because of the complex relationship between different layers. Unlike these multilayer models, we do not consider the relationships between different layers to reduce the complexity of the algorithm. Instead we calculate the likelihood term for each layer and then combine them together rather than using the connections between layers. Now the main computational cost of the proposed method focuses on the pixel layer since the number of nodes in it is quite large. Whereas our experiments show that the pixel layer is not necessary in likelihood estimation because we use pairwise and higher order smooth terms to fine tune the boundary consistency, as Fig.1 and Fig.5 shown. Thus, we propose to deduct it to further reduce the computations and it allows our approach to segment images efficiently.

For layer t, let π_{kl}^t denotes the likelihood of assigning label l to region k. Then we formulate the cost function for the likelihood as follows:

$$J(\pi_l^t) = \frac{1}{2} \sum\nolimits_{k,q}^{|R^t|} \omega_{kq}^{\mathrm{t}} (\pi_{kl}^t - \pi_{ql}^t)^2 + \frac{1}{2} \sum\nolimits_k^{|R^t|} \lambda_k^{\mathrm{t}} (\pi_{kl}^t - \pi_{kl}^{t*})^2 \tag{7}$$

Here π_{kl}^{t*} is the region-seed likelihood. More specifically, for region k in layer t, if there are seeded pixels and all seeded pixels belong to label l, then we set $\pi_{kl}^{t*} = 1$. Otherwise, if there are no seeded pixels in region k or the seeded pixels belong to different labels, then $\pi_{kl}^{t*} = 0$. λ_k^{t} is the parameter for seeded regions. It is set to be λ if region k is a seed and 0 otherwise. The first term in (7) is the consistent constraint which encourages nearby regions to have the same label. The second term is the constraint which assumes that each region tends to have the user-input label.

Equation (7) is reformulated into a matrix form as follows:

$$J(\pi_l^t) = \frac{1}{2} \pi_l^{tT} (D^t - W^{\mathrm{t}}) \pi_l^t + \frac{1}{2} (\pi_l^t - \pi_l^{t*})^T \Lambda^{\mathrm{t}} (\pi_l^t - \pi_l^{t*}) \tag{8}$$

Here $\pi_l^t = [\pi_{kl}^t]_{|R^t| \times 1}$ denotes the matrix form of the region likelihoods. $W^{\mathrm{t}} = [\omega_{kq}^{\mathrm{t}}]_{|R^t| \times |R^t|}$ represent the edge weight matric, $\mathrm{D}^{\mathrm{t}} = diag\left(\left[d_1^t, \ldots, d_{|R^t|}^t\right]\right)$, $d_k^t = \sum_{q=0}^{|R^t|} \omega_{kq}^{\mathrm{t}}$ and $\Lambda^{\mathrm{t}} = diag(\left[\lambda_1^{\mathrm{t}}, \ldots, \lambda_{|R^t|}^{\mathrm{t}}\right])$. To get the likelihood π_l^t that

minimize the cost function J, we differentiate (8) with respect to π_l^t, and set it to zero:

$$\frac{\partial J\left(\pi_l^t\right)}{\partial \pi_l^t} = \left(D^t - W^{\mathrm{t}}\right)\pi_l^t + \Lambda^{\mathrm{t}}\left(\pi_l^t - \pi_l^{t*}\right) = \left(D^t - W^{\mathrm{t}} + \Lambda^{\mathrm{t}}\right)\pi_l^t - \Lambda^{\mathrm{t}}\pi_l^{t*} = 0 \quad (9)$$

Since $B = D^t - W^{\mathrm{t}} + \Lambda^{\mathrm{t}}$ is positive definite, the region likelihood for layer t is given by:

$$\pi_1^t = B^{-1}\Lambda^{\mathrm{t}}\pi_1^{t*} \quad (10)$$

$$\pi_0^t = B^{-1}\Lambda^{\mathrm{t}}\pi_0^{t*} \quad (11)$$

We calculate the probability for each label. In binary segmentations, the probability of y_k^t belong to the foreground and the background are given by:

$$\xi(y_{k1}^t) = \frac{\pi_{k1}}{\sum_l \pi_{kl}} = \frac{\pi_{k1}}{\pi_{k1} + \pi_{k0}} \quad (12)$$

$$\xi(y_{k0}^t) = \frac{\pi_{k0}}{\sum_l \pi_{kl}} = \frac{\pi_{k0}}{\pi_{k1} + \pi_{k0}} \quad (13)$$

Each layer's probabilities have been obtained from (12) and (13), now we combine them layer by layer to get a probability map Q for pixels. A pixel's probability to label l is combined by its corresponding region probability from all layers. It is formulated by $q_{il} = \sum_t \rho_i^t \xi(y_{il}^t)$. The weight $\rho_i^t = \sigma_i^t \Big/ \sum_t \sigma_i^t$ is used to measure the quality of regions obtained by unsupervised algorithms at each layer. For layer t, if pixel i belongs to region k, then σ_i^t is the variance of pixel color values in region k. The data term for the energy function can be set $D(\mathrm{x}) = -\sum_{i\in\Omega} \ln q_i$. Algorithm 1 shows our data term estimation step.

Algorithm 1 Data Term Estimation

Input: Image I, region layer array R, initial seeds x^*
Output: Probability P to pixels in I

1. Set the edge weights for the data term.
2. for i=l: the size of R
3. Construct the D, W, Λ matrices for (8).
4. Calculate the B matric by (9).
5. Compute the probability for each label by (12) and (13).
6. end
7. For each pixel in I, combining its corresponding regions of each layer to estimate its probability.

C. SMOOTH TERM ESTIMATION

In order to fine tune the segmentation results and avoid object inconsistence in regions caused by inaccurate unsupervised region segmentation, we use the smooth term in our model to encourage the label consistency among similar pixels. It is presented by pairwise potentials and higher order potentials in this research for local similarity constraints and global similarity constraints, respectively.

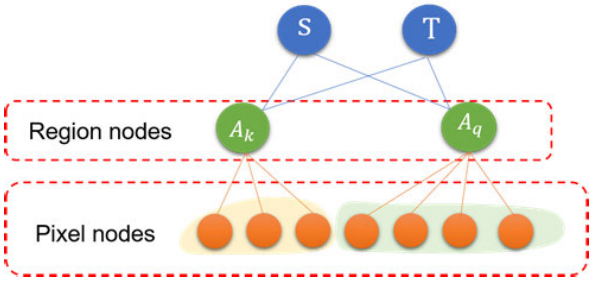


FIGURE 4. Graph construction for higher order potentials.

1) PAIRWISE POTENTIALS

The local similarity constraints measure the penalty of neighboring pixels taking different labels. It helps the image to be partitioned as close as possible to the boundaries. The form of the constraints can be expressed by pairwise potentials as:

$$E_p = \sum\nolimits_{\{i,j\}\in\mathbb{N}} \psi_{ij}\left(x_i, x_j\right) \quad (14)$$

$$\psi_{ij}\left(x_i, x_j\right) = \omega_{ij}^{\mathrm{x}} \left|x_i - x_j\right| \quad (15)$$

where ω_{ij}^{x} is the edge weight between pixel i and pixel j, which is usually obtained by applying a Gaussian kernel to the distance between them. Here we follow [22], [23] to use L_2 based Gaussian kernel $\omega_{ij}^{\mathrm{x}} = \exp(-\frac{d^2(i,j)}{\sigma^2})$, $d(i,j)$ is the Euclidean distance between the colors of pixel i and j, σ denotes the average $d^2\,(i,j)$ over all neighboring pixel pairs in the image. It can penalize discontinuities a lot between pixel i and pixel j when $d(i,j) < \sigma$. However, if pixels are very different, the penalty is small.

2) HIGHER ORDER POTENTIALS

We extend the pairwise smooth model by incorporating the global constraint E_c defined over regions we got in section III-A. E_c is defined over a set of regions to capture higher order cues for pixels. It is converted to a sum of the unary potential over region node c and the pairwise potentials between c and its corresponding pixel i ($i \in \Omega_c$) which encourage pixel i to take the dominant label of region c. In this case, it can be optimized using traditional pairwise CRF algorithms such as graph cut. The form of E_c is given by:

$$E_c = \sum\nolimits_c \varphi_c\left(\mathrm{x}_c\right) \quad (16)$$

$$\varphi_c\left(\mathrm{x}_c\right) = \left(f_u\left(y_c\right) + \sum\nolimits_{i\in\Omega_c} f_p\left(x_i, y_c\right)\right) \quad (17)$$

where f_u refers to the likelihood of region $C, f_p\left(x_i, y_c\right) = \delta\left(x_i \neq y_c\right)$, $\delta(\cdot)$ is an indicator function which takes 1 for true input and 0 otherwise.Ω_c denotes the set of all pixels in region c.

We use auxiliary nodes to construct the graph for higher order potentials of our model. Fig. 4 shows the graph construction. We add auxiliary nodes (such as A_k, A_q) to represent the regions we generated. S, T refer to the label nodes which represent foreground label and background label, respectively. The edges between region nodes and label

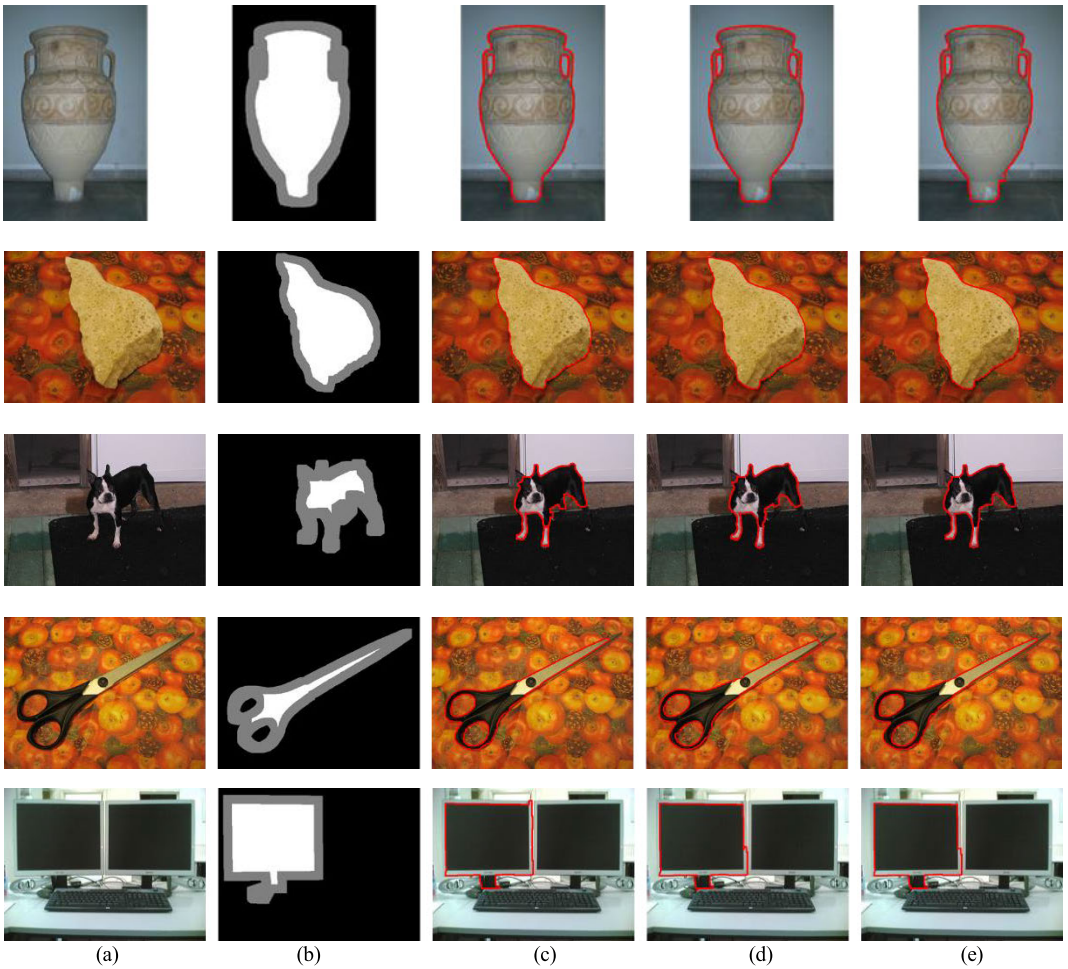


FIGURE 5. Sample results with trimaps using different layers for data term estimation: (a)Input images (b) trimaps (c) combining one pixel layer and three region layers, (d) combining three region layers, and (e) using one region layer.

nodes reflect the unary potentials f_u, which is based on the calculations we discussed in section III-B. Edges between region nodes and their corresponding pixel nodes are utilized to encourage the label consistency. The labelling results can be optimized by the max-flow/min-cut algorithm [31]. Algorithm 2 shows the proposed segmentation method.

Algorithm 2 The Proposed Segmentation Algorithm

Input: Image I.User scribbles S, number of region layers N
Output: Label X of pixels in I

1. Generate seed x* from 5.
2. Construct N region layers R by mean shit
3. Estimate data term from algorithm I based on /, R. x*.
4. Set the edge weights for smooth term.
5. Construct the graph. Local connected edges are used in the pixel layer.
6. Use auxiliary nodes to construct the higher order potentials.
7. Optimize the energy function and get X by the max-flow algorithm.

IV. EXPERIMENTS

We compare our performance with the state-of-the-art methods including OneCut [22], GBMR [17], and NHO [24] in this section. All of these algorithms are implemented based on the public codes provided by the authors. Firstly, region layers were obtained by the mean shift algorithm [25]. The segmentation performance is related with the region quality. For example, smaller regions are able to capture more detailed boundaries while larger regions can obtain more structure features. Thus, we generated three region layers by varying the algorithm's parameters to employ different scales of features. The spatial bandwidth parameter H_s and the range bandwidth parameter H_r in mean shift are set to {(10,7), (10,10), (10,15)} as [24]. For the constant λ which is usually bigger than $max_{i\in\Omega}\sum_{j:\{i,j\}\in\mathbb{N}}\omega_{ij}$, is set to 1000. The values of $\beta = 0.38$, $\theta_1 = 20$, $\theta_2 = 33$ and $\theta_3 = 3$ are empirically set followed by [28] for all experiments in this research.

The experiments were conducted on the GrabCut dataset [15], the Berkeley dataset [32], and the PASCAL VOC dataset [33]. The GrabCut dataset is a commonly

FIGURE 6. Sample segmentations with trimaps: (a)Input images (b) One Cut (c) GBMR, (d) NHO, (e) Ours, and (f) Ground truth.

used benchmark for interactive segmentation problems and consists of 50 images with ground truth masks and trimaps. We also used the test set of the Berkeley dataset which contains 100 single object images. It includes images with similar foreground/ background appearances, complex textures and so on, so it is usually used to represent the challenges in interactive segmentation problems. In addition, we selected 400 images from the PASCAL VOC dataset. All the object classes have been included in our dataset.

A. EFFECTIVENESS OF MULTILAYER RELATIONS

The pixel layer and three region layers are combined for calculating the likelihood term. In order to analyze the impact of these layers on the segmentation results, we used different layers to segment the images. Fig. 5 demonstrates the influence of each layer in likelihood estimation step for pixels. Fig. 5(a) is the input image. Fig. 5(b) is the trimaps. Fig. 5(c) shows the results of combining the pixel layer and three region layers together. Fig. 5(d) gives the results by using the three region layers. Fig. 5(e) shows the results of using only one region layer which was generated by the bandwidth values of (10, 10). It can be found that since we employ pairwise and high order smooth constraints as fine-tuning, the pixel layer containing the most nodes is not necessary for segmentations. This can help us reduce the calculation time of the algorithm. We built two architectures in our experiment. The first one does not contain the pixel layer, which means we only use three region layers to estimate the data term, and the second one combines the pixel layer and the region layers together. The quantitative results provided by these two architectures are shown in Table 1.

B. RESULTS COMPARISONS

To measure the performance of the segmentations, we employ two metrics: error rate and F-measure. The error rate takes the form:

$$\gamma_{error} = \frac{|x_{mis}|}{|x_{unlabel}|} \tag{18}$$

where $|\boldsymbol{x_{mis}}|$ denotes the number of misclassified pixels, $|\boldsymbol{x_{unlabel}}|$ refers to the number of all unclassified pixels.

F-measure is defined as the weighted harmonic mean value of precision and recall:

$$F_\alpha = \frac{(\alpha^2+1)\text{Precision} \times \text{Recall}}{\alpha^2\text{Precision} + \text{Recall}} \tag{19}$$

where α states the relative importance between precision and recall and is set to 0.3 as [22].

TABLE 1. Comparison results of error rate and f-measure among the three datasets.

Method	GrabCut		Berkely		PASCAL VOC	
	$\gamma_{error}(\%)$	F_α	$\gamma_{error}(\%)$	F_α	$\gamma_{error}(\%)$	F_α
OneCut	5 ± 3.8	0.90	6.5 ± 4.8	0.85	7.4 ± 6.2	0.765
GBMR	3.9 ± 3.2	0.91	4.6 ± 4.2	0.867	6.7 ± 5.6	0.784
NHO	2.8 ± 2.1	0.934	3.1 ± 2.8	0.882	4.2 ± 4.1	0.84
Ours (no pixel layer)	2.2 ± 1.9	0.94	2.65 ± 2.5	0.914	3.8 ± 2.7	0.90
Ours	2.12 ± 1.8	0.942	2.5 ± 2.2	0.92	3.6 ± 2.6	0.912

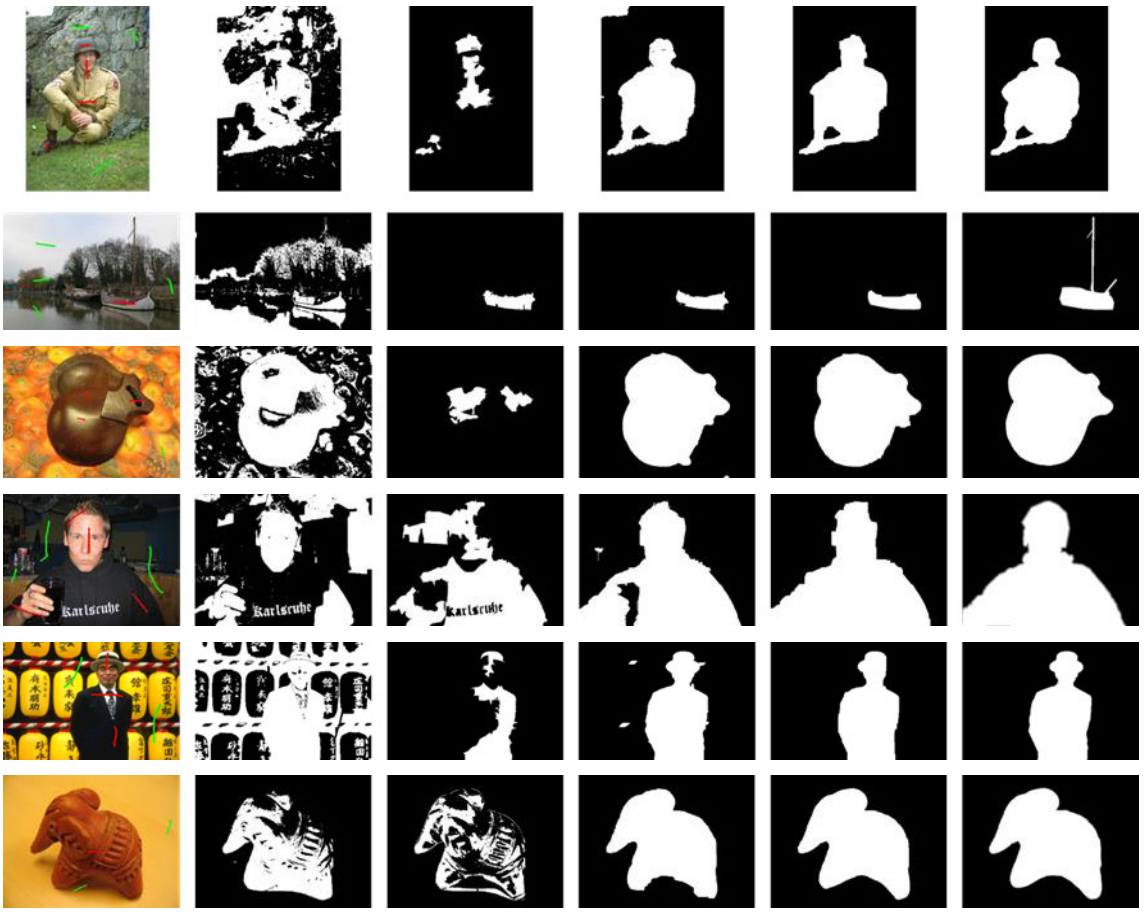


FIGURE 7. Segmentations with a few scribbles: (a) Input images with scribbles, (b) One Cut, (c) GBMR, (d) NHO, (e) Ours, and (f) ground truth.

We got initial seeds from the trimaps. The GrabCut dataset uses its own trimaps, while the Berkeley dataset and the PASCAL VOC dataset use the trimaps generated by dilation and erosion the ground truth masks with the radius of 10 pixels. The quantitative results of different methods are summarized in Table 1. It can be found that the proposed method has more satisfactory performance than OneCut and GBMR. That is because we take advantage of the multiple region layers which can model expressive feature representations of the images. Compared with NHO, which also introduced multilayer-based model, our method provide better results by combining nonparametric model and higher order smooth constrains together. Some segmentation examples achieved by the state-of-the-art methods and the proposed method are given by Fig. 6.

The user inputs play an important role in interactive segmentation problems. For qualitative comparison, we used mouse to draw strokes as the initial seeds. Not like many algorithms utilizing long strokes along object boundaries to get good performances, we only use a few scribbles and it is not necessary to be placed near the boundaries. To give a fair comparison, the seeds for each method are the same. The segmentation results with a few user scribbles are illustrated in Fig. 7. It is observed that OneCut is not able to achieve satisfactory results with a small amount of inputs because it requires enough initial seeds to build the parametric model. The foreground objects in the GBMR method are discontinuous due to the wrong region partition. NHO obtains better performance for using multiple region-level features. However, the results they get sometimes have small ''islands''

TABLE 2. Error rates for different combinations of each part in the proposed method.

Data term	Pairwise smooth	Higher order smooth	γ_{error} (%)		
			GrabCut	Berkely	PASCAL VOC
✓	×	×	2.9	3.3	4.5
✓	✓	×	2.35	2.78	4.05
✓	✓	✓	2.2	2.65	3.8

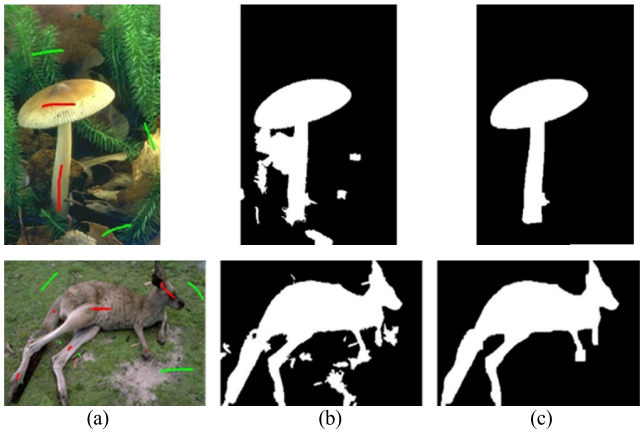


FIGURE 8. Effectiveness of the smooth term. (a) shows the input image with a few scribbles, (b) shows the results without using the smooth term, and (c) shows the results refined by the smooth term.

and the boundaries are inconsistent. The proposed method gets more robust and satisfactory results compared with these approaches. It can be seen that the higher order smooth term help to improve the performance of our method.

C. EFFECTIVENESS OF THE SMOOTH TERM

We combine the nonparametric model with the CRF model to improve the segmentation accuracy. The data term is estimated by the region-based nonparametric model. The smooth term is used to fine-tune the segmentation results and avoid object inconsistence in regions caused by inaccurate unsupervised region segmentation. The pairwise and higher order smooth terms are calculated by (14) and (16), respectively. To justify the proposed architecture, we analyze the effect of each part in our method. We only used region layers to test the results and the initial seeds were generated from the trimaps. The error rate for different combinations of each part is summarized in Table 2. It can be observed that adding smooth term enhances our performance. We get better results when using higher order and pairwise smooth terms together. Fig. 8 gives a visual comparison of the segmentation results to illustrate the effects of the smooth term. Fig.8 (b) and (c) show the results without and with the smooth term. Here we used both higher order and pairwise smooth terms to refine the outputs. As shown in Fig.8 (b), the outputs often have some small isolated regions and the boundaries are coarser. Using smooth terms to refine the results, we obtain more satisfactory results.

D. INITIAL SEEDS SENSITIVITY

Although initial seeds are necessary for the interactive image segmentations, a good algorithm should not be sensitive to the quantity and location of the seeds. We further analyze the proposed method's robustness to user scribbles compared with the state-of-the-art methods. We generated seeds from the trimaps firstly. The positive seeds and negative seeds were randomly sampled from the object region and the background region, accounting for 1% to 100% of the total seed amount, respectively. These seeds can be regarded as the user inputs. Fig. 9 and Fig.10 demonstrate the average error rate and F-measure of our method and other methods at different seed counts, respectively. It can be noticed that OneCut and GBMR are very sensitive to the number of seeds since they have less powerful representations of images. When the number of seeds is less than 40% of the total, the segmentation error is very large. This is because the above algorithms often need significant user interactions to estimate the object distribution. In contrast, we use nonparametric models to calculate likelihood term which can simplifies user inputs to a few scribbles. NHO achieves better results compared with the first two models. The proposed method can obtain the best performance when the number of initial seeds varies. Even with only 10% of the seeds, we can still get satisfactory results. That is because we do not need initial seeds to build the appearance model of the image and the smooth terms help us improve the label consistency along the object boundaries. Therefore, our method can obtain more stable segmentation results, which are not sensitive to the initial seeds.

E. COMPUTATIONAL COST

For running time, we conducted our experiments on a PC with an Intel Core i7 running at 3.7GHz and 16 GB of RAM. Table 3 displays the average running time for different methods on our testing dataset. The region generation time for the proposed method and the NHO method is not included in the running time. It is noted from Table 3 that the NHO algorithms needs the most time, more than 10 seconds, to deal with an image because its graph construction is complicated for modelling the relationship between different layers. Our method (no pixel layer) can segment images in about 0.3 second, which is much faster than the NHO method. That is because instead of using the relationship between the layers, we calculate the likelihoods for each layer separately and then combine them together. In addition, we also find that pixel layer with the most nodes and the highest computational cost has little effect on performance as we use higher order smooth term to refine the results. Therefore, we further remove the pixel layer in the data estimation step, which makes our algorithm more computationally competitive. The proposed method of combining the pixel layer with the region layer

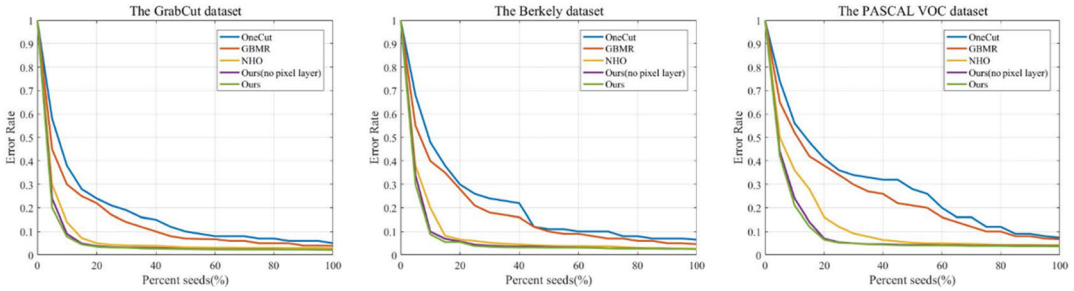


FIGURE 9. The error rate against the percentage of the total number of seeds.

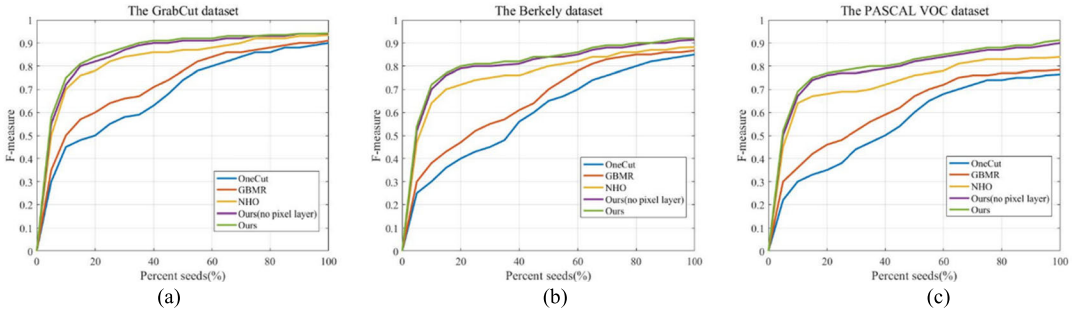


FIGURE 10. The F-measure against the percentage of the total number of seeds.

TABLE 3. Comparison results of running time(s).

Method	GrabCut			Berkely			PASCAL VOC		
	Avg.	Max.	Min.	Avg.	Max.	Min.	Avg.	Max.	Min.
OneCut	0.56	4.32	0.24	0.52	3.75	0.22	0.62	5.84	0.27
GBMR	1.2	7.32	0.52	1.08	8.37	0.46	2.3	8.74	0.62
NHO	9.68	14.8	6.52	9.42	15.7	5.65	9.68	9.68	9.68
Ours (no pixel layer)	0.22	2.34	0.08	0.22	2.78	0.07	0.35	3.2	0.13
Ours	1.32	4.6	0.62	1.34	5.3	0.79	1.62	5.83	0.82

requires about 1.3 to 2 seconds to segment the image, which is similar to the GBMR method but with higher accuracy. When we remove the pixel layer, our method requires the least time to process an image and still achieve satisfactory results.

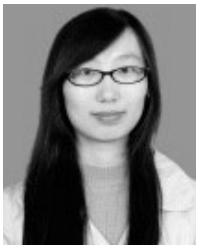
V. CONCLUSION

In this paper, a novel interactive segmentation algorithm incorporating a nonparametric model into the CRF framework is introduced. Multiple region layers are utilized in the nonparametric model for data term estimation to overcome bad regions generated by unsupervised methods and increase label consistency. The densely connected connections are employed in the region layers. Unlike other models that use the relationship between pixel and region layers, we calculate the likelihood of each layer separately to reduce computational cost. At the same time, we analyze that the pixel layer has little effect on the data estimation since we use smooth term to fine tune the results. Therefore we remove the pixel layer to further reduce the complexity of the algorithm. The results show that our algorithm can segment images accurately and efficiently and is robust to the initial seed. The running time of our algorithms is competitive compared to other region-based methods.

REFERENCES

[1] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient N-D image segmentation," *Int. J. Comput. Vis.*, vol. 70, no. 2, pp. 109–131, Nov. 2006.
[2] H. Zhou, J. Zheng, and L. Wei, "Texture aware image segmentation using graph cuts and active contours," *Pattern Recognit.*, vol. 46, no. 6, pp. 1719–1733, Jun. 2013.
[3] S. Jegelka and J. A. Bilmes, "Graph cuts with interacting edge weights: Examples, approximations, and algorithms," *Math. Program.*, vol. 162, nos. 1–2, pp. 241–282, Mar. 2017.
[4] Y. Li, J. Sun, C. Tang, and H. Shum, "Lazy snapping," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 303–308, Aug. 2004.
[5] Y. Li and X. Li, "A background correction method based on lazy snapping," in *Proc. 7th Int. Conf. Image Graph.*, Jul. 2013, pp. 144–148.
[6] L. Grady, "Random walks for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1768–1783, Nov. 2006.
[7] J. Shen, Y. Du, W. Wang, and X. Li, "Lazy random walks for superpixel segmentation," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1451–1462, Apr. 2014.
[8] T. H. Kim, K. M. Lee, and S. U. Lee, "Generative image segmentation using random walks with restart," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 264–275.
[9] X. Bai and G. Sapiro, "Geodesic matting: A framework for fast interactive image and video segmentation and matting," *Int. J. Comput. Vis.*, vol. 82, no. 2, pp. 113–132, 2009.

[10] J. Mille, S. Bougleux, and L. D. Cohen, "Combination of piecewise-geodesic paths for interactive segmentation," *Int. J. Comput. Vis.*, vol. 112, no. 1, pp. 1–22, Mar. 2015.
[11] F. Meng, H. Li, G. Liu, and K. N. Ngan, "Object co-segmentation based on shortest path algorithm and saliency model," *IEEE Trans. Multimedia*, vol. 14, no. 5, pp. 1429–1441, Oct. 2012.
[12] J. Liew, Y. Wei, X. Wei, S.-H. Ong, and J. Feng, "Regional interactive image segmentation networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2746–2754.
[13] Y. Hu, A. Soltoggio, R. Lock, and S. Carter, "A fully convolutional two-stream fusion network for interactive image segmentation," *Neural Netw.*, vol. 109, pp. 31–42, Jan. 2018.
[14] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp, "Image segmentation with a bounding box prior," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 277–284.
[15] C. Rother, V. Kolmogorov, and A. Blake, ""GrabCut": Interactive foreground extraction using iterated graph cuts," in *Proc. ACM SIGGRAPH*, Aug. 2004, pp. 309–314.
[16] P. Kohli, L. Ladický, and P. H. S. Torr, "Robust higher order potentials for enforcing label consistency," *Int. J. Comput. Vis.*, vol. 82, no. 3, pp. 302–324, May 2009.
[17] H. Li, W. Wu, and E. Wu, "Robust interactive image segmentation via graph-based manifold ranking," *Comput. Vis. Media*, vol. 1, no. 3, pp. 183–195, 2015.
[18] T. Wang, Q. Sun, Z. Ji, Q. Chen, and P. Fu, "Multi-layer graph constraints for interactive image segmentation via game theory," *Pattern Recognit.*, vol. 55, pp. 28–44, Jul. 2016.
[19] E. Erdil, M. U. Ghani, L. Rada, A. O. Argunsah, D. Unay, T. Tasdizen, and M. Cetin, "Nonparametric joint shape and feature priors for image segmentation," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5312–5323, Nov. 2017.
[20] P. Márquez-Neila, P. Kohli, C. Rother, and L. Baumela, "Non-parametric higher-order random fields for image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2014.
[21] S. Vicente, V. Kolmogorov, and C. Rother, "Joint optimization of segmentation and appearance models," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 755–762.
[22] M. Tang, L. Gorelick, O. Veksler, and Y. Boykov, "GrabCut in one cut," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1769–1776.
[23] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," 2012, *arXiv:1210.5644*. [Online]. Available: https://arxiv.org/abs/1210.5644
[24] T. H. Kim, K. M. Lee, and S. U. Lee, "Nonparametric higher-order learning for interactive segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3201–3208.
[25] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.
[26] T. Wang, Z. Ji, Q. Sun, Q. Chen, and X.-Y. Jing, "Interactive multilabel image segmentation via robust multilayer graph constraints," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2358–2371, Dec. 2016.
[27] X. Chen, F. He, and H. Yu, "A matting method based on full feature coverage," *Multimedia Tools Appl.*, vol. 78, no. 9, pp. 11173–11201, May 2019.
[28] S. Niu, Q. Chen, L. de Sisternes, Z. Ji, Z. Zhou, and D. L. Rubin, "Robust noise region-based active contour model via local similarity factor for image segmentation," *Pattern Recognit.*, vol. 61, pp. 104–119, Jan. 2017.
[29] H. Yu, F. He, and Y. Pan, "A novel region-based active contour model via local patch similarity measure for image segmentation," *Multimedia Tools Appl.*, vol. 77, no. 18, pp. 24097–24119, Sep. 2018.
[30] M. M. Cheng, V. A. Prisacariu, S. Zheng, P. H. S. Torr, and C. Rother, "DenseCut: Densely connected CRFs for realtime GrabCut," *Comput. Graph. Forum*, vol. 34, no. 7, pp. 193–201, 2015.
[31] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," presented at the Proc. 3rd Int. Workshop Energy Minimization Methods Comput. Vis. Pattern Recognit., 2001.
[32] K. Mcguinness and N. E. O'Connor, "A comparative evaluation of interactive segmentation algorithms," *Pattern Recognit.*, vol. 43, no. 2, pp. 434–444, Feb. 2010.
[33] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2009. doi: 10.1007/s11263-009-0275-4.

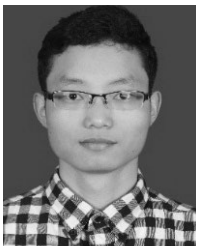


DAN WANG received the master's degree in mechanical engineering from Xiamen University, Xiamen. She is currently pursuing the Ph.D. degree with the South China University of Technology, Guangzhou, Guangdong, China. Her research interests include computer vision, and developing automated control systems based on image understanding and researching the machine vision-based mechanical automated processing and inspection systems.

GUOQING HU received the M.S. degree from Northwestern Polytechnical University, China, and the Ph.D. degree from Sichuan University, China.

He was a Professor with Xiamen University, China. He was an Advanced Visiting Scholar with The Chinese University of Hong Kong and the University of Nottingham. He is currently a Professor and a Ph.D. Student Supervisor with the School of Mechanical and Automotive Engineering, South China University of Technology, China. He completed and participated in more than 90 projects including the National 863 project, the National Natural Science Foundation Project, the National Major Projects, International Cooperation Projects, Provincial Key Projects, and the Province Fund Cooperation Projects. He was published 248 papers, 22 patents and two textbooks. His research interests include amphibious flying machine, intelligent robot, industrial image processing, automation and industrial robot, electromechanical integration, and advanced sensor technology.

QIANBO LIU received the M.S. degree from the South China University of Technology, China, in 2004, where he is currently pursuing the Ph.D. degree with the School of Mechanical and Automotive Engineering. His research interests include visual tracking, deep learning, and object detection.



CHENGZHI LYU received the B.E. degree in electromechanical engineering from the Hubei University of Technology, Wuhan, Hubei, in 2013. He is currently pursuing the Ph.D. degree with the South China University of Technology, Guangzhou, Guangdong, China. His research interests include artificial intelligent algorithms and machine automation.

MD MOJAHIDUL ISLAM received the B.Sc. and M.Sc. degrees in computer science and engineering from Islamic University, Bangladesh. He is currently pursuing the Ph.D. degree with the School of Mechanical and Automotive Engineering, South China University of Technology, China. From 2010 to 2015, he was an Assistant Professor with the Department of Computer Science and Engineering, Islamic University, Bangladesh. His research interests include computer vision, object tracking, pattern recognition, multimedia analysis, and machine learning.