IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 10, 2019, accepted November 29, 2019, date of publication December 4, 2019,
date of current version December 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2957504

Foreground Extraction Combining Graph Cut and Histogram Shape Analysis

KUN HE[1], DAN WANG[1], BIN WANG[1], BEN FENG[1], AND CHENYU LI[2]
[1]School of Computer Science, Sichuan University, Chengdu 610065, China
[2]School of Statistics and Mathematics, Nanjing Audit University, Nanjing 211899, China

Corresponding author: Dan Wang (535459443@qq.com)

This work was supported by the Sichuan Province Science and Technology Support Program of China under Grant 2016JZ0014.

ABSTRACT This paper presents a foreground extraction model which develops the Grab Cut model by applying the histogram shape analysis method. In this model, the foreground extraction is formulated as an inference problem based on edges and appearance models. The inference is solved via a minimum cut/maximum flow algorithm scheme, which allows incorporation of edge information and automatic tuning of parameters in appearance models. We use the histogram shape analysis method to analyze the intensity distribution of an image, and estimate the optimal number of regions in order to best model appearances of the foreground and background. The appearance models are defined as a maximum likelihood form instead of the original Gaussian Mixture Models in order to distinguish the small regions in an image. Numerical experimental results on the Berkeley segmentation database and Weizmann horse's database indicate that, compared to existing foreground extraction models, the proposed model provides comparable performance in terms of segmentation metric and computational cost, while being insensitive to the small region in an image.

INDEX TERMS Foreground extraction, inference, histogram shape analysis, and maximum likelihood.

I. INTRODUCTION

The foreground, user's interested objects, provides useful information for image analysis and comprehension. Foreground extraction is a task of distinguishing between the specific object and background [1]. Existing methods rely on one or more low-level features, such as the intensity distributions [2], edges, and region connectivity [3], where the overall aim is to achieve an accurate object with minimal user interaction. Owning to the ambiguity of features for segmentation, foreground extraction is still a challenging task in the computer vision community.

Among existing foreground extraction techniques, the methods based on the graph cut can effectively extract the foreground according to edges and appearance models [4]. An appearance model is a statistical model for the intensity/color distributions. The existing appearance models are broadly categorized into the local histogram and the Gaussian Mixture Models (GMMs), the former can explicitly represent the intensity /color distributions of the user-labeled pixels [5], [6]; however, the estimation accuracy of appearance parameters varies with the user interaction. The latter makes use of the all pixels to estimate appearance parameters in a better way [7]–[9]. The accuracy of the GMMs depends on i) the number of Gaussian in each GMM; and ii) the representation form. The original GMM is defined as the weight sum of Gaussians, which cannot effectively distinguish the small regions in the foreground and background [10].

In this work, we propose a novel appearance model which is formulated as a maximum likelihood rather than the weight sum of Gaussians. In this appearance model, the optimal number of Gaussians is estimated by the histogram shape analysis method, in which the number is automatically adjusted according to the intensity distributions of an image. Combining edges and appearance models, the foreground extraction is formulated as a joint optimization for the foreground extraction and appearance parameters. The proposed model is experimentally shown to compare favorably with contemporary foreground extraction models.

This paper is organized as follows: In Section 2, we review and discuss related foreground extraction methods based on the graph cut and machine learning. In Section 3, the proposed model is detailed. Finally, the experimental results and conclusions are presented in Section 4 and 5, respectively.

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

II. RELATION WORKS

The definition of foreground in an image varies with the individual cognition. For different users, the foreground in an image is different. To extract the foreground, it is necessary to specify the foreground with additional information [11], such as the user interaction or the prior information. The Magic Wand [12] is a simpler foreground extraction technique based on the user interaction, and computes a set of pixels that are similarity with the user-labeled pixels. The method can successfully extract foreground from cartoon images, however, the performance is poor for images with intensity overlapping regions between the foreground and background [13].

Boykov and Jolly [5] propose a generative Markov random field model for the foreground extraction. The foreground extraction is formulated as a graph partitioning problem according to edges and appearance models. The existing appearance models fall into the local histogram [5], [6] and the GMMs [7]–[9]. The local histogram explicitly describes the intensity distributions of the user-labeled foreground and background. The foreground extraction models based on the local histogram, such as the Graph Cut [5] and One Cut [6], often produce good results for cartoon images. However, the local histogram only represents distributions of local pixels, and the accuracy of the appearance models depends on the user input information.

The GMM, a parametric representing form of the appearance model, is formulated as the weight sum of several Gaussians. In the appearance model, assuming that an image consists of homogeneous regions, color distributions of each region may be approximated as one Gaussian. Given the number of regions in the foreground and background, appearances of the foreground and background are formulated as the GMMs with the fixed number of Gaussians [7]. Compared to the local histogram, this model makes use of all pixels of an image to learn appearance parameters, and achieves good results with little user interaction. For real images, the accuracy of this model depends on i) the inhomogeneity, such as texture; ii) the number of regions in the foreground and background; and iii) the representation form. For i), the inhomogeneity leads to the poor estimation accuracy of appearance parameter. To remove the negative effect of inhomogeneity, the Super Cut improves the estimated precision of appearance parameters by introducing super-pixel appearances [8]. For ii), The improved Grab Cut [9] uses an unsupervised algorithm [13] to analyze the foreground and background pixels and estimates the optimal number of regions. For iii), Heimowitz and Keller formulate the foreground extraction as a probabilistic inference problem [14]. The aim is to estimate the marginal assignment probabilities of image pixels by the Kullback-Liebler divergence between the super-pixel distribution and the appearance models.

The performance of foreground extraction based on the graph cut varies with features such as the appearance differences between the foreground and the background. The ambiguity of appearance differences, such as the color overlapping regions between the foreground and background, leads to incorrect results [15]. The fully convolution networks (FCNs) [16] can automatically extract features by deep learning [17] on the training set of the specific foreground. Adding an extra information (the user interaction) as the input of a convolutional neural network [18], [19], the foreground is extracted. Compared to the methods using the *artificial* features such as edges and intensity/color distributions, the performances of the FCNs are competitive for extracting subordinate-level categories that may appear in the training set, such as handwritten numeral recognition [20], and traffic congestion identification [21]. However, those are poor for extracting an object that is not a member of the training set.

III. THE FOREGROUND EXTRACTION MODEL

According to edges and appearances, we propose a novel foreground extraction model which combines the histogram shape analysis and the graph cut into a unified model. In this model, the histogram shape analysis method is used to analyze the intensity distributions of an image, and estimate the optimal number of regions in the foreground and background. Given an initial boundary box, an image $\boldsymbol{u}$ with pixels $N = W \times H$ is divided into a background T_B and a foreground T_F where there are some pixels of background. The unknown foreground mask is expressed as an array of variables $\boldsymbol{x} = (x_1, \cdots x_i, \cdots x_N)$, where $x_i \in \{0, 1\}$, with 0 for the foreground and 1 for the background. This foreground extraction task is to infer the unknown variables according to the appearances and edges, and an energy functional for the foreground extraction is formulated as the following form:

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x},\,\boldsymbol{\omega}} \left\{ U(\boldsymbol{x}, \boldsymbol{\omega}, \boldsymbol{u}) + V(\boldsymbol{x}, \boldsymbol{u}) \right\}. \tag{1}$$

Here $\boldsymbol{\omega}$ denotes appearance parameters that describe color distributions of the foreground and background. In (1), the term $U(\boldsymbol{x}, \boldsymbol{\omega}, \boldsymbol{u})$ evaluates the fitness of the variables $\boldsymbol{x}$ to the image $\boldsymbol{u}$ according to the given appearance parameters, and the term $V(\boldsymbol{x}, \boldsymbol{u})$ measures the fitness of extracted foreground boundaries to edges. When boundaries of the foreground locate at the high gradients, the term $V(\boldsymbol{x}, \boldsymbol{u})$ reaches minimum, which defined as

$$V(\boldsymbol{x}, \boldsymbol{u}) = \sum_{i=1}^{N} \sum_{(i,j)\in\Lambda_i} \frac{\gamma\left[x_i \neq x_j\right]}{dis(i,j)} \exp(-\frac{1}{\lambda}\left\|\boldsymbol{u}_i - \boldsymbol{u}_j\right\|_2^2). \tag{2}$$

Here Λ_i is a set of pixel-pairs. According to the spatial continuity of boundaries of the foreground, the set is composed of the pixel i and adjacent pixels of 8-way connectivity such as the horizontal, vertical, and diagonal directions. The factor $dis(\bullet)$, the Euclidean distance of a pixel-pair, helps the term $V(\boldsymbol{x}, \boldsymbol{u})$ to approximate a geometric length of the foreground boundaries. The constant γ is set as 30 in order to relax the tendency to the high gradient. To ensure that the exponential term in (2) can appropriately switch between the high and low gradients, the constant λ is defined as:

$$\lambda = \frac{2}{M} \sum_{i=1}^{N} \sum_{(i,j)\in\Lambda_i} \left\|\boldsymbol{u}_i - \boldsymbol{u}_j\right\|_2^2 \tag{3}$$

where M denotes the number of pixel pairs over an image, and it is computed as

$$M = 4W \times H - 3(W + H) + 2. \tag{4}$$

A. AN OPTIMIZATION APPEARANCE MODEL

The extracted foreground, an optimal result of (1), depends on not only image edges, but also the differences of appearances between the foreground and background. Assuming that an image consists of the homogeneous regions, the color distributions of each region are compact and may be approximated by a Gaussian [7]. Take the m$^{-\text{th}}$ region for example:

$$G(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m, \boldsymbol{y}) = \frac{\exp\left(-0.5(\boldsymbol{y} - \boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_m)\right)}{\sqrt{(2\pi)^3 \det(\boldsymbol{\Sigma}_m)}}. \tag{5}$$

Here $\boldsymbol{y} = (y_R, y_G, y_B)$ denotes the color value of each pixel in the region; $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$, respectively denoting the mean and covariance matrix, describe the statistical information of the region. Known the number of regions in the foreground and background (K_0 and K_1), we represent the appearances of the foreground and background as the mixture Gaussian models, and parameters are:

$$\boldsymbol{\omega} = \{\boldsymbol{\omega}^0, \boldsymbol{\omega}^1\}. \tag{6}$$

Here $\boldsymbol{\omega}^0$ and $\boldsymbol{\omega}^1$ denote the appearance parameters of the foreground and background, respectively. They comprise the variables:

$$\begin{cases} \boldsymbol{\omega}^0 = \{K_0, \pi_m^0, \boldsymbol{\mu}_m^0, \boldsymbol{\Sigma}_m^0, m = 1 \cdots K_0\} \\ \boldsymbol{\omega}^1 = \{K_1, \pi_m^1, \boldsymbol{\mu}_m^1, \boldsymbol{\Sigma}_m^1, m = 1 \cdots K_1\} \end{cases} \tag{7}$$

Here π_m, a weight coefficient, is the fraction of pixels in the m$^{-\text{th}}$ region which are assigned to the foreground (or background).

In an image, the foreground and background may be composed of one or more regions. The unsuitable number of regions (K_0 and K_1) leads to negative effects on the appearance parameters estimation [9]. For a cartoon image, the intensities of a region could be clustered around a certain intensity-level, and the histogram shape exhibits more than one peak where each peak corresponds to a region. Thus, the optimal number of regions may be estimated by analyzing the histogram shape.

For an image with inhomogeneity such as the texture, the histogram $h(z)$, where z denotes the intensity level of an image, has many local maxima which form the pseudo-peaks. To remove pseudo-peaks, we use the median filtering to smooth the histogram. The smoothed histogram $\bar{h}_s(z)$ is defined as:

$$\bar{h}_s(z_i) = median\left\{h(z_{i-s/2}), \cdots, h(z_i), \cdots, h(z_{i+s/2})\right\}. \tag{8}$$

Here s is the sampling scale of the median filter.

To estimate the number of peaks, we analyze the smoothed histogram shape and apply the sign of differences of the smoothed histogram $\bar{h}_s(z)$ to detect valleys. The sign of difference of $\bar{h}_s(z)$ is defined as

$$\hat{h}_s(z) = \begin{cases} -1, & \delta(\bar{h}_s(z)) < 0 \\ 0, & \delta(\bar{h}_s(z)) = 0 \\ 1, & \delta(\bar{h}_s(z)) > 0 \end{cases} \tag{9}$$

Here $\delta(\bullet)$ denotes the center difference operator. With consideration for zeros, the valley v_i is defined as:

$$v_i = \begin{cases} z_i, & if\ \hat{h}_s(z_{i-1}) = -1\ and\ \hat{h}_s(z_{i+1}) = 1 \\ \dfrac{z_j + z_k}{2}, & if\ \hat{h}_s(z_j) = -1,\ \hat{h}_s(z_k) = 1, \\ & \hat{h}_s(z_l) = 0, and\ l, i \in (j, k) \end{cases} \tag{10}$$

Given valleys $\boldsymbol{v} = \{v_0, \cdots, v_i, \cdots, v_K\}$, the image is divided into K regions, and the intensity distributions of each region correspond to one peak in the smoothed histogram.

Since the foreground and background are subsets of an image, the number of the foreground and background (K_0 and K_1) is estimated by combining K and the initial boundary box, and it satisfies $K_0 \le K$, and $K_1 \le K$. Given $\boldsymbol{\omega}^0$ and $\boldsymbol{\omega}^1$, the term $U(\boldsymbol{x}, \boldsymbol{\omega}, \boldsymbol{u})$ in (1) is formulated for all pixels in T_F:

$$U(\boldsymbol{x}, \boldsymbol{\omega}, \boldsymbol{u}) = -\sum_{i \in T_F} \left(\log L_F(\boldsymbol{x}, \boldsymbol{\omega}^0, \boldsymbol{u}_i) + \log L_B(\boldsymbol{x}, \boldsymbol{\omega}^1, \boldsymbol{u}_i)\right). \tag{11}$$

Here $L_F(\boldsymbol{x}, \boldsymbol{\omega}^0, \boldsymbol{u}_i)$ and $L_B(\boldsymbol{x}, \boldsymbol{\omega}^1, \boldsymbol{u}_i)$ denotes the maximum likelihood that $\boldsymbol{u}_i$ respectively belongs to the foreground and background, which are defined as

$$\begin{cases} L_F(\boldsymbol{x}, \boldsymbol{\omega}^0, \boldsymbol{u}_i) = \max\{\pi_m^0 G(\mu_m^0, \Sigma_m^0, \boldsymbol{u}_i), m = 0, \cdots, K_0\} \\ L_B(\boldsymbol{x}, \boldsymbol{\omega}^1, \boldsymbol{u}_i) = \max\{\pi_m^1 G(\mu_m^1, \Sigma_m^1, \boldsymbol{u}_i), m = 0, \cdots, K_1\}. \end{cases} \tag{12}$$

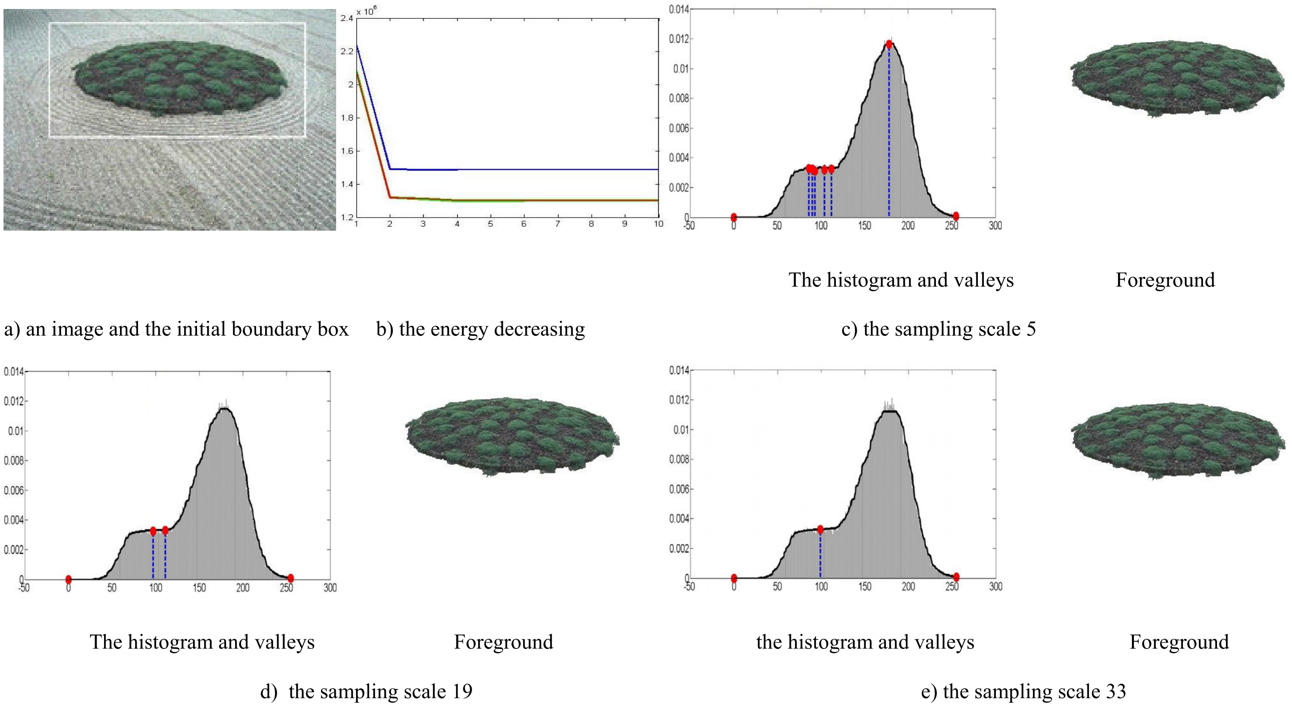
The (11) has the same form as that in [7]–[9]. The difference is that it depends on the maximum rather than the weighted sum of Gaussians.

B. THE FOREGROUND EXTRACTION ALGORITHM

The proposed model is formulated as a joint optimization for the foreground extraction and appearance parameters, where the extraction result relies on appearance parameters. In practice, appearance parameters are unknown before the foreground extraction. The optimal result of (1) is achieved by the alternate optimization for $\boldsymbol{x}$ and $\boldsymbol{\omega}$, so (1) is written as:

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}}\{\min_{\boldsymbol{\omega}}(U(\boldsymbol{x}, \boldsymbol{\omega}, \boldsymbol{u}) + V(\boldsymbol{x}, \boldsymbol{u}))\}. \tag{13}$$

The term $V(\boldsymbol{x}, \boldsymbol{u})$ in (13), varying with gradients, is computed once and reuse. Owing to the presence of a part of background pixels in the foreground, the term $U(\boldsymbol{x}, \boldsymbol{\omega}, \boldsymbol{u})$ should be updated during the segmentation process. Thus, the optimal result of (13) is obtained by iteratively performing the following operations: i) Given appearance parameters



The histogram and valleys

Foreground

a) an image and the initial boundary box

b) the energy decreasing

c) the sampling scale 5

The histogram and valleys

Foreground

the histogram and valleys

Foreground

d) the sampling scale 19

e) the sampling scale 33

FIGURE 1. Foreground extraction process for the simple scene image. a) the original image and the initial boundary box, b) the energy of this model decreases over 10 iterations (the blue, green, and red curves denote the energy decreasing with the sampling size of the medial filter 5, 19, and 33, respectively), c-e) Foreground and the histogram (the gray area denotes the histogram, the black curve the smoothed histogram, and red dots valleys) using a median filter with sampling sizes: 5, 19, and 33.

$\boldsymbol{\omega}$, the local optimal result $\boldsymbol{x}$ can be found by the standard minimum cut/ maximum flow algorithm [22]; ii) Given the current $\boldsymbol{x}$, $\boldsymbol{\omega}$ are updated. The appearance parameters $\boldsymbol{\omega}$ comprises the number of regions and the statistical information of each region. The number of regions in the foreground and background is estimated by combining the current $\boldsymbol{x}$ and the number of regions in an image, and the mean and covariance matrix of each region are recalculated using an expectation-maximization (EM)-style procedure. In each iteration, some background pixels in the foreground will be correctly classified, the estimation accuracy of appearance parameters is improved step by step. So it is guaranteed not to increase the energy of (13). The above process repeats until convergence. In practice, we simply stop after five iterations.

The pseudocode for this algorithm is shown as

```
*Require*: $T_F$ and $T_B$ using bounding box.
1: Initialize $\boldsymbol{x}$, $x_i = 0$ for $i \in T_F$ and $x_i = 1$ for $i \in T_B$.
2: Compute the term $V(\boldsymbol{x}, \boldsymbol{u})$ using the formula (2).
3. Estimate the number of regions $K$ in an image by the histogram shape analysis.
4: N: = 1
    5: Repeat
    6: Update $\boldsymbol{x}$, given the current $\boldsymbol{\omega}$ using the standard minimum cut/ maximum flow algorithm.
    7: Estimate $K_0$ and $K_1$, combining $K$ and the current $\boldsymbol{x}$.
    8: Update $\boldsymbol{\omega}$ , given current $\boldsymbol{x}$ using EM.
    9: Compute the term $U(\boldsymbol{x}, \boldsymbol{\omega}, \boldsymbol{u})$ using the formula (11).
10: Until: N>5
11: Output the foreground $T_F$.
```

IV. NUMERICAL EXPERIMENTS

The experiments of this study were conducted using VC 6.0 on a PC with Intel-Core i7CPU @ 3.40 GHz and 4 GB of RAM without any particular code optimization. We used two widely segmentation metrics: the intersection over union (IOU) metric [23] and F-measure. The latter was computed by the following:

$$F - measure = \frac{2 \times precision \times recall}{precision + recall}. \tag{14}$$

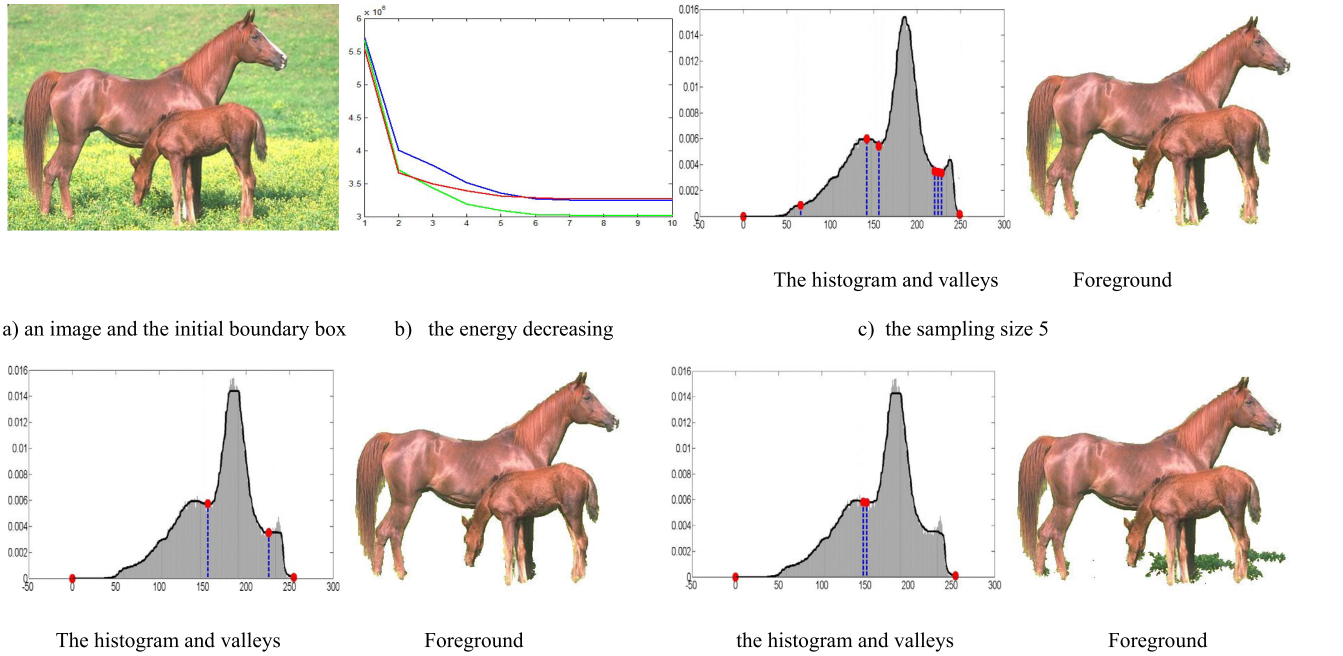
where

$$precision = \frac{F(s) \cap F(g)}{F(g)}, recall = \frac{F(s) \cap F(g)}{F(s)}. \tag{15}$$

Here $F(s)$ and $F(g)$ denote the extracted foreground and the ground truth, respectively.

A. PARAMETER DISCUSSION

The performance of the proposed model varied with appearance parameters, which depended on the number of regions in the foreground and background. To estimate the optimal number of regions, we used the histogram shape analysis method to analyze the shape of the intensity histogram. For an image, there are pseudo-peaks in the histogram because of the texture. To remove the pseudo-peaks, we used the median filtering to smooth the histogram, and analyzed the shape of the smoothed histogram. The extraction results for the simple and complex scene images, with the different sampling sizes of the median filtering, were shown in the Fig.1 and 2, respectively.



a) an image and the initial boundary box b) the energy decreasing c) the sampling size 5

d) the sampling size 19 e) the sampling size 23

FIGURE 2. **Foreground extraction process for the complex scene image. a) the original image and the initial boundary box, b) the energy of this model decreases over 10 iterations (the blue, green, and red curves denote the energy decreasing with the sampling size of the medial filter 5, 19, and 23, respectively), c-e) the foreground and the histogram (the gray area denotes the histogram, the black curve the smoothed histogram, and red dots valleys) using a median filter with sampling sizes: 5, 19, and 23.**

TABLE 1. **Segmentation metrics with different sampling sizes on Fig.1 and 2 (Iterator 5).**

images	*sampling size*	*Precision*	*Recall*	*F-Measure*	IOU
Fig.1	5	0.987	0.979	0.983	0.967
	19	0.993	0.978	0.985	0.971
	33	0.990	0.979	0.984	0.970
Fig.2	5	0.961	0.971	0.966	0.935
	19	0.982	0.963	0.972	0.946
	23	0.931	0.966	0.948	0.901

TABLE 2. **Segmentation metrics on images for the BSD300 and the weizmann horse's database.**

methods	*F-measure*		IOU	
	range	mean	range	mean
The Berkeley segmentation database (BSD300)				
This model	0.615~0.989	0.886	0.518~0.978	0.823
Super Cut	0.603~0.985	0.841	0.453~0.975	0.796
Improved Grab Cut	0.554~0.985	0.796	0.403~0.960	0.736
Grab Cut	0.412~0.980	0.682	0.319~0.935	0.618
Li model	0.237~0.872	0.401	0.201~0.802	0.337
The Weizmann horse's database				
This model	0.604~0.985	0.815	0.542~0.973	0.776
Super Cut	0.558~0.975	0.779	0.403~0.951	0.706
Improved Grab Cut	0.502~0.975	0.692	0.334~0.950	0.647
Grab Cut	0.362~0.960	0.574	0.246~0.920	0.539
Li model	0.188~0.853	0.384	0.161~0.724	0.301

A simple scene image consisted of homogeneous regions, and the histogram shape exhibited multi-model with a few pseudo-peaks. The number of peaks in the histogram changed slightly with the sampling size of the median filter. The extracted foreground was insensitive to the sampling size, shown as the Fig.1. However, textures in a complex scene led to many pseudo-peaks, shown as the Fig.2. For a small sampling size, the remnant pseudo-peaks caused the poor generalization of appearance parameters because of the over-fitting. The extracted foreground contained some background pixels, shown as the Fig.2c). If the histogram was smoothed with a larger sampling size, some valleys were removed. It led that the accuracy of appearance parameters was poor because the color distributions of some regions were wide, shown as Fig.2e).

The segmentation metrics on the Fig.1 and 2, with different sampling sizes, were listed in Table 1. For a small sampling size, an image was over-divided because of pseudo-peaks. The appearance differences among regions were minor, and the segmentation metrics were lower. For a large sampling size, an image was under-divided, and accuracy of appearance parameters was poor. In this work, the sampling size of the median filter was set as 19 by evaluated against the ground truth over 150 images.

B. COMPARISON AND ANALYSIS

To evaluate the foreground extraction performance, experiments were conducted to compare this model with comparable models based on user interaction, such as the graph

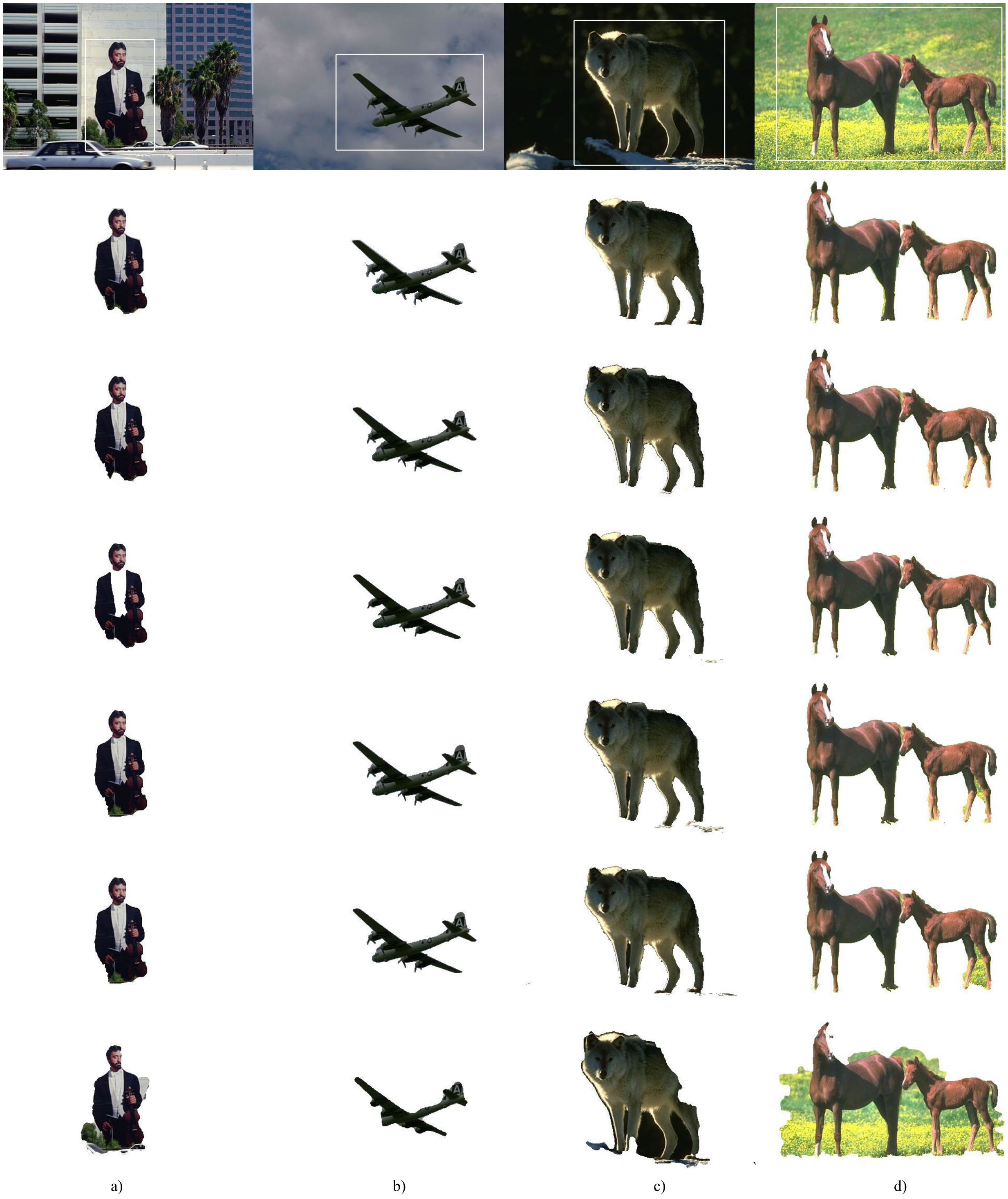


FIGURE 3. Comparison of the proposed model with the Grab Cut, the Super Cut, the improved Grab Cut and the Li model on the partial images in the BSD 300. Row 1: original images and initial bounding box, Row 2: the ground truth, Row 3: this method, Row 4: the Super Cut, Row 5: the improved Grab Cut, Row 6: the Grab Cut, Row 7: the Li model.

cut [7]–[9] and the active contour. The Li model [24], one of the active contour models, extracted the foreground from images by the level set method within the constraint of edges. In the Li model, the Gaussian filtering with $\sigma = 1.5$ was used to smooth inhomogeneity, and removed the negative effects of inhomogeneity on edges. The models based on the graph cut extracted the foreground according to edges and appearances, such as the Grab Cut [7], the Super Cut [8], and the improved Grab Cut [9]. In those models, appearances of the foreground and background were represented as the weight sum of serval Gaussians. The latter two models further extended the Grab Cut model by the super-pixel appearances

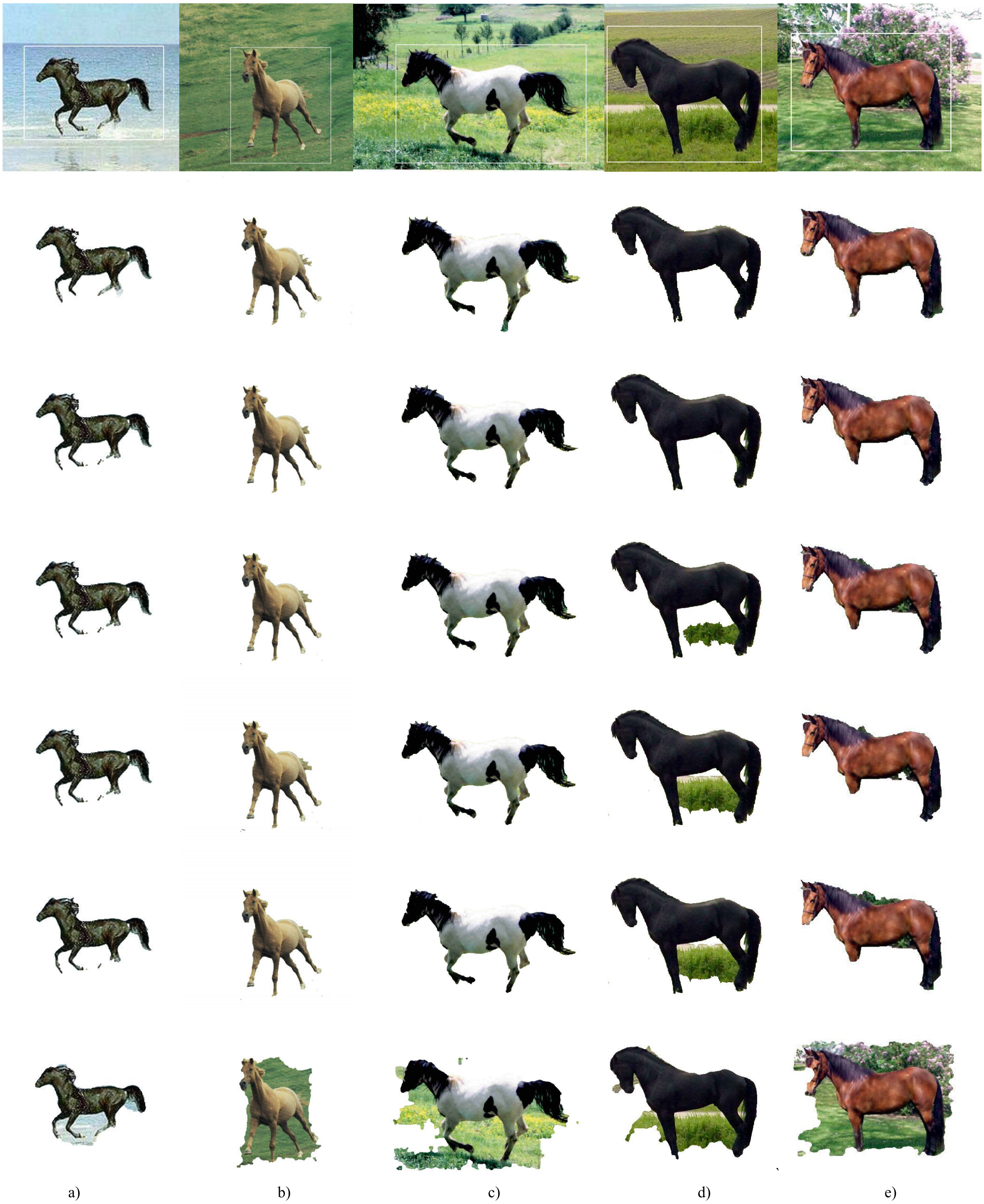


FIGURE 4. Comparison of the proposed model with the Grab Cut, the Super Cut, the improved Grab Cut and the Li model on the partial images in the Weizmann horse's database. Row 1: original images and initial bounding box, Row 2: the ground truth, Row 3: this method, Row 4: the Super Cut, Row 5: the improved Grab Cut, Row 6: the Grab Cut, Row 7: the Li model.

and optimizing the number of Gaussian in the appearance models, respectively. The Super Cut used the super-pixels of an image to remove the negative effect of inhomogeneity, and the improved Grab Cut applied the CLUSTER optimization technique to estimate the optimal number of regions, and well modeled the foreground and background.

In this work, the tested images arrived from the Berkeley segmentation database (BSD300) and the Weizmann horse's database. The BSD300 included 300 images and manually labeled boundaries of the foreground, and this database was divided into a training set containing 200 images and a test set including 100 images. The Weizmann horse's database

TABLE 3. The segmentation metrics and computational cost on the Fig.3 and 4.

Segmentation methods	*Fig .3a*	*Fig .3b*	*Fig.3c*	*Fig.3d*	*Fig.4a*	*Fig .4b*	*Fig.4c*	*Fig .4d*	*Fig .4e*
	480×320	480×320	480×320	480×320	376×358	800×771	752×503	448×434	400×330
This model									
Precision	0.980	0.996	0.973	0.981	0.948	0.927	0.985	0.987	0.973
Recall	0.960	0.936	0.986	0.958	0.958	0.976	0.951	0.984	0.947
F-Measure	0.970	0.966	0.980	0.969	0.953	0.951	0.968	0.985	0.959
IOU	0.942	0.934	0.960	0.941	0.910	0.905	0.937	0.971	0.922
Computational cost (s)	2.993	2.169	3.284	3.237	2.912	9.125	8.934	2.845	2.412
The Super Cut [8]									
Precision	0.975	0.972	0.977	0.971	0.951	0.925	0.985	0.865	0.957
Recall	0.960	0.909	0.981	0.959	0.954	0.975	0.950	0.990	0.945
F-Measure	0.968	0.939	0.979	0.965	0.952	0.949	0.967	0.923	0.951
IOU	0.938	0.885	0.959	0.932	0.909	0.904	0.935	0.858	0.906
Computational cost (s)	7. 923	6.645	6. 931	6.582	7. 185	14.83	16.45	7.719	7.657
The improved Grab Cut [9]									
Precision	0.969	0.972	0.961	0.992	0.945	0.923	0.986	0.780	0.969
Recall	0.963	0.906	0.998	0.929	0.958	0.979	0.945	0.991	0.919
F-Measure	0.965	0.938	0.979	0.960	0.952	0.951	0.965	0.873	0.943
IOU	0.933	0.883	0.959	0.922	0.908	0.905	0.932	0.775	0.893
CPU-time(s)	6.984	5.216	7.433	7.524	7.381	12.39	13.28	6.536	8.312
The Grab Cut [7]									
Precision	0.968	0.986	0.974	0.946	0.960	0.925	0.992	0.778	0.929
Recall	0.962	0.866	0.978	0.959	0.940	0.976	0.921	0.990	0.941
F-Measure	0.966	0.922	0.976	0.953	0.950	0.950	0.955	0.871	0.935
IOU	0.934	0.855	0.952	0.910	0.904	0.904	0.914	0.772	0.878
Computational cost (s)	4.397	4.273	5.237	6.178	5.835	10.03	11.66	6.115	8.954
The Li Model[24]									
Precision	0.791	0.987	0.763	0.511	0.706	0.442	0.512	0.712	0.451
Recall	0.944	0.670	0.861	0.941	0.936	0.987	0.945	0.935	0.973
F-Measure	0.861	0.799	0.809	0.662	0.805	0.610	0.664	0.809	0.616
IOU	0.756	0.665	0.679	0.495	0.673	0.439	0.497	0.679	0.446
Computational cost (s)	12.43	10.67	13.69	15.18	13.38	19.49	19.12	13.98	10.67

included 328 color images on horses and corresponding ground truth. The segmentation metrics on both databases were listed in Table 2. The IOU and *F-measure* indicated that the performance of the proposed model was superior to that of the other models.

The partial results of both databases were shown in the Fig. 3 and 4, respectively. For simpler images where there were the significant differences among regions and approximated to homogeneity within a region, the results using the proposed model were visually similar to those by the graph cut [7]–[9], shown as the Fig. 3a), b), Fig. 4a), and b). However, the proposed model was competitive for images with substantial inhomogeneity. In this work, we extended the Grab Cut from the following: i) the histogram shape analysis method was used to analyze the intensity distributions and estimate the optimal number of regions in an image, instead of the fixed number. Compared to the improved Grab Cut [9], the number of regions was estimated from an image rather than the foreground and background. It improved the accuracy of appearance models and the foreground extraction performance, shown as the Fig.3 c), d) and Fig.4 c). ii) The appearances of the foreground and background were formulated as a maximum likelihood form rather than the GMMs [7]–[9], which was favor to distinguish the small regions. The performance was better than that of the Super Cut, shown as the Fig.4d).

The performances of the foreground extraction models should not only vary with the models themselves, but also depend on features of an image. The Li model [24], assuming that boundaries of the foreground were smoothing, can extract the foreground by the level set method within the constraint of edges, such as the Fig.3 a). In the model, edges in the background led that the level set converged to the local minimum. The performances of the proposed model, combining edges and appearances, were better than those of the Li model.

The computational cost and segmentation metrics, on images in the Fig. 3 and 4, were listed in Table 3. For images with homogeneity, the IOU and F-measure using the models on the graph cut were similar. For images with substantial inhomogeneity, the segmentation metrics using the proposed model were higher than those of the other models. In this model, the histogram shape analysis method was used to estimate the number of regions according to the intensity distribution. Among regions, there were significant differences in the intensity but not the color space. Compared to the ground truth, the performance of the proposed model was poor for images with a color overlapping region between the foreground and background, shown as in the Fig. 4e). The computational cost of this model was lower than those of the other models. The reason is that the number of regions was computed once rather than iterations of the K-means clustering [25].

V. CONCLUSION

In this paper, we present two modifications of the Grab Cut to improve foreground extraction performance. The histogram shape analysis method is shown to improve the segmentation performance by removing the negative effect of the unsuitable number of regions. In addition, appearances of the foreground and background are formulated as a maximum likelihood form rather than the GMMs, which helps to extract the foreground with the small regions. However, appearance parameters are estimated in the RGB color space without the visual perception of R, G, and B. To improve performance, visual perception appearance models will be introduced to extract the foreground.

REFERENCES

[1] I. N. Junejo and N. Ahmed, "Foreground extraction for freely moving RGBD cameras," *IET Comput. Vis.*, vol. 12, no. 3, pp. 322–331, Apr. 2018.

[2] P. P. Tunga and V. Singh, "Extraction and description of tumour region from the brain MRI image using segmentation techniques," in *Proc. IEEE Int. Conf. Recent Trends Electron., Inf. Commun. Technol. (RTEICT)*, May 2017, pp. 1571–1576.

[3] R. Hettiarachchi and J. F. Peters, "Voronoï region-based adaptive unsupervised color image segmentation," *Pattern Recognit.*, vol. 65, pp. 119–135, May 2017.

[4] B. Peng, L. Zhang, and D. Zhang, "A survey of graph theoretical approaches to image segmentation," *Pattern Recognit.*, vol. 46, no. 3, pp. 1020–1038, Mar. 2013.

[5] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Jul. 2001, pp. 105–112.

[6] M. Tang, L. Gorelick, O. Veksler, and Y. Boykov, "GrabCut in one cut," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2013, pp. 1769–1776.

[7] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.

[8] S. Wu, M. Nakao, and T. Matsuda, "SuperCut: Superpixel based foreground extraction with loose bounding boxes in one cutting," *IEEE Signal Process. Lett.*, vol. 24, no. 12, pp. 1803–1807, Dec. 2017.

[9] D. Chen, B. Chen, G. Mamic, C. Fookes, and S. Sridharan, "Improved GrabCut segmentation via GMM optimisation," in *Proc. Digit. Image Comput., Techn. Appl.*, Dec. 2008, pp. 39–45.

[10] A. Heimowitz and Y. Keller, "Image segmentation via probabilistic graph matching," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4743–4752, Oct. 2016.

[11] R. Hebbalaguppe, K. McGuinness, J. Kuklyte, G. Healy, N. O'Connor, and A. Smeaton, "How interaction methods affect image segmentation: User experience in the task," in *Proc. 1st IEEE Workshop User-Centered Comput. Vis. (UCCV)*, Jan. 2013, pp. 19–24.

[12] *Adobe Photoshop User Guide*, Adobe Syst., San Jose, CA, USA, 2002.

[13] C. A. Bouman, "Cluster: An unsupervised algorithm for modeling Gaussian mixtures," USA. Accessed: Aug. 16, 2015. [Online]. Available: https://engineering.purdue.edu/~bouman/software/cluster/manual.pdf

[14] C. G. Bampis, P. Maragos, and A. C. Bovik, "Graph-driven diffusion and random walk schemes for image segmentation," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 35–50, Jan. 2016.

[15] F. L. Yi and I. Moon, "Image segmentation: A survey of graph-cut methods," in *Proc. Int. Conf. Syst. Inform.*, May 2012, pp. 1936–1941.

[16] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[17] V. Shreyas and V. Pankajakshan, "A deep learning architecture for brain tumor segmentation in MRI images," in *Proc. IEEE 19th Int. Workshop Multimedia Signal Process.*, Oct. 2017, pp. 1–6.

[18] N. Xu, B. Prics, S. Cohen, J. Yang, and T. Huang, "Deep GrabCut for object selection," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2017, pp. 1–12.

[19] K.-K. Maninis, S. Caelles, J. Pont-Tuset, and L. Van Gool, "Deep extreme cut: From extreme points to object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 616–625.

[20] L. Guo and S. Ding, "A hybrid deep learning CNN-ELM model and its application in handwritten numeral recognition," *J. Comput. Inf. Syst.*, vol. 11, no. 7, pp. 2673–2680, 2015.

[21] H. Cui, Y. Liu, X. Song, and L. Pannong, "Traffic image congestion identification based on CNN deep learning model," *Technol. Innov. Appl.*, vol. 4, no. 224, pp. 19–22, 2018.

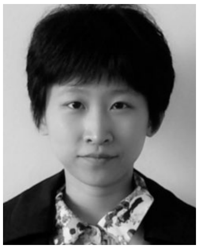
[22] V. Kolmogorov and R. Zabin, "What energy functions can be minimized via graph cuts?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 147–159, Feb. 2004.

[23] J. Pont-Tuset and F. Marques, "Measures and meta-measures for the supervised evaluation of image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2131–2138.

[24] C. Li, C. Xu, C. Gui, and M. D. Fox, "Level set evolution without re-initialization: A new variational formulation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 430–436.

[25] S. H. Kang, B. Sandberg, and A. M. Yip, "A regularized k-means and multiphase scale segmentation," *Inverse Problems Imag.*, vol. 5, no. 2, pp. 407–429, May 2011.

KUN HE received the Ph.D. degree in electrical and computer engineering from Sichuan University, in 2006. Since 2006, he has been a Professor Research Fellow with the School of Computer Science, Sichuan University. His research interests include pattern recognition, computer vision, and image processing.



DAN WANG received the bachelor's degree in software engineering from Sichuan University, in 2014, and the master's degree from the Key National Defense Laboratory of Visual Synthesis Graphic and Image, in 2017. She is currently pursuing the Ph.D. degree with the Computer Science Department, Sichuan University. Her major work was pattern recognition, image processing, and medical image analysis.

BIN WANG received the bachelor's degree in computer science and technology from Sichuan University, in 2018. He is currently pursuing the master's degree with the School of Computer Science, Sichuan University. His research interests include image processing, pattern recognition, and video processing.

BEN FENG is currently pursuing the bachelor's degree in computer science with Sichuan University. His research interests include image processing, pattern recognition, and computer vision.

CHENYU LI is currently pursuing the bachelor's degree with the School of Statistics and Mathematics, Nanjing Audit University. His research interests include the application of mathematical statistics and uncertainty inference probability.