

Interactive stereo image segmentation via adaptive prior selection

Wei Ma¹ · Yue Qin¹ · Shibiao Xu² · Xiaopeng Zhang²

Received: 28 June 2017 / Revised: 14 March 2018 / Accepted: 29 April 2018
© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract Interactive stereo image segmentation (i.e., cutting out objects from stereo pairs with limited user assistance) is an important research topic in computer vision. Given a pair of images, users mark a few foreground/background pixels, based on which prior models are formulated for labeling unknown pixels. Note that color priors might not help if the marked foreground and background have similar colors. However, integrating multiple types of priors, e.g., color and disparity in segmenting stereo pairs, is not trivial. This is because differing pairs of images and even differing pixels in the same image might require different proportions of the priors. Besides, disparities of natural images are too noisy to be directly used. This paper presents a method that can adaptively determine the proportion of the priors (color or disparity) for each pixel. Specifically speaking, the segmentation problem is defined in the framework of MRF (Markov Random Field). We formulate an MRF energy function which is composed of clues from the two types of priors, as well as neighborhood smoothness and stereo correspondence constraints. The weights of the color and disparity priors at each pixel are treated as variables which are optimized together with the label (foreground or background) of the pixel. In order to overcome the noise problem, the weight of the disparity prior is controlled by a confidence value learned from data. The energy

✉ Shibiao Xu
shibiao.xu@ia.ac.cn

✉ Xiaopeng Zhang
xiaopeng.zhang@ia.ac.cn

¹ Faculty of Information Technology, Beijing University of Technology, No. 100 Pingleyuan Street, Chaoyang District, Beijing 100124, China

² National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

function is optimized by using multi-label graph cut. Experimental results show that our method performs well.

Keywords Stereo image segmentation · Interactive segmentation · Prior selection · Multi-label MRF · Graph cut

1 Introduction

Stereo cameras are widely used in many fields, e.g., 3D movies and Virtual Reality (VR). Therefore, various research topics on stereo image editing appear in recent years [11, 14, 15, 22], including interactive stereo segmentation [8, 16, 20], i.e., cutting out foreground objects from a pair of stereo images with limited user assistance. Specifically, users mark a few foreground and background pixels in either of the two views with strokes in different colors, respectively. A segmentation algorithm is then employed to separate foreground and background in the pair, simultaneously, based on user-provided prior clues. If the users are not satisfied with the results, more interaction could be involved to provide more clues. The interaction might be replaced by automatic localization of foreground objects, with assumption that salient parts belong to foreground [1]. However, it is hard to automatically generate satisfactory results [19] in natural images. In this paper, we focus on interactive segmentation of stereo images.

One key problem in interactive stereo segmentation is how to utilize user-marked pixels as priors to label unmarked ones as foreground or background. The prior models are various but usually defined in color spaces in traditional methods. For example, Ju et al. [8] adopted color GMM (Gaussian Mixture Model) to model foreground/background priors. Ma et al. [17] used clusters in the RGB color space to represent the priors. However, color priors might not help if the foreground and background have similar colors while the visual appearances of the foreground/background in different viewpoints might be different [28]. Since depth/disparity is computable from stereo images, we can formulate two types of priors in color and depth/disparity spaces, respectively. Combining color and disparity priors in stereo image segmentation has been tried in [16]. However, this attempt [16] adopted a fixed proportion of the constraints from color and disparity priors. In this paper, we present a method with adaptive proportions. Using multiple types of clues with adaptive weights has been proved essential in other fields, e.g., image retrieval [24, 27].

We hold that during stereo segmentation, different pairs of images and even different pixels in the same image need specially weighted priors. We exemplify the argument in Fig. 1 with three synthetic pairs of stereo images and their depth maps. For simplicity, only single color and depth views of each pair are given. In these images, the triangles are ground-truth foreground. In order to perform segmentation, users mark some foreground and background in the color views with red and blue strokes, respectively. The strokes in the color views are automatically propagated to the disparity maps. The color and disparity prior models of foreground and background for each pair are built by referring to the user marked pixels. Since the marked foreground/background pixels in each pair share the same color or disparity value in this illustration, the priors could be defined as single colors or disparity values, which are visualized in the circles on the top of Fig. 1. We then classify those user-unmarked pixels (e.g., A, B, C, and D in Fig. 1a, b and c) as foreground or background, based on the similarities between these pixels and their prior models in color and disparity. With traditional methods [16], the weights of the color and disparity ingredients in the

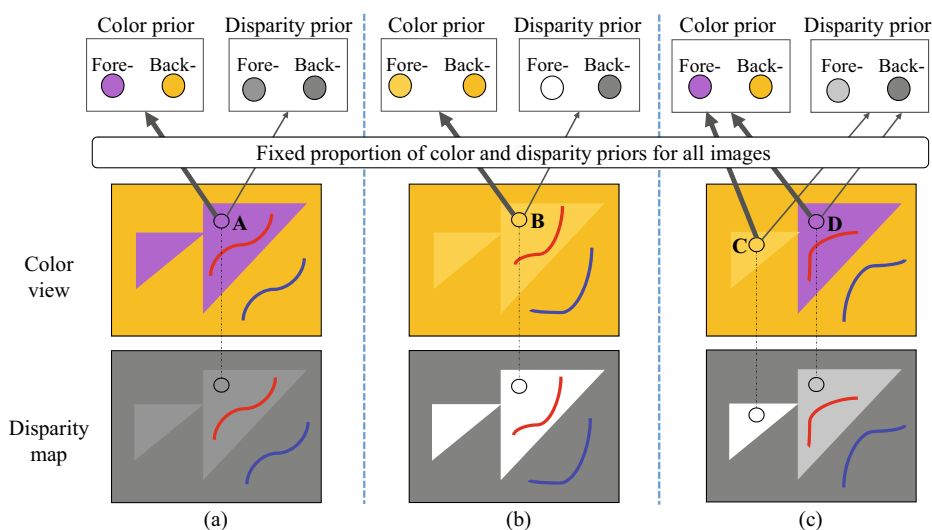


Fig. 1 Illustration on segmentation with a fixed proportion of color and disparity priors. (a), (b) and (c) are three synthetic images to be segmented and their disparity maps. The triangles in the images are ground-truth foreground. The priors, modeled as single colors and disparity values in this illustration, are determined by pixels covered by user input strokes in each color view, red for foreground and blue for background. The strokes in the color views are automatically propagated to the disparity maps. A, B, C, and D are unmarked pixels. The width of the gray arrows linking the images to the priors represents the proportions of color and disparity ingredients in the similarity metric between unmarked pixels and the prior models. The proportions are fixed for all images and even pixels in the same image in this illustration

similarity metric are fixed for labeling every unknown pixels in all image pairs, as illustrated by the width of the arrows linking the images to the priors in Fig. 1. In this example, color is weighted more than disparity. This proportion is good for pixel A, since color priors are more helpful than disparity in classifying A as its ground-truth label, i.e., foreground. However, the situation is totally inverse for pixel B. Moreover, even pixels in the same image, e.g., C and D with ground-truth labels of foreground in Fig. 1c, require different types of priors for correct decision. This is because D is closer to the foreground prior in color space, while C is in disparity dimension as seen from Fig. 1. Therefore, it is essential to adaptively select priors for different images and even different pixels in the same image during segmentation.

In view of the above analysis, we propose a segmentation method that can select prior constraints adaptively for each pixel. The method treats the weights of the two types of priors at each pixel as variables, which are optimized together with the label (foreground or background) of the pixel by using multi-label graph cut. Meanwhile, considering that the given disparities of natural stereo images are noisy, a data-learned threshold is adopted to restrict the confidence on disparity. In the experiments, we validate the effectiveness of the disparity confidence control and demonstrate the good performance of the proposed method in segmentation of stereo images.

The contributions of the paper include: 1) We present an argument that the proportions of different types of priors for classifying different pixels should be adaptive during interactive stereo image segmentation; 2) A novel algorithm is proposed to improve the accuracy of interactive stereo segmentation by adaptively weighting color and disparity priors for each

pixel. The argument and the proposed algorithm are proved via experimental comparison (in Section 4.4) between our method and the baseline method in [16] which used a fixed proportion of color and disparity priors; 3) A confidence control on disparity priors is introduced and experimentally validated (in Section 4.3) to avoid the influences from noises; 4) Experiments, including error rate analysis on two evaluation metrics, are given in Section 4 to compare our method with a range of recent state-of-the-art methods.

The arrangement of the remaining is as follows. In Section 2, we review existing literatures related to our work. Then, our method is described in detail in Section 3. Section 4 presents experiments to validate the proposed method. Conclusions are given in Section 5.

2 Related work

In the following, we introduce the state-of-the-arts in interactive stereo image segmentation. Most of these methods modeled priors only in color spaces. However, depth/disparity is considered helpful and has been widely used in segmentation of RGBD or depth images. Therefore, we will also review the utilization of depth/disparity in segmentation and analyze the roles of depth/disparity in stereo segmentation.

Interactive stereo image segmentation Interactive segmentation of single images has been relatively well studied in the last two decades [2, 5, 7, 18, 19, 25]. In recent years, along with the popularity of stereo media in the field of 3D TV/movie, VR, etc., interactive segmentation of stereo images became a hot research topic. Many popular frameworks for single image segmentation (e.g., snakes [10] and graph cut [3]) have been adjusted to be applied to segmentation of stereo images. For example, Ju et al. [9] proposed StereoSnakes by integrating consistency between stereo contours in the framework of snakes. In 2011, Price et al. [20] successfully integrated color priors with neighborhood smoothness and stereo correspondence constraints in MRF (Markov Random Field) and sought labels of unknown pixels via binary graph cut. Many subsequent methods were proposed for better segmentation performances within Price's framework [8, 16, 17]. For example, Ju et al. [8] introduced the iterative graph cut optimization process from GrabCut [21] into the framework. Ma et al. [17] replaced dense correspondences with sparse ones to speed up the segmentation.

Although a lot of methods have been developed, many issues remain open in interactive stereo segmentation. A specific one is the utilization of user-provided foreground/background clues. Users generally mark few pixels as foreground and background. The attributes of the foreground and background could be modeled by these pixels and then used for labeling unknown pixels. Most existing methods for interactive stereo segmentation modeled the prior attributes in color spaces [8, 17, 20].

Utilization of depth/disparity in segmentation Depth/disparity is experimentally proved effective in segmentation of depth or RGBD images [23, 26]. A direct and typical example is the successive separation of foreground from background using only depth values in Kinect applications [23]. Moreover, Zhang et al. [26] used depth to form silhouette consistency for joint multi-view segmentation of RGBD images. Feng et al. [6] exploited depth information and normals derived from the depth to help interactive segmentation of RGBD images.

However, full exploitation of depth/disparity clues in interactive stereo image segmentation remains an open issue. Most related work simply used disparity [8, 17, 20] to correlate

left and right views for simultaneous and consistent stereo segmentation. Depth/disparity was also involved in modeling the priors of foreground and background [16], together with RGB color clues. However, in this method, the authors fixed the proportion of the two types of priors during segmentation of all stereo pairs, which is not reasonable as we analyzed in the introduction part. Besides, disparity maps are generally obtained by stereo matching which is not well solved and suffered from noises. In this paper, we make the prior selection adaptive and solve the noise problem.

3 Method

We follow the general pipeline of MRF-based interactive segmentation [5, 16, 18, 20]. Given a pair of images, users draw few strokes to indicate parts of foreground and background. The interaction way could be replaced by the others, e.g., marking trimaps [2]. Next, the prior models are built by the labeled pixels. Then, an energy function with prior clues and constraints of smoothness and consistency is formulated. Finally, the energy function is optimized by using graph cut to obtain the label of each unknown pixel, foreground or background [5]. If the users are not content with the results, more interaction could be involved to enrich the prior clues and trigger a new cycle of update and optimization of the energy function. In the following, we will explain the definition of our prior models. Then, we present the formulation and optimization of the energy function that could select priors adaptively for each pixel, while determining its foreground/background label. For easy understanding of the proposed method, we describe all involved notations in Table 1.

3.1 Prior models

The prior models of foreground/background are built by user marked pixels in a given pair of stereo images. Instead of using only color clues as done in most of traditional methods [8, 17, 20], we use both color and disparity for priors. In implementation, we use k-means algorithm to build clusters of the foreground/background pixels in RGB and disparity spaces, respectively. The two sets of cluster centers, as shown in Fig. 2 (red for foreground and blue for background), are defined as the prior models of foreground and background, respectively. For clear illustration, we horizontally stretch the disparity clusters which should be distributed on the disparity axis. During interactive segmentation, once a new stroke is added to mark more foreground/background pixels, the foreground/background prior model would be updated by k-means clustering based on the extended set of pixels.

Given a new pixel to be classified as foreground or background, its distance to the nearest foreground/background clusters in RGB and disparity spaces are considered important clues for classification. As we mentioned before, RGB and disparity clues play different roles in classifying different unknown pixels. For example, to label the unknown pixel whose attributes are visualized in Fig. 2 (the stars), color should be chosen over disparity, since the differences between its distances to foreground and background priors are more distinct in the RGB space than in the disparity dimension. In the following, we explain how to compute the probability of each pixel belonging to foreground and background by its distances to the priors in the attribute spaces. The attribute selection is done by energy minimization in next section. Note that the work in [16], which used a fixed proportion of the color and disparity clues for segmentation, is our baseline method. For easy explanation of the novel parts in this paper, parts of notations or equations in the following are similar with [16].

Table 1 Notation table

I	A pair of stereo images.
I^l	Left view of I .
I^r	Right view of I .
p_i	A pixel in I .
p_i^l	A pixel in I^l .
p_i^r	A pixel in I^r .
x_i	Label(foreground or background) of pixel p_i .
c_i	Color of pixel p_i .
C^f	Set of foreground cluster centers in color space.
C^b	Set of background cluster centers in color space.
df_i^c	The minimum distance from p_i to C^f .
db_i^c	The minimum distance from p_i to C^b .
$P_D^c(.,.)$	Probability that a pixel belongs to foreground/background in color space.
$P_D^d(.,.)$	Probability that a pixel belongs to foreground/background in disparity dimension.
N_B	Set of four-neighborhood pixels in I .
N_C	Set of corresponding pixels in I .
$f_D(.,.,.)$	Data term encoding clues from priors in MRF energy function.
$f_B(.,.)$	Term of smoothness constraints in MRF energy function.
$f_C(.,.)$	Term of consistency constraints in MRF energy function.
λ_B	Weight of $f_B(.,.)$
λ_C	Weight of $f_C(.,.)$.
$f_D^c(.,.,.)$	Clues from color priors in $f_D(.,.,.)$.
$f_D^d(.,.,.)$	Clues from disparity priors in $f_D(.,.,.)$.
w_i^c	Weight of $f_D^c(.,.,.)$.
w_i^d	Weight of $f_D^d(.,.,.)$.
$f_B^c(.,.)$	Term of color gradient in $f_B(.,.)$.
$f_B^d(.,.)$	Term of disparity variance in $f_B(.,.)$.
λ^c	Weight of $f_B^c(.,.)$.
λ^d	Weight of $f_B^d(.,.)$.
$S(.,.)$	Similarity between two pixels.

We use $I = \{I^l, I^r\}$ to represent the pair of stereo images, where I^l and I^r are the left and right views, respectively. p_i denotes a pixel in I . $x_i \in \{1, 0\}$ (foreground and background) is the label of p_i .

The minimum distances from p_i to the foreground and background priors in RGB color space are given by,

$$\begin{aligned} df_i^c &= \min(\|c_i - C_j^f\|^2), \quad j = 1, \dots, k \\ db_i^c &= \min(\|c_i - C_j^b\|^2), \quad j = 1, \dots, k \end{aligned} \quad (1)$$

Here, c_i denotes the RGB color of p_i . $\{C^f\}$ and $\{C^b\}$ are the foreground and background color cluster centers. k is the number of clusters in the prior models. As done in [13, 17], $k = 64$ in our experiments. In the color space, the probabilities that p_i belongs to foreground and background are determined by its normalized distances to the priors, which are given

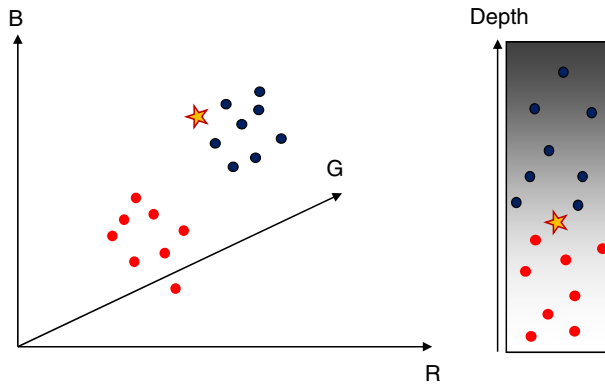


Fig. 2 Prior models (red spots are cluster centers of foreground and blue ones are those of background) in the RGB color space and disparity dimension (horizontally stretched in this illustration for clear illustration). The stars represent an unknown pixel in the attribute spaces

by,

$$P_D^c(p_i, x_i) = \begin{cases} 1 - \frac{df_i^c}{df_i^c + db_i^c}, & x_i = 1 \\ 1 - \frac{db_i^c}{df_i^c + db_i^c}, & x_i = 0 \end{cases} \quad (2)$$

Similarly, the probabilities in disparity dimension is defined as,

$$P_D^d(p_i, x_i) = \begin{cases} 1 - \frac{df_i^d}{df_i^d + db_i^d}, & x_i = 1 \\ 1 - \frac{db_i^d}{df_i^d + db_i^d}, & x_i = 0 \end{cases} \quad (3)$$

Here, df_i^d and db_i^d are the minimum distances from p_i to the foreground and background in the disparity dimension, whose computation is the same with those in the RGB color space (refer to (1)).

3.2 Energy function

In this part, we explain how to achieve adaptive prior selection for each pixel and integrate constraints extracted from the image pairs, including smoothness and correspondence constraints. We use N_B to denote the set of four-neighborhood pixels in I . p_i with an extra superscript l or r indicates the specific view that p_i belongs to. N_C is the set of corresponding pairs between the left and right views. The correspondences are determined by double-checking: if p_i^l 's corresponding pixel is p_j^r , and p_j^r 's is p_i^l , they are collected into N_C . Given a pixel, its corresponding pixel in the other view is determined by its disparity value computed by [12].

An MRF energy function for segmentation is given by,

$$E(x) = \sum_{p_i \in I} f_D(w_i^c, w_i^d, x_i) + \lambda_B \sum_{(p_i, p_j) \in N_B} f_B(x_i, x_j) + \lambda_C \sum_{(p_i^l, p_j^r) \in N_C} f_C(x_i^l, x_j^r) \quad (4)$$

f_D is a data term encoding clues from priors in RGB color and disparity spaces. f_B and f_C are smoothness constraints between neighbor pixels and consistency constraints between stereo corresponding points, respectively. λ_B and λ_C are weights for balancing the three terms.

Clues from priors The data term measures the disagreement between the labeling and the priors. It is given by,

$$f_D(w_i^c, w_i^d, x_i) = w_i^c f_D^c(p_i, x_i) + w_i^d f_D^d(p_i, x_i) \quad (5)$$

Here, $f_D^c(x_i) = 1 - P_D^c(p_i, x_i)$ and $f_D^d(x_i) = 1 - P_D^d(p_i, x_i)$. P_D^c and P_D^d are given in (2) and (3), respectively. From (5), it can be seen that the data term combines the clues from the color and disparity priors. w_i^c and w_i^d are the weights at p_i . $w_i^c + w_i^d = 1$. As we mentioned before, different pixels require different proportions of the clues from the two types of priors. The weights will be inferred along with the label x_i .

Smoothness and consistency constraints f_B , encoding smoothness constraints between neighboring pixels in each view, is given by

$$f_B(x_i, x_j) = \begin{cases} 0, & x_i = x_j \\ \lambda^c f_B^c(p_i, p_j) + \lambda^d f_B^d(p_i), & x_i \neq x_j \end{cases} \quad (6)$$

f_B^c and f_B^d are constraints defined by the color gradient and disparity variance at each pixel. They take the same forms with those in [16]. λ^c and λ^d are weights for balancing f_B^c and f_B^d . $\lambda^d = 1 - \lambda^c$.

f_C , encoding consistency constraints between corresponding points, is given by

$$f_C(x_i^l, x_j^r) = \begin{cases} 0, & x_i^l = x_j^r \\ S(p_i^l, p_j^r), & x_i^l \neq x_j^r \end{cases} \quad (7)$$

$S(p_i^l, p_j^r)$ represents the similarity between the pair of corresponding points in RGB color space (refer to [16] for similarity definition).

3.3 Optimization

In our method, beside the label $x_i \in \{1, 0\}$ of p_i , its prior weights, w_i^c and w_i^d in (5), are also variables to be optimized. The two weights are both in the range of $[0, 1]$. Since $w_i^c + w_i^d = 1$, only one weight needs to be decided. It is worth noting that a pixel would choose color or disparity as the dominant prior in practical experiments, for the reason that the energy function is linear to the weights at each pixel. Therefore, during optimization, the weights of the priors would prefer the two terminal states in the range of $[0, 1]$, i.e., $w_i^c, w_i^d \in \{0, 1\}$. Furthermore, considering that disparities obtained via stereo matching are noisy, we use $\{0, \delta\}$ as possible solutions of w_i^d , where $0 < \delta < 1$ to mitigate the influence of noisy disparities. The value of δ , depending on the confidence of disparities, is learned from data, which will be explained in the experiment part.

Different from the baseline method [16] and the other traditional methods [8, 17, 20], in which only a single variable (the label of foreground or background) was involved, we have two variables (the label and the prior weight) for each pixel, i.e., (x_i, w_i^c) . We use L to denote the set of the possible states of (x_i, w_i^c) . $L = \{(1, 1), (0, 1), (1, 1 - \delta), (0, 1 - \delta)\}$. Since our energy function has more than two states and satisfies the *metric* conditions given in [4], we use the α -expansion algorithm [4] to minimize it.

4 Experiments

In this part, we first describe the dataset used for testing the proposed method. Then, we introduce the evaluation metrics used in the experiments and how to determine the parameters involved in our method. Next, we validate our method by comparing it with the baseline method in [16] and several other state-of-the-art methods, based on the dataset and the evaluation metrics. Finally, we analyze the computational complexity of the proposed method.

4.1 Dataset

We use the dataset, composed of 37 pairs of stereo images, given in [16] to test our method and compare it with state-of-the-art ones. Besides the stereo image pairs, the dataset also provides user-input strokes (red for foreground and blue for background as shown in Fig. 6), disparity maps, and ground-truth segmentation results (refer to Fig. 6). The disparity maps are obtained by the algorithm in [12]. As seen from Fig. 6, the disparities are noisy and thereby should be used carefully. Note that interactive image segmentation is an iterative process involving gradual user interaction and optimization. In order to qualitatively compare different methods, we fixed the user input strokes to be the provided ones for all methods involved in the comparison.

4.2 Evaluation metrics

We use ER-F (Error Rate-Foreground) to express the ratio between the number of mis-labeled pixels and the total number of ground-truth foreground pixels in each pair of images. A pixel is mis-labeled if it is assigned a label inconsistent with its ground-truth one. Note that, different from the other methods [16] which used the total number of ground-truth background pixels as denominator, we adopt that of the ground-truth foreground pixels in ER-F. The reason is that the number of mis-labeled pixels are often small while background generally occupies a large part in images (refer to Fig. 6). It will be easier to differentiate the performances of different methods using foreground as denominator.

The variances in ER-Fs of different image pairs are generally large. For example, the ER-F of pair No. 22 is around ten times that of No.5 in Fig. 6. Therefore, the average ER-F over 37 pairs would bias to a few large ER-Fs, since a large ER-F might beat the sum of ten small ER-Fs. Therefore, we use the numbers of image pairs with the minimal ER-F among those obtained by different methods (Num_MinER-F for short) as one of the metrics.

Moreover, average ER-F Decline over the 37 pairs is also adopted for comparison evaluation. Given the ER-F of a pair of views obtained by a state-of-the-art method, ER-F Decline denotes the ER-F proportion decreased by our method. It is computed as follows. First, we subtract the ER-F of our method from that of the state-of-the-art method. ER-F Decline is the ratio of the subtraction result to the ER-F obtained by the state-of-the-art method. ER-F Decline mitigates the influences of large ER-F variances among different pairs via normalization.

4.3 Parameter selection and validation on disparity confidence control

There are four free parameters in our method, including λ_B and λ_C in (4), δ in (5), and λ^c in (6). λ_B and λ_C are used to balance the data term, smoothness and correspondence constraints. They take the same values as those in the baseline method [16]. δ is in the

range of $(0, 1)$. If the disparity/depth maps are accurate enough, δ could be close to 1. Unfortunately, disparity/depth maps generated by stereo matching are usually noisy (as seen from Fig. 6). We learn the value of δ from data. As δ , λ^c is also related to balancing color and disparity ingredients. Therefore, instead of directly using the value in [16], we adjust λ^c as well.

At first, we initialize λ^c with the value in [16]. Then, we test the possible values of δ on our dataset. Note that the metrics of Num_MinER-F and average ER-F Decline mentioned above are for comparison of different methods rather than influences of parameters on segmentation. Therefore, although easy to result in bias, we still have to use the average ER-F (AER-F for short) over 37 pairs as metric here. The AER-F at possible values (with an interval of 0.1) of δ are given in Fig. 3. A large δ means high confidence on disparity during segmentation. The AER-F is the lowest at $\delta = 0.5$ (refer to Fig. 3). Therefore, we set $\delta = 0.5$. Next, we fix δ and find the best $\lambda^c=0.5$ as done for δ .

Since there are only 37 pairs in the dataset, we use all of them to learn δ and λ^c . If a large dataset is provided, a small part could be used for the parameters. $\delta=0.5$ works for natural image pairs with disparities obtained by the stereo matching algorithm in [12]. If the acquisition means of depth data changes, users are suggested to learn a new δ .

Besides parameter selection, Fig. 3 also proves the effectiveness of the proposed confidence control on the disparity priors. From Fig. 3, we can see that the AER-F continuously decreases until $\delta = 0.5$. However, after $\delta > 0.5$, the AER-F grows rapidly. Therefore, priors in disparity dimension would help in stereo image segmentation, but the confidence on them should be controlled.

4.4 Accuracy comparison and analysis

We compare our method with four state-of-the-art methods, including StereoCut [20], Stereo GrabCut [8], FastCut [17] and RGB-D StereoCut [16]. StereoCut, which fixed the proportion of the color and disparity priors for segmentation of all image pairs, is treated as the baseline method of the proposed one. RGB-D StereoCut and FastCut are run with their original settings. RGB-D StereoCut has the same data and smoothness constraints with our method. In the implementation of StereoCut, we replace its original data and smoothness constraints with those of RGB-D StereoCut to prevent bias. The parameters that balance the data, smoothness, and correspondence constraints in StereoCut are the same with those

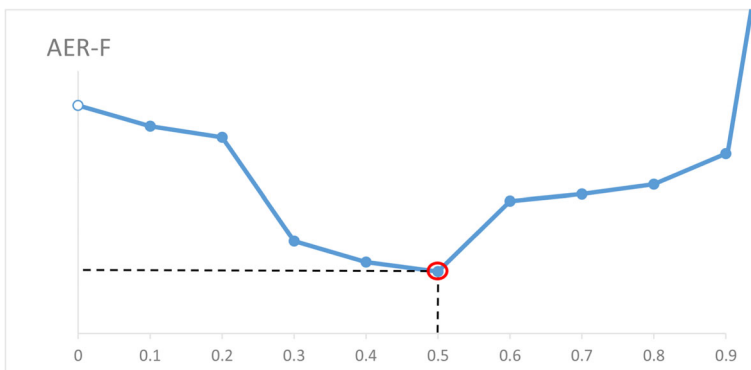


Fig. 3 The AER-Fs vary along the change of δ which is uniformly sampled in its possible range, $(0, 1)$, with an interval of 0.1. The AER-F is the lowest at $\delta = 0.5$

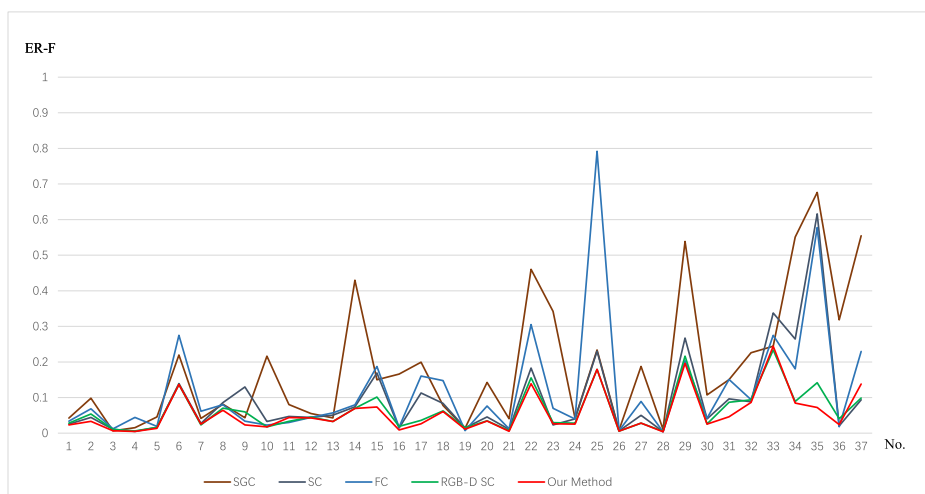


Fig. 4 The ER-Fs of Stereo GraphCut(SGC) [8], StereoCut(SC) [20], FastCut(FC) [17], RGB-D StereoCut(RGB-D SC) [16] and our method on each stereo pair

in RGB-D StereoCut. All the methods adopt the same user-input strokes provided in the dataset.

Figure 4 presents the error rates of the methods on each pair of stereo images. Results of several pairs of images obtained by the five methods are visualized in Fig. 6. From the figures, we can see that our method obtains the lowest ER-Fs in segmenting most of the image pairs. In all, our method gains on 27 pairs, i.e., having the lowest error rates on them among all the involved methods, while the others win at most 5 pairs (refer to the Num_MinER-F values in Fig. 5a). Even on pairs where our method fails, the gaps between our error rates and the best ones are tiny (refer to Figs. 4 and 6).

We compute the ER-F Declines of our method against the other four methods on every pair. The average ER-F Decline over the 37 pairs of stereo images against each of the four methods is given in Fig. 5b. From the figure, we can see that our method decreases 57.09%,

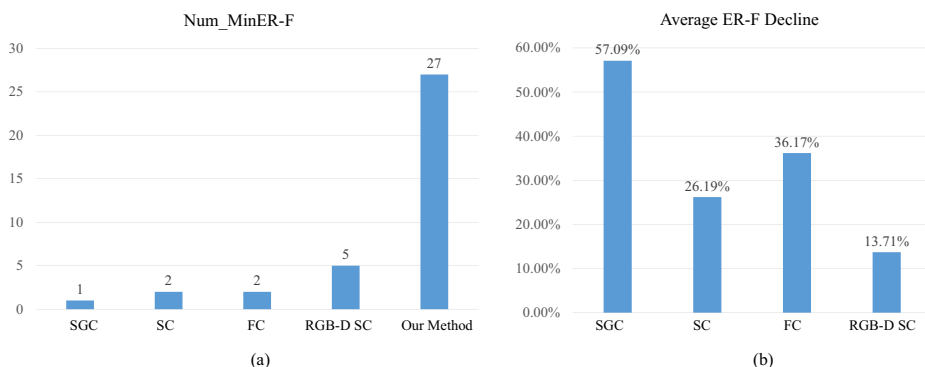


Fig. 5 **a** shows the Num_MinER-F values achieved by Stereo GrabCut(SGC) [8], StereoCut(SC) [20], FastCut(FC) [17], RGB-D StereoCut(RGB-D SC) [16] and our method; **b** is the average ER-F Decline of our method against the other four methods


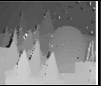







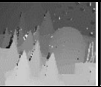
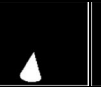





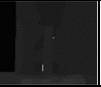







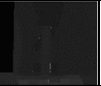






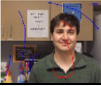







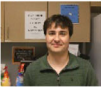








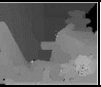

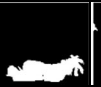
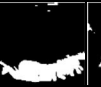

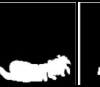


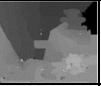







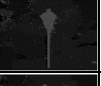
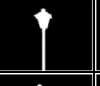

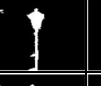





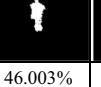
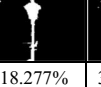
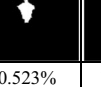
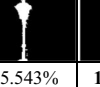
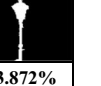











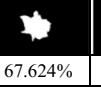
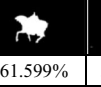
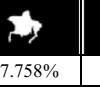
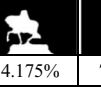

No.	Input	Disparity Map	Ground Truth	SGC	SC	FC	RGB-D SC	Our Method
3								
								
	ER-F			0.654%	1.078%	1.202 %	1.079%	0.667%
5								
								
	ER-F			4.539%	1.6233%	1.909%	1.553%	1.343%
10								
								
	ER-F			21.633%	3.294%	2.284%	1.745%	1.753%
15								
								
	ER-F			15.001%	16.896%	18.734%	10.149%	7.340%
22								
								
	ER-F			46.003%	18.277%	30.523%	15.543%	13.872%
35								
								
	ER-F			67.624%	61.599%	57.758%	14.175%	7.198%

Fig. 6 Input and segmentation results of stereo pair No. 3, 5, 10, 15, 22, and 35. The second column presents the original stereo image pairs and the input strokes marked on the left view (red for the foreground and blue for the background). Disparity maps and ground truth segmentation results are shown in the third and fourth columns, respectively. The fifth to tenth columns are the results and their ER-Fs obtained by Stereo GraphCut (SGC) [8], StereoCut (SC) [20], FastCut (FC) [17], RGB-D StereoCut (RGB-D SC) [16], and our method

26.19%, 36.17%, and 13.71% ER-F on average, against Stereo GrabCut [8], StereoCut [20], FastCut [17], and RGB-D StereoCut, respectively. Both the maximum Num_MinER-F and the average ER-F Declines demonstrate the good performances of the proposed method.

The highest Num_MinER-F and the positive ER-F Declines over the baseline method in [16] straightforwardly prove our argument that the prior selection at each pixel should be adapted.

4.5 Computational complexity analysis

Compared with the baseline method RGB-D StereoCut [16] and the other traditional methods, e.g., StereoCut [20], which solved the segmentation problem by binary graph cut, our method involves four solution states. Therefore, we use α -expansion [3] to find the solution, which is composed of $|L|$ cycles of binary graph cut [3]. L is the label set and $|L| = 4$. During optimization, segments would be refined by increasing the number of cycles of α -expansion until no refinement happens [3]. However, more iterations will induce higher time costs. Through experiments, we find that the energy function stops decreasing after two cycles. Therefore, we set the number of cycles in α -expansion to two. In all, our method is around $2|L| = 8$ times the complexity of binary graph cut [3]. Compared to the exponential growth of computing ability (by Moore's Law), the increase by a small constant factor in our computation complexity could be ignored.

5 Conclusion

The paper presented an interactive stereo image segmentation method which is capable of selecting priors adaptively for different pixels. Given a pair of stereo images, users are allowed to draw strokes to mark a few foreground and background pixels. Then, the proposed method formulated priors in both RGB color space and disparity dimension. Finally, constrained by neighborhood smoothness and stereo correspondences, the prior selection was solved together with the inference of each pixel's label (foreground or background) in the framework of multi-label MRF. Considering that disparity maps are noisy, a confidence control strategy was proposed and integrated in the framework. Through experiments, we validated the confidence control strategy and evaluated the proposed method with multiple metrics. Our method was experimentally proved to perform better than state-of-the-art methods.

Acknowledgments This research is supported by National Natural Science Foundation of China (61771026, 61379096, 61671451, 61502490), Scientific Research Project of Beijing Educational Committee (KM201510005015), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) and Beijing Municipal Natural Science Foundation (4152006). Great thanks to Dr. Xing Su and Dr. Tong Li for helping proofread the paper.

References

1. Achanta R, Estrada F, Wils P, Susstrunk S (2008) Salient region detection and segmentation. In: International conference on computer vision systems. Springer, pp 66–75
2. Blake A, Rother C, Brown M, Perez P, Torr P (2004) Interactive image segmentation using an adaptive gmmrf model. In: European conference on computer vision. Springer, pp 428–441

3. Boykov Y, Veksler O, Zabih R (2001) Interactive graph cuts for optimal boundary region segmentation of objects in n-d images. In: International conference on computer vision. IEEE, pp 105–112
4. Boykov Y, Veksler O, Zabih R (2002) Fast approximate energy minimization via graph cuts. *IEEE Trans Pattern Anal Mach Intell* 23(11):1222–1239
5. Boykov Y, Kolmogorov V (2004) An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans Pattern Anal Mach Intell* 26(9):1142–1137
6. Feng J, Price B, Cohen S, Chang S (2016) Interactive segmentation on rgbd images via cue selection. In: Computer vision and pattern recognition. IEEE, pp 156–164
7. Giró-Nieto X, Martos M, Mohedano E, Pont-Tuset J (2015) From global image annotation to interactive object segmentation. *Multimed Tools Appl* 70(1):475–493
8. Ju R, Xu X, Yang Y, Wu G (2013) Stereo GrabCut: Interactive and consistent object extraction for stereo images. In: Pacific-rim conference on advances in multimedia information processing. IEEE, pp 418–429
9. Ju R, Ren T, Wu G (2015) Stereosnakes Contour based consistent object extraction for stereo images. In: International conference on computer vision. IEEE, pp 1724–1732
10. Kass M, Witkin A, Terzopoulos D (1988) Snakes Active contour models. *Int J Comput Vis* 1(4):321–331
11. Kim Y, Winnemoller H, Lee S (2014) WYSIWYG stereo painting with usability enhancements. *IEEE Trans Vis Comput Graph* 20:957–969
12. Kolmogorov V, Zabih R (2002) Multi-camera scene reconstruction via graph cuts. In: European conference on computer vision. IEEE, pp 82–96
13. Li Y, Sun J, Tang CK, Shum HY (2004) Lazy snapping. *ACM Trans Graph* 23(3):303–308
14. Lo W, Baar J, Knaus C, Zwicker M (2010) Stereoscopic 3d copy & paste. *ACM Trans Graph* 29:147:1–147:10
15. Luo S, Shen I, Chen B, Cheng W, Chuang Y (2012) Perspective-aware warping for seamless stereoscopic image cloning. *ACM Trans Graph* 31:182:1–182:8
16. Ma W, Qin Y, Yang L, Xu S, Zhang X (2016) Interactive stereo image segmentation with rgb-d hybrid constraints. *IEEE Signal Process Lett* 23(11):1533–1537
17. Ma W, Yang L, Zhang Y, Duan L (2016) Fast interactive stereo image segmentation. *Multimed Tools Appl* 75(18):10,935–10,948
18. Ma W, Zhang Y, Yang L, Duan L (2016) Graph-cut based interactive image segmentation with randomized textron searching. *Comput Animat Virtual Worlds* 27(5):454–465
19. Ning J, Zhang L, Zhang D, Wu C (2010) Interactive image segmentation by maximal similarity based region merging. *Pattern Recogn* 43(2):445–456
20. Price B, Cohen S (2011) StereoCut: Consistent interactive object selection in stereo image pairs. In: International conference on computer vision. IEEE, pp 1148–1155
21. Rother C, kolmogorov V, Blake A (2004) Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Trans Graph* 23(3):309–314
22. Smith B, Zhang L, Jin H (2009) Stereo matching with nonparametric smoothness priors in feature space. In: Computer vision and pattern recognition. IEEE, pp 485–492
23. Xia L, Chen CC, Aggarwal JK (2011) Human detection using depth information by kinect. In: Computer vision and pattern recognition workshops. IEEE, pp 15–22
24. Xie L, Zhu L, Chen G (2016) Unsupervised multi-graph cross-modal hashing for large-scale multimedia retrieval. *Multimed Tools Appl* 75(15):9185–9204
25. Xu N, Price B, Cohen S, Yang J, Huang T (2016) Deep interactive object selection. In: Computer vision and pattern recognition. IEEE, pp 373–381
26. Zhang C, Li Z, Cai R, Chao H, Rui Y (2016) Joint multiview segmentation and localization of rgb-d images using depth-induced silhouette consistency. In: Computer vision and pattern recognition. IEEE, pp 4031–4039
27. Zhu L, Jin H, Zheng R, Feng X (2014) Weighting scheme for image retrieval based on bag-of-visual-words. *IET Image Process* 8(9):509–518
28. Zhu L, Shen J, Liu X, Xie L, Nie L (2016) Learning compact visual representation with canonical views for robust mobile landmark search. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, pp 3959–3965



Wei Ma received her Ph.D. degree in Computer Science from Peking University, in 2009. She is currently an Associate Professor at Faculty of Information Technology, Beijing University of Technology. Her current research interests include image processing, computer vision and their applications in the field of protection and exhibition of Chinese ancient paintings, and Information Content Security.



Yue Qin received her Bachelor degree in 2015. She is currently a Master Student at Faculty of Information Technology, Beijing University of Technology. She has been doing research in the field of Image Processing and Computer Vision. Her research interests are Stereo Image Editing and Repairing of Ancient Paintings.



Shibiao Xu received the B.S. degrees in Information Engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 2009, and the Ph.D. degree in Computer Science from Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2014. He is currently an associate professor in the Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include image based three-dimensional scene reconstruction and scene semantic understanding.



Xiaopeng Zhang received the B.S. degree and M.S. degree in Mathematics from Northwest University, Xi'an, China, in 1984 and 1987 respectively, and the Ph.D. degree in Computer Science from Institute of Software, Chinese Academy of Sciences, Beijing, China, in 1999. He is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His main research interests are computer graphics and computer vision.