Interactive Stereo Image Segmentation With RGB-D Hybrid Constraints

Wei Ma, *Member, IEEE*, Yue Qin, *Member, IEEE*, Luwei Yang, *Member, IEEE*, Shibiao Xu, *Member, IEEE*, and Xiaopeng Zhang, *Member, IEEE*

***Abstract*—This letter presents an approach to extracting a target object interactively from a given pair of stereo images. First, a user marks a few parts of the object and background in either of the two views with strokes. The marked pixels are used to generate the prior models of the foreground and background. Second, a graph is constructed with constraints formulated by the priors of foreground/background, similarities between intraview neighbor pixels and correspondences between interview pixels. Third, two segments of the foreground are extracted from the two views by optimization of the graph via graph cut. Traditional methods generally define the priors and neighbor similarities in RGB space. Differently, the proposed method integrates disparity distributions of foreground/background to enrich the priors and defines the similarity metric between neighbor pixels in RGB-D space. The proposed method that utilizes RGB-D hybrid constraints generates stereo segments with accuracies higher than those obtained by state-of-the-art methods.**

***Index Terms*—Disparity, graph cut, interactive segmentation, prior model, similarity, stereo image segmentation.**

I. Introduction

THE rapid increase in the amount of stereoscopic 3-D data has promoted research on efficient editing of stereo images [1], [2]. Extracting objects from stereo images is a key step that precedes many other operations, e.g., stereo image composition [1], [2] and 3-D modeling [3]. In this letter, we focus on the task of selecting objects interactively, i.e., performing object-background segmentation with user assistance, in stereo images. After segmentation, each pixel in the images will have a single label, object or background. A good method of interactive stereo segmentation should involve the smallest possible amount of user interaction. The object segments from left and right views should be consistent with each other [4].

A pair of stereo images can be handled separately with interactive segmentation methods for single images [5]–[7] because research on single-image segmentation began early and has already produced several practical tools (e.g., *Adobe Photoshop*). However, the amount of interaction in this case may double. In addition, the consistency between the segments from the two views cannot be easily guaranteed. Interactive cosegmentation methods can also be utilized for stereo image segmentation. For example, iCoseg, which was proposed by Batra *et al.* [8], [9], adopts a common foreground appearance model for a pair/group of images. However, the constraints of sharing the same foreground are not adequately strong to generate consistent results for stereo images [4].

Considering that a pair of images presents the same scene in two close views, researchers have attempted to relate the pair via correspondences for efficient and consistent segmentation [4], [10]. Tasli and Alatan [10] expanded the user-marked information of background and foreground in one view to the other through sparse SIFT correspondences to avoid doubled user interaction [11]. Ju *et al.* [12] presented a contour-based method to segment stereo images. This method has high efficiency because its searching space is restricted in boundaries of the stereo objects, which are correlated to one another by disparities. Ma *et al.* [13] performed stereo segmentation under the constraints of SIFT correspondences. These methods share limited information between left and right views for high efficiency [12], [13]. Price and Cohen presented StereoCut (SC) [4]. In this method, dense correspondences are formulated into interview constraints in a graph defined on a given pair of stereo images. Foreground/background color priors constructed by user-marked pixels, and color similarities between neighbor pixels are also adopted to define constraints in the graph. The segments are obtained by optimization on the graph through graph cut. Based on the same framework, Ju *et al.* [1] expanded the GrabCut method [14] for single images to cut out objects from stereo images.

In this study, we solved the problem of interactive stereo segmentation in the framework of graph cut, as done in SC [4]. Different from SC, the roles of dense correspondences are believed to be far more than merely relating a pair of images. In fact, we can obtain disparity/depth values through correspondences [15]. Similar to color and texture, disparity/depth contains important clues for segmentation. For example, Mutto *et al.* [16] performed unsupervised segmentation by clustering color and depth feature spaces. Kowdle *et al.* [17] extracted

Manuscript received May 5, 2016; revised August 29, 2016; accepted August 30, 2016. Date of publication September 1, 2016; date of current version September 14, 2016. This research was supported in part by Beijing Municipal Natural Science Foundation under Grant 4152006, in part by the Scientific Research Project of Beijing Educational Committee under Grant KM201510005015, in part by Ri-Xin Talents Project under Grant 2014-RX-L06in part by Seed Funding for International Cooperation of Beijing University of Technology, and in part by the National Natural Science Foundation of China under Grant 61332017 and Grant 61271430. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Giuseppe Scarpa. *(Corresponding authors: Dr. Shibiao Xu; Dr. Xiaopeng Zhang.)*

W. Ma, Y. Qin, and L. Yang are with Beijing University of Technology, Beijing 100871, China (e-mail: mawei@bjut.edu.cn; qinyue1992@emails.bjut.edu.cn; mluweiyang@163.com).

S. Xu and X. Zhang are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: shibiao.xu@ia.ac.cn; xiaopeng.zhang@ia.ac.cn).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2016.2605133

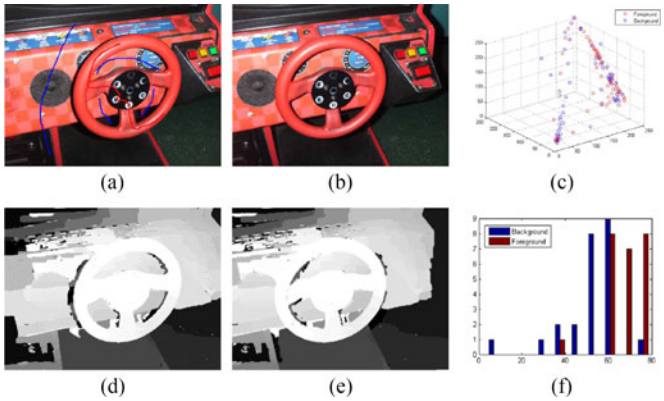

Fig. 1. (a) and (b) are an input pair with user-marked foreground and background. (d) and (e) are their disparity maps. (c) and (f) are the color and disparity distributions of the marked foreground/background, respectively.

objects from multiple views automatically and simultaneously by using both appearance and depth cues. Different from these studies, the current study focused on interactive segmentation. We analyzed the functions of disparities and proposed a new hybrid prior model for the foreground/background and a similarity measure for neighbor pixels in the framework of graph cut. Experimental results demonstrate that the proposed method can obtain segments more accurately than state-of-the-art methods with the same amount of user interactions.

II. Proposed Method

The proposed method aims to solve the stereo image segmentation problem in the framework of graph cut. First, we utilized the algorithm in [18], whose source codes are available in http://www.cs.cornell.edu/People/vnk/recon.html, to obtain disparity maps offline. Although targeting at multiview stereo, this algorithm can generate satisfactory results on various natural images of two views. Second, after importing a pair of images and their disparity maps, users can indicate the target foreground and background with strokes [e.g., blue for background and red for foreground, as shown in Fig. 1(a)], in either of the two image views. Third, the proposed method builds a graph with RGB and disparity constraints (RGB-D constraints for short). Establishing this graph is the key part of the proposed method. Finally, segmentation results were obtained by applying standard graph cut optimization on the entire graph. If users are not satisfied with the results, they can label additional foreground/background parts by referring to the results, and a new cycle of graph construction and optimization process would be triggered.

We used $I = \{I^l, I^r\}$ to represent the left and right views of a given pair of stereo images. A graph $G = (v, \varepsilon)$ was constructed for I. In the graph, v is the set of nodes, and each node represents a pixel of I. Object segmentation aims to assign each node $p_i \in v$ a label $x_i \in \{1, 0\}$ representing the foreground and background. Variable ε is the set of edges connecting intraview four neighbors, interview corresponding pixels, and each pixel to two terminal nodes representing the foreground and background. The definition of the three types of edge weights, namely, intraview neighbor, interview correspondence, and terminal edges, determines the performance of segmentation. In the following text, we explain the formulations of the three weights from the view angle of the energy function.

A. The Proposed Energy Function

The energy function defined on the graph is provided by

$$E(x) = \sum_{p_i \in I} f_D(p_i, x_i) + \lambda_B \sum_{(p_i, p_j) \in N_B} f_B(p_i, p_j)|x_i - x_j| + \lambda_C \sum_{(p_i^l, p_j^r) \in N_C} f_C(p_i^l, p_j^r)|x_i^l - x_j^r| \quad (1)$$

where $f_D(p_i, x_i)$ is a unary term representing the similarity of pixel p_i to foreground and background prior models. $f_B(p_i, p_j)$ denotes the differences of pixels with their neighbors in each view of the image. Thus, it is also called the intraview binary term. N_B is a set recording the neighborhood of every pixel in both left and right views. $f_C(p_i^l, p_j^r)$ is referred to as the interview binary term because it denotes the matching results of the correspondent points. N_C is the set of corresponding pairs. λ_B and λ_C are user-defined weights to balance the three terms.

1) Unary Term: In traditional methods, unary term generally encodes constraints from the color priors of foreground and background [19], [4]. In this letter, we presented a hybrid form composed of a traditional color term and constraints from foreground/background disparity distributions (DispDis for short), given by

$$f_D(p_i, x_i) = \lambda_c(1 - P_c(x_i|c_i)) + \lambda_d(1 - P_d(x_i|d_i)) \quad (2)$$

where $P_c(x_i|c_i)$ denotes the probability of labeling pixel p_i, whose color is c_i, as foreground ($x_i = 1$) or background ($x_i = 0$). Similarly, $P_d(x_i|d_i)$ represents the probability of x_i taking the foreground or the background label given the disparity value d_i of pixel p_i. λ_c and λ_d as weights balancing the color and disparity constraints satisfy $\lambda_c + \lambda_d = 1$.

Four k-means clustering procedures were performed to generate the color and DispDis of foreground and background with user-marked pixels, as done in [19]. Each distribution is composed of a set of clusters. We used $\{C^f\}$ and $\{C^b\}$ to represent the centers of the foreground and background color clusters, respectively. The color term $P_c(x_i|c_i)$ in (2) is given by

$$P_c(x_i|c_i) = \begin{cases} \dfrac{d_i^{\min}}{s_i^{\min} + d_i^{\min}}, & x = 1 \\ \dfrac{s_i^{\min}}{s_i^{\min} + d_i^{\min}}, & x = 0 \end{cases}$$

$$\text{where,} \quad \begin{cases} s_i^{\min} = \min(||c_i - C_j^f||^2), & j = 1, ..., k \\ d_i^{\min} = \min(||c_i - C_j^b||^2), & j = 1, ..., k. \end{cases} \quad (3)$$

In this study, k was set to 64, as recommended in [19]. $s_i^{\min}$ and $d_i^{\min}$ denote the smallest distances from the color value c_i of pixel p_i to the background/foreground color prior models, respectively. The computation of disparity term $P_d(x_i|d_i)$ in (2) is similar to that of the color term.

2) Intraview Binary Term: The intraview binary term encodes similarities between neighbor pixels. In traditional methods, similarities are generally defined as being inversely proportional to color differences [4], [13]. The proposed method considers an RGB-D hybrid form provided by

$$f_B(p_i, p_j) = \lambda_c f_G(p_i, p_j) + \lambda_d f_V(p_i) \tag{4}$$

where $f_G(p_i, p_j)$ encodes color differences and takes a traditional form as in [20], which is provided by

$$f_G(p_i, p_j) = \frac{1}{(1 + ||c_i - c_j||^2)}, (p_i, p_j) \in N_B. \tag{5}$$

$f_V(p_i)$ encodes the disparity difference between p_i and p_j. The obtained disparities are too noisy to be utilized for similarity measure; thus, we introduce disparity variance (DispVar for short) to define $f_V(p_i)$, which is given by

$$f_V(p_i) = Z\left(1 - \text{var}^{0.2}_{l(p_i)}\right) \tag{6}$$

where $l(p_i)$ indicates the local area centered at p_i. $\text{var}_{l(p_i)}$ denotes the variance in the local area. The superscript 0.2 was determined through experiments. For computation efficiency, we divided a disparity map into square areas with a size of 8×8. The variance value of each area was precomputed after importing the disparity map. The constant Z was experimentally set to 0.0005 for all the experiments.

Note that we took λ_c and λ_d, originally used in (2), to balance the two terms on the right-hand side of (4), which means the roles of color and disparity in (2) and (4) have the same proportion.

3) Interview Binary Term: The interview binary term causes the corresponding pixels in the two views to have the same label (background or foreground). In this study, this term takes a form similar to that in SC [4], which is provided by

$$f_C(p_i, p_j) = \frac{C(p_i^l, p_j^r) + C(p_j^r, p_i^l)}{2}. \tag{7}$$

In this equation, $C(p_i^l, p_j^r)$ is defined as

$$C(p_i^l, p_j^r) = P(M(p_i^l) = p_j^r) P(x_i^l | M(p_i^l) = p_j^r, x_j^r) \tag{8}$$

where $M(p_i^l) = p_j^r$ represents the mapping relationship between left and right view corresponding points. $P(M(p_i^l) = p_j^r)$ is the matching probability function decided by a consistent delta function, as performed in [4]. $P(x_i^l | M(p_i^l) = p_j^r, x_j^r)$ represents the similarity between p_i^l and p_j^r. It is provided by

$$P(x_i^l | M(p_i^l) = p_j^r, x_j^r) = \frac{1}{||c_i^l - c_j^r||^2 + 1} \tag{9}$$

where c_i^l is the color value of the left view pixel p_i^l and c_j^r is the color value of the left view pixel p_j^r. The computation of $C(p_j^r, p_i^l)$ is similar to that of $C(p_i^l, p_j^r)$.

B. Effectiveness of the Two Novel Constraints

We analyzed the effectiveness of the two novel constraints formulated by DispDis and DispVar, respectively.

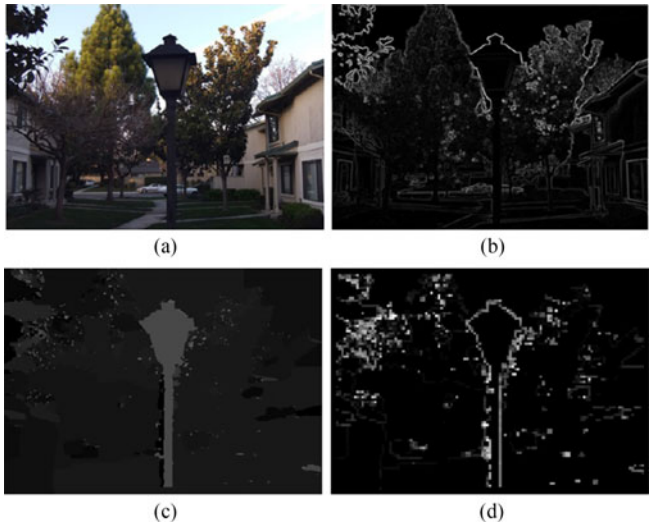

Fig. 2. (a) shows an input image. (b) shows the edge map of (a). (c) shows the disparity map of (a). (d) shows the map of disparity variances.

1) Disparity Distribution Constraint: Traditional segmentation methods define the priors of the foreground/background based on color [4], [13]. However, the same colors may appear on background and foreground of many natural images, as illustrated in Fig. 1(c). In this case, the unary term is ineffective. Fig. 1(f) shows the disparity distribution histograms of the user-marked background and foreground pixels. The disparity distribution of the foreground and that of the background vary significantly, thereby inspiring us to integrate disparity distribution in the prior models for effective constraints. The disparity is 1D, so the histograms were utilized for illustration in this study. In the implementation, we used clusters to represent DispDis given that clusters are powerful, particularly for images with large variances in disparity values.

2) Disparity Variance Constraint: Traditional binary term generally encodes gradients among neighbor pixels [4], [13]. However, many natural images do not have clear edges along the border of objects and even have large gradients within the object. As shown in Fig. 2(b), the upper part of the lamppost has clear edges unlike the supporting rod. In this case, the boundaries of the rod cannot be easily located by depending only on the color gradients. In this study, we integrated the DispVar near each pixel into the binary term, as explained earlier. Fig. 2(d) presents a DispVar map. The boundaries of the rod in this map are considerably clearer than those in the edge map [see Fig. 2(b)]. This observation is consistent with the fact that target objects in natural images generally lie in deep layers different from their background. The disparity variances are not pixel-level precise. Therefore, in this study, we combined the color gradients with disparity variances to measure the similarities among neighbor pixels.

III. Experimental Results

A. Parameter Selection

We used a dataset of 37 pairs of stereo images (available at http://ilabqy.github.io/index.html) to test our method. The

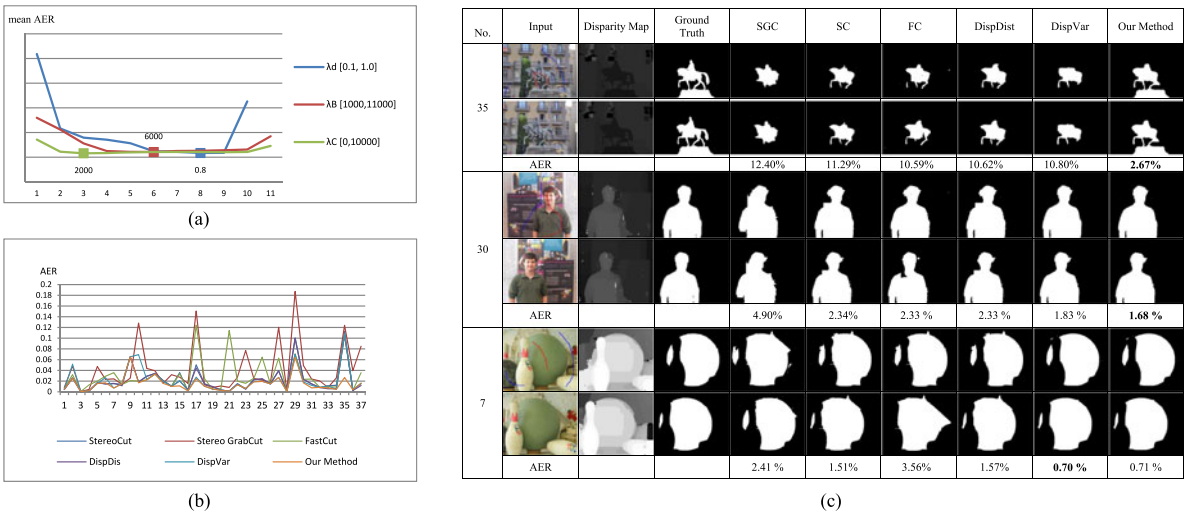

Fig. 3. (a) shows the mean average error rates (AERs) obtained by the proposed method with different parameters for parameter selection. (b) presents the AERs of the six methods in each stereo pair. (c) shows the segmentation results of the six methods for stereo pairs No. 7, 30, and 35. The input strokes are marked on the left view of each pair (red for the foreground and blue for the background). The disparity and the ground truth are shown in the third and fourth columns, respectively. The fifth to tenth columns are results, and their AERs are obtained via the six methods.

TABLE I
AERs OF THE SIX DIFFERENT METHODS

SGC [1]	SC [4]	FC [13]	DispDis	DispVar	Our Method
4.159%	2.185%	2.919%	2.067%	2.071%	**1.519%**

TABLE II
AVERAGE TIME COSTS OF SC, DISPDIS, DISPVAR, AND THE PROPOSED METHOD IN BUILDING GRAPHS FOR THE 37 PAIRS

SC [4]	DispDis	DispVar	Our Method
0.308 s	0.326 s	0.322 s	**0.340 s**

dataset is composed of stereo images, input strokes, disparity maps, and ground truth segmentation results. We used the average error rate (AER) over the two views of a pair of images to denote the segmentation accuracy of the pair. The error rate in one view is defined as the ratio of the number of pixels incorrectly labeled to the total number of ground-truth foreground pixels in the view. If the assigned label of a pixel is consistent with its ground truth, then this pixel is considered correctly labeled [13]. The mean AER of the 37 pairs was used to represent segmentation performance.

The proposed method has three free parameters: λ_d, λ_C, and λ_D. In the experiment, the optimal parameters were selected via linear searching [21]. Linear search is performed by fixing all but one parameter and by changing the parameter over a predefined range [21]. The variation in the mean AER with the change in the parameters is visualized in Fig. 3(a). The optimal values of the parameters were set to $\lambda_d = 0.8$, $\lambda_B = 6000$, and $\lambda_C = 2000$, which minimize the mean AER.

B. Performance Analysis

We compared our method to three state-of-the-art methods, namely, SC [4], Stereo GrabCut (SGC) [1], and FastCut (FC) [13]. In implementing SC, we replaced its original region and boundary constraints [4] with ours, i.e., $1 - P_c(x_i|c_i)$ in (2) and f_G in (4), to prevent bias. The parameters that balance the region, boundary, and interview constraints are the same with ours. The mean AERs of the three methods given in Table I show that our method has the best performance because its mean AER is lower than that of SGC [1] (by 2.64%), SC [4] (by 0.666%), and FC [13] (by 1.4%). Even with only DispDis or DispVar, our method obtains better performance than state-of-the-art methods, which can be seen from Table I. Fig. 3(b) presents the AERs of the methods in each pair of images. Only some of the detailed data of the pairs are given in Fig. 3(c) because of page limitation. More results are presented at http://ilabqy.github.io/index.html. Fig. 3 shows that our method has the lowest AERs in nearly every pair of images. However, in the case that the disparities of foreground and background are similar, as those of pair No. 7 in Fig. 3(c), DispDis plays a negative role [refer to the AERs in Fig. 3(c)]. DispVar might decrease the accuracy of our method if the disparities at boundaries are very noisy, e.g., those of pair No. 6 (refer to the last figure on the webpage).

Unlike SC [4], our method involves two extra constraints; thus, the time cost of our method in building graphs for graph cut optimization [18] is more than that of SC (by only 0.032 s as we can see from Table II). The time costs are tested on a PC with Intel(R) Core(TM) i5-4590 CPU @3.30GHz. The costs of the other two state-of-the-art methods are not given because they have already been presented in [13] together with that of SC.

IV. CONCLUSION

We presented an interactive stereo image segmentation method that uses RGB-D hybrid constraints to build an energy function. The constraints encode DispDis in prior models of the foreground and background, and disparity variances in the similarities between neighbor pixels. The good performance of our method and the proposed constraints was proven by comparing them with the state-of-the-art methods.

REFERENCES

[1] R. Ju, X. Xu, Y. Yang, and G. Wu, "Stereo grabcut: Interactive and consistent object extraction for stereo images," in *Pacific-Rim Conf. Multimedia*, 2013, vol. 8294, pp. 418–429.

[2] W.-Y. Lo, J. van Baar, C. Knaus, M. Zwicker, and M. Gross, "Stereoscopic 3d copy & paste," *ACM Trans. Graph.*, vol. 29, no. 6, pp. 147:1–147:10, 2010.

[3] T. H. Kim, K. M. Lee, and S. U. Lee, "A unified probabilistic approach to feature matching and object segmentation," in *Proc. Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 464–467.

[4] B. Price and S. Cohen, "Stereocut: Consistent interactive object selection in stereo image pairs," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1148–1155.

[5] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary region segmentation of objects in n-d images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2001, vol. 1, pp. 105–112.

[6] D. P. Huttenlocher and P. F. Felzenszwalb, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, 2004.

[7] A. Seitz, L. Grady, and M.-P. Jolly, "Segmentation from a box," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, vol. 1, pp. 367–374.

[8] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "iCoseg: Interactive co-segmentation with intelligent scribble guidance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3169–3176.

[9] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "Interactively co-segmentating topically related images with intelligent scribble guidance," *Int. J. Comput. Vis.*, vol. 93, no. 3, pp. 273–292, 2011.

[10] H. Tasli and A. Alatan, "User assisted stereo image segmentation," in *3DTV-Conf.: True Vis.-Capture, Transmission Display 3D Video*, Oct. 2012, pp. 1–4.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[12] R. Ju, T. Ren, and G. Wu, "Stereosnakes: Contour based consistent object extraction for stereo images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1724–1732.

[13] W. Ma, L. Yang, Y. Zhang, and L. Duan, "Fast interactive stereo image segmentation," *Multimedia Tools Appl.*, vol. 75, no. 18, pp. 1–14, 2016.

[14] 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.

[15] P. Kamencay, M. Breznan, R. Jarina, P. Lukac, and M. Zachariasova, "Improved depth map estimation from stereo images based on hybrid method," *Radioengineering*, vol. 21, no. 1, pp. 70–78, 2012.

[16] C. D. Mutto, P. Zanuttigh, G. Cortelazzo, and S. Mattoccia, "Scene segmentation assisted by stereo vision," in *Proc. Int. Conf. 3D Imaging, Model., Process., Vis. Transmission*, May 2011, pp. 57–64.

[17] A. Kowdle, S. N. Sinha, and R. Szeliski, "Multiple view object cosegmentation using appearance and stereo cues," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 789–803.

[18] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 82–96.

[19] Y. Li, J. Sun, C.-K. Tang, and H.-Y. Shum, "Lazy snapping," *Trans. Graph.*, vol. 23, no. 3, pp. 303–308, 2004.

[20] W. Ma, J. Liu, L. Duan, and X. Zhang, "Image segmentation with automatically balanced constraints," in *Proc. IEEE 2nd IAPR Asian Conf. Pattern Recognit.*, 2013, pp. 557–561.

[21] P. Kohli, H. Nickisch, C. Rother, and C. Rhemann, "User-centric learning and evaluation of interactive segmentation systems," *Int. J. Comput. Vis.*, vol. 100, no. 3, pp. 261–274, 2012.