RGBD Co-saliency Detection via Bagging-Based Clustering

Hangke Song, Zhi Liu, *Senior Member, IEEE*, Yufeng Xie, Lishan Wu, and Mengke Huang

***Abstract*—With the additional depth information, RGBD co-saliency detection, which is an emerging and interesting issue in saliency detection, aims to discover the common salient objects in a set of RGBD images. This letter proposes a novel RGBD co-saliency model using bagging-based clustering. First, candidate object regions are generated based on RGBD single saliency maps and region pre-segmentation. Then, in order to make regional clustering more robust to different image sets, the feature bagging method is introduced to randomly generate multiple clustering results and the cluster-level weak co-saliency maps. Finally, a clustering quality (CQ) criterion is devised to adaptively integrate the weak co-saliency maps into the final co-saliency map for each image. Experimental results on a public RGBD co-saliency dataset show that the proposed co-saliency model significantly outperforms the state-of-the-art co-saliency models.**

***Index Terms*—Clustering quality (CQ), co-saliency detection, feature bagging, RGBD image.**

I. Introduction

SALIENCY detection simulates the mechanism of human visual attention to select the most salient objects that stand out from the other parts in the scene. Since the seminal work in [1], a huge number of saliency models have been proposed for single image, such as [2]–[6], just to name a few. Rather than focusing on such a relatively mature research topic, in this letter, we focus on a novel branch of saliency detection, i.e., co-saliency detection for a set of RGBD images.

Different from the traditional saliency detection, co-saliency detection aims at discovering the common and salient objects in a set of relevant images, and can be widely used in many applications such as image/video co-segmentation [7], [8], object colocalization [9], and weakly supervised object detection [10]. The concept of co-saliency was first defined as calculating saliency of image pixels in the context of other related images in [11]. After that, some models started to discover co-salient objects from image pairs [12] and extended to image set with more than two related images [13]–[17]. Contrast cue, spatial cue, and corresponding cue are measured at the cluster level to generate co-saliency maps in [13]. In [14], a general saliency map fusion framework, which exploits the relationship of multiple saliency cues to obtain the self-adaptive weights, is proposed to generate the combined co-saliency map. In [15], based on hierarchical segmentation, regional similarity, contrast, and object prior are integrated with global similarity to obtain co-saliency maps. In [16], co-salient exemplars are generated by similarity of color and SIFT features, and exploited to recover co-saliency maps by propagating saliency values of exemplars. In [17], multiscale segmentation based intra-image saliency is integrated with region match based inter-image saliency for co-saliency detection.

With the popularity of stereo cameras, depth cameras and Kinect sensors, depth information has been proven to be powerful for saliency detection in RGBD images as well as stereoscopic images [18]–[20]. In [21], for the purpose of co-segmentation, several existing saliency models and co-saliency models for color images as well as saliency models for RGBD images are fused to generate co-saliency maps for RGBD images, but the relevance of depth information among different RGBD images is not exploited to highlight the co-salient objects. In other words, co-saliency detection via the use of depth information has not been effectively explored yet.

Therefore, this letter proposes a novel RGBD co-saliency detection framework using the bagging-based clustering, as shown in Fig. 1. First, RGBD single saliency maps and region pre-segmentation are used to generate candidate object regions. Then regional features of color, depth, and geometric properties are randomly selected via the bagging method to perform multiple clustering processes and generate the cluster-level weak co-saliency (WCS) maps. Finally, a clustering quality (CQ) criterion is proposed to discover the most effective clustering results, and multiple WCS maps are adaptively fused in a discriminative way to generate the final co-saliency map for each image. Compared with the co-saliency model in [13], which mainly uses the pixel clustering as a pre-processing step and exploits feature contrast and spatial prior for measuring co-saliency, our model regards the clustering process, which leverages feature bagging on the basis of candidate object regions, as the key of co-saliency detection, and effectively utilizes multiple clustering results for measuring co-saliency. Our main contribution lies in the following three aspects.

1) In addition to color cues, the relevance of depth information among different images in the RGBD image set is utilized to highlight the co-salient objects.
2) In order to discover the most discrinimative features for clustering, a feature bagging method is proposed to perform clustering multiple times with different features, and to effectively enhance the overall reliability of regional clustering for different scenes compared to the single clustering with the fixed features.

Manuscript received July 1, 2016; revised August 18, 2016; accepted October 3, 2016. Date of publication October 4, 2016; date of current version October 26, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant 61471230 and Grant 61171144, and in part by the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zhu Liu. *(Corresponding author: Zhi Liu.)*

The authors are with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: hksong0209@163.com; liuzhisjtu@163.com; xieyufeng0227@163.com; wlsxxrs@163.com; mengkehuang1@163.com).

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.2016.2615293

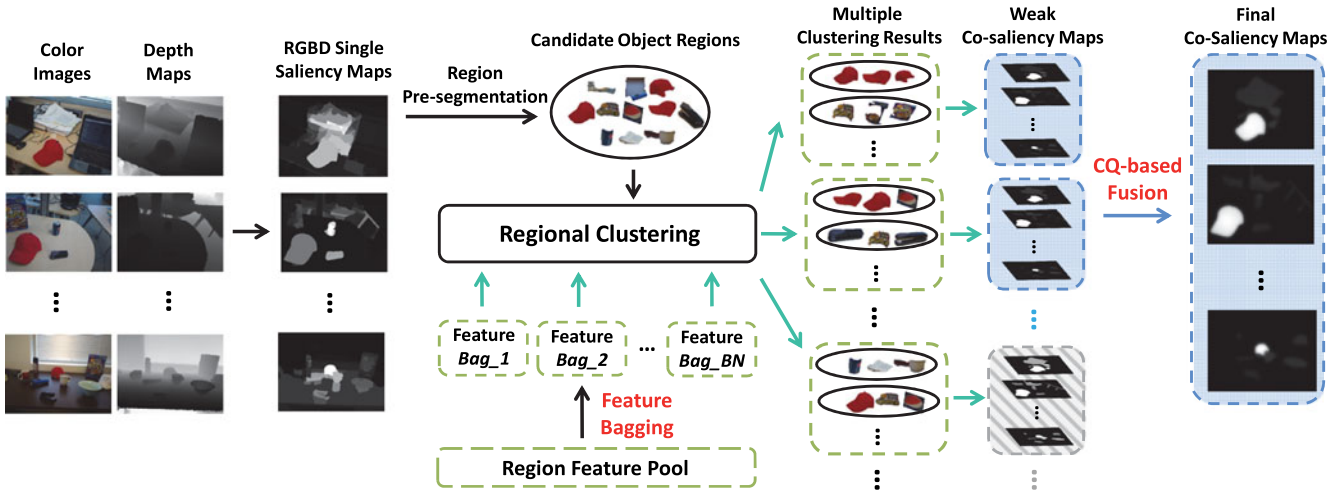

Fig. 1. (Better viewed in color) Framework of RGBD co-saliency model via bagging-based clustering.

3) A CQ criterion and a CQ-based fusion method are proposed to discriminatively integrate multiple cluster-level WCS maps to achieve the higher co-saliency detection performance.

II. Proposed RGBD Co-saliency Model

A. Generate Candidate Object Regions

Given a set of RGBD images $\{I_m\}_{m=1}^{M}$, the candidate object regions are generated based on RGBD single saliency maps. First, we use our previous work [20] to generate for each image the RGBD single saliency map, which is normalized into the range of [0, 1]. Second, following [20], the *gPb–owt–ucm* method [22] is performed on the color image to obtain a moderate segmentation result with Q regions. Each region in the image I_m is denoted as $\{R_i^m\}_{i=1}^{Q}$, and the RGBD single saliency of R_i^m is assigned with the mean of RGBD single saliency values of all pixels in R_i^m. Finally, R_i^m is selected as a candidate object region for clustering if its RGBD single saliency is greater than the threshold T, which is empirically set to 0.25 for a relatively higher recall rate.

B. Regional Clustering via Feature Bagging

In this letter, we perform co-saliency detection by measuring the saliency of each cluster generated from the image set $\{I_m\}_{m=1}^{M}$. On the basis of each candidate object region, we extract various color, depth, and geometric features for clustering. The details of these features are as follows (along with the dimension of feature in the parenthesis): average color of each channel in the RGB, HSV, and L*a*b* color spaces (9); absolute responses (3) of Leung and Malik's (LM) filters [23] on the color image; average depth value and depth range (2); histograms of oriented gradients (HOG) [24] (3) on the depth map; region area and perimeter (2); the width and height of the bounding box (2); the length of the major/minor axis and the eccentricity of the bounding ellipse (3). Note that for texture features including absolute responses of LM filters and HOG histograms, their dimensions are actually reduced by the principal component analysis method, to avoid the distraction of high-dimensional features in the clustering process. The depth range is defined as the absolute difference between the minimum depth value and the maximum depth value of the pixels in each region. The total dimension of all the above features is 24.

However, the discriminability of individual feature always changes over the image set with different scenes, and the challenge is how to choose the appropriate ones among all the features for obtaining accurate clustering results with different scenes. To address this challenge, we introduce the bagging method into the clustering process to try different feature subsets and perform multiple clustering processes, so as to have the chance to obtain some relatively accurate clustering results. For the *n*th clustering process, we first generate a random number $FN^n (1 \le FN^n \le 24)$ to randomly select a total of FN^n features and concatenate them to form a FN^n-dimensional feature vector for each region. Then, for each cluster C_j^n in the *n*th clustering result, the cluster-level co-saliency (CCS) is defined as follows:

$$\mathrm{CCS}_j^n = MS_j^n \cdot \exp(1 - \mathrm{NCD}_j^n) \cdot \exp\left(\mathrm{CCO}_j^n / M\right) \quad (1)$$

where MS_j^n is the mean of RGBD single saliency values of all regions contained in C_j^n. The Euclidean distance between the feature vector of each region in C_j^n and the cluster center of C_j^n is calculated, and the mean of all such Euclidean distances in C_j^n is defined as the cluster dispersion measure CD_j^n. CD_j^n is normalized as $\mathrm{NCD}_j^n = (\mathrm{CD}_j^n - \mathrm{CD}_{\min}^n)/(\mathrm{CD}_{\max}^n - \mathrm{CD}_{\min}^n)$ with the range of [0, 1], where $\mathrm{CD}_{\min}^n/\mathrm{CD}_{\max}^n$ is the minimum/maximum cluster dispersion measure in the *n*th clustering result. A lower value of NCD_j^n indicates the higher similarity among the regions in C_j^n. The cluster co-occurrence rate CCO_j^n is represented by the number of images involved in C_j^n. Therefore, using (1), the cluster in which the regions are salient in individual image, show the higher similarity and occur in more images will be evaluated with the higher co-saliency. In our implementation, the number of clustering processes, *BN*, is set to 150, for a balance between the runtime and clustering performance. The maximum number of clusters for each clustering process is set to 20. The fuzzy *c*-means clustering method [25] is used to perform each clustering process, and its time cost is lower on average when compared with the widely used *k*-means clustering method [26].

As shown in Fig. 1, with a total of BN groups of features via bagging for clustering the candidate object regions, a total of BN clustering results are generated. The cluster-level co-saliency CCS_j^n is assigned to each region in the cluster C_j^n and termed as the WCS measure. Note that the WCS measures of regions that are not selected as the candidate object regions are uniformly set to 0. As shown in Fig. 1, for each image I_m, a total of *BN* WCS maps $\{\mathrm{WCS}_m^n\}_{n=1}^{BN}$ are generated for *BN* clustering results.

C. CQ-Based Fusion

Because of the randomness of feature bagging for each clustering process, discriminative feature subsets or misleading feature subsets could come out together. Therefore, how to evaluate each clustering result with different features is the key to find out the most effective WCS maps. For this purpose, we propose a criterion to evaluate the quality of clustering result, and use it to calculate the fusion weights for adaptively integrating multiple WCS maps. For the *n*th clustering process, the CQ is defined as follows:

$$\mathrm{CQ}^n = \mathrm{SR}_J^n \cdot \exp\left(\frac{\mathrm{CCO}_J^n/M}{RN^n} - DA_J^n\right) \tag{2}$$

where $\mathrm{SR}_J^n = \sum_{j \neq J} [\mathrm{CCS}_J^n / \mathrm{CCS}_j^n]$ is the cluster separation rate, and *J* is the index of the cluster C_J^n, which has the largest CCS value. A larger value of SR_J^n indicates that the object regions are highlighted better from background regions with the *n*th clustering process. DA_J^n is defined as the variance on the number of regions belonging to the cluster C_J^n in each image. A small value of DA_J^n means that all images tend to have the same number of the highlighted object regions, and this indicates the higher reliability of the *n*th clustering process. By using the adaptive thresholding method [27], each WCS map with the *n*th clustering process is thresholded to generate the binary object map, and the average number of regions in all binary object maps is denoted as RN^n. A small value of RN^n indicates that object regions are more likely to be concentrated and background regions are uniformly suppressed in WCS maps, and this indicates the better quality of WCS maps. Using (2), for the clustering process with the higher reliability and the corresponding WCS maps with the better quality, its CQ value is higher.

Based on the CQ values, the top $TN = BN/3$ clustering results are utilized to generate the final region-level co-saliency map for each image I_m via the adaptive fusion as follows:

$$\mathrm{FCS}_m = \sum_{n=1}^{TN} \mathrm{CQ}^n \cdot \mathrm{WCS}_m^n \tag{3}$$

which is convolved with a small Gaussian kernel to generate the final pixel-level co-saliency map with the better visualization as suggested in [28]. As shown in Fig. 1, based on the CQ values, the accurate results with high quality such as the top two sets of WCS maps (shown with blue background color) are used for CQ-based adaptive fusion, while the inaccurate results such as the bottom set of WCS maps (shown with gray shade lines) are abandoned. For each image, the final pixel-level co-saliency map by combining the top *TN* WCS maps can uniformly highlight the co-salient object regions and effectively suppress irrelevant regions.

Fig. 2. Examples of co-saliency detection on four image sets.

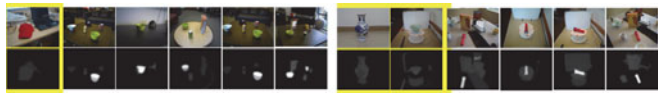

Fig. 3. (Better viewed in color) Examples of irrelevant images involved (yellow: irrelevant images).

III. Experimental Results

We performed experiments on a public dataset [21], which consists of 183 RGBD images from 16 object classes. This dataset is suitable for evaluating the co-saliency detection performance on RGBD images. We compared our co-saliency model with the state-of-the-art co-saliency models including a RGBD co-saliency model, fusion of five saliency models (FFS) [21], as well as four high-performing co-saliency models, including cluster-based model (CB) [13], self-adaptively weighted co-saliency model (SACS) [14], object discovery and recovery model (ODR) [16], and hierarchical segmentation model (HS) [15], for color images.

For a subjective comparison, co-saliency maps for several image sets are shown in Fig. 2. For a fair comparison, all co-saliency maps are normalized into the same range of [0, 255]. As can be seen in Fig. 2, for some simple scenes such as the first example with red caps, most models can highlight the co-salient objects. But compared with other models, our model can suppress background regions more effectively. Some complicated image sets, in which some background regions show similar colors with the co-salient object such as the second and third example in Fig. 2, make all the other co-saliency models quite confusing. Nonetheless, the geometric features as well as the depth features between the co-salient objects and those background regions are different, and thus our model can effectively suppress those background regions. For heterogeneous co-salient objects such as the person in the rightmost example of Fig. 2, our model can effectively exploit the relevance of depth cues to cluster different co-salient object regions and highlight the complete co-salient objects with well-defined boundaries. Besides, we randomly added several irrelevant images to each image set in the dataset, and Fig. 3 shows the co-saliency detection results of two image sets. Although the irrelevant images

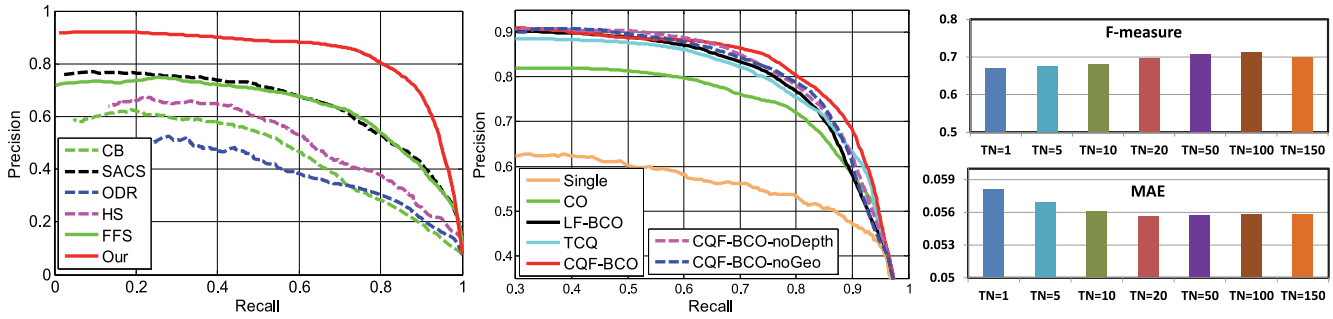

Fig. 4. (Better viewed in color) (a) PR curves of different co-saliency models, (b) performance contribution of each component in our model, and (c) the effect of *TN* on the performance of our model.

TABLE I
F-MEASURE (LARGER IS BETTER) AND MAE SCORE (SMALLER IS BETTER) OF DIFFERENT CO-SALIENCY MODELS

Model	CB	SACS	ODR	HS	FFS	Our
F-measure	0.451	0.473	0.386	0.480	0.531	**0.709**
MAE	0.119	0.200	0.176	0.134	0.141	**0.056**

contain other objects, all regions are effectively suppressed in their co-saliency maps, indicating the robustness of our model to irrelevant images.

We evaluated the co-saliency detection performance using the commonly used precision-recall (PR) curve, the F-measure by setting its coefficient β^2 to 1, and the MAE score which is defined as the average absolute difference between the obtained saliency map and the ground truth. As shown in Fig. 4(a), our model consistently outperforms all the other models with large margins. Besides, as shown in Table I, our model achieves the highest F-measure and the lowest MAE score. This demonstrates that our model can achieve the better co-saliency detection performance than other models for RGBD image sets.

We further evaluated the performance contribution of different parts in our model. Several different versions of our model are denoted as follows.

1) CO: CCS using (1) based on single clustering with all features (a total dimension of 24).
2) LF–BCO: bagging-based final co-saliency map by linearly fusing multiple WCS maps using the same weight.
3) TCQ: CCS using (1) based on the feature subset selected by the largest CQ value.
4) CQF–BCO: bagging-based final co-saliency map by CQ-based fusion using (2) and (3), i.e., the final output of our model.

The PR curves of the above versions are shown in Fig. 4(b) and some analyses are presented as follows.

1) The PR curve of CO is obviously higher than that of RGBD single saliency [20]. It indicates that with the relatively low-quality RGBD single saliency maps, the co-salient regions can be united together via the clustering process and then effectively highlighted by using (1).
2) The PR curve of LF–BCO is higher than that of CO. As the feature bagging can randomly generate multiple feature subsets and different clustering results, a simple linear fusion strategy using the same weight can boost the performance of single clustering (CO). Such a performance elevation demonstrates the effectiveness of our bagging-based clustering method.
3) The PR curves indicate that TCQ outperforms CO and CQF–BCO outperforms LF–BCO. We can conclude that the proposed CQ criterion can find out the most discriminative features for clustering in different scenes, and helps to adaptively weight more the better ones of WCS maps to achieve the performance improvement.

Besides, we also tested the versions of our model without depth features and geometric features, respectively. As show in Fig. 4(b), both versions have performance degradation. Besides, the F-measures/MAE scores without depth features (CQF–BCO-noDepth) and without geometric features (CQF–BCO-noGeo) are 0.649/0.061 and 0.656/0.059, respectively. Compared with the F-measure/MAE score of our model (CQF–BCO), 0.709/0.056, we can also see the performance degradation of both versions. This verifies the importance of depth features and geometric features for RGBD co-saliency detection.

The effect of *TN* on the performance of our model is shown in Fig. 4(c). We gradually increase the number of *TN* from 1 to 150 ($BN = 150$) in our implementation. It can be seen from Fig. 4(c) that the performance of our model achieves the best performance around $TN = 50$.

Our model is implemented using MATLAB on a PC with an Intel Core i7 4.0 GHz CPU and 16-GB RAM. The average processing time per image with a resolution of 640×480 is 5.58 s, including 3.56 s for generating candidate object regions, 1.95 s for bagging-based clustering and 0.07 s for CQ-based fusion.

IV. CONCLUSION

This letter proposes a novel RGBD co-saliency model via bagging-based clustering. RGBD single saliency maps and region pre-segmentation are utilized to generate candidate object regions. Then the clustering via feature bagging is performed repeatedly to calculate multiple WCS co-saliency measures at cluster level. Finally, the CQ is evaluated for adaptively fusing multiple WCS maps in a discriminative way. Experimental results on the RGBD co-saliency dataset demonstrate that our model consistently outperforms the state-of-the-art co-saliency models.

REFERENCES

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[2] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. 2013 IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2083–2090.

[3] Z. Liu, W. Zou, and O. Le Meur, "Saliency tree: A novel saliency detection framework," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 1937–1952, May 2014.

[4] M. M. Cheng, N. J. Mitra, X. Huang, P. Torr, and S. M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[5] K. Fu, C. Gong, I. Y.-H. Gu, and J. Yang, "Normalized cut-based saliency detection by adaptive multi-level region merging," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5671–5683, Dec. 2015.

[6] C. Gong, D. Tao, W. Liu, S. J. Maybank, M. Fang, K. Fu, and J. Yang, "Saliency propagation from simple to difficult," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2531–2539.

[7] Z. Wang and R. Liu, "Semi-supervised learning for large scale image cosegmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2012, pp. 393–400.

[8] W. Wang, J. Shen, X. Li, and F. Porikli, "Robust video object cosegmentation," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3137–3148, Oct. 2015.

[9] M. Cho, S. Kwak, C. Schmid, and J. Ponce, "Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1201–1210.

[10] P. Siva and T. Xiang, "Weakly supervised object detector learning with model drift detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 343–350.

[11] D. E. Jacobs, D. B. Goldman, and E. Shechtman, "Cosaliency: Where people look when comparing images," in *Proc. 23nd Annu. ACM Symp. User Interface Softw. Technol.*, Oct. 2010, pp. 219–228.

[12] H. Li and K. Ngan, "A co-saliency model of image pairs," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3365–3375, Dec. 2011.

[13] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3766–3778, Oct. 2013.

[14] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng, "Self-adaptively weighted co-saliency detection via rank constraint," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4175–4186, Sep. 2014.

[15] Z. Liu, W. Zou, L. Li, L. Shen, and O. Le Meur, "Co-saliency detection based on hierarchical segmentation," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 88–92, Jan. 2014.

[16] L. Ye, Z. Liu, J. Li, W.-L. Zhao, and L. Shen, "Co-saliency detection via co-salient object discovery and recovery," *IEEE Signal Process. Lett.*, vol. 22, no. 11, pp. 2073–2077, Nov. 2015.

[17] H. Li, F. Meng, and K. N. Ngan, "Co-salient object detection from multiple images," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1896–1909, Dec. 2013.

[18] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *Proc. 12th Eur. Conf. Comput. Vis.*, Oct. 2012, pp. 101–115.

[19] R. Ju, Y. Liu, T. Ren, L. Ge, and G. Wu, "Depth-aware salient object detection using anisotropic center-surround difference," *Signal Process. Image Commun.*, vol. 38, no. 10, pp. 115–126, Oct. 2015.

[20] H. Song, Z. Liu, H. Du, and G. Sun, "Depth-aware saliency detection using discriminative saliency fusion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2016, pp. 1626–1630.

[21] H. Fu, D. Xu, S. Lin, and J. Liu, "Object-based RGBD image co-segmentation with mutex constraint," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4428–4436.

[22] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.

[23] T. K. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *Int. J. Comput. Vis.*, vol. 43, no. 1, pp. 29–44, Feb. 2001.

[24] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.

[25] J. Bezdek, *Pattern Recognition With Fuzzy Objective Function Algorithms*. New York, NY, USA: Plenum, 1981.

[26] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math, Statist., Probability*, Jun. 1967, pp. 281–297.

[27] N. Otsu, "A threshold selection method from gray-level histograms," *Trans. Syst. Man Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[28] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *Proc. 2012 IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 478–485.