Salient Object Detection Using Array Images

Tingtian Li and Daniel P. K. Lun
Department of Electronic and Information Engineering
The Hong Kong Polytechnic University
Kowloon, Hong Kong
tingtianpolyu.li@connect.polyu.hk, enpklun@polyu.edu.hk

***Abstract*—Most existing saliency detection methods utilize low-level features to detect salient objects. In this paper, we first verify that the foreground objects in the scene can be an effective cue for saliency detection. We then propose a novel saliency detection algorithm which combines low level features with high level object detection results to enhance the performance. For extracting the foreground objects in a scene, we first make use of a camera array to obtain a set of images of the scene from different viewing angles. Based on the array images, we identify the feature points of the objects so as to generate the foreground and background feature point cues. Together with a new K-Nearest Neighbor model, a cost function is developed to allow a reliable and automatic segmentation of the foreground objects. The outliers in the segmentation are further removed by a low-rank decomposition method. Finally, the detected objects are fused with the low-level object features to generate the saliency map. Experimental results show that the proposed algorithm consistently gives a better performance compared to the traditional methods.**

I. INTRODUCTION

Saliency detection is a fundamental problem in computer vision. Traditional methods focus on estimating salient objects using the low-level features such as texture and position [1, 2]. An alternative approach namely, co-saliency detection [3-5], attempts to estimate saliency maps from two or more images that contain similar foregrounds. Because these methods utilize inter-image information, they are able to simultaneously estimate the salient regions for multiple similar images [3-5]. In [4], multiple fixed windows are used to assist with object identification in different images, and a voting strategy is used to generate the saliency map. In [3], a low-rank model is used to determine the weights for combining different single-image saliency detection results. Recently, it is found that the depth map can be used to assist in saliency map estimation [6, 7]. Since the depth maps are estimated from multiple images, these methods can also be regarded as a special kind of multiple-image-based saliency detection approach. However, it remains a challenge to obtain highly accurate depth estimates solely from multi-view images, especially at the object boundaries.

The increasing availability of low-cost cameras motivate the use of camera arrays for various applications [8]. Through the redundancy obtained from the captured multi-view images, it is possible to obtain more accurate depth information of a scene so that many three-dimensional (3D) operations can be efficiently carried out. In this paper, a novel saliency detection algorithm using multiple images captured by a camera array as shown in Fig. 1 is proposed. The new algorithm is based on the idea that the foreground objects have a high probability to be the salient ones in a scene. In this paper, we first verify such idea by studying two image databases commonly used for saliency detection research. For better extracting the foreground objects, a new optimization cost function based on the foreground and background feature point cues as well as a K-Nearest Neighbor (KNN) model is proposed. We then use a low-rank decomposition method to refine the result and fuse with the low-level features to detect the salient objects. Experimental results show that the new algorithm improves over the traditional methods consistently.

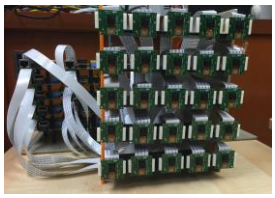


Fig. 1. The 5x5 camera array system developed by our team.

II. THE FOREGROUND OBJECTS AND THE SALIENT OBJECTS

From many experimental results, we observe that foreground objects often become the saliency of an image since they show distinct appearance from the background. To verify it, we analyze two databases CSSD [9] and SED1 [10] commonly used in saliency detection research. Since these two datasets contain ground truth salient objects, we can compare the ground truth with the foreground objects in each image. The analysis results are shown in Table I. We find that over 90% of the images have the ground truth and the foreground objects totally or partially matched. Actually, the reason for most totally non-matched cases is there are no foreground objects. The result inspires us that we can include the foreground objects as a kind of high-level cue with the traditional low-level features in the saliency detection algorithms.

	Totally matched	Partially matched	Totally non-matched
CSSD	71%	26%	3%
SAD1	78%	13%	9%

Table I. The extent of matching between foreground objects and ground truth salient objects.

III. FOREGROUND OBJECT EXTRACTION

Accurately extracting the shape of foreground objects is a challenging task. Traditional methods often cut along edges of

large contrast but can miss the true object boundaries. Here we propose an efficient object segmentation algorithm based on a new energy cost function. The new cost function makes use of the feature point cues obtained from the array images and a new KNN model that allow the shapes of the objects to be more reliably detected.

A. *Feature Point Cues*

Given an array image set, we have shown in [11] that the feature points of the background and foreground objects in a scene can be detected by applying the Random Sample Consensus (RANSAC) algorithm to the array images. To remove the outliers when classifying the feature points, the Ordering Points to Identify the Clustering Structure (OPTICS) algorithm [12] is adopted. Fig. 2 shows an example of the classification result. It can be seen that the feature points of the foreground are well detected. Based on the detected feature points, we define the foreground feature point cue as follows:

$$C_{FFP}(p) = \alpha \cdot \sum_j g(p, f_j), \tag{1}$$

$$g(p, f_j) = \begin{cases} \exp(-d(p, f_j)/2\delta^2) & if\ d(p, f_j) < \rho \\ 0 & else \end{cases} \tag{2}$$

where $C_{FFP}(p)$ denotes the foreground feature point cue for pixel p of an image selected from the array image set of which the foreground object is to be detected. $d(p, f_j)$ represents the Euclidean distance between pixel p and a foreground feature point f_j; α and δ are two strength-controlling constants, and ρ is a threshold. The above foreground feature point cue is defined based on the assumption that the region around a foreground feature point has a high probability of also within the foreground. However, this probability should decrease as the distance from the feature point increases. Based on the same principle, we can define the background feature point cue $C_{BFP}(p)$ in a similar way. The foreground and background cues are used in the proposed energy cost function described below.

B. *New Energy Cost Function*

For foreground object segmentation, the Markov Random Field (MRF) technique has been generally adopted. However, traditional MRF based algorithms often fall into the trap of cutting along wrong high-contrast boundaries thus some dense outliers can connect to the segmentation target. Additional manual pixel labeling is needed to correct the errors. Similar to the traditional MRF, the proposed energy function in (3) is defined so that its minimum corresponds to a good segmentation of the foreground object. However, we make use of the array images to generate the feature point cues as in (1) for using in the data term D_p of the cost function. Besides, as inspired by the recent research on KNN matting [13], we define the smoothness term $V_{p,q}$ of the cost function in a way that the coherence of a few nearby and similar pixels is considered instead of restricting to the neighboring pixels as in MRF.

$$\arg\min_L E(L) = \sum_{p \epsilon P} D_p(L_p) + \sum_{p,q \epsilon SKNN} V_{p,q}(L_p, L_q) \tag{3}$$

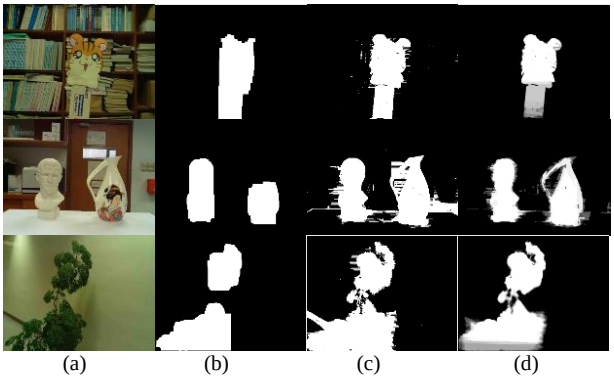
$$D_p(L_p) = C_{FFP}(p)(1 - L_p) + C_{BFP}(p) L_p \tag{4}$$

$$V_{p,q}(L_p, L_q) = \gamma \exp(-\|s(p) - s(q)\|_2^2)[L_p \neq L_q] \tag{5}$$

(a) (b) (c)

Fig. 2. (a) The original image. (b) Initial result of foreground feature point detection. (c) Result after further refinement with OPTICS.

In (3), L and P are the label set and pixel set of the entire image. $L_p \in L$ is the label of pixel $p \in P$. We set $L_p = 0$ and 1 if p is classified as background and foreground respectively. As shown in (4), the data term D_p will penalize the cost function if a wrong classification is made. In (5), $[L_p \neq L_q] = 1$ if $L_p \neq L_q$; and 0 otherwise. Thus the term $V_{p,q}$ of two pixels p and q will be zero if they have the same label. Otherwise, $V_{p,q}$ is evaluated based on their difference in s, where $s(p) = [R(p)\ G(p)\ B(p)\ x(p)\ y(p)]^T$ is a vector of the RGB value and position of p. Note that $E(L)$ in (3) is evaluated by accumulating $V_{p,q}$ for all pixel pairs $\{p, q\}$ within the *SKNN* set, which is defined as the set of K nearest neighboring pixels of p measured by the similarity in RGB values and distance. Normally, all pixels within the *SKNN* set should have the same label due to the smoothness of object texture. If a pixel q within the set is wrongly classified, the classification of p will still follow the majority in the set since $V_{p,q}$ is small. In the situation that p is wrongly classified such that it is different from most other in the set, a large sum of $V_{p,q}$ will be generated. It penalizes the cost function and forces the label of p to change. The proposed KNN model is particularly effective at object boundaries where incorrect classifications happen frequently. Since the KNN model will cluster similar pixels, a pixel at the boundary tends to follow the classification of the same object. It is on contrary to the traditional MRF which considers only the neighboring pixels where the classification can be rather random at the object boundaries. The minimization of the cost function in (3) can be achieved by using the max-flow/min-cut methods. Fig. 3(a) and (b) show the results using the traditional MRF and the proposed approach. It is seen that the shape of the objects is better retrieved by the proposed approach. However, some outliers appear which can be further removed using a low-rank decomposition method as described below.



(a) (b) (c) (d)

Fig. 3. (a) Input image. (b) Segmentation results using the traditional neighboring graph structure method. (c) Segmentation results using the proposed method. (d) Results refined by the low rank decomposition.

IV. Low-Rank Decomposition to Refine the Results

To remove the outliers, we first align the detected masks to the perspective of the reference view. It is achieved by means of the homographies estimated from the object feature points obtained in Section III. Because of parallax, alignment of the foreground implies misalignment of the background. So the foreground part of the aligned masks will be very similar, thus having low rank; the background outliers however can be quite different, thus having high rank. By retaining only the low rank part of the aligned masks, we can remove the background outliers. More specifically, let $M = [vec(W_1A_1), vec(W_2A_2) \dots vec(W_nA_n)]$, where W_i is the alignment operator for the ith array image; A_i is the ith mask estimated in Section III. $vec(.)$ is an operator that converts an image into a column vector. So M is a matrix of which each column represents an aligned foreground mask developed in Section III. An objective function is then defined as follows:

$$\arg\min_{Z} J(Z,S) = \|Z\|_* + \xi\|S\|_1, \quad \text{s.t. } M=Z+S \tag{6}$$

where Z is the low-rank part of M, which is the foreground region that we desire; S is the sparse background outliers; and ξ is a constant. Rather than directly minimizing J according to the rank of Z, we consider the nuclear norm of Z and the l_1-norm of S in (6) to allow the problem to be solved using a convex approach. We use the Augmented Lagrange Multiplier (ALM) method plus the Alternating Direction Method (ADM) to solve (6). Fig. 3(d) shows the refined segmentation results. It can be seen that almost all outliers are removed.

We also define a confidence value in (7) for measuring the certainty of each detected object,

$$C = \begin{cases} mean(\tilde{Z})exp(-var(\tilde{Z})/\tau_1^2) & if\ \tilde{Z} \neq \emptyset \\ 0 & if\ \tilde{Z} = \emptyset \end{cases} \tag{7}$$

$$\tilde{Z} = \{Z | Z > 0\}$$

where τ_1 is a constant. $\tilde{Z}$ with large mean and low variance mean the intensity is high and stable. Note that $C = 0$ if the scene consists of only a plane without any foreground object.

V. Fusion with Low Level Features

With the refined foreground objects Z and the confidence value C, the final saliency map can be obtained by fusing with some low level features used in the traditional Robust Background Detection (RBD) [14] saliency detection method.

$$\min O(s_i) = \sum wb_i s_i^2 + \sum wf_i (s_i - 1)^2 + \sum w_{ij}(s_i - s_j)^2, \tag{8}$$

$$wb_i = (1 - \tilde{\xi}_w) w_{Bnd}(s_i) + \tilde{\xi}_w (1 - F(s_i)) \tag{9}$$

$$wf_i = (1 - \tilde{\xi}_w) w_{Ctr}(s_i) + \tilde{\xi}_w F(s_i) \tag{10}$$

$$\tilde{\xi}_w = \min(\gamma C,\ 0.8) \tag{11}$$

In (8), s_i represents the results of region division using SLIC [15]. $F(s_i)$ is the sum of the values of Z in region s_i. wb_i and wf_i are the weights for non-salient and salient regions, respectively. $w_{Bnd}(s_i)$ measures the extent to which region s_i touches the image boundary and $w_{Ctr}(s)$ is a measure of the region contrast as discussed in RBD. In (10), $\tilde{\xi}_w$ is an adaptive weight that is used to balance the low-level and object-level features. It has a large value if the object confidence is high (after scaling by a constant γ), and is capped to 0.8 to ensure the low-level features can assist in detecting the salient objects.

VI. Experiments and Evaluation

A series of experiments were carried out to compare the performance of the proposed algorithm with the state-of-the-art saliency detection methods. Since there is no existing multi-view image dataset for saliency detection, we captured 30 groups of images using our camera array, and manually labeled the ground truths. We compared our method with 8 single-image saliency detection methods. They include: context-aware saliency detection (CA) [16], dense and sparse reconstruction (DSR) [17], absorbing Markov chain (MC) [18], PCA [19], RBD, sparse signal mixing (SS) [20], region covariances (COV) [21], and hierarchical saliency model (HS) [9]. Two co-saliency detection methods were also compared, which include: self-adaptively weighted co-saliency detection (SWC) [3] and cluster-based co-saliency detection (CO) [5]. We also compared with the depth-enhanced saliency detection method (DES) [6] with the depth map obtained using the classic method described in [22]. Fig. 4 shows the quantitative comparison results evaluated in the situation that foreground objects exist. Two criteria are considered: the precision-recall (PR) curve, and the receiver operator characteristic (ROC) curve. It can be seen that the proposed method always achieve the best results. Fig. 5 shows the qualitative comparison results evaluated in the situations that foreground objects exist. It is seen that the proposed method enhances the intensities of the salient object and has much fewer background outliers in the saliency map. Note that in the situations that the scene has no foreground objects, the proposed algorithm will be similar to RBD according to (8)-(11). Fig. 6 shows the results for scenes without foreground objects to verify this.

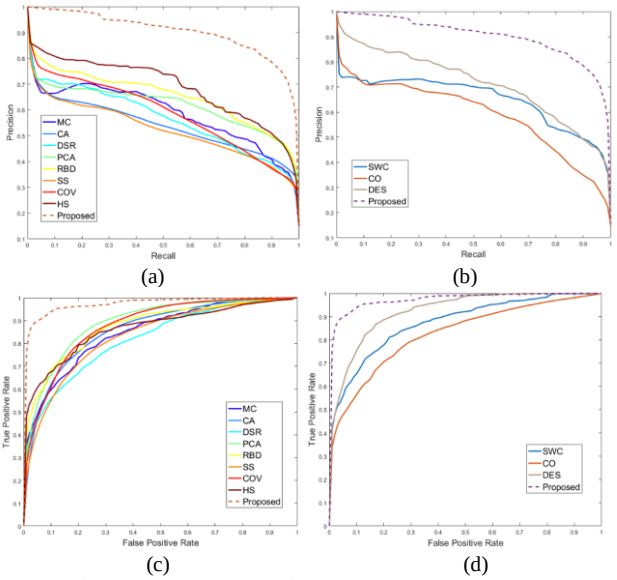


Fig. 4. Left: PR and ROC curves of single-image-based methods. Right: PR and ROC curves of multiple-image and depth-based methods.

Fig. 6. Saliency detection results for scenes without foreground objects.

VII. Conclusions

In this paper, we investigated the relationship between the foreground objects and salient objects, and proposed a new saliency detection algorithm by combing the high level foreground object detection result with the low-level object features. For foreground object segmentation, we proposed the reliable feature point cue and a novel KNN model to better the process. A low-rank decomposition method is then applied to remove the outliers. After fusing with low-level features, the proposed approach showed impressive performances in identifying the salient objects over the traditional approaches.

Acknowledgment

This research work was fully supported by the Hong Kong Polytechnic University under research account G-YBK8.

References

[1] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," *ACM MM*, pp. 374-381, 2003.

[2] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *PAMI*, vol. 20, pp. 1254-1259, 1998.

[3] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng, "Self-adaptively weighted co-saliency detection via rank constraint," *TIP*, vol. 23, pp. 4175-4186, 2014.

[4] H. Li, F. Meng, and K. N. Ngan, "Co-salient object detection from multiple images," *IEEE TMM*, vol. 15, pp. 1896-1909, 2013.

[5] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *TIP*, vol. 22, pp. 3766-3778, 2013.

[6] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," *ICIMCS*, p. 23, 2014.

[7] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," *CVPR*, pp. 454-461, 2012.

[8] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, *et al.*, "High performance imaging using large camera arrays," *ACM TOG*, pp. 765-776, 2005.

[9] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," *CVPR*, pp. 1155-1162, 2013.

[10] S. Alpert, M. Galun, A. Brandt, and R. Basri, "Image segmentation by probabilistic bottom-up aggregation and cue integration," *PAMI*, vol. 34, pp. 315-327, 2012.

[11] T. Li and D. P. K. Lun, "Super-resolution imaging with occlusion removal using a camera array," *ISCAS*, pp. 2487-2490, 2016.

[12] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: ordering points to identify the clustering structure," in *ACM Sigmod Record*, pp. 49-60, 1999.

[13] Q. Chen, D. Li, and C.-K. Tang, "KNN matting," *PAMI*, vol. 35, pp. 2175-2188, 2013.

[14] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *CVPR*, pp. 2814-2821, 2014.

[15] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *PAMI*, vol. 34, pp. 2274-2282, 2012.

[16] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *PAMI*, vol. 34, pp. 1915-1926, 2012.

[17] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," *ICCV*, pp. 2976-2983, 2013.

[18] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, "Saliency detection via absorbing markov chain," *ICCV*, pp. 1665-1672, 2013.

[19] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?," *CVPR*, pp. 1139-1146, 2013.

[20] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *PAMI*, vol. 34, pp. 194-201, 2012.

[21] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *Journal of Vision*, vol. 13, pp. 11-11, 2013.

[22] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," *CVPR*, pp. 807-814, 2005.

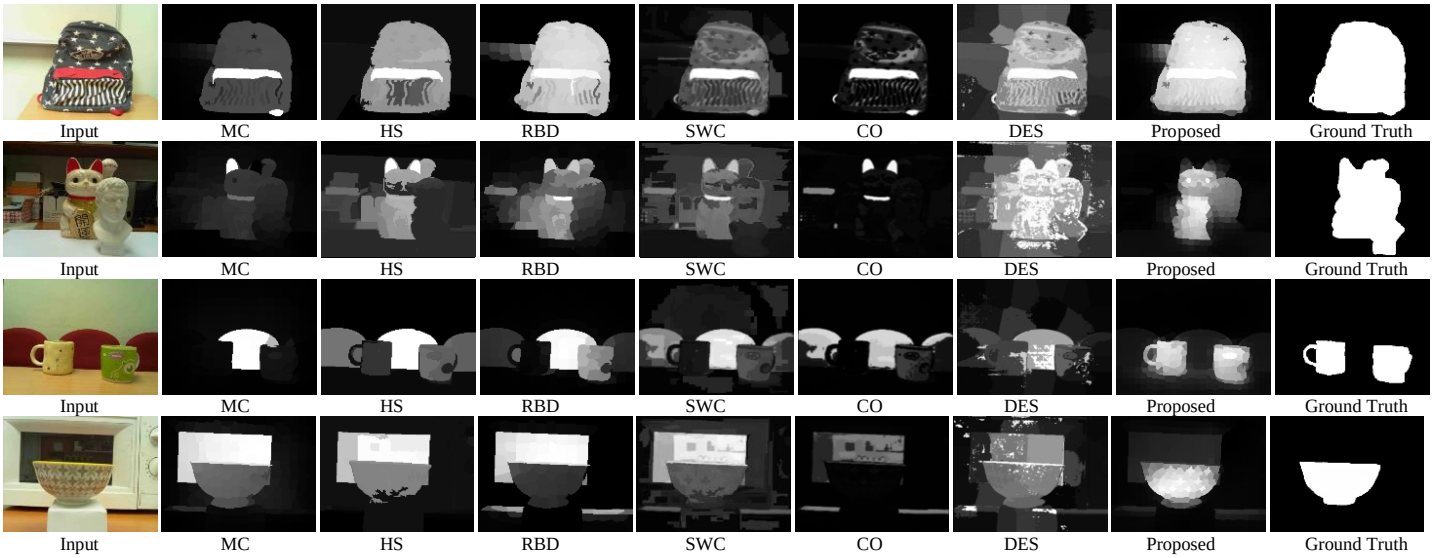


Fig. 5. Examples of the saliency maps obtained using various state-of-the-art methods and our proposed method in situations with foreground objects.