Vis Comput (2016) 32:111–121
DOI 10.1007/s00371-014-1059-6

ORIGINAL ARTICLE

Depth incorporating with color improves salient object detection

Yanlong Tang · Ruofeng Tong · Min Tang · Yun Zhang

Published online: 20 January 2015

Abstract Detecting salient objects in challenging images attracts increasing attention as many applications require more robust method to deal with complex images from the Internet. Prior methods produce poor saliency maps in challenging cases mainly due to the complex patterns in the background and internal color edges in the foreground. The former problem may introduce noises into saliency maps and the later forms the difficulty in determining object boundaries. Observing that depth map can supply layering information and more reliable boundary, we improve salient object detection by integrating two features: color information and depth information which are calculated from stereo images. The two features collaborate in a two-stage framework. In the object location stage, depth mainly helps to produce a noise-filtered salient patch, which indicates the location of the object. In the object boundary inference stage, boundary information is encoded in a graph using both depth and color information, and then we employ the random walk to infer more reliable boundaries and obtain the final saliency map. We also build a data set containing 100+ stereo pairs to test the effectiveness of our method. Experiments show that our depth-plus-color based method significantly improves salient object detection compared with previous color-based methods.

Keywords Salient object detection · Depth information · Color information · Stereo images

1 Introduction

Saliency detection is a process that computer imitates HVS (human visual system) to understand scenes in images. When human observes an image, he/she always focus on a subset of the whole image. Saliency detection is a mechanism to filter out irrelevant information and highlight most noticeable foreground regions. Driven by different applications, saliency detection can be divided into two categories: fixation prediction and salient object detection. The former aims at predicting locations in a scene that human eyes may fixate and could be used in active gaze control [3,4], robot localization [48], recognition [44] and advertising [43]. A series of works [5,19,23,25,27,45] develop various models for fixation detection. The latter aims to automatically identify the most salient object, which is mainly applied in object segmentation [14,22,29], object recognition [2], image retrieval [7,8,26,33] and image editing [37]. Many works [6,11,20,28,31,35,36,47,51] devote to detecting salient object in a scene more accurately and efficiently. However, recent work [51] states that previous methods may be not robust enough in challenging cases, as illustrated in Fig. 1.

Previous monocular methods often utilize 2D features, such as color, orientation to measure saliency, without introducing depth feature. Lang et al. [30] point out that there is slight difference between human visual saliency and 2D

Y. Tang (✉) · R. Tong · M. Tang
The State Key Lab of CAD&CG, Institute of Artificial Intelligence,
Zhejiang University, Hangzhou 310027, China
e-mail: yanlongtang@gmail.com

R. Tong
e-mail: trf@zju.edu.cn

M. Tang
e-mail: tang_m@zju.edu.cn

Y. Zhang
Zhejiang Institute of Radio and TV Technology, Zhejiang University of Media and Communications, Hangzhou 310018, China
e-mail: zhangyun_zju@zju.edu.cn

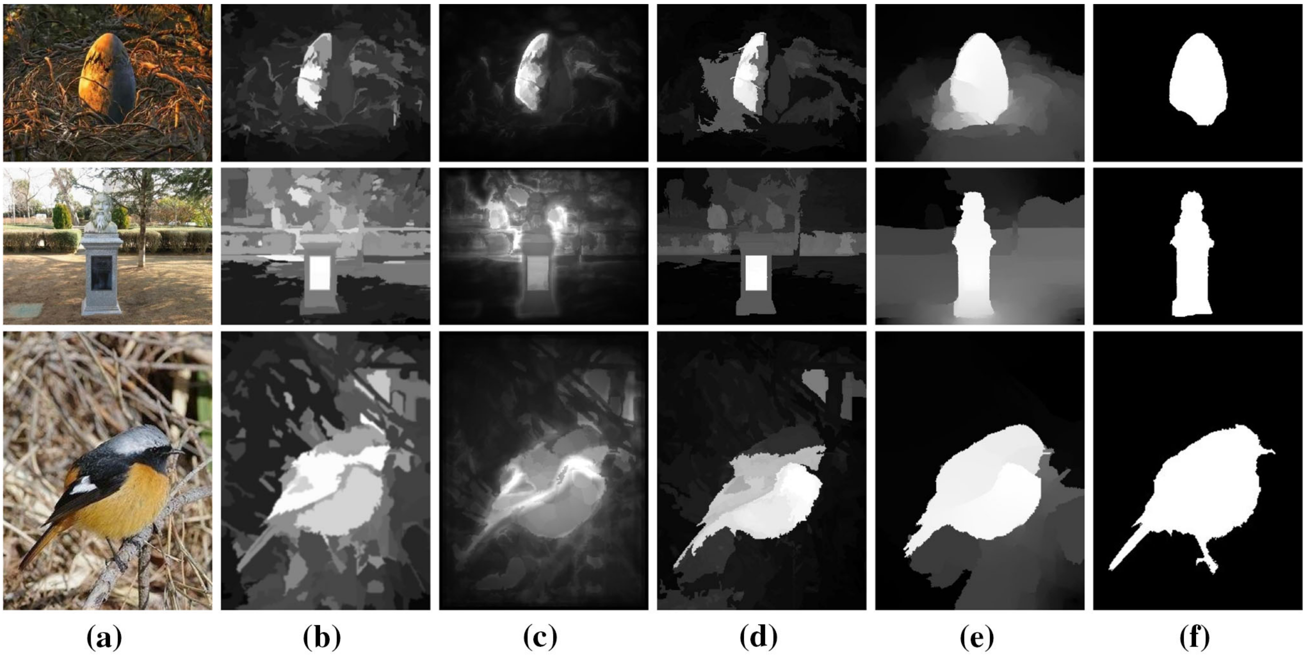


Fig. 1 Salient object detection in challenging images, which contain complex patterns in the background and internal color edges in the foreground. **a** Original image. **b** RC [11]. **c** PCA [36]. **d** HS [51]. **e** Ours. **f** Ground truth

saliency metric, and human always focuses on the depth information when evaluating saliency in a scene. Stereopsis not only attracts ever-increasing attention in stereo image/video editing and analysis [15,16,38,39,50], but also attracts more and more attention in saliency detection, as it supplies additional depth cue. There are many attempts in saliency detection by taking depth into consideration. In the fixation prediction field, Lang et al. [30] study the discrepancies in the eye fixation data when human viewing 2D/3D scenes and propose a model to improve the saliency detection with depth priors. Fang et al. [17] improve the saliency detection for stereo images using four features (color, luminance, texture and depth) extracted from DCT coefficients. Ciptadi et al. [12] measure saliency with the constructed 3D layout and shape features from depth maps. In the salient object detection literature, Niu et al. [40] define saliency using depth information computed from stereo images, and their results show that stereo saliency is a useful component to previous visual saliency analysis. However, they do not consider the color feature which is very important in saliency analysis. In this paper, we combine color and depth feature to significantly improve saliency detection for challenging scenes. Desingh et al. [13] propose a learning-based model to fuse depth information captured by depth sensor and color information to measure saliency. This model improves salient region detection in indoor settings, in which the ground truth contains multiregions instead of a single unambiguous object.

In this work, we aim at salient object detection. Different from works [13,40], we introduces both depth and color features to detect a single unambiguous object in a scene. We use depth information in a different way basing on the characteristics of salient object [28]:

1. The salient object is always different from its surrounding context.
2. The salient object is most probably placed near the center of an image.
3. A salient object is located in a limited depth range in a scene (Our observation).
4. A salient object always has a well-defined closed boundary.

In our saliency model, we combine color and depth features to detect salient object. We notice that the depth map computed by stereo matching [49] can provide depth layering and objects' shape information. The former can help to locate salient objects more accurately and the latter can be used to control and refine the border of salient objects.

Our salient object detection includes two stages: object location and object boundary inference. Both stages combine color and depth features for better results according to the 4 characteristics of salient objects. The detection result is given by a saliency map, in which the intensity (normalized to [0, 1]) of each pixel denotes the probability of belonging to the salient region, and the final saliency quality is evaluated by precision rate and recall rate. Experiments show that our results can achieve high precision rate and recall rate at the same time.

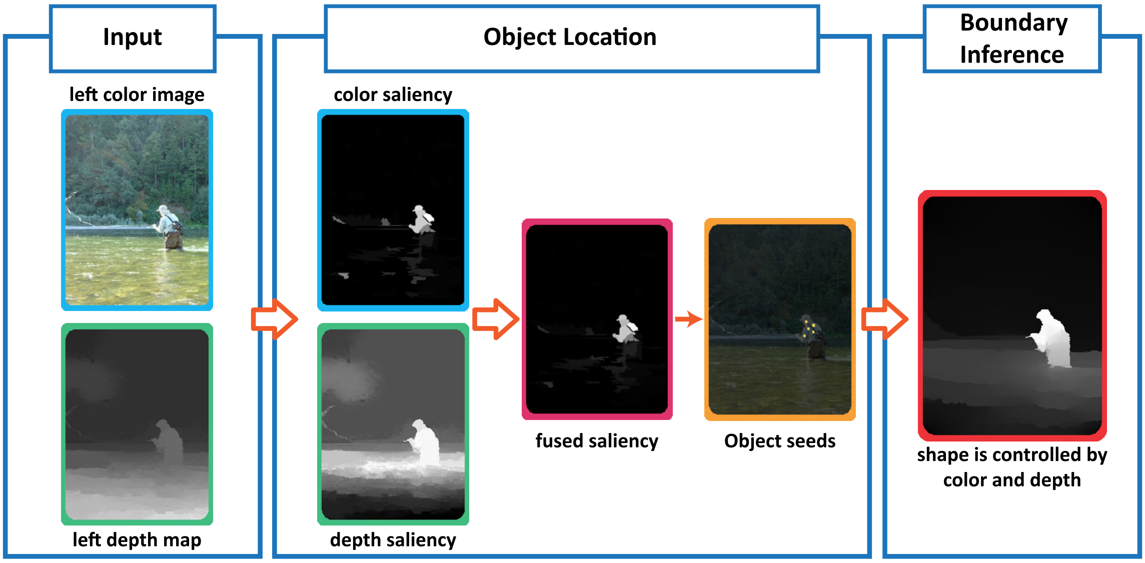


Fig. 2 Framework of our method. The *left* color image and *left* depth map which is precomputed from stereo images are considered as inputs. In the salient object location stage, we sample a few seeds from the extracted salient patch to identify the object location. Firstly, a color saliency map is produced from *left* color image based on color contrast observation. Then, we calculate the depth saliency map from the *left* depth image, according to the depth layering information. The two maps are finally combined to produce a fused saliency map, from which the salient patch is correctly extracted. In the object boundary inference stage, with the predetermined seeds, the random walk [21] is employed to produce the final saliency map, in which plausible object boundary is given

The contribution of this work is that we propose a model that improves salient object detection by incorporating both depth and color information. Our method is divided into two independent stages: object location and object boundary inference. In the first stage, we produce a noise-free salient patch by fusing the color saliency map and depth saliency map. In the second stage, we define a graph using both depth and color features, and then extend the random work [21] to inferring more reliable object boundary.

2 Methodology

2.1 Overview

In this work, we mainly explore how color and depth information work together to produce better saliency maps, because depth map supplies additional layering information and more reliable boundary. In this work, the depth map is produced from stereo images that describe the same scene by stereo matching [49], and our method consists of two stages: object location and object boundary inference. Both stages use color and depth information to improve accuracy. The workflow is shown in Fig. 2.

- In the object location stage, a noise-filtered salient patch is produced for sampling a few object seeds. This salient patch is generated from two initial saliency maps, which are generated by applying color and depth information, respectively.
- In the object boundary inference stage, we represent the image as a rectangular plane with different thermal conductivity at different positions. With the object seeds' temperature (saliency value) known as 1 and image corners' temperature assumed to be 0, our task is to infer the temperature of the rest positions, which is formulated as a Combinatorial Dirichlet problem.

2.2 Object location

We locate the salient object by sampling a few seeds in a small patch detected as part of the salient object. Firstly, an initial rough estimate of the 2D position (in the form of salient patch) of the object is given using color feature. Then, object's depth location (*salient depth*) in depth domain is estimated. Finally, the 2D position estimation is refined by the estimated depth location.

2.2.1 2D location

Inspired by the 1st characteristic of salient object, we use color contrast to locate the salient object. We first compute the saliency using the spatially weighted region contrast [9–11]:

The input color image is firstly segmented into regions using the graph-based method [18]. When measuring the weight of each edge in the segmentation, we consider both color difference in $L^*a^*b^*$ space and depth difference.

Then, the color distance between two regions r_k and r_i is defined as:

$$D_{\mathrm{r}}(r_k, r_i) = \sum_{s=1}^{n_k} \sum_{t=1}^{n_i} f(c_{k,s}) f(c_{i,t}) D(c_{k,s}, c_{i,t}) \tag{1}$$

where $f(c_{k,s})$ is the probability of the sth color $c_{k,s}$ among all n_k colors in the kth region r_k. $D(c_{k,s}, c_{i,t})$ is the color difference in $L^*a^*b^*$ space.

Finally, The spatially weighted region contrast saliency is:

$$S(r_k) = w_{\mathrm{s}}(r_k) \sum_{r_k \neq r_i} \mathrm{e}^{\frac{D_{\mathrm{s}}(r_k, r_i)}{-\sigma_S^2}} w(r_i) D_{\mathrm{r}}(r_k, r_i) \tag{2}$$

where $D_{\mathrm{r}}(r_k, r_i)$ measures the region color contrast, $w(r_i)$ is a weighting term that emphasizes color contrast to bigger regions, $\mathrm{e}^{\frac{D_{\mathrm{s}}(r_k, r_i)}{-\sigma_S^2}}$ refers to the spatial weight which weights more if the region distance $D_{\mathrm{s}}(r_k, r_i)$ is smaller, and $w_{\mathrm{s}}(r_k)$ is the center bias term, indicating that regions around the center are more likely to be salient.

$S(r_k)$ is further improved using non-salient regions prior and color space smoothing operation. Border regions are finally assigned the saliency value 0, as they are typically non-salient background regions, and this operation helps to improve the precision of the saliency map. Color space smoothing, means to replace the saliency value of each color by the weighted average of the saliency values of similar colors. This operation can reduce quantization artifacts and can uniformly highlight the entire salient object. More details of the two improving operations are elaborated in Sections 4.3 and 3.2 of the work [9].

The saliency map produced by the color contrast method gives an initial estimation of the salient object's location (See Fig. 3b). We select pixels with saliency values ranking top 1 % in this map as the salient patch. Observing from challenging images, we find that the selected patch may contain a lot of undesired background parts, thus we aim to further filter out background noises using the depth information in the following steps.

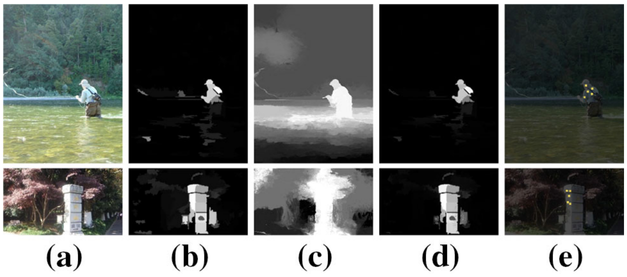


Fig. 3 Object location estimation. **a** Original image. **b** Initial location estimation (color saliency map). **c** Depth saliency mask (map) **d** Refined location (fused saliency map) using depth. **e** Seed points extracted from salient patch

2.2.2 *Depth location*

Each pixel of the extracted candidate patch has a 3D position, $P(x, y, d)$. According to the 3rd characteristic, we give an estimation of where the object locates in the depth domain. In this estimation step, the center prior proposed in the 2nd characteristic is used as follows:

$$d_{\mathrm{s}} = f(P) = \sum_{i=1}^{N} \phi_i d_i \tag{3}$$

$$\phi_i = \frac{\mathrm{e}^{-\sigma P_i}}{\sum_{j=1}^{N} \mathrm{e}^{-\sigma P_j}} \tag{4}$$

$$P_i = \sqrt{(x_i - x_{\mathrm{c}})^2 + (y_i - y_{\mathrm{c}})^2} \tag{5}$$

where $P = (X, Y, D)$ is an $N \times 3$ vector indicating the positions of pixels in the candidate patches. $X = (x_1, x_2, \ldots, x_N)^T$, $Y = (y_1, y_2, \ldots, y_N)^T$, $D = (d_1, d_2, \ldots, d_N)^T$. $(x_{\mathrm{c}}, y_{\mathrm{c}})$ is the image center. Free parameter σ equals 1.0. This depth location d_{s} is named *salient depth.*

2.2.3 *Refined location*

To remove background noises in the initial salient patch, we give a saliency model in the depth domain based on the *salient depth*:

$$S_d(i, j) = \mathrm{e}^{-\frac{\|d(i,j) - d_{\mathrm{s}}\|_1}{\lambda^2}} \tag{6}$$

where $d(i, j)$ is the depth value of location (i, j) and λ is a regulation parameter with default value 0.3. According to this model, saliency value decreases when the location deviates from the *salient depth*. Although this model can highlight the whole object and darken most part of the background, it also introduces undesired noises into background regions at the similar depth level (see Fig. 3c).

To further filter out the undesired background noises, we propose a *multiplication* operation on the color saliency map and depth saliency map.

$$S_p(i, j) = S_c(i, j) \times S_d(i, j) \tag{7}$$

Thus, the salient patch (Fig. 3d) contains much less noise and can be correctly extracted, then we sample 1–5 seeds (see Fig. 3e) from the patch.

2.3 Object boundary inference

With the location of the object seeds, a rough estimate of object location is given, but the boundary of the object is still unknown. Following the 4th characteristic of salient object, we consider the shape information (specially the boundary information) when estimating salient object. As we aim to measure the object saliency not segment objects, it is not necessary to give a strong border of the object. Thus in this stage, the object boundary is inferred in a probabilistic way, meaning that locations with sharp jumps in probabilities (saliency values) are more likely to be boundaries.

Furthermore, depth and color information are combined to infer the border of the salient object. Then a diffusion method (it is a special case of diffusion [21]) is applied to producing the final saliency map, which also displays the inferred object boundary. Our boundary inference is based on the theory of heat diffusion, which is elaborated as follows. The seed points can be treated as a heat source with fixed temperature value 1, and the four corner positions (belongs to salient object with tiny probability) have fixed temperature value 0. The image is represented as a kind of inhomogeneous medium with various thermal conductivities at various positions. With these fix-value seeds, a steady heat map is obtained, which can represent the final saliency map. This problem has been formulated as a Combinatorial Dirichlet problem [21], and the key of this problem is to define an inhomogeneous medium with the object boundary information implicitly encoded.

2.3.1 Boundary information encoding

An instinctive idea of defining this inhomogeneous medium is to assign different thermal conductivity values at different positions. Positions with lower thermal conductivity are more probably considered as the boundary that interdicts the conduction of heat. We use local difference to define local thermal conductivity. Specially, we represent the whole image as a four-connected and undirected graph form $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. $\mathcal{V}$ indicates pixels and $\mathcal{E}$ represents the weights between adjacent nodes. w_{ij} refers to the weight which denotes the similarity between v_i and v_j. The feature of each node is a vector that is composed of color and depth features. Instead of using direct L_2 form to measure the vector difference, we define the feature difference in the following form [see Eq. (7)]. In this metric, both depth discontinuity and color difference are important to represent the similarity of pixel pairs, and the former weights much more. For example, large color difference may indicate internal textures or edges, which are actually not object boundaries.

$$w_{ij} = \mathrm{e}^{-\frac{\|df_i - df_j\|_2 + \beta\|cf_i - cf_j\|_2}{\alpha^2}} \tag{8}$$

where i and j are adjacent pixels (nodes), and df_i, cf_i are the depth and color feature, respectively. For each pixel, depth feature refers to the depth value and color feature is the pixel value in $L^*a^*b^*$ space after the color quantization. α and β are free parameters with experimentally selected value 0.1. The $\|\cdot\|_2$ is normalized to [0, 1].

2.3.2 Infer boundary by diffusion

After obtaining the object's seeds in Sect. 2.2, we set their saliency to 1 and corner points' saliency to 0, and our final step is to decide the saliency of unknown positions in the graph. This problem can be formulated as a steady source heat diffusion, with the medium pre-defined in the previous step. Further, this problem is actually a specific case of the combinational Dirichlet problem [21], as the heat field (saliency map) is steady. Thus, our task is to decide a harmonic function that satisfies the Laplace equation:

$$\Delta S(i, j) = 0 \tag{9}$$

with boundary (seeded positions and corner positions) values 1 and 0.

The harmonic function is formulated as the minimization of a Dirichlet integral:

$$D[S] = \frac{1}{2}\int_{\Omega} |\nabla S|^2 \mathrm{d}\Omega \tag{10}$$

In our case, the combinatorial Laplace matrix is defined as:

$$L_{ij} = \begin{cases} d_i & \text{if } i = j \\ -w_{ij} & \text{if } v_i \text{ and } v_j \text{ are adjacent nodes} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Following the derivation in [21], the Dirichlet integral can be further formulated as:

$$D[S] = \frac{1}{2}S^T L S = \frac{1}{2}\sum_{e_{ij}\in E} w_{ij}(S_i - S_j)^2 \tag{12}$$

According to the saliency values, vertices are divided and reorganized into two groups S_Y and S_N, which are seed and non-seed points, and Eq. 11 can be decomposed as follows:

$$\begin{aligned} D\begin{bmatrix} S_Y \\ S_N \end{bmatrix} &= \frac{1}{2}\begin{bmatrix} S_Y^T & S_N^T \end{bmatrix}\begin{bmatrix} L_Y & B \\ B^T & L_N \end{bmatrix}\begin{bmatrix} S_Y \\ S_N \end{bmatrix} \\ &= \frac{1}{2}(S_Y^T L_Y S_Y + 2S_N^T B^T S_Y + S_N^T L_N S_N) \end{aligned} \tag{13}$$

By taking the partial of $D\begin{bmatrix} S_Y \\ S_N \end{bmatrix}$ with respect to S_N, we obtain:

$$L_N S_N = -B^T S_Y \tag{14}$$

The saliency values S_N of all points in the graph can be calculated by solving the above linear equation system, and the final saliency map S is obtained.

3 Experiments and evaluation

Data collection. To validate our approach, we build a data set of 103 stereo pairs (left and right images). As each image of the pair represents the same scene, we measure saliency on the left image. The depth (disparity) maps are precomputed using stereo matching [49]. The ground truth is a binary map containing a region of interest. We present a statistical analysis of the distribution of salient objects' location in our data set (Fig. 4a), which is also performed in the work [41] (Fig. 4b) to evaluate data set bias. Both data set shows center bias of the distribution. This is because human naturally frames an object of interest near the center of the image when taking pictures, as is stated in [41]. Other public data sets such as MSRA [34] and SSB [12] also show such center bias. Part of this collection is from the Internet and the rest are shot by us in real scenes. In general, challenging images are widely existed in the Internet, and they usually contain complex structures in salient regions or the background, e.g., small-scale high-contrast patterns in the background [51]. In this paper, we test the effectiveness of our saliency detection in challenging images.

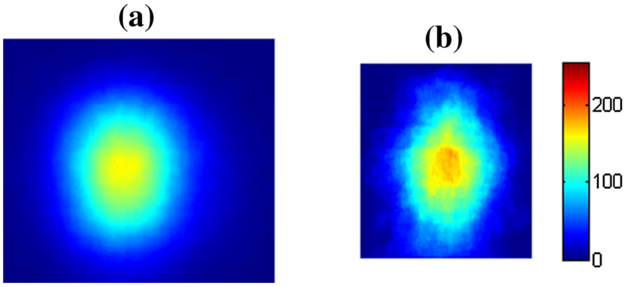


Fig. 4 Location distribution of salient objects. **a** RGBD data set [41]. **b** Our data set. Both data sets show the center bias of salient object distribution

Quantitative evaluation. In the salient object detection literature, different measures from fixation prediction are proposed to evaluate saliency maps. Our method is evaluated by precision–recall curve and F-measure, which are widely used standard evaluation metrics [1,11,51]. The F-measure is formulated as:

$$F_\beta = \frac{(1+\beta^2)\cdot \text{precision}\cdot \text{recall}}{\beta^2\cdot \text{precision} + \text{recall}} \tag{15}$$

In this measure, the range of β^2 is set to (0, 1.0), because recall rate is not as important as precision rate in saliency detection [34]. We set $\beta^2 = 0.3$ in our experiments.

We compare our method with state-of-the-art monocular methods. Like [32], we do not intend to emphasize that our approach outperforms state-of-the-art monocular methods as we use additional depth information. Rather, the goal of the comparison is to verify that additional depth information can greatly improve salient object detection. The target methods to be compared with ours are chosen based on three considerations: (1) State-of-the-art methods: HS [51], PCA [36], GBMR [52], BMS [54] and RC2 [9], which appeared in 2013 and 2014. (2) Methods (RC [11] and CB [28]) that outperform previous ones on the same data set. (3) Other classic methods (SF [42], LR [46], HC [11], FT [1], LC [53]and SR [24]) with high citation rate. To reduce errors, we use the authors' original codes to implement corresponding methods.

The quantitative evaluation results are shown in Fig 5. The precision–recall curve demonstrates that our method can achieve high precision and recall rate at the same time, which indicates that our color + depth based method greatly outper-

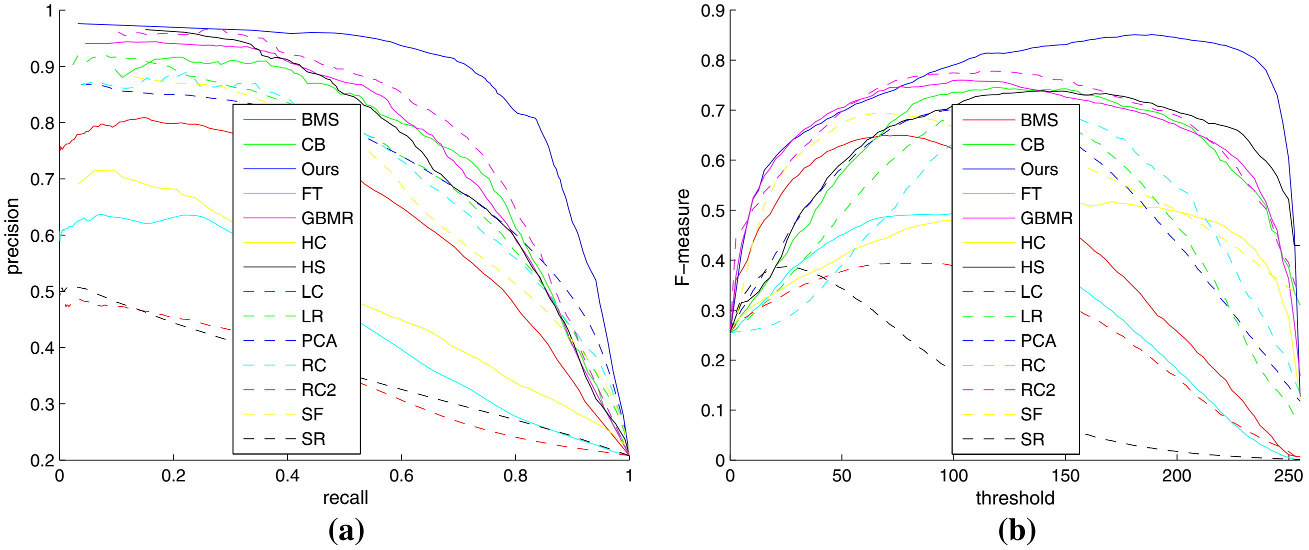


Fig. 5 Quantitative evaluation by comparing with state-of-the-art monocular methods. **a** Precision–recall curve. **b** F-measure

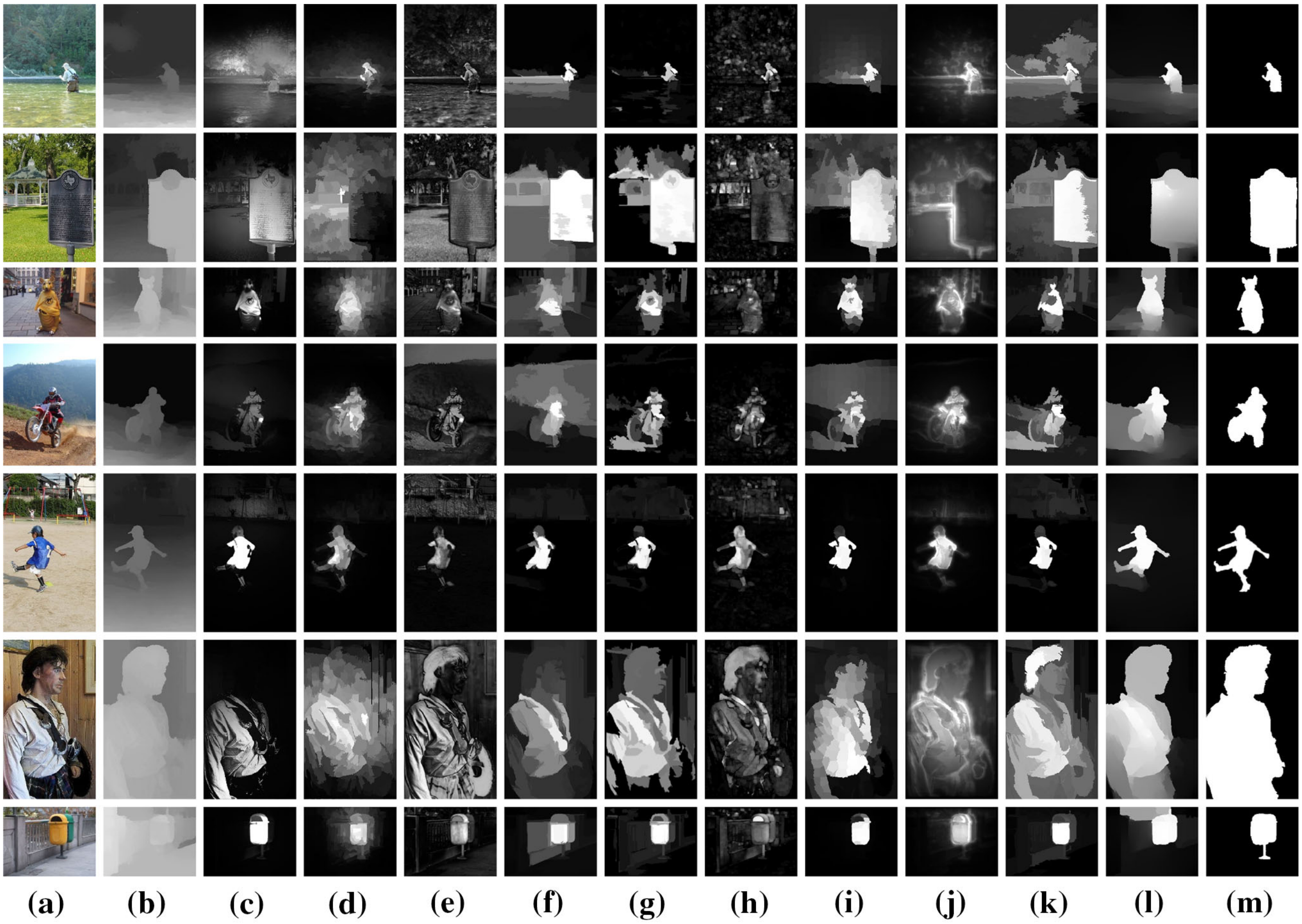


Fig. 6 Visual comparison of saliency maps produced by various methods. **a** Original image. **b** Depth map. **c** SF [42]. **d** LR [46]. **e** FT [1]. **f** CB [28]. **g** RC2 [9]. **h** BMS [54]. **i** GBMR [52]. **j** PCA [36]. **k** HS [51]. **l** Ours. **m** Ground truth

forms methods without introducing depth information. When the precision rate equals 0.9, our method can achieve recall rate 0.72, while other methods can achieve only 0.5 at most. On the other hand, if the recall rate is fixed at 0.9, the precision rate of our method reaches 0.76, while the maximum precision rate of other methods is only 0.5. The F-measure also shows the advantages of our method. Especially, when the threshold is greater than 100, our method significantly exceeds others. In Fig. 6, visual comparisons are given, which vividly demonstrate the advantages of our approach.

Necessity of both stages. To demonstrate that both stages of the proposed pipeline are necessary, we illustrate the performance of initial, intermediate and final results on our data set, which is shown in Fig. 7. Initial results are produced by the initial color saliency method—RC2 [9]; intermediate results, indicated by $S_p(i, j)$ in Eq. (7), are generated from the first stage—Object location; final results are produced by Stage 1 + Stage 2. The precision–recall curves in Fig. 7 demonstrate that both stages make some improvement.

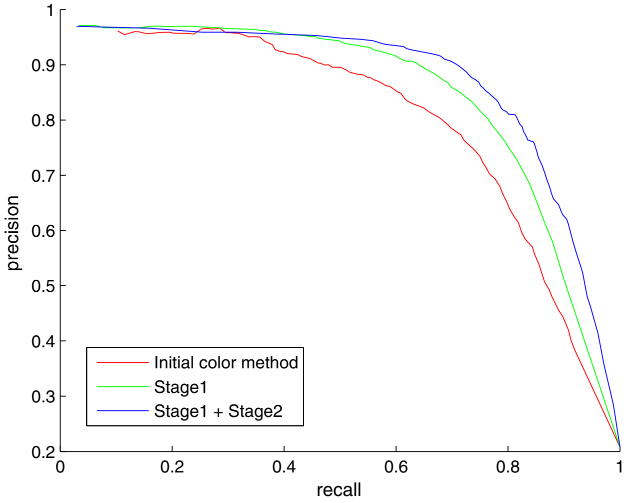


Fig. 7 The performance of initial (RC2), intermediate (Stage 1) and final (Stage 1 + Stage 2) results. It shows the necessity of both stages, which both make some improvement

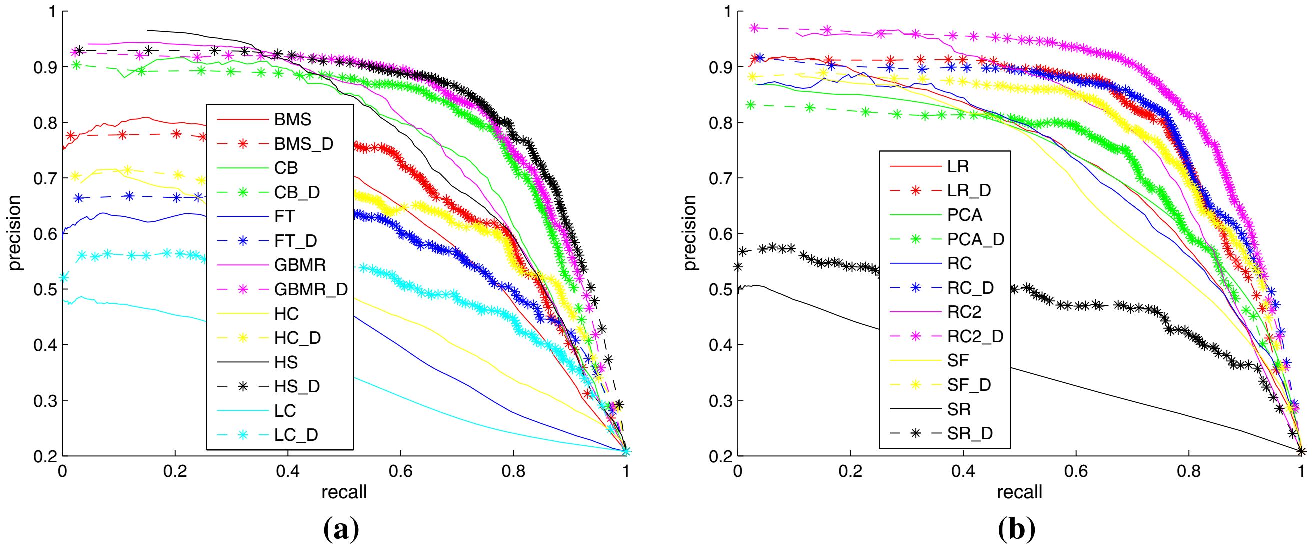


Fig. 8 Comparison of each *color* saliency method (BMS, CB, . . ., SR) and corresponding color–depth saliency method (BMS_D, CB_D, . . .,SR_D). RC2_D is the proposed pipeline and it achieves the best performance

Scalability. The proposed framework is scalable, and can be reconstructed by integrating other color saliency models. To demonstrate the effectiveness of the proposed method, we replace RC2 [9] by other color saliency methods, and perform additional experiments. The results are shown in Fig. 8. The precision–recall curve shows that each color–depth pipeline (BMS_D, CB_D, . . ., SR_D) can significantly improve the corresponding color-only method (BMS, CB, . . ., SR). And the proposed pipeline (RC2_D) achieves the best performance among all the color–depth pipelines. The reason for the excellent performance of color–depth pipeline lies in that each color method provides some location cue of the salient object (Stage 1), which can be used to infer more reliable object boundaries of the salient object (Stage 2). We propose RC2_D in this work because it provides the best location cue (most parts of the extracted salient patch are located within the salient object) and achieves the best performance.

4 Limitations

Currently, salient object detection is mainly applied in Internet (monocular or binocular) images. Our method can deal with stereo (binocular) images, which widely exist in the Internet, such as Flickr. Recently, a few works [12,13,41] explore how to improve saliency detection using captured depth and an RGBD data set [41] captured by Kinect is released for salient object detection in the work [41]. Although, our pipeline is not designed for Kinect 3D point cloud data, we applied our method on this data set, by simply treating the 3D data as a depth image. Results of our method and other methods (LS [12], SVR [13], RGBD [41]) which are designed for Kinect 3D data are illustrated in Fig. 9.

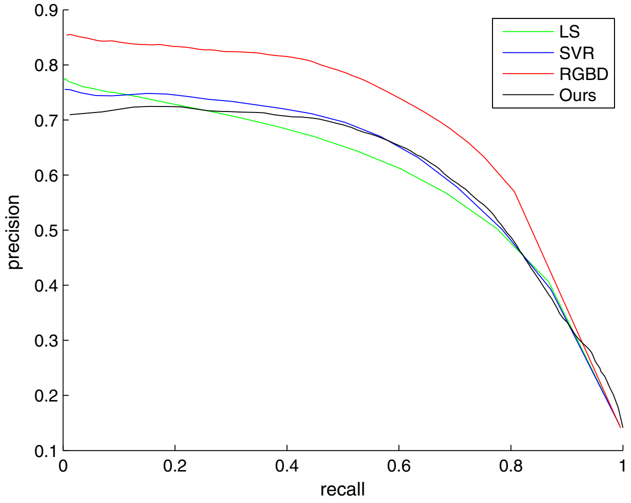


Fig. 9 Comparison results on the RGBD data set [41]

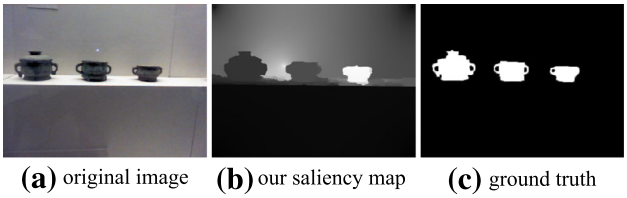


Fig. 10 Failure example of multiple salient objects detection

In Fig. 9, it shows that LS [12], SVR [13] and our method have comparable performance, while RGBD [41] outperforms the others. Reasons for our unsatisfying performance include: First, our method is designed for detecting single unambiguous salient objet, while a series of images in RGBD data set [41] contain multiple salient objects. In this case, our method fails to produce good saliency map, as is shown in Fig. 10. Second, different from the other three methods, our method is not designed for Kinect 3D point cloud data. We only use relative depth to measure saliency while other features, such as surface normal, are not introduced. Also, the proposed algorithm needs further modification to be applied on Kinect depth data.

5 Conclusions and future work

We present a two-stage framework for detecting salient objects in challenging images and each stage combines color and depth features, which is designed based on the characteristics of salient objects. In the first stage, to locate the salient object, we combine color and depth features for robustness, as color contrast-based methods are usually vulnerable in challenging scenarios, which may introduce background noises in the salient patch. In the second stage, the salient object is entirely highlighted through inferring plausible object boundaries. The inference process is formulated as a random walk problem on a graph, which is defined using both depth and color information. We also construct a stereo image data set for evaluating salient object detection in challenging images. This data set contains 100+ image pairs and each pair has only one salient object. We test our method on this data set and the evaluation results show that the proposed color-plus-depth based method significantly improves salient object detection compared to previous color-based methods, which indicates the usefulness of depth in salient object detection.

In the future, we will further explore how to improve saliency detection by combining depth and other features. Also, this work utilizes inaccurate depth information which is acquired by time-consuming computation from stereo images. We will further attempt to use the accurate depth information captured by sensors (such as Microsoft Kinect) to measure saliency, like the work [41]. As current pipeline is not suitable for multiple salient objects detection (Fig. 10), we will design new RGBD pipeline for this task in the future.

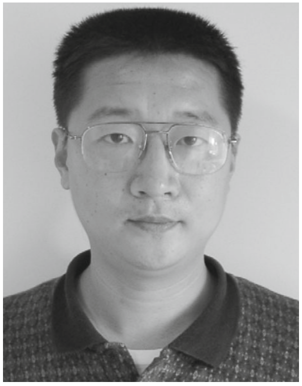
Acknowledgments We thank Prof. Jianjun Zhang and Jian Chang from Bournemouth University for their valuable discussion. We also thank all anonymous reviewers for their valuable comments. This work was supported by National High-Tech Research and Development Program of China (No. 2013AA013903), National Basic Research Program of China (No. 2011CB302205), the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA Grant agreement No. [612627]-'AniNex', Zhejiang Provincial Natural Science Foundation of China (No. LY14F020050).

References

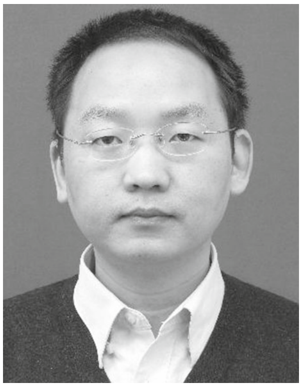
1. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1597–1604 (2009)
2. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2189–2202 (2012)
3. Backer, G., Mertsching, B., Bollmann, M.: Data-and model-driven gaze control for an active-vision system. IEEE Trans. Pattern Anal. Mach. Intell. **23**(12), 1415–1429 (2001)
4. Borji, A., Ahmadabadi, M.N., Araabi, B.N., Hamidi, M.: Online learning of task-driven object-based visual attention control. Image Vis. Comput. **28**(7), 1130–1145 (2010)
5. Bruce, N., Tsotsos, J.: Saliency based on information maximization. Adv. Neural. Inf. Process. Syst. **18**, 155 (2006)
6. Chang, K.Y., Liu, T.L., Chen, H.T., Lai, S.H.: Fusing generic objectness and visual saliency for salient object detection. In: IEEE International Conference on Computer Vision (ICCV), pp. 914–921 (2011)
7. Chen, T., Cheng, M.M., Tan, P., Shamir, A., Hu, S.M.: Sketch2photo: internet image montage. ACM Trans. Graph. **28**(5), 124:1–124:10 (2009)
8. Cheng, M.M., Mitra, N.J., Huang, X., Hu, S.M.: Salientshape: group saliency in image collections. Vis. Comput. **30**(4), 1–11 (2013)
9. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H.S., Hu, S.M.: Global contrast based salient region detection. IEEE TPAMI (2014). doi:10.1109/TPAMI.2014.2345401
10. Cheng, M.M., Warrell, J., Lin, W.Y., Zheng, S., Vineet, V., Crook, N.: Efficient salient region detection with soft image abstraction. In: IEEE International Conference on Computer Vision (ICCV), pp. 1529–1536 (2013)
11. Cheng, M.M., Zhang, G.X., Mitra, N.J., Huang, X., Hu, S.M.: Global contrast based salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 409–416 (2011)
12. Ciptadi, A., Hermans, T., Rehg, J.M.: An in depth view of saliency. In: BMVC, pp. 1–11 (2013)
13. Desingh, K., Krishna, K.M., Rajan, D., Jawahar, C.: Depth really matters: improving visual salient region detection with depth. In: BMVC, pp. 1–11 (2013)
14. Donoser, M., Urschler, M., Hirzer, M., Bischof, H.: Saliency driven total variation segmentation. In: IEEE International Conference on Computer Vision (ICCV), pp. 817–824 (2009)
15. Du, S.P., Hu, S.M., Martin, R.R.: Changing perspective in stereoscopic images. IEEE Trans. Vis. Comput. Graph. **19**(8), 1288–1297 (2013)
16. Du, S.P., Masia, B., Hu, S.M., Gutierrez, D.: A metric of visual comfort for stereoscopic motion. ACM Trans. Graph. **32**(6), 222 (2013)
17. Fang, Y., Wang, J., Narwaria, M., Le Callet, P., Lin, W.: Saliency detection for stereoscopic images. In: Visual Communications and Image Processing (VCIP), pp. 1–6. IEEE (2013)
18. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. Int. J. Comput. Vis. **59**(2), 167–181 (2004)
19. Garcia-Diaz, A., Fdez-Vidal, X.R., Pardo, X.M., Dosil, R.: Decorrelation and distinctiveness provide with human-like saliency. In: Advanced Concepts for Intelligent Vision Systems, pp. 343–354. Springer, Berlin (2009)
20. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. IEEE Trans. Pattern Anal. Mach. Intell. **34**(10), 1915–1926 (2012)
21. Grady, L.: Random walks for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **28**(11), 1768–1783 (2006)

22. Han, J., Ngan, K.N., Li, M., Zhang, H.J.: Unsupervised extraction of visual attention objects in color images. IEEE Trans. Circuits Syst. Video Technol. **16**(1), 141–145 (2006)
23. Harel, J., Koch, C., Perona, P., et al.: Graph-based visual saliency. Adv. Neural Inf. Process. Syst. **19**, 545 (2007)
24. Hou, X., Zhang, L.: Saliency detection: a spectral residual approach. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2007)
25. Hou, X., Zhang, L.: Dynamic visual attention: searching for coding length increments. In: NIPS, vol. 5, p. 7 (2008)
26. Hu, S.M., Chen, T., Xu, K., Cheng, M.M., Martin, R.: Internet visual media processing: a survey with graphics and vision applications. Vis. Comput. **29**(5), 393–405 (2013)
27. Itti, L., Koch, C., Niebur, E., et al.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
28. Jiang, H., Wang, J., Yuan, Z., Liu, T., Zheng, N., Li, S.: Automatic salient object segmentation based on context and shape prior. In: BMVC, vol. 3, p. 7 (2011)
29. Ko, B.C., Nam, J.Y.: Object-of-interest image segmentation based on human attention and semantic region clustering. JOSA A **23**(10), 2462–2470 (2006)
30. Lang, C., Nguyen, T.V., Katti, H., Yadati, K., Kankanhalli, M., Yan, S.: Depth matters: influence of depth cues on visual saliency. In: ECCV, pp. 101–115. Springer, Berlin (2012)
31. Li, H., Ngan, K.N.: A co-saliency model of image pairs. IEEE Trans. Image Process. **20**(12), 3365–3375 (2011)
32. Li, N., Ye, J., Ji, Y., Ling, H., Yu, J.: Saliency detection on light field. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
33. Liu, H., Zhang, L., Huang, H.: Web-image driven best views of 3d shapes. Vis. Comput. **28**(3), 279–287 (2012)
34. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.Y.: Learning to detect a salient object. IEEE Trans. Pattern Anal. Mach. Intell. **33**(2), 353–367 (2011)
35. Liu, Z., Xue, Y., Shen, L., Zhang, Z.: Nonparametric saliency detection using kernel density estimation. In: IEEE International Conference on Image Processing (ICIP), pp. 253–256 (2010)
36. Margolin, R., Tal, A., Zelnik-Manor, L.: What makes a patch distinct? In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1139–1146 (2013)
37. Margolin, R., Zelnik-Manor, L., Tal, A.: Saliency for image manipulation. Vis. Comput. **29**(5), 381–392 (2013)
38. Mu, T.J., Sun, J.J., Martin, R.R., Hu, S.M.: A response time model for abrupt changes in binocular disparity. Vis. Comput. 1–13 (2014)
39. Mu, T.J., Wang, J.H., Du, S.P., Hu, S.M.: Stereoscopic image completion and depth recovery. Vis. Comput. **30**(6–8), 833–843 (2014)
40. Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging stereopsis for saliency analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 454–461 (2012)
41. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: Rgbd salient object detection: a benchmark and algorithms. In: ECCV, pp. 92–109 (2014)
42. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: contrast based filtering for salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 733–740 (2012)
43. Rosenholtz, R., Dorai, A., Freeman, R.: Do predictions of visual perception aid design? ACM Trans. Appl. Percept. (TAP) **8**(2), 12 (2011)
44. Rutishauser, U., Walther, D., Koch, C., Perona, P.: Is bottom-up attention useful for object recognition? In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. II-37. IEEE (2004)
45. Seo, H.J., Milanfar, P.: Static and space–time visual saliency detection by self-resemblance. J. Vis. **9**(12), 15 (2009)
46. Shen, X., Wu, Y.: A unified approach to salient object detection via low rank matrix recovery. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 853–860 (2012)
47. Shi, Y., Yi, Y., Yan, H., Dai, J., Zhang, M., Kong, J.: Region contrast and supervised locality-preserving projection-based saliency detection. Vis. Comput. 1–15 (2014)
48. Siagian, C., Itti, L.: Biologically inspired mobile robot vision localization. IEEE Trans. Robot. **25**(4), 861–873 (2009)
49. Smith, B.M., Zhang, L., Jin, H.: Stereo matching with nonparametric smoothness priors in feature space. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 485–492 (2009)
50. Tong, R., Zhang, Y., Cheng, K.L.: Stereopasting: interactive composition in stereoscopic images. IEEE Trans. Vis. Comput. Graph. **19**(8), 1375–1385 (2013)
51. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1155–1162 (2013)
52. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3166–3173 (2013)
53. Zhai, Y., Shah, M.: Visual attention detection in video sequences using spatiotemporal cues. In: Proceedings of the 14th Annual ACM International Conference on Multimedia, pp. 815–824. ACM (2006)
54. Zhang, J., Sclaroff, S.: Saliency detection: a boolean map approach. In: IEEE International Conference on Computer Vision (ICCV), pp. 153–160 (2013)

Yanlong Tang is a Ph.D. candidate of Zhejiang University. He received his B.S. from Shandong University, China in 2013. His research interests include image and video processing, computer graphics.



Ruofeng Tong is a professor in Department of Computer Science, Zhejiang University, China. He received his B.S. from Fudan University, China in 1991, and a Ph.D. from Zhejiang University, China in 1996. His research interests include image and video processing, computer graphics, and computer animation.



Min Tang is a professor in the college of computer science at Zhejiang University, China, since 2000. He received his B.S., M.S., and Ph.D. from Zhejiang University in 1994, 1996, and 1999, respectively. From June 2003 to May 2004, he was a visiting scholar at Wichita State University. From April 2007 to April 2008 he was a visiting scholar at the University of North Carolina at Chapel Hill. His research interests include geometry modeling, collision detection, and GPU-based algorithm acceleration.

Yun Zhang is an assistant professor of Zhejiang University of Media and Communications. He received his B.S. and M.S. from Hangzhou Dianzi University, China in 2006 and 2009, and a Ph.D. from Zhejiang University, China in 2013. His research interests include image / video editing and computer vision.