Multimed Tools Appl (2018) 77:14655–14672
DOI 10.1007/s11042-017-5052-8

Accurate saliency detection based on depth feature of 3D images

Haoqian Wang [1,2] **· Bing Yan** [1] **· Xingzheng Wang** [1,2] **·**
Yongbing Zhang [1,2] **· Yi Yang** [3]

Received: 13 October 2016 / Revised: 20 May 2017 / Accepted: 24 July 2017 /
Published online: 16 August 2017

Abstract This paper presents an accurate saliency detection algorithm customized for 3D images which contain abundant depth cue. Firstly, depth feature is calculated based on the sharp regions' positions within the focal stack. Then, we compute the coarse saliency map by subtracting the background region from the all-focus image according to the depth feature. Finally, we employ the contrast information in the coarse saliency map to obtain the final result. Experiments on light field dataset demonstrate that our approach favorably outperforms five state-of-the-art methods in terms of precision, recall and F-Measure. Moreover, the depth feature is validated to be a valuable complement to existing visual saliency analysis under the circumstance that the background regions are complex or similar to salient object regions.

Keywords Saliency detection · Depth feature · Background region · Light field

1 Introduction

Saliency detection [2], which aims to detect the salient attention-grabbing objects in a scene and segment the whole objects thereby represent and describe the object regions in characteristic features, has been a hot topic for nearly 20 years since it is an indispensable stage in computer vision tasks including object-of-interest image segmentation [12], image retrieval [23], object recognition [24] and retargeting [14]. However, almost all of methods are for 2D images. With the rapid development of visual reality technique, an

✉ Xingzheng Wang
xingzheng.wang@sz.tsinghua.edu.cn

[1] Key Lab of Broadband Network and Multimedia, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China

[2] Shenzhen Institute of Future Media Technology, Shenzhen 518071, China

[3] LUSTER LightTech Co., Ltd., Beijing 100094, China

increasing number of 3D images is arising. New techniques aiming at processing stereoscopic images are urgently needed to arise. Therefore, this paper presents an accurate saliency detection method on 3D images. The concept of saliency detection was first proposed by Itti et al. [13] in 1998, whose followers continued his concept to conduct saliency detection via local method. In addition, Itti's center-surround operation is typically center-prior based, which is simply less efficient in terms of computing expense. Later, Hou's spectral residual based approach [25] is characterized by low computational complexity, while it has difficulty in dealing with sophisticated images. Recently Zhang et al. [10] proposed the usefulness of surroundness for eye fixation prediction by proposing a Boolean Map based Saliency model. These are typical saliency detection approaches whose performance are excellent for 2D images.

The objective of this paper is to achieve high accuracy of saliency detection task on stereoscopic images by the aid of exploiting the depth information. The images on which we conduct saliency detection are light field images, which consist of depth cues and exactly meet the demand of this paper. Compared with traditional 2D images, light field images are characterized by the property of refocusing. After a range of post-processing, a series of refocused images (shown in Fig. 1 (a), the regions surrounded by black windows are clear and focused areas) are available, which is the focal stack. The foreground and background regions are separated based on the locations of sharp regions (illustrated in Fig. 1 (b)). Besides, there is an all-focus image, all of whose pixels are focused, i.e., both the foreground and background of the image is clear. The focal stack and the all-focus image are our test images. Lots of existing methods focused on finding features that can express the salient object better. Just like the image emotion prediction in [30] shows, the emotions that are evoked in viewers by an image are highly subjective and different. We try to make use of the background information offered by the focal stack instead of analyzing the object information directly. Here are the procedures of our approach. Firstly, the depth information of the input images is calculated via analyzing the focal stack; and then background and foreground of the image are separated coarsely through background prior and the coarse saliency map is obtained. Thereafter, the final result is computed by calculating feature contrast globally based on coarse saliency map. The pipeline of our method is illustrated in Fig. 1(c). The vital step is extracting depth information.

In summary, this paper has the following contributions

1. It proposes a novel algorithm customized for light field images containing abundant depth cue. By extracting the depth features hided in the focal stack of the light field image, our method manages to accurately distinguish the background layers from the foreground. The step insures the accuracy of the result. Our model creates a new perspective for saliency detection, and achieves the state-of-the-art performance;
2. The texture feature contrast is employed in our method, which is seldomly exploited in the previous algorithms. We avoid the center prior since salient objects usually appear at the image border, and it is not a universal prior.

The remainder of this paper is organized as follows. Section 2 introduces a brief review of the previous work related to existing saliency detection models and that based on 3D images. In Section 3, we elaborate our algorithm in detail. We provide experimental results for real images in Section 4. Lastly, concluded remarks are made in Section 5.

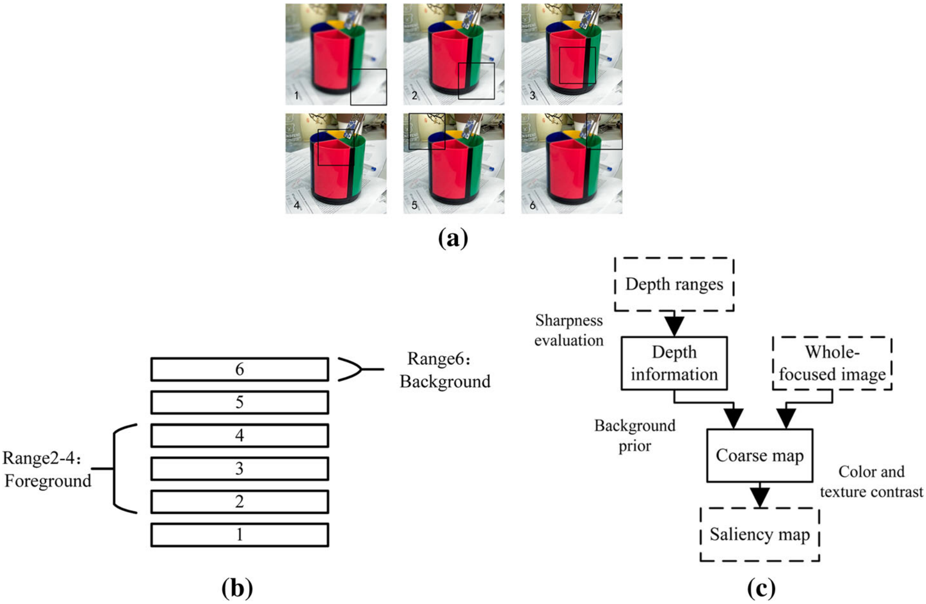


Fig. 1 **a** Focal stack. **b** The depth feature of focal stack. **c** The framework of our method

2 Related work

2.1 The existing saliency detection models

Since saliency detection was presented by Itti et al. [13] in 1998, the concept has attracted lots of attention. It is an indispensable stage for lots of image processing tasks. For example, the applications of saliency maps consist of object aware image retargeting [11], image editing techniques [8, 15], and object detection [6] and etc. In the early time, researchers were influenced deeply by Itti, so they were used to compute the center-surround contrast. Nevertheless, the conspicuous objects are quite different in different scenes, which makes it difficult to fix a universal model. Hou [25] provided a new perspective for saliency detection. He put forward the new way creatively to locate the background regions first; thus the salient region is obtained after subtracting the background from the image. Inspired by the background prior, many methods are proposed, such as [27, 31]. These methods using background prior knowledge inevitably assume that the boundary regions belong to background regions. However, some fail to find the whole salient object when there are foreground noises in the boundary regions. Thus, the background prior is not universal though it works well in many cases. Some complements are needed to achieve high accuracy.

Visual saliency can be viewed from other perspectives. Contrast-based methods are the most popular ones. Local contrast can be used to detect low-level saliency [13], and global contrast such as color contrast [16] could suppress the background better. Recently, the combination [20] of local and global contrast has been represented and the result shows high accuracy. However, these are approaches for traditional 2D images. In the next subsection, we will discuss the methods for 3D images.

2.2 3D image based saliency detection methods

As noted in [1], human beings live in a three-dimensional real world, where stereoscopic information makes it easier to locate the objects we are interested in. In the last few years, researchers have exploited depth cues from the input images with their disparity images [28]. It combined the global contrast and the salient measure computed from domain knowledge in stereoscopic photography. The result of this method is not as good as the state-of-the-art methods, for the fact that some prior knowledge such as background prior is not utilized. Besides, the performance of the method depends highly on the quality of the disparity map. Zhao et al. [29] proposed a feature fusion method based on multi-modal graph learning for view-based 3D object retrieval, which used several visual features to conduct feature fusion and the final result is satisfactory. In the paper, we also utilize feature fusion after locating the foreground regions in order to obtain accurate results.

With the rapid development of light field cameras, such as Lytro and Raytrix, the merits of light field images have attracted an increasing number of researchers. Meanwhile, light field imaging offers new possibilities for many computer vision tasks which are confronted with the bottleneck. The light field data benefit saliency detection in various ways [19]. Firstly, light field images could be refocused to any depth of the scene, which could provide both depth and focus cues. Secondly, the focus cues could be transformed into depth cues through a series of steps. This is the core of our method. Li [19] uses the light field images' focusness measure to detect salient regions. However, there are some shortcomings of this model: firstly, center prior is not always effective since salient objects do not always appear at the image center [27]. Moreover, global-based color contrast manipulation is far from enough. Some saliency maps of [19] could not distinguish the salient regions from background when their appearances are similar. In our method, we introduce the feature descriptor Local Binary Pattern histograms (LBP) [18] into the contrast operation. The performance of our method achieves much higher accuracy because LBP is a classical texture detector and it is an effective complement to color contrast. In this paper, we exploit the depth feature of the input 3D images to represent the salient regions accurately. As described in [5], humans fixate preferentially at closer depth ranges, i.e., objects popping out from the screen tend to be salient [28].

3 Accurate saliency detection based on the depth feature of 3D images

The proposed approach is an accurate saliency detection based on the depth feature of 3D images. This section will explain the method in detail. The input images of our algorithm are focal stack (depth number $d = 1, 2, \ldots, L$) and the all-focus image. The focal stack images are utilized to obtain depth information, and thereby extracting background regions. Accordingly, the background regions in the corresponding all-focus image are obtained. Next, saliency maps are computed based on feature contrast manipulation.

3.1 Extracting depth information from the focal stack

From Fig. 1(a) we can see, the location of the clear region is moving from front to behind as the number of depth slice increases. In the recent light field-based saliency detection work [19], the focusness is measured by analyzing the image statistics in the frequency domain. However, traversal operations with the sliding window is not efficient enough. In this paper we employ the gradient operator to measure clarity of the region. Let (j, k) denote the pixel in the image from the focal stack. $j = 1, 2, ..., w$; $k = 1, 2, \ldots, h$ (w, h are the width and height of input image respectively). *Gray* (j, k) is the grayscale value of the pixel (j, k) calculated by weighting the RGB value of (j, k):

$$Gray(j,k) = R(j,k) \times 0.299 + G(j,k) \times 0.587 + B(j,k) \times 0.114. \tag{1}$$

Next, gradient value of the focal stack image along the coordinate axis x is computed as:

$$G_x(j,k) = \begin{cases} Gray(j,k+1) - Gray(j,k) \ , \ k \ = \ 1 \\ Gray(j,k) - Gray(j,k{-}1) \ , \ k \ = \ end \\ \dfrac{Gray(j,k+1){-}Gray(j,k{-}1)}{2}, otherwise \end{cases}. \tag{2}$$

Then sharpness matrix could be defined as:

$$G(j,k) = \sqrt{{G_x}^2(j,k) + {G_y}^2(j,k)}. \tag{3}$$

$G_y(j, k)$ represents the gradient value along coordinate axis y. Next we compute the sharpness for each region ri using:

$$G(r_i) = \frac{1}{N_{r_i}} \sum_{(j,k)\in r_i} G(j,k), \tag{4}$$

where N_{r_i} is the number of pixels within the region ri. Next, we calculate the sharpness of the image horizontally and vertically:

$$\begin{aligned} D_j &= \frac{1}{\lambda} \sum_{k=1}^{h} G(j,k) \\ D_k &= \frac{1}{\lambda} \sum_{j=1}^{w} G(j,k) \, . \\ \lambda &= \sum_{j=1}^{w} \sum_{k=1}^{h} G(j,k) \end{aligned} \tag{5}$$

There is little change in each row at some of the background areas, such as the ground. Thus, the average operation hardly influences the sharpness of the background regions. Thereafter, we utilize Gaussian filtering to select the image whose clear region is background, such as the last one in Fig. 1(a).

$$\begin{aligned} Gauss(x) &= 1{-}\exp^{-\frac{(x-\mu_x)^2}{2\sigma^2}} \\ \mu_x &= \frac{\mu_c + \mu_p}{2} \end{aligned} , \tag{6}$$

where μ_c represents the center coordinate of j(or k) and μ_p is the peak location of D_j(orD_k). σ controls the band width of the filter. Therefore we could compute the background measurement:

$$BM = \left(Gauss(j){\cdot}D_j + Gauss(k){\cdot}D_k\right){\cdot}\exp\left(\frac{\eta{\cdot}d}{L}\right). \tag{7}$$

where η represents the influence of the depth. The value of background measurement is ranging between 0 and 1. The layer with the highest background measurement is chosen as the background layer. The sharpness matrix of background layer is denoted as$G_B(r)$.

3.2 Detecting the foreground regions coarsely

Now, we use the sharpness measure of the selected background layer to distinguish the foreground from background regions in the all-focus image. By means of the mean-shift algorithm [7], the all-focus image is segmented into N superpixels *ri*, *i* = 1, 2, ..., *N*. Here we utilize the object-biased Gaussian model to analyze the background layer. Therefore background cue is computed as:

$$
\begin{aligned}
BC(r) &= G_B(r)\cdot\left(1-\exp^{-\frac{(r-r_o)^2}{2\sigma_r{}^2}}\right) \\
r_o &= \sum_i \frac{G(r_i)}{\sum G(r_i)}
\end{aligned}
\qquad (8)
$$

where $G_B(r)$ represents the region sharpness of the background layer, *r* is the superpixel. *ro* denotes object center derived from the region sharpness $G(r_i)$. Thereafter, we threshold background cue to separate foreground from background regions in the all-focus image. Virtually the original saliency map has been obtained yet, though the pixel values are binary.

3.3 Conducting feature contrast manipulation between foreground and background

Feature contrast is exploited extensively in saliency detection models such as [16, 31]. However, nearly all of the methods extract color (RGB and CIELab color features) and location feature descriptor. While texture feature descriptor is seldomly utilized. In this paper, we employ both color (RGB pixels) and texture feature (the Local Binary Pattern histograms [18]) to further represent salient regions within an image. The feature contrast is based on the coarse saliency map, i.e., to compute the contrast between foreground and background regions. $F(r)$represents the R, G and B values of the selected foreground regions respectively. $F(r')$ represents the corresponding R, G and B values of background regions respectively. After calculating the color distance between $F(r)$ and $F(r')$, they are added together to get the final color contrast $c(r, r')$.

$$
c\left(r,r'\right) = \sum_{F\in\{R,G,B\}} \left|F(r)-F\left(r'\right)\right|_2^2. \qquad (9)
$$

where *r* denotes the salient region and *r'* denotes the background region. For each *r*, we calculate *c* (*r*, *r'*) with respect to all the background regions in R, G and B channels. Then we use harmonic variance to better express color contrast:

$$
H(r) = K\left(\sum_{r'=1}^{K} \frac{1}{c(r,r')}\right)^{-1}. \qquad (10)
$$

Where K denotes the number of background regions.

Next we use the texture feature descriptor Local Binary Pattern histograms (LBP) to express the texture disparity. According to [18], we construct an LBP histogram for each superpixel, i.e., a vector of 59 dimensions ($\{h_i\}$, $i = 1, 2, ...59$, where h_i is the value of the i-th bin in an LBP histogram. For each foreground region r, the texture contrast is computed with respect to all the K background regions in the following:

$$t(r) = \frac{1}{N_D}\sqrt{\sum_{r'=1}^{K} |h_i(r) - h_i(r')|_2^2}. \tag{11}$$

Where $\{h_i\}$ is a 59-dimensional matrix which records the texture feature of the image. N_D is 59. Based on the above analysis, we linearly combine color contrast with texture contrast:

$$Con(r) = (1-\rho)\cdot H(r) + \rho\cdot t(r). \tag{12}$$

Where ρsuppresses background pixels to be detected. If ρ is large, the background pixel values cannot be suppressed successfully.

3.4 Obtaining the final saliency map

The final saliency map is an optimized map of (12). The weight is computed by the background measurement.

$$S(r) = \left[(1-Gauss(j))\cdot D_j + (1-Gauss(k))\cdot D_k\right]\cdot con(r)\cdot\alpha, \tag{13}$$

where α is a constant between 0 and 1.

4 Experimental comparisons

We conducted the saliency detection experiment on the light field dataset provided by Li et al. [19]. The dataset contains 100 light field images. To compare with previous approaches, we use all-focus images as input to run their open source code. We compared with 5 state-of-the-art methods including algorithms based on spectral residual (SR [25]), Frequency-Tuned (FT [21]), Context-Aware SaliencyDetection (CA [22]), Low Rank Matrix Recovery (LR [26]) and Saliency Detection with Multi-Scale Superpixels (MS [17]).

4.1 Parameter setting

Based on the saliency detection algorithm proposed in Section 3, we have conducted experiments utilizing light field (LF) image whose size is 360 × 360. According to the size, some parameters in Section 3 could be determined here. In our algorithm, since we make use of background prior to separate the foreground and background regions in the all-focus image, the important step is to select the background layer in the focal stack. From Fig. 1 (a) we could see, if the layer is deeper in the focal stack, its boundary regions are sharper. Thus the background layer could be selected by high-pass filter, which is illustrated in eq. 6. The parameter σ controls the band width of high-pass filter. In order to filter the foreground regions which are not sharp in the background layer and maintain the remaining regions as much as possible simultaneously, σ should be somehow smaller than the length of the image. Consequently, we set σ as 40, which is also smaller than the size of estimated salient object. By means of the high-pass filter, we could obtain the background measurement which represents the possibility that a layer is background layer. In eq. 7, $d$denotes the number of layer in the focal stack and L denotes the total number of layers. Since almost all the focal stacks have around 10 layers, we select η as 10 empirically. (see Table 1).

After determining the background layer, then we need to leverage it to distinguish the salient regions from background ones in the all-focus image. As a matter of fact, the region division of the background layer is consistent with that of all-focus image. According to eq. 4, the region sharpness of background layer could be represented as $G_B(r)$, whose sharpness values are larger in the boundary regions. Here we adopt the location prior to help in selecting the background regions, since the salient object tend to be located at the center of the image without touching the image boundary. As shown in eq. 8, we also choose High-pass filter to obtain the background cue. Here, σ_r controls the band width of high-pass filter and we set it as $0.25\times w$ in order to make the boundary regions to be detected as background. In fact, when salient object is not located at image center, the choice could also ensures the accuracy of the detection.

After thresholding background cue, the original foreground areas are located simply. Then feature contrast manipulation including RGB color and Local Binary Pattern histograms between foreground and background will be conducted. Hence feature confusion should consider the proportion between these two features. In our implementation, when color contrast is dominant in the process of feature fusion, the final result is better. Consequently, we set the parameter ρ as 0.05 thereby emphasize the effect the feature of color contrast.

4.2 Evaluation metrics

In this paper, we utilize both qualitative and quantitative comparisons to demonstrate the accuracy and robustness of our algorithm. Firstly, qualitative comparisons contain the visual contrast of final results, the efficiency of depth feature and texture feature, Comparisons with previous method based on light field, and the performance of edge detection. These comparison results are conducted by the vision of human thereby are intuitive. Thereafter, if the difference is far from obvious, the qualitative comparison result is not accurate. Accordingly, quantitative comparisons are necessary since the comparison results have specific values. There are three popular quantitative metrics for saliency detection: Precision, Recall and F-Measure. Besides, Mean absolute error (MAE) and Running time are also adopted to illustrate the accuracy of our approach. Next we will introduce these metrics in detail.

$$\begin{cases} precision = \dfrac{sum(S, GT)}{sum(S)} \\ recall = \dfrac{sum(S, GT)}{sum(GT)} \end{cases}, \tag{14}$$

Table 1 Parameter setting in experiments

Equation	(6)	(7)	(8)	(12)
Parameter	σ	η	σ_r	ρ
Value	40	10	$0.25\times w$	0.05

where *sum* means the sum of pixels whose values are 1. (S, GT) represents the and-operation of saliency map and the corresponding ground truth.

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

where β controls the weight of precision value. In our implementation, β^2=0.4.

As usual we take the P-R curve and F-Measure curve as the measure to evaluate our method with other five methods. These two curves (in Fig. 6) are illustrated by changed-threshold, whose threshold is between 0 and 255. Moreover, another fixed-threshold segment experiment is conducted in order to obtain the P-R-F-measure histogram (in Fig. 7(a)).

MAE is adopted as another evaluation criterion [9]. It is defined as the average pixelwise absolute difference between the binary ground truth GT and the saliency map S:

$$MAE = \frac{1}{w\times h}\sum_{j=1}^{w}\sum_{k=1}^{h}|S(j,k)-GT(j,k)|. \tag{16}$$

4.3 Qualitative results

Visual comparison of different saliency detection algorithms vs. our algorithm We demonstrate some results generated by six methods for qualitative comparison in Fig. 3. From left to right are input image, SR (Spectral Residual), FT(Frequency-Tuned), CA(Context-Aware Saliency Detection), LR(Low Rank Matrix Recovery), MS (Saliency Detection with Multi-Scale Superpixels), our method and ground truth. Our method could highlight the salient object integrally, which is a considerably challenging problem for previous methods such as CA and MS. Figure 2 demonstrates that performance of our algorithm is much better when the input image is characteristic of the foreground/background similarity. For instance, in the 3rd row, only our approach can distinguish the blue paper flower from background. And the 2nd

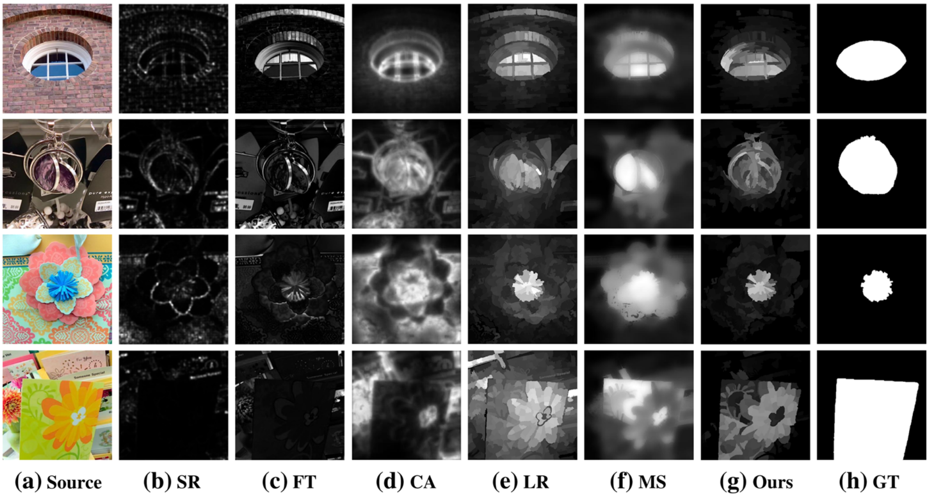


Fig. 2 Visual comparison of different saliency detection algorithms vs. our algorithm. **a** Input image. **b** SR [25]. **c** FT [21]. **d** CA [22]. **e** LR [26]. **f** MS [17]. **g** Ours. **h** Ground truth

row could also prove that our algorithm does better than the other methods under the circumstances of complex background textures. The previous methods such as SR and FT are not able to highlight the salient regions. The results of SR are blurry and some local contrasts are detected (shown in Fig. 2 (b)). FT could not suppress the features which appear frequently (illustrated in Fig. 2 (c)). Like other early models, CA exactly highlights the edge of salient objects rather than the whole salient objects. Although MS works well with lots of scenes, the results in Fig. 2 (f) are blurry and indistinguishable.

The efficiency of depth feature and texture feature The key step of our method is extracting depth feature, which ensures the detection of salient objects in complicated scenes. Furthermore, texture feature is also an efficient assistance when tackling these problems. Some comparisons of results with and without texture feature descriptor are illustrated in Fig. 3. It shows that the LBP descriptor helps to improve the performance considerably. As shown in Fig. 3, the stuff above the door is wrongly highlighted without LBP descriptor. The 4th column illustrates the results calculated via absorbing Markov chain [3]. For simple input images (the 2nd row), the state-of-the-art method MC works better than ours. However, the reason is that there are only 3 depth layers in the second input image, i.e., the depth of this input image is shallow. The depth information is less useful under the circumstance. As for the 3rd row in Fig. 3, there are 11 layers in the focal stack. So the traditional approach works worse than our

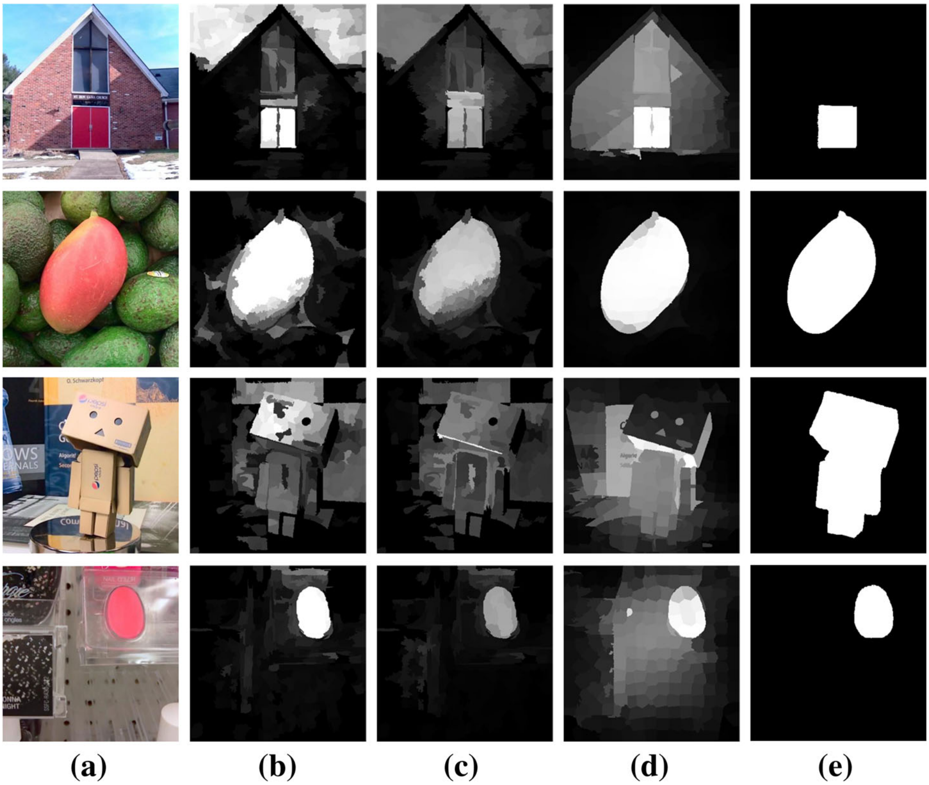


Fig. 3 Saliency maps with/without texture feature extraction. **a** Source image. **b** Output with texture feature. **c** Output without texture feature. **d** MC [3]. **e** Ground truth

method. Therefore, in most cases, it is essential to utilize depth feature in order to detect the salient objects correctly.

Comparisons with previous method based on depth information In Fig. 4 we compare our method with LFS [19]. Results intuitively demonstrate that LFS could not suppress the influence of disordered background, thus there are a number of regions with mid-level saliency, which is unfavorable. However, our algorithm overcomes the shortcomings and the saliency maps express the real saliency values well. For example, in the first row, the blue flower could be detected successfully by our approach. While LFS could not handle the cases with similar foreground and background. The reason could be concluded as follows. In the process of leveraging depth information, we only calculate the background measure in order to locate the background layer in the focal stack; while LFS computes the foreground cue simultaneously thereby select the foreground layers. Nevertheless, the number of foreground layers in the focal stack is not sure since it is corresponding to the size of salient area. In fact, the formula $FLS > 0.7 \times \max(FLS)$is doubtful since the author tells nothing about the parameter 0.7. Thus, the foreground cue is inaccurate, which influenced the final result a lot. Besides, the running time of LFS on 4 images is 23.59 s, while the time of ours on 4 same images is 6.20 s. The two algorithms were both tested on an Intel i5 3.10GHz CPU with 8GB RAM.

A discussion about edge detection Later we introduce the edge detection into our method. The effects are favorable (see Fig. 5). The algorithm with edge detection could do better than without it on the first two input images. However, for the last two input images, edge detection could not suppress the background areas and our algorithm outperforms the edge detection obviously. The reason lies in the own properties of edge detection. Since edge detection locates and represents the pixels whose brightness changes violently. Of course, the boundary of salient object will be detected. Whereas, the isolated unique regions belonging to background could also be found by edge detection. Hence, it is inferred that edge detection could not perform well on images with similar foreground and background. However, Fig. 5 (a) shows

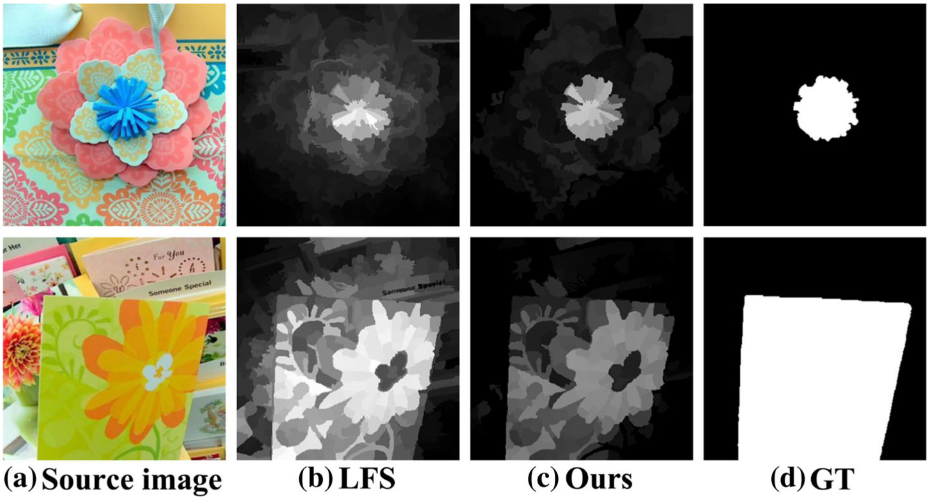


Fig. 4 Comparison of LFS and ours. **a** Source image. **b** The method based on light field [19]. **c** Our method. **d** Ground truth

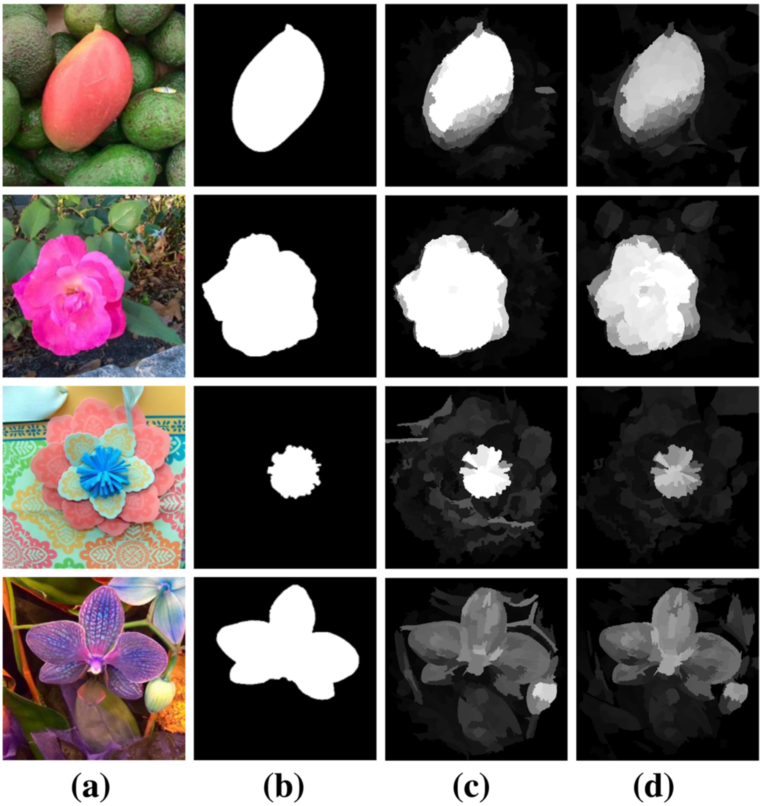


Fig. 5 Comparison of ours and edge detection. **a** Input image. **b** Ground truth. **c** Edge detection. **d** Ours

that if foreground and background have different colors edge detection would successfully detects the right edges. Therefore, in order to obtain high robustness, we think the original algorithm without edge detection better.

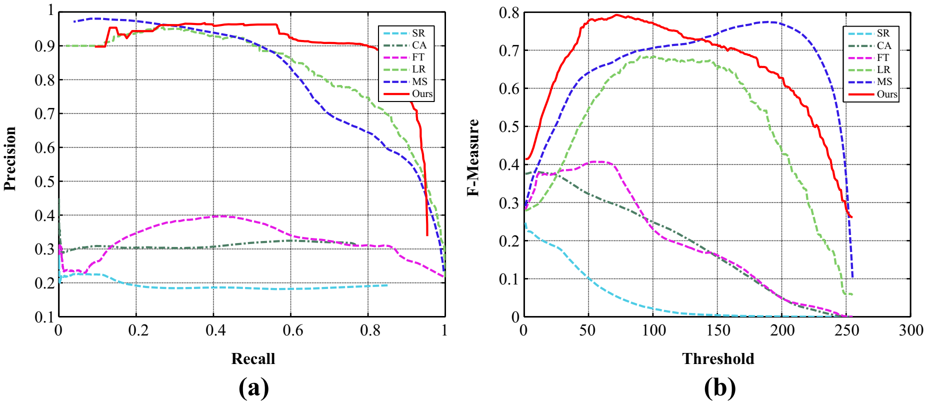
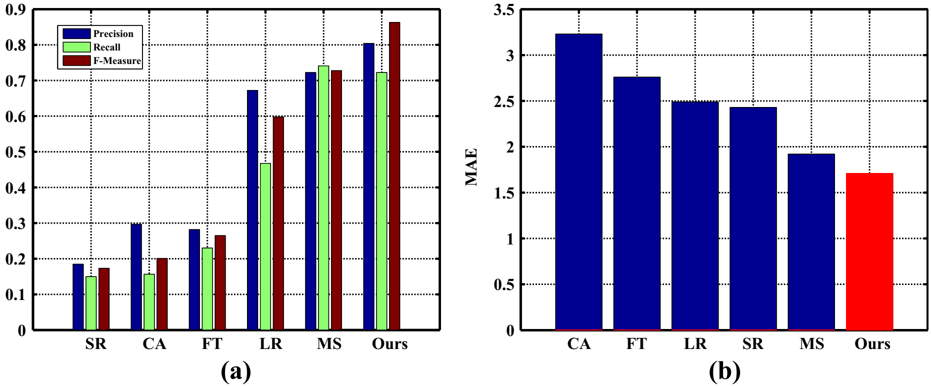


Fig. 6 Average precision-recall curve and F-Measure curve in comparison with 5 state-of-the-art methods. The black curve is our method. **a** Precision-recall curve. **b** F-Measure curve. The two figures show that our method considerably outperforms other methods in precision, recall and F-Measure values



(a)

(b)

Fig. 7 Quantitative comparison of saliency maps generated from 6 different methods on light field dataset. **a** the comparison of P-R and F-Measure values. **b** the comparison of MAE values

4.4 Quantitative results

Precision-recall curve and F-Measure curve Since the saliency map we have calculated is a grayscale image, not a binary one. Before the comparison with ground truth, the saliency map should be transformed to binary segmentation of salient objects via thresholding the saliency map with a threshold. Thus we obtain 255 binary masks, furthermore 255 pairs of average P-R values of all the images included in the test dataset. Based on these data, the P-R curve is pictured in Fig. 6(a). It shows that our method outperforms the others greatly. In addition, the minimum recall values of our method are obviously higher than those of the other methods, because our maps contain more pixels with the saliency value 255. Additionally, the F-Measure is utilized to measure the quality of saliency map. Figure 6(b) shows the F-Measure curve calculated based on the precision-recall values in Fig. 6(a).These two curves illustrates that our algorithm performs the other five methods in accuracy. As for the F-Measure curve, when the threshold is larger than 150, the blue curve that represents MS method is higher than our curve. The reason is that the saliency maps obtained via MS have larger saliency values than our saliency maps, which is clear in Fig. 2. However, that does not mean accuracy, which could be demonstrated by another evaluation metric, shuffled AUC (sAUC) [4]. In Table 2, the sAUC values are illustrated, which could prove the accuracy of our model further.

Notice that the Precision-Recall curve and F-Measure curve are less smooth than they appear in conventional saliency works whose algorithms are tested on the datasets containing more than 10,000 images. However, the light field dataset consists of only 100 images, so the curves seem to be unsmooth. In the future work, we will test the algorithm on larger datasets in order to improve its performance.

Besides, we adopt the fixed-threshold method to calculate precision-recall values and F-Measure values. The segmentation threshold is twice the average value of the whole saliency map. The performance is illustrated in Fig. 7(a).

Table 2 Shuffled AUC of 5 state-of-the-art methods vs. ours

Methods	SR	FT	CA	LR	MS	Ours
sAUC	0.52	0.58	0.64	0.67	0.69	0.78

Table 3 Comparison of running time (seconds per image)

Method	SR	CA	FT	LR	MS	Ours
Running time(s)	0.09	76.41	0.12	20.19	5.11	1.13
Code	Matlab	Matlab	Matlab	Matlab	Matlab	Matlab

Mean absolute error MAE is illustrated in Fig. 7(b)). It is obvious that the MAE result of our algorithm is the smallest compared with other five models, which demonstrates the high accuracy of our model.

Running time In Table 3, we compare average running time on light field dataset with other state-of-the-art algorithms mentioned above. We use the authors' code for the other five algorithms. All of the 6 algorithms are tested on an Intel i5 3.10GHz CPU with 8GB RAM. It shows that our method is faster than CA, LR and MS method. The faster method is SR, whose code contains only 5 sentences. However, SR could not handle the images with complicated background and foreground. Consequently, considering both the accuracy and the processing time, our approach has the best performance among all of the approaches.

5 Conclusion

In this paper, we proposed a saliency detection method tailored specifically for light field images based on depth feature. In contrast with traditional methods, our approach performs much better on 3D images. Through extracting depth feature hided in the focal stack, our method locates the background layer accurately. This step insures the accuracy of the result. Besides, the texture feature is extracted to improve the performance of contrast cues, which is hardly made use of before. Compared with the 5 state-of-the-art saliency detection models mentioned above, our method defeats them on light field image dataset with the help of depth feature. This paper is to share the novel saliency detection algorithm for 3D images, which will have broad prospect in the future. Achieving higher accuracy and robustness on other 3D image datasets are left as future works.

Acknowledgements This work is partially supported by the NSFC fund (61571259, 61531014, 61471213), and Shenzhen Research fund (JCYJ20160331185006518, GGFW2017040714161462).

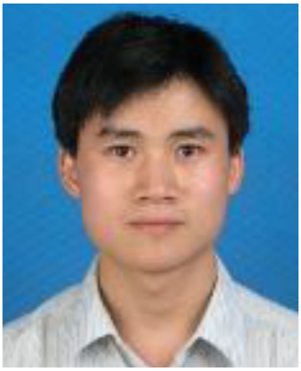
References

1. Ali B, Ming-Ming C, Huaizu J, Jia L (2014) Salient object detection: A survey, arXiv preprint arXiv: 1411.5878
2. Ali B, Ming-Ming C, Huaizu J, Jia L (2015) Salient object detection: A benchmark. IEEE Trans Image Process 24:5706–5722
3. Bowen J, Lihe Z, Huchuan L, Chuan Y, Ming-Hsuan Y (2013) Saliency detection via absorbing markov chain. Proceedings of the IEEE International Conference on Computer Vision(ICCV), 1665-1672
4. Bylinskii Z, Judd T, Oliva A, Torralba A, Durand F (2016) What do different evaluation metrics tell us about saliency models?, arXiv preprint arXiv:1604.03605

5. Congyan L, Nguyen TV, Harish K, Karthik Y, Mohan K, Shuicheng Y (2012) Depth matters: Influence of depth cues on visual saliency. Comput Vis-ECCV 2012:101–115
6. Dashan G, Sunhyoung H, Nuno V (2009) Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. IEEE Trans Pattern Anal Mach Intell 31:989–1005
7. Dorin C, Peter M (2002) Mean shift: A robust approach toward feature space analysis. IEEE Trans Pattern Anal Mach Intell 24:603–619
8. Guo-Xin Z, Ming-Ming C, Shi-Min H, Martin RR (2009) A Shape-Preserving Approach to Image Resizing. Comput Graph Forum 28:1897–1906
9. Hornung A, Pritch Y, Krahenbuhl P, et al (2012) Saliency filters: Contrast based filtering for salient region detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 733-740
10. Jianming Z, Stan S (2016) Exploiting surroundedness for saliency detection: aBoolean map approach. IEEE Trans Pattern Anal Mach Intell 38:889–902
11. Jin S, Haibin L (2011) Scale and object aware image retargeting for thumbnail browsing. Proceedings of the IEEE International Conference on Computer Vision(ICCV), 1511-1518
12. Junwei H, Ngi NK, Mingjing L, Hong-Jiang Z (2006) Unsupervised extraction of visual attention objects in color images. IEEE Trans Circ Syst Video Technol 16:141–145
13. Laurent I, Christof K, Ernst N, others (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 20:1254-1259
14. Michael R, Ariel S, Shai A (2008) Improved seam carving for video retargeting. ACM Trans Graph (TOG) 27(16):1–9
15. Ming-Ming C, Fang-Lue Z, Mitra NJ, Xiaolei H, Shi-Min H (2010) RepFinder: finding approximately repeated scene elements for image editing. ACM Trans Graph (TOG) 29:83
16. Ming-Ming C, Mitra NJ, Xiaolei H, Torr Philip HS, Shi-Min H (2015) Global contrast based salient region detection. IEEE Trans Pattern Anal Mach Intell 37:569–582
17. Na T, Huchuan L, Lihe Z, Xiang R (2014) Saliency detection with multi-scale superpixels. IEEE Signal Process Lett 21:1035–1039
18. Na T, Huchuan L, Xiang R, Ming-Hsuan Y (2015) Salient object detection via bootstrap learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 1884-1892
19. Nianyi L, Jinwei Y, Yu J, Haibin L, Jingyi Y (2014) Saliency detection on light field. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2806-2813
20. Qinmu P, Yiu-ming C, Xinge Y, Yan TY (2017) A Hybrid of Local and Global Saliencies for Detecting Image Salient Region and Appearance. IEEE Trans Syst Man Cybernet: Syst 47(1):86–97
21. Radhakrishna A, Sheila H, Francisco E, Sabine S (2009) Frequency-tuned salient region detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 1597-1604
22. Stas G, Lihi Z-M, Ayellet T (2012) Context-aware saliency detection. IEEE Trans Pattern Anal Mach Intell 34:1915–1926
23. Tao C, Ming-Ming C, Ping T, Ariel S, Shi-Min H (2009) Sketch2Photo: internet image montage. ACM Trans Graph (TOG) 28(124):1–10
24. Ueli R, Dirk W, Christof K, Pietro P (2004) Is bottom-up attention useful for object recognition?, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2, II-37
25. Xiaodi H, Liqing Z (2007) Saliency detection: A spectral residual approach. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 1-8
26. Xiaohui S, Ying W (2012) A unied approach to salient object detection via low rank matrix recovery. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 853-860
27. Xiaohui L, Huchuan L, Lihe Z, Xiang R, Ming-Hsuan Y (2013) Saliency detection via dense and sparse reconstruction. Proceedings of the IEEE International Conference on Computer Vision(ICCV), 2976-2983
28. Yuzhen N, Yujie G, Xueqing L, Feng L (2012) Leveraging stereopsis for saliency analysis. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 454-461
29. Zhao S, Yao H, Zhang Y, Wang Y, Liu S (2015) View-based 3D object retrieval via multi-modal graph learning[J]. Signal Process 112(C):110–118
30. Zhao S, Yao H, Gao Y, Ji R, Ding G (2016) Continuous Probability Distribution Prediction of Image Emotions via Multi-Task Shared Sparse Regression[J]. IEEE Transactions on Multimedia, PP(99):1-1
31. Zhu W, Liang S, Wei Y, Sun J (2014) Saliency optimization from robust background detection. Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR), 2814-2821

Haoqian Wang received the B.S. and M.E. degrees from Heilongjiang University, Harbin, China, in 1999 and 2002, respectively, and the Ph.D. degree from the Harbin Institute of Technology, Harbin, in 2005. Since 2005, he has been a faculty member of the Graduate School at Shenzhen, Tsinghua University, Shenzhen, China, and is currently an Associate Professor. His current research interests include video communication and signal processing.

Bing Yan received the B.S. degree from Xidian University, Xi'an, China, in 2015. She is currently working toward the M.E. degree in Graduate School at Shenzhen, Tsinghua University, Shenzhen, China. Her research interest is in saliency detection.



Xingzheng Wang received the B.Sc. degree in mechanical engineering and automation from the North China University of Technology, Beijing, China, in 2004, the M.Sc. degree in mechanical engineering from Tsinghua University, Beijing, in 2007, and the Ph.D. degree in computer science from Hong Kong Polytechnic University, Hong Kong, in 2013. He has been a Post-Doctoral Fellow with the Graduate School at Shenzhen, Tsinghua University, Shenzhen, China, since July 2013. His research interests include computer vision, pattern recognition, and video processing.

Yongbing Zhang received the B.A. degree in English and the M.S. and Ph.D degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2004, 2006, and 2010, respectively. He joined Graduate School at Shenzhen, Tsinghua University, Shenzhen, China in 2010, where he is currently an associate professor. He was the receipt of the Best Student Paper Award at IEEE International Conference on Visual Communication and Image Processing in 2015. His current research interests include video processing, image and video coding, video streaming, and transmission.

Yi Yang received the B.S. degree from Department of Engineering Physics, Tsinghua University, Beijing, in 1992. After graduation she worked in the Ministry of Nuclear Industry 261 plant (Beijing). In 1994, she entered the United States EG & G China Representative Office. From 1996 to present, she has founded LUSTER LightTech Co., Ltd. in Beijing, mainly responsible for image and machine vision product sales and business development.