Pattern Recognition 47 (2014) 1740–1750

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Salient object detection via local saliency estimation and global homogeneity refinement

Hsin-Ho Yeh [a], Keng-Hao Liu [a,*], Chu-Song Chen [a,b,**]

[a] Institute of Information Science, Academia Sinica, Taipei, Taiwan
[b] Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 19 February 2013
Received in revised form
29 October 2013
Accepted 13 November 2013
Available online 23 November 2013

Keywords:
Salient object detection
Local contrast
Global homogeneity

ABSTRACT

This paper presents a new hybrid approach for detecting salient objects in an image. It consists of two processes: local saliency estimation and global-homogeneity refinement. We model the salient object detection problem as a region growing and competition process by propagating the influence of foreground and background seed-patches. First, the initial local saliency of each image patch is measured by fusing local contrasts with spatial priors, thereby the seed-patches of foreground and background are constructed. Later, the global-homogeneous information is utilized to refine the saliency results by evaluating the ratio of the foreground and background likelihoods propagated from the seed-patches. Despite the idea is simple, our method can effectively achieve consistent performance for detecting object saliency. The experimental results demonstrate that our proposed method can accomplish remarkable precision and recall rates with good computational efficiency.

1. Introduction

Identifying the visually attentive area in an image without a priori knowledge about the scene is the fundamental to our vision. Let us take a look at the image shown in Fig. 1(a) as an example. In this figure, people can easily identify the most salient words appeared. This image was generated by wordle.net, a web-tool converting a textfile into a figure. It extracts the keywords frequently occurred in the textfile and generates a figure consisting of those words for visualization, where a higher-frequency word is generated to occupy a larger area with more distinctive colors or brightness, which thus presents higher "visual saliency." Note that such a salient-words synthesis process can be seen as an inverse problem of salient object detection. Since the textfiles to be submitted are arbitrarily composed of web users, there is no prior knowledge about the frequency of the words (almost all words, except several stop words, have the same opportunities of becoming the keywords). Nevertheless, due to the bottom-up nature of saliency region detection, people can still find the salient objects even when there is no high-level or prior object-categorical information.

Salient regions, referred to as the image part attracting most human's attention, are fundamental for many high-level tasks in computer vision. Recent advance shows that salient regions provide useful information for image segmentation [1], object recognition [2,3], and motion detection [4]. They also contribute to improve the performance of many image quality metrics [5] and aesthetics value assessment [6–8]. Early studies of computational saliency [9,10] used biology-inspired models to simulate the selection mechanism of Human Vision System (HVS). In such model, visual input is decomposed into a set of feature maps by several pre-attentive filters based on image features such as colors or intensity. The saliency in a feature map is regarded as the competition result of neurons (or locations) in the spatial domain, where the prominent neurons significantly differ from their surrounds can survive. Then, the saliency intensities in all feature maps are integrated into a final map.

Later, several studies were introduced to enhance the saliency detection performance of biology-inspired models. Gopalakrishnan et al. [11] and Wang et al. [12] employed random walk to model the competition process. Valenti et al. [13] designed several new pre-attentive filters based on edge and color responses. Wang et al. [14] learned the dictionary from a large-scale set of images for estimating the intrinsic and the extrinsic saliency. Murray et al. [15] explored the integration step of the feature maps for further refinement purpose.

In practice, the performance of most biology-inspired methods is still limited, which could be owing to the difficulties encountered in choosing or learning the filters. Thus, various approaches

* Corresponding author.
** Principal corresponding author at: Institute of Information Science, Academia Sinica, Taipei, Taiwan. Tel.: +886 2 27883799x1310.
E-mail addresses: hhyeh@iis.sinica.edu.tw (H.-H. Yeh), keng3@iis.sinica.edu.tw (K.-H. Liu), song@iis.sinica.edu.tw (C.-S. Chen).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.11.015

Fig. 1. What are the salient keywords? (a) wordle.net provides us an efficient tool to summarize frequently appeared topic words in a document. Such a saliency-synthesis procedure poses an inverse problem of saliency detection. (b) The saliency detection result (per-pixel saliency degree) obtained by our proposed method. (c) The objective ground truth.

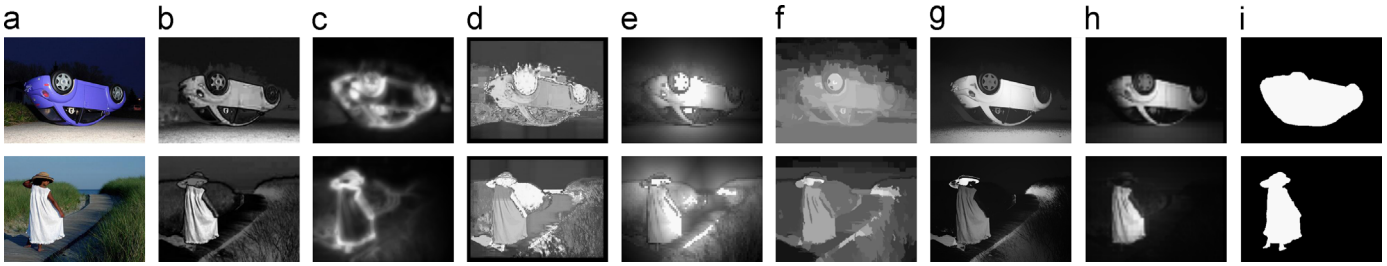

Fig. 2. Examples of salient object detection. (a) shows the original images. (b)–(f) demonstrate the state-of-the-art saliency maps from FT [16], CA [20], SS [18], CAS [21], RC [22], and SF [23]. The map obtained by our method LGA(h) is plotted as well. The (i) shows the ground truths (GT). From these results, the propose salient methods not only detect the salient regions well, but also reduce the noises better in the background clutter.

adopting other kinds of models or frameworks were proposed to resolve this dilemma. For instance, Achanta et al. [16] designed a band-pass filter to preserve a reasonable range of spatial frequency for detecting the salient regions. Lu et al. [17] found that the shape information, such as convexity and concavity, can reveal significant clues for locating the salient regions. In general, these approaches employ global information via frequency or shape analysis. However, they would often fail to detect small objects due to the fact that distinctive smaller regions are easily overlooked.

In order to solve the certain issues suffered from the global-based methods, another technique, referred to as local contrast, was proposed. It was derived from the concept of the so-called "center-surround" concept. Center-surround contrast has already been employed by Itti et al. [9,10], which reflects the local spatial-discontinuities of visual contrast. Local contrast, inspired from this mechanism, was thus developed to emphasize the uniqueness of a certain region by accentuating the contrasts to neighbors. It has been widely exploited by window-based approaches in [18,19]. Rahtu et al. [18] conducted an effective Bayesian formulation to measure the salient regions from color. To present the regionality, they separate a sliding window into the inner and collar sub-windows. Then a pixel's saliency is directly determined by the color contrast between the two sub-windows. This approach achieves nice results owing to the appropriate consideration of local contrasts. Klein et al. [19] further extended it by using Kullback–Leibler divergence. However, without knowing the object size, the methods of local contrast have to change their window size to locate the saliency for different scales of object. The problem caused by object-sized variation would degrade their performance significantly in many cases.

Based on the above-mentioned reasons, how to relax the spatial constraints of window-based approaches could be a main issue of improving the performance. Goferman et al. [20] presented the spatially weighted color contrast (or called the surrounding contrast) for saliency detection. The surrounding contrast of a pixel is measured from the weighted distances (the color distances multiplied by the inverse of spatial distances) to the other patches. The saliency values of different scales are then averaged to obtain a single saliency map. Although their results are useful for certain applications such as image re-targeting, this approach usually suffers from the problem that the detection results are sensitive to the edges in an image, and tends to overlook homogeneous foreground regions.

From the above point of view, an ideal contrast-driven saliency detector should take both local perspective and global-homogeneous property into consideration. Many recent studies focused on the local saliency estimation based on the principles such as Rareness [18,24], Contrast [10,23], and Center-Surrounding [19,25]. They might neglect that a salient region usually differs from its neighborhood outside the region besides containing homogeneous parts inside the region. How to link locally distinctive patches/regions that could be homogeneously associated to the foreground is a crucial key.

In this paper, we propose to detect the salient objects from an initial local saliency estimation process, where several seed regions can be selected from the initialization. Then a homogeneity-growing competition process is designed to precisely locate the salient foreground region. Our approach can perform better than state-of-the-art results (such as [16,18,20–23], and is efficient to implement. Fig. 2 shows the examples of salient object detection results obtained by various methods.

This paper is organized as follows. Some of the related works are discussed in Section 2. The problem formulation and the proposed method are introduced in Sections 3 and 4, respectively. Experimental results and comparisons are shown in Section 5. Finally, conclusions and future works are drawn in Section 6.

2. More of the related works

In addition to the above-mentioned works, new methods about saliency detection is proceeding to be explored. For instance, Duan et al. [26] further addressed this problem by employing the PCA

to reduce the patches' dimensionality. Thus some noises are eliminated in the calculation of saliency. Yeh et al. [21] proposed a contrast-aware method that combines local and global contrasts to achieve the compactness nature of salient objects. Huang et al. [25] presented a unifying perspective that the saliency area of an image is described by center-surround divergence. Vikram et al. [27] computed image saliency map based on local saliencies among random rectangular regions of interest.

Recently, using a prerequisite technique to increase the accuracy of saliency detection became a main trend. Feng et al. [28] presented a method for detecting window saliency from the composition prospective, in which a segmentation method is employed to decompose image. They assume that the cost for composing a region with visual uniqueness is more than that of composing an ordinary region. Cheng et al. [22] developed a region-based saliency detector by combining the regional representation with the histogram-based acceleration, where the spatial weighted contrast and image pre-segmentation are adopted. Perazzi et al. [23] took similar idea to introduce a contrast-based-filtering method based on super-pixel pre-segmented abstraction images.

Despite the above works [22,23,28] reach good success in both accuracy and efficiency, using a pre-processing technique (i.e. image segmentation) as a prerequisite undoubtedly brings additional computational cost. Furthermore, the pre-segmentation would be sensitive to different image structure and noises, which could make it difficult or inflexible to accommodate on various kinds of datasets. Our work does not rely on any complicated pre-processing techniques, which still produces comparative saliency results.

3. Problem formulation and our framework

We propose a two-stage approach where a local contrast measure is implemented to detect the initial saliency seeds, then a global-homogeneity propagating method is followed to generate and refine the saliency map.

A salient object detector f takes an $n \times m$ input image I and generates a saliency map $SM \in \mathbb{R}^{n \times m}$, where the salient value for a pixel x (i.e. $SM(x) \in [0,1]$) indicates its saliency degree. The higher the $SM(x)$ is, the more salient the pixel x is. In this paper, we use CIELab color space[1] to represent the visual input since it has shown the high efficiency for detecting image saliency [18,20].

Our method adopts patch-based representation instead of pixel-wise process to reduce computational complexity. An input image I is divided into r non-overlapped square patches $\{P_1, P_2, \ldots, P_r\}$ where each patch P_i is located in 2D spatial location $\mathbf{l}_i \in \mathbb{R}^2$, for $1 \le i \le r$. Each patch presents a $w \times w$ sub-image, thus its colors can be vectorized into $\mathbf{c}_i \in \mathbb{R}^{w \times w \times 3}$. We set $w=5$ in the implementation.

A patch is then represented as a vector $P_i = [\mathbf{l}_i, \mathbf{c}_i]^T$ by concatenating its spatial location and colors. We further define $d_{spa}(P_i, P_j) = \|\mathbf{l}_i - \mathbf{l}_j\|_2$ and $d_{col}(P_i, P_j) = \|\mathbf{c}_i - \mathbf{c}_j\|_2$ to represent the distance between patches P_i and P_j in the spatial and color domains, respectively, where the $\|.\|_2$ denotes the L2 norm.

3.1. Overview of our approach

An overview of proposed framework is given at first. Each patch P_i will be associated with a *local saliency* (LS) *strength* (denoted as $LS(P_i) \in [0,1]$) in advance, and how to compute it will be detailed later. Later, we select some patches as the *foreground seeds* and some others as the *background seeds*, where the former consist of patches of higher LS strengths, and the latter consist of patches of lower LS strengths.

In our empirical study, no matter how a method is designed to estimate the LS strengths, the method based simply on such local (or relatively local) information around a patch can only provide us several isolated spots. It will then fail to detect the entire salient object. Hence, instead of using the LS information only, we choose the foreground seeds that are highly probable to be part of the salient object at first. These foreground seeds present the 'distinctive' or 'uniqueness' spots/locations in an image. Then, we seek to refine the results by finding the other salient patches of the object based on their similarities to the foreground seeds in the space-color domain. An example of the foreground seeds conducted by our approach is shown in Fig. 3(b) (green dots).

However, locating the salient objects based only on the foreground seeds would be difficult to yield a discriminating result. Hence, we further consider the background seeds that are the most 'repetitive' patterns which are almost impossible to be part of the salient object. An example of the background seeds is shown in Fig. 3(b) (red dots). Regions grown from the foreground and background seeds based on the space-temporal similarities then form a competitive process, yielding a more discriminating decision of the salient detection results.

Our proposed framework thus includes two stages. Firstly, the local saliency seeds are estimated by the uniqueness property via a local contrast measure $LS(\cdot)$. Secondly, the global-homogeneity seed-propagating method is used to refine the saliency map. The details are given in the following.

3.2. Seed propagation process and saliency map presentation

In this section, we assume that the foreground and background seeds have been given, and focus on the problem of constructing the saliency map based on these seeds. Denote S_{FG} and S_{BG} to be the patch sets consisting of the foreground and background seeds, respectively. How to construct S_{FG} and S_{BG} will be presented in Section 4. Basically, these seeds are determined so that they belong to the foreground (i.e., saliency region) or background (i.e., non-saliency region) with high confidence as discussed above. We formulate the salient object detection as a process of propagating the strengths of the foreground and background seeds and then construct the saliency map $SM(\cdot)$ based on the propagated strengths.

More specifically, given a foreground seed patch $Z \in S_{FG}$, its propagated strength to an arbitrary patch $P \in \{P_1, P_2, \ldots, P_r\}$ is defined by the distance between P and Z in the space-color domain and the local saliency degree of P, and is formulated by the following distribution:

$$Prob_{FG}(P|Z) = C_F \exp\left(-\frac{d_{cor}(P,Z)}{\sigma_{cor}}\right) \exp\left(-\frac{d_{spa}(P,Z)}{\sigma_{spa}}\right) LS(P), \quad (1)$$

where C_F is a normalization constant making the summation of the probabilities $Prob_{FG}(P|Z)$ over $P \in \{P_1, P_2, \ldots, P_r\}$ equal to 1.

We employ a mixture distribution to superimpose the influence of seeds. The foreground likelihood is then defined as the mixture distribution centered at the foreground seeds with equal weights:

$$Prob(P|S_{FG}) = \sum_{Z \in S_{FG}} \frac{1}{|S_{FG}|} Prob_{FG}(P|Z), \quad (2)$$

where $|S_{FG}|$ is the cardinality of S_{FG}.

Similarly, for a background seed $Z \in S_{BG}$, its propagated strength to an arbitrary patch $P \in \{P_1, P_2, \ldots, P_r\}$ is defined by the distance

[1] We normalize the values into [0,1] for each channel in CIELab color space.

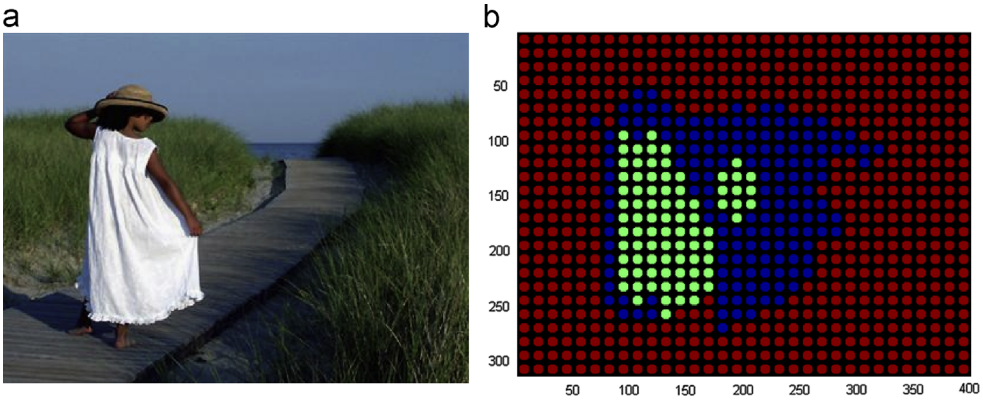

Fig. 3. Example of the foreground and background seeds. (a) shows the original image. (b) shows the geographical location of foreground seeds (green), background seeds (red), and indeterminate region (blue). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

between P and Z and the local non-saliency degree of P:

$$Prob_{BG}(P|Z) = C_B \exp\left(-\frac{d_{cor}(P,Z)}{\sigma_{cor}}\right) \exp\left(-\frac{d_{spa}(P,Z)}{\sigma_{spa}}\right)(1-LS(P)), \tag{3}$$

where C_B is a normalization constant making the summation of the probabilities $Prob_{BG}(P|Z)$ over $P \in \{P_1, P_2, \ldots, P_r\}$ equal to 1. The background likelihood is then defined as a mixture distribution given the background seeds:

$$Prob(P|S_{BG}) = \sum_{Z \in S_{BG}} \frac{1}{|S_{BG}|} Prob_{BG}(P|Z), \tag{4}$$

We use the likelihood ratio $Prob(P|S_{FG})/Prob(P|S_{BG})$ of the foreground and background to present the saliency degree of image patch P. We deploy a alternative form which can limit the saliency output for better visual representation as the following form:

$$SM(P) = \frac{1}{1+\frac{Prob(P|S_{BG})}{Prob(P|S_{FG})}}. \tag{5}$$

In our implementation, the scale parameters σ_{cor} and σ_{spa} above are set as 1 and 200, respectively, for a 400×300 image in CIELab color space. How to evaluate the local saliency $LS(P)$ and build the seeds S_{FG} and S_{BG} remain an issue, which will be introduced in the next section.

4. Local saliency estimation and seeds generation

Estimating the local contrast property is regarded as the fundamental step in many works for visual-attention evaluation [9,10,21,23,29]. We name the saliency generated from such property as local saliency. The local saliency can catch the regional uniqueness, which is easily ignored by global-contrast-based saliency detector. According to a recent study conducted by Huang et al. [25], most of the existing bottom-up saliency estimation methods can be explained by a unified center-surround principle. From this perspective, most approaches share the same idea that the saliency degree is determined from the dissimilarity between the center and surround regions, which can be jointly represented in a mathematical framework called center-surround divergence in [25]. Different approaches are simply variations of computing the center-surround divergences.

In our work, without loss of generality, we enhance the idea of one of the approaches, surround-contrast (SC) in [20], for local saliency degree estimation. We introduce a new method called Surrounding Contrast with Consistency (SCC) that employs further the patch-consistency information to construct an improved local contrast measure. In the following, we first review the SC in Section 4.1, and then introduce the SCC in Section 4.2.

4.1. Review of surrounding contrast (SC)

Salient regions usually contain higher local contrasts since they are distinctive from their neighborhoods. Thus, an appropriate contrast detector is required to measure such 'uniqueness' property. In traditional methods, the local contrast is merely measured by color distance so that it makes no difference when any two patches are spatially far-away from or close to each other. Because the salient regions tend to group together in spatial domain, the spatial distance should be further considered for better local contrast presentation.

To do so, the surrounding contrast (SC) was proposed [20]. It assumes that the nearby patches of current location play more importance than the far-away ones in local contrast acquisition. By simply weighting with the inverse of spatial distance, the 'color gaps' between nearby patches and current patch become more conspicuous than distant ones. The SC is defined as

$$SC(P_i) = \sum_{P_j, \forall j} \frac{d_{col}(P_i, P_j)}{1+\lambda \times d_{spa}(P_i, P_j)}, \tag{6}$$

where λ is use to control the proportion of color/spatial weighting. (We use $\lambda = 1$ in our implementation.)

4.2. Surrounding contrast with consistency (SCC)

Although SC is useful to highlight local salient regions, the salient response of object boundaries is often too strong, as the results shown in [20]. The reason is that the patches located inside the same object may have similar colors with their neighbors. It is often the case that they are failed to be highlighted by their surrounding patches. In this case, Eq. (6) fails to highlight the salient regions inside an object. A vivid example is shown in Fig. 4b where the patches inside the yellow leaf were failed to be detected.

To address this issue, we consider again that the contrasts for those color-similar patches inside a salient object should be consistent; the local contrast of a certain image patch P_j is determined by its similar neighbors in color domain instead of purely by itself. This newly designed contrast, called surrounding contrast with consistency (SCC), is calculated by the weighted sum of SC values of k color-similar patches of current patch. The SCC is defined as

$$SCC(P_i) = \sum_{P_j \in N_k(P_i)} Consistency(P_i, P_j) \times SC(P_j), \tag{7}$$

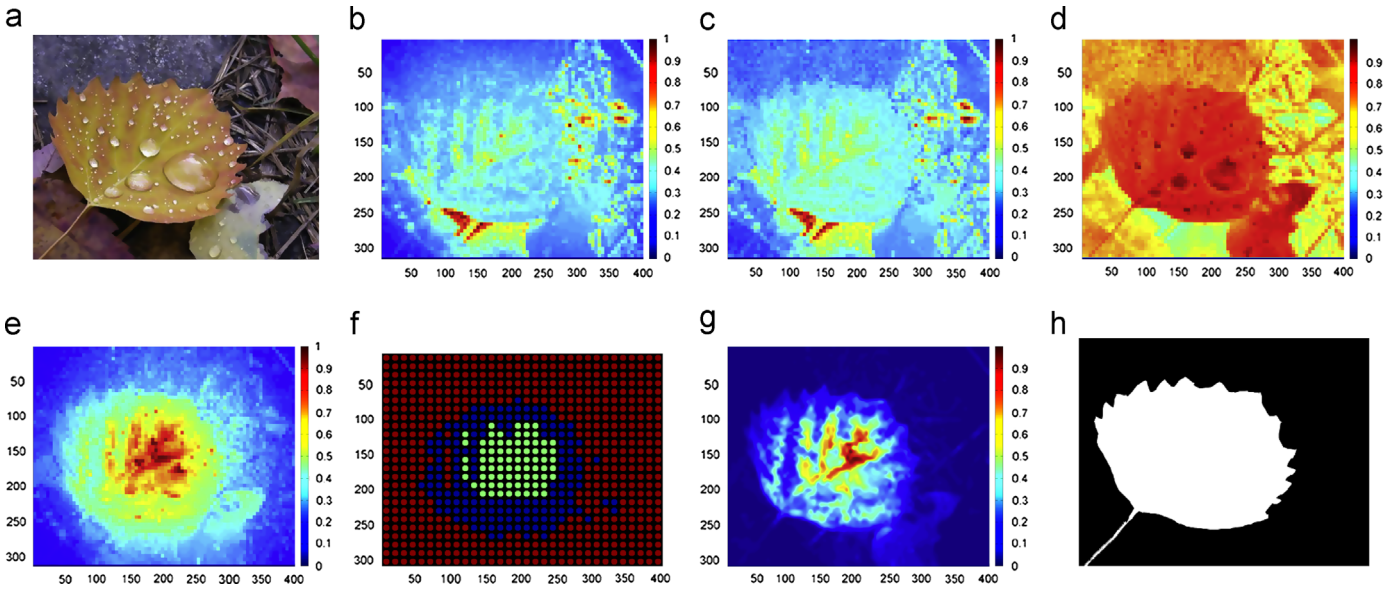

Fig. 4. (a) shows the original image. (b) and (c) show the local contrast results of SC and SCC, respectively. (d) shows the IP for the given image. With the combination of (c) and (d) with SCP, the local saliency is obtained and plotted in (e). Based on the local saliency, the geographical location of foreground seeds (green), background seeds (red), and indeterminate seeds (blue) are depicted in (f). (g) shows the saliency map detected by the proposed method, and (h) is the given ground truth. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

where $N_k(P_i)$ contains the k-nearest neighbors for current patch P_i in terms of color distance. The term *Consistency* measures the weighted color similarity between patches P_i and its neighbor P_j:

$$Consistency(P_i,P_j)=\frac{\exp(-d_{cor}(P_i,P_j))}{\sum_{P_h\in N_k(P_i)}\exp(-d_{cor}(P_i,P_h))}. \tag{8}$$

By taking this design in (7), the issue of inconsistent contrast values inside an object can be resolved. One should note that SC is a special case of SCC when $k=1$. As k is growing, the salient values of the patches inside the same object are thus able to be highlighted. However, too large k will over-smooth the saliency between the foreground and the background patch. In our implementation, we adopt $k=8$. A visualization result is illustrated in Fig. 4c. According to the result, the regions near leaf's veins were able to be filled by SCC. It shows that SCC can consistently emphasize the salient regions for the patches located either in the salient region or in its border.

4.3. Fusion with prior information and seeds generation

In general, the patches with higher luminance easily arouse the audience's attention [30], and people naturally select the regions of interest close to the image center [31]. Thus, we simply use Intensity Prior (IP) and Spatial Center Prior (SCP) as supportive prior information to enhance the saliency results:

$$IP(P_i)=\exp(mean_L(P_i)), \tag{9}$$

where $mean_L$ calculates the averaged value of P_i respect to L channel, and

$$SCP(P_i)=\exp\left(-\frac{d_{spa}(P_i,\overline{P})}{\sigma_{spa}}\right), \tag{10}$$

where $\overline{P}$ denotes the center patch of image and σ_{spa} is the same as that in Eqs. (1) and (3).

We integrate the above formulation to acquire our local saliency (LS) measure. The LS for patch P_i is defined by combining Eqs. (7), (9) and (10):

$$LS(P_i)=SCC(P_i)\times IP(P_i)\times SCP(P_i). \tag{11}$$

Finally, the values of $LS(P_i)$ are limited to a certain range [0,1]. (i.e. Those values that are larger than one are assigned one, and negative values are assigned zero.) A simple result of LS map is shown in Fig. 4e.

To generate foreground and background seeds, the image patches are first ranked by their degrees of LS $(P_i), \forall i\le r$. Then we simply assign the foreground seeds (S_{FG}) which contain the patches with higher LS degrees, and background seeds (S_{BG}) which contain those with lower LS degrees. More specifically, the S_{FG} collects the patches which are ranked in top $\gamma_f\%$, and the BS gathers the patches ranked in the lowest $\gamma_b\%$:

$$S_{FG}=\{P_j|\phi(P_j)\le r\times\gamma_f\%\}, \tag{12}$$

$$S_{BG}=\{P_h|\phi(P_h)\ge r\times(1-\gamma_b\%)\}, \tag{13}$$

where $\phi(\cdot)$ is the sorting function in descending order according the LS values.

The boundaries between foreground and background that cannot be explicitly defined are referred to as an 'indeterminate' or 'unknown' class, such as well-known Trimap Segmentation [32]. Similarly, such a gap remains between S_{FG} and S_{BG} (i.e., $\gamma_f\le(1-\gamma_b)$) to avoid grading the performance caused by such uncertain regions. The ambiguity is then solved by the seed region propagation process from the mixture distributions of the foreground and background introduced in Section 3.2.

After the $LS(\cdot)$ of each patch is estimated and S_{FG} and S_{BG} are generated, we then use Eqs. (2) and (4) to compute the foreground and background likelihoods, respectively. Finally, the saliency map is determined by Eq. (5). Our algorithm is named as LG and its procedure is given as follows.

Algorithm 1. Local seed global homogeneity (LG).

Input A color image.

1: Input image separates into patches $\{P_1,P_2,\ldots,P_r\}$.
2: For each patch, compute the local saliency $LS(P_i)$ using Eq. (11).
3: Select foreground seeds and background seeds using Eqs. (12) and (13).
4: Calculate foreground and background likelihoods using Eqs. (2) and (4), respectively.

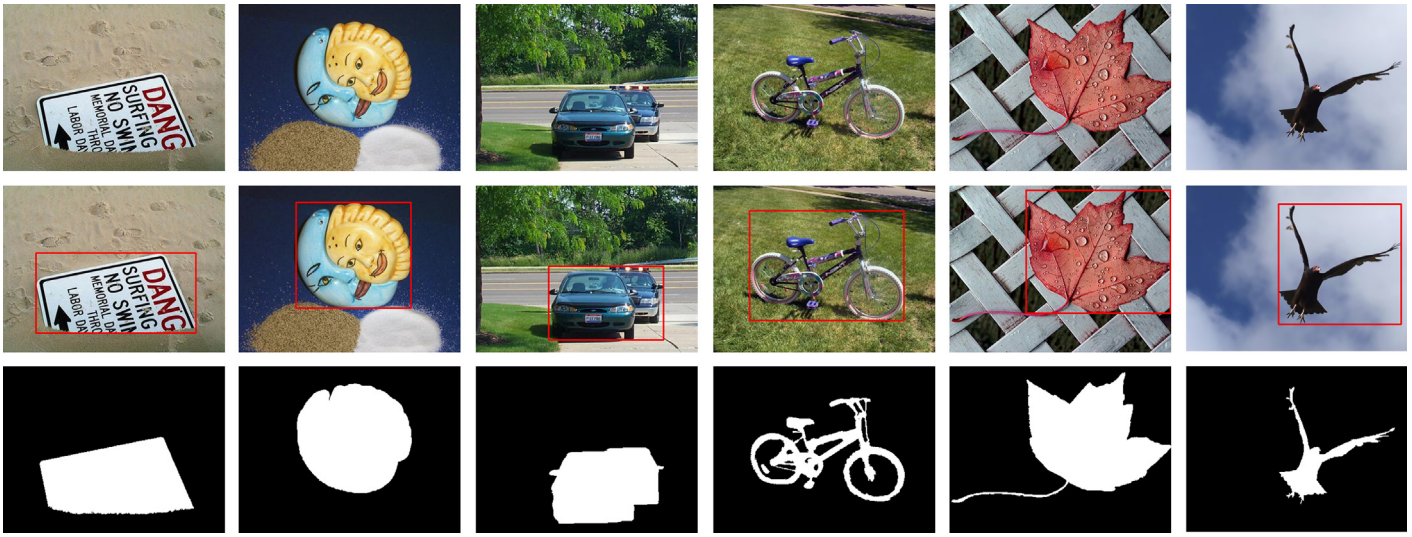

Fig. 5. The examples and the corresponding ground truth of MSRA-*B* and MSRA-ASD datasets. The first row shows the original sample images. The second row displays the window-based ground truth images (for MSRA-*B*). The last row shows the pixel-wise ground truth images (for MSRA-ASD).

5: Compute saliency degree $SM(P_i)$ using Eq. (5).
Output: A saliency map reconstructed for all $SM(P_i)$, $i=1,2,\ldots,r$.

5. Experimental results

5.1. Datasets

Our experiments are implemented on two benchmark datasets which are widely used in related works. According to the comprehensive investigation in [33], there are up to five benchmark datasets for salient object detection. Among those datasets, we choose the two largest ones, MSRA-*B* and MSRA-ASD, for our experiments because they contain larger amounts of images and are regarded as the primarily comparative and representative datasets in most works.

MSRA-B[2]: This dataset contains 5000 images, and each image was labeled by 9 users requested to draw a rectangle bounding the salient object [29] as the multiple ground truth. Its sample images are shown in the first row of Fig. 5 and the corresponding window-based ground truth is shown in the second row of Fig. 5.

MSRA-ASD: It is a pixel-wise salient object dataset released by Achanta et al. [16]. They thought the window-based ground truth is imprecise for evaluating the performance of object detection. They thus refined the ground truths of 1,000 images selected from the MSRA-*B* dataset into a pixel-wise matter[3] as shown in the last row of Fig. 5.

Undoubtedly, the ground truth of MSRA-ASD is annotated more accurately, and thus MSRA-ASD should be a benchmark for evaluating the performance of salient object detection. Nevertheless, MSRA-ASD contains a relatively small portion of images, which may degrade its reliability. Therefore, we use both MSRA-ASD and MSRA-*B* databases to do a more comprehensive evaluation in the experiments.

5.2. Implementation

Except for the implementation of proposed LG, we further conduct an approximated version of LG, named as LGA, to increase computational efficiency in local saliency estimation.

Once patch representation is done by step 1 in Algorithm 1, the image is further represented by non-overlapping square blocks with size $(2l+1)w \times (2l+1)w$, where w denotes the patch size defined in Section 4. For each block, the step 2 in Algorithm 1 is simplified by which we only compute the local saliency value $LS(\cdot)$ for the center patch (of size $w \times w$), and then duplicate the value to the other patches in the block. In the experiments, we use two versions of the algorithm by setting $l=0$ and $l=1$, which are referred to as LG (w/o local saliency approximation) and LGA (with local saliency approximation), respectively. Both of them are compared to recent studies including FT [16], CA [20], RC [22], SS [18], BITS [19], SF [23], CAS [21] and ICC [13] on the MSRA-ASD and MSRA-*B* datasets. The results are shown below.

In addition, we set $\gamma_b = 70$ and $\gamma_f = 10$ for the experiments conducted in Sections 5.3 and 5.4.

5.3. Performance evaluation using MSRA-ASD database

MSRA-ASD database provides 1000 images, where each image has its corresponding pixel-wise binary ground truth (i.e. foreground/background). We compare the segmentation performance in a fixed threshold condition, referred to as precision-and-recall, which is popularly used in related literature. Since the ground truth is binary, the saliency map should be evaluated by varying the threshold T from 0 to 255. More specifically, for each threshold T, we compute its averaged precision and recall value among all 1000 images. Then those values are collected and used to plot a PR curve as shown in Fig. 6. Such evaluation setting is same as the compared methods: FT [16],[4] CA [20],[5] RC [22],[6] SS [18], BITS [19], SF [23],[7] and CAS [21].

Fig. 6(a) shows the comparative quantitative performance using local saliency w/o and with global-homogeneity refinement. Apparently, the refinement process can improve the precision performance. It implies that global refinement process indeed smoothens the salient spots in the homogeneous regions very well. Fig. 6(b) further displays the comparative analysis between

[2] http://research.microsoft.com/en-us/um/people/jiansun/SalientObject/salient_object.htm
[3] http://ivrgwww.epfl.ch/supplementary_material/RK_CVPR09/index.html

[4] This result is from their project page, one should note the erratum.
[5] Since Goferman et al. do not use this dataset for comparison, we use the results generated by their released code instead.
[6] This performance is obtained from the saliency map released by the authors. See http://cg.cs.tsinghua.edu.cn/people/~cmm/saliency/ for details.
[7] The results of [18,19,23] are reported from their papers directly.

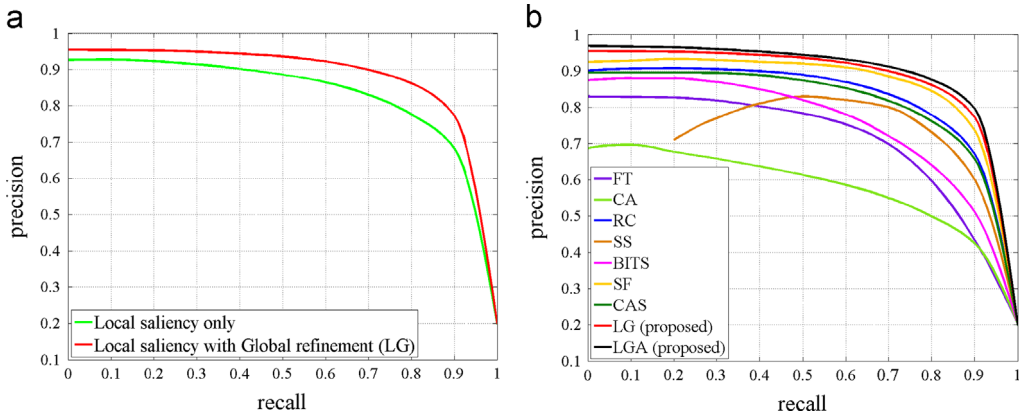

Fig. 6. The comparative performance of averaged precision and recall curves among 1000 images in MSRA-ASD database. (a) The performance using the proposed local saliency only, and local saliency with global refinement (LG). (b) The exhaustive comparison of our proposed methods to state-of-art approaches. The compared methods are FT [16], CA [20], RC [22], SS [18], BITS [19], CAS [21], SF [23], LG (the proposed method w/o approximation), and LGA (the proposed method with approximation).

Table 1
The p values of paired t-test implemented on state-of-arts and the proposed LGA at different recall settings.

Method	FT [16]	CA [20]	SS [18]	CAS [21]	RC [22]	SF [23]
Recall = 0.1	4.2e − 092	3.3e − 155	7.8e − 066	5.8e − 034	3.2e − 015	6.0e − 004
Recall = 0.3	3.0e − 114	2.9e − 193	3.3e − 052	6.9e − 031	3.5e − 017	3.0e − 005
Recall = 0.5	5.3e − 139	9.1e − 211	4.3e − 050	1.6e − 029	2.5e − 015	3.1e − 007
Recall = 0.7	1.5e − 183	3.9e − 214	2.1e − 055	1.7e − 035	1.3e − 010	3.2e − 007
Recall = 0.9	2.8e − 264	1.5e − 199	5.2e − 095	2.6e − 062	4.5e − 010	1.4e − 007

state-of-art methods and the proposed LG and LGA. As the results shown in the figure, LG demonstrates remarkable performance. It significantly outperforms almost compared methods, and achieves roughly 80% precision while maintaining 90% recall under a specific threshold. It suggests that with a simple threshold, the segmented foreground and background regions are highly consistent with human's experience. Surprisingly, LGA further improves LG. The possible reason that LGA surpasses LG is that the additional approximation process further shares the saliency information spatially in an appropriate object scale so that it makes local salient values more uniformly distributed inside the object instead of being accumulated in border.

To further identify whether the performance improvement is significant, we analyze the experimental results conducted on MSRA-ASD dataset in a statistical way. We follow the method in [34] to apply the paired t-test between each state-of-art method and the proposed LGA. (i.e., we implement t-test on all the pairs [LGA, X], where X is selected as FT, CA, SS, CAS, RC, and SF). To do so, for each obtained saliency result of each method, we fix the recall at a ceratin value and calculate the corresponding precision value for the testing purpose. So the saliency results of each method are represented by 1000 values for a fixed recall. Table 1 lists the p values generated by the paired t-test with the recall values setting on 0.1, 0.3, 0.5, 0.7, and 0.9. All the p values shown in the Table are very small, and all the null hypothesis are thus valid under the significant level 0.05. Statistically, this test verified the effectiveness of proposed method.

The visual results are shown in Fig. 7 for additional qualitative comparison. Apparently, our proposed methods LG and LGA own better capability either in salient object detection or background suppression. For instance, LG and LGA can detect the leaves of flower shown in the fifth row of Fig. 7 while others barely detect the borders or the center part of flower. Note that the results of LG and LGA are similar but actually different. Compared to LG, the LGA can produce more uniform and consistent saliency values inside the objects.

On the other hand, the computational efficiency is another point worthy of being noted. Table 2 tabulates the average computation time of some state-of-arts and the proposed methods processing on MSRA-ASD images. Those time were calculated by the average cost of ten runs among all images with the PC of Intel Core i7 2.50 GHz, 12GB RAM and Matlab2011b. The results of RC are directly reported by [23] since we could not implement its code in a batch mode. Based on Table 2, our proposed LG and LGA require less time to process MSRA-B images than SS, CA, and CAS do. Meanwhile, they retain better precision-and-recall performance as shown in Fig. 6(b). Additionally, it should be noted that both RC and SF are implemented in C++ so that they require much less processing time than those implemented by Matlab. The difference in computing time required by Matlab and C++ can be observed in the case of FT. Hence, we expect that the cost of LG and LGA can be significantly reduced by using C++. In conclusion, our proposed LG and LGA reach a good balance between detection performance and computational efficiency among all the compared methods.

5.4. Performance evaluation using MSRA-B database

Unlike the wide use of MSRA-ASD dataset in the literature, there are only very few works (such as the ICC [13] and [29]) showing the experiments on MSRA-*B* dataset because it is more challenging due to its larger data volume. Since there is no standard way to evaluate the saliency performance for MSRA-*B*, we follow the criterion in [13] for the comparison. The procedure is described as follows.

5.4.1. Ground truth preparation

As mentioned in Section 5.1, each image in MSRA-*B* owns nine window-based ground truths that locate the salient objects. To unify them as single one, we generate an averaged ground truth by

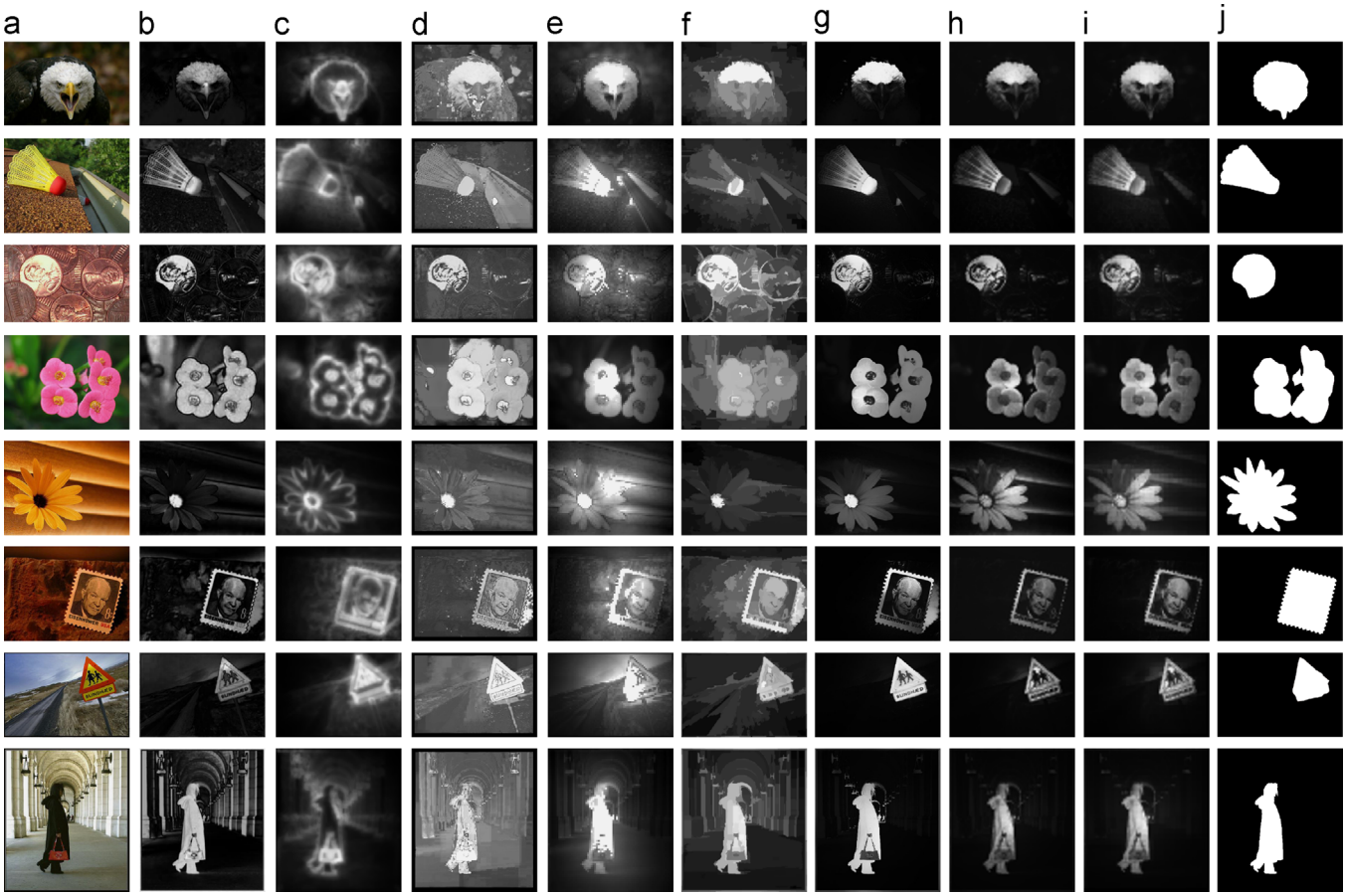

Fig. 7. (a) shows some example images. (b)–(g) demonstrate the detected results from FT [16], CA [20], SS [18], CAS [21], RC [22], and SF [23]. (h) and (i) plot the image saliency by the proposed LG and LGA respectively. (j) presents the ground truths.

Table 2
The comparison of average computational time processing on single image in MSRA-ASD database.

Method	FT [16]	CA [20]	SS [18]	CAS [21]
Time (s)	0.13 (0.012*)	43.9	8.5	6.6
Code	Matlab (C++)	Matlab	Matlab	Matlab
Method	RC [22]	SF [23]	LG	LGA
Time (s)	0.14*	0.18 (0.15*)	4.1	2.5
Code	C++	C++	Matlab	Matlab

* The results with ∗ are reported by [23].

using the labeling consistency metric [29]:

$$G(x) = \frac{1}{g}\sum_{z=1}^{g} A_z(x), \tag{14}$$

where g is the number of users, and $A_z(x) \in \{0,1\}$ is the annotated binary label for user z at pixel x. We regard the regions with $G \geq 0.5$ as foreground, and the rest is background. Finally, we draw a rectangular window that tightly bound all the foreground regions as the ground truth.

5.4.2. Saliency map to window conversion

Since the ground truth is a rectangular region, we must convert the obtained saliency map SM to a rectangular window (denoted as W_{SM}) for the comparison purpose. Intuitively, if the saliency result is accurate, the detected window's location should be precise as well. There are four parameters (i.e. window's 2D position: horizontal coordinate, vertical coordinate, width, and height, denoted as w_x, w_y, w_w, and w_h, respectively) needed to be optimized for the obtained saliency map. To this end, we perform exhaustive search on SM to find such four parameters which reach the maximum saliency response. To reduce the computational complexity, we assume that the size of detected window is equal to the size of ground truth. In this case, the location of the detected window is only the variable required to be considered, and can be simply retrieved by applying sliding window approach on saliency map. This setting is also adopted in ICC [13]. Intuitively, if the saliency result is accurate, the detected window's location should be precise as well.

In this section, we adopt LGA as the saliency map generator, and name the LGA detection followed by the saliency map to window conversion as LGAW.

Table 3
The comparative performance in MSRA-*B* salient object dataset.

Method	Precision (%)	Recall (%)	*F*-measure (%)
ICC [13]	85.77	85.28	85.61
LGAW (proposed)	85.64	85.63	85.64
GC-LGAW (proposed)	**87.78**	**87.76**	**87.77**

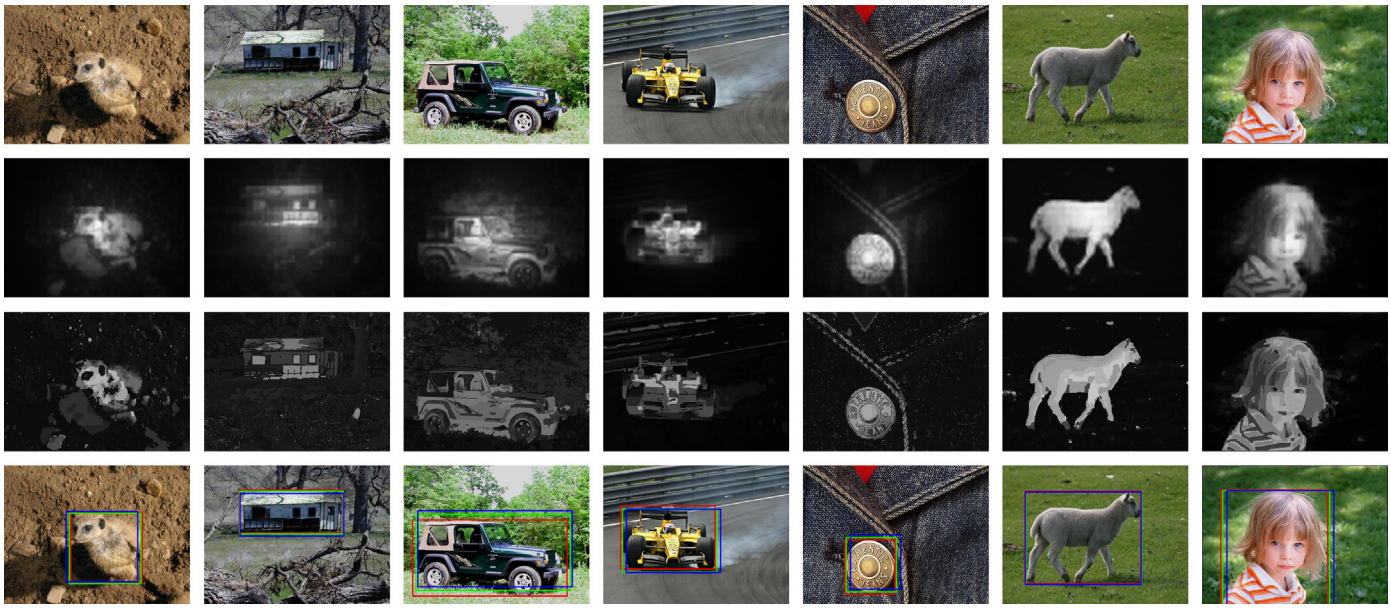

Fig. 8. The first row shows some examples from MSRA-*B* dataset. The saliency maps obtained by our method are shown in the second row. The third row shows the saliency maps obtained by the postprocessing of Graph Cut. The last row shows the detection results: ground truth (red), the LGAW (blue), and GC-LGAW (green). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

5.4.3. Graph Cut pre-processing

Since ICC [13] employs Graph Cut as post-processing to increase performance, we do the same way for a more fair comparison. We further consider another method that includes Graph Cut as the pre-processing on LGAW method, referred to as GC-LGAW, for the experiments.

The GC-LGAW is performed by the following steps: An input image is first segmented into α regions by using Graph Cut. Second, implementing LGA on input image to produce the saliency map. Once the saliency map is obtained, each segmented region is weighted by the averaged salient values inside it. Finally, exhaustive window search is performed on the weighted regions to find the best rectangle $(w_x^*, w_y^*, w_w^*, w_h^*)$. In our experiment, we use the source code[8] and set $\alpha = 30$.

5.4.4. Evaluation

Assume that the optimal rectangle window for obtained saliency map is W_{SM}^* $(=w_x^*, w_y^*, w_w^*, w_h^*)$. Let D be the subimage obtained by cropping the saliency map SM with W_{SM}^*. To evaluate the performance of detection result D respect to ground truth G, we define the precision and recall as

$$precision = \frac{\sum_x G(x)D(x)}{\sum_x D(x)}, \quad recall = \frac{\sum_x G(x)D(x)}{\sum_x G(x)}. \tag{15}$$

The weighted harmonic mean of precision and recall, referred to as *F*-measure, is also computed:

$$F\text{-}measure = \frac{(1+\beta) \times precision \times recall}{\beta \times precision + recall}, \tag{16}$$

where $\beta = 0.5$ is used.

Table 3 shows the comparative performance of ICC, LGAW, and GC-LGAW in MSRA-B dataset. It can be seen that the LGAW produces nearly identical result with ICC, despite ICC has already employed Graph Cut and others clues for enhancing its performance on salient object detection.

With the assistance of Graph Cut, our GC-LGAW considerably surpasses ICC and LGAW. It suggests that locating salient regions/objects becomes easier with the combination of Graph Cut and the saliency map generated by our proposed method, which is also demonstrated by the qualitative results shown in Fig. 8.

5.5. Discussion

There are a few parameters required to be set in the proposed method. Among those parameters, determining the proportion of foreground/background seeds is particularly worth being noted. To obtain S_{FG} and S_{BG}, we have experimented with different settings of parameters in a wide range, $\gamma_f = \{5, 10, 20\}$ and $\gamma_b = \{10, 30, 50, 70\}$, on MSRA-ASD dataset. (The reason why the proportion of foreground is selected as smaller number since the area of salient object appearing in an image is usually relatively smaller than background regions.) Fig. 9(a) shows the precision-and-recall of LGA among those parameters, and Fig. 9(b) further plots corresponding averaged

[8] http://www.wisdom.weizmann.ac.il/~bagon/matlab.html

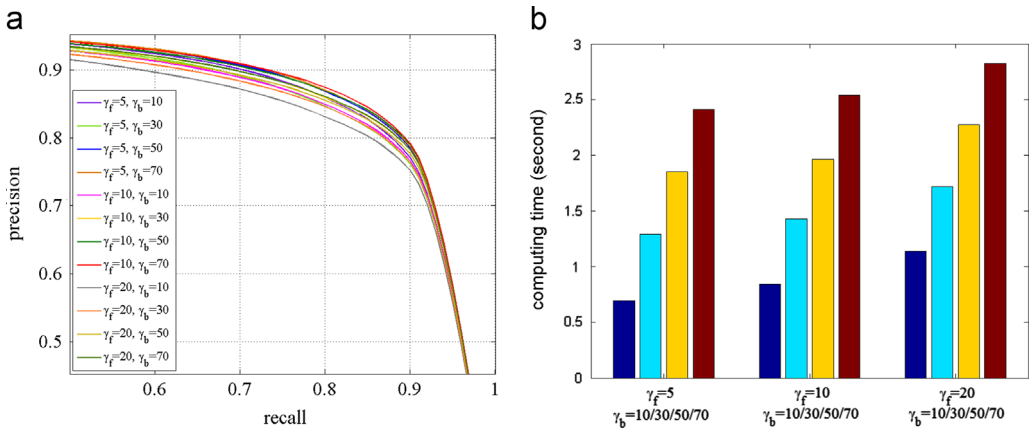

Fig. 9. The comparative performance of LGA by using different parameters γ_b and γ_f. (a) The precision and recall performance in range [0.5 1]. (b) The corresponding computing time for processing single image (in seconds).

computation time of processing single image. The following observations can help us to conclude our findings:

1. It seems that the performance is not sensitive to the population of the selected seeds. One of the possible reasons is that we take the concept in "Trimap Segmentation" to avoid selecting inappropriate seeds. The spirit of trimap is to further define an "intermediate" or fuzzy region locating in the boundary of foreground and background. (For instance, we set the patches of top 10% LS degree as foreground seeds, and those of below 70% LS degree as background seeds in LGA. So there remains around 20% intermediate patches not being occupied as effective seeds for the global-homogeneity refinement). Therefore, for the cases of retaining the intermediate patches, we believe that the selected seeds are more appropriate to represent foreground and background.
2. There is significant color discrepancy between foreground and background in MSRA images. In this case, the few amount of well-selected foreground (or background) seeds are able to represent the foreground (or background) statistics.
   Since the visual difference between foreground and background in MSRA dataset is usually significant, we thought it is another reason why the performance of proposed method is not sensitive to seed population. On the contrary, there would not have been such discrepancy in ordinary images such as mobile photos, which could be a study-worthy direction for the future works.
3. The required computation time of LGA highly depends on the population of seeds. For instance, it cost average 0.69 s with $\gamma_f = 5$ and $\gamma_b = 10$ while the required running time rapidly increases to 2.4 s with $\gamma_b = 70$. Based on our study, calculating the global-homogeneity requires most computing power because large amounts of similarity measures between image patches must be carried out. This would be a bottleneck in computational efficiency where we will address this issue as our future direction.

5.6. Summary

From the above-mentioned contents, the main features of proposed methods are the following:

1. An efficient framework is introduced. We use mixture of distributions to describe the local distinctness linked with global-homogeneity characteristics to obtain an accurate saliency map.
2. Unlike some modern works, our method does not rely on any pre-segmentation as a pre-processing step to precisely describe the discrepancy of foreground and background or object boundaries. For instance, [22,23] are considered as two most comparative works in this paper (see Fig. 6). They use Superpixel Segmentation as a pre-processing to abstract the image into perceptually uniform regions. This step does provide enormous benefits in obtaining accurate saliency boundaries, however, it also increases burdens of computation and empirical parameter settings, which would make it more difficult to fulfill saliency detection on a variety of situations. In addition, the performance of proposed method even surpasses those comparative methods.
3. Finally, our method is based on patch-based representation for balancing performance and computational efficiency. In general, providing a full or fine resolution saliency map requires a lot of time. However, it is believed that the initial objective of saliency detection is finding the salient/distinctive locations or spots in the scene. In this sense, providing a full resolution saliency map is not absolutely necessary. In fact, our proposed saliency method has provided precise salient objects' locations which can be used as priors for further post-processing stage (such as super-pixel segmentation) to generate a full resolution salient region.

6. Conclusions and future works

In this paper, we present a hybrid approach to detect salient objects. Our proposed saliency approach is realized by measuring the local saliency then followed by a global-homogeneous refinement process, which is described by a seed-propagating process and formulated by mixture distributions. The experimental results demonstrate that our method can achieve significant saliency results on well-known benchmark datasets in quantitative analysis. In future work, we will further extend to detect salient objects in videos with motion information consideration, and explore the possibility of implementing instant saliency detection on portable devices to fulfill real-time applications.

Conflict of interest statement

None declared.

Appendix A. Supplementary data

Supplementary data associated with this paper can be found in the online version at http://dx.doi.org/10.1016/j.patcog.2013.11.015.

References

[1] Q. Li, Y. Zhou, J. Yang, Saliency based image segmentation, in: ICMT, 2011, pp. 5068–5071.
[2] U. Rutishauser, D. Walther, C. Koch, P. Perona, Is bottom-up attention useful for object recognition?, in: CVPR, 2004, pp. 37–44.
[3] D. Gao, S. Han, N. Vasconcelos, Discriminant saliency the detection of suspicious, coincidences, and applications to visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 989–1005.
[4] C. Liu, P.C. Yuen, G. Qiu, Object motion detection using information theoretic spatio-temporal saliency, Pattern Recognit. 42 (2009) 2897–2906.
[5] U. Engelke, H. Kaprykowsky, H.-J. Zepernick, P. Ndjiki-Nya, Visual attention in quality assessment, IEEE Signal Process. Mag. 28 (2011) 50–59.
[6] X. Sun, H. Yao, R. Ji, S. Liu, Photo assessment based on computational visual attention model, in: ACM MM, 2009, pp. 541–544.
[7] L.-K. Wong, K.-L. Low, Saliency-enhanced image aesthetics class prediction, in: ACM MM, 2009, pp. 993–966.
[8] H.-H. Yeh, C.-Y. Yang, M.-S. Lee, C.-S. Chen, Video aesthetic quality assessment by temporal integration of photo- and motion-based features, IEEE Trans. Multimedia 15 (2013) 1944–1957.
[9] L. Itti, C. Koch, Computational modelling of visual attention, Nat. Rev. Neurosci. 2 (2001) 194–203.
[10] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 1254–1259.
[11] V. Gopalakrishnan, Y. Hu, D. Rajan, Random walks on graphs to model saliency in images, in: CVPR, 2009, pp. 1698–1705.
[12] W. Wang, Y. Wang, Q. Huang, W. Gao, Measuring visual saliency by site entropy rate, in: CVPR, 2010, pp. 2368–2375.
[13] R. Valenti, N. Sebe, T. Gevers, Image saliency by isocentric curvedness and color, in: ICCV, 2009, pp. 2185–2192.
[14] M. Wang, J. Konrad, P. Ishwar, K. Jing, H.A. Rowley, Image saliency: from intrinsic to extrinsic context, in: CVPR, 2011, pp. 417–424.
[15] N. Murray, M. Vanrell, X. Otazu, C. Parraga, Saliency estimation using a non-parametric low-level vision model, in: CVPR, 2011, pp. 433–440.
[16] R. Achanta, S. Hemami, F. Estrada, S. Süstrunk, Frequency-tuned salient region detection, in: CVPR, 2009, pp. 1597–1604.
[17] Y. Lu, W. Zhang, H. Lu, X. Xue, Salient object detection using concavity context, in: ICCV, 2011, pp. 233–240.
[18] E. Rahtu, J. Kannala, M. Salo, H.J., Segmenting salient objects from images and videos, in: ECCV, 2010, pp. 366–379.
[19] D. Klein, S. Frintrop, Center-surround divergence of feature statistics for salient object detection, in: ICCV, 2011, pp. 2214–2219.
[20] S. Goferman, L. Zelnik-Manor, A. Tal, Context-aware saliency detection, in: CVPR, 2010, pp. 2376–2383.
[21] H.-H. Yeh, C.-S. Chen, From rareness to compactness: contrast-aware image saliency detection, in: ICIP, Orlando, Florida, USA, 2012.
[22] M.-M. Cheng, G.-X. Zhang, N.J. Mitra, X. Huang, S.-M. Hu, Global contrast based salient region detection, in: CVPR, 2011.
[23] F. Perazzi, P. Krahenbuhl, Y. Pritch, A. Hornung, Saliency filters: contrast-based filtering for salient region detection, in: CVPR, 2012, pp. 733–740.
[24] L. Zhang, T.K. Marks, M.H. Tong, H. Shan, G.W. Cottrell, Sun: a Bayesian framework for saliency using natural statistics, J. Vis. 8 (2008) 1–20.
[25] J.-B. Huang, N. Ahuja, Saliency detection via divergence analysis: a unified perspective, in: ICPR, 2012.
[26] L. Duan, C. Wu, J. Miao, L. Qing, Y. Fu, Visual saliency detection by spatially weighted dissimilarity, in: CVPR, 2011, pp. 437–480.
[27] T.N. Vikram, M. Tscherepanow, B. Wrede, A saliency map based on sampling an image into random rectangular regions of interest, Pattern Recognit. 45 (2012) 3114–3124.
[28] J. Feng, Y. Wei, L. Tao, C. Zhang, J. Sun, Salient object detection by composition, in: ICCV, 2011, pp. 1028–1035.
[29] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H.Y. Shum, Learning to detect a salient object, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2010) 353–367.
[30] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, Hum. Neurobiol. 4 (1985) 219–227.
[31] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: ICCV, 2009, pp. 2106–2113.
[32] C. Rhemann, C. Rother, A. Rav-Acha, T. Sharp, High resolution matting via interactive trimap segmentation, in: CVPR, 2008, pp. 1–8.
[33] A. Borji, N. Sihite, L. Itti, Salient object detection: a benchmark, in: ECCV, 2012, pp. 414–429.
[34] J.-H. Zhai, Fuzzy decision tree based on fuzzy-rough technique, Soft Comput. 15 (2011) 1087–1096.

Hsin-Ho Yeh received the B.S. degree in Computer Science from National Chung Cheng University, in 2007 and the M.S. degree from Computer Science department of National Cheng Kung University, in 2009, respectively. He is currently a Research Assistant at the Institute of Information Science, Academia Sinica, Taipei, Taiwan (R.O.C.). His research interests include image processing, computer vision, and data mining.

Keng-Hao Liu received the B.S. degree in mathematical sciences from National Chengchi University, Taipei, Taiwan (R.O.C.), and the M.S. and Ph.D. degrees in Electrical Engineering from University of Maryland, Baltimore County, Baltimore, in 2009 and 2011, respectively. He is currently a Postdoctoral Fellow in Institute of Information Science, Academia Sinica, Taipei, Taiwan (R.O.C.). His research interests include multi/hyperspectral image processing, pattern recognition, computer vision, and machine learning.

Chu-Song Chen received a B.S. degree in Control Engineering from National Chiao-Tung University, Taiwan, in 1989. He received an M.S. degree in 1991 and a Ph.D. degree in 1996 both from the Department of Computer Science and Information Engineering, National Taiwan University. He is now a deputy director of Research Center for Information Technology Innovation (CITI), Academia Sinica, and a research fellow of Institute of Information Science (IIS), Academia Sinica, Taiwan. He is also an adjunct professor of the Graduate Institute of Networking and Multimedia, National Taiwan University. Dr. Chen's research interests include pattern recognition, computer vision, signal/image processing, and multimedia analysis.