IPSJ Transactions on Mathematical Modeling and Its Applications Vol.9 No.1 13–22 (Feb. 2016)

Regular Paper

Salient Region Detection by Enhancing Diversity of Multiple Priors

Masao Yamanaka[1,a)]

Received: September 7, 2015, Revised: October 18, 2015/October 31, 2015,
Accepted: November 16, 2015

Abstract: Detection of salient regions in images is useful for many computer vision applications. However, existing methods tend to perform poorly in complex background because lower-level priors involved in the evaluation of visual saliency are not reliable. I propose a salient region detection method by enhancing diversity of multiple priors. The method naturally integrates conspicuity maps generated with the multiple priors in order to detect salient regions in images. Extensive experiments show that the method can comfortably achieve comparable performance to the existing methods even without the help from machine learning techniques. The combination with a simple machine learning technique further improves the performance that outperforms the state-of-the-art, when evaluated using one of the largest publicly available data sets.

Keywords: visual saliency, human perception, salient region detection, super-pixel, logistic regression

1. Introduction

Detecting salient regions in images has been extensively investigated in many computer vision applications such as object detection and recognition [17], [18], [25], image editing [5], [9], [13], image segmentation [14], and adaptive compression of images [6]. Here, a salient region indicates a region in an image that visually stands out from its surroundings, as illustrated in **Fig. 1**. Key properties that make a region salient are the visual difference from the background and the uniqueness which attracts human attention.

Methods of salient region detection can be roughly divided into two categories: bottom-up approach and top-down approach. Bottom-up approaches are data-driven based on lower-level priors (e.g., *contrast prior*, *low-rank prior*, and *boundary prior*). Both local [11], [22] and global [1], [5] contrast priors have been proposed so far. For example, center-surrounding operators [11] are performed on feature maps to obtain the local maxima of visual saliency, and regional contrast features such as center-surround histograms [15] and center-surround divergence of feature statistics [4], [23] have been also introduced. Recently, Refs. [18], [24] proposed a *low-rank prior*, which decomposes an image into a low-rank matrix representing the background and a sparse noise matrix indicating the salient regions by low-rank matrix recovery, and Ref. [21] proposed a *boundary prior*, which assumes the image boundary is mostly background for the evaluation of visual saliency. Thus, bottom-up approaches make use of the visual difference from the background. However, while the salient regions are mostly unique, the inverse might not necessarily be true. Not all unique regions have the visual difference from the background, and a small region with high contrast might be recognized as meaningless noise by human. Furthermore, natural images usually exhibit cluttered backgrounds, so models that make simplified assumptions, such that the background lies in a low-dimensional feature space, might not perform well in practice. Thus, using only a lower-level prior has shortcomings.

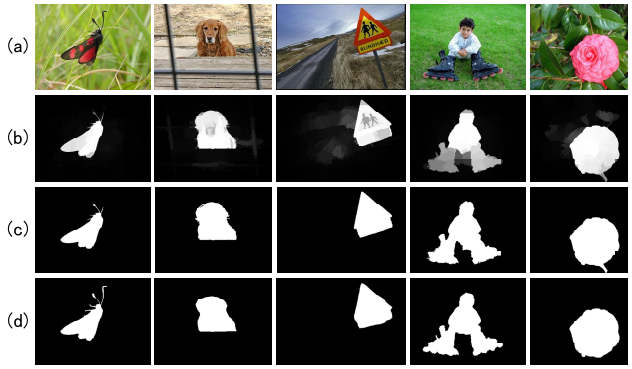

Fig. 1 Examples of the salient region extracted with proposed method on the MSRA-1000 [1]. (a) Input Images; (b) Saliency maps; (c) Salient region detection results based on (b) by adaptive thresholding [16]; (d) Ground truth.

On the other hands, top-down approaches are goal-driven based on higher-level knowledge about interesting objects to identify salient regions. A variety of top-down methods have been also proposed so far. For example, Ref. [25] learns directly features of interesting regions by dictionary learning and then generates the saliency map by modeling spatial consistency with conditional random field, and Ref. [8] proposed a top-down saliency algorithm by selecting discriminant features from a pre-defined filter bank. However, the performance of the top-down approaches depends heavily on the quality and quantity of ground truth data

[1] Department of Physics, Chuo University, Bunkyo, Tokyo 112–8551, Japan
[a)] sonicbowy@gmail.com

used for supervised learning, and gathering a large number of high-quality training data is costly. Furthermore, adding a new object category is not straightforward because human subjectivity often causes ambiguity.

In addition, unlike previous approaches that are purely bottom-up or top-down, combination approaches that integrate multiple conspicuity maps generated with different priors or features was demonstrated to be a promising alternative to the previous approaches. For example, Ref. [3] combined multi-scale saliency, color contrast, edge density, and superpixel straddling in a Bayesian framework, Ref. [15] integrated multi-scale contrast, center-surround histograms, and color spatial-distributions with conditional random field, and Ref. [13] fused several conspicuity maps by weighted averaging, where the weights learned by support vector machine. However, the combination methods tend to identify regions with various structures as salient, which is not always appropriate in practice.

In this paper, I propose a new combination method to detect salient regions in images. The combination method integrates multiple conspicuity maps generated with not only *lower-level priors* but also *higher-level priors*. I define three types of lower-level priors (i.e., *local contrast prior*, *global contrast prior*, and *boundary prior*) which are motivated by early human visual system, and four types of higher-level priors (i.e., *face prior*, *color prior*, *closedness prior*, and *center prior*) which are motivated by human perceptions. The combination method is based on the standard structure of cognitive visual attention models [11], [19], where the saliency computation consists of following four steps.

First, an input image is segmented into small image patches called *super-pixels* by using SLIC [2]. The super-pixels are less likely to cross object boundaries, which lead to more accurately segmented salient regions. In practice, because it may be difficult to determine the appropriate super-pixel size in advance, I generate four types of segmented images, each consisting of about 25, 50, 75, 100 super-pixels. Then, in the second step, the segmented images are converted into *feature maps* in which each super-pixel is assigned a conspicuity strength. Here, the procedures to compute the conspicuity strength for each prior are described in following sections. Furthermore, in the third step, the feature maps which are generated with every segmented images in each prior are fused into a single conspicuity map by linear combination. Here, coefficients of the linear combination are determined by the inverse of the number of feature maps in each prior. Finally, in the fourth step, the multiple conspicuity maps which are generated with both of the lower-level and higher-level priors are integrated into a single *saliency map* by a simple machine learning technique (i.e., *logistic regression*). The above formulation is summarized in **Fig. 2**, which illustrates the schematic overview of my salient region detection system.

Through extensive experiments, I demonstrate that the proposed method can comfortably achieve comparable performance to the existing methods even without the help from the machine learning technique, and the combination with the machine learning technique further improves the performance that outperforms the state-of-the-art, when evaluated using one of the largest publicly available data sets [1].

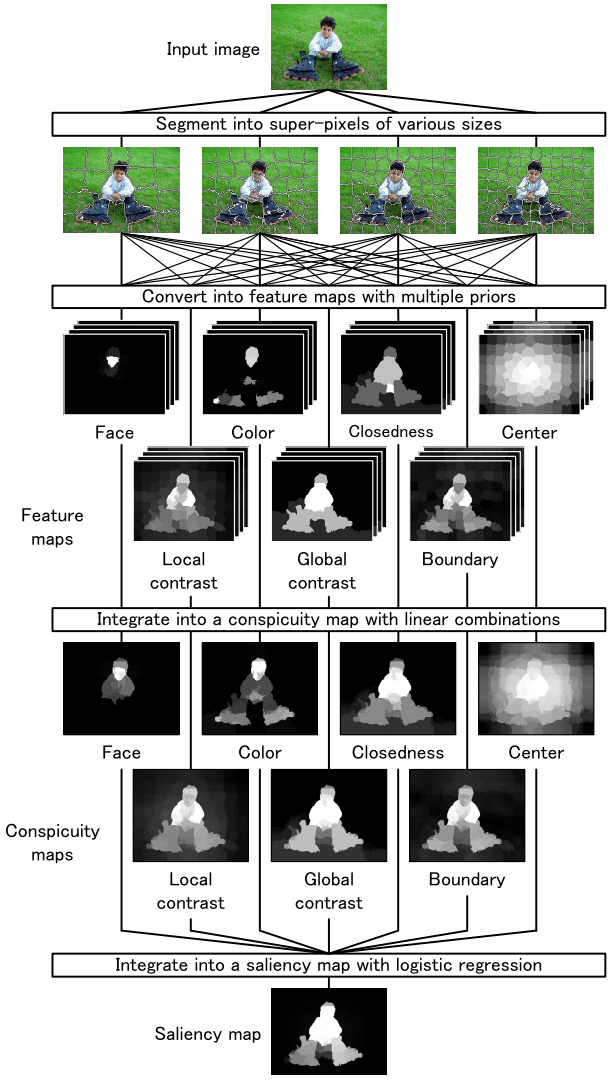

Fig. 2 Schematic overview of my salient region detection system.

2. Multiple Priors

In this section, I introduce how the conspicuity strength for each prior is evaluated from the image segmented into super-pixels.

2.1 Local-contrast Prior (LC)

Reference [9] proposed a local conspicuity measure based on the differences of the color and position between small image elements. Based on the local conspicuity measure, I define a *local-contrast prior* at the i-th super-pixel sp_i in the image segmented into N super-pixels as:

$$LC(sp_i) = \sum_{j=1}^{N} \frac{n(sp_j) \cdot d_{\text{color}}(sp_i, sp_j)}{1 + c \cdot d_{\text{position}}(sp_i, sp_j)},$$

where $n(sp_j)$ is the number of pixels in the j-th super-pixel sp_j, and $d_{\text{color}}(sp_i, sp_j)$ is the Euclidean distance between the color features of super-pixels sp_i and sp_j which are vectorized with the arithmetic mean pixel value in HSV color space, and $d_{\text{position}}(sp_i, sp_j)$ is the Euclidean distance between the gravity positions of super-pixels sp_i and sp_j, and c is the median of the Euclidean distances between the gravity positions of every pair of

super-pixels.

The local-contrast prior is proportional to the difference in visual appearance and *inverse* proportional to the positional distance between super-pixels. Thus, the i-th super-pixel sp_i is considered as visually salient when it is highly dissimilar to the surrounding super-pixels.

2.2 Global-contrast Prior (GC)

Reference [12] proposed a global conspicuity measure based on the differences of the color and size between segmented region types, which favors region types with a unique visual appearance within the entire image. By simplifying the global conspicuity measure, I define a *global-contrast prior* at the i-th region type r_i in the image segmented into M region types as:

$$GC(r_i) = \sum_{j=1}^{M} n(r_j) \cdot w(r_i) \cdot d_{\text{color}}(r_i, r_j), \tag{1}$$

where $n(r_j)$ is the number of pixels in the j-th region type r_j, and $w(r_i)$ is the *weight* for the i-th region type r_i, and $d_{\text{color}}(r_i, r_j)$ is the Euclidean distance between the color features of region types r_i and r_j which are vectorized with the arithmetic mean pixel value in HSV color space. The remaining questions in Eq. (1) are how to segment into M region types r_i ($i \in [1, M]$) and compute the weight $w(r_i)$ for the i-th region type r_i, which are described in detail below.

My segmentation technique consists of following four steps. First, an input image is split into CIE L^*a^*b color channels. Then, in the second step, the channel images are converted into average maps in which each super-pixel is assigned the arithmetic mean pixel value. Furthermore, in the third step, the average maps are binarized by adaptive thresholding [16]. Finally, the input image is segmented into several region types with the binary patterns. Here, the binary patterns consist of the eight types that $(0, 0, 0)$, $(0, 0, 1)$, $(0, 1, 0)$, $(0, 1, 1)$, $(1, 0, 0)$, $(1, 0, 1)$, $(1, 1, 0)$, and $(1, 1, 1)$. Therefore, the number of segmented region types is eight at most. The above formulation is summarized in **Fig. 3**, which illustrates the schematic overview of my segmentation technique. In addition, examples of the image segmented with my segmentation technique are shown in **Fig. 4**. You can see that the images are roughly divided into several region types while maintaining the global structural features.

Furthermore, I define the weight $w(r_i)$ for the i-th region type r_i as:

$$w(r_i) = 1 - \frac{|r_i \cap B|}{|B|}, \tag{2}$$

where B is the *boundary region* on the segmented image as shown in **Fig. 5**. In my experiments, I define the boundary size L as:

$$L = 0.05 \cdot \min(W, H),$$

where W and H are the horizontal and vertical image size, respectively. The weight $w(r_i)$ is *inverse* proportional to the overlap rate between the i-th region type r_i and the boundary region B. Therefore, when the overlap rate is large, $GC(sp_i)$ is rather suppressed as a whole even if $d_{\text{color}}(r_i, r_j)$ is large. On the other hand, when the overlap rate is small, $GC(sp_i)$ is rather enhanced as a whole even if $d_{\text{color}}(r_i, r_j)$ is small.

Thus, the i-th region type r_i is considered as visually salient when it is highly dissimilar to the other regions and not almost

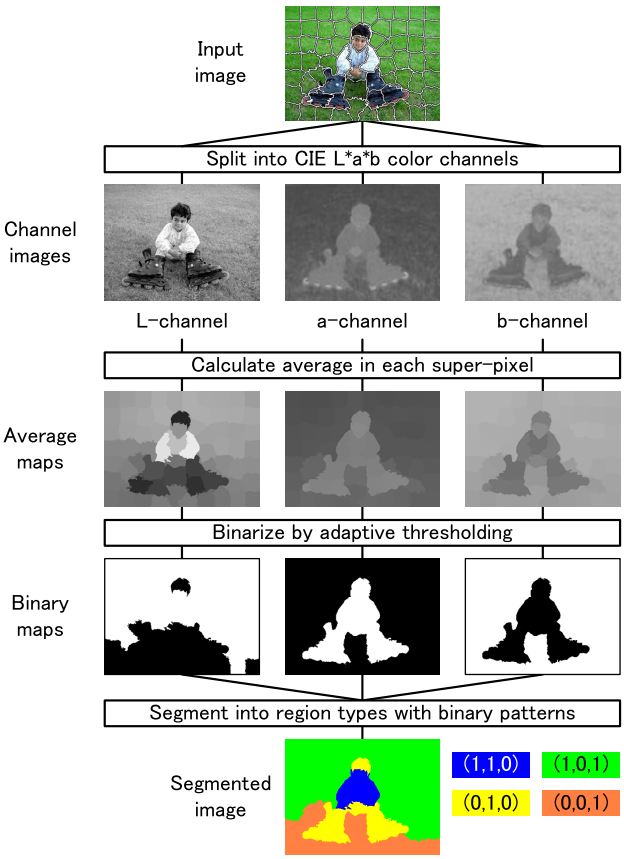

Fig. 3 Schematic overview of my segmentation technique.

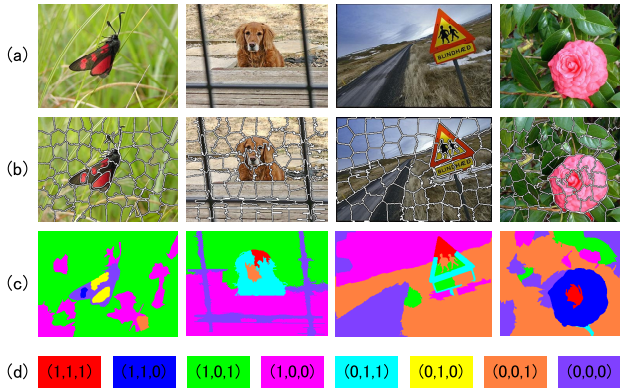

Fig. 4 Examples of the image segmented with my segmentation technique. (a) Input images; (b) Images segmented into super-pixels; (c) Images segmented into region types; (d) Colors of region type and their binary patterns.

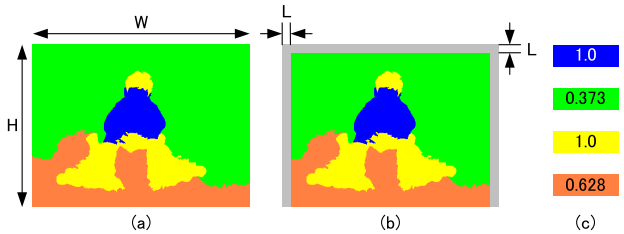

Fig. 5 Definition of boundary region on the segmented image. (a) Segmented region types (blue, green, yellow, and orange); (b) Boundary region (gray); (c) Colors of region type and their weights with Eq. (2) on the segmented image.

overlap with the boundary region B on the segmented image.

2.3 Boundary Prior (BD)

Reference [21] proposed a boundary conspicuity measure due to the assumption that the image boundary is mostly background. Based on the boundary conspicuity measure, I define a *boundary prior* at the i-th super-pixel sp_i in the image segmented into N super-pixels as:

$$BD(sp_i) = \sum_{j=1}^{\mathrm{M}} W(r_j) \cdot d_{\mathrm{color}}(sp_i, r_j \cap B),$$

where r_j and B are the j-th region generated with my segmentation technique and the boundary region on the segmented image as described above, and $d_{\mathrm{color}}(sp_i, r_j \cap B)$ is the Euclidean distance between the color features of regions sp_i and $r_j \cap B$ which are vectorized with the arithmetic mean pixel value in HSV color space, and $W(r_j)$ is the *weight* for the j-th region r_j defined by

$$W(r_j) = \frac{|r_j \cap B|}{|B|}.$$

The weight $W(r_j)$ is the overlap rate between j-th region r_j and the boundary region B. Therefore, when the overlap rate is small, $BD(sp_i)$ is rather suppressed as a whole even if $d_{\mathrm{color}}(sp_i, r_j \cap B)$ is large. Similarly, when the overlap rate is large, $BD(sp_i)$ is rather enhanced as a whole even if $d_{\mathrm{color}}(sp_i, r_j \cap B)$ is small.

Thus, the i-th super-pixel sp_i is considered as visually salient when it is highly dissimilar to the regions which extensively overlap with the boundary region B on the segmented image.

2.4 Face Prior (FC)

People pay more attention to certain semantic objects such as faces even without specific purposes. Therefore, I perform face detection on the images as is the case with Refs. [9], [12], [13]. In my implementation, I incorporated the face detection algorithm [20] which is commonly used in computer vision community. More specifically, when the face of size $W_{\mathrm{f}} \times H_{\mathrm{f}}$ is detected at the position $(X_{\mathrm{f}}, Y_{\mathrm{f}})$ in an image, I define a *face prior* at the i-th super-pixel in the image segmented into N super-pixels as:

$$FC(sp_i) = \sum_{j=1}^{n(sp_i)} \exp\left(-\frac{(X_{\mathrm{f}} - x_j)^2}{2\sigma_{\mathrm{fx}}^2} - \frac{(Y_{\mathrm{f}} - y_j)^2}{2\sigma_{\mathrm{fy}}^2}\right), \tag{3}$$

where $n(sp_i)$ is the number of pixels in the i-th super-pixel sp_i, X_f and Y_f are respectively the horizontal and vertical center position of detected face in the image, x_j and y_j are respectively the j-th pixel positions of horizontal and vertical direction in the i-th super-pixel sp_i, σ_{fx} and σ_{fy} are respectively the Gaussian sizes of horizontal and vertical direction defined by

$$\sigma_{\mathrm{fx}} = \frac{W_{\mathrm{f}}}{4.0},\ \sigma_{\mathrm{fy}} = \frac{H_{\mathrm{f}}}{4.0}.$$

Example of the feature map generated with the face prior is shown in **Fig. 6**. You can see that the super-pixels near the detected face are assigned higher conspicuity strength according to Eq. (3). Thus, the face prior makes it possible to generate the feature map while maintaining the face contour features as shown in Fig. 6.

Fig. 6 Example of the feature map generated with the face prior in Eq. (3). (a) Face detection result; (b) Image segmented into super-pixels; (c) Feature map.

Fig. 7 Example of the feature map generated with the color prior in Eq. (4). (a) Input image; (b) Image segmented into super-pixels; (c) Feature map.

2.5 Color Prior (CL)

From our daily experience, we find that warm color regions such as red and yellow are more pronounced as pointed out in Refs. [9], [13], [18]. To use such information, I define a *color prior* at the i-th super-pixel sp_i in the image segmented into N super-pixels as:

$$CL(sp_i) = \sum_{j=1}^{n(sp_i)} \frac{\max\left(R_j^{'} - G_j^{'}, 0\right)}{n(sp_i)}, \tag{4}$$

where $n(sp_i)$ is the number of pixels in the i-th super-pixel sp_i, $R_j^{'}$ and $G_j^{'}$ are respectively the j-th normalized color elements of R and G in the i-th super-pixel sp_i defined by

$$R_j^{'} = \frac{R_j}{R_j + G_j + B_j},\ G_j^{'} = \frac{G_j}{R_j + G_j + B_j}.$$

Here, R_j, G_j, and B_j are respectively the j-th color elements of R, G, and B in the i-th super-pixel sp_i.

Examples of the feature maps generated with the color prior are shown in **Fig. 7**. You can see that the warm color regions (i.e., *triangle traffic sign*) are popped out while maintaining the structural features. Thus, the color prior favors the regions which are more attractive to humans and eliminates the background color such as road, sky, and cloud.

2.6 Closedness Prior (CD)

A region which has a wider spatial distribution is typically less salient than regions which have small spatial spread as pointed out in Refs. [10], [15]. On the basis of the knowledge, I define a *closedness prior* at the i-th region r_i in the image segmented into M regions as:

$$CD(r_i) = \left(1 - \frac{D(r_i)}{\max_{i \in \mathrm{M}} D(r_i)}\right) w(r_i), \tag{5}$$

where r_i is the i-th region segmented with my segmentation technique as described above, and $w(r_i)$ is the weight for the i-th region r_i defined by Eq. (2), and $D(r_i)$ is the average of the Euclidean distances between the gravity positions of every pair of super-pixels which are labeled as i-th region r_i with my segmentation technique.

Fig. 8 Example of the feature map generated with the closedness prior in Eq. (5). (a) Input image; (b) Image segmented into regions; (d) Feature map.

Fig. 9 Example of the feature map generated with the center prior in Eq. (6). (a) Input image; (b) Image segmented into super-pixels; (d) Feature map.

Example of the feature map generated with the closedness prior is shown in **Fig. 8**. You can see that the closed regions (i.e., *dog*) are popped out while maintaining the contour features. Thus, the closedness prior favors the regions with small spatial variance and eliminates the background color of large variance.

2.7 Center Prior (CT)

People taking photographs generally frame an interest object near the image center. Therefore, similarly as in Ref. [13], I generate a center prior using a Gaussian distribution based on the distances between the gravities of super-pixels and the image center. More specifically, I define a *center prior* at the i-th super-pixel sp_i in the image segmented into N regions as:

$$CT(sp_i) = \sum_{j=1}^{n(sp_i)} \exp\left(-\frac{(X_g - x_j)^2}{2\sigma_{gx}^2} - \frac{(Y_g - y_j)^2}{2\sigma_{gy}^2}\right), \qquad (6)$$

where $n(sp_i)$ is the number of pixels in the i-th super-pixel sp_i, x_j and y_j are respectively the j-th pixel positions of horizontal and vertical direction in the i-th super-pixel sp_i, X_g and Y_g are respectively the horizontal and vertical center position of the image, σ_{gx} and σ_{gy} are respectively the Gaussian sizes of horizontal and vertical direction defined by

$$\sigma_{gx} = \frac{W}{2.0},\ \sigma_{gy} = \frac{H}{2.0}.$$

Here, W and H are the horizontal and vertical image size, respectively.

Example of the feature map generated with the center prior is shown in **Fig. 9**. You can see that the super-pixels near the image center are assigned higher conspicuity strength according to Eq. (6). Thus, the center prior favors the regions near the image center and eliminates the regions near the image corner.

Overall, it is possible to generate the seven types of feature maps with the proposed lower-level and higher-level priors. Then, the feature maps are normalized into $[0, 255]$ range as shown in **Fig. 10**. Furthermore the feature maps are integrated into a single conspicuity map by linear combinations. Finally, the conspicuity maps are also normalized into $[0, 255]$ range as shown in **Fig. 11**.

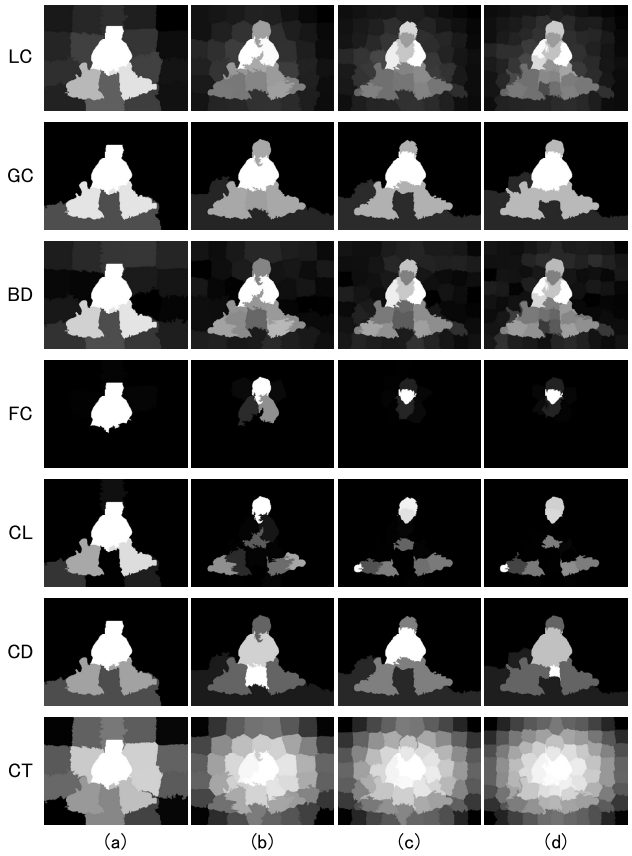

Fig. 10 Feature maps generated with the proposed lower-level and higher-level priors, each consisting of about (a) 25, (b) 50, (c) 75, (d) 100 super-pixels.

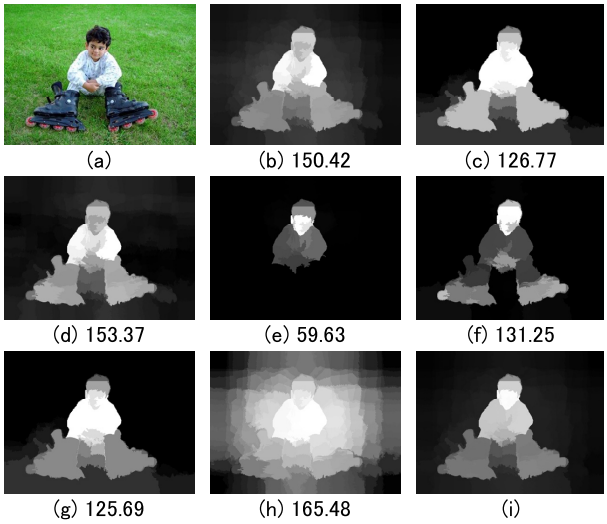

Fig. 11 Conspicuity maps generated with linear combinations between feature maps in Fig. 10. (a) Input image; (b) LC; (c) GC; (d) BD; (e) FC; (f) CL; (g) CD; (h) CT; (i) Saliency map integrated with linear combination. The number below each conspicuity map is the weight value β_k in Eq. (7).

3. Saliency Map Construction

In this section, I introduce how the conspicuity maps are integrated into a single saliency map. As the details are shown in below, there are two different approaches to construct the saliency map.

3.1 Linear Combination Based Approach

First, I introduce a linear combination based approach to construct a saliency map. Specifically, I define a saliency strength at the position (x, y) in the saliency map as:

$$ss(x, y) = \sum_{k=1}^{K} \frac{cs_k(x, y)}{\beta_k}, \tag{7}$$

where K is the number of conspicuity maps (i.e,. $K = 7$), $cs_k(x, y)$ is the conspicuity strength at the position (x, y) in the k-th conspicuity map, and β_k is the sum of weighted standard deviations for the horizontal and vertical direction in the k-th conspicuity map defined by

$$\beta_k = \sqrt{\frac{\sum_{x=1}^{W}\sum_{y=1}^{H}\left((X_k - x)^2 + (Y_k - y)^2\right) \cdot cs_k(x, y)}{\sum_{x=1}^{W}\sum_{y=1}^{H} cs_k(x, y)}}. \tag{8}$$

Here, W and H are respectively the horizontal and vertical image size. Furthermore, X_k and Y_k are respectively the weighted averages for the horizontal and vertical direction in the k-th conspicuity map defined by

$$\begin{cases} X_k = \dfrac{\sum_{x=1}^{W}\sum_{y=1}^{H} x \cdot cs_k(x, y)}{\sum_{x=1}^{W}\sum_{y=1}^{H} cs_k(x, y)}, \\ Y_k = \dfrac{\sum_{x=1}^{W}\sum_{y=1}^{H} y \cdot cs_k(x, y)}{\sum_{x=1}^{W}\sum_{y=1}^{H} cs_k(x, y)}. \end{cases}$$

An example of the saliency map generated with linear combination based approach is shown in Fig. 11. Here, The number below each conspicuity map is the weight value β_k in Eq. (8). You can see that the conspicuity maps with large spacial dispersion (e.g., (h) CT and (d) BD) are assigned higher weight value β_k. Therefore, the conspicuity maps with large spacial dispersion are rather suppressed as a whole even if the conspicuity strengths are large. Finally, the saliency map is normalized into $[0, 255]$ range as shown in Fig. 11, which is for the visualization and the evaluation with ground truth [1].

3.2 Machine Learning Based Approach

In the second approach, a popular machine learning technique called *logistic regression* is adopted to integrate multiple conspicuity maps into a single saliency map. Specifically, the model of the logistic regression is that

$$\log \frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})} = \boldsymbol{\alpha} \cdot \boldsymbol{x} + \beta,$$

where $\boldsymbol{x}$ is a seven dimensional feature vector in which elements are consist of conspicuity strengths that are computed with multiple priors (i.e., LC, BC, BD, FC, CL, CD, and CT), $p(\boldsymbol{x})$ is the probability density function in the particular category that the feature vector $\boldsymbol{x}$ is extracted from salient region in an image, β and $\boldsymbol{\alpha}$ are the constant of the model and the coefficient of the predictor variables, respectively.

When β and $\boldsymbol{\alpha}$ are determined by learning performed using positive samples (i.e., approximately 2×10^5 samples randomly extracted from salient regions) and negative samples (i.e., approximately 9×10^5 samples randomly extracted from its background)[*1], the probability density for the feature vector $\boldsymbol{x}'$ extracted from a test image is estimated by

[*1] The details of the learning method leave it out in this paper.

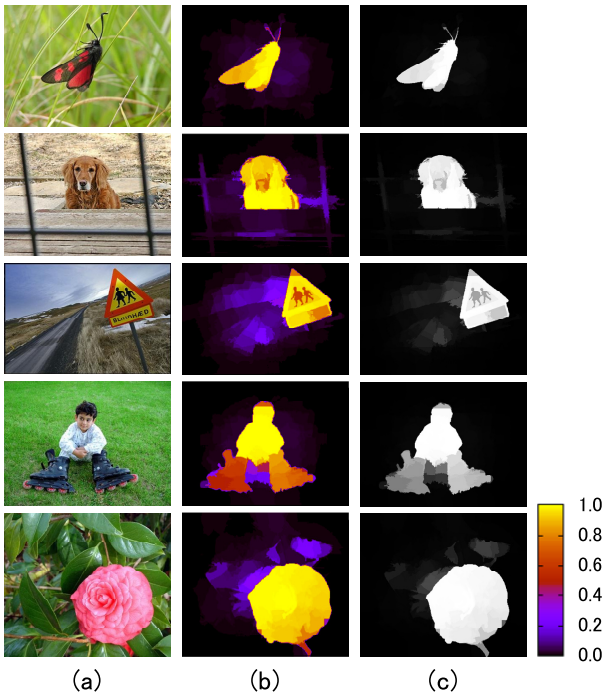

Fig. 12 Examples of the probability density estimated with Eq. (9). (a) Input images; (b) Estimated probability densities; (c) Normalized images into $[0, 255]$ range based on (b).

$$\hat{p}(\boldsymbol{x}') = \frac{1}{1 + \exp\left(-\left(\hat{\boldsymbol{\alpha}} \cdot \boldsymbol{x}' + \hat{\beta}\right)\right)}, \tag{9}$$

where $\hat{\boldsymbol{\alpha}}$ and $\hat{\beta}$ are the model parameters obtained through learning. Examples of the probability density estimated with Eq. (9) are shown in **Fig. 12** (b). Finally, the saliency map is also normalized into $[0, 255]$ range as shown in Fig. 12 (c).

4. Experiments

I evaluate my salient region detection method quantitatively on the publicly available data set [1]. Instead of using a bounding box for the salient region, accurate human-marked labels are provided as ground truth in this 1000-image data set. Here, a test image is excluded when learning the model parameters (i.e., $\hat{\boldsymbol{\alpha}}$ and $\hat{\beta}$) in Eq. (9), which are determined on other images from the MSRA data set.

I follow a general methodology [1] to evaluate the accuracy of the detected salient region. In the evaluation, the image is segmented with a fixed threshold according to the saliency values. Given a threshold $T \in [0, 255]$, the regions whose saliency values are higher than T are marked as foreground (i.e., *salient region*). Then, the segmented image is compared with the ground truth mask to obtain the precision and recall. When the threshold T varies from 0 to 255, different precision-recall pairs are obtained and a precision-recall curve can be drawn. The average precision-recall curve is generated by combing the results from all the 1000 test images.

First, I compare the performance of my proposed method with three different approaches: linear combination based approach, machine learning based approach, and the results from each prior (i.e., LC, BC, BD, FC, CL, CD, and CT). The average precision-recall curves are shown in **Fig. 13**. By integrating

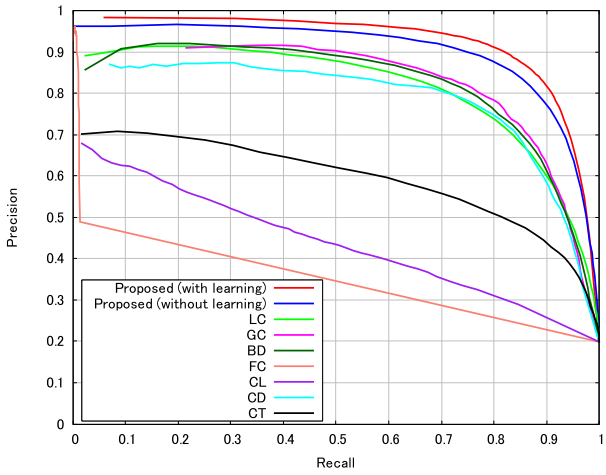

Fig. 13 Average precision-recall curves on the 1000-image data set [1]. By combining the multiple priors, the salient region detection performance is significantly improved.

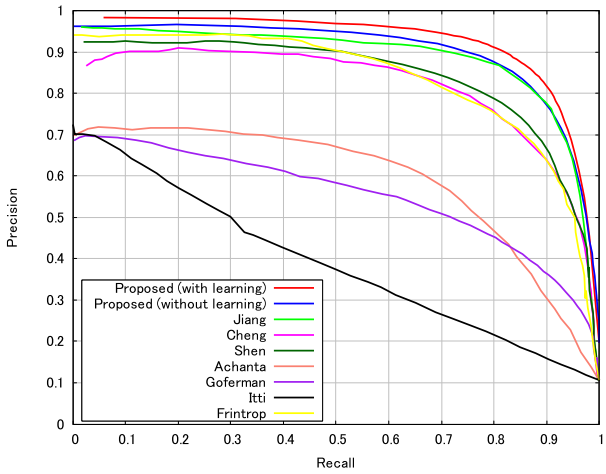

Fig. 14 Average precision-recall curves on the 1000-image data set [1]. Compared with seven state-of-the-arts (i.e., Jiang [12], Cheng [5], Shen [18], Achanta [1], Goferman [9], Itti [11], and Frintrop [7]), my proposed linear combination based approach nearly equivalent to Jiang [12], and machine learning based approach achieved the best performance.

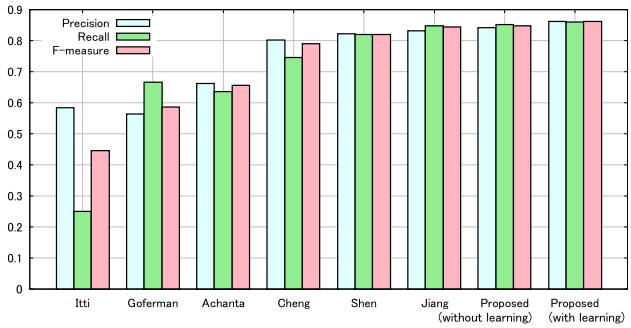

Fig. 15 Average precision, recall and F-measure using different approaches with adaptive thresholding. my proposed method achieves the best precision, recall and F-measure.

the multiple priors, the salient region detection performance is significantly improved.

In the second, I compare the average precision-recall curves obtained with my proposed methods with Cheng [5], Achanta [1], Goferman [9], Itti [11], and two recently proposed methods Jiang [12], Shen [18] and Frintrop [7]. My proposed linear combination based approach is comparable to Jiang [12] and outperforms the other methods as shwon in **Fig. 14**. Furthermore, my proposed machine learning based approach achieves the highest average precision and recall among all methods. The average precision, recall and F-Measure using different approaches with adaptive thresholding are shown in **Fig. 15**. Among all approaches, my proposed method achieves highest precision, recall and F-Measure values[*2]. **Figure 16** shows some examples of

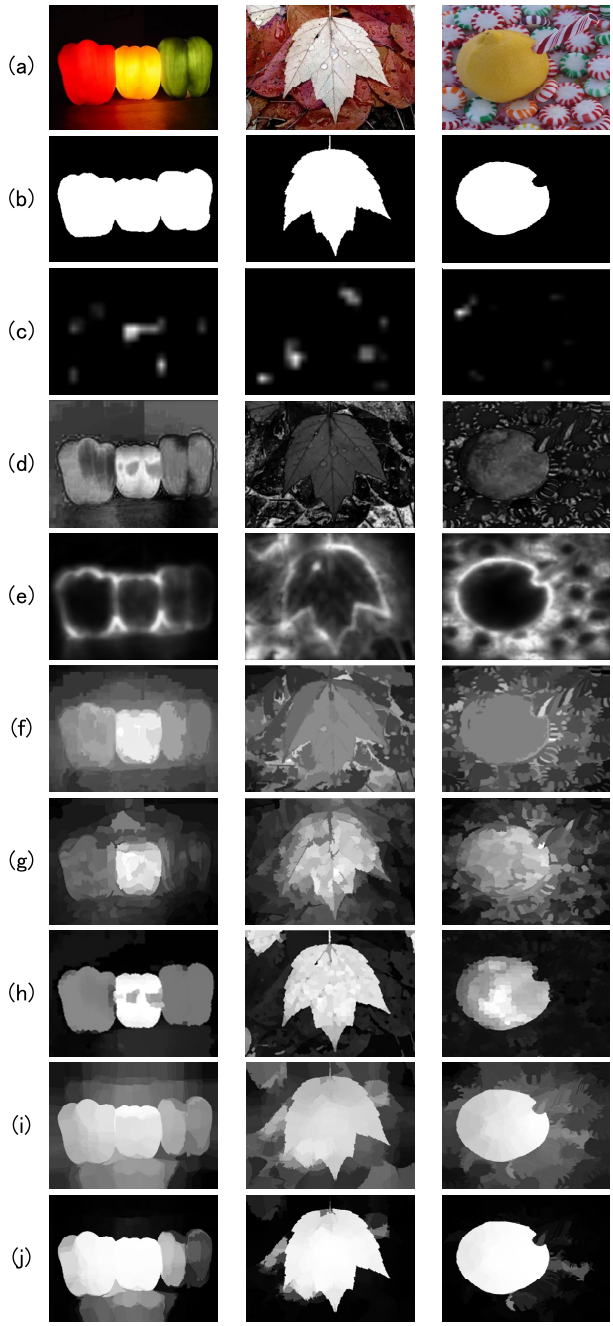

Fig. 16 Examples of saliency map construction using different methods on the MSRA-1000 database [1]. (a) Input images; (b) Ground truth; (c) Itti [11]; (d) Achanta [1]; (e) Goferman [9]; (f) Cheng [5]; (g) Shen [18]; (h) Jiang [12]; (i) My proposed linear combination based approach; (j) My proposed machine learning based approach.

[*2] Here, I did not apply any special post-processing techniques (e.g., *SaliencyCut*) in order to make a fair comparison. Therefore, the f-measure in this paper differ from that of published in Cheng [5]. Actually, an appropriate post-processing technique is applied to my proposed method so that the performance can be significantly improved.

saliency map construction result using my proposed method and Itti [11], Achanta [1], Goferman [9], Cheng [5], Shen [18], and Jiang [12].

The processing time necessary for making a saliency map, composed of 400 by 300 [pixels] images, was about 7 [sec] on an Intel 2.53 [GHz] machine with 4.0 [GB] RAM memory. This is approximately equal to the processing time of the Jiang [12], but about 10 times of the processing time of the Itti [11]. However, this weakness about computational time can be also overcome by applying parallel computation and coarse-to-fine strategy.

5. Discussion

In this section, I discuss which priors worked better in what types of images and why they worked better. I also refer to failure cases to specify the limitations of my proposed method.

Examples of the conspicuity map generated with the LC (CD) are shown in **Fig. 17**. You can see LC (CD) success and failure case in Fig. 17. Here, the salient region in image (a) locally stands out from the background and is surrounded by another region which is composed of the color (i.e., *black*) different from that (i.e., *silver*) of the salient region. In such a case, it is possible to accurately detect the salient region in an image with LC. On the other hand, because the salient region in image (b) consists of several conspicuous parts (i.e., *orange and yellow regions*), it is difficult to pop out simultaneously both of them with LC. However, CD can easily pop out whole salient region in image (b), although CD cannot detect accurately the salient region in image (a). Thus, LC and CD are in a complementary relationship mutually in these images.

Examples of the conspicuity map generated with the GC (LC) are shown in **Fig. 18**. You can see GC (LC) success and failure case in Fig. 18. Here, salient region in image (a) is globally different from the background and is composed of monotonous color (i.e., *black shadow*). In such a case, it is possible to accurately detect the salient region in an image with GC. On the other hand, because salient region in image (b) is similar to a part of the background (i.e., *mountain*), it is impossible to pop out not the background but only the salient region with GC. However, LC can easily pop out only the salient region in image (b), although LC cannot detect accurately the salient region in image (a). Thus, GC and LC are in a complementary relationship mutually in these images.

Examples of the conspicuity map generated with the BD (CL) are shown in **Fig. 19**. You can see BD (CL) success and failure case in Fig. 19. Here, the salient region in image (a) is significantly different from the boundary region (i.e., *black boundary*). In such a case, it is possible to accurately detect the salient region in an image with BD. On the other hand, because not only the salient region (i.e., *red strawberry*) in image (b) but also its surrounding region (i.e., *pink food wash dishpan*) stands out from the boundary region, it is difficult to pop out only the salient region with BD. However, CL can easily pop out only the salient region in image (b), although CL cannot detect accurately the salient region in image (a). Thus, BD and CL are in a complementary relationship mutually in these images.

Examples of the conspicuity map generated with the FC (CT) are shown in **Fig. 20**. You can see FC (CT) success and failure case in Fig. 20. Here, the face detection is succeeded in image (a), while the face detection is failed in image (b). Naturally enough, if the face detection is failed in an image, FC cannot work effectively. However, CT can roughly detect the salient region (i.e., *baby face*) in image (b), although CT cannot work well in image (a). Thus, FC and CT are in a complementary relationship mutually in these images.

Examples of the conspicuity map generated with the CL (BD)

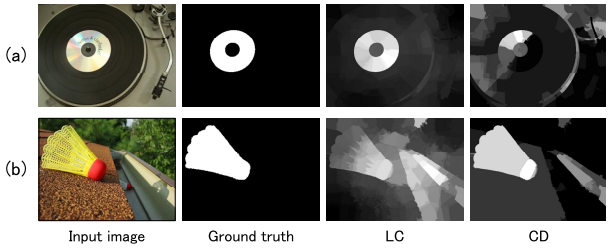

Fig. 17 Examples of conspicuity map generated with LC (CD). LC success (CD failure): salient region in image (a) locally stands out from the background. LC failure (CD success): salient region in image (b) consists of several conspicuous parts (i.e., *orange and yellow regions*).

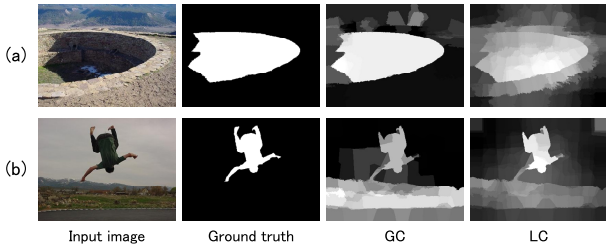

Fig. 18 Examples of conspicuity map generated with GC (LC). GC success (LC failure): salient region in image (a) is globally different from the background. GC failure (LC success): salient region in image (b) is similar to a part of the background (i.e., *mountains*).

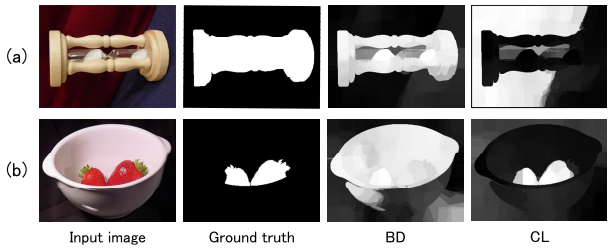

Fig. 19 Examples of conspicuity map generated with BD (CL). BD success (CL failure): salient region in image (a) is different from the boundary region. BD failure (CL success): not only salient region in image (b) but also its surrounding region is different from the boundary region.

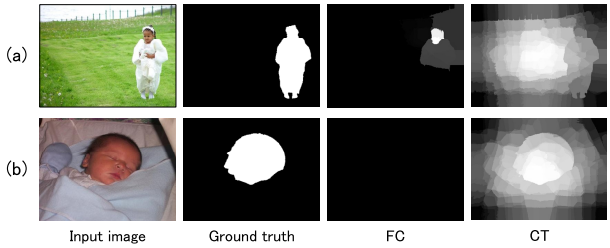

Fig. 20 Examples of conspicuity map generated with FC (CT). FC success (CT failure): image (a) which is succeeded in detecting the face region. FC failure (CT success): image (b) which is failed in detecting the face region.

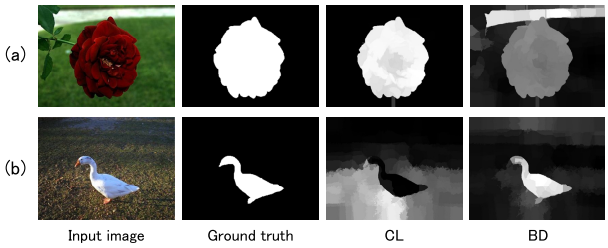

Fig. 21 Examples of conspicuity map generated with CL (BD). CL success (BD failure): salient region in image (a) is mainly composed of warm color. CL failure (BD success): not salient region but background region in image (b) is rather composed with warm color (i.e., ground lighted by the sunlight).

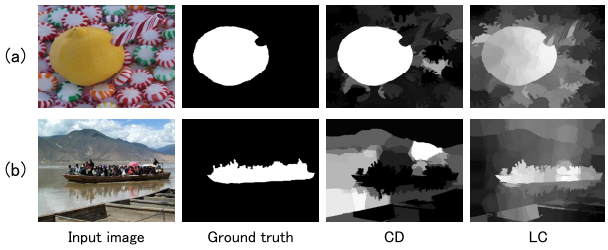

Fig. 22 Examples of conspicuity map generated with CD (LC). CD success (LC failure): salient region in image (a) has a small spatial distribution with monotonous color. CD failure (LC success): salient region in image (b) has a wide spatial distribution with several colors.

are shown in **Fig. 21**. You can see CL (BD) success and failure case in Fig. 21. Here, the salient region in image (a) is mainly composed of warm color (i.e., *red rose*). In such a case, it is possible to accurately detect the salient region in an image with CL. On the other hand, because not the salient region but the background in image (b) is rather composed of warm color (i.e., *ground lighted by the sunlight*), it is impossible to pop out the salient region with CL. However, BD can detect accurately the salient region in image (b), although BD cannot work well in image (a). Thus, CL and BD are in a complementary relationship mutually in these images.

Examples of the conspicuity map generated with the CD (LC) are shown in **Fig. 22**. You can see CD (LC) success and failure case in Fig. 22. Here, the salient region in image (a) has a small spatial distribution with monotonous color (i.e., *yellow*). In such a case, it is possible to accurately detect the salient region in an image with CD. On the other hand, because the salient region in image (b) has a wide spatial distribution with cluttered colors (i.e., *gray, white, and pink*), it is difficult to pop out the salient region with CD. However, LC can detect accurately the salient region in image (b), although LC cannot perform well in image (a). Thus, CD and LC are in a complementary relationship mutually in these images.

Examples of the conspicuity map generated with the CT (GC) are shown in **Fig. 23**. You can see CT (GC) success and failure case in Fig. 23. Here, the salient region in image (a) is near the image center, while the salient region in image (b) is near the image corner. Naturally enough, if the salient region in an image is not near the center, CT cannot work effectively. However, GC can detect accurately the salient region in image (b), although GC cannot perform well in image (a). Thus, CT and GC are in a complementary relationship mutually in these images.

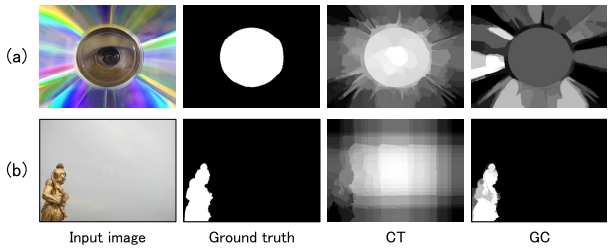

Fig. 23 Examples of conspicuity map generated with CT (GC). CT success (GC failure): Salient region in image (a) is near the image center. CT failure (GC success): Salient region in image (b) is near the image corner.

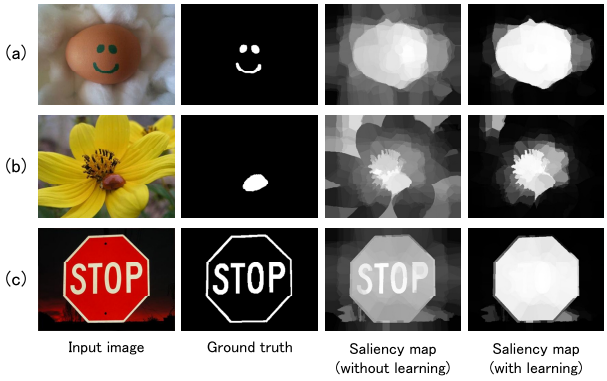

Fig. 24 Examples of failure image. (a) *smiley* written on the egg; (b) *insect* on the flower; (c) *character string* on the sign.

Finally, examples of failure image are shown in **Fig. 24**. You can see representative failure images in Fig. 24. Here, while my proposed method intuitively seems to work well, the ground truth in image (a) is not the egg itself but the *smiley* written on the egg, and the ground truth in image (b) is actually not the flower itself but the *insect* on the flower. Furthermore, the ground truth in image (c) is not the STOP sign itself but rather the character string (i.e., *S, T, O, and P*) on the sign. To overcome this problem based on the uncertainty by human subjectivity, it will be necessary to define more advanced higher-level priors.

6. Conclusion

In this paper, I presented a salient region detection method by enhancing diversity of multiple priors. In my proposed method, the three types of lower-level priors (i.e., *local contrast prior*, *global contrast prior*, and *boundary prior*) which are motivated by early human visual system, and the four types of higher-level priors (i.e., *face prior*, *color prior*, *closedness prior*, and *center prior*) which are motivated by human perceptions are defined. By integrating the higher-level and lower-level priors, the salient region detection results are significantly improved and more consistent with human intelligent vision system. Experimental results indicate that the algorithm outperforms existing salient region detection methods including Jiang [12], Cheng [5], Shen [18], Achanta [1], Goferman [9], Itti [11], Jiang [12] and Frintrop [7]. Furthermore, my proposed saliency computation system can be used as a prototype model in task-dependent computer vision applications by integrating more advanced higher-level priors, which merits further need to study in future.

References

[1] Achanta, R., Hemami, S., Estrada, F. and Susstrunk, S.: Frequency-tuned salient region detection, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.1597–1604 (2009).

[2] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P. and Susstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.34, No.11, pp.2274–2282 (2012).

[3] Alexe, B., Deselaers, T. and Ferrari, V.: What is an object?, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.73–80 (2010).

[4] Alexe, D., Deselaers, T. and Ferrari, V.: Center-surround divergence of feature statistics for salient object detection, *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, pp.2214–2219 (2011).

[5] Cheng, M., Zhang, G., Mitra, N., Huang, X. and Hu, S.: Global contrast based salient region detection, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.409–416 (2011).

[6] Christopoulos, C., Skodras, A. and Ebrahimi, T.: The jpeg2000 still image coding system: an overview, *IEEE Trans. Consumer Electronics*, Vol.46, No.4, pp.1103–1127 (2002).

[7] Frintrop, S.: Traditional saliency reloaded: A good old model in new shape, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.82–92 (2015).

[8] Gao, D. and Vasconcelos, N.: Discriminant saliency for visual recognition from cluttered scenes, *Proc. Neural Information Processing Systems*, pp.481–488 (2004).

[9] Goferman, S., Zelnik-Manor, L. and Tal, A.: Context-aware saliency detection, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.2376–2383 (2010).

[10] Gopalakrishnan, V., Hu, Y. and Rajan, D.: Salient region detection by modeling distributions of color and orientation, *IEEE Trans. Multimedia*, Vol.11, No.5, pp.892–905 (2009).

[11] Itti, L., Koch, C. and Niebur, E.: A model of saliency-based visual attention for rapid scene analysis, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.20, No.11, pp.1254–1259 (1998).

[12] Jiang, Z. and Davis, L.: Submodular salient region detection, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 23-28, pp.2043–2050 (2013).

[13] Judd, T., Ehinger, K., Durand, F. and Torralba, A.: Learning to predict where humans look, *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, pp.2106–2113 (2009).

[14] Ko, B. and Nam, J.: Object-of-interest image segmentation based on human attention and semantic region clustering, *The Journal of the Optical Society of America A*, Vol.23, No.10, pp.2462–2470 (2006).

[15] Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X. and Shum, H.Y.: Learning to detect a salient object, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.33, No.2, pp.353–367 (2011).

[16] Otsu, N.: A threshold selection method from gray-level histograms, *IEEE Trans. Systems, Man, and Cybernetics: Systems*, Vol.9, No.1, pp.62–66 (1974).

[17] Rutishauser, U., Walther, D., Koch, C. and Perona, P.: Is bottom-up attention useful for object recognition?, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.11–37 (2004).

[18] Shen, X. and Wu, Y.: A unified approach to salient object detection via low rank matrix recovery, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.853–860 (2012).

[19] Treisman, A.M. and Gelade, G.: A feature-integration theory of attention, *Cognitive Psychology*, Vol.12, No.1, pp.97–136 (1980).

[20] Viola, P. and Jones, M.: Rapid object detection using a boosted cascade of simple featuers, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.511–518 (2002).

[21] Wei, Y., Wen, F., Zhu, W. and Sun, J.: Geodesic saliency using background priors, *Proc. 12th European Conference on Computer Vision — Volume Part III*, pp.29–42 (2012).

[22] Yamanaka, M., Matsugu, M. and Sugiyama, M.: Automatic detection of regions of interest using multiple visual saliency measures based on density ratio estimation, *Proc. Vision Engineering Workshop 2010 (ViEW2010)*, Dec. 9-10, pp.7–8 (2010).

[23] Yamanaka, M., Matsugu, M. and Sugiyama, M.: Salient object detection based on direct density-ratio estimation, *IPSJ Trans. Mathematical Modeling and Its Applications*, Vol.6, No.2, pp.78–85 (2013).

[24] Yan, J., Zhu, M., Liu, H. and Liu, Y.: Visual saliency detection via sparsity pursuit, *IEEE Sigunal Processing Letters*, Vol.17, No.8, pp.739–742 (2010).

[25] Yang, J. and Yang, M.: Top-down visual saliency via joint crf and dictionary learning, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.2296–2303 (2012).

Masao Yamanaka has been engaged in the development of image information processing techniques, such as face recognition in still image, pose estimation in depth image, and anomaly detection of human actions in video so far. His recent research interests include not only a wide range of image information processing techniques but also multi modal sensing using advanced machine learning.