Journal of Electronic Imaging

JElectronicImaging.org

Visual saliency region detection by combination of soft- and hard-segmentation-wise approaches

Kanghan Oh
Kwanjong You

SPIE. IS&T imaging.org

Kanghan Oh, Kwanjong You, "Visual saliency region detection by combination of soft- and hard-segmentation-wise approaches," *J. Electron. Imaging* **27**(5), 051204 (2018), doi: 10.1117/1.JEI.27.5.051204.

Visual saliency region detection by combination of soft- and hard-segmentation-wise approaches

Kanghan Oh[a] **and Kwanjong You**[b,*]
[a]Chonbuk National University, Division of Computer Science and Engineering, Jeonju, Republic of Korea
[b]Chosun University, Department of ICT Convergence, Gwangju, Republic of Korea

Abstract. Recent studies in saliency detection have exploited contrast value as a main feature and background prior as a secondary feature. To apply the background prior, most approaches are based on soft- or hard-segmentation mechanisms, and a significant improvement is seen. However, because of contrast feature usage, the soft-segmentation (SS)-wise models have many technical challenges when a high interobject dissimilarity exists. Although hard-segmentation-wise saliency models intuitively use the background prior without usage of the contrast feature, this model suffers from local noises due to undesirable discontinuous artifacts. By analyzing the drawbacks of the existing models, a combination saliency model, reflecting both soft- and hard-segmentation techniques is shown. The proposed model consists of the following three phases: SS-wise saliency, hard-segmentation-wise saliency, and a final saliency combination. In particular, we proposed an iterative reweighting processing for which an influence of outlier segmentation maps is decreased to improve the hard-segmentation-wise saliency. As shown in the experimental results, the proposed model outperforms the state-of-the-art models on various benchmark datasets, which consist of single, multiple, and complex object images. [DOI: 10.1117/1.JEI.27.5.051204]

Keywords: salient-object detection; visual attention model; object segmentation; objectness proposal.

Paper 170806SS received Sep. 20, 2017; accepted for publication Jan. 9, 2018; published online Feb. 12, 2018.

1 Introduction

The human-vision system is able to automatically identify objects in static and dynamic scenes; this fundamental capability allows individuals to automatically concentrate on attractive and important targets in complex scenes. In the computer-vision community, the subject that simulates the human-visual attention system is referred to as salient region detection;[1] the aim of the saliency model is to create an intensity map that represents its probability corresponding objectness. Since the estimated saliency is a higher level feature map, the model can be used for various image-processing and pattern-recognition applications, such as visual tracking,[2] object segmentation,[3,4] object recognition,[5,6] image matching,[7] and image/video compression.[8,9,10,11]

Although the study of saliency region detection is quite extensive and diverse, a common feature among most existing studies[12,13,14,15,16,17] is that the models have been dependent on the contrast feature. Because the contrast feature reflects the human-visual system that automatically concentrates on uniqueness and rarity,[1] it has been widely used for the detection of the salient region. To improve saliency map quality, recent saliency models have begun to employ simple spatial features such as boundary prior or background information as the secondary feature, leading to significantly better performance compared with that of previous models. However, the use of simple boundary prior as the secondary feature is very simple, fragile, and their integration process is mostly heuristic.[17] To address these issues, soft-segmentation (SS)-wise saliency detection models[15,17,18] were proposed, and significant progress has been made compared with those of other saliency models. The point of SS models is that an object's saliency is interpreted by considering a homogeneous-region-level spatial model, which is also called "boundary connectivity (BC);" in these models, the undirected-weighted-graph model is employed to construct spatial weights between each super pixel. During the color contrast computation between patches (or super pixels), these spatial weights are used to weigh similar colors, and the weights on a constructed graph can be regarded as SS information. The models are quite solid compared with prior models for which simple background clues are considered; this is because they are considering cluster- (or segmentation) level background clues. Intuitively, the models are reasonable and robust because the human-visual system does not use only pixel-level clues to identify objects. However, the approach is still not enough to represent the human-visual system; for this reason, limitations are commonly observed when a high dissimilarity between objects (pixel-inside features) exists due to contrast feature usage. Note that the detailed description regarding the drawbacks of contrast feature was addressed in this study.[19] Despite these limitations, the SS-wise models are designed to use the contrast as a main feature; at the same time, the BC model is used to assist the contrast feature. Although their background model can be used directly with hard-segmentation clues, they proposed a "soft" approach because an image segmentation itself is an unsolved problem.[17]

Aiming to solve the problem, hard-segmentation (HS)-wise saliency detection models have been presented in Refs. 1–19. The models have shown that the spatial background clues based on the hard-segmented regions can be

[*]Address all correspondence to: Kwanjong You, E-mail: youkwanjong@hanmail.net

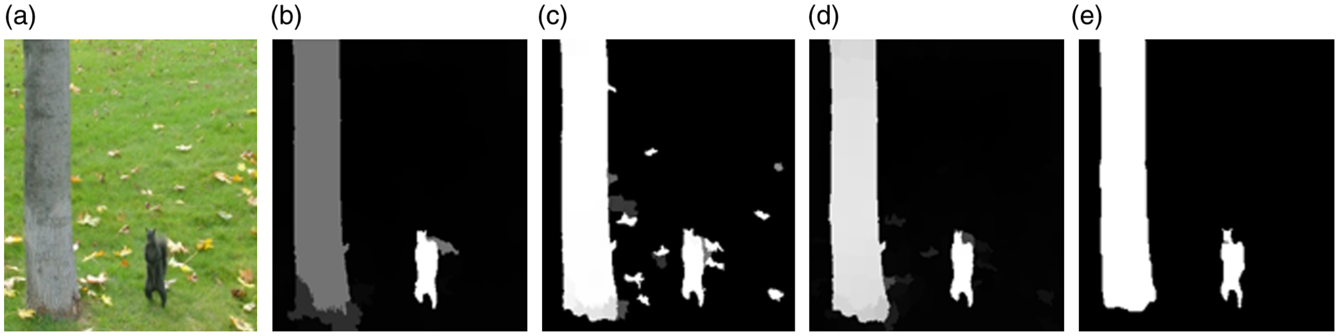

Fig. 1 Limitations of the existing methods and the authors' contributions: (a) original image, (b) SO,[17] (c) RRFC,[19] (d) proposed method, and (f) ground truth. The SS-wise model (b) loses the information of the objects (left tree), and the hard-segmentation-wise model (c) remains local noises, while the proposed model (d) combines both models and generates favorable salient regions.

well expressed in terms of objectness instead of contrast feature. In the HS-wise models, multilevel hard-segmentation maps were constructed, and then the models computed spatial saliency in regard to the segmented maps using the robust background measurement; the models were significantly robust in the limitation of contrast feature usage. However, due to undesirable discontinuous artifacts, the HS-wise saliency model suffers from local noises. In this field, the mentioned difficulties are endemic and universal issues; a few examples are shown in Fig. 1. For the second example, we can see that the SS-based models tend to lose the foreground information of the object (left tree) because of their dependency on the contrast features. For the third example, although object regions are well defined, many local noise blobs are observed in the HS-based model due to undesirable discontinuous artifacts.

In this paper, we proposed a combination model reflecting both soft- and hard-segmentation techniques. The motivation behind such combination process was to overcome the above-mentioned limitations caused by existing models. Our proposed model has the following contributions: (1) a combination system that encompasses both hard and soft techniques is proposed here for the first time. It outperformed techniques of existing models; and (2) to achieve reliable hard-segmentation results, an iterative reweighting process, for which an influence of outlier segmentation maps is decreased, is proposed here for the first time. In addition, SS-wise saliency clues were employed as prior knowledge to improve the quality of segmentation maps.

This paper is organized as follows: in Sec. 2 related works are briefly described with its advantages and disadvantages; the details of the proposed model are described in Sec. 3; in Sec. 4, the proposed methods are evaluated against state-of-the-art approaches with four benchmark datasets; and in Sec. 5, a conclusion and some future work are presented.

2 Related Works

Over the previous decades, a considerable number of studies regarding the visual-saliency model have been proposed based on various mechanisms and extensive reviews can be found in Refs. 21 and 22. In this section, we briefly review the related works based on several viewpoints.

Although handcrafted-saliency-detection models have been quite successful, its heuristic rules still present a limitation for a variety of challenging cases. Aiming to overcome this limitation, deep-learning[23,24,25,26] based saliency models have recently been proposed. The common mechanism of deep-learning based models is that a discriminative feature between the foreground and background is automatically extracted and interpreted during the deep-learning training phase, and then the trained-network model is employed to compute the visual saliency. Note that the convolution neural network (CNN), which is effective for an image analysis, was usually used as the deep-learning algorithm. The CNN-based models have achieved better performance than the handcrafted-saliency models in a variety of challenging cases; however, a sufficient training dataset, a high-quality GPU, and considerable time are required for the learning part, and a failure-cause analysis is very difficult.[19]

Most saliency approaches[12,15,16,17,18,10] were designed to employ contrast value as a main feature. The contrast-based saliency models consist of the following two types: global- and local-contrast-based models. The main mechanism of the global-contrast models computes the object's saliency through the computation of the color contrast between each of the pixels and the mean value of an entire image. Although the global-contrast models are effective to detect salient regions of simple pattern images, these models have a limitation in a poor global contrast and a complex pattern image. The local-contrast-based models have been proposed to overcome the drawbacks of the global-contrast models. These models compute a salient region by considering the local neighborhoods of the pixels. Although these models are useful to an object's saliency, they suffer from local noises when computing complex pattern images. Moreover, the window (kernel) sizes for different objects at different scales must be modified to optimize final salient region.[1] As mentioned previously, the contrast that reflects a human's visual attention system has been commonly employed as a standard feature for the most saliency models, but its extreme dependency on the most-highlighted region causes drawbacks when the object dissimilarity is high.[19]

Recently, the SS-wise saliency models were proposed and have shown excellent performance among the handcrafted-saliency models.[15,17,18] The undirected weighted graph was constructed to obtain weight values between super pixels, and a robust boundary measure was employed as the spatial prior. In the SS-wise saliency model, the constructed graph can be regarded as soft clustering information, and it has the similar effect of analyzing hard-segmentation results.

A graph model was presented in Ref. 15; the model incorporates local and global contrast, and these clues are combined by exploiting a robust background measure. Unlike the above models, the method of Ref. 16 directly builds hierarchical hard-segmentation maps using hierarchical-clustering techniques with three-heuristic thresholds; the saliency scores are calculated using both the local contrast clues and consistent-inference methods.[16] Although various techniques to compute visual saliency have been applied, the soft segmentation-based models are consequentially based on the contrast feature. For this reason, the models suffer from contrast limitation.

In contrast to the SS-wise models, the aim of the hard-segmentation-wise saliency models[1–19] is to detect the salient regions without the contrast based on multilevel hard-segmentation maps; for this model, only the spatial features that represent the pixel variation and the location clues were adopted for the saliency-score computation. For this reason, the models are quite solid in contrast limitation. However, for the hard-segmentation phase, because heuristic or simple parameter selection techniques were adopted, undesirable-outlier segmentation maps are often generated, and it led to poor performance. In addition, the works[1–19] have used a simple and heuristic integration method to generate the final-fused saliency map without robust optimization processing. Although a optimization process, which is called "recursive processing," was used to optimize the saliency map in the RRFC,[19] this model is very time consuming and suffers from local noise due to its recursive process, which tends to reinforce the local noise when an initial saliency map has relatively strong noise saliency.

3 Proposed Approach

The proposed salient-object detection model is summarized in Fig. 2 and is fully presented in this section. The proposed model consists of the following four phases: (1) preprocessing, (2) SS-wise saliency, (3) hard-segmentation-wise saliency, and (4) saliency optimization. In the preprocessing phase, an input image is abstracted as a set of super pixels using the simple linear iterative clustering (SLIC) algorithm;[27] given a set of super pixels S, we mainly employed two types of regional features, which are average color (CIELAB) and centroid coordinates of super pixel patches. In the second and third stages, the SS- and HS-wise saliency clues are computed; in particular, to acquire a reliable-hard-segmented region, SS-wise saliency clues were employed as a priori knowledge and the iterative reweighting process is implemented to weight favorable segmentation maps during the HS-wise saliency model computation. In the last stage, the saliency clues were optimized using the objective function containing a robust background measure.

3.1 *Background Prior Model*

To compute object saliency corresponding to each segmented region or image patch, a robust background-measurement model called BC proposed in Ref. 17 is considered. The definition of BC is more robust compared with those of other boundary prior-based models that are heuristic, simple, and fragile. The definition of the BC method can be written as follows:

$$\mathrm{BC}(R) = \frac{Np(p|p \in Bnd, p \in R)}{\sqrt{Np(p|p \in R)}}, \tag{1}$$

where Bnd is the boundary patches, R is the observed cluster region, p is the image patch, and $Np(.)$ is a function to count the image patches. Figure 3 shows the definition of the BC; the example has four clustering regions, and we can easily identify the foreground and background clusters. By the BC definition, the blue and red clusters have 0.83 and 0.63, respectively, and the white and gray clusters have 2.41 and 2.80, respectively. The model computes the cluster-based connection strength with the image boundary, and it returns higher values to background clues. In summary, the salient

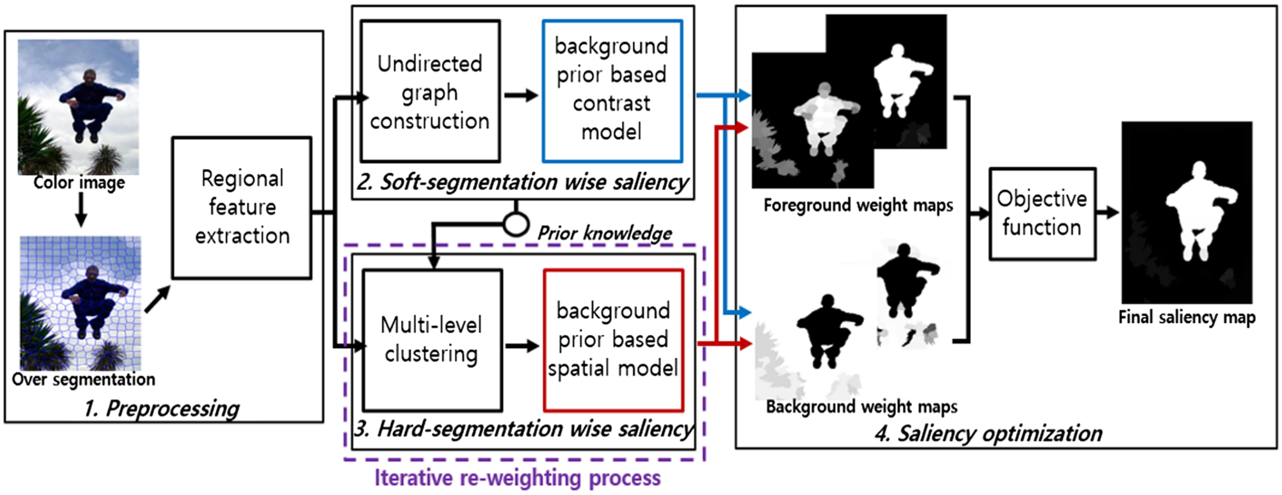
Color image
Over segmentation
1. Preprocessing
Regional feature extraction
Undirected graph construction
background prior based contrast model
2. Soft-segmentation wise saliency
Prior knowledge
Multi-level clustering
background prior based spatial model
3. Hard-segmentation wise saliency
Iterative re-weighting process
Foreground weight maps
Background weight maps
Objective function
Final saliency map
4. Saliency optimization

Fig. 2 Framework of the proposed method. The overall framework consists of the following four phases: preprocessing, SS-wise saliency, hard-segmentation-wise saliency, and saliency optimization. After preprocessing and soft- and hard-segmentation-wise saliency computation, foreground and background weight maps are obtained from each stream. They are then fused by background measure-based optimization function.

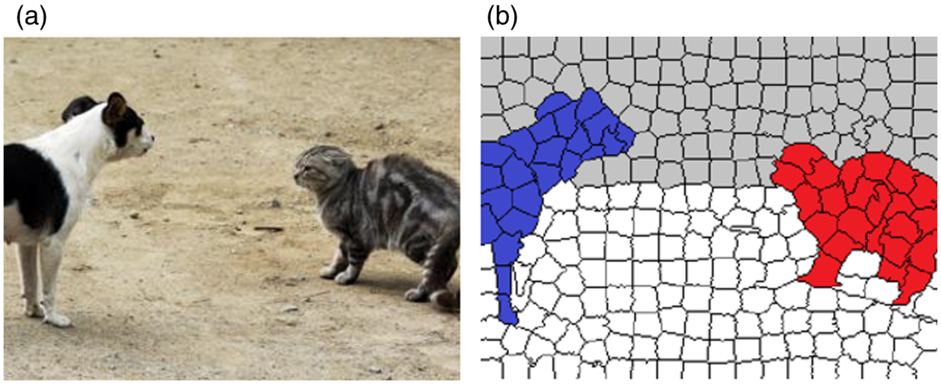

Fig. 3 Example of background prior model: (a) original image and (b) BC.

regions are much less connected to image borders than the background elements.

3.2 *Soft-Segmentation-Wise Saliency*

In this phase, the SS-wise saliency[17] is computed using both the undirected-weighted-graph theory and the BC definition. In the first stage, the undirected weighted graph is constructed by connecting all adjacent super pixels; the "spanning area" of each super pixel p is defined by the following equation:

$$E(p_i) = \sum_{j=1}^{N} exp\left[-\frac{D_{\text{geo}}(p_i, p_j)}{2\sigma_{\text{area}}^2}\right], \tag{2}$$

where the result of Eq. (2) is a soft-segmented area of the region that p_i belongs to, p is a super pixel and N is its total number, $D_{\text{geo}}(p_i, p_j)$ is the geodesic distance between two super pixels in the CIELab color domain and is the accumulated edge weights along their shortest path on the constructed graph by computing the Euclidean distance between their average colors. For this reason, as the frequency of color similar to patch p_i increases, the result of $E(p_i)$ is also increased. The BC value is computed by the following equation:

$$Bndcon(p_i) = \frac{E(p_i|p_j\epsilon bnd)}{\sqrt{E(p_i)}}, \tag{3}$$

where $E(p_i|p_j\epsilon bnd)$ means that it only considers the patches p_j located on the image borders when computing function $E(.)$. To compute Eq. (3), the shortest paths between super pixels are calculated using Johnson's algorithm;[28] based on the above definition, the background weight can be written as

$$B_{ss}(p_i) = 1 - \exp\left[-\frac{Bndcon(p_i)^2}{2\sigma_{bndcon}^2}\right]. \tag{4}$$

When BC is large, it is close to 1, and its result represents a background probability. The foreground weight, which is called background-weighted contrast, is defined as

$$F_{ss}(p_i) = \sum_{i=1}^{N} D_{\text{contrast}}(p_i, p_j) D_{\text{spatial}}(p_i, p_j) B_{ss}(p_i), \tag{5}$$

where $D_{\text{contrast}}(.)$ is the color contrast between super pixels, and $D_{\text{spatial}}(.)$ is the spatial distance between super pixels. Equation (5) was designed to receive high B_{ss} for the foreground patches, and its contrast is enhanced. In summary, although SS-wise saliency is based on the global-contrast model, robust background measurement called "BC" was applied as an additional feature during the contrast computation, and it led to high performance. The visual example is shown in Fig. 4.

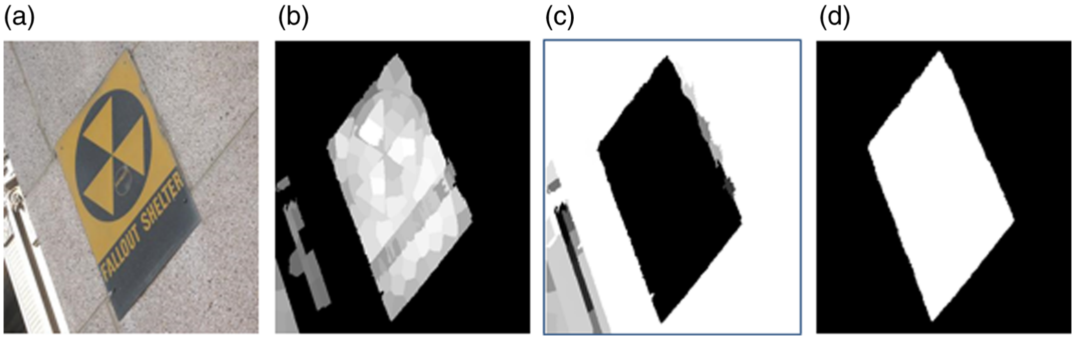

Fig. 4 Example of SS-wise saliency: (a) original image, (b) foreground weight map in Eq. (5), (c) background weight map in Eq. (4), and (d) ground truth.

3.3 *Hard-Segmentation-Wise Saliency*

Zhu et al.[17] have mentioned that the BC is intuitive, but it is difficult to compute directly because an image segmentation itself is a challenging and unsolved problem (i.e., parameter selection). For this reason, the study[17] does not use the definition directly but applies it as a weight of the color contrast computation. However, as mentioned previously, the color contrast feature has a limitation when a high inter-object dissimilarity exists. To overcome the limitation, the hard-segmentation-wise saliency models[1–19] were proposed, and the process of these models usually consists of three phases: multilevel segmentation-region construction, spatial-saliency computation, and optimization.

In this study, we use the hierarchical-clustering algorithm for the multilevel segmented-region construction; in consideration of time computation, this way is more effective than mean-shift[29] usage, for which multilevel kernels[30] are considered. To construct reliable segmented regions, we have considered the foreground weight F_{ss} as a sixth regional feature, and it led to improved segmentation quality. In the hierarchical-clustering process, threshold values (number of class) should be defined for hard-segmented region construction, and we empirically set its thresholds at $T = [2,3\ldots 8]$ in the experiment. After constructing hard-segmentation maps, we computed corresponding saliency maps using the BC. Unlike the SS-wise models in which an input unit is used for the patches, the robust background model is directly calculated without the color contrast computation, and the input unit in the HS-wise process is the hard-segmented regions R; so, it can be expressed as $R_k = \{p_1, p_2, \ldots, p_n\}$. To directly apply the background model to our work, the super pixel S and the clustering region C are now considered for the patch p and the observed hard-segmented region R, respectively[20] in Eq. (1), and the HS-wise initial saliency map can be defined by the following equation:

$$HS_{\text{map}(o)} = BC(R_k), \qquad [k = 1, \ldots, n], \tag{6}$$

where R_i means an i'th segmented region, and n is the number of segmented regions in a hierarchical-clustering map; we can see its results, $HS_{\text{map}} = \{HS_{\text{map}(1)}, HS_{\text{map}(2)}, \ldots, HS_{\text{map}(7)}\}$ from the second row in Fig. 5, and the multilevel saliency maps HS_{map} (in Fig. 5, second row) are linearly integrated to acquire HS_{Imap}. The visual results are shown in Fig. 5, where we can see the well-defined clustering maps regardless of the parameter T changes.

For the optimization process, just like the SS-wise saliency process, the results should be expressed as two maps representing foreground and background-weight maps. In the proposed method, the sigmoid function was employed to obtain the continuous (soft) weights; sigmoid functions to build the foreground F_{hs} and background B_{hs} maps are given by the following equation:

$$F_{hs}(p_i) = \frac{1}{1 + e^{-a[HS_{\text{Imap}}(p_i) - \mu]}}, \tag{7}$$

$$B_{hs}(p_i) = \frac{1}{1 + e^{a[HS_{\text{Imap}}(p_i) - \mu]}}, \tag{8}$$

where a is a curve gradient, and μ and HS_{Imap} are the harmonic mean value and an integrated saliency map, respectively; $HS_{\text{Imap}}(p_i)$ denotes patch p_i wise average value on the overlapping region between HS_{Imap} and super pixel p_i. Note that the proposed HS-wise processing consumes just 0.28 (average) seconds per 400×300 image, and this is significantly fast compared with the existing models,[1–19] which require a processing time about 2 to 4 s.

3.4 *Iterative Reweighting Process*

In the HS-wise stream, performance is significantly controlled by the segmentation map's quality; we, therefore, attempt to reduce influence of outlier segmentation elements, which cause performance degradation during the iterative processing. The pseudocode of the iterative processing is described in Fig. 6, and its process consists of the following phases.

1. Similarity scores between the HS_{Imap} and each HS-wise saliency map $HS_{\text{map}(o)}$ is computed using the 2-D correlation coefficient, and these scores are regarded as weight values for each $HS_{\text{map}(o)}$.

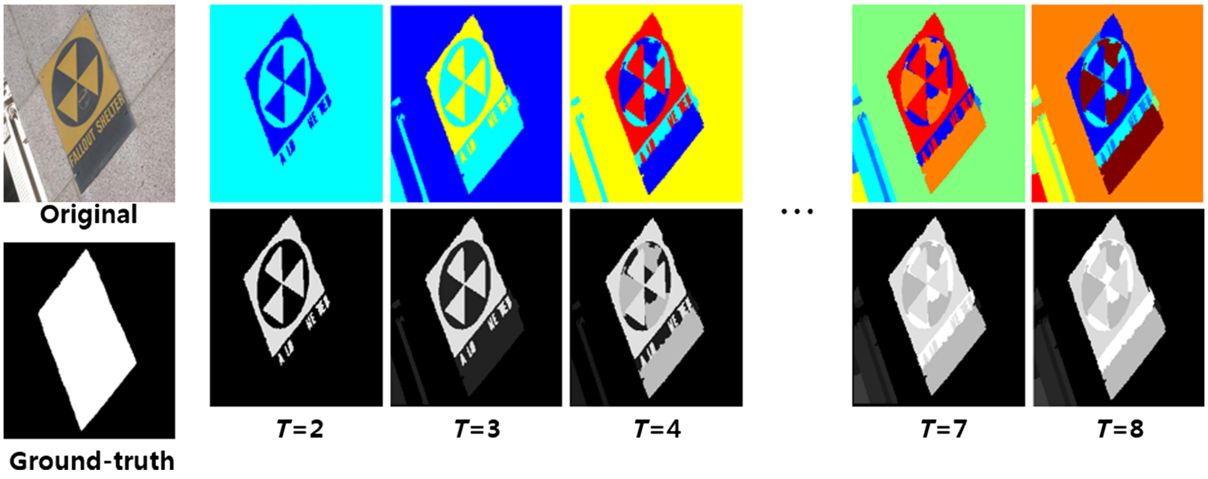

Fig. 5 Example of hard-segmentation-wise saliency: first row represents the segmented maps using the hierarchical-clustering algorithm, and its corresponding saliency maps are illustrated in second row.

Algorithm 1: Iterative reweighting procedure

```
Input : HS wise saliency maps $\boldsymbol{HS_{map}} = \{ HS_{map(1)}, HS_{map(2),\dots}, HS_{map(N)} \}$
Output : Integrated HS wise saliency map $HS_{Imap}$
$wHS_{map} \leftarrow \emptyset,\ HS_{Imap} \leftarrow \emptyset$ // initialization
$HS_{Imap} \leftarrow \frac{1}{N}\sum_{o=1}^{N} HS_{map(o)}$ // Linear fusion
Repeat {
  for ($o$ =1 to $N$ ){
      $\rho_i = \frac{(HS_{map(o)} - \overline{HS_{map(o)}})\,(HS_{Imap} - \overline{HS_{Imap}})}{\sqrt{(HS_{map(o)} - \overline{HS_{map(o)}})^2\,(HS_{Imap} - \overline{HS_{Imap}})^2}}$ //weighting value computation
  }
  $wHS_{Imap} \leftarrow \frac{1}{N}\sum_{i=1}^{N} HS_{map(i)}\,\rho_i$ // Weighted fusion
  If $HS_{Imap} \cong wHS_{Imap}$ then repeat break
  else $HS_{Imap} \leftarrow wHS_{Imap}$ end //Updating integrated saliency map
}
$HS_{Imap} \leftarrow$ Normalization $(HS_{Imap})$
Return $HS_{Imap}$
```

Fig. 6 Pseudocode of iterative reweighting procedure.

2. We multiply each HS-wise saliency map by the corresponding weight, and then they are fused to update HS_{Imap} (weighted fusion).
3. Processes (1) to (2) are repeated until there no significant changes remain between the current and previous sources.

The proposed iterative reweighting process encourages statistical consistency, leading to decrease in the influence of outlier segmentation maps during the fusion process. The goal is to weight good segmentation maps and reject irregular sources; thus, the proposed iterative processing is not compute-expensive compared with the existing model,[19] in which the mean-shift algorithm is repeatedly executed to improve segmentation results. Figure 7 shows that the visual performance of our saliency maps is enhanced with an increasing number of iterations.

3.5 *Saliency Optimization*

In prior works, the saliency clues computed from the multilevel phases are combined heuristically using weighted summation or multiplication.[1–19] In the proposed method, we have employed a cost-function that is based on the error-minimization technique to optimize the final saliency region. Given the foreground and background weight maps, the objective cost function is defined by the following equation:[31,17,15]

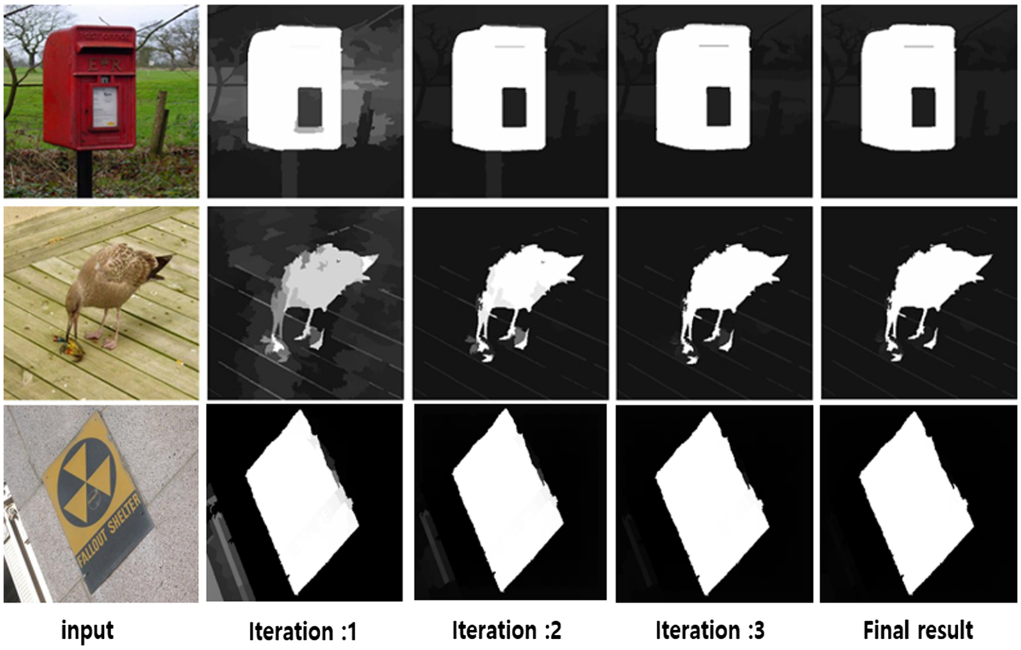

Fig. 7 Visual example of iterative reweighting process.

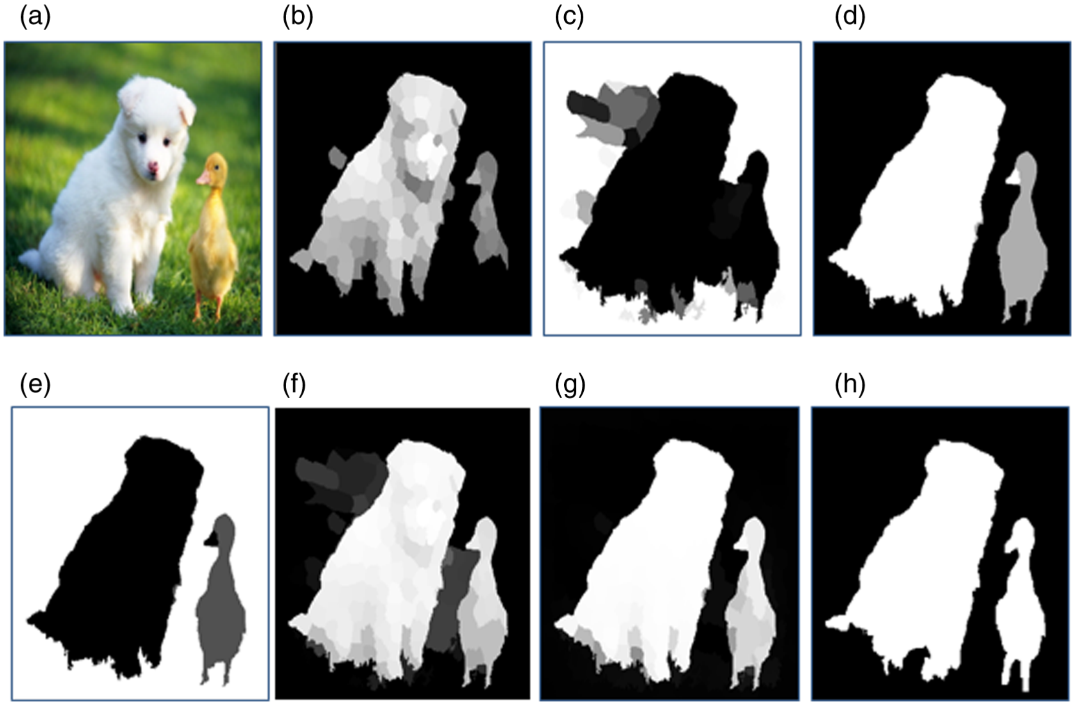

Fig. 8 Illustrations of our process. (a) Original image, (b) SS-wise foreground weight map in Eq. (5), (c) SS-wise background weight map in Eq. (4), (d) HS-wise foreground weight map in Eq. (7), (e) HS-wise background weight map in Eq. (8), (f) simple linear integration result, (g) optimized saliency map by minimizing Eq. (9), and (h) ground-truth.

$$\arg\min_s \left\{ \sum_{i=1}^{N} [F_{ss(i)} + F_{hs(i)}](s_i - 1)^2 + \sum_{i=1}^{N} [B_{ss(i)} + B_{hs(i)}] s_i^2 + \sum_{i,j} \omega_{ij} (s_i - s_j)^2 \right\}, \tag{9}$$

where F and B are foreground and background weights. ω_{ij} is the smoothness term, and it is effective to eliminate small noises in both foreground and background; three terms in Eq. (8) are all squared errors, and the optimal saliency map is computed by the least square method.[17] High F encourages saliency s_i to take the saliency value close to 1, and B encourages saliency s_i to move close to 0. The last smoothness term encourages continuous saliency values, and it is effective to remove local noise in both foreground and background regions. For every adjacent super pixel pair (i, j), the smoothness term is defined by the following equation:

$$\omega_{ij} = \exp\left[-\frac{D_{\text{contrast}}(p_i, p_j)}{2\sigma_{\text{smooth}}^2} \right]. \tag{10}$$

The results of the optimization process are shown in Fig. 8; as can be seen in the following figure, the overall salient region is enhanced after the optimization process is implemented, and a significant improvement exists when comparing with a simple integration result [Fig. 8(f)].

4 Experimental Results

The experiments were conducted on an Intel(R) Core(TM) i5 4670 with a CPU of 3.40 GHz and 12 GB of memory. The proposed model was evaluated on three benchmarks: MSRA, ECSSD, and MSOD; our performance is compared with those of the state-of-the-art methods, such as CHS,[2] RC,[12] SO,[17] RFC,[20] and RRFC,[19] respectively. The relevant competitive models were selected based on the citations and their high performance. The MSRA-ASD[12] dataset includes 1000 single-object images with a pixel-wise ground truth that is indicated from the MSRA10K dataset; the dataset is the most commonly used for the evaluation of salient-detection performance. The ECSSD[16] contains 1000 images with complex patterns in both the foreground and background. The SED2[32] is a multiple-salient object benchmark, which consists of 100 images with more than two objects with a higher dissimilarity.

4.1 *Setup and Evaluation Methods*

The precision, recall, and F-measure (F_β), which are commonly used for a quantitative comparison of different models, were considered for the performance evaluation. For a reliable comparison of the various saliency-detection methods, the salient regions should be evaluated with a variation of the fixed-threshold values from 0 to 255; here, precision represents the percentage of salient pixels that correspond to the ground truth, whereas recall represents the ratio of the salient pixels that belong to the total number of ground truths. As discussed in Refs. 33, 19, and 34, the true negative counts are not considered for either the precision or the recall measure, and this means that these measures cannot be used for an evaluation of the nonsalient regions. For the quantitative comparison, we, therefore, used the F-measure curve, for which various thresholds are considered, and the AUC, which is the area under the F-measure curve, instead of the precision–recall curve; the F-measure is calculated using the following equation:

$$F_{\beta} = \frac{(1+\beta^2)\text{Precision} \times \text{Recall}}{\beta^2\text{Precision} + \text{Recall}}. \tag{11}$$

The recall metric detects the percentage of true positive pixels in the saliency map through the use of the total number of true positives in the ground truth, and the precision metric provides the percentage of detected true positives as compared with the total number of positive pixels in the detected binary-motion mask. Since $\beta^2 = 0.3$ is set for most of the existing methods,[12,1–34,16,17] more weighting of the precision rather than the recall, $\beta^2 = 0.3$, was also set for the quantitative-performance comparison involving the state-of-the-art methods. The performances of the saliency models were also evaluated according to the average precision, recall, and β^2, which are commonly used in related areas to evaluate the performance; here, an image-dependent adaptive threshold value[35] that is computed as twice the mean saliency was used to perform the saliency-map binarization. For a more comprehensive comparison, we also evaluated the saliency-detection models using the mean absolute error (MAE), whereby a result for the similarity between the continuous-saliency map $\overline{S}$ and the ground truth $\overline{Gt}$, both of which had been normalized from 0 to 1, was provided. The MAE score is defined by the following equation:

$$\text{MAE} = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |\overline{S}_{xy} - \overline{Gt}_{xy}|. \tag{12}$$

4.2 *Quantitative Performance Comparison*

Our model was evaluated on the four datasets: MSRA-ASD, MSRA10K, ECSSD, and SED2 and our performance is compared with those of the state-of-the-art methods, such as CHS,[16] RC,[12] SO,[17] RFC,[20] and RRFC,[19] respectively. The relevant competitive models were selected based on their performance. Note that we selected the parameters of the compared saliency models in accordance with their parameter settings that were already noted in the existing manuscripts.[12,20,19,16,17] The quantitative comparisons are presented in Fig. 9, and their row and column reflect the benchmarks (from top to bottom: MSRA-ASD, MSRA10K, ECSSD, and SED2) and the evaluation methods (from left to right: F-measure curve, F-measure, and MAE), respectively.

Since the saliency map consisted of continuous intensity values, it should be evaluated with a variation of the fixed-threshold values from 0 to 255; we, therefore, used F-measure curve for which various thresholds are considered. The proposed method is evaluated on both the MSRA-ASD and MSRA10K datasets. The MSRA-ASD dataset includes 1000 single-object images with a pixel-wise ground truth that is indicated from the MSRA10K dataset. Notably, even though the MSRA-ASD is made up of simple foreground and background images, in recent years, it is the most commonly used dataset for the evaluation of the salient-detection performance. To obtain more extensive experimental results, the MSRA10K dataset, which is composed of 10,000 single-object images with the pixel-level ground truth, is also used in this test. In Fig. 9 (MSRA-ASD and 10 K), in terms of the F-measure curve, our model and the RRFC outperform those of other models; in the specific range between 0.3 and 0.8, our model is clearly outstanding. In particular, our model clearly outperforms those of the existing models in F_{β} (second column, MSRA-ASD and 10K). The RRFC and our model have achieved favorable performance rates regarding the MAE, and this means that the model results in a well-defined background.

To overcome the simplicity of the MSRA, an ECSSD containing 1000 images with complex patterns in both the foreground and background is proposed in Ref. 16; however, although this dataset includes many semantically meaningful images, the images are structurally complex for a performance evaluation. In Fig. 9 (ECSSD-left), our curve is consistently higher than those of the existing models in the specific range between 0.2 and 0.9, and our model is also outstanding in terms of F_{β}. However, our model results in an ordinary performance in the ECSSD when considering the MAE, and this result shows that the proposed model tends to fail the background region detection in the complex pattern image.

This dataset[32] consists of 100 images containing exactly two objects, and the pixel-wise ground truth is also provided. In particular, some of the images have two challenging tasks as follows: first, the properties of the objects are radically different; and second, the objects are located in the image borders. The SED2 evaluations, therefore, allow for an immediate identification of the limitations of the existing approaches. Considering F-measure curves, our model was clearly included in the high-performance group, but it is very ambiguous. However, we can see that the proposed model outperforms others in both the MAE and F_{β}.

In consideration of the performance, the proposed model and the RRFC have achieved outstanding performance compared with those of other models; there may be a debate, however, regarding which model is better. Note that the proposed model is highly competitive when compared with the RRFC; the RRFC is very computation-intensive because of its recursive process. Given a typical 400 × 300 image, the RRFC takes 4 s for testing; in addition, the time consumption of the RRFC is significantly irregular because more than average processing time is often required to reach a convergence state according to image states (i.e., some cases take 8 to 15 s for testing). Unlike the RRFC, the proposed model consumes just 0.35 s per 400 × 300 image, and its overall processing time is more regular than that of the RRFC. The processing time results regarding the saliency models are shown in Table 1. In particular, the performance of the proposed model is generally outstanding in terms of the F-measure, and this phenomenon means that the model successfully detects the foreground region and the respective spatial locations in the scenes.

The visual comparisons regarding the four benchmarks (MSRA-ASD, 10K, ECSSD, and SED2) are shown in Fig. 10. The results of the proposed model are relatively accurate compared with those of the existing model; in particular, a great improvement is evident when a comparison is made with its previous models (SO and RFC). In relation to the SED2 benchmark, Fig. 10 shows that our model correctly and uniformly highlights multiple salient objects regardless of both the higher object dissimilarity and the number of objects, whereas the object regions of the existing models are not uniformly highlighted. In terms of the complex-pattern image, the proposed method not only successfully detects the object region, but it also clearly eliminates the

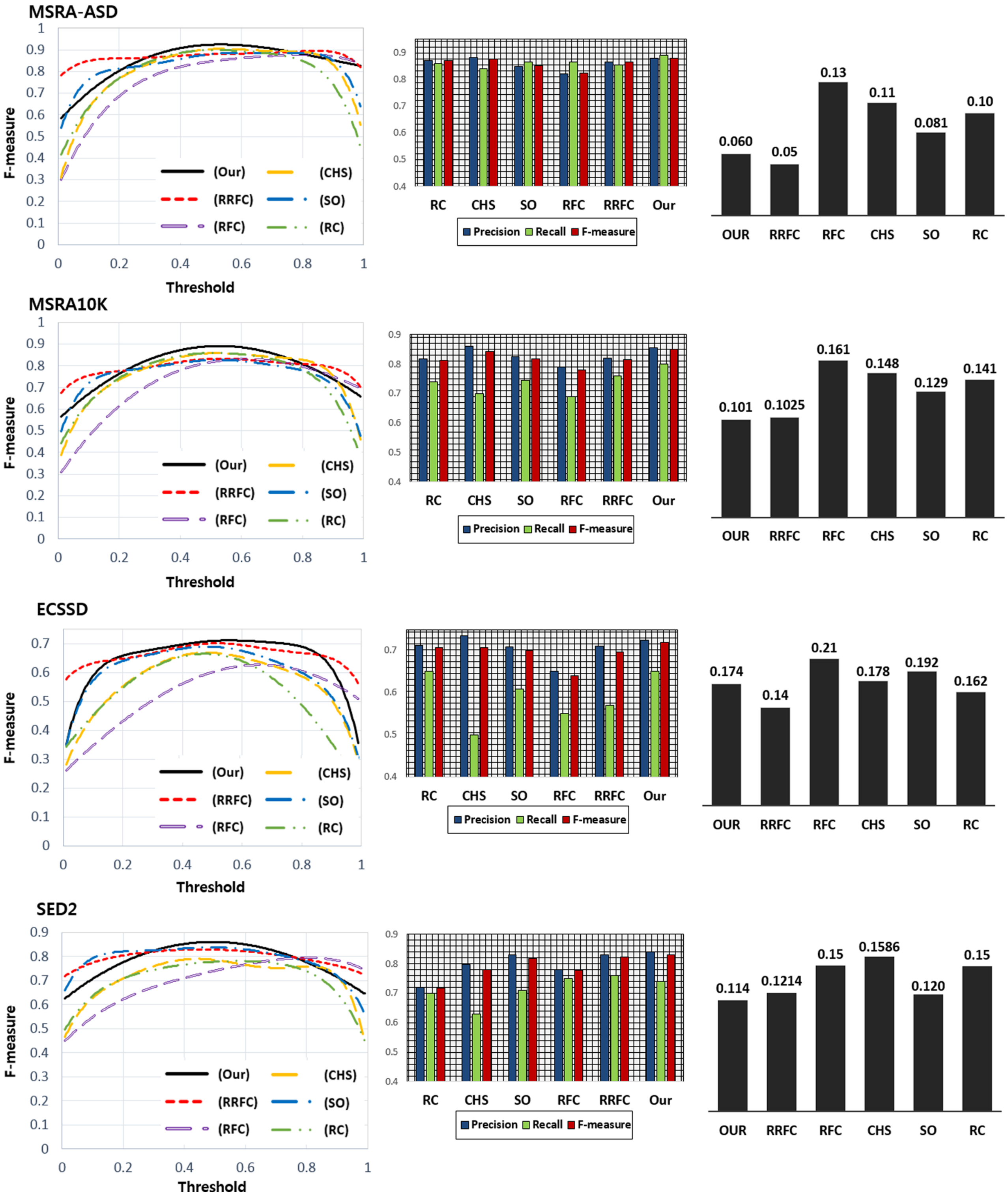

Fig. 9 Quantitative comparison of salient-object detections using *F*-measure curve, precision, recall, and MAE.

Table 1 Computation time comparisons.

	SO[17]	RC[12]	CHS[16]	IEFC[20]	RFC[1]	RRFC[19]	Our
Code	MATLAB	C	C	MATLAB	MATLAB	MATLAB	MATLAB
Runtime (s)	0.21	0.11	0.45	2.20	1.54	4.31	0.35

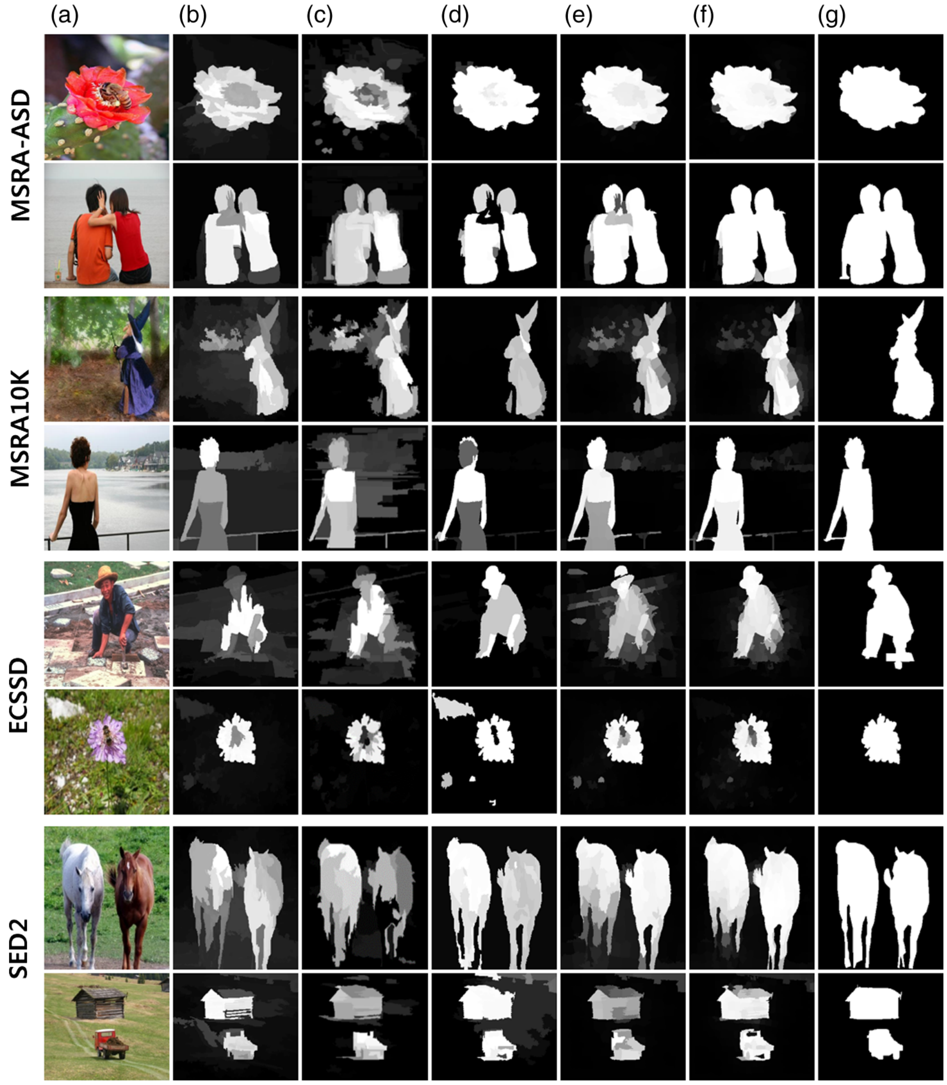

Fig. 10 Visual comparison of salient-object detection results: (a) original, (b) CHS,[16] (c) RC,[12] (d) RRFC,[19] (e) SO,[17] (f) ours, and (g) ground truth.

background. In summary, the proposed model is generally superior regardless of the benchmark type; in particular, the outstanding F_{β}rates show that the foreground clues of the proposed model are well highlighted compared with those of the existing models.

4.3 *Analysis of Proposed Model*

In this section, the manner in which the accuracy of the proposed model is affected by both the parameters and the partial functions is further analyzed. In the first stage, the performances are described in Table 2 according to the number of super pixels. As the super pixel number increases, the processing time also increases, and the results show that our model has archived a favorable performance of between 300 and 400 super pixels. Generally, the number of super pixels does not have a major influence on the final results. The results regarding the influence of the gradient value for normalization are described in Table 3, where the gradient $a = 7$ is the most proper for the benchmark, and "**B**" represents

Table 2 Performance comparison according to super pixel numbers.

Super pixels	Precision	Recall	*F*-measure	Runtime (s)
50	0.822	0.741	0.814	0.177
100	0.829	0.758	0.822	0.198
200	0.827	0.790	0.824	0.230
300	0.828	0.801	0.825	0.276
400	0.835	0.809	**0.833**	0.335
500	0.830	0.813	0.829	0.409
600	0.821	0.820	0.821	0.480
700	0.818	0.824	0.819	0.598

Note: Bold value represents the best performance.

Table 3 Performance comparison according to gradient value of sigmoid function.

Gradient values	Precision	Recall	*F*-measure
1	0.827	0.809	0.825
3	0.827	0.809	0.825
5	0.833	0.808	0.831
7	0.835	0.809	**0.833**
10	0.834	0.809	0.832
20	0.834	0.807	0.832
B	0.820	0.799	0.818

Note: Bold value represents the best performance.

Table 4 Performance comparison according to cluster numbers in the HS-wise process.

Cluster numbers	Precision	Recall	*F*-measure
K = [2 to 4]	0.833	0.792	0.829
K = [2 to 8]	0.835	0.809	**0.833**
K = [2 to 12]	0.827	0.814	0.826
K = [2 to 16]	0.820	0.819	0.820
K = [2 to 20]	0.816	0.822	0.816

Note: Bold value represents the best performance.

Table 5 Performance comparison according to combination process.

	Precision	Recall	*F*-measure	MAE	Runtime (s)
SS-wise stream	0.815	0.783	0.812	0.120	0.173
HS-wise stream	0.829	0.788	0.825	0.112	0.128
Combination	0.835	0.809	**0.833**	**0.107**	0.301

Note: Bold values represent the best performance.

the result using the harmonic mean binary maps without the sigmoid function usage; as can be seen from the results, the continuous maps by the sigmoid functions are relatively advantageous for obtaining foreground and background weights compared with the harmonic mean value usage. As the gradient value was decreased, we can see that our model tended to output a favorable performance with the lower threshold values and a poor performance with the higher threshold values, whereas the opposite effect was achieved with the larger gradient values. This finding means that the saliency values of the foreground for which the small gradient values are considered were formed at a low-intensity position. The performance comparisons according to hierarchical cluster-number changes are described in Table 4, and these results show that the use of too many clusters causes performance degradation. Through analysis of the precision scores, we can easily assume that this phenomenon is because of over-segmented regions caused by higher threshold values. The results in Table 5 show that the combination result of two streams is advantageous over the independent use of each stream. In particular, the recall score is greatly improved after the combination process, and the HS-wise stream shows better performance than the SS-wise stream. From the experimental results, we can easily confirm the synergy effect by the proposed combination mechanism. With an increasing number of iterations, the performance of our model was exponentially improved (Fig. 11), leading to a convergence of performance regardless of any further increase. A considerable performance enhancement was observed between the first and second iterations in terms of *F*-measure curve and MAE. In addition, a result stopped by the proposed stop condition (adaptive threshold) of recursive processing was clearly included in the convergence domain.

5 Conclusion

In this paper, we proposed a combination model reflecting both soft- and hard-segmentation techniques. In particular, in the HS-wise stream, the iterative reweighting process was proposed to decrease influence of outlier segmentation maps, and the prior knowledge generated from the SS-wise stream was employed to enhance the segmentation map's quality; the proposed model provides a favorable result compared with the existing model in terms of both performance

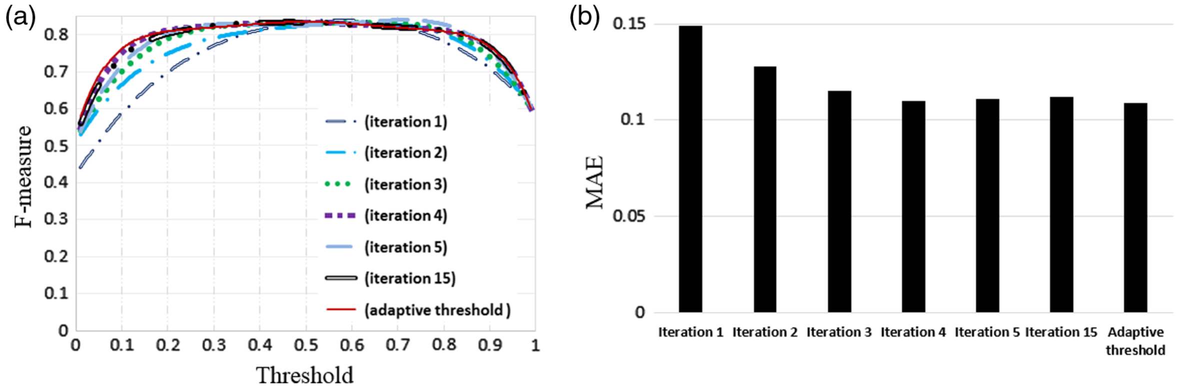

Fig. 11 Performance comparison according to iterative reweighting processing. (a) *F*-measure curve and (b) MAE.

and processing time. In addition, in the combination phase, the robust optimization function was used to fuse results from the two streams, and the result shows that the combination of two streams outperforms the independent use of each stream. The experimental results demonstrate that our model achieved superior performance in terms of the efficiency of the MAE and the superior F-measure on benchmarks, which reflect simple, complex, and multiple objects. In terms of the limitations of the proposed model, the final result obtained using the iterative processing is heavily reliant on its initial state, and the weighted fusion method is very simple; furthermore, the hierarchical-clustering algorithm occasionally failed to detect optimal clusters when there was an insufficient feature distribution with unclear gradient, leading to poor segmentation results. For a future work, we plan to improve the performance of the proposed model using an adaptive fusion method[36,37] and multiple clustering algorithms (ensemble technique); in addition, a theoretical analysis of the proposed model needed to be conducted.

Disclosures

The authors declare that there are no conflicts of interest related to this article.

Acknowledgments

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2017-2015-0-00378) supervised by the IITP (Institute for Information & communications Technology Promotion).

References

1. K. H. Oh et al., "Detection of multiple salient objects through the integration of estimated foreground clues," *Image Vision Comput.* **54**, 31–44 (2016).
2. S. Stalder, H. Grabner, and L. Van Gool, "Dynamic objectness for adaptive tracking," in *Asian Conf. on Computer Vision*, pp. 43–56 (2012).
3. J. Chanho and K. Changick, "A unified spectral-domain approach for saliency detection and its application to automatic object segmentation," *IEEE Trans. Image Process.* **21**(3), 1272–1283 (2012).
4. B. C. Ko and J. Y. Nam, "Object of interest image segmentation based on human attention and semantic region clustering," *Pattern Recognit.* **23**(10), 2462–2470 (2006).
5. S. Frintrop and P. Jensfelt, "Attentional landmarks and active gaze control for visual SLAM," *IEEE Trans. Rob.* **24**(5), 1054–1065 (2008).
6. U. Rutishauser et al., "Is bottom up attention useful for object recognition?" in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 37–44 (2004).
7. L. Zheng, S. Wang, and Q. Tian, "Coupled binary embedding for large scale image retrieval," *IEEE Trans. Image Process.* **23**(8), 3368–3380 (2014).
8. Z. Li, S. Qin, and L. Itti, "Visual attention guided bit allocation in video compression," *Image Vision Comput.* **29**(1), 1–14 (2011).
9. Y. F. Ma et al., "A user attention model for video summarization," in *ACM Int. Conf. on Multimedia*, pp. 533–542 (2002).
10. Y. Fang et al., "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Trans. Image Process.* **21**(9), 3888–3901 (2012).
11. Y. Fang et al., "A video saliency detection model in compressed domain," *IEEE Trans. Circuits Syst. Video Technol.* **24**(1), 27–38 (2014).
12. M. M. Cheng et al., "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(3), 569–582 (2015).
13. D. Klein and S. Frintrop, "Center-surround divergence of features statistics for salient object detection," in *IEEE Int. Conf. on Computer Vision (ICCV 2011)*, pp. 2214–2219 (2011).
14. G. Kotstra, B. Boer, and L. R. B. Schomaker, "Predicting eyes on complex visual stimuli using local symmetry," *Cognit. Comput.* **3**(1), 223–240 (2011).
15. W. Qiaosong, Z. Wen, and P. Robinson, "Grab: visual saliency via novel graph model and background prior," in *IEEE Conf. on Computer Vision and Pattern Recognition* (2016).
16. J. Shi et al., "Hierarchical image saliency detection on extended CSSD," *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(4), 717–729 (2016).
17. W. Zhu, S. Liang, and J. Sun, "Saliency optimization from robust background detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2814–2821 (2014).
18. R. S. Srivatsa and R. V. Babu, "Salient object detection via objectness measure," in *IEEE Int. Conf. on Image Processing*, pp. 4481–4485 (2015).
19. K. H. Oh, M. E. Lee, and Y. R. Lee, "Salient object detection using recursive regional feature clustering," *Inf. Sci.* **387**, 1–18 (2017).
20. K. H. Oh et al., "Detection of multiple salient objects by categorizing regional features," *KSII Trans. Internet Inf. Syst.* **10**(1), 272–287 (2016).
21. A. Borji et al., "Salient object detection: a benchmark," *IEEE Trans. Image Process.* **24**(12), 5706–5722 (2015).
22. A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(1), 185–207 (2013).
23. T. Chen et al., "DISC: deep image saliency computing via progressive representation learning," *IEEE Trans. Neural Networks Learn. Syst.* **27**(6), 1135–1149 (2016).
24. X. Li et al., "DeepSaliency: multi-task deep neural network model for salient object detection," *IEEE Trans. Image Process.* **25**(8), 3919–3930 (2016).
25. G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *IEEE Conf. on Computer Vision and Pattern Recognition* (2016).
26. N. Liu et al., "Predicting eye fixations using convolutional neural networks," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 362–370 (2015).
27. R. Achanta et al., "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(11), 2274–2282 (2012).
28. D. B. Johnson, "Efficient algorithms for shortest paths in sparse networks," *J. ACM* **24**(1), 1–13 (1977).
29. D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(5), 603–619 (2002).
30. M. Kristan, A. Leonardis, and D. Skocaj, "Multivariate online kernel density estimation with Gaussian kernels," *Pattern Recognit.* **44**, 2630–2642 (2011).
31. O. Le Meur and Z. Liu, "Saliency aggregation: does unity make strength?" in *Asian Conf. on Computer Vision*, pp. 18–32 (2014).
32. S. Alpert et al., "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1–8 (2007).
33. J. Liu and S. Wang, "Salient region detection via simple local and global contrast representation," *Neurocomputing* **147**(5), 435–443 (2015).
34. F. Perazzi et al., "Saliency filters: contrast based salient region detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 733–740 (2012).
35. R. Achanta et al., "Frequency tuned salient region detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1597–1604 (2009).
36. Y. Fang et al., "Visual attention modeling for stereoscopic video: a benchmark and computational model," *IEEE Trans. Image Process.* **26**(10), 4684–4696 (2017).
37. Y. Fang et al., "Video saliency incorporating spatiotemporal cues and uncertainty weighting," *IEEE Trans. Image Process.* **23**(9), 3910–3921 (2014).

Kanghan Oh received his BS degree in computer science from Honam University, Korea, in 2010 and his PhD in electronic and computer engineering from Chonnam National University, Korea, in 2017. Currently, he is a postdoctoral researcher of Division of Electronics and Computer Engineering at Chonbuk National University, Korea. His research interests are object detection, neuroimaging, and document image processing.

Kwanjong You graduated from Chosun University, Department of Mechanical Engineering, in February 1988. In February 1991, he graduated from Inha University with a master's degree in energy engineering. In August 2005, he received his PhD in engineering from Mokpo National University Graduate School of Engineering. Currently, he teaches as a professor at Chosun University Future Society Convergence University.