Computer Science and Information Systems 18(2):379–399 https://doi.org/10.2298/CSIS200120001K

A Novel Distant Target Region Detection Method Using Hybrid Saliency-Based Attention Model under Complex Textures

Jaepil Ko[1] and Kyung Joo Cheoi[2, *]

[1]Department of Computer Engineering, Kumoh National Institute of Technoligy,
Daehak-ro 61, Gumi-si, Gyeongbuk, 39177 Korea
nonzero@kumoh.ac.kr
[2]Department of Computer Science, Chungbuk National University,
Chungdae-ro 1, Seowon-gu, Cheongju-si, Chungbuk, 28644 Korea
kjcheoi@chungbuk.ac.kr

Abstract. In this paper, a hybrid visual attention model to effectively detect a distant target is proposed. The model employs the human visual attention mechanism and consists of two models, the training model, and the detection model. In the training model, some of the features are selected to train in the process of extracting and combining the early visual features from the training image of the target by bottom-up manner, and these features are trained and accumulated as trained data. When the image containing the target is input into the detection model, a task of selectively promoting only features of the target using pre-trained data is performed. As a result, the desired target is detected through the saliency map created as a result of the feature combination. The model has been tested on various images, and the experimental results demonstrate that the proposed model detected the target more accurately and faster than other previous models.

Keywords: target, hybrid, saliency, attention

1. Introduction

So far, lots of various studies on target detection have been conducted. Many of the studies on target detection have shown attempts to use human visual attention mechanisms for target detection [1, 2].In particular, in many intelligent robotic systems developed to assist humans [3], selecting only information useful for the current task from a large number of image information is very important in terms of efficient use of information as well as computational efficiency. To select such useful information, a various method has been proposed to increase the efficiency of computation by adaptively selecting only the high-priority information related to the current work from the numerous information of the input image using the selective visual attention mechanism of a human being. The mechanism of selective visual attention has long been first studied by cognitive scientists and has attracted a lot of interest in the computer vision area because this capability helps find conspicuous objects or regions.

[*] Corresponding author

Treisman suggested that the attention region is determined by integrating various features in the input image. Based on this feature integration theory [4], visual attention studies have been done largely into two approaches, a bottom-up approach, and a top-down approach. The first computational model for human visual attention was the model presented in [4] and was inspired by the feature integration theory [4]. Numerous successful bottom-up computational models have been developed based on this model [6-10]. This visual attention model has been expanded to be used for videos using motion information which is one of the temporal features, as a bottom-up clue [19-22]. In addition, a model using depth information was proposed to apply to the 3D image [23-26]. The main differences between the most successful computational models proposed so far are in methods of extracting features and generating the saliency map [27]. The word 'Saliency' means the intensity of a feature of an input image and can be described as a difference between a pixel and its surrounding neighborhood [1]. An area with high saliency is an area attracting attention.

Many successful bottom-up visual attention models have been proposed, but these models have a limitation that the attention region in the image does not always match the desired target. Detecting conspicuous objects or regions still remains a difficult issue. Therefore, a number of studies based on the top-down approach have been conducted to solve these kinds of issues. The top-down approach is based on the fact that a human pays attention to an object or part of an image that he or she already knows before another object [28]. This approach extracts and combines the features in the same process as the bottom-up approach, but makes the objects that want to be more noticeable by applying high-level knowledge to the extracted basic features [29]. High-level knowledge used here is feature information such as color, shape, and intensity learned about the object and can be obtained by human learning function. This learning knowledge simplifies information about objects according to certain rules. For example, consider the case that we are learning about a beverage can. If the cans of beverages are red and blue, and red has a relatively higher visual attention than blue, we do not learn both blue and red, but rather simplify them to red. Actually, Coca-Cola cans are mixed with white and red, but we just cognize that “Coca-Cola” cans are red because of these reasons.

Although a number of models have been proposed, the fast and accurate detection of salient regions remains a challenge in target detection, particularly in cases of complex textures. In this paper, a hybrid visual attention model for distant target detection that is robust to the color environment is proposed. The model proposed here is designed to extend the capabilities of the previous model that uses high-level information on the target. It overcomes the limitations of the bottom-up visual attention approach, which does not accurately detect the desired object, and the limitation of the top-down visual attention approach, which has only the feature values related to the object to be detected.

This paper is organized as follows. In Section 2, previous studies of target detection using top-down information that shares the basic frame with the proposed model are presented, and the proposed model is presented in Section 3. In Section 4, experimental results and discussions were described to evaluate the performance of the model. Finally, conclusions were drawn with some general observations and recommendations for ongoing work in Section 5.

2. Related Works

Saliency detection technique using a visual attention mechanism is widely used in the fields of target detection [30, 31]. However, as mentioned in the previous section, the bottom-up model often does not find the desired target. Most conventional detecting models are based on training to detect a specific target by the difference of saliency in the local context of image [1].

The model in [8], a top-down attention model that searches and learns the optimal set of linear map weights for a given object in an image was proposed. The model in [32] used mixture information of bottom-up and top-down information, but this model has a limitation from the fact that these two kinds of information were combined with a fixed weight. In addition, if a shape feature (eg, a circle shape) is selected as top-down information, other objects of the same shape are difficult to find. In the model of [33], features were extracted by the method in [6] to search for the object, and the values of the most prominent part of each of the 42 feature maps of the training image were learned with the naive Bayesian network. When detecting a specific object, multiple features maps were filtered by applying the pre-learned subbands, and then center-surround operator was applied to feature maps to enhance the region which is much different to surroundings. And these multiple maps were simply multiplied to make one single saliency map. This model has a limitation that all values other than the learned feature values are lost and that the relationship between features cannot be maintained because all feature maps are multiplied when generating the saliency map. In [34], a top-down attention model for robot navigation was proposed. This model calculates the robot's position by learning Gist features and landmarks. The biggest limitation of this model is that it is practically impossible to apply because it has to actually visit the place and learn. In [35], a top-down attention model that can be applied to finding a person's face was proposed, but it needs to select the color of the clothes manually whenever a top-down saliency map is generated. The model in [36] extracted the attention area by inputting top-down information to the model in [6]. In this model, color, direction, and shape features were extracted, the candidate areas were compared to the similarity of the target, and the candidate areas were weighted according to the results of the similarity comparison. This model does not use learning processing, and only the comparison of one specified object by specifying the search object area in the image. Xiao [37] stored basic features such as lines, points, and circles that make up a target and extracted a target from the input image using them, but it has limitations in that the standard of features for representing objects is not accurately presented.

Recently, deep learning has been successfully utilized in target detection. Automatic feature extraction methods with convolutional neural networks combined with transfer learning achieved top level performance on saliency estimation [38-47] have been proposed. CNN (Convolutional Neural Networks) capture typical high-level features to detect salient objects that are prominent in a particular size and category and achieved an advanced performance for saliency detection issues. Though these deep learning models have shown preferable results, they extract features on special levels, and all levels of information are significant. It also needs a supervised learning process. The existing problem is how to choose the network layers, although each network layer is significant, nevertheless, full convolution network layers increase workload. The other problem is what kind of low-level features to integrate and how to integrate all level features into multiple resolutions. Moreover, these models do not produce a temporal sequence of eye movements, which can be very important not only in developing a system that deals with video streams, but also in understanding human vision.

Although lots of models have been proposed, the fast and accurate detection of salient regions remains a challenge in target detection, particularly in cases of complex textures. Summarizing previous studies that explored objects using the top-down visual attention approach discussed so far, there are some limitations in the method of saliency map generation and performance evaluation. First, in most models, weight, color, intensity, and shape feature maps are integrated with the same weighted without considering the relative differences between features. In this case, if a specific feature value of the search object is changed in the search image, the search becomes difficult. Let's suppose that a red can has been trained, but when the model needs to detect a darker red can. When the top-down information is input, the intensity and shape features become more prominent than the color feature of the red can. Therefore, if there are cans of different colors of the same size in the search image if the intensity and shape features are the same, the color features are pushed out to other colors, so that the red cans cannot be easily found. Second, in the previous model, a saliency map was made by selecting only some feature maps from the extracted feature map such as intensity and color. Since humansHumans take all of the various features into account when paying attention, the method of considering only some features does not fully mimic human visual processing. Third, the kind of experimental images used in the performance evaluation experiments has been very monotonous. If the feature values of the training objects are similar when learning, other objects other than the target to be searched for may be found. Performance evaluation experiments were performed in many models using only experimental images contained only one trained object. Therefore, the performance of the interference between the trained objectscannot be evaluated.

In this paper, a model with the following characteristics is proposed in order to overcome the above-mentioned limitations of the existing model. To overcome the first described limitations, trained data is generated by training the relative differences between features. This trained data can be used in a detection model to detect targets more efficiently. To overcome the second described limitations, a saliency map that takes into account all the features is made. This eliminates the loss of features, allowing the proposed model to handle a wider variety of information, and also allow the model to have a structure similar to the human visual process. To overcome the third described limitation, experiments were performed on images containing a number of trained objects. Through these experiments, the proposed model can be shown to detect the desired target well in the image containing other trained objects.

3. The Methodology

The proposed model was developed to effectively detect a distant target using the top-down information by expanding the bottom-up visual attention model proposed in [27]. To detect targets in the proposed model, the training process is required before searching for targets. Pre-trained data causes bias in the feature extraction and combining phases to find the desired target.

The proposed model consists of the training model and detection model as shown in Fig. 1. In the training model, training images for the target are input and processed to generate trained data. The training model has two processes, feature extraction, and training. Such early visual features such as color, intensity, and form-orientation are extracted from the input image, and information of each feature is selectively selected and trained, and then trained data is generated. For training, the naive Bayesian network was used. In the detection model, a target is detected using the trained data from an image containing the target to be detected. In the detection model, the early visual feature extraction and saliency map generation process goes through the same process as that of the training model. In the training model, the feature values for training are newly calculated and selected during this process, but in the detection model, weights that are calculated using pre-trained data are given to feature maps.

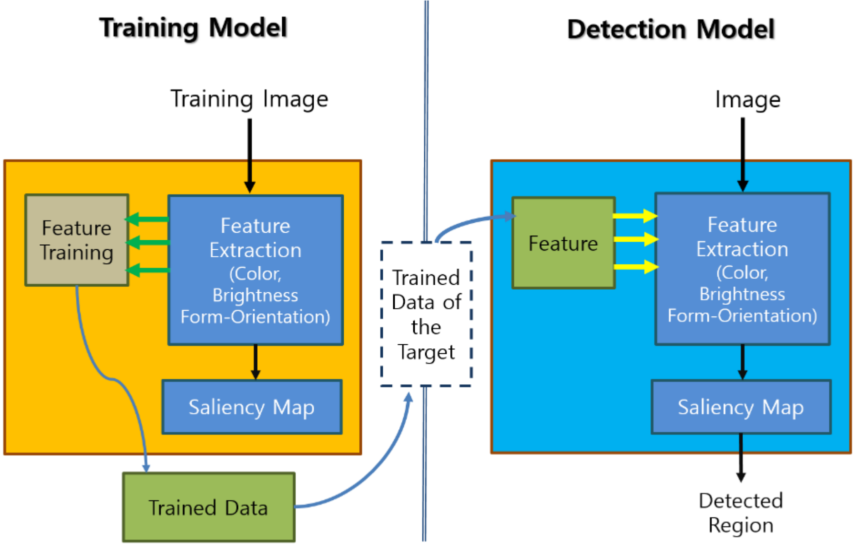
Training Model
Detection Model
Training Image
Image
Feature Training
Feature Extraction (Color, Brightness Form-Orientation)
Saliency Map
Trained Data of the Target
Feature
Feature Extraction (Color, Brightness Form-Orientation)
Saliency Map
Trained Data
Detected Region

Fig. 1. The overall process of the proposed model

3.1. Training Model

In the training model, the general features of the training image of the target are extracted, and the unique attributes of the target are trained.

Feature extraction and saliency map generation

The basic bottom-up process of extracting the early visual features and weighting them together to produce the saliency map proceeds in the same way as in the model of [27]. In the proposed training model, some of the features extracted in this process are selected and used for training. The overall process of extracting early visual features and integrating them into a saliency map by weighting them is shown in Fig. 2.Feature extraction and saliency map generation process in “Training” component can be described detail in 5 steps as follows.

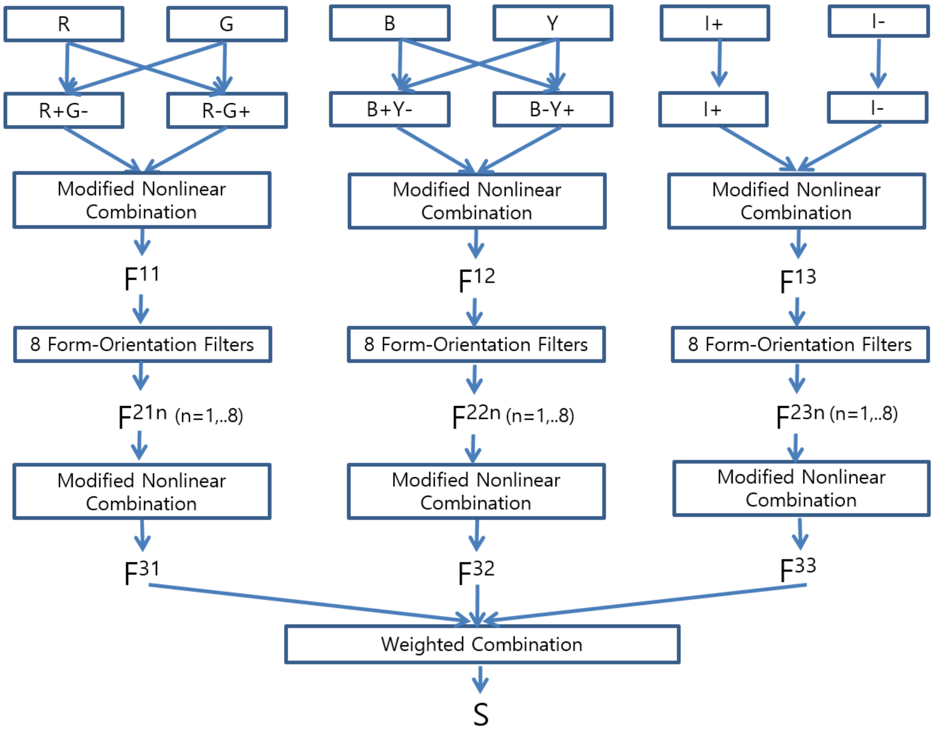
R
G
B
Y
I+
I-
R+G-
R-G+
B+Y-
B-Y+
I+
I-
Modified Nonlinear Combination
Modified Nonlinear Combination
Modified Nonlinear Combination
F11
F12
F13
8 Form-Orientation Filters
8 Form-Orientation Filters
8 Form-Orientation Filters
F21n (n=1,..8)
F22n (n=1,..8)
F23n (n=1,..8)
Modified Nonlinear Combination
Modified Nonlinear Combination
Modified Nonlinear Combination
F31
F32
F33
Weighted Combination
S

Fig. 2. Features extracted from the proposed training model

First, early visual features, R(red), G(green), B(blue), Y(yellow), I+(ON-intensity) and I-(OFF-intensity) are extracted by equation (1) from input image.

$$\begin{gathered} R = r - \frac{g+b}{2}, \quad G = g - \frac{r+g}{2}, \;\; B = b - \frac{r+g}{2} \\ Y = r + g - (2(|r-g| + b), \;\; I+ \;\; = \frac{r+g+b}{3}, \quad I- \;= 1 - \frac{r+g+b}{3} \end{gathered} \tag{1}$$

In equation (1), R is the red channel of the image, G is the green channel, and B is the blue channel. Since the intensity features seen by the human eye may have a high saliency either on the bright part of the image or on the dark part on the contrary, ON intensity features (I+) with high feature values for bright parts and OFF intensity features (I-) with high feature values for dark parts are generated.

Second, early visual features are then reorganized into R+G-, R-G+, B+Y-, B-Y+, which are relative color pairs for R/G, B/Y colors extracted based on an opponent-process theory of color vision [48]. These opposite features are generated by equation (2).

$$R + G -= R - G, R - G += G - R, \;\; B + Y -= B - Y, B - Y += Y - B \tag{2}$$

Through an improved nonlinear combining method shown in equation (3), R+G- and R-G+ are combined into F^{11}, and B+Y- and B-Y+ are combined into F^{12}, and I+ and I-features are combined into F^{13}.

$$
\begin{gathered}
F_{x,y}^{k} = \frac{F_{x,y}^{k} - MinF}{MaxF - MinF}, \\
F_{x,y}^{k} = F_{x,y}^{k} \times Diff(F^{k}), \qquad Diff(F^{k}) = (MaxF^{k} - AveLF^{k})^{2} \\
MaxF^{k} = \max\left(F_{x,y}^{k}\right), AveLF^{k} = average\left[\text{local} \max\left(F_{x,y}^{k}\right)\right], \\
MaxF = \max\left(F_{x,y}^{1}, \ldots, F_{x,y}^{k}\right), MinF = \min\left(F_{x,y}^{1}, \ldots, F_{x,y}^{k}\right)
\end{gathered} \tag{3}
$$

In equation (3), k is the number of imput map, and *Diff*(F^{k}) is the relatvive activity value of F^{k}. At this point, 'activity' is a unit that indicates the intensity of attention of the input image, which means that the higher the amount of activity, the more noticeable features are included than the surroundings.

Third, three form-orientation feature maps (F^{21} for R/G color, F^{22} for B/Y color, F^{23} for intensity) are generated by extracting form-orientation features (F^{21n}, F^{22n}, F^{23n}, n=1~8) that have eight orientations from the extracted two color features (F^{11}, F^{12}) and intensity feature (F^{13}) respectively. Center-surround computations with 8 orientations (0π/8, 1 π /8, … 7π/8) [27] are used as form-orientation filter.and combining them using an improved nonlinear combining method.

Fourth, achieved form-orientations are combined by nonlinear combination method into of F^{31}, F^{32}, and F^{33}.

Finally, saliency map (S) is generated by weighted combination method by equation (4). In each process, all input images are normalized between 0 and 1.In equation (5) k is the number of feature maps, W^{k} is the weight of feature map, and *Diff*(F^{k}) is the relative activity value of F^{k}.

$$
\begin{gathered}
S = W_{1} \times F^{31} + W_{2} \times F^{32} + W_{3} \times F^{33}, \\
W_{k} = \frac{\sum_{i=1}^{n} Diff(F^{i})}{\sum_{i=1}^{n} Diff(F^{i}) - Diff(F^{k})}
\end{gathered} \tag{4}
$$

Selected feature values for training

In the proposed training model, various feature values were involved in training to learn the unique features of the object. The feature values in the various feature maps generated from the training image and the relative feature values between the feature maps were all considered, and the following four types of feature values shown in Table 1 were calculated and participated in the training.

Table 1. Selected feature values for training

Selected feature value	Reason
R, G, B, Y, I +, I- feature map values corresponding to maximum saliency location of F^{11}, F^{12}, F^{13} feature map	To find out the intensities of the unique colors of the target.
Relative activity amount of each pair of R+ G-, R-G+, B+Y-, B-Y+, I+, and I- feature maps	To find out which features of the input map were reflected in creating F^{11}, F^{12}, and F^{13} feature maps.
Relative activity amount of each of F^{21n}, F^{22n}, F^{23n} feature maps.	To find out which features of the input map were reflected in creating F^{21}, F^{22}, F^{23} feature maps.
Weights given to the F^{21}, F^{22}, F^{23} feature maps used when generating the saliency map (S)	To find out which features of the input map were reflected in creating a saliency map.

Fig. 3 shows the example of a training feature selection mechanism considering the features in the various feature maps and Fig. 4. shows the example of a training feature selection mechanism considering the features between the various feature maps. In Fig. 3, R, G feature map values corresponding to maximum saliency location of F^{11} feature map were selected, and in Fig. 4, the relative activity amount of orientation feature maps of F^{21n} was selected.

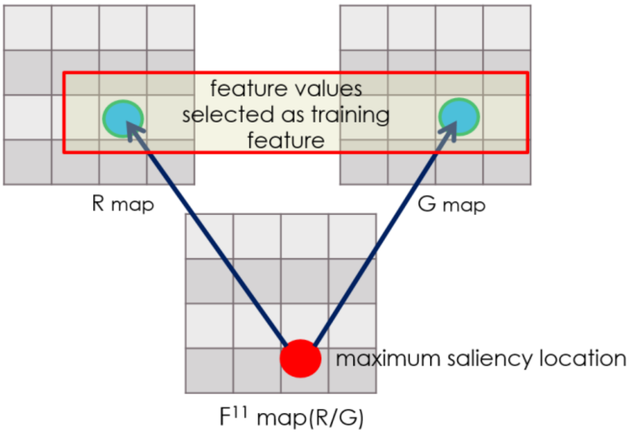

Fig. 3. The example of a training feature selection mechanism considering the features in the various feature maps

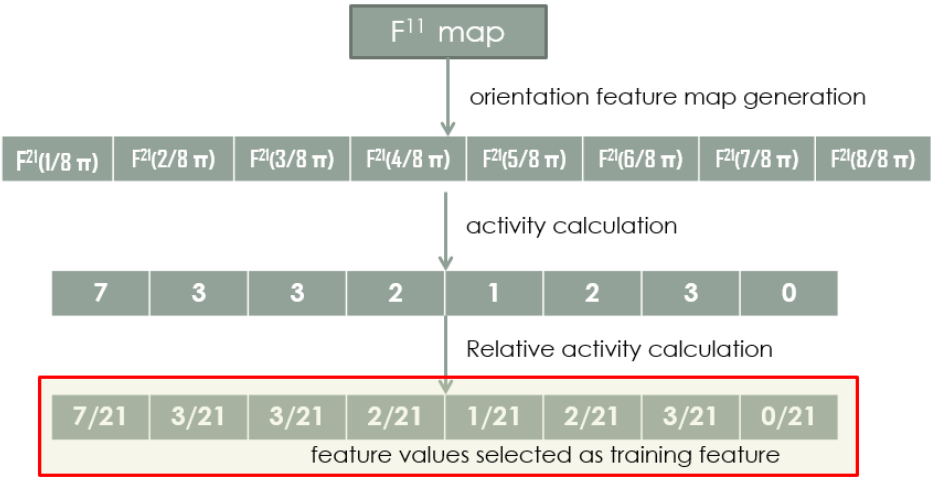

Fig. 4. The example of a training feature selection mechanism considering the features between the various feature maps

Trained Data

The features extracted and selected from each category's training images were stored as mean($\boldsymbol{\mu}$) and standard deviation($\boldsymbol{\sigma^2}$) as shown in equation (5) through naive bayesian. In equation (1), is the value of input data and is the number of data input. In this way, a plurality of images is trained to construct trained data. The trained data is represented by a probability equation as shown in equation (6), where $N()$ is a normal distribution curve, $p()$ is a probability, and n is the total number of features to be trained. That is, the probability that the value '$\boldsymbol{\theta}$' emerges from the feature 'F' is the result value when 'theta' is input into a normal distribution curve composed of the mean($\boldsymbol{\mu}$) and standard deviation($\boldsymbol{\sigma^2}$) of the trained '*F*'.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i\,, \qquad \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu)^2. \tag{5}$$

$$\prod_{j=1}^{n} p(F_j \,|\, \theta_j) \propto N\big(F_j\,;\mu_j\,;\sigma_j\big). \tag{6}$$

3.2. Detection Model

In the detection model, trained data is used to detect the target object. The early visual feature extraction and saliency map generation process goes through the same process as that of the training model. In the training model, the feature values for training are newly calculated and selected during this process, but in the detection model, weights that are calculated using pre-trained data are given to feature maps.

Biasing the feature values of R, G, B, Y, I+, I-

In order to make the region corresponding to target in the extracted R, G, B, Y, I +, I- feature maps to have high feature values, high weight is assigned to each feature map that is similar to the maximum feature values of R, G, B, Y, I+, and I- of the trained data. The weighting method is as follows. First, a Gaussian distribution curve is drawn using the mean ($\boldsymbol{\mu}$) and standard deviation ($\boldsymbol{\sigma^2}$) of the trained data. And then the values of the feature map are increased by the result value of passing each value of the feature map through a gaussian distribution curve sd shown in equation (7). In equation (7), x means the feature values in the feature map, and y means the value of the results when the x was inputted to the gaussian distribution curve. Feature value x is modified by adding the result value y with x itself. This adjusts the value of the feature similar to the target to a higher value.

$$y = gaussian(x), \qquad x = x + y \tag{7}$$

At this time, the value of the curve is adjusted to match the maximum value of the y-axis of the gaussian distribution curve with the maximum value of the feature map so that the feature values similar to the trained data in the output feature map should be the maximum feature values.

Biasing in nonlinear combination process for generation of F^{11}, F^{12}, and F^{13}

When top-down information is input to the R, G, B, Y, I+, and I- feature maps, F^{11}, F^{12}, and F^{13} are generated with an improved nonlinear combination. In the nonlinear combination process, additional activity amount of trained data is input to select the salient feature map. Fig, 5 shows the basic mechanism of biasing activity values of feature maps. In Fig, 5, the biggest of the calculated activity values of the input maps is 15 of the first map, but in trained data, the third of the input maps has the largest value of 9. So, it is biased by the value of the third map of the input maps.

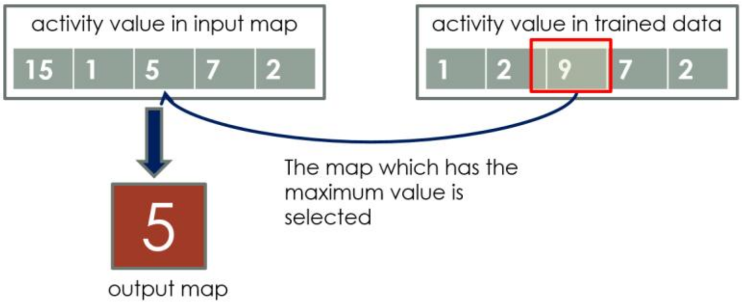
activity value in input map
15
1
5
7
2
activity value in trained data
1
2
9
7
2
The map which has the maximum value is selected
5
output map

Fig. 5. The basic mechanism of biasing activity values of feature maps

The nonlinear combination reflecting the top-down information can be expressed as equation (8). In equation (8), *C* means input map, *CN* means a map with assigned weight, and T_a means activity amount of *C* map in the trained data.

$$CN_{x,y}^{k} = C_{x,y}^{k} \times T_{a}^{k}. \tag{8}$$

Biasing in nonlinear combination process for generation of F^{31}, F^{32}, and F^{33}

The eight form-orientation feature maps F^{21n}, F^{22n}, $F^{23n}$are generated for each color and intensity feature by performing a center-surround operation that mimics the cellular reactivity seen in the "ON-centered, OFF-surround" receptive field of humans [5]. The eight form-orientation feature maps are integrated into the orientation with the largest response because one feature was divided into eight orientations. At this time, in the nonlinear combining process, the top-down information is input to the form-orientation activity of each direction in the trained data as the top-down information to generate an output map having the prominent direction of the target.

Biasing in weighted combination process for generation of the saliency map

When weighted combining is performed, the weight of trained data is added as top-down information to generate a saliency map that has the relationship between the color and intensity of the target. The saliency map is generated by equation (9) and L_1, L_2,L_3 is the weight corresponding to each map in the trained data. The maximum salient location of the saliency map generated by the weighted combination is the portion that matches the target.

$$S = L_1 \times F^{31} + L_2 \times F^{32} + L_3 \times F^{33}. \tag{9}$$

4. Experiments and the results

To evaluate the performance of the model, the model was applied to the problem of detecting targets, such as a pen, triangle-shaped safety sign, and beverage can. In addition, to carry out a quantitative evaluation, the proposed model was compared with the model in [5], [7], and [26].

Most models that use top-down information are difficult to compare because of the lack of training and test images. However, because the model of [8] provides both training image, test image, and general evaluation criteria, it is possible to compare the performance with the proposed model. The proposed model was also compared with previous bottom-up models in [6] and [27]. The model of [6] is the most frequently referred to in the bottom-up model studies, and it was selected for the performance comparison with other models in the future. The model in [27] was selected because it has the same feature extraction method as the proposed model, but not use top-down information.

4.1. Training images

For 3 types of detection experiments and robustness experiments of color and brightness, images of red, blue, and black pens with strong color and orientation features commonly found in the real-world were created as training images. In addition, 32 triangle-shaped safety signs and 45 red beverage cans taken at random campus locations and times were used as training images. The training image was made by 8 angular changes for one training image for one object, and 12 brightness changes and 10 step size changes for each angular change. The brightness change was changed in 5 steps from -30 to +30, and the change in size was changed in 0.2 steps from 1.2 to 3.0 times. In summary, as shown in Table 2, a total of 176 training images was created for the training image for one object.

Table 2. Number of newly Selected feature values for training

composition of training images for each object	variation1		variation2		Total No.
	rotation	8	brightness	12	96
			size	10	80
					176

4.2. Test images

Test images used in specific pen detection experiments. Test images used in the experiment for detecting a specific pen is a scene containing the one, two, and three red, blue, and black pens that are trained as shown in Table 3. A total of 147 test images were used. Test images used in triangular-shaped safety signs and beverage can detection experiments. 32 triangular-shaped safety signs and 59 red cans were used as test images in triangular-shaped safety signs and beverage can detection experiments. An example of the test image used in the detection experiment is shown in Fig. 6.

Table 3. Test image used in specific pen detection experiments

	1 trained object	2 trained object	3 trained object	Total
Total	63	63	21	147

Test images used in color robustness experiments

The robustness test for color is to see how the color of distractors and the background of the scene with the object effects the target detection. The images with a red, blue, gray, sky blue, and brown background and the images with red, blue, and green distractors among images used in specific pen detection experiments were used as the Test images used in color robustness experiments. A total of 100 images were used.

Test images used in brightness robustness experiments

The robustness test for brightness is to check the effect of brightness change on target detection. 294 images with brightness variation from +30 to –30 applied to the original test images used in specific pen detection experiments were used as test images in brightness robustness experiments.

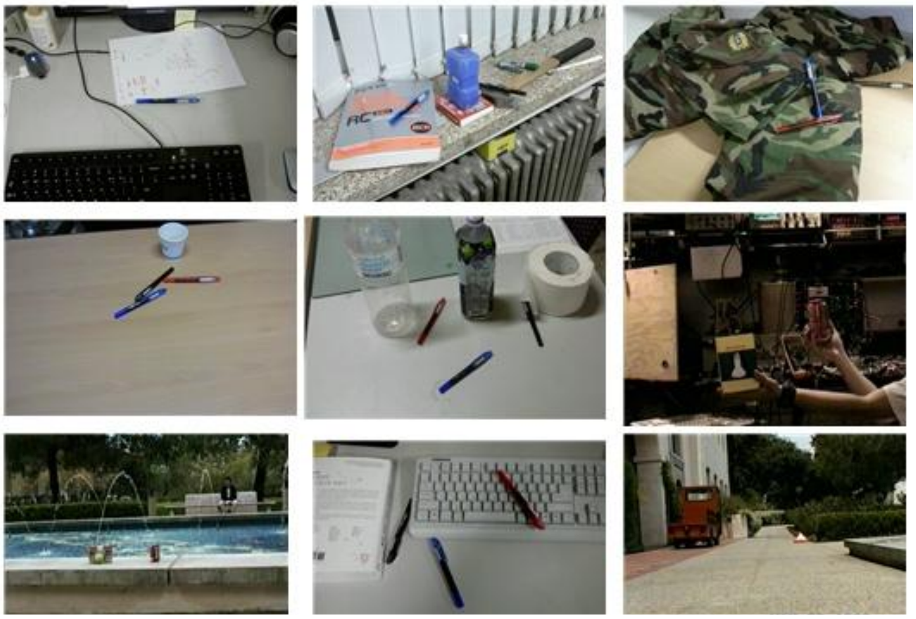

Fig. 6. An example of the test image used in the detection experiment

4.3. Performance Evaluation Criteria

Two performance evaluation criteria, i) The successful target detection rate (p) using equation (10) for the images detected within the specified number of times in the test image, ii) the average viewpoint variation number until the target found, were used.

$$p = \frac{r}{r + q} \tag{10}$$

In equation (10), r is the number of images that successfully detected the target, q is the number of images that failed at detecting the target. If a model detects a target in 3 images out of 5 test images, the successful target detection rate (p) is 0.6. Among successfully detected images, if this model detected the target in the third search from the first image, the third search from the second image, the second-search in the third image, then the average viewpoint variation number until target found is 2.6. The successful target detection rate (p) can be used to evaluate the detection efficiency in the entire test image, and the average viewpoint variation number until target found can be used to evaluate how fast the model detected the target. The higher the successful target detection rate, the lower the viewpoint variation number until the target found, the better the performance.

4.4. Results

Fig. 7 shows an example of the result of a saliency map of the proposed model using top-down information and the result that is not. The target to be detected in the input image is a blue pen. Without using trained data, with only bottom-up features, the proposed model detects the red pen at first-search. The saliency map on the left side of Fig. 7 shows that the red pen is the most salient object. But what if we want to detect a blue pen, not a red pen? The proposed model allows you to find blue pens at once. The saliency map on the right side of Fig. 7 was made using trained data, showing that the blue pen, not the red one, is the most salient object.

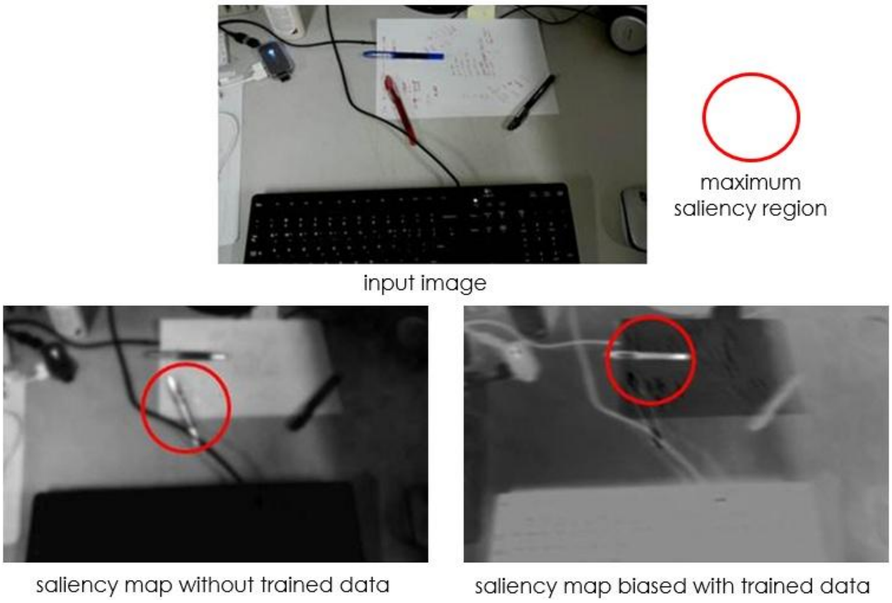
maximum saliency region
input image
saliency map without trained data
saliency map biased with trained data

Fig. 7. Example of the result of the saliency map without top-down information (left) and the result with top-down information (right)

Fig. 8 demonstrates the example of detection results for 3 types of target detection experiments in the proposed method. The three images on the left side of Fig. 8 are examples of results from specific pen detection experiments. The image on the first line is an image where two trained objects (red and blue pens) are placed on a bookshelf with several books. The target is a blue pen. The proposed model detected the target after the second-search but the other models failed at detecting the target within the specified number of search-times. The image in the second row on the left of Fig. 8 shows a yellow alarm clock on a light brown table, two trained objects (a red pen and a blue pen), and a gray digital clock. The target is a blue pen. The proposed model and the models of [6] and [8] detected the target at once, but the model of [27] failed at detecting the target within the specified number of search-times. The image in the third row on the left of Fig. 8 is a test image with a military uniform on a brown table and three trained objects (red pen, blue pen, and black pen) placed on the uniform. The target is a red pen. The proposed model detected the target at once, but the other models failed at detecting the target within the specified number of search-times.

The three images in the middle of Fig. 8 are examples of the results from triangular-shaped safety sign detection experiments, and the three images on the right are examples of the results from beverage can detection experiments. We can see that the proposed model detects the target in various environments in the first-search.

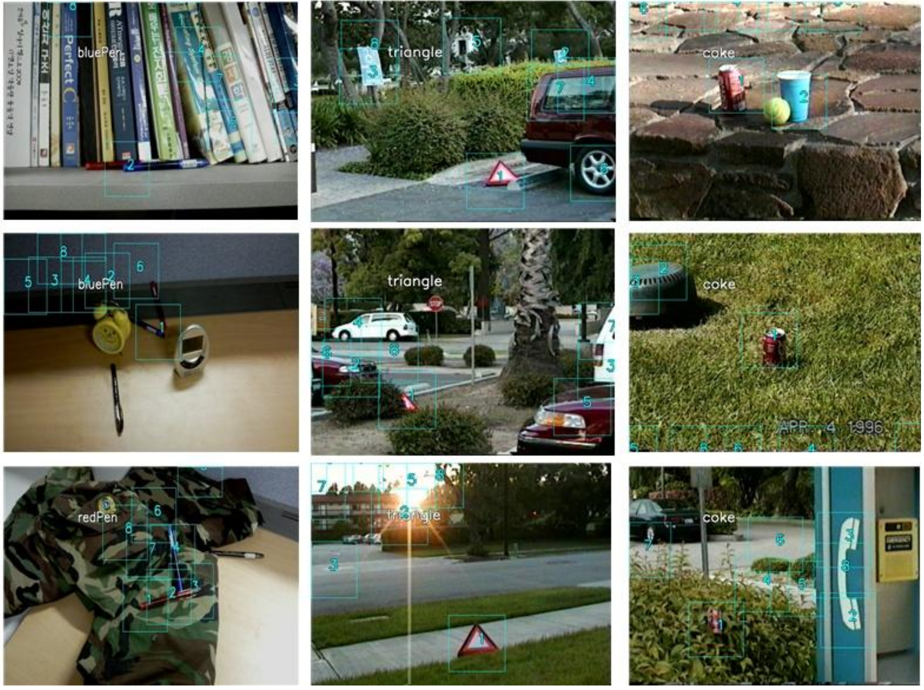

Fig. 8. Example results on 3 types of detection experiments: pen detection (left), triangular-shaped safety sign detection (middle), beverage can detection (right)

Table 4 and Table 5 summarizes the performance evaluation result of the proposed model with the comparison result of 3 models. Overall, the proposed model achieved the successful target detection rate with an average of 94.33% on 3 types of target detection experiments, and 97% on color and brightness robustness experiments. An average viewpoint variation number until the target found from the proposed model is 1.37 on 3 types of target detection experiments, and 1.33 on color and brightness robustness experiments. And these results are the best performance among other comparison models.

The proposed model detected the target almost at first-search, but other models didn’t. The model in [8] has the best performance except for the proposed model, and it detected the target on average 1.67th search on 3 types of target detection experiments and 1.87th search on color and brightness robustness experiments. However, the proposed model has better performance with 1.37 on 3 types of target detection and 1.33 on color and brightness robustness experiments. Even the successful target detection rate is superior to that of [8].

These results demonstrate that the proposed model outperforms other previous models. The color robustness test and the brightness robustness test were experiments to confirm how robust the proposed model was for the brightness variation and the disturbing color. Previous models were difficult to secure versatility only by experimenting on the images of the monotonous environment. The color robustness test results and the brightness robustness test results demonstrate that the proposed model is very robust to color and brightness. In particular, even if the background color of the test image is similar to the target and there are some distractors, the performance of the model did not deteriorate much.

Table 4. Overall results on 3 types of target detection experiments

	The successful target detection rate				Viewpoint variation number until the target found			
	This Model	[8]	[27]	[6]	This Model	[8]	[27]	[6]
Pen	89	75	67	60	1.21	1.8	3.57	4.31
Safty sign	99	90	79	75	1.59	1.8	4.02	5.6
Beverage can	95	89	82	67	1.31	1.35	5.3	6.5
average	94.33	84.67	76	67.33	1.37	1.67	4.3	5.47

Table 5. Results on color and brightness robustness experiments

		The successful target detection rate				viewpoint variation number until the target found			
		This Model	[8]	[27]	[6]	This Model	[8]	[27]	[6]
Color	Red background	100	100	92	75	1	1.25	2.5	2.13
	Bluebackground	100	100	83	50	1	1.40	2.2	2.80
	Gray background	100	100	75	67	1	1.14	3.0	2.28
	Sky blue Background	100	92	83	83	1	2.25	2.38	2.25
	Brown background	100	92	75	75	1	1.43	2.5	2.29
	Red distractor	100	100	83	75	1.14	2.14	3.71	3.29
	Blue distractor	100	100	100	100	1.67	1.83	4.08	3.42
	Green distractor	100	100	92	83	1.11	1.67	2.5	3.11
Bright-ness	+30	89	75	65	63	2.06	2.9	4.15	4.12
	-30	91	73	60	61	2.27	2.73	4.95	4.19
average		97	93.2	80.8	73.2	1.33	1.87	3.20	2.99

5. Conclusions

In this paper, a robust distant target detection model employing the human visual attention mechanism was proposed. The proposed model consisted of a training model for training top-down information and a searching model for detecting targets using top-down information. In the training model, some primitive features of training objects extracted in a bottom-up manner were selected and trained. The detection model biases feature maps and adjust the saliency map to detect desired targets. In the process of detecting, the bottom-up saliency and top-down saliency were both considered. The proposed model trained the relationship information between color, intensity, form-orientation features in order to overcome the limitations of the previous training model that could not accurately express the features of an object because of the relationship between color and brightness was not considered. Also, all training images with changes under certain criteria were used for training in order to overcome the limitations of the previous model which did not produce reliable results by selectively using the training images. The proposed model detected the target using all features in order to overcome the limitations of previous detection models which used only a few features to detect targets which could cause detection failures due to loss of information.

The entire process of the proposed detection model is similar to that of the training model, but instead of training selected features, top-down information was input and biased. Top-down information is input in three phases: first, the color and brightness values of the trained data were input in the process of generating early visual feature maps of color and intensity. Among the early feature values of the color and brightness feature maps, the values similar to the color and brightness values in the trained data were modified to be more salient. In addition, the activity information of the trained data is used in a nonlinear combination process, and it leads to making features of the target objects more salient. Second, relative activity values of feature maps in the trained data were input in the nonlinear combination process of form-orientation feature maps. The detailed process of the nonlinear combination method was modified by trained data. Finally, the weights of the form-orientation feature maps in the trained data were input in the weighted combination process.

To evaluate the performance of the proposed model, experiments were conducted to apply the proposed model to detection problems such as detecting a specific pen, triangular safety objects, and can. Also, color robustness and brightness robustness experiments were additionally performed. To evaluate the performance of the proposed model, experiments were conducted to apply the proposed model to detection problems such as detecting a specific pen, triangular safety objects, and can. Also, color robustness and brightness robustness experiments were additionally performed. In addition, to carry out a quantitative evaluation, the proposed model was compared with the previous model. The previous model only used the images contains only one trained object so that they failed to evaluate performance considering the effects between different trained objects. In this experiment, to overcome these limitations of the previous model, an image containing several trained objects was used as the experimental image. The color robustness test and the brightness robustness test are experiments to confirm how robust the proposed model is for the brightness variation and the disturbing color. Existing models were difficult to secure versatility only by experimenting on the images of the monotonous environment. These experiments were conducted to confirm whether the proposed model overcomes the limitations of these existing models. Quantitative and qualitative analyses of the experimental results showed that the proposed model outperformed the traditional model in target detection and was comparable to the state-of-the-art models.

References

1. Borji, A. Itti, L.: State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 35, No. 1, 185–207. (2013)
2. Bi, Z., Dou, S., liu, Z., Li, Y.: A Recommendations Model with Multiaspect Awareness and Hierarchical User-Product Attention Mechanisms. Computer Science and Information Systems, Vol. 17, No. 3, 849–865. (2020)
3. Paliwal, N., Vanjani, P., Liu, J., Saini, S., Sharma, A.: Image processing-based intelligent robotic system for assistance of agricultural crops. International Journal of Social and Humanistic Computing. Vol. 3, No. 2, 191-204. (2019)
4. Treisman A.M., Gelad, G.: A Feature-integration Theory of Attention. Cognitive Psychology, Vol. 12, No. 1, 97-136. (1980)
5. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology, Vol. 4, 219-227. (1985)
6. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, 1254–1259. (1998)
7. Itti, L., Koch, C.: Computational modeling of visual attention. Nature Reviews, Neuroscience, Vol. 2, 194–203. (2001)
8 Itti, L., Koch, C.: Feature combination strategies for saliency-based visual attention systems. Journal of Electronic Imaging, Vol. 10, No. 1, 161-169. (2001)
9. Dhavale, N., Itti, L.: Saliency-based multifoveated MPEG Compression, In Proceedings of IEEE International Symposium on Signal Processing and its Applications, Paris, France, France,229-232. (2003)
10. Yun, Z., Shah. M.: Visual Attention Detection in Video Sequences Using Spatiotemporal Cues. In Proceedings of 14th annual ACM international conference on Multimedia, Santa Barbara, CA, USA,815-824. (2006)
11. Torralba, A., Oliva, A., Castelhano, M. S., Henderson, J. M.: Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. Psychological Review, 113, 766–786. (2006)
12. Li, S., Lee, M.: An Efficient Spatiotemporal Attention Model and Its Application to Shot Matching. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 17, No. 10, 1383-1387. (2007)
13. Li, H., Su, X., Wang, J., Kan, H., Han, T., Zeng, Y., Chai, X.: Image processing strategies based on saliency segmentation for object recognition under simulated prosthetic vision. Artificial Intelligence in Medicine, Vol. 84, 64-78. (2018)
14. Li, H., Han, T., Wang, J., Lu, Z., Cao, X., Chen, Y., Li, L., Zhou, C., Chai, X.: A real-time image optimization strategy based on global saliency detection for artificial retinal prostheses, Information Sciences 415–416, 1-18. (2017)
15. Lei, J., Wang, B., Fang, Y., Lin, W., Callet, P.L., Ling, M., Hou, C.: A universal framework for salient object detection. IEEE Transactions on Multimedia, Vol. 18, No. 9, 1783–1795 (2016).
16. Wang, Z., Xiang, D., Hou, S., Wu, F., Background-driven salient object detection. IEEE Transactions on Multimedia, Vol. 19, No. 4, 750–762. (2017).
17. Park, M., Cheoi, K.: Selective Visual Attention System Based on Spatiotemporal Features, Lecture Notes in Computer Science, Vol. 5068, 203-212. (2008)

18. Cheoi, K., Park, M.: Visual Information Selection Mechanism Based on Human Visual Attention. Journal of Korea Multimedia Society, Vol. 14, No. 3, 378-391. (2011)
19. Chen, C., Li, S., Wang, Y., Qin, H., Hao, A. Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion. IEEE Transactions on Image Processing, Vol. 26, No. 7, 3156–3170. (2017).
20. Xi, T., Zhao, W., Wang, H., Lin, W.: Salient object detection with spatiotemporal background priors for video. IEEE Transactions on Image Processing, Vol. 26, No. 7, 3425–3436. (2017).
21. Liu, Z., Li, J., Ye, L., Sun, G., Shen, L. Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 27, No. 12, 2527-2542. (2017).
22. Wang, W., Shen, J., Yang, R., Porikli, F.: A unified spatiotemporal prior based on geodesic distance for video object segmentation. IEEE Transactions on Pattern Analasis and Machine Intelligence, Vol. 40, No. 1, 20–33. (2018).
23. Wang J., Silva M., Callet P., Ricordel V.: Computational model of stereoscopic 3D visual saliency. IEEE Transactions on Image Processing, Vol. 22, No. 6, 2151–2165. (2013)
24. Feng, D., Barnes, N., You, S., McCarthy, C.: Local background enclosure for RGB-D salient object detection. In Proceedings of International Conference on Computer Vision and Pattern Recognition, 2343–2350. (2016).
25. Song, H., Liu, Z., Du, H., Sun, G.: Depth-aware saliency detection using discriminative saliency fusion. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 1626–1630. (2016).
26. Wang, A., Wang, M.: RGB-D salient object detection via minimum barrier distance transform and saliency fusion. IEEE Signal Processing Letters, Vol. 24, No. 5, 663–667. (2017).
27. Cheoi, K., Kim, M.: Adaptive Spatiotemporal Feature Extraction and Dynamic Combining Methods for Selective Visual Attention System. Wireless Pers. Commun. Vol. 98, 3227–3243. (2018).
28. Oliva, A., Torralba, A., Castelhano, M., Henderson, J.: Top-down control of visual attention in object detection. Proceedings of International Conference on Image Processing (pp. 253–256).Barcelona, Catalonia: IEEE Press. (2003).
29. Torralba, A., Oliva, A., Castelhano, M., Henderson, J.: Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. Psychological Review, 113, 766-786. (2006)
30. Borji, A., Cheng, M., Hou, Q., Jiang H., Li, J.: Salient object detection: A survey. Computational Visual Media. Vol. 5, No. 7, 117-250. (2019)
31. Ren, Z., Gao, S., Chia, L., Tsan,g U.: Region-based saliency detection and its application in object recognition. IEEE Transactions on Circuits & Systems for Video Technology, Vol. 24, No. 5, pp. 769-779. (2013)
32. Yu, Y., Mann, G., Gosine, R.: An Object-Based Visual Attention Model for Robotic Applications. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), Vol. 40, No. 5, 1398-1412. (2010)
33. Elazary, L., Itti, L.: A Bayesian model for efficient visual search and recognition. Vision Research, Vol. 50, No. 14, 1338-1352. (2010)
34. Siagian, C., Itti, L.: Biologically Inspired Mobile Robot Vision Localization. IEEE Transactions on Robotics, Vol. 25, No. 4, 861-873. (2009)
35. Lee, K., Buxton, H., Feng, J.: Cue-guided search: a computational model of selective attention. IEEE Transactions on Neural Networks, Vol. 16, No. 4. (2005)
36. Zhang, J. Li, Z., Jingjing, G., Zhixing, L.: A Study of Top-down Visual Attention Model Based on Similarity Distance. In Proceedings of 2nd International Congress on Image and Signal Processing, Tianjin, China, 1-5. (2009)
37. Xiao, J., Cai, C., Ding, M., Zhou, C.: The Application of Novel Target Region Extraction Model Based on Object-accumulated Visual Attention Mechanism. In Proceedings of Fourth International Conference on Natural Computation, Jinan, China, 116-120. (2008)

38. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.: Learning to Detect a Salient Object. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, No. 2, 353–367. (2011)
39. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 5455-5463. (2015)
40. Wang, L., Lu, H., Ruan, X., Yang, M.: Deep networks for saliency detection via local estimation and global search. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 3183-3192. (2015)
41. Li, X., Zhao, L., Wei, L., Yang, M., Wu, F., Zhuang, Y., Ling, H., Wang, J: DeepSaliency: Multi-Task deep neural network model for salient object detection. IEEE Transactions on Image Processing. Vol, 25, No. 8, 3919-3930. (2016)
42. Wang, L., Lu, H., Zhang, P., Ruan, X.: Saliency Detection with Recurrent Fully Convolutional Networks. In European Conference on Computer Vision, 825-841. (2016)
43. Zhang, P., Zhuo, T., Huang, W., Chen, K., Kankanhalli, M.: Online object tracking based on CNN with spatial-temporal saliency guided sampling. Neurocomputing. Vol. 257, 115-127. (2017)
44. Zhang, J., Li, B., Y. Dai, F. Porikli, He, M.: Integrated deep and shallow networks for salient object detection. In Proceedings of IEEE International Conference on Image Processing, 271–276. (2017)
45. Zhang, P., Wang, D., Lu, H., Wang, H., Yin, B: Learning uncertain convolutional features for accurate saliency detection. In Proceedings of Interational Conference on Computer Vision, 212–221. (2017)
46. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan,X.: Learning to detect salient objects with image-level supervision. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 3796–3806. (2017)
47. Zhang, J., Zhang, T., Dai, Y., Harandi, M., Hartley, R.: Deep unsupervised saliency detection: A multiple noisy labeling perspective. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,9029–9038. (2018)
48. Hurvich, L., Jameson, D.: An opponent-process theory of color vision. Psychological Review, Vol. 64, No. 6, 384-404. (1957)

Kyung Joo Cheoi received Ph.D. degrees in Computer Science from Yonsei University, Korea in 2002. During 2002-2005, she worked as a research engineer in TI-specialist-Tech. of LG CNS, Korea. She is currently an associate professor of the Department of Computer Science in Chungbuk National University. Her research interests include computer vision, image processing, and machine learning.

Jaepil Ko received Ph.D. degree in Computer Science from Yonsei University, Korea in 2004. He is currently a faculty member of computer engineering department at Kumoh National Institute of Technology. He has served as a director of the Artificial Intelligence Society of KIISE. His research interests include pattern recognition, computer vision and image processing.

Received: February 10, 2020; Accepted: January 24, 2021.