IEICE TRANS. INF. & SYST., VOL.E98–D, NO.12 DECEMBER 2015

PAPER

Top-Down Visual Attention Estimation Using Spatially Localized Activation Based on Linear Separability of Visual Features

Takatsugu HIRAYAMA[†,††a)], ***Member***, **Toshiya OHIRA**[†b)], ***Nonmember***, ***and*** **Kenji MASE**[†c)], ***Fellow***

SUMMARY Intelligent information systems captivate people's attention. Examples of such systems include driving support vehicles capable of sensing driver state and communication robots capable of interacting with humans. Modeling how people search visual information is indispensable for designing these kinds of systems. In this paper, we focus on human visual attention, which is closely related to visual search behavior. We propose a computational model to estimate human visual attention while carrying out a visual target search task. Existing models estimate visual attention using the ratio between a representative value of visual feature of a target stimulus and that of distractors or background. The models, however, can not often achieve a better performance for difficult search tasks that require a sequentially spotlighting process. For such tasks, the linear separability effect of a visual feature distribution should be considered. Hence, we introduce this effect to spatially localized activation. Concretely, our top-down model estimates target-specific visual attention using Fisher's variance ratio between a visual feature distribution of a local region in the field of view and that of a target stimulus. We confirm the effectiveness of our computational model through a visual search experiment.

key words: *human visual attention, visual search, saliency map, activation map, linear separability*

1. Introduction

Many researchers have focused on human vision to develop advanced information systems that interact with humans because gazing at something implies human cognitive states such as interest and intent. A human-friendly robot requires not only verbal communication but also nonverbal communication such as eye contact and mutual gaze [1]. In order to establish natural joint attention between a person and a robot, the robot should estimate when and on what the person will focus [2]. Intelligent driving support systems should also be able to estimate driver's visual attention. The systems help the driver recognize driving cues such as signboards and guide plates on the roads. The visibility of these objects differs in varying traffic conditions [3] and the driver's cognitive state. The systems are effective if they can estimate the visibility according to the situations and make the driver aware of their locations. Human visual attention is important for designing rich human-computer interaction.

Visual attention is a built-in mechanism of the human visual system and is used to quickly focus one's attention on a region in a visual scene that is most likely to contain objects of interest. Visual attention is classified as either bottom-up or top-down [4]. When only visual stimuli activate visual attention in a scene, this is known as bottom-up processing. In contrast, when people view a scene with intention, such as searching for a target or driving a car, they shift their visual attention in a top-down manner.

In recent years, computing visual saliency and simulating visual attention have attracted much attention in the field of robotics and computer vision. Itti *et al.* proposed a representative computational model of visual saliency [5]. They incorporated a bottom-up computational process into their proposed saliency map model based on the feature integration theory [6] and multi-resolution structure [7]. Other bottom-up visual attention models, many of which are the derivatives of the saliency map model, have been developed by other researchers [8].

On the other hand, computational models of top-down visual attention are still not well studied. However, many psychophysical findings and conceptual models on task-oriented visual attention have been reported [9]. Our research focuses on estimating top-down visual attention activated during visual search tasks. In this paper, we define this attention as target-specific visual attention. We propose a novel computational model based on psychophysical findings of visual search, which estimates target-specific visual attention using spatially localized activation based on linear separability between visual features of a target and the others.

This paper is organized as follows. The next section defines the visual search task treated in this work. Section 3 covers the related work and identifies the problem. Section 4 describes the proposed model to estimate target-specific visual attention. Section 5 reports and discusses experimental results for investigating the effectiveness of the proposed method. Section 6 concludes the paper.

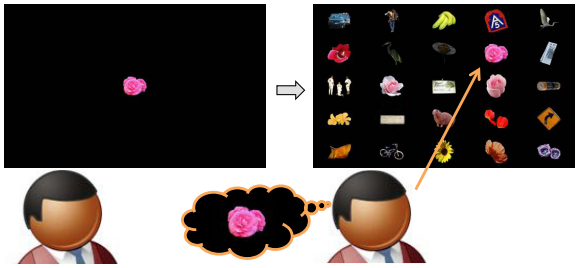
2. Visual Search Task

Conventional psychophysical studies on visual search employ simple geometric images. Recently, some researchers have used natural object images with more complicated textures [10]. According to them, we design a visual search task using natural object images as follows: (1) An ex-

Manuscript received May 7, 2015.
Manuscript revised August 19, 2015.
Manuscript publicized September 10, 2015.

[†]The authors are with the Graduate School of Information Science, Nagoya University, Nagoya-shi, 464–8603 Japan.
[††]The author is with the Graduate Program for Real-World Data Circulation Leaders, Nagoya University, Nagoya-shi, 464–8603 Japan.
a) E-mail: hirayama@is.nagoya-u.ac.jp
b) E-mail: ohira@cmc.ss.is.nagoya-u.ac.jp
c) E-mail: mase@nagoya-u.jp

DOI: 10.1587/transinf.2015EDP7170



(a) Target presentation phase (b) Target search phase

Fig. 1 Example of a pair of images used for the target presentation phase and the target search phase.

perimenter presents a target image as a cue to an experimental participant at the center of the display field for several seconds (Fig. 1 (a)). (2) The experimenter then presents a search image where the target and distractor objects are aligned and asks the participant to search for the target object (Fig. 1 (b)). Note that the target image is the same as the target object included in the search image.

3. Related Work

In this section, we discuss researches related to top-down visual attention in search tasks. Figure 2 outlines a typical computational process that modulates the weights of visual features [4]. Some researchers have proposed the computational models based on this weight modulation process. Elazary *et al.* proposed a Bayesian model called SalBayes, which regards the posterior probability that the visual features extracted from an image region belong to an object class as saliency [10]. The model needs to recognize the object to detect the target by means of maximum a posteriori probability estimation.

For visual search tasks, it is important to consider the relationship between targets and distractors. The guided search model gives more weight to feature channels that uniquely represent the target [11]. The weighted response of each channel to the target is compared with its average response to the distractors. The channel with the greatest positive difference is selected to compute the top-down attention map. The discriminant saliency model proposed by Gao *et al.* [12] compares entropy of each visual feature extracted from training images of a target with that from distractors and then selects the feature channels with positive difference. The model also computes mutual information of class labels, i.e., target and distractor, and the visual feature extracted from a search image when the posterior probability that the visual feature is in the target is larger than that in the distractors. This process is applied to the selected feature channels. Finally, the top-down attention map is computed by accumulating the mutual information.

Signal-to-noise ratio (SNR) is effective for controlling the weight of each visual feature, which is the ratio between target salience and distractors salience (or background salience). Navalpakkam *et al.* improved Itti's original saliency map model [5] by using the maximum SNR as an objective function for weight modulation [13]. The calculation of SNR depends on the mean features of the target and distractors. Frintrop *et al.* proposed a weight modulation model that directly applies SNRs computed from training image features to the modulation weights [14]. A top-down saliency map is generated by taking the difference between the excitation and inhibition maps. The excitation map consists of the weighted responses of feature channels with SNR > 1, whereas the inhibition map consists of the responses with SNR < 1. A target-specific visual attention map is produced by combining the top-down and bottom-up saliency maps. As a result, Frintrop *et al.* developed a highly accurate visual attention system named VOCUS to search for specific targets [15].

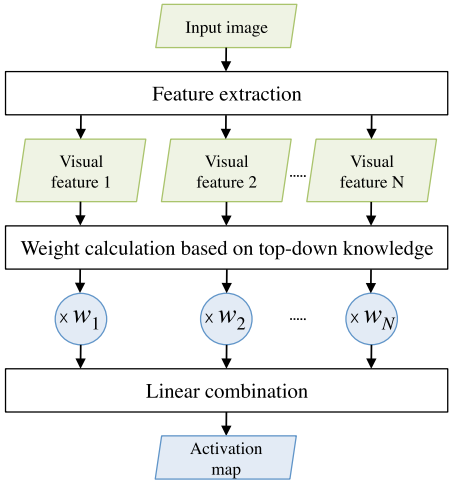


Fig. 2 Computational model of top-down visual attention that modulates weights of visual features [4].

These weighted modulation models, however, have the following problem:

- Because these methods employ the representative value, such as mean, of visual feature extracted from the target and that extracted from all other stimuli regions (or background region) to compute the weight, they only work well for uniform distributions and cannot focus on a relationship in visual feature between any local region in the search image and the target image.

Our goal is to resolve this issue. We use spotlighting to perform the conjunction search of existing feature integration theory [6], and calculate spatially localized weight for a region based on the relationship between the visual feature extracted from the local region in the search image and from the target image. In particular, we pay special attention to the linear separability effect on visual search tasks to calculate these weights. Hodsoll *et al.* found that people can easily find a target which is linearly separable from distractors in feature space [16].

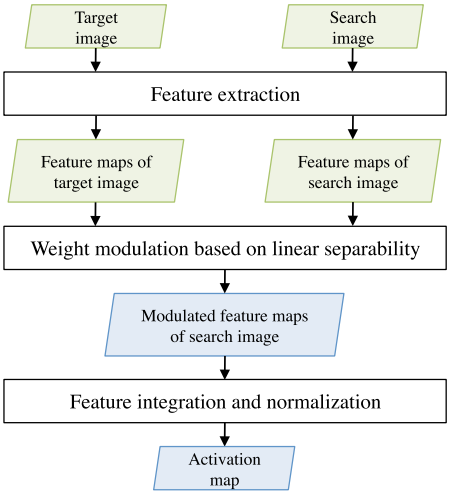


Fig. 3 Process flow of our proposed model.

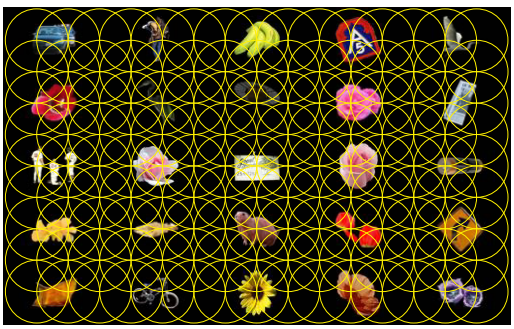


Fig. 4 Spotlight regions in a search image. Yellow circles whose diameter is visual angle θ_s represent spotlight regions, which are put at spatial intervals $d_s \sim \theta_s/2$.

4. Proposed Method

In this section, we describe our top-down model for computing target-specific visual attention. We extend the original saliency map model proposed by Itti *et al.* to this top-down model. Figure 3 shows the process flow of our model. As mentioned in Sect. 3, we consider the relationship between the visual feature extracted from each local region in the search image and from the target image. To compute the spatially localized weights based on the psychophysical findings, we first put some local circular regions of interest whose diameter is visual angle θ_s, at spatial intervals d_s on the search image as shown in Fig. 4. We refer to this region as spotlight region for the remainder of this paper. We then calculate the weights for each spotlight region. For this process, we utilize the linear separability effect as described by Hodsoll *et al.* [16].

An activation map that estimates target-specific visual attention is computed using the following three processes: (1) extraction of early visual features from the target image and entire search image, (2) calculation of weights for each spotlight region and modulation of the features using the weights, and (3) integration and normalization of the features.

4.1 Extraction of Early Visual Features from Search Image and Target Image

We create feature maps that Itti *et al.* [5] have proposed to compute the saliency map based on early visual features. First, nine images with varying scales ($v \in 0 \ldots 8$) are created using Gaussian pyramids that progressively filter out higher frequencies and subsample the image. Red (r), green (g), and blue (b) channels are extracted from the images. An intensity image (I) and four broadly-tuned color images (R, G, B, and Y) are created according to

$$I(v) = (r(v) + g(v) + b(v))/3, \tag{1}$$
$$R(v) = r(v) - (g(v) + b(v))/2, \tag{2}$$
$$G(v) = g(v) - (r(v) + b(v))/2, \tag{3}$$
$$B(v) = b(v) - (r(v) + g(v))/2, \tag{4}$$
$$Y(v) = (r(v) + g(v))/2 - |r(v) - g(v)|/2 - b(v). \tag{5}$$

Four local orientation images $O(v, \theta)$ ($\theta \in 0°, 45°, 90°, 135°$) are created from I using oriented Gabor pyramids as follows:

$$O(v, \theta) = I(v) * \phi(\theta), \tag{6}$$

where $*$ denotes a convolution and ϕ denotes a Gabor filter.

Next, a set of feature maps are created from six patterns of center-surround differences between a "center" finer scale c ($\in 2, 3, 4$) and a "surround" coarser scale s ($= c + \delta$ ($\delta \in 3, 4$)) as follows:

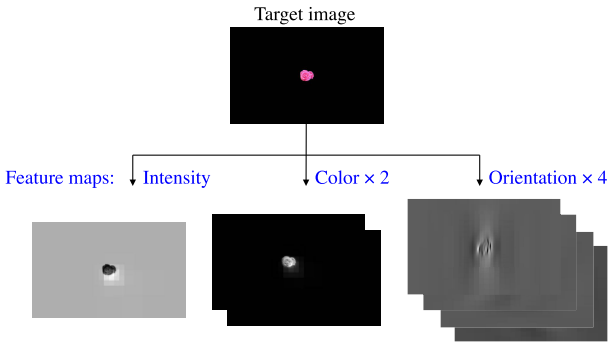
$$\mathcal{F}_I = \sum_{c=2}^{4} \sum_{s=c+3}^{c+4} I(c) \ominus I(s), \tag{7}$$
$$\mathcal{F}_{RG} = \sum_{c=2}^{4} \sum_{s=c+3}^{c+4} (R(c) - G(c)) \ominus (G(s) - R(s)), \tag{8}$$
$$\mathcal{F}_{BY} = \sum_{c=2}^{4} \sum_{s=c+3}^{c+4} (B(c) - Y(c)) \ominus (Y(s) - B(s)), \tag{9}$$
$$\mathcal{F}_O(\theta) = \sum_{c=2}^{4} \sum_{s=c+3}^{c+4} O(c, \theta) \ominus O(s, \theta), \tag{10}$$

where "$\ominus$" denotes interpolation to the finer scale and point-by-point subtraction.

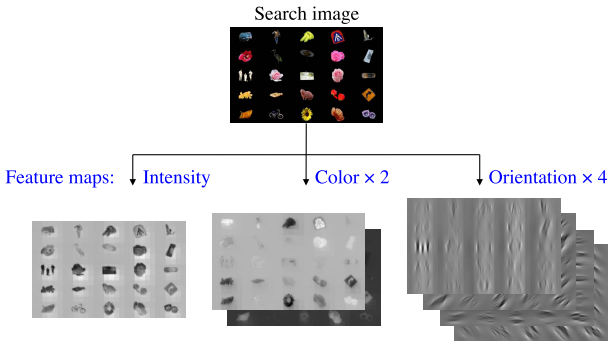
Figure 5 shows the feature maps. The upper part is an example of feature maps computed from a target image. The bottom part is an example from a search image. The seven feature maps created are as follows: one intensity map, two color maps (RG, BY), and four orientation maps (0°, 45°, 90°, 135°).

4.2 Weight Modulation Based on Linear Separability between Feature Distributions

We weight each feature map within each spotlight region

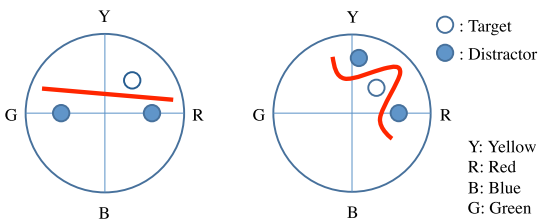


(a) Feature maps extracted from target image



(b) Feature maps extracted from search image

Fig. 5 Examples of feature maps.



(a) Linearly separable (b) Nonlinearly separable

Fig. 6 Examples of linearly and nonlinearly separable targets [16].

based on linear separability. We first extract the distributions of visual features on the feature maps, and then compute variance ratios between the distributions extracted from each spotlight region in the search image and from a spotlight region at the center of the target image using principles of linear separability.

4.2.1 Psychophysical Findings on Visual Search

Hodsoll *et al.* suggested that the difficulty of visual search is dependent on whether or not a target is linearly separable from other objects within a feature space [16]. If the feature distribution of the target is linearly separable from that of the distractors as shown in Fig. 6 (a), it is easy to locate the target. In contrast, if the feature distributions of the target and the distractors are nonlinearly separable as shown in Fig. 6 (b), a serial search is required to locate the target by shifting one's spotlight of attention in the conjunction search

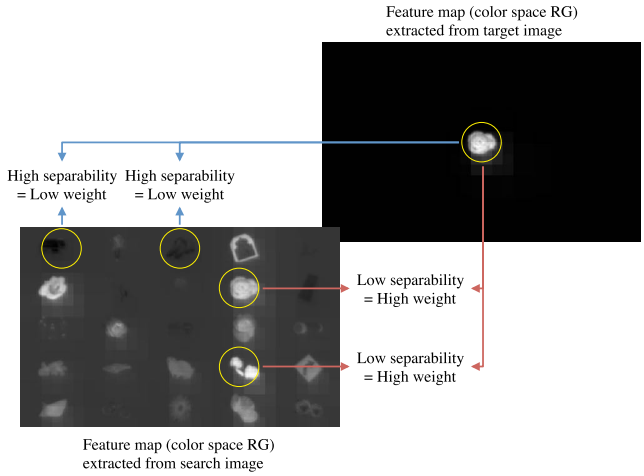


Fig. 7 Weight calculations for some spotlight regions surrounded by yellow circles.

manner [6]. In the case of Fig. 6 (a), the color feature is important unlike in Fig. 6 (b). In accordance to the findings, we consider that the linear separability in the conjunction search manner modulates target-specific visual attention. We assume that Fisher's variance ratio of between-class variance to within-class variance is fit to simulate the linear separability effect. The ratio is a measure of linear separability [17]. The linear separability effect also exists for the other feature spaces as well as the color space.

4.2.2 Linear Separability of Feature Distributions

We employ Fisher's variance ratio $J_{i,j}$ of between-class variance $\sigma^2_{B_{i,j}}$ to within-class variance $\sigma^2_{W_{i,j}}$ as a measure of the linear separability between the visual feature distribution in the i-th feature map extracted from a spotlight region χ_j in the search image ($l = 1$) and from the target image ($l = 2$). These variances are defined as follows:

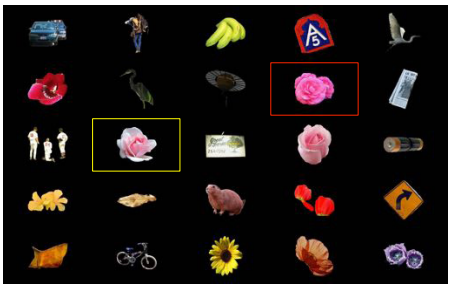
$$J_{i,j} = \frac{\sigma^2_{B_{i,j}}}{\sigma^2_{W_{i,j}}}, \tag{11}$$

$$\sigma^2_{B_{i,j}} = \sum_{l=1}^{2} (m_l - m)^2, \tag{12}$$

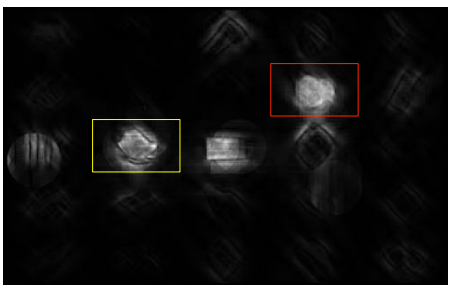
$$\sigma^2_{W_{i,j}} = \frac{1}{n} \sum_{l=1}^{2} \sum_{x,y \in \chi_j} \left(\mathcal{F}_{i,j}(x,y) - m_l \right)^2, \tag{13}$$

where $\mathcal{F}_{i,j}$ is the feature for each pixel (x, y) within the spotlight region χ_j of the i-th feature map, m_l is the centroid of the feature distribution, m is the centroid of m_l, and n is the number of pixels in the spotlight region. We apply the above calculation to each feature map computed by equations (7) – (10).

Figure 7 shows an example of calculating weights within some spotlight regions on the feature map. Note that the spotlight regions are circled in yellow.



(a) Search image



(b) Activation map

Fig. 8 A search image and the corresponding activation map estimated by our proposed model. The pink flower which is surrounded by a red rectangle is a target object.

4.2.3 Weight Modulation of Feature Maps

We modulate the feature maps within each spotlight region using the weights based on the variance ratio. Although the target image had precisely the same appearance as an object in the search image, they generated different feature maps owing to the center-surround computation using the Gaussian pyramid. Hence, linear separability between the target image and any spotlight region is nonzero.

The variance ratio is calculated for each spotlight region of the seven feature maps. If the variance ratio is low, the region in the search image is similar to the target image on the feature map, and hence, a higher weight should be given to the region on the feature map. If the variance ratio is high, the region is not similar to the target image on the feature map. In this case, the region on the feature map should be given a lower weight. To accomplish this, we apply the reciprocal of the variance ratio to the weight, $w_{i,j} = 1/J_{i,j}$, that is multiplied by the i-th feature map within the spotlight region χ_j as follows: $w_{i,j}|\mathcal{F}_{i,j}(x, y)|$. Depending on the visual angle θ_s of the spotlight region and the interval d_s between the regions, a spotlight region overlaps the surrounding spotlight regions. In this case, we multiply the original feature map of the overlapping region by the weight for each of the spotlight region, and then calculate the mean of the modulated feature maps.

4.3 Feature Integration and Normalization

We integrate the seven feature maps into an activation map using the same process as saliency map computation [5]. First, the seven feature maps are normalized with respect to each modality and integrated into three conspicuity maps of intensity, color, and orientation. Then, the conspicuity maps are normalized and integrated into the activation map. The local maximum of the activation map is regarded as the most attracted location of the target search task. Figure 8 (a) is a search image, and the pink-red flower which is surrounded by a red rectangle is a target object. Figure 8 (b) is the activation map created from the search image. It shows that the other pink flower surrounded by a yellow rectangle is activated as with the target object unlike another pink flower and other red flowers on the map.

5. Experiment and Result

5.1 Data Set

We employed the MSRA Salient Object Database [18] that includes 1000 images and their binary mask images. To create each search image, we selected 25 images from the dataset in a random manner and placed the object regions extracted using their mask images in a 5×5 grid pattern on a black background. The target image included an object selected from them in a random manner. We conducted the experiment with various target and search images to minimizing the effects of visual memory. Therefore, we did not put same target objects in different search images or different target objects in same search images. The size of the search image and the target image was 1920×1200 pixels.

Ten participants (nine males and one female) with normal vision, whose ages ranged between 22 and 24 years, participated in our experiment. Each participant sat with his/her chin on a padded rest in front of a 24.1-inch screen. The distance between the participant and the screen was 600 mm. They were instructed first to observe a target image for five seconds and then shift their gaze to the center of the screen once, and next search for the target object in the paired search image until they found it. We treated this procedure as a trial of our visual search experiment. We conducted 100 trials for each participant. All participants were given the 100 combinations of target and search images in the same order. We recorded their eye movements using a Tobii X60 Eye Tracker (data rate: 60 Hz, accuracy: typical 0.5 degrees, precision: typical 0.5 degrees) installed below the screen during the trials. With regard to the visual angle of spotlight region θ_s, we consider the error range of the eye tracker (accuracy + precision: 1.0°) and central area of vision (foveal area + para-foveal area: 5.0°), i.e., θ_s is the total range (6.0°). We investigate the estimation accuracy for three kinds of interval between spotlight regions: $d_s \sim \theta_s/4, \theta_s/2, 3\theta_s/4$.

5.2 Comparative Models

We employ three conventional models to evaluate our proposed model. One is the bottom-up visual attention estimation model, i.e. saliency map model, proposed by Itti *et*

al. [5]. Another is the target-specific visual attention model proposed by Frintrop *et al.* [14]. In this paper, we use the same visual feature set used by the saliency map model for Frintrop's model and for our proposed model. As mentioned in Sect. 3, Frintrop's model learns the optimum weight of each feature channel from training images. In this experiment, the weight of each target was learned using all combinations of the target image and the search images. We assumed that the limited dataset would cause overtraining. Alternatively, to avoid learning, we calculate the weights from a target image and the paired search image and apply them to the feature maps of the search image to estimate a top-down saliency map S_T. We call the third model SNR based model. As with Frintrop's model, the top-down saliency map is integrated with the bottom-up saliency map S_B to estimate a global activation map S_A. The contribution of each map is adjusted by a top-down factor $w_T \in [0 \dots 1]$:

$$S_A = (1 - w_T) * S_B + w_T * S_T. \tag{14}$$

For $w_T = 0.5$, bottom-up and top-down cues are evenly regarded, whereas for $w_T = 1.0$, only top-down cue is considered. We employ two activation maps with $w_T = 0.5$ and $w_T = 1.0$ as the comparative models for the evaluation.

5.3 Evaluation Approach

To quantify how well our estimations match the participants' actual fixation positions[†], we use the normalized scanpath saliency (NSS) [20], which is defined as the response value at the current fixation position $\boldsymbol{x}_{human} = (x_{human}, y_{human}) \in \mathbb{Z}^2$ in a visual attention map S that has been normalized to have zero mean and unit standard deviation:

$$\mathrm{NSS} = \frac{1}{\sigma_S}(S(x_{human}, y_{human}) - \mu_S) \tag{15}$$

where μ_S and σ_S^2 are the mean and variance of the visual attention map. A larger NSS score means a better fit, whereas a zero NSS score means that the model was no better than chance at attractive location.

When we calculate NSS, we consider the error range of the eye tracker and central area of vision as with the visual angle of spotlight region, and regard the total range (visual angle: 6.0°) as a fixation area. Figure 9 shows an example of gaze positions and fixation areas. Each green point shows the gaze position, which was recorded through a visual search trial. Each white circle, whose center is the green point, shows the fixation area. In this paper, we exploit the mean NSS within the circles to evaluate the models.

5.4 Experimental Result

Figure 10 shows the mean NSS across all visual search tasks, i.e., 100 trials × 10 participants. The score for our proposed model $d_s \sim \theta_s/4$ was 0.479 ± 0.129, which was higher than the other models (our proposed models $d_s \sim \theta_s/2$: 0.433 ± 0.136, $d_s \sim 3\theta_s/4$: 0.344 ± 0.114, Itti's saliency map: 0.253 ± 0.084, Frintrop's models $w_T = 0.5$: 0.271 ± 0.084, $w_T = 1.0$: 0.301 ± 0.100, SNR based models $w_T = 0.5$: 0.398 ± 0.115, $w_T = 1.0$: 0.449 ± 0.154). Especially, the Friedman test and the multiple pairwise comparison revealed that the score for our proposed model $d_s \sim \theta_s/4$ was significantly higher than for Itti's saliency map with $\chi^2(7) = 29.0, p < .01$ and for Frintrop's models $w_T = 0.5$ with $\chi^2(7) = 23.4, p < .01$ and $w_T = 1.0$ with $\chi^2(7) = 19.2, p < .01$, and the score for $d_s \sim \theta_s/2$ was significantly higher than for Itti's saliency map with $\chi^2(7) = 16.1, p < .05$. Also, the score for the SNR based model $w_T = 1.0$ was significantly higher than for Itti's saliency map with $\chi^2(7) = 17.6, p < .05$.

Figure 11 shows the mean response value in false es-

[†]A relatively stable eye position within some threshold of dispersion (2.5° in this experiment) over some minimum duration (100 msec) [19].

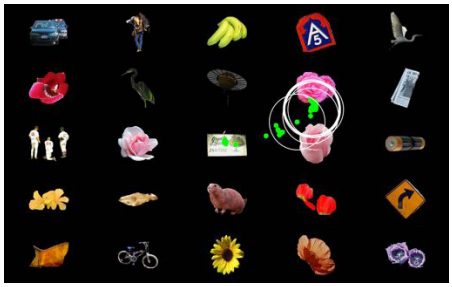


Fig. 9 Gaze positions (green points) measured using the eye tracker and fixation areas (white circles) considered the error range of eye tracker and central area of vision.

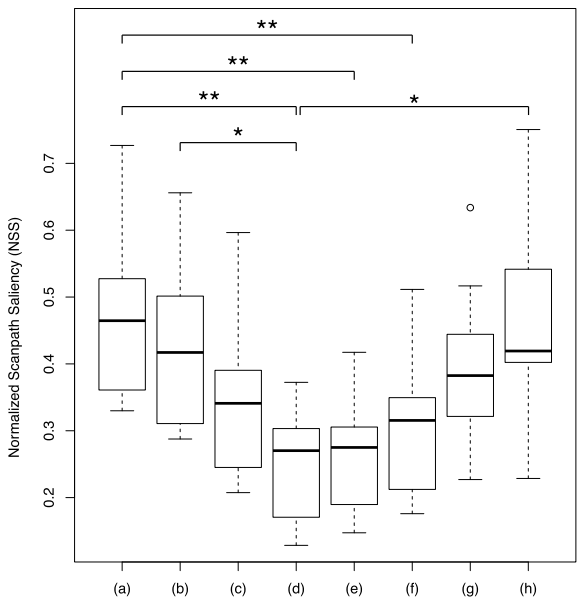


Fig. 10 The boxplots of mean NSS. (a) proposed models $d_s \sim \theta_s/4$, (b) $d_s \sim \theta_s/2$, (c) $d_s \sim 3\theta_s/4$, (d) Itti's saliency map, (e) Frintrop's models $w_T = 0.5$, (f) $w_T = 1.0$, (g) SNR based models $w_T = 0.5$, (h) $w_T = 1.0$. A higher score means a better estimation. The lower edge of the box is the lower quartile and the upper edge is the upper quartile. The circle indicates the outlier.

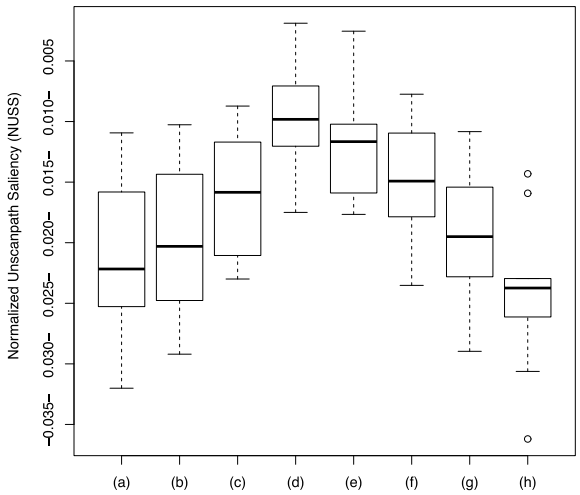
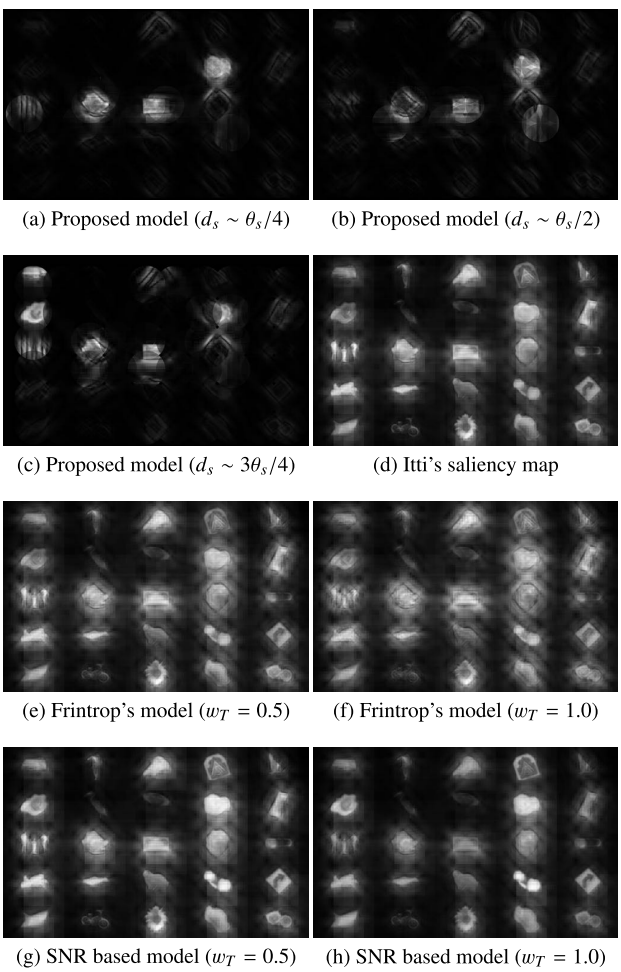


Fig. 11 The boxplots of mean NUSS. (a) proposed models $d_s \sim \theta_s/4$, (b) $d_s \sim \theta_s/2$, (c) $d_s \sim 3\theta_s/4$, (d) Itti's saliency map, (e) Frintrop's models $w_T = 0.5$, (f) $w_T = 1.0$, (g) SNR based models $w_T = 0.5$, (h) $w_T = 1.0$. A lower score means a less false estimation. The lower edge of the box is the lower quartile and the upper edge is the upper quartile. The circle indicates the outlier.



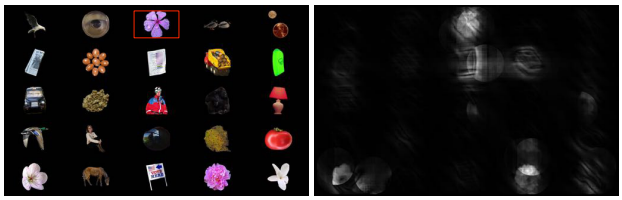
(a) Proposed model ($d_s \sim \theta_s/4$) (b) Proposed model ($d_s \sim \theta_s/2$)

(c) Proposed model ($d_s \sim 3\theta_s/4$) (d) Itti's saliency map

(e) Frintrop's model ($w_T = 0.5$) (f) Frintrop's model ($w_T = 1.0$)

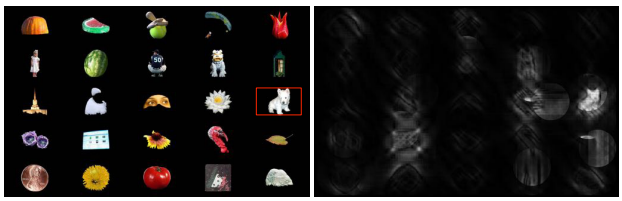
(g) SNR based model ($w_T = 0.5$) (h) SNR based model ($w_T = 1.0$)

Fig. 12 Activation maps that our proposed model and the conventional models estimated the visual attention to the search image shown in Fig. 8.



(a) Search image (b) Activation map

Fig. 13 Visual attention searching for the violet flower by the proposed model ($d_s \sim \theta_s/4$).



(a) Search image (b) Activation map

Fig. 14 Visual attention searching for the white dog by the proposed model ($d_s \sim \theta_s/4$).

timation areas, i.e., ungazed areas. We call the response value the normalized unscanpath saliency (NUSS). Our activation maps had much the same NUSS as the other models (our proposed models $d_s \sim \theta_s/4$: -0.021 ± 0.006, $d_s \sim \theta_s/2$: -0.020 ± 0.006, $d_s \sim 3\theta_s/4$: -0.016 ± 0.005, Itti's saliency map: -0.010 ± 0.005, Frintrop's models $w_T = 0.5$: -0.012 ± 0.005, $w_T = 1.0$: -0.015 ± 0.005, SNR based models $w_T = 0.5$: -0.019 ± 0.006, $w_T = 1.0$: -0.024 ± 0.006) compared with NSS. These figures suggest that the activation map estimated by our proposed model was broadly consistent with the actual focused area. We can also confirm that shorter interval between the spotlight regions resulted in better estimation. Such dense spotlighting simulates target-specific visual attention. However, there are a trade-off between the accuracy and the computational cost.

Figure 12 shows experimental results obtained from the target image and the paired search image shown in Fig. 8. This figure shows that our proposed model estimated top-down visual attention with high recall and precision. The mean NSS of each model was 0.920 (our proposed model $d_s \sim \theta_s/4$), 1.075 ($d_s \sim \theta_s/2$), 0.583 ($d_s \sim 3\theta_s/4$), 0.688 (Itti's saliency map), 0.741 (Frintrop's model $w_T = 0.5$), 0.780 ($w_T = 1.0$), 0.703 (SNR based model $w_T = 0.5$), and 0.704 ($w_T = 1.0$) in the case of Fig. 12. Also, the mean NUSS of each model was -0.052 (our proposed model $d_s \sim \theta_s/4$), -0.060 ($d_s \sim \theta_s/2$), -0.033 ($d_s \sim 3\theta_s/4$), -0.036 (Itti's saliency map), -0.038 (Frintrop's model $w_T = 0.5$), -0.040 ($w_T = 1.0$), -0.037 (SNR based model $w_T = 0.5$), and -0.038 ($w_T = 1.0$). Figures 13 and 14 show the other examples of activation maps estimated by our proposed model with $d_s \sim \theta_s/4$. The mean NSS of Figs. 13 (b) and 14 (b) were 1.304 and 0.697, respectively. Also, the mean NUSS of Figs. 13 (b) and 14 (b) were -0.055 and -0.042, respectively.

In situations where the saliency of a target object was

low and the saliencies of other objects around the target object were adequately high, participants might focus their attention on the target object even if saliency of target object was extremely low. This phenomenon suggests a need to modulate the weights focusing not only on spatially localized features, but also on more global features of the entire feature map. Further, participants might search for a target object with their peripheral vision, especially when the target object was located near the center of the search image. In this case, an accurate evaluation of top-down visual attention may not be achieved using any eye tracker. Thus, redesigning the layout of objects on the search image or conducting an experiment to measure covert attention [21] would be helpful to alleviate this problem.

6. Conclusion

In this paper, we focused on the effect of linear separability between the visual feature distributions of a target object and a local region in field of view on the visual target search task and proposed a computational model that estimates the target-specific top-down visual attention. Through the experiment, we confirmed the effectiveness of our computational model.

In the future, we plan to verify our model with natural images that contain complicated background. We will also propose to calculate weights focusing not only on local saliency but also on global saliency in consideration of the spatial relationship between visual features of the focused object and the neighboring objects. Further, we will extend the model to estimate visual attention dynamics.

Acknowledgments

This work is partially supported by a Grant in Aid for Scientific Research from MEXT (Ministry of Education, Culture, Sports, Science, and Technology) of Japan under Contract no. 26730119 and 26280074, and the Graduate Program for Real-World Data Circulation Leaders, Nagoya University.

References

[1] T. Yonezawa, H. Yamazoe, A. Utsumi, and S. Abe, "Gaze-communicative behavior of stuffed-toy robot with joint attention and eye contact based on ambient gaze-tracking," Proceedings of the 9th International Conference on Multimodal Interfaces, pp.140–145, 2007.
[2] Y. Nagai, "From bottom-up visual attention to robot action learning," Proceedings of the 8th IEEE International Conference on Development and Learning, pp.1–6, 2009.
[3] R. Sato, K. Doman, D. Deguchi, Y. Mekada, I. Ide, H. Murase, and Y. Tamatsu, "Visibility estimation of traffic signals under rainy weather conditions for smart driving support," Proceedings of the 15th International IEEE Conference on Intelligent Transportation Systems, pp.1321–1326, 2012.
[4] A. Kimura, R. Yonetani, and T. Hirayama, "Computational models of human visual attention and their implementations: A survey," IEICE Trans. Inf. & Syst., vol.E96.D, no.3, pp.562–578, 2013.
[5] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.20, no.11, pp.1254–1259, 1998.
[6] A. Treisman and G. Gelade, "A feature-integration theory of attention," Cognitive Psychology, vol.12, no.1, pp.97–136, 1980.
[7] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," Matters of Intelligence, pp.115–141, 1987.
[8] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.35, no.1, pp.185–207, 2013.
[9] S. Frintrop, E. Rome, and H.I. Christensen, "Computational visual attention systems and their cognitive foundations: A survey," ACM Transactions on Applied Perception, vol.7, no.1, pp.1–46, 2010.
[10] L. Elazary and L. Itti, "A bayesian model for efficient visual search and recognition," Vision Research, vol.50, no.14, pp.1338–1352, 2010.
[11] J. Wolfe, "Guided search 2.0: A revised model of visual search," Psychonomic Bulletin & Review, vol.1, no.2, pp.202–238, 1994.
[12] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.31, no.6, pp.989–1005, 2009.
[13] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp.2049–2056, 2006.
[14] S. Frintrop, G. Backer, and E. Rome, "Goal-directed search with a top-down modulated computational attention system," Pattern Recognition, pp.117–124, 2005.
[15] S. Frintrop, "Vocus: a visual attention system for object detection and goal-directed search," Lecture Notes in Artificial Intelligence (LNAI), Springer, 2006.
[16] J. Hodsoll and G. Humphreys, "Driving attention with the top down: The relative contribution of target templates to the linear separability effect in the size dimension," Perception & Psychophysics, vol.63, no.5, pp.918–926, 2001.
[17] R. Fisher, "The use of multiple measurements in taxonomic problems," Annals of Eugenics, vol.7, no.2, pp.179–188, 1936.
[18] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.33, no.2, pp.353–367, 2011.
[19] R.J.K. Jacob and K.S. Karn, "Eye tracking in human-computer interaction and usability research: Ready to deliver the promises," in J. Hyönä, R. Radach, H. Deubel (Eds.), The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research, pp.573–605, Elsevier, Amsterdam, 2003.
[20] R.J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," Vision Research, vol.45, no.18, pp.2397–2416, 2005.
[21] D. Engel and C. Curio, "Detectability prediction in dynamic scenes for enhanced environment perception," Proceedings of IEEE Intelligent Vehicles Symposium, pp.178–183, 2012.

Takatsugu Hirayama received the M.E. and D.E. degrees in Engineering Science from Osaka University, Japan, in 2002 and 2005, respectively. From 2005 to 2011, he was a research assistant professor at the Graduate School of Informatics, Kyoto University. He is currently a designated associate professor at the Graduate School of Information Science, Nagoya University. His research interests include computer vision, human vision, human communication, and human-computer interaction. He has received two best paper awards from IEICE ISS and at the 16th IEEE Symposium on Multimedia, and the best short paper award at the 8th ACM Symposium on Eye Tracking Research and Applications in 2014. He is a member of IEICE, the Information Processing Society of Japan (IPSJ), the Human Interface Society of Japan, and ACM.

Toshiya Ohira received the B.E. degree in Information Engineering and the M.S. degree in Information Science from Nagoya University, in 2012 and 2014, respectively. He is with Nippon Telegraph and Telephone (NTT) West Corporation. His research interests include human visual attention and visual search.

Kenji Mase received the B.S. degree in Electrical Engineering and the M.S. and Ph.D. degrees in Information Engineering from Nagoya University in 1979, 1981 and 1992 respectively. He became a professor of Nagoya University in August 2002. He is now with the Graduate School of Information Science, Nagoya University. He joined the Nippon Telegraph and Telephone Corporation (NTT) in 1981 and had been with the NTT Human Interface Laboratories. He was a visiting researcher at the Media Laboratory, MIT in 1988-1989. He has been with ATR (Advanced Telecommunications Research Institute) in 1995-2002. His research interests include gesture recognition, computer graphics, artificial intelligence and their applications for computer-aided communications. He is a member of the Information Processing Society of Japan (IPSJ), Japan Society of Artificial Intelligence (JSAI), Virtual Reality Society of Japan, Human Interface Society of Japan and ACM, a senior member of IEEE Computer Society and a fellow of IEICE.