# Top-down Visual Attention Computational Model Using Visual Feature Distribution of Search Target

Toshiya Ohira
Nagoya University
Nagoya, Aichi, Japan
ohira@cmc.ss.is.nagoya-u.ac.jp

Takatsugu Hirayama
Nagoya University
Nagoya, Aichi, Japan
hirayama@is.nagoya-u.ac.jp

Shohei Usui
Nagoya University
Nagoya, Aichi, Japan
usui@cmc.ss.is.nagoya-u.ac.jp

Shota Sato
Nagoya University
Nagoya, Aichi, Japan
sato@cmc.ss.is.nagoya-u.ac.jp

Kenji Mase
Nagoya University
Nagoya, Aichi, Japan
mase@nagoya-u.jp

## ABSTRACT

Advanced information systems captivate people's attention. Examples of such systems include advanced driver support cars 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 a person's visual attention while carrying out a visual target search task. Existing models estimate visual attention using the mean difference between the visual feature distribution of a target stimulus and other stimuli. This model is limited, however, in that for difficult search tasks, a better performance is not often achieved. For such tasks, the linear separability effect of a visual feature distribution must be considered. We incorporate this effect into our proposed model that estimates target-specific visual attention using the Fisher's variance ratio[1] between a local visual feature distribution of a target stimulus and each of the other stimuli. We confirm the effectiveness of our computational model using a visual search experiment.

## Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous

## General Terms

Theory

## Keywords

Human visual attention, Visual search, Top-down human visual attention, Target-specific attention map, Linear separability, Attention map, Computational model

## 1. INTRODUCTION

Many researchers have focused on gaze to develop advanced information systems that interact with humans. To gaze at something implies human cognitive states such as interest and intent. For example, a human-friendly robot not only requires verbal communication, but also requires nonverbal communication such as eye contact and mutual gaze[13]. 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[8]. The advanced driving support system of a vehicle should also be able estimate a driver's visual attention. This system helps the driver recognize driving cues such as signboards and guide plates on the roads. The visibility of these objects differs in varying conditions such as daytime/nighttime, roads with/without obstacles, and urban/rural roads[11]. The support system is effective if it can estimate the visibility according to their environments 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. When only visual stimuli activate visual attention in a scene, this is known as bottom-up processing. In contrast, when a person views 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, simulating visual attention and computing visual saliency have attracted much attention in the field of robotics and computer vision. Itti et al.[4] proposed a representative computational model of visual saliency. They incorporated a bottom-up computational process into their proposed saliency map model based on the feature integration theory of Treisman and Gelade[12] and multi-resolution structure of Koch and Ullman[6]. Other bottom-up visual attention models, which are the derivatives of Itti's model, have been developed by other researchers. On the other hand, computational models of top-down visual attention are not well studied. However, many psychophysical find-

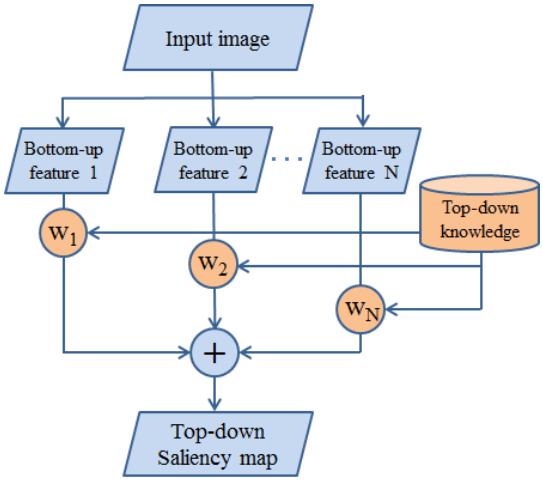
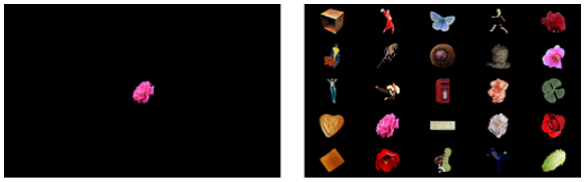



Figure 1: Computational model of top-down visual attention that modulates weights of visual features[5].



(a)Target image (b)Panel image

Figure 2: Example of a pair of images used for the visual search task.

ings and conceptual models on task-oriented visual attention have been reported. Our research focuses on estimating a 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.

## 2. RELATED WORKS

In this section, we discuss works related to top-down visual attention in search tasks. Figure 1 outlines a typical computational process that modulates the weights of visual features[5]. Some researchers have proposed the computational models based on this weight modulation process. For visual search tasks, it is important to consider the relationship between targets and distractors. Navalpakkam et al.[9] improved Itti's original saliency map model[4] by using the maximum signal-to-noise ratio (SNR) as an objective function for weight modulation. The signal-to-noise ratio is the ratio between target salience and distractor salience and is effective for controlling the weight of each visual feature. The calculation of SNR depends on the average feature distribution 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[2]. 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, top-down, visual attention system named VOCUS to search for specific targets.

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

- Because the optimal weight of each visual features is extracted from a finite set of training images, they are prone to over-fitting to the learned dataset.
- A relationship between the target and each distractor cannot be modulated because a weight is calculated between each visual feature of the target and all other stimuli.
- Because these methods employ the mean or the expected value of each visual feature distribution extracted from the object regions to compute the weight, so they only work well for uniform distributions.

Our goal is to resolve each of these issues. First, we calculate weights from visual features extracted from only a pair of images in a visual search task. These images consist of a target image and a search image that includes the target object and other distractor objects. Second, we use spotlighting to perform the conjunction search of existing feature integration theory[12], and calculate spatially localized weights to modulate the relationship between the target and each object. Third, we calculate the weights according to the distributions of visual features extracted from each object region. In particular, we pay special attention to the linear separability effect on visual search tasks to the calculation of these weights. Hodsoll et al.[3] found that people can easily find a target which is linearly separable from distractors in feature space.

In this paper, we propose a top-down model for computing target-specific visual attention by considering the dispersion of visual features of each object.

## 3. VISUAL SEARCH TASK

Conventional psychophysical studies on visual search employ simple geometric images. We use natural images with more complicated textures as shown in Figure 2. We design a visual search task as follows: (1) An experimenter presents a target image to an experimental participant at the center of the display field for several seconds (Figure 2(a)). (2) The experimenter presents a panel image where the target and distractor objects are aligned (Figure 2(b)) and asks the participant to search for the target. Note that the target image is the same as the target included in the panel image.

## 4. PROPOSED METHOD

In this section, we describe our 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 Section 2, we consider the linear separability between the visual feature distributions of each object in the panel image

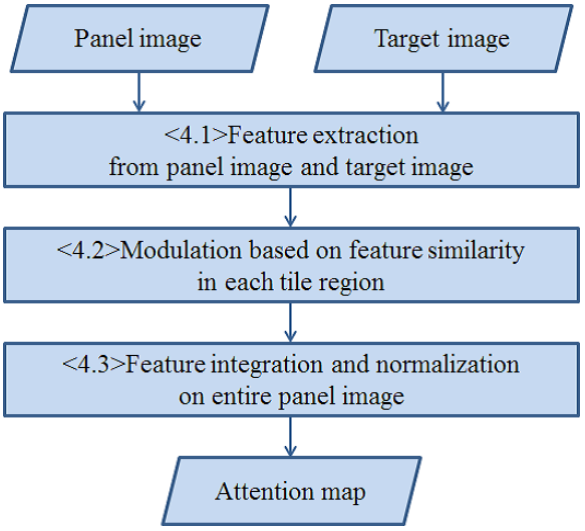



Figure 3: Process flow of our proposed model.

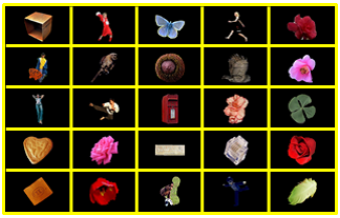


Figure 4: Segmented panel image.

and the target. To compute the weights based on linear separability, we first segment the panel image into equal subregions as shown in Figure 4. Each subregion contains an object. We then calculate the weights for each subregion. We refer to this region as the tile region for the remainder of this paper. Weights are computed using the following two processes: (1) extraction of early visual features from the target image and entire panel image and (2) modulation of weights for each tile region. For the latter process, we utilize the linear separability effect as described by Hodsoll et al.[4].

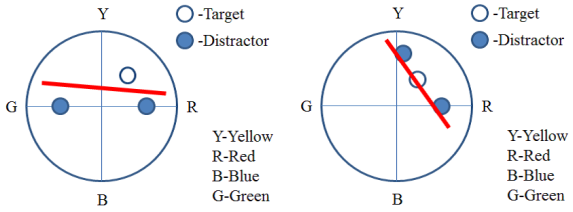
## 4.1 Extraction of early features from panel image and target image

Similar to Itti et al.[4], we employ early visual features to compute the visual attention map. First, nine images with varying scales ($c \in 0 \ldots 8$) are created using Gaussian pyramids that progressively filter out higher frequencies and subsample the images. 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

$$I(c) = (r(c) + g(c) + b(c))/3, \quad (1)$$

$$R(c) = r(c) - (g(c) + b(c))/2, \quad (2)$$

$$G(c) = g(c) - (r(c) + b(c))/2, \quad (3)$$




(a)linearly separable (b)nonlinearly separable

Figure 5: Examples of linearly or nonlinearly separable targets[3].

$$B(c) = b(c) - (r(c) + g(c))/2, \quad (4)$$

$$Y(c) = (r(c) + g(c))/2 - |r(c) - g(c)|/2 - b(c). \quad (5)$$

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

$$O(c, \theta) = I(c) * \phi(\theta), \quad (6)$$

where $*$ means a convolution and $\phi$ means a Gabor filter.

Next, a set of feature maps are created in six patterns of center-surround differences. The across-scale difference between two maps, denoted by "$\ominus$" below, is obtained by interpolating the finer scale and point-by-point subtraction. The center-surround differences between a "center" fine scale $c$ ($\in 2, 3, 4$) and a "surround" coarser scale $s$ ($= c + \delta$ ($\delta \in 3, 4$)) yield the feature maps as follows:

$$I(c, s) = |I(c) \ominus I(s)|, \quad (7)$$

$$RG(c, s) = |(R(c) - G(c)) \ominus (G(s) - R(s))|, \quad (8)$$

$$BY(c, s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|, \quad (9)$$

$$O(c, s, \theta) = |O(c, \theta) \ominus O(s, \theta)|. \quad (10)$$

## 4.2 Modulation of weights based on linear separability between feature distributions

We modulate the weight of each tile region on each feature map based on linear separability. We first extract the distributions of visual features on the feature maps, and then, we compute variance ratios between the distribution of each object in the panel image and 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 particular feature space. If the feature distribution of the target is linearly separable from that of the distractors (as shown in Figure 5(a)) , it is easy to locate the target[12]. In contrast, if the feature distribution of the target and the distractors is nonlinearly separable (as shown in Figure 5(b)), a serial search is required to locate the target by shifting one's spotlight of attention in the conjunction search manner. In the case of Figure 5(a), the color feature is important unlike in Figure 5(b). The linear separability effect exists for other feature spaces in addition to the color space. In accordance to others' findings, we consider that linear separability modulates

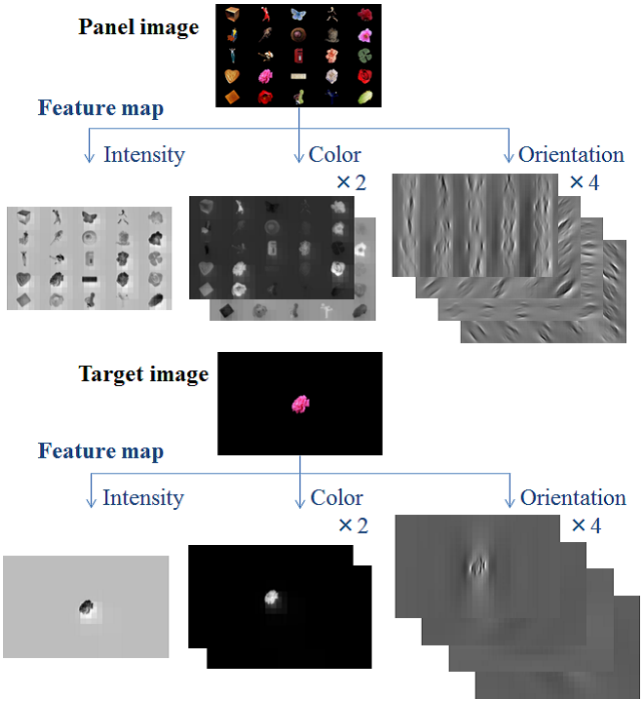



Figure 6: Feature maps for calculation linear separability.

target-specific visual attention. We assume that the 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[1].

#### 4.2.2 *Linear separability of feature distributions*

We employ $J_\sigma$ the Fisher's variance ratio, (11), of between-class variance, (12), to within-class variance, (13), as a measure of the linear separability between the visual feature distribution of each object in the panel image and the target. These variances are defined as follows:

$$J_\sigma = \frac{{\sigma^2}_B}{{\sigma^2}_W}, \tag{11}$$

$${\sigma^2}_B = \frac{1}{n}\sum_{i=1}^{c} n_i (m_i - m)^2, \tag{12}$$

$${\sigma^2}_W = \frac{1}{n}\sum_{i=1}^{c}\sum_{x\in\chi_i} (x - m_i)^2, \tag{13}$$

where $x$ is the feature for each pixel within an object region $\chi_i$ of the feature map, $m_i$ is the centroid of the feature distribution, and $m$ is the centroid of $m_i$. We apply each feature distribution computed by equations $(7)-(10)$ excluding the calculation of absolute value to $x$ because the sign of each feature space is important for measuring the separability of the distributions. Figure 6 shows the feature maps. The upper part of Figure 6 is an example of feature maps computed from a panel image. The bottom part of Figure 6 is an example of feature maps computed from a target image. The seven feature maps created are as follows: one intensity map, two color maps ($RG$, $BY$), and four orientation maps ($0°$, $45°$, $90°$, $135°$).

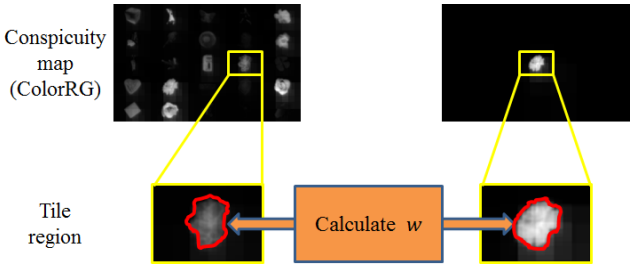
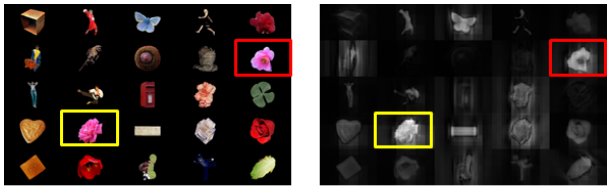



Figure 7: Calculation of weights.



(a) Panel image (b)Attention map

Figure 8: Target-specific visual attention map.

#### 4.2.3 *Weight modulation of feature maps*

We modulate the weights of feature maps within each tile region based on the variance ratio. Figure 7 shows an example of calculating weights within a tile region on the feature map. We calculate the variance ratio between the feature distribution within an object region on the panel image and within the object region on the target image. Note that each object region is circled in red as shown in Figure 7. Although the target image had precisely the same appearance as an object in the panel image, they generated different feature maps owing to the center-surround computation using the Gaussian pyramid. Hence, linear separability between the target image and the object image is nonzero. The variance ratio is calculated for each of the seven feature maps. If the variance ratio is low, the focused object in the panel image is similar to the target on the feature map, and hence, a higher weight should be given to the feature map. If the variance ratio is high, the focused object in the panel image is not similar to the target on the feature map. In this case, the feature map should be given a lower weight. To accomplish this, we apply the reciprocal of the variance ratio to the weight that is multiplied by the entire feature map within a given tile region, including the object in focus.

### 4.3 Feature integration and normalization

We integrate the seven feature maps into an activation map using the same process as saliency map computation[4]. 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 target-specific visual attention map. The local maximum of the attention map is regarded as the most attracted location of the target search task. Figure 8(a) is a panel image, and the pink flower which is surrounded by a yellow rectangle is a target object. Figure 8(b) is the target-specific visual attention map created

from the panel image. It shows that the other pink flower surrounded by a red rectangle is activated as with the target unlike other red flowers on the map.

# 5. EXPERIMENT AND RESULT

## 5.1 Data set

We employed the MSRA Salient Object Database[7] that includes 1000 images and their binary mask images. We selected 25 images from the dataset in a random manner to create each panel image and placed the object images extracted using their mask images in a $5 \times 5$ grid pattern on a black background. The target image included an object selected from them in a random manner. The size of the panel image and the target image was $1920 \times 1200$ pixels.

Ten participants (nine males and one female) with normal vision, whose ages ranged between 22 and 24 years, participated in our experiment. They were instructed first to observe a target image for five seconds and then search for the target in the paired panel image until they found it. We treated this procedure as a trial of our visual search experiment. We conducted one hundred trials for each participant and recorded their eye movements using a Tobii X60 Eye Tracker (data rate: 60 Hz, accuracy: typical 0.5 degrees) during the trials.

## 5.2 Comparative models

We employed and reimplemented two conventional models to evaluate our proposed model. One was the bottom-up visual attention estimation model, i.e. saliency map model, proposed by Ittiet al.[4]. The other was the target-specific visual attention model proposed by Frintrop et al.[2]. 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 Section 2, Frintrop's model learns the optimum weight of each feature channel from training 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 panel image and apply them to the feature maps of the panel image to estimate a top-down saliency map $S_{td}$ so that we obtain the upper bound performance of Frintrop's model. The top-down saliency map is integrated with the bottom-up saliency map $S_{bu}$ to estimate a global target-specific visual attention map $S_A$. The contribution of each map is adjusted by a top-down factor $w \in [0 \dots 1]$:

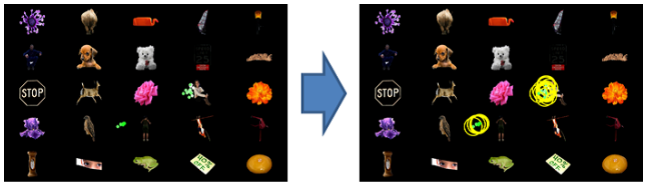
$$S_A = (1 - w) * S_{bu} + w * S_{td} \quad (0 \leq w \leq 1). \tag{14}$$

For $w = 0.5$, bottom-up and top-down cues are evenly regarded, whereas for $w = 1.0$, only top-down cue is considered. We employ two target-specific visual attention maps with $w = 0.5$ and $w = 1.0$ as the comparative models for evaluating our approach.

## 5.3 Evaluation approach

To quantify how well our estimations match the participants' actual eye positions, we use the normalized scanpath saliency (NSS)[10], which is defined as the response value at the current eye position $\vec{x}_{human} = (x_{human}, y_{human}) \in \mathbb{Z}^2$ in a visual attention map $S$ that has been normalized to have zero mean and unit standard deviation:

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



(a)Gaze points (green points) (b)Gaze areas(yellow circles)

Figure 9: Gazed areas considered the error range of eye tracker and central fovea area.

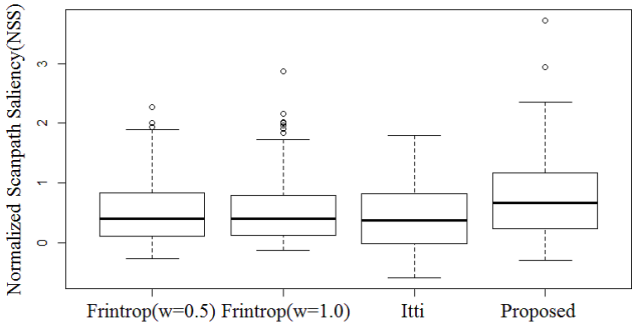



Figure 10: The boxplots of average NSS. The lower edge of the box is the lower quartile and the upper edge is the upper quartile. The circle indicates the outlier.

where $\mu_S$ and $\sigma_S^2$ 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 (visual angle:1.0°) and central fovea area (2.0°). We consider the total range (3.0°) as a gazed area. Figure 9 shows an example of gazed areas. Each green point shows the gazed position, which was recorded through a visual search trial. Each yellow circle, whose center is the green point, shows the gazed area. In this paper, we exploit the average NSS within the circles to evaluate the models.

## 5.4 Experimental result

Figure 10 shows the average NSS across all visual search tasks, i.e., 100 trials $\times$ ten participants. The score for our proposed model was $0.804 \pm 0.072$, which was higher than the other models (bottom-up model: $0.443 \pm 0.057$, Frintrop $w = 0.5$: $0.557 \pm 0.058$, Frintrop $w = 1.0$: $0.573 \pm 0.059$). The average response value in false detection areas, i.e., ungazed areas, for our visual attention map was $1.337 \pm 0.024$, which was lower than that for the other models (bottom-up model: $1.644 \pm 0.011$, Frintrop $w = 0.5$: $2.125 \pm 0.019$, Frintrop $w = 1.0$: $2.136 \pm 0.015$). These results suggest that the visual attention map estimated by our proposed model was broadly consistent with the actual focused area.

Figure 11 shows examples of experimental results obtained from a particular participant. This figure shows that the proposed model estimated top-down visual attention with high accuracy. In the case of Figure 11, each mean NSS of

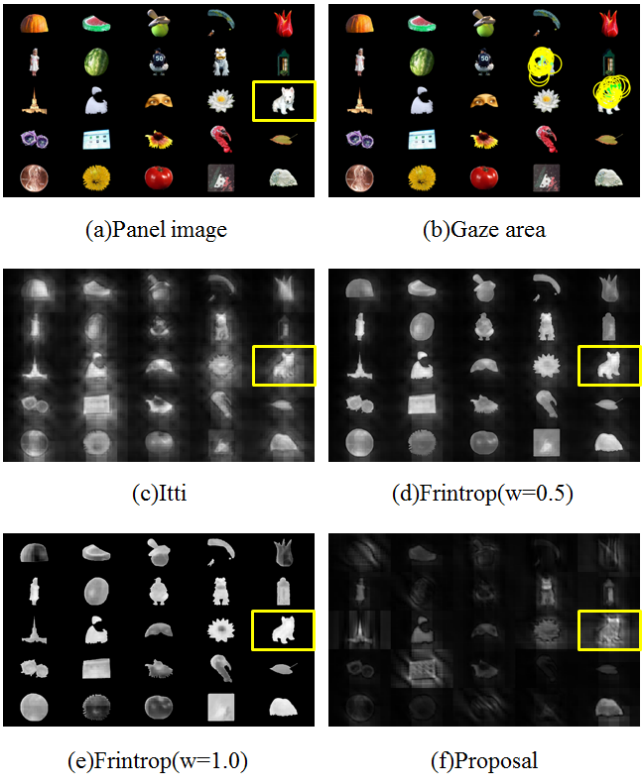



Figure 11: Examples of estimation results with higher NSS.

the model was 0.185 (bottom-up model), 1.520 (Frintrop, $w = 0.5$), 1.432 (Frintrop, $w = 1.0$), and 2.392 (our proposed model).

In situations where the saliency of a target object was low and the saliency of other objects around the target were adequately high, participants might focus their attention on the target even if saliency of target is extremely low. This phenomenon suggests a need to modulate the weights focusing not only on local saliency features, but also on global saliency features of the entire feature map. Further, participants might search for a target with their peripheral vision, especially when the target was located near the center of the panel image. In this case, an accurate estimate of top-down visual attention may not be achieved. Thus, redesigning the layout of objects on the panel image would be helpful to alleviate this problem.

## 6. CONCLUSION

In this paper, we focused on the effect of linear separability between the visual feature distribution of a target object and each of other objects on the visual target search task and proposed a computational model which estimates the target-specific top-down visual attention. We confirmed the effectiveness of our computational model.

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

## Acknowledgment

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.23700168.

## 7. REFERENCES

[1] R. Fisher, "The use of multiple measurements in taxonomic problems," Annals of Eugenics, 7(2):179–188, 1936.

[2] S. Frintrop, G. Backer, and E. Rome. "Goal-directed search with a top-down modulated computational attention system," Pattern Recognition, 3663:117–124. 2005.

[3] 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, 63(5):918–926, 2001.

[4] L. Itti, C. Koch, and E. Niebur. "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. on Pattern Analysis and Machine Intelligence, 20(11):1254–1259, 1998.

[5] A. Kimura, R. Yonetani, and T. Hirayama. "Computational models of human visual attention and their implementations: A survey," IEICE Trans. on Information and Systems, 96(3):562–578, 2013.

[6] C. Koch and S. Ullman. "Shifts in selective visual attention: towards the underlying neural circuitry," Matters of Intelligence, pages 115–141. 1987.

[7] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum. "Learning to detect a salient object," IEEE Trans. on Pattern Analysis and Machine Intelligence, 33(2):353–367, 2011.

[8] Y. Nagai. "From bottom-up visual attention to robot action learning," The 8th IEEE International Conference on Development and Learning, pages 1–6, 2009.

[9] V. Navalpakkam and L. Itti. "An integrated model of top-down and bottom-up attention for optimizing detection speed," IEEE Conference on Computer Vision and Pattern Recognition, pages 2049–2056, 2006.

[10] R. Peters, A. Iyer, L. Itti, and C. Koch. "Components of bottom-up gaze allocation in natural images," Vision Research, 45(18):2397–2416, 2005.

[11] 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," The 15th IEEE International Conference on Intelligent Transportation Systems, pages 1321–1326, 2012.

[12] A. Treisman and G. Gelade. "A feature-integration theory of attention," Cognitive psychology, 12(1):97–136, 1980.

[13] 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, pages 140–145, 2007.