CGI2018 manuscript No.
(will be inserted by the editor)

Visual Attention Prediction for Images with Leading Line Structure

Issei Mochizuki · Masahiro Toyoura · Xiayang Mao

Abstract Researchers have proposed a wide variety of visual attention models, ranging from models that use local, low-level image features to recent approaches that incorporate semantic information. However, most models do not account for the visual attention evident in images with certain global structures. We focus specifically on "leading line" structures, in which explicit or implicit lines converge at a point. Through this study, we have conducted the experiments to investigate the visual attentions in images with leading line structure and propose new models that combine the low level feature of center-surround differences of visual stimuli, the semantic feature of center bias and the structure feature of leading lines. We also create a new data set from 110 natural images containing leading lines and the eye-tracking data for 16 subjects. Our evaluation experiment showed that our models outperform the existing models against common indicators of saliency-map evaluation, underscoring the importance of leading lines in the modelling of visual attention.

Keywords visual attention model · saliency map · structure information · leading lines

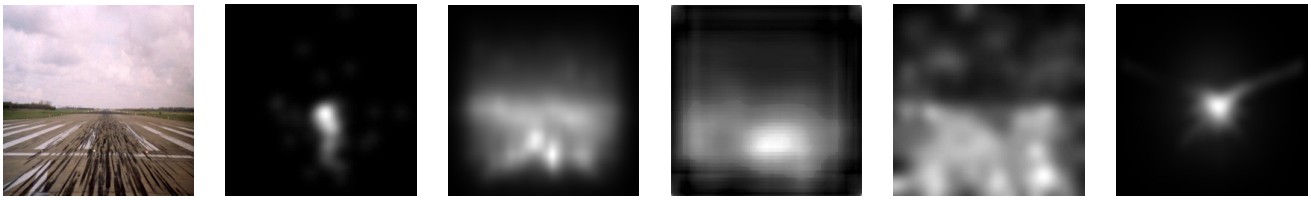
1 Introduction

As a resource for enabling effective image processing and image synthesis that reflect the mechanisms of human visual perception, the visual attention model has drawn considerable interest in graphics rendering, robot vision, advertising design and many other fields of modern visual computing. Researchers have thus far proposed numerous computational models for predicting visual attention. The initial approaches concentrated on "bottom-up" saliency, operating on the biological premise that locations with features distinguishing them from surroundings are more prone to draw people's attention [9,10]. Many studies explored the idea of combining bottom-up features and top-down factors to better predict visual attention [1,4,5,15,18,21,22]. More recent studies have applied deep-learning algorithms to visual saliency models to automatically learn the requisite features for gaze prediction [8,13,17,19,23,24].

Meanwhile, cognitive psychologists have shown that human eyes gravitate towards certain special structures occurring globally in images, such as parallel pairs of lines or lines that converge on a single point. In photographs, structures formed by explicit or implicit lines that converge on a specific point are called "leading lines." Painters and professional photographers often place leading lines in their compositions intentionally to draw the viewer's attention to a particular location. Borji et al. [2] showed that vanishing points found in perspective-projected roads, railroad tracks and tunnels attract attention through two types of eye-tracking experiments. Liang et al. [12] noticed the attraction effect of several kinds of structures, including horizontal line, convex part and vanishing point, and proposed the methods for computing the saliency map by detecting those structures individually. The Liang et al. method, however, fails to predict the effect of leading lines in many cases due to the underlying vanishing point detecting algorithm.

In this paper, we propose new visual attention models for images containing leading lines. Fig. 1(a) shows an image with leading lines, and Fig. 1(b) is the fixation map generated from the eye-tracking data of 16 subjects. Fig. 1(c)–(f) are saliency maps generated via the Itti et al. method [9], Pan et al. method [17], Liang et al. method [12] and our method, respectively. The

I. Mochizuki · M. Toyoura · X. Mao
University of Yamanashi,
4-3-11, Kofu, Yamanashi, 400-8511, Japan
E-mail: mao@yamanashi.ac.jp



(a)Input (b)Eye-tracking (c)Itti (d)SalNet (e)Liang (f)Ours

Fig. 1 An image with leading lines and saliency maps generated with existing methods and the proposed method

proposed method predicts the eye-fixation map in (b) more accurately than the other three existing methods, which fail to predict the accumulation of eye fixations around the leading line convergence. In our evaluation experiment, the models that we proposed and implemented demonstrated higher levels of accuracy than the existing methods did in terms of five common indicators of saliency-map evaluation.

Our study offers the following contributions:

(1) Through our eye-tracking experiment, we investigated the ways in which images containing leading lines attract human attention.
(2) We proposed and implemented a method for generating maps predicting the attention-attraction effects of leading lines.
(3) We proposed new models for integrating the synergy among the bottom-up features from the traditional methods, image-center bias and leading lines.

2 Related Work

Since Itti et al. proposed the first computational saliency map model based on the feature integration theory in 1998, researchers have conducted a wide range of studies aiming to improve and apply the model [3]. The Itti et al. model [9] involves calculating the center-surround differences for three visual features—brightness, color and orientation—at different resolutions, creating a map for each feature and integrating the maps via non-linear weights to produce a saliency map that makes the salient portions of the image more prominent. The Graph-Based Visual Saliency (GBVS) model [7] proposed by Harel et al. involves creating a complete graph with a node for every pixel in an image, using Markov chains that define the weights between nodes in terms of image-feature dissimilarity and analysing the distributions of the Markov chains to calculate the center-surround differences for the brightness, color, and direction. Although many other approaches have been developed, we use Itti et al.'s model and the GBVS model for integrating the bottom-up saliency feature in our proposed visual attention models, as these are widely referred to as the representative bottom-up saliency models.

While traditional methods apply a bottom-up approach, some studies have discussed top-down models that focus on visual attention to high-level factors. These studies assume that attentions first focus on bottom-up visual saliency and then move on to higher-level factors such as objects [1,5,15], actions [18] and events [4, 20]. Marat et al. [16] looked at the effects of semantic information by focusing on the priority attention that subjects paid to faces and center bias. State-of-the-art machine learning techniques have been used to learn the weights for integrating different features or learn from the eye-tracking data the effect of semantic features for attracting visual attentions [21,22].

Recently, several studies have been conducted to investigate how structure features can affect visual attention. Borji et al. [2] conducted two types of gaze-tracking experiments and showed that vanishing points attract attentions in both free-viewing tasks and visual search tasks. Liang et al. [12] conducted an eye-tracking experiment using 500 images of special structures and found that in addition to the vanishing point, the horizontal line and convex part also attract attention. They also proposed a method for generating the saliency map by detecting these features individually and combining them with the weights obtained through multi-kernel learning. Their method detects vanishing points by first detecting the horizon and finding the point with a geometric feature most similar to the predefined feature of a vanishing point. Their method assumes vanishing points are on the boundary between sky and ground, which is not true for many cases. As shown in Fig. 1(e), their method may also fail for leading lines which are vague or not continuous. Our method extended the adaptive soft voting technique in [11] to directly create maps reflecting the degree of attention caused by leading lines. It does not impose the assumption on the presence of sky region or horizon, but works for any images with lines converging at a point, even if those lines are not clearly delineated or continuous.

Another popular research topic in recent years has been applying deep learning to saliency models. Us-

ing eye-tracking data, Liu et al. [13] directly trained a Multiresolution Convolutional Neural Network (Mr-CNN) from fixation area and non-fixation area. Similarly, Huang et al. [8] repurposed a convolutional neural network (CNN) for object recognition as a system for saliency. Pan et al. [17] proposed the SalNet approach, which formulates saliency estimation as an end-to-end regression. To the best of our knowledge, all the current approaches based on deep learning, however, could not produce good results for images with leading line structures, which is probably due to the lack of train images containing leading line structures. We chose to design the proposed method with handcrafted features so that we could better elucidate the relationships between leading line structures and other features.

3 Proposed Visual Attention Models

3.1 Visual Attention in Images with leading lines

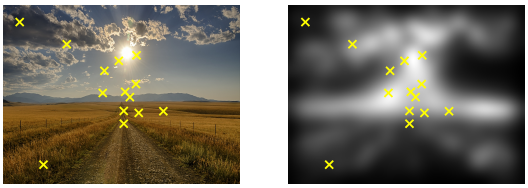
Our experiments have revealed the following features of visual attention for images with leading lines.

Attraction Effects of leading lines and Low-Level Features. Borji et al. [2] showed that vanishing points found in perspective-projected roads, railroad tracks and tunnels attract attention through two types of eye-tracking experiments. Liang et al. [12] noticed the attraction effect of several kinds of structures, including horizontal lines, convex parts and vanishing points. The images on the left sides (a) of Fig. 2 and Fig. 3 are superimpositions showing the fixation points of 16 subjects. The images on the right sides (b) of Fig. 2 and Fig. 3, meanwhile, are the Itti et al. model-based saliency maps for their corresponding source images. A look at Fig. 2(a) shows that the subjects' fixations tended to gather around the convergence of the leading lines, a trend that the Itti et al. model-based saliency map did not anticipate. The Itti et al. model evidently produced high values in locations that had different directional properties or stronger contrast levels than their surrounding areas. Some of the subjects looked at the sun above the convergence of the leading lines in Fig. 3(a). In other words, some of the fixation positions matched the salient positions in the Itti et al. saliency map. These experiment results showed that improving the accuracy of fixation estimations would require taking both the effects of leading lines and the effects of center-surround differences into account.

Synergy between leading lines and the Center-Surround Differences of Visual Stimuli. We have observed in our experiments that subjects have a particularly strong tendency to focus their attention on locations where leading lines and visual stimuli with large center-surround differences overlap. In Fig. 4(a), where the convergence of leading lines and the areas with large center-surround differences are in different locations, the fixation distribution extends to locations without any especially prominent features. In Fig. 4(b), however, the leading

(a)Fixation points (b)Itti at el. saliency map for (a)

Fig. 2 Attention-attraction effect at the convergence of leading lines



(a)Fixation points (b)Itti at el. saliency map for (a)

Fig. 3 The effects of bottom-up features. Both leading lines and strong center-surround differences attract attention.

(a)leading lines converge at a point far away from the area with large center-surround difference (b)leading lines converge at a point close to the area with large center-surround difference

Fig. 4 The synergy effect of leading lines and bottom-up features. Attention-attraction effect is enhanced when the convergence of leading lines overlaps with the large center-surround difference area of the visual stimulus.

(a) Leading lines convergence at the center of the image (b) Leading lines converge at a position away from the center of the image

Fig. 5 The synergy effect of leading lines and image-center bias. The attention-attraction effect is enhanced when leading lines converge at the center of the image.

line and the areas with considerable center-surround differences are relatively close together. The subjects' fixations thus congregated in that area.

The Effect of Center Bias. Several studies on visual attention have established that human fixations tend to align with the center of an image[20]. In the study by Tatler et al. [20], who conducted an eye-tracking experiment using a set of 120 landscape images with a distribution of features across a wide array of different locations, the results showed that humans tend to concentrate their visual perception on the center of an image regardless of the nature of the experiment task. As the experiment results in Fig. 5 suggest, leading lines converging near the center of an image have stronger attention-attraction properties.

3.2 Proposed Visual Attention Models

Based on the above-mentioned observations, our study proposes the following three visual attention models to elucidate the relationships between leading lines and other types of features:

1. A model that integrates the effects of leading lines and the effects of center-surround differences
2. A model that accounts for the synergy of leading lines and center-surround differences
3. A model that factors in the additional element of center bias

The following section provides details on each model.

3.2.1 Integrating a leading line Saliency Map and a Center-Surround Difference Saliency Map

Operating on the knowledge that both areas around the convergence of leading lines and areas with large center-surround differences attract attention, we first generate a leading line saliency map (M_l) and a center-surround difference saliency map (M_s) for an input image. We then weight the two maps via Formula (1) below to create saliency map M.

$$M = kM_s + (1-k)M_l \tag{1}$$

Here, $0 \leq k \leq 1$ represents a coefficient for determining the weights of the center-surround difference saliency map and the leading line saliency map in the resulting saliency map. For our study, we determine the optimal value of coefficient k by conducting and eye-tracking experiment with images containing leading lines, generate fixation map M^E based on the eye-tracking data and then use fixation map M^E as example data. Then k can be computed as follows:

$$k = \arg\min_{k} \sum_{n=1}^{N} |kM_{sn} + (1-k)M_{ln} - M_n^E| \tag{2}$$

n represents the number of images in the example data set.

3.2.2 Adding Synergy

Operating on the assumption that the spatial consistency of leading lines and center-surround differences boosts saliency in a non-linear fashion, we thus incorporate the product of the center-surround difference saliency map M_s and leading-line saliency map M_l into the model of Formula (2) to create a model that reflects the synergy between the two features:

$$M = k_1M_s + k_2M_l + (1-k_1-k_2)\sqrt{M_sM_l} \tag{3}$$

k_1 and k_2 are coefficients for determining the respective weights of the individual maps in the integrated map. Like the models that we noted above, we used fixation map M^E, which we created based on the subjects' fixations, for our training data, then used Formula (4) below to solve for the minimum value of a constrained nonlinear multivariate ($0 \leq k \leq 1$) and thereby estimated the optimal value of $\boldsymbol{k} = (k_1, k_2)$.

$$\begin{aligned}\boldsymbol{k} = \arg\min_{\boldsymbol{k}} \sum_{n=1}^{N} &(k_1M_{sn} + k_2M_{ln} \\ &+ (1-k_1-k_2)\sqrt{M_{sn}M_{ln}} - M_n^E)^2\end{aligned} \tag{4}$$

3.2.3 Adding Synergy and Center Bias

Our experiments observed that fixations would gather at the center of an image containing leading lines. To consider this effect, we generated center-bias map M_c, where locations nearer the center of the map have higher levels of saliency. We then integrated map M_c with leading line map M_l and center-surround difference map M_s via Fomula 5 to reflect the effects of both synergy and center bias.

$$\begin{aligned}M =& k_1M_s + k_2M_l + k_3M_c \\ &+ (1-k_1-k_2-k_3)\sqrt[3]{M_sM_lM_c}\end{aligned} \tag{5}$$

As we did for the model that incorporated synergy, we find the value of weight $\boldsymbol{k} = (k_1, k_2, k_3)$ by solving for the minimum value of a constrained nonlinear multivariate ($0 \leq k \leq 1$).

$$\begin{aligned}\boldsymbol{k} = \arg\min_{\boldsymbol{k}} \sum_{n=1}^{N} &(k_1M_{sn} + k_2M_{ln} + k_3M_{cn} \\ &+ (1-k_1-k_2-k_3)\sqrt[3]{M_{sn}M_{ln}M_c} - M_n^E)^2\end{aligned} \tag{6}$$

4 Generating Maps

4.1 Generating the leading line Map with Adaptive Soft Voting

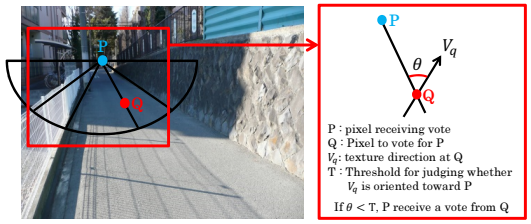
Our eye-tracking experiment showed that fixations tend to gather towards the convergence of leading lines. To represent the attention-attraction effect of leading lines, then, one needs to generate a map in which saliency increases in proportion to the proximity to the leading line convergence point. We extended the vanishing point detection algorithm that Kong et al. [11] applied to detect road regions in perspective projections in order to create leading line saliency map M_l.

Given the difficulties of identifying roads—especially off-road paths—in an image as uninterrupted lines or curves, Kong et al. forgoes trying to detect straight lines and curves and instead selects points at which the orientations of numerous pixels are pointing as vanishing point candidates. The direction of the texture at each pixel can be detected with Gabor filter banks. Fig. 6 provides a more detailed look at the process. Pixel P is a vanishing point candidate, while Q represents the pixels within a large semicircle with pixel P at its center. If the angle θ formed by the orientation V_q at a pixel Q and straight line QP is within a given threshold, pixel P receives a vote. If a certain pixel has a much higher vote count than the other pixels in a given image, then the texture directions of many pixels in the image are pointing towards that particular pixel—or, in other words, the vanishing point. Because the voting value each pixel receives shows how many explicit or implicit lines are oriented towards that pixel and it should have a significant correlation with the attraction effect of leading lines. Thus, our study uses a voting map (see Fig. 6(b)) obtained via voting processing to create a map representing the attraction effect of leading line.

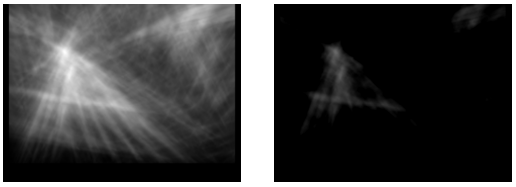
As Fig. 6(b) shows, a voting map renders the leading line convergence points in higher pixel values while also assigning certain values to orientations in other areas that are explicitly or implicitly heading towards convergence points. To emphasise the nonlinear property of visual attention, we apply a DoG (Difference-of-Gaussian) filter to the map of voting results to produce a map that not only emphasises the saliency of leading line convergence points in a nonlinear manner but also suppresses the saliency of other areas.

4.2 Generating the Center-Surround Difference Map

To create the center-surround difference saliency map for our study, we used the Itti et al. model and Harel et al. model [7]. For Harel et al.'s model, their GBVS tool [6] is used to generate center-surround difference saliency map M_s (Fig. 7(c)).



(a)Adaptive soft voting



(b)Voting map (c)leading line map:M_l

Fig. 6 leading line map generation using adaptive soft voting.

4.3 Generating the Center-Bias Map

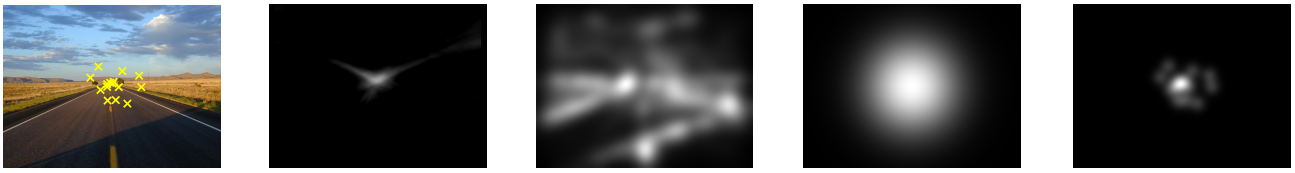
Marat et al. [16] showed that an image with Gaussian kernels at the center and a variance (σ) that matches a viewing angle of 10° is an effective rendering of a saliency map illustrating center bias. Taking our experiment environment into account, we generated center-bias map M_c (Fig. 7(d)), which has a σ of 186 pixels.

4.4 Generating the Fixation Map

To create fixation map M^E, which would serve as our training data for learning weights for integrating different features, we placed Gaussian kernels at the fixation locations from the eye-tracking experiment at a variance of σ so that the half width at half maximum (HWHM) would match the range of the fovea (viewing angle: 2°). Fig. 7(e) is an example of a fixation map. In our experiment environment, the on-screen width of an image at a viewing angle of 2° is 37.5 pixels. We thus set σ to 32 pixels.

5 Collecting Eye-Tracking Data

We used an eye tracker (Tobii X2-60) to record the eye movements of 16 subjects. As Fig. 8(a) shows, we had the subjects sit 60 cm from the display and look freely at the images without immobilising the subjects' heads. The stimuli that we presented to the subjects comprised 110 images containing leading lines and 88 dummy stimulus images. We collected the 110 images containing leading lines based on two conditions. First, our image set needed to have leading lines that converged not only on non-central locations but also on

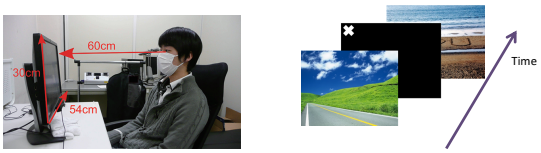


(a)Original image (b)leading line map: M_l (c)Center-surround difference map:M_s (d)Center bias map: M_c (e)Fixation map:M^E

Fig. 7 The constituting feature maps of the proposed saliency models

a wide-ranging distribution of points throughout the images; Fig. 8(b) is one example. By incorporating different convergence locations, our approach obtains eye-tracking data that factors in the effects of location on directing observer gaze. The other condition was that our images needed to include bottom-up salient areas (areas with significant center-surround differences) besides the leading lines, as Fig. 3(a) shows. Using these types of images makes it possible to compare the effects of leading lines and center-surround differences. We also observed that seeing multiple images containing leading lines in succession could induce the fixations towards leading lines. To prevent the experiment methods from interfering with our objectives, we displayed images containing leading lines and dummy images in a random order for 3 seconds each. After each stimulus image, the subjects saw a visual calibration image with an "x" mark in one corner randomly chosen from the four corners for another 3 s before the next stimulus image appeared (see Fig. 8(b)). The tasks that we gave to the subjects were to look freely at the stimulus images and focus on the "x" marks when the visual calibration images were visible. By thus detecting fixation data away from the "x" mark in the gaze image, we were able to cancel out the effects on the initial position of the subjects' fixations.

The saliency map is for modelling the visual attention in a short period, for example, within less than 200 ms, after a stimulus is presented. In principle, therefore, one would need to use fixations from within a given period of time (generally 200 ms) after the subject sees the stimulus. However, initial fixations for a given image are susceptible to the after-image effects of the previous stimulus. The time it takes to follow a scan path to areas of interest in an image also varies from person to person, making it difficult to set a threshold for the initial time period. To address these problems, our study uses the second fixations obtained after stimulus presentation. Fig. 9(a) shows a visualization of ordered fixation data. The figure shows that the first fixations of most subjects tended to congregate at the top-left area of the image—the location of the "x" mark in the visual calibration image (Fig. 9(b))—that the subjects



(a)Observation environment (b)Stimulus

Fig. 8 Experiment environment

(a)Order of fixations (b)Visual calibration image

Fig. 9 Fixations and their relation to the visual calibration image

saw immediately before the stimulus image. The second fixations, however, almost always moved away from the "x" mark.

6 Experiment

6.1 Obtaining Learning Results and Generating Saliency Maps

To learn weight k for the proposed models and evaluate the models, we used the fixations that the 16 subjects exhibited in response to 110 images containing leading lines. We divided the 110 images into 11 groups of 10 images each and conducted 11-fold cross-validation, using 10 sets for learning and 1 set for evaluation. Table 1 shows the learnt average values of coefficient k for each model. The Our1 model corresponds to the weighted integration of the leading line and center-surround difference maps given by Formula (1) in Section 3.2.1. The Our2 model, which is represented with Formula (3) in Section 3.2.1, adds the element of synergy to Our1, and, finally, Our3 represented with Formula (5) in Section 3.2.3, incorporates the element of center bias into Our2. The parenthetical information indicates the saliency maps that we used for center-surround differences. The learning results in Table 1 shows that the

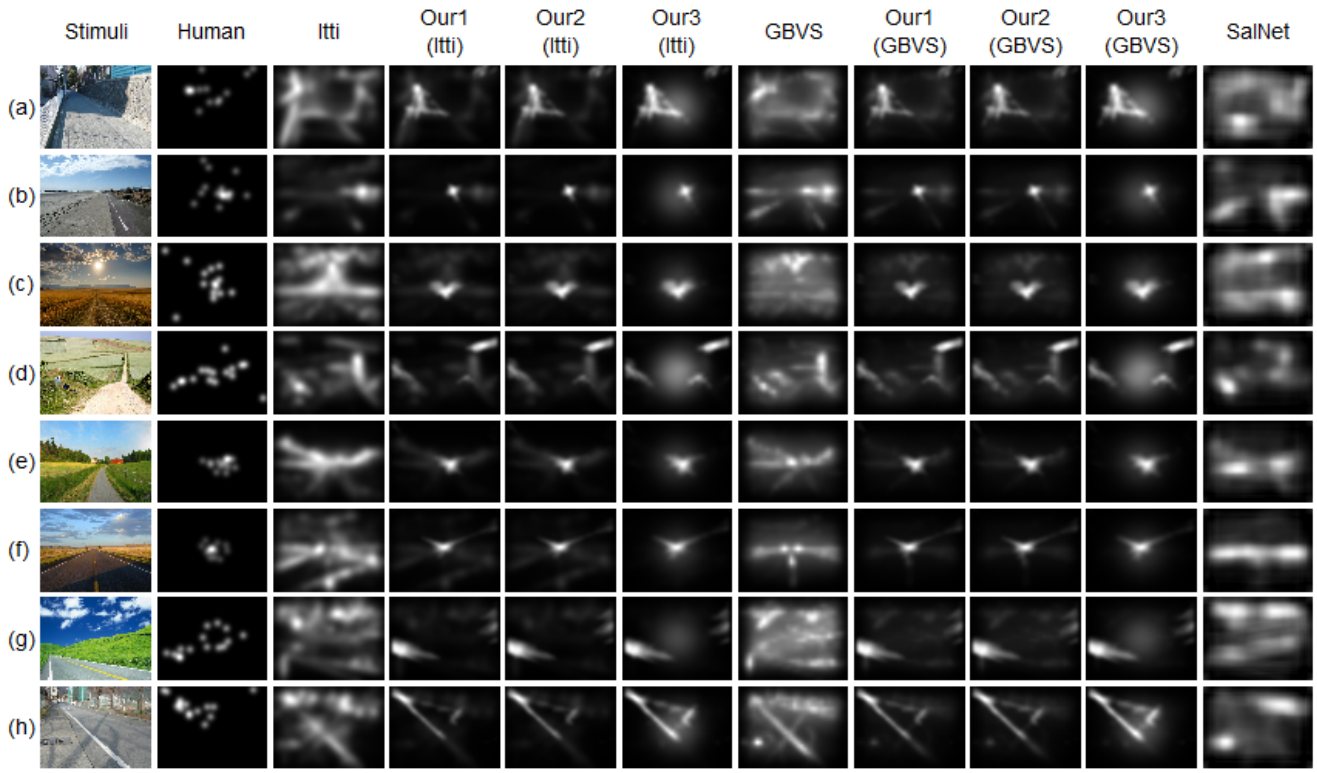


Fig. 10 Saliency maps generated via existing models and the proposed models for images containing leading lines

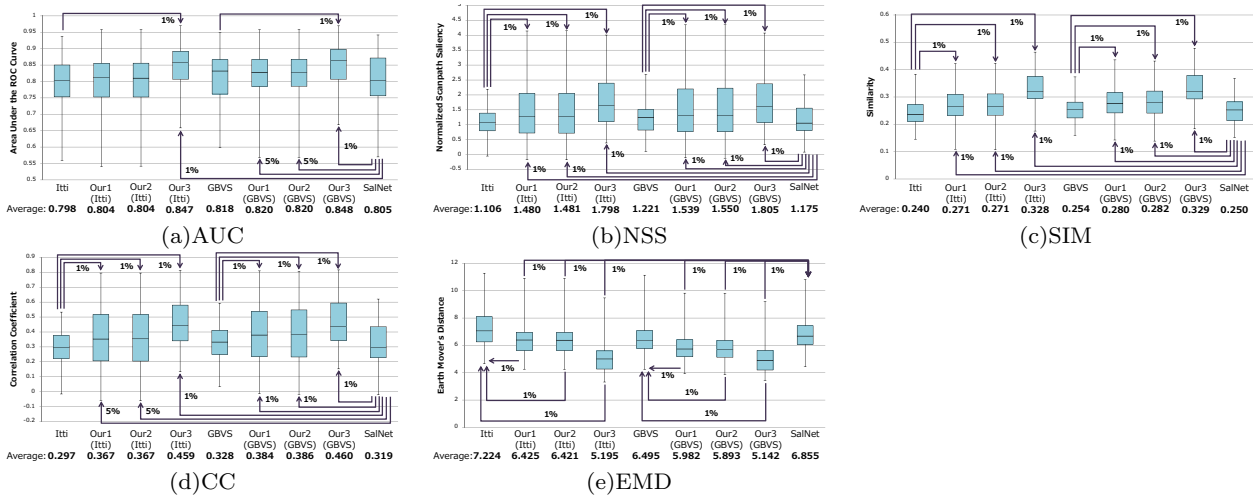


Fig. 11 Comparison of the proposed method and existing methods by MIT Saliency Benchmark [15]

coefficients for the leading line maps are higher than their counterparts in all of the proposed models. In the Our2 model, the value of the coefficient for the synergy term was low. On the other hand, the value is higher than that of the coefficient for the center-surround difference map in the Our3 model that accounted for center bias. These findings suggest that leading line convergence points and areas with large center-surround differences attract more attention when they lie near the center of a given image.

Fig. 10 shows saliency maps created via the proposed method and the existing methods, providing an intuitive comparison of the various approaches. The "Human" column in Fig. 10 shows fixation maps of the subjects. Looking at Fig. 10, one can see that the proposed models were more accurate in predicting the subjects' actual fixations than the Itti et al. model [9],

Table 1 Average weights

	Itti	Leading Line	Center Bias	Synergy
Our1(Itti)	0.0985	0.9015		
Our1(GBVS)	0.1167	0.8833		
Our2(Itti)	0.0980	0.8843		0.0176
Our2(GBVS)	0.1136	0.7846		0.1019
Our3(Itti)	0.0166	0.6779	0.1247	0.1809
Our3(GBVS)	0.0242	0.6526	0.1191	0.2040

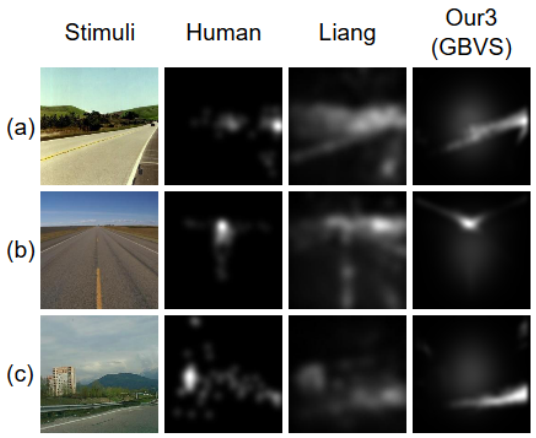


Fig. 12 Comparison with the method of Liang et al. [12]

the Harel et al. GBVS model [6] and the Pan et al. SalNet model [17]. When dealing with images containing salient objects such as the utility pole in Fig. 10(a) and the sun in (c), the existing methods predict that these areas will draw the most attention. The actual fixation map, however, shows that people's gazes fell primarily around the corresponding leading line convergence points. In Fig. 10(d) and (g), the leading lines converge in a location outside the center of the respective images. While Our1 and Our2 failed to predict the locations of the subjects' fixations in the center of the image, Our3 identified the image-center location as a more salient area. Incorporating the element of center bias clearly enabled Our3 to predict that the subjects' fixations would fall more towards the center.

The time required for computing a leading line map for a 240×180 image using adaptive soft voting is 82.15 s on a computer with an Intel Corei7-5600U 2.6 GHz CPU, 16 MB main memory and a GeForce 940M GPU. The time increases drastically when the size of the image increases and is the major part that impacts the time performance of the proposed models. Currently, we perform the adaptive soft voting on a smaller image and upscale the resulting map when integrating it with bottom-up and center-bias maps.

6.2 Quantitative Evaluation

To conduct a quantitative comparison of the proposed method and the existing methods, we used the evaluation technique from the MIT Saliency Benchmark [14]. Fig. 11 shows the average 11-fold validation scores and box-plots for five common indicators of saliency-map quality. The arrows in the figures provide a comparison of the average values of one-tailed t-tests. For each indicator, the AUC represents the area under the ROC curve where the detection rate is the percentage of pixels that are (1) equal to or greater than the threshold and (2) on a fixation location. The false detection rate is the percentage of pixels not on a fixation location. NSS (Normalized Scanpath Saliency) corresponds to the average values of the pixels on the fixations in a normalized saliency map. SIM (Similarity) represents the similarity of the saliency map and fixation map distributions, and CC (Correlation Coefficient) is an assessment of the correlation between the saliency map and fixation map. EMD (Earth Mover's Distance) is a measure of the distance between two distributions. For our study, EMD is the optimal amount of work for matching the two maps in terms of pixel-to-pixel distance and the amount of weight in the corresponding movement, with the pixel values in the maps acting as weights.

A high-accuracy model has high AUC, NSS, SIM and CC values and a low EMD value. The five indicators fall into two general categories: approaches that focus on fixation location (AUC and NSS) and approaches that concentrate on fixation distribution (SIM, CC and EMD). As Fig. 11 shows, the proposed models demonstrated better performance than the Itti et al. method and GBVS method in terms of all five evaluation indicators. In the results for the NSS, SIM and EMD indicators, the proposed models exhibited significant differences at a significance level of 1%. All of the proposed models also showed differences in terms of CC, which suggests that accounting for leading lines enables estimations that better approximate actual human fixations. Our3, which incorporated the element of center bias, produced the highest-precision results. As Fig. 11(a) suggests, center bias plays a pivotal role in boosting fixation-location detection rates. Fig. 11(c), (d) and (e) also reveal important findings. While the AUC results for Our1 and Our2 failed to produce better evaluation values than the conventional center-surround difference method at any degree of significance, the results for SIM, CC and EMD—indicators that operate on fixation distribution—all gave the proposed models significantly better assessments. As our maps focused on leading lines not only in terms of their convergence

(a) Stimulus (b) Saliency map (c) Human

Fig. 13 Example of prediction failure. (Our3 with GBVS)

points but also in light of their attraction effects at the pixel level, the salient areas in the maps took on a slightly larger scope. The SIM, CC and EMD results reflect our approach's ability to deliver accurate estimates of fixation distributions, a capability that comes from that expanded range. Compared to the Pan et al. SalNet method, which applies deep-learning techniques, Our1, Our2, and Our3 all had significantly higher evaluation scores for all the indicators other than AUC. Although not applicable to all of our proposed models, the AUC results still exhibited a significant preference for the proposed model with a GBVS model map serving as the source of center-surround difference effects. This is likely due to two factors. First, SalNet and the many other saliency models based on deep learning have network designs that focus on two levels—bottom-up features and top-down factors—and not special structures such as leading lines. The second contributing factor is that the sets of learning images in deep-learning models do not contain many leading lines; thus, the models encounter difficulties learning the attention-attraction effects of leading line structures.

6.3 Comparison with the Method of Liang et al. [12]

Figure 12 shows the saliency maps generated with the proposed method and the method of Liang et al. [12]. For the images in (a) and (b), the maps by Liang et al. failed to predict the fixations around the convergence of leading lines. Their method detects a vanishing point by first finding the highest point on the horizon, and hence the method may fail when the vanishing point is not on the horizon (Fig. 12(a)) or when the horizon is relative flat (Fig. 12(b)). However, Liang et al. also considered the attention-attraction effect of convex parts. Their method can predict fixations better than our method for the images containing convex parts, such as the building in Fig. 12(c).

7 Limitation and Future Work

Fig. 13 shows two examples with which our method failed to predict the visual attentions correctly. No fixations fall around the convergence of leading lines in Fig. 13(a), although the saliency map generated with our method (Our3 with GBVS) shows a high value at that area (Fig. 13(b)). We found that the effect of leading lines becomes weak when the structure is far away from the center of the image. In this case, the semantic feature (human) and the structural feature of the waiting space and the pole may also strongly attract the subjects' attention. In Fig. 13(c), one can see that most fixations fall on the object in front of the convergence of leading lines. Actually, we have observed that as attentions are attracted towards the convergence along the leading line, if there is some object in front of the convergence, attentions tend to be drawn to that object. We need to experiment with more images to develop a more sophisticated model for predicting such features accurately.

One major limitation of the proposed method is that it cannot be applied to images without leading lines. An easy way to alleviate this limitation is to first examine whether the image contains leading lines and then apply either the proposed models (if the image contains leading lines) or conventional models (if the image does not contain leading lines). Our study revealed the relationship among leading line structure, bottom-up feature of center-surrounding and image-center bias by using handcraftec features and learning the contributions of each feature from eye-tracking data. Fusing our findings into end-to-end models that utilise recent deep-learning techniques would make it possible to develop a more general saliency model capable of predicting the comprehensive effects of bottom-up, structural and semantic features.

Compliance with Ethical Standards

Funding: This study was funded by JSPS Grants-in-Aid for Scientific Research (Grant No. 17H00738, 16K12459).

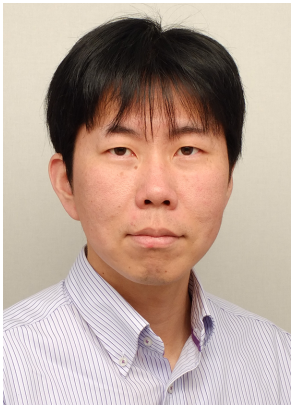
Conflict of Interest: Issei Mochizuki declares that he has no conflict of interest. Masahiro Toyoura declares that he has no conflict of interest. Xiaoyang Mao declares that she has no conflict of interest.

References

1. A. Borji. Boosting bottom-up and top-down visual features for saliency estimation. In CVPR, 2012.
2. A. Borji, M. Feng, and H. Lu. Vanishing point attracts gaze in free-viewing and visual search tasks. J. Vision, 16(14):18, 2016.

3. A. Borji, M. M. Cheng, H. Jiang, J. Li, Salient object detection: A survey, arXiv preprint arXiv:1411.5878, 2014.
4. R. Carmi, and L. Itti. Visual causes versus correlates of attentional selection in dynamic scenes. Vision Research, 46(26):4333-4345, 2006.
5. M. Cerf, J. Harel, W. Einhäuser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. In NIPS, 2008.
6. J. Harel, A Saliency Implementation in MATLAB: http://www.klab.caltech.edu/ harel/share/gbvs.php
7. J. Harel, C. Koch, and P. Perona, Graph-Based Visual Saliency, Proceedings of Neural Information Processing Systems (NIPS), pp.545-552, 2006.
8. X. Huang, C. Shen, X. Boix, and Q. Zhao, Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks, In ICCV, pp.262-270, 2015.
9. 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.
10. L. Itti, and C. Koch, Feature combination strategies for saliency-based visual attention systems, Journal of Electronic Imaging, pp.161-169, 2001.
11. H. Kong, J. Y. Audibert, and J. Ponce, Vanishing point detection for road detection, In CVPR, pp.96-103, 2009.
12. H. Liang, M. Jiang, R. Liang, and Q. Zhao. Saliency Prediction with Scene Structural Guidance. 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Banff Center, Banff, Canada, October 5-8, 2017.
13. N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, Predicting eye fixations using convolutional neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.
14. T. Judd, F. Durand, and A. Torralba, A benchmark of computational models of saliency to predict human fixations, MIT Technical Report, MIT-CSAIL-TR-2012-001, pp.545-552, 2012.
15. T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In ICCV, 2009.
16. S. Marat, A. Rahman, D. Pellerin, N. Guyader, and D. Houzet, Improving Visual Saliency by Adding 'Face Feature Map' and 'Center Bias', Cognitive Computation, vol.5, no.1, pp.63-75, 2013.
17. J. Pan, E. Sayrol, X. Giro-i-Nieto, K. McGuinness, and N. E. O'Connor, Shallow and deep convolutional networks for saliency prediction. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
18. S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua. An eye fixation database for saliency detection in images. In ECCV, 2010.
19. C. Shen, M. Song, and Q. Zhao. Learning high-level concepts by training a deep network on eye fixations. In NIPS Deep Learning and Unsup Feat Learn Workshop, 2012.
20. B. W. Tatler, The Central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions, J. Vision, vol.7, issue 14, article 4, 2007.
21. J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, Predicting human gaze beyond pixels, J. Vision, vol.14, no.1, pp.28-28, 2014.
22. Q. Zhao, and C. Koch. Learning Visual Saliency By Combining Feature Maps in a Nonlinear Manner Using AdaBoost, J. Vision, vol.14, issue 6, article 22, pp.1-15, 2012.
23. W. Wang, J. Shen, L. Shao, Video salient object detection via fully convolutional networks. IEEE Transactions on Image Processing, vol.27, no.1, pp.38–49, 2018.
24. W. Wang, J. Shen, Deep visual attention prediction. IEEE Transactions on Image Processing, vol. 27, pp. 2368–2378, 2018.

Issei Mochizuki received the B.Sc. degree in Engineering from University of Yamanashi, Japan. He is currently a Master Course Student at Department of Computer Science and Engineering, University of Yamanashi, Japan. His research interest includes visual attention.



Masahiro Toyoura received the B.Sc. degree in Engineering, M.Sc. and Ph.D. degrees in Informatics from Kyoto University in 2003, 2005 and 2008 respectively. He is currently an Associate Professor at Department of Computer Science and Engineering, University of Yamanashi, Japan. His research interests are augmented reality, computer and human vision. He is a member of ACM and IEEE Computer Society.

Xiaoyang Mao received her B.Sc. in Computer Science from Fudan University, China, M.Sc. and Ph.D. in Computer Science from University of Tokyo. She is currently a Professor at Department of Computer Science and Engineering, University of Yamanashi, Japan. Her research interests include texture synthesis, non-photo-realistic rendering and their application to scientific visualization. She is a member of ACM SIGRRAPH and IEEE Computer Society.