Incorporating Visual Field Characteristics into a Saliency Map

Hideyuki Kubota, Yusuke Sugano,
Takahiro Okabe, Yoichi Sato
The University of Tokyo
{hk99,sugano,takahiro,ysato}@iis.u-tokyo.ac.jp

Akihiro Sugimoto
National Institute of Informatics
sugimoto@nii.ac.jp

Kazuo Hiraki
The University of Tokyo
khiraki@idea.c.u-tokyo.ac.jp

Abstract

Characteristics of the human visual field are well known to be different in central (fovea) and peripheral areas. Existing computational models of visual saliency, however, do not take into account this biological evidence. The existing models compute visual saliency uniformly over the retina and, thus, have difficulty in accurately predicting the next gaze (fixation) point. This paper proposes to incorporate human visual field characteristics into visual saliency, and presents a computational model for producing such a saliency map. Our model integrates image features obtained by bottom-up computation in such a way that weights for the integration depend on the distance from the current gaze point where the weights are optimally learned using actual saccade data. The experimental results using a large number of fixation/saccade data with wide viewing angles demonstrate the advantage of our saliency map, showing that it can accurately predict the point where one looks next.

Keywords: saliency, eye movements, visual fields, fovea vision, peripheral vision, learning

1 Introduction

For many applications including human-computer interaction and multimedia systems, understanding where humans look in a scene is essential. Human eye movement is believed to be driven not only by top-down cues that depend on one's intention and purpose of actions but also by bottom-up cues that depend on visual stimuli coming from the surrounding environment.

To predict eye movement, many studies on bottom-up approaches have investigated what kinds of visual stimuli attract human gaze/visual attention. The most popular hypothesis is that humans look at a salient region whose low-level image features are significantly different from those in other regions. Itti *et al*. [1998] proposed a model for computing how likely each point in an image attracts gaze/visual attention, *i.e*., a saliency map. Their original model first extracts low-level image features such as intensity, color, and orientation from a given image. These low-level features are then converted to feature maps by using center-surround filtering with different spatial scales, and a saliency map is computed by integrating feature maps.

Extensions from the seminal work by Itti *et al*. [1998] have been intensively studied through the decade. Alternatives to center-surround mechanisms such as information maximization [Bruce and Tsotsos 2006] and graph representation [Harel et al. 2007] were proposed. Itti *et al*. [2003] extended the model to be applicable to a video by incorporating low-level dynamic features such as motions and flickers. Cerf *et al*. [2008] proposed combining face detection with a saliency map computed from low-level features. However, the mechanism of the bottom-up visual attention has not yet fully understood, and the fixation prediction accuracy of these models is not necessarily high. In this work, we focus on one of the important factors to be considered, *i.e*., characteristics of the human eyes.

Biological studies show that the density of a photoreceptor decreases the further away from the fovea it gets [Curcio et al. 1990]. Therefore, the spatial resolution of a perceived image is highest in the fovea while it is lower in the peripheral visual fields. Groll and Hirsch [1987] reported that the field of view of the fovea with the highest spatial sensitivity corresponds to four degrees. Virsu and Rovamo [1979], on the other hand, measured the contrast sensitivity for discriminating the direction of movent or orientation of sinusoidal gratings both in central and peripheral vision. Their result indicates that the cut-off frequencies of all tasks decrease and the location peaks of the contrast sensitivity functions shift towards lower spatial frequancies as going away from the fovea. Boynton *et al*. [1964] reported that differences exist in the hue measurements, in which the absolute color-naming procedure is utilized, depending on the region of the field of view.

Most of existing saliency map models do not take into account the differences between a fovea and a peripheral visual fields and handle the entire visual field uniformly, with few exceptions. Parkhurst *et al*. [2002] proposed a model that assigns lower saliency values as the distance from the current fixation point becomes larger. In Vincent *et al*.'s model [2007], input images are converted to retinal images, and feature maps are computed from retinal images. However, feature dependency of the visual field characteristics is not fully explored in these relatively simple models.

In contrast, we propose a novel computational model of a saliency map based on the characteristics of the human visual field. In the proposed model, feature maps are integrated with varying weights determined based on the distance from the current fixation point according to the visual field characteristics. Since it is not always easy to assign the optimal weights according to the visual field characteristics, we tackle this problem using a data-driven approach following prior works [Kienzle et al. 2007; Judd et al. 2009; Zhao and Koch 2011]. Optimal set of weights for image features are learned by using a large set of actual saccade data. This leads to a more accurate saliency map model to predict the next fixation location, and it has a great potential to applications like attentive user interfaces and navigation systems.

The main contribution of this work is two-fold: (1) We propose a novel computational model for a saliency map that incorporates the characteristics of the human visual field. (2) We make our own eye movment dataset, in which the actual fixation data by 15 subjects are included, available online for further studies.

ETRA 2012, Santa Barbara, CA, March 28 – 30, 2012.
© 2012 ACM 978-1-4503-1225-7/12/0003 $10.00

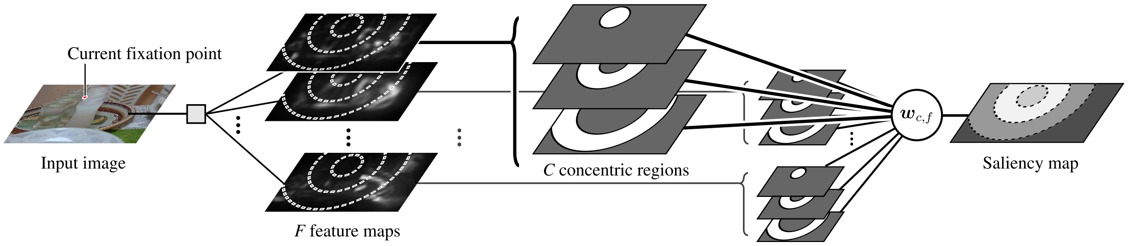

Figure 1: *Flow of the proposed framework. F feature maps are divided into C concentric regions around the current fixation point. Then, these maps are integrated into a final saliency map with different weights for each region of each map.*

Figure 2: *Samples of the images included in our dataset.*

2 Learning Saliency Map with Visual Field Characteristics

The purpose of our work is to explore a computational model of a visual saliency map in which the characteristics of the human visual field are taken into account. Given the current fixation point, our model computes a saliency map for better prediction of the next fixation point. Figure 1 illustrates the flow of our model. Unlike existing models that use spatially flat weights to integrate different feature maps, we propose using varying feature weights according to the current fixation point. A total of F feature maps are divided into C concentric regions centered around the current fixation point. By varying weight values to integrate these concentric regions into the final saliency map, we optimize the model according to the visual field characteristics of humans using a dataset consists of actual saccade data.

2.1 Dataset

The field of view of the human eye is approximately 200 degrees left to right and 120 degrees up and down [Henson 1993]. However, existing eye movement data sets were collected using a relatively narrow field of view (at most 36 degrees in viewing angle) [Judd et al. 2009; Cerf et al. 2009; Ramanathan et al. 2010; Bruce and Tsotsos 2009]. Thus, they cannot be used for our purpose because this viewing angles are too small to consider the characteristics of the human visual field. Accordingly, we built our own dataset of human eye movements with 57 degrees of horizontal viewing angle, which is significantly wider than that of the existing datasets. All of the images and the gaze recording data will be made available online for further studies[1].

Figure 2 shows some sample images from our dataset. We collected 400 random images from Flickr Creative Commons. Gaze position data for these images were acquired binocularly at 60 Hz using a Tobii TX300 eye tracker [2011]. Fifteen test subjects participated in the gaze data collection. They were asked to sit approximately 133 cm away from a 65-inch (143.5 × 80.2 cm) display in a dark room and look at images displayed on the full screen with a resolution of 1366 × 768, and a chin rest was used to stabilize the subjects' heads. Each image was shown for four seconds in a randomized order, and a white cross mark on a black background was shown for two seconds at each interval.

To calibrate the position of the first saccade, test subjects were instructed to look at the cross mark at first and then at anywhere in an image freely. As a dummy task to motivate the subjects, they were also asked to evaluate with three levels how interesting the previous image was by pressing a numerical keypad.

Fixations were obtained from gaze data based on their velocity. When the velocity becomes higher than a predefined threshhold (22 degrees/sec in our current setting), it is considered a saccadic movement. Gaze data is divided into two fixations according to before and after each saccade, and the coordinates of each fixation point are computed as the average between the coordinates of each divided data. On average, five saccades were extracted from one subject on one image.

2.2 Saliency Map Model

We employed graph-based visual saliency (GBVS) model [Harel et al. 2007] as the baseline model of a bottom-up saliency map. In this model, the input image is decomposed into several types of visual feature images using simple linear filters as in the basic saliency model [Itti et al. 2003]. Three basic features—intensity, color, and orientation—are employed in GBVS. They are first extracted at multiple scales (three scales in our study, $1/4, 1/8, 1/16$) in a Gaussian pyramid from the original image. In GBVS, feature maps are computed as equilibrium distributions in a Markov chain. Pixels in the feature images are treated as nodes of a graph, and transition probabilities between nodes are defined based on their distance and dissimilarity. The equilibrium distribution in this way represents uniqueness and saliency in each image location. Feature maps are computed from each of 3×3 feature images.

Following Cerf *et al.* [2008], we used a feature map based on face detection. Facial locations are estimated using a face detector from Face.com [2011], and facial feature maps are computed as a Gaussian distribution with respect to the center of the detected face. We defined the variance σ_f of the Gaussian distribution as the size of each detected face: $\sigma_f = \sqrt{(w/2)^2 + (h/2)^2}$, where w and h are the width and height of the detected face. Of the 400 images in our dataset, faces were detected in 133, and only a few false positive detections were included.

These feature maps are resized to a fixed resolution (228×128), and we achieve $F = 3{\times}3{+}1$ feature maps $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_F\}$. As described above, the feature maps are then divided into C concentric regions according to the current fixation location $\boldsymbol{p}_{\text{cur}}$. A mask map $\boldsymbol{d}_c$ corresponding to the c-th region is defined as

$$\boldsymbol{d}_c(\boldsymbol{p}) = \begin{cases} 1 & \text{if } d'_{c-1} \le |\boldsymbol{p} - \boldsymbol{p}_{\text{cur}}| < d'_c, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

[1] http://www.hci.iis.u-tokyo.ac.jp/datasets/

where $\boldsymbol{d}_c(\boldsymbol{p})$ indicates the value of $\boldsymbol{d}_c$ at $\boldsymbol{p}$. In our current implementation, maps are divided into six regions with $d'_c = \{5.3, 7.3, 9.0, 11.1, 14.3, \infty\}$ so that the number of saccade targets becomes the same across regions.

Then the final saliency map $\boldsymbol{s}$ is computed as a weighted sum of $F \times C$ maps as

$$\boldsymbol{s}(\boldsymbol{w}) = \sum_c \sum_f w_{c,f}(\boldsymbol{d}_c \circ \boldsymbol{s}_f), \tag{2}$$

where $\circ$ denotes an element-wise Hadamard product, and $w_{c,f}$ indicates the weight for the c-th region of the f-th feature. Here, $\boldsymbol{w} = \{w_{c,f}\}$ is the $F \times C$ dimensional weight vector.

2.3 Learning Optimal Weights

The weight vector $\boldsymbol{w}$ is learned using the actual human saccade data. Constrained linear least squares regression is employed to optimize the weights.

Given a ground-truth saccade from a current fixation $\boldsymbol{p}_{\text{cur}}$ to the next fixation $\boldsymbol{p}_{\text{next}}$, the ideal saliency map may have a single peak around $\boldsymbol{p}_{\text{next}}$. The target map $\boldsymbol{t}$ that represents the ideal saliency map is defined as

$$\boldsymbol{t}(\boldsymbol{p}) = \begin{cases} 1 & \text{if } |\boldsymbol{p} - \boldsymbol{p}_{\text{next}}| \leq t', \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $t' = 1$ degree indicates a fixation area threshold that is defined according to the central visual field of humans.

Then $\boldsymbol{w}$ is optimized by minimizing the error between $\boldsymbol{s}$ and $\boldsymbol{t}$ with nonnegative constraints [Lawson and Hanson 1974]:

$$\boldsymbol{w} = \arg\min_{\boldsymbol{w}} \sum_i \sum_j \left\| \boldsymbol{s}_{(i,j)}(\boldsymbol{w}) - \boldsymbol{t}_{(i,j)} \right\|^2, \tag{4}$$

subject to

$$\boldsymbol{w} \geq \boldsymbol{O},$$

where (i, j) denotes the j-th ground-truth saccade data in the i-th image. For computational efficiency, Eq. (4) is solved with random sampling. Positions to evaluate the error were randomly chosen from each saccade data so that the number of chosen positions from the region with $\boldsymbol{t}(\boldsymbol{p}) = 0$ becomes 20 times larger than that from the region with $\boldsymbol{t}(\boldsymbol{p}) = 1$. If the distance to $\boldsymbol{p}_{\text{next}}$ is between 1 and 4 degrees, the position is ignored since they were considered unreliable. The total number of chosen positions were 50,000 from $\boldsymbol{t}(\boldsymbol{p}) = 1$.

3 Experimental Results

We evaluated the performance of our saliency map model by using a set of 6,000 scan-path data collected for 400 visual stimuli by 15 subjects. The scanpath data were randomly divided in half into two groups: one for training and the other for testing.

Unlike the existing models, our saliency map model takes into account the non-uniformity existing in human visual fields and depends on a current fixation point. The performance of our proposed model was evaluated by assessing the prediction accuracy of the next fixation point given a current fixation point. A saliency map is computed based on our model for a given current fixation point, and the saliency value at the next fixation point is evaluated.

Our proposed model was compared with three other saliency models referred as the GBVS+Face model [Cerf et al. 2008], the GBVS model [Harel et al. 2007] and the undivided-learning model. The undivided-learning model is computed by our proposed model with $C = 1$. This is essentially the same as the learning-based saliency map model introduced by Zhao and Koch [2011]. The key difference between our proposed model and the undivided-learning model is that whether entire images are treated equally without being divided into multiple regions based on the current fixation point. The GBVS+Face model and the GBVS model were used as a baseline of performance evaluation in our experiments.

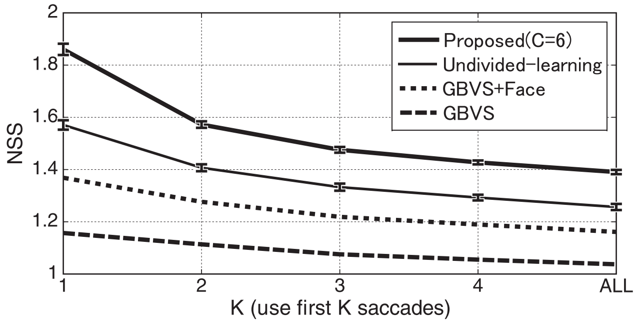

Figure 3: *NSS performance of saliency models when the first K saccades were used for training and testing. Error bars indicate standard deviations. Our proposed model archived higher performance than any of the other existing models.*

3.1 Results

Several performance metrics of saliency maps have been proposed. In this work we employed normalized scan-path salience (NSS) [Peters et al. 2005], which is known to be better suited to scan-path evaluation. To compute NSS, salience maps are linearly normalized to have zero mean and unit standard deviation. The normalized salience values are then extracted from the ground-truth fixation points, and NSS is computed as the mean of these values. Hence zero NSS value indicates no correspondence between saliency maps and fixation points, and higher NSS values indicate greater correspondence.

A comparison of avarage NSS values over 10 trials is shown in Figure 3. The bold line corresponds to the proposed learning model, the thin line corresponds to the undivided learning model, the dotted line corresponds to GBVS+Face model, and the dashed line corresponds to the GBVS model. It is clear that our model achieved higher performance than any of the existing models.

The weight values optimized in the training process are shown in Figure 4. They are computed as the averages over 10 trials using the first saccade data ($K = 1$). The right-most data corresponds to the undivided-learning model, and the six others correspond to each concentric region of the proposed model. F, C, I and O indicate face, color, intensity, and orientation respectively, and the subscript indicates the scale in an image pyramid where a larger number means a lower scale.

On the whole, orientation and face were the two most dominant features, and optimized weights were relatively larger for these two features as the distance from the fovea becomes larger. This is consistent with the results by Zhao and Koch [2011]. The weights of the orientation map for higher scales became smaller in areas away from the current fixation point. This means that the orientation features of small details becomes less important in peripheral vision, and agrees with the known characteristics of human visual fields. While the weights for color and intensity maps become smaller

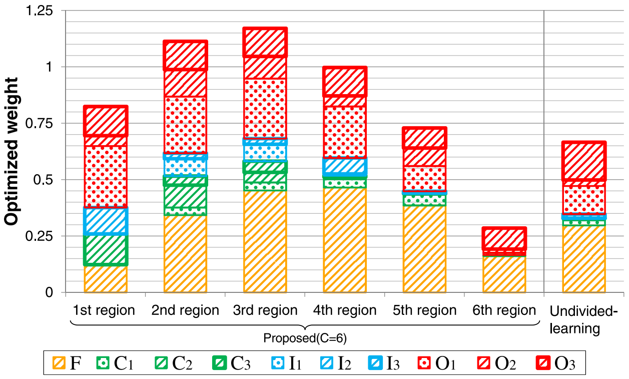

Figure 4: *Averages of optimized weights through 10 trials using the first saccade data ($K = 1$). F, C, I and O indicate face, color, intensity, and orientation respectively, and the subscript indicates the scale in an image pyramid where a larger number means a lower scale.*

with larger distances, no strong dependence on image scales could be observed on them. This is mainly because these features themselves do not have a strong dependence on image scales, and there are no major differences between feature maps on different scales. Large weights were assigned to the facial feature map regardless of the distance, except the first and sixth region. A possible reason for this is that faces are hardly recognized in the peripheral (sixth) region and facial regions often have been already focused in the foveal (first) region.

4 Conclusion

We proposed a novel model for computing a saliency map that incorporates human visual field characteristics. Unlike existing models that use a constant weight in integrating image features regardless of their spatial positions, our model uses different weights depending on the distance from the current fixation point. The weights are learned using actual saccade data. The experimental results indicated that the gaze point estimation performance improved by using our model.

Our introduced weights are designed to change in distance from the current fixation point. However, our visual field characteristics differ from not only the distance from the fovea but also its direction. For example, the characteristics are not the same between the horizontal and vertical directions. Therefore, weights that are more closely adapted to real characteristics will be needed.

Acknowledgments

This research was supported by CREST, JST.

References

BOYNTON, R. M., SCHAFER, W., AND NEUN, M. E. 1964. Hue-wavelength relation measured by color-naming method for three retinal locations. *Science 146*, 3644, 666–668.

BRUCE, N., AND TSOTSOS, J. 2006. Saliency based on information maximization. In *Proc. NIPS 18*, 155–162.

BRUCE, N., AND TSOTSOS, J. 2009. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision 9*, 3, 1–24.

CERF, M., HAREL, J., EINHÄUSER, W., AND KOCH, C. 2008. Predicting human gaze using low-level saliency combined with face detection. In *Proc. NIPS 20*, 241–248.

CERF, M., FRADY, E. P., AND KOCH, C. 2009. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision 9*, 12, 1–15.

CURCIO, C. A., SLOAN, K. R., KALINA, R. E., AND HENDRICKSON, A. E. 1990. Human photoreceptor topography. *Journal of Comparative Neurology 292*, 4, 487–523.

FACE.COM, 2011. Face.com face recognition api. http://face.com/. [Online; accessed 11-November-2011].

GROLL, S. L., AND HIRSCH, J. 1987. Two-dot vernier discrimination within 2.0 degrees of the foveal center. *Journal of the Optical Society of America A 4*, 8, 1535–1542.

HAREL, J., KOCH, C., AND PERONA, P. 2007. Graph-based visual saliency. In *Proc. NIPS 19*, 545–552.

HENSON, D. 1993. *Visual Fields*. Oxford University Press, 2–3.

ITTI, L., KOCH, C., AND NIEBUR, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transaction on Pattern Analysis and Machine Intelligence 20*, 11, 1254–1259.

ITTI, L., DHAVALE, N., AND PIGHIN, F. 2003. Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Proc. SPIE 48th Annual International Symposium on Optical Science and Technology*, 64–78.

JUDD, T., EHINGER, K., DURAND, F., AND TORRALBA, A. 2009. Learning to predict where humans look. In *Proc. ICCV 2009*, 2106–2113.

KIENZLE, W., WICHMANN, F. A., SCHÖLKOPF, B., AND FRANZ, M. O. 2007. A nonparametric approach to bottom-up visual saliency. In *Proc. NIPS 19*, 689–696.

LAWSON, C. L., AND HANSON, R. J. 1974. Solving least square problem. In *Prentice-Hall*. ch. 23, 161.

PARKHURST, D., LAW, K., AND NIEBUR, E. 2002. Modeling the role of salience in the allocation of overt visual attention. *Vision Research 42*, 1, 107–123.

PETERS, R. J., IYER, A., ITTI, L., AND KOCH, C. 2005. Components of bottom-up gaze allocation in natural images. *Vision Research 45*, 18, 2397–2416.

RAMANATHAN, S., KATTI, H., SEBE, N., KANKANHALLI, M., AND CHUA, T.-S. 2010. An eye fixation database for saliency detection in images. In *Proc. ECCV 2010: Part IV*, 30–43.

TOBII TECHNOLOGY INC., 2011. Tobii TX300 eye tracker. http://www.tobii.com/en/eye-tracking-research/global/products/hardware/tobii-tx300-eye-tracker/.

VINCENT, B. T., TROSCIANKO, T., AND GILCHRIST, I. D. 2007. Invistigating a space-variant weighted salience account of visual selection. *Vision Research 47*, 13, 1809–1820.

VIRSU, V., AND ROVAMO, J. 1979. Visual resolution, contrast sensitivity, and the cortical magnification factor. *Experimental Brain Research 37*, 3, 475–494.

ZHAO, Q., AND KOCH, C. 2011. Learning a saliency map using fixated locations in natural scenes. *Journal of Vision 11*, 3, 1–15.