Visual Attention with Contextual Saliencies of a Scene

Sukhan Lee
College of Information and Communication Engineering
Department of Interaction Science
Sungkyunkwan University
Gyeonggi-do, South Korea
82-31-299-6471
lsh@ece.skku.ac.kr

HyunKook Ahn
College of Information and Communication Engineering
Sungkyunkwan University
Gyeonggi-do, South Korea
82-31-299-6572
tomc4you@naver.com

JooYun Han
Department of Interaction Science
Sungkyunkwan University
Seoul, South Korea
82-31-299-6572
hanjooyun@gmail.com

Yu-Bu Lee
College of Information and Communication Engineering
Sungkyunkwan University
Gyeonggi-do, South Korea
82-31-299-6487
basilia@skku.edu

Abstract

This paper presents an examination of the possible competition and cooperation that may take place in human visual attention, between the bottom-up saliencies incurred by photometric signatures and the top-down saliencies incurred by the primary context of a scene. It is found that the strength of the primary context of a scene represents a dominant guiding factor for determining the visual fixations for attention: in the case where there exists a strong context in a scene, the objects and/or regions that are tightly coupled with the context dominate for defining the saliencies that guide fixations for attention. It appears that, in human visual perception, a higher priority is assigned to the efficient understanding of a visual context than the direct response to photometric saliencies not supported by the context. The claims described above are derived from the experimental verification of the following conjectures: 1) There is a tendency for the bottom-up saliencies to be more significant when the context of the scenes observed is either weak or nonexistent. 2) For the scene of a strong context, the top-down context saliencies such as the objects and regions that are associated with understanding the present context tend to dominate over the bottom-up saliencies. 3) When the scene of a strong context includes both positive and negative saliencies, where the positive/negative contextual saliencies are referred to here as those saliencies significant for understanding the context yet well-expected/unexpected for the given context in terms of the prior knowledge, the negative saliencies are assigned a higher priority than the positive saliencies for attention.

Categories and Subject Descriptors

G.4 [**Data Analysis**]: Fixation data analysis

General Terms

Experimentation

ICUIMC(IMCOM)'13, January 17–19, 2013, Kota Kinabalu, Malaysia.

Keywords

Eye movement, Context saliency, Bottom up and top down process, fixations

1. Introduction

The human eye is one of the most important sense organs. In fact, approximately 90% of the information that humans take from their surroundings is obtained through the eyes (Drury & Clement 1978). Since humans need to move their pupils constantly in order to recognize specific objects, eye and pupil movements can be tracked to ascertain the target of an individual's gaze. Research of this kind has identified two types of eye movement fixations and saccades (the combination of these movement types is known as a scan-path). A fixation is when the eyes remains fixed at a specific position, and a scan-path is the instantaneous movement that occurs during a fixation. One of the psychological mechanisms that govern sight stimuli and eye movement is attention (Parkhurst et al., 2002). Two theories have been proposed regarding how this attention is generated. According to the bottom-up attentional selection theory, since sight pays attention to salient stimuli, attributes such as color, luminance, and direction determine eye movement (Peters et al., 2005; Lee et al., 1999; Itti & Koch, 2001; Parkhurst et al., 2002; Underwood et al., 2006). In contrast, the top-down attentional selection theory holds that an individual's pattern of fixation and scan-path will vary depending on the knowledge system of the individual, such as their personal experiences or memories. The sense of sight is not comprised simply of eye movements; rather, it is a selective and active sense in which eye movement acts as a control process that connects sight with perceptual, cognitive, and behavioral activities in real time (Henderson & Hollongworth, 1998; Henderson, 2003; Yantis & Egeth, 1999).

Previous studies have indicated that human eye fixations and scan-paths will occur differently depending on the context information represented in the image. For example, when seeing an image of road environment, an individual who has a knowledge system that is familiar with a road environment unconsciously generates a projected image based on this prior knowledge, and this unconscious projection overlaps with the actual image. Once this process has occurred, the individual

focuses on the most interesting part of the actual image. At this point, while the individual remains aware of the overall context-specific information, such as other vehicles and the road, context saliency is enacted so that the individual focuses on only one truck among numerous vehicles (of a similar kind) on the road. For another example, an image which has only one black person with numerous white persons can be also context saliency. That is, the context dependency in eye fixation and scan-path is first to grasp the scene context as a whole by categorization, followed by the understanding of the details of contextual significances through a series of fixations on context saliencies. These observations indicate that the movement of the human eye is not simply directed only by bottom-up saliencies; rather, it seems that the fixation and scan-path occur in different ways in such a way to understand context information effectively. In terms of the efficacy of understanding the context, the initial categorization of the context triggers the top-down feedback of interesting objects and/or regions to be verified, referred to here as positive context saliencies, which will dominate the process of visual fixations for attention. However, under the existence of contextually salient objects and/or regions, seen as peculiar relative to the primary context of the scene that disturb the expectations from the categorization, referred to here as negative context saliencies, the visual attention puts an even higher priority to such negative contextual saliencies. To further investigate and verify our observations as described above, we formally propose the following conjectures in order to verify their validity based on extensive experimentations: 1) There is a tendency for the bottom-up saliencies to be more significant when the context of the scenes observed is either weak or nonexistent. 2) For the scene of a strong context, the top-down context saliencies such as the objects and regions that are associated with understanding the present context tend to dominate over the bottom-up saliencies. 3) When the scene of a strong context includes both positive and negative saliencies, where the positive/negative contextual saliencies are referred to here as those saliencies significant for understanding the context yet well-expected/unexpected for the given context in terms of the prior knowledge, the negative saliencies are assigned a higher priority than the positive saliencies for attention.

2. Stimuli and Experimental Design

The eye tracker model used in this study was ViewPoint GigE-60 from Arrington Research Inc. Monocular for mobile was also used. The six participants in this study were all between the ages of 20 and 40. A 17-inch monitor with a 1024×768 resolution was used to present the images. A distance of was 0.827m was kept between the monitor and the person. Each participant's gaze was tracked using the eye tracker while they remained stationary; a chin fixing plate helped keep their head in place. First, the process of a 9-point calibration was performed using the ViewPoint program. 9 images were presented in random order to subjects for 3s each. Each image was followed by a screen with a white cross in the middle of a black background blank screen for 2s so that the initial eye movements occurring in response to an image were not biased by the previous image.

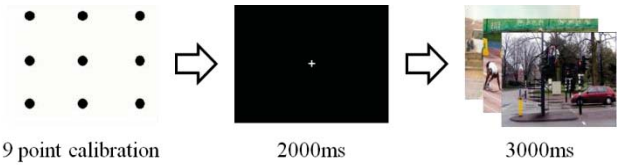


Figure 1. Procedure of the eye tracking experiment.

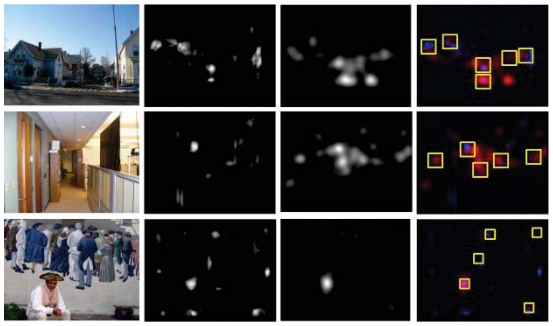
This process was repeated until nine images were presented. We used the experimental images with three different categories as shown in Figure 1. Two experiment images from the ill-contextual group were randomly set to lower bottom-up saliency by darkening the brightness of the original images.

Figure 2. Example of images used in the eye tracking experiment: (a) images with weak or nonexistent context correlation, (b) images with "positive context", and (c) images with "negative context".

3. Analysis

To analyze three different categories, the following experiment methods were used in relation to a main image focusing on an office presentation scenario. (1) For analyzing relation between context and human attention on image where context was weak or nonexistent, we compare the correlation between human fixation map and computational saliency maps: Itti's [6], Hou's saliency maps [12]. We also count the number of fixations included in the salient region as shown in Figure 3. Itti et al. [6] created a computational model of contrast-based spatial attention that is derived from a biologically plausible architecture. They compute saliency maps for luminance, color, and orientation on different scales that aggregate and combine information about each location in an image, which is then fed into a combined saliency map. Hou and Zhang have explored the role of spectral components in an image to detect saliency. The gist of a scene is represented by an averaged Fourier envelope and the differential spectral components are used to extract salient regions. However, humans pay more attention to areas corresponding to meaningful objects, such as faces and text, even when those objects are not visually obtrusive. (2) For "positive context" images which have specific context such as meeting scene, we generated a human fixation map for each of the six participants and compared human fixation map with computational saliency maps. The subjects intentionally

or unintentionally see certain area when they faced the images. In order to figure out on which objects are focused, we analyzed the fixation points whether it stayed on salient area or not and the targeted objects which have a contextual role in the images. (3) For "negative context" images, situations or regions that were not related to the original experiment images were set in the experiment images as an ill-contextual region.



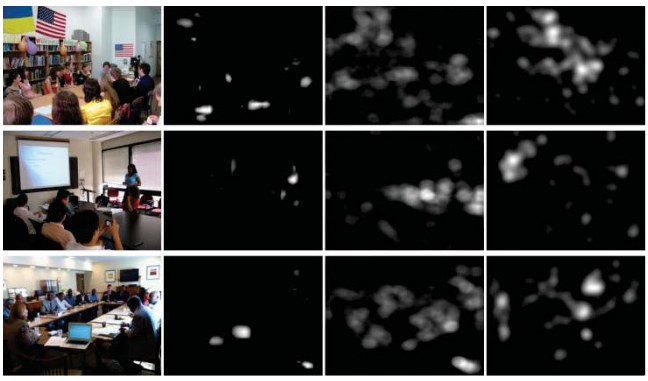
(a) (b) (c) (d)

Figure 3. Analysis of fixations included in the salient region: (a) original image, (b) Itti's saliency map, (c) human fixation map, and (d) correlation map represented fixations included in the salient region (yellow boxes).

In addition, we compute the proportion of fixations that focused on the ill-contextual region was calculated.

4. Experimental Results

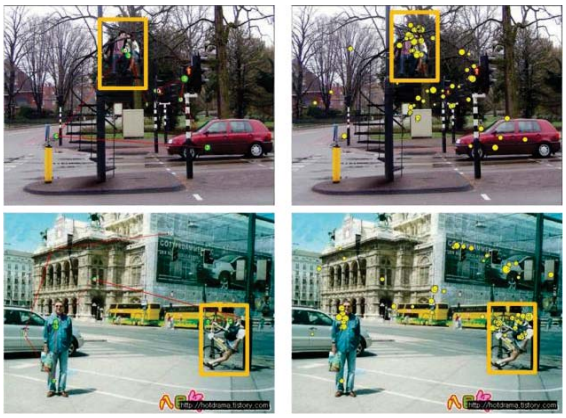
For the images which have weak or non contextual objects to the original images, the analysis revealed in Figure 7 that the percentage of fixations on the salient regions were as follows: subject 1: 53%, subject 2: 53%, subject 3: 50%, subject 4: 50%, subject 5: 51%, and subject 6: 48% this result shows that the eye movements were mostly guided by bottom-up saliency first in weak or non correlated context. In the experiment of "positive context" images (refer to Figure 4) regarding the priority of 'bottom-up' and 'top-down' processes, first five fixations located in salient area were counted and this rate is indicated in Figure 8. During the free-viewing task, 9.23% of subjects fixated their focus on salient area for the first fixation. It was 9.23% to locate in salient part for the second fixation as well. For the third and fourth fixation rates were 14.87% and 13.85%. On the fifth fixation, it rates 13.85%.



(a) (b) (c) (d)

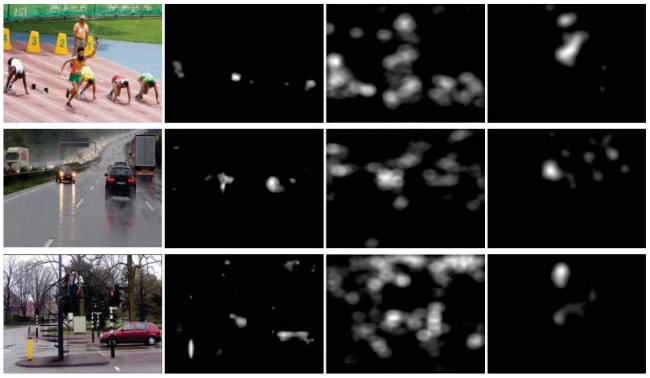
Figure 4. Comparison of the correlation between human fixation map and computational saliency maps from "positive context" experiment: (a) original image, (b) Itti's saliency map, (c) Hou's saliency map, and (d) human fixation map generated by convolving a Gausssian over the fixation locations obtained from eye tracking data for all subjects.

Figure 5 show the analysis of the results of the "negative context" experiment. Figure 5(a) and (b) show the scan path and the distribution of human fixations for 3s of viewing time, respectively. Figure 5(b) shows that the fixations of all the participants tended to cluster around the ill-contextual region relatively frequently. Table 1 illustrates how many fixations the participants experienced before they directed their gaze toward the ill-contextual region of each experiment image. The analysis revealed that participants tended to direct their gaze toward the ill-contextual region after one to four fixations. The X marks in Table 1 indicate errors that occurred during the experiment; these results were excluded from the experiment result. These errors came about when no accurate scan-path could be obtained—either because the participant's head moved after calibration, or because of a program error. In order to represent the experiment results in more detail, we counted the percentage of fixations between the first and the fifth fixations focused on the ill-contextual region, which is indicated by orange squares as shown in Figure 5.



(a) (b)

Figure 5. Analysis from the "negative context" experiment: (a) scan path (b) the distribution of human fixations for 3s of viewing time.



(a) (b) (c) (d)

Figure 6. Comparison of the correlation between human fixation map and computational saliency maps from "negative context" experiment: (a) original image, (b) Itti's saliency map, (c) Hou's saliency map, and (d) human fixation map generated by convolving a Gausssian over the fixation locations obtained from eye tracking data for all subjects.

Table 1. Number of ill-contextual regions found (No. of Fixations found/No. of total fixations)

Subject	Img1	Img2	Img3	Img4	Img5	Img6	Img7	Img8	Img9
1	1/7	1/10	2/3	1/9	1/8	1/7	2/4	1/7	1/7
2	1/9	1/9	1/10	1/10	1/9	X	1/10	X	1/6
3	1/10	1/10	2/10	2/11	1/5	2/10	2/10	1/10	4/10
4	1/10	4/10	3/11	1/12	2/12	1/14	1/10	1/11	1/9
5	8/9	1/9	1/10	2/11	X	X	8/10	10/12	1/7
6	1/10	1/11	2/12	1/11	1/10	1/13	1/10	1/11	1/9

Figure 9 shows the objects on which the subjects fixated their focus and its rate. Each image had different context and various objects, but there were common targets attract the human fixation in the context of meeting room scene in general. First, the subjects mostly focused on people and its face, about 26%. The table was secondly focused for 22% in the meeting room scenes; there were some objects on the table such as cups, notes, papers etc. Wall rates about 15% that it was not specific objects, but some images have decorations or frames on the wall or subjects fixated on it by the process of scanning. Most presenters were standing in the scene; subjects naturally fixated on speaker for 10%. And then, followed by bookshelf/closet for 6%, equipment for 5%, decoration and ceiling.

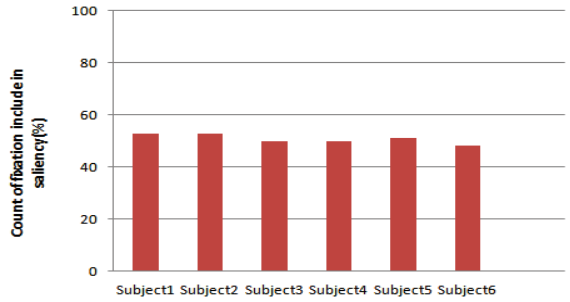


Figure 7. The number of fixations included in the salient region.

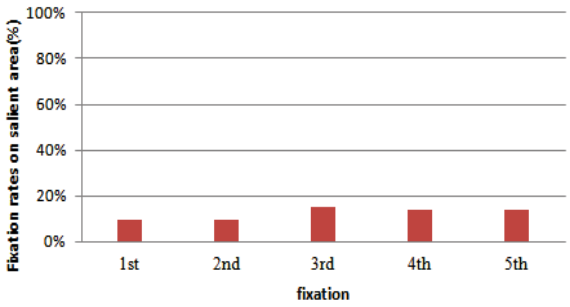


Figure 8. Percentage of fixation included in the salient area between the first and the fifth fixations from "positive context" experiment.

Figure 10 shows that the percentage of first fixations that focused on the ill-contextual region was 68.5%, this fell to 16.6% for the second fixation, 0% for the third fixation, 1.8% for the fourth fixation, and 0% for the fifth fixation. These results were consistent with those shown in Table 1, which indicate that the participants' gaze tended to focus on the ill-contextual region by the time of the fourth fixation. Moreover, the high values of the first and second fixations indicate that the participants recognized the ill-contextual region quickly.

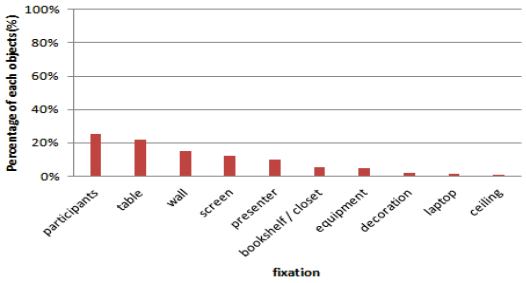


Figure 9. Percentage fixated on each object in "positive context" images.

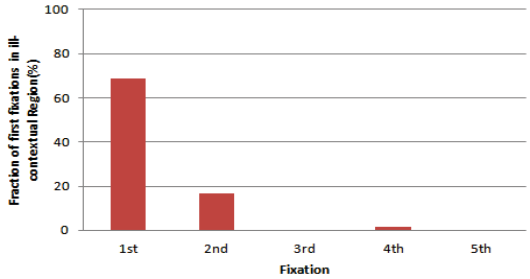


Figure 10. Percentage of fixations included in the ill-contextual region between the first and the fifth fixations.

From Itti's and Hou's saliency map, it is clear that an individual's fixation shifts focus according to changes in color, intensity, and orientation. However, the results suggest that, with regard to the fixations of the participants, bottom-up directed gazes were infrequent. To analyze this, we compared an human fixation map with two computational saliency maps. To compare the human fixation map more objectively with the computational saliency maps, we measured the correlation coefficients between them. Table 2 shows the correlation coefficient between the Itti's saliency map and the human fixation map. We see that the participants gazed at the detailed objects in order to understand context saliency. That is, the participants did not concentrate on bottom-up saliency; rather, their fixations and scan-path were directed to the ill-contextual region, primarily in order to understand context information. The coefficient criteria in the correlation map was determined that as more overlapping is found between the human fixation map and Itti's saliency map, the value become closer to one, and as the overlapping is less, the value is

closer to zero [3].

Table 2. Correlation coefficient between the Itti's saliency map and the human fixation map

Image	1	2	3	4	5	6	7	8	9	Mean
Correlation Coefficient	0.108	0.324	0.1	0	0.141	0	0.025	0	0.19	0.098

5. Conclusion

This may be linked to the nature of human visual processing, in which both photometric and contextual perception processes take place in parallel, meaning that the contextual feedback pathways in visual processing that allows initial categorization of the context may experience only a short time delay from the bottom-up detection of photometric saliencies.

This study evaluated whether the fixation and scan path of the human eye occurs differently depending on the context information shown in the images upon which the eye is gazing. To evaluate this, three experiment methods were used. First, for images where the relationship of the context images with the main image was weak or nonexistent, the fixation of the eye was found to be guided by bottom-up saliency, and focused on the salient part of the image. In contrast, in the case where the relationship between the image and the context was "positive" and where it was "negative," fixation was guided primarily by the context that related to the image, rather than by bottom-up saliency. On the basis of these results, the following conclusions can be made. In terms of understanding the context efficiently, it appears that the objects and/or regions that are seen as particularly relevant to the primary context of the scene (referred to here as context saliencies) are given higher priority for attention. In this case, the initial categorization of the context triggers the top-down feedback of interesting objects and/or regions to be verified, referred to here as positive context saliencies, which strongly influence the process of visual fixations for attention. However, under the existence of contextually salient objects and/or regions that disturb the expectations toward the categorized context (referred to here as negative context saliencies), the visual attention tends to prioritize the understanding of such contextual saliencies and, thus, quickly move its focus toward the contextual saliencies. Also it appears that, in visual perception, a higher priority is assigned to the efficient understanding of the visual context than is assigned to the direct photometric saliencies that are not supported by the context. This may be linked to the nature of human visual processing, in which both photometric and contextual perception processes take place in parallel, meaning that the contextual feedback pathways in visual processing that allows initial categorization of the context may experience only a short time delay from the bottom-up detection of photometric saliencies.

6. ACKNOWLEDGMENTS

This research was supported by WCU(World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology(R31-10062), by the KORUS-Tech program (kt-2010-SW-AP-FS0-0004), by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (2012-008188),by PRCP through NRF of Korea funded by MEST (2012-0005861), and by MKE, Korea under ITRC NIPA-2012-(H0301-12-3001).

7. REFERENCES

[1] Cerf, M., Harel, J., Huth, A., Einhauser, W., and Koch, C. (2008). Decodingg what people see form where they look: Predicting visual stimuli from scanpaths. In L. Paletta & J.K. Tsotso(Eds.), Lecture Notes in Artificial Intelligence, LNAI 539. Heidelberg: Springer-Verlag Berlin.

[2] Dickinson, S., Christensen, H., Tsotsos, J., and Olofsson, G. (1994). Active object recognition integrating attention and viewpoint control. Proceedings of the Third European Conference on Computer Vision, Stockholm, Sweden.

[3] Dooseok Kang, Sukhan Lee, and Yu-Bu Lee (2011). Human visual attention with context-specific top-down saliency. Proceedings of IEEE International Conference on Robotics and Biomimetics, 2055 – 2060.

[4] Einhäuser, W., Kruse, W., Hoffmann, K., and König, P. (2006). Differences of monkey and human overt attention under natural conditions. Vision Research, 46, 1194–1209.

[5] Foulsham, T. and Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. Journal of Vision, 8(2):6, 1–17.

[6] Hou, X. and Zhang L. (2007). Saliency detection: A spectral residual approach. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1–8.

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

[8] Itti, L., and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research, 40, 1489–1506.

[9] Mack, A., Pappas, Z., Silverman, M., and Gay, R. (2002). What we see: Inattention and the capture of attention by meaning. Consciousness and Cognition, 11, 488–506.

[10] N. Ouerhani, R. von Wartburg, H. Hügli, and R. M. Müri. (2004). Empirical validation of the saliency-based model of visual attention, Electronic Letters on Computer Vision and Image Analysis, 3(1), 13–24.

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

[12] Yu-Bu Lee and Sukhan Lee (2011). Robust face detection based on knowledge-directed specification of bottom-up saliency. ETRI Journal, 33(4) 600-610.