Analysis of Human Attention toward Context-related Images for Understanding Visual Information Processing

Joo Yun Han
Department of Interaction Science
Sungkyunkwan University
Seoul, Korea
hanjooyun@gmail.com

Yu-Bu Lee
College of Information & Communication Engineering
Sungkyunkwan University
Gyeonggi-do, Korea
basilia@skku.edu

Sukhan Lee
Department of Interaction Science,
College of Inforamtion & Communication Enginnering
Sungkyunkwan University
Seoul/Gyeonggi-do, Korea
lsh1@skku.edu

***Abstract*—The purpose of this study is to understand human's visual information processing mechanisms. We emphasize the importance of "context" for attracting visual attention as an effect of top-down process. We created indices for saliency index by degree of saliency using Itti's saliency model and for context index by object importance from pre-experimental survey for analyzing the present data. In terms of bottom-up process, the effect of bottom-up saliency was insignificant and correlation between the saliency map and human fixation map was rather low. We compared the context index within both salient and non-salient area that top-down processing is a strong attentional guidance. If the contextually meaningful objects and visually salient area was overlapped, it robustly attracts attentions. Furthermore, the counts of fixations on objects and the order in sequence of eye-movement showed that the subjects tended to fixate more often and more frequently toward more importantly considered objects.**

Keywords—context-related saliency; human attention; bottom-up saliency; top-down process; fixation

I. INTRODUCTION

Like mise-en-scène, used in film studies to make meaning by the visual arrangement of elements in a scene or frame [1], objects can explain the context of an environment or scene. We observe several pieces of visual information and understand the context around many objects. At the same time, we unconsciously and simultaneously repeat the process of seeing and understanding visual information through saccadic eye movements [2]. The goal of the current study is to identify what draws our attention (information we naturally focus on) in order to predict the eye-movement or attention pattern by understanding visual information processing.

Several researchers examining visual recognition have been using a variety of approaches for determining aspects of visual information processing: analysis of reading patterns [3], precise scene or object recognition [4-5], and the construction of visual saliency algorithms [6-7]. In particular, the saliency model suggested by Itti and Koch is representative of the bottom-up-based computational model; this model offers predictions on where attention is likely to be attracted. This model is effective in situations in which an image has "little semantic information" and there is no defined task for the observer [8]. In other words, this model does not consider contextual importance for top-down based visual information processing. Itti and Elazary investigated contributions of low-level saliency on subjective object choice via the "LabelMe" database. This process reveals which objects may be selected as more interesting through low-level visual features rather than cognitive, top-down processes [9]. On the other hands, several researchers have considered examining the role of semantic content on visual information processing [8]. One of research has observed strong patterns of attentional deployment for faces and text within natural scenes. This work has shown that faces and text are difficult to ignore, even if there is a conflict between the task (top-down inputs) and the facial and textual features of the stimuli (bottom-up saliency). It has been suggested that the improved "saliency-driven attentional deployment" model predicts an attended area under the receiver operating characteristic (ROC) curve more accurately, than actual eye-movement for images in which there are faces or text [10]. Recent studies have demonstrated contextual influences on human attention, such as contextual cueing guiding visual behavior – contextual guidance model – which combines bottom-up saliency and top-down mechanisms within real-world scenes, was proposed based on an emphasis for contextual information [11]. Torralba and colleagues present the "approach model," which consists of two pathways with salient and global features on the basis of a Bayesian framework [8]. Spain and Perona proposed the "urn models" for measuring and predicting the "importance of objects" based on statistical data assigned by human observers [12].

This research is derived from an effort to understand the processing mechanisms of visual attention in humans. We emphasize the importance of "context" for attracting attention as top-down process. We considered that objects in the images represented and composed the "context" and the contextual importance was examined by the objects. We also examined the relationship between context-driven eye fixations and bottom-up saliency when viewing context-related scenes. To measure saliency and contextual importance, indices were prepared using Itti's saliency model and a pre-experimental survey. Each fixation data clustered against saccadic eye

978-1-4799-0652-9/13 $31.00 © 2013 IEEE
DOI 10.1109/SMC.2013.531

movements was investigated by these standardized indices. We applied three approaches to analyze fixation data: the effect of the saliency model on early stage visual processing and the correlation between human fixation map and saliency map; contextual importance of fixations in salient and non-salient area, fixation counts and fixated order in sequence of eye-movement by importance of objects.

II. METHOD

A. Eye-tracking experiment and data acquisition

The experiment was designed to track subjects' eye-movements when viewing context-related images. Ten graduate students (6 male and 4 female) participated in the experiment. All subjects had normal or corrected to normal vision. Eye-tracking data was gathered with Arrington ViewPoint™ Eye Tracker, which uses an infrared camera to conduct corneal/pupil reflection eye tracking. All subjects wear the glasses-type eye tracker and hold their head position by using a forehead and chin rest. 17-inch screen was placed 0.97m from the participants and thirty-five images (1024 × 768 pixels) were presented for 3 seconds. The experiment was conducted in darkroom for eliminating other environmental effects such as light intensity busy brain e.g. on the experiment and its results.

The fixation data is accumulated with 0.5° accuracy. Because raw eye-tracking data were noisy and massive, it was applied dispersion-based algorithm to identify and differentiate the valuable fixation and saccadic eye-movement. This algorithm used parameters including spatial concept by 0.5-1 degree of visual angle as a threshold and temporal feature of 50ms duration time. Fig. 1 shows that by applying the dispersion threshold, the window size is extended until the dispersion exceeds the threshold [13]. The dispersion is indicated by (1) refereeing to Fig. 1[14].

$$\text{Dispersion(D)} = [\max(x) - \min(x)] + [\max(y) - \min(y)] \quad (1)$$

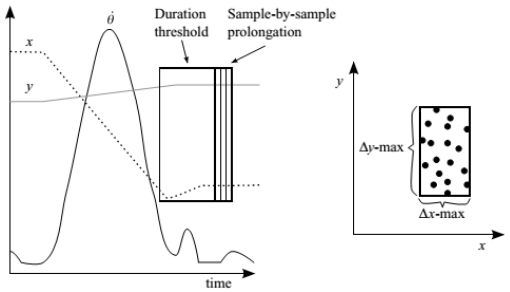

Fig. 1. The principle for caculation of duration and dispersion by dispersion algorithm. The window extended depends on the dispersion inside like as right graph [13]

Each image had a certain context that people might encounter in everyday life and familiar objects were contained in the scene: meeting room scene, road scene, and library scene and so on. Meanwhile, the composition and size of salient areas were also considered for the experiment. Eye movements were determined through a calibration procedure before beginning the experiment that subjects were instructed to move their focus around 9 points located on computer screen. After the calibration, subjects fixed their focus on a white cross at the center of a black screen. This section was used throughout the experiment for refresh the memory and remove afterimage, in between image stimuli. Subjects were instructed to see the images freely.

B. Ittis' saliency map and human fixation map

For detecting salient areas within the image, we applied Itti's saliency model. The saliency map suggests certain salient areas driven by features such as color, intensity, and orientation. These features would be expected to attract attention in a bottom-up processing. Human fixation map is accumulation of all subjects' fixation data. To create the human fixation map, we convolved the spatial distribution with a 2D Gaussian filter according to computation method by Ouerhani et al [14]. The standard deviation was 37 according to fovea size [15]. Fig. 2 shows the original image, saliency map and human fixation map.

Fig. 2. (a) original images (b) saliency map (c) all the fixation points are printed on the original image (d) human fixation map

C. Indices

We prepared standardized measures by creating a saliency and context indices. The saliency and context indices showed saliency of a certain region and how contextually important an object is, respectively. For the saliency index, an attended map constructed by the Saliency Toolbox 2.2 was used: as shown in Fig. 3, this map contained nine regions, each marked with the "attended locations" within them in an ordinal fashion. The first to the ninth attended locations became an indicator of the saliency index.

Fig. 3. Salient areas from first to eighth are marked on the original images calculated by Itti's saliency model

For the context index, seven subjects who did not participate in the experiment were surveyed. They were askedto write down five objects by each image in importance order for understanding the context. Since we consider the objects compose the "context" of image, it was based on participants' subjective evaluation; therefore, the results and the created index were reflected by top-down processing.

The context index was created by tabulating the objects mentioned by order and frequency. Fig. 4 is the example of context index and related image. Context index 5 is the most important object and index 1 is the least important object in the image.

Objects in image	Context index
Conductor	5
Audiences	4
Violinist	3
Violin, music score	2
Chair	1

Fig. 4. Example of original image and context index

The first to the tenth fixation points (the initial fixation point was eliminated because it was considered to be a reflection of the refresh screen with the white cross on the black screen) were consigned to coordinates by both the saliency and context indices.

III. RESULTS

The results reveal that (i) effect of bottom-up saliency (ii) contextual importance of fixations in salient and non-salient area (iii) fixation counts and fixated order in sequence by importance of objects.

A. *Effect of bottom-up saliency on early fixations and correlation between human fixation map and saliency map*

The first to third fixation points were examined to assess early fixation deployment. Fig.5 shows that first, second, and third fixations tended to be within non-salient area rather than salient area, unlike predicted by Itti's saliency map: this illustrates the tendencies of early-stage eye-movements. For the first fixation, 73% of fixations were within non-salient areas, and 27% were within salient areas. The rate of the fixations moved to non-salient areas for the second fixations was 69% and the other rate for non-salient area was 31%. For the third fixation, 69% were within salient area and 31% were within non-salient area as well.

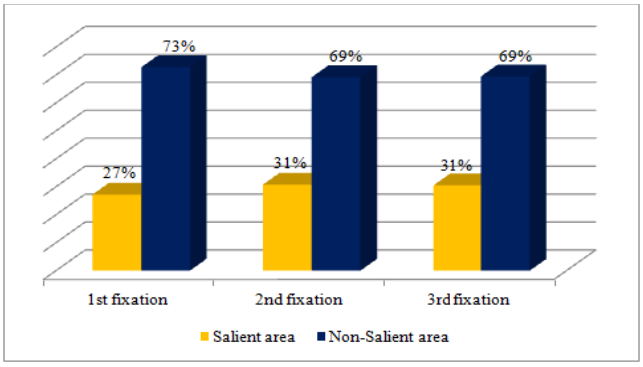

Fig. 5. The comparison of the rate by the location of fixation point whether it is stayed in salient area or non-salient area

The effect of visual saliency was not strong during early visual processing; thus, we compared the human fixation map with the saliency map for investigating all fixation data not only early stage eye-movement. The correlation coefficient of the two maps is computed according to (2). $M_h(x)$ represents human fixation map, and $M_c(x)$ means saliency map. μ_h and μ_c represents the mean value of the two maps, $M_h(x)$ and $M_c(x)$ respectively.

$$Corr = \frac{\sum_x [M_h(x)-\mu_h)\cdot(M_c(x)-\mu_c)]}{\sqrt{\sum_x (M_h(x)-\mu_h)^2 \cdot \sum_x (M_c(x)-\mu_c)^2}} \quad (2)$$

The correlation was rather small, suggesting no significant relationship. The correlation between the human fixation map and saliency maps was 0.14 (SD = .141). As shown in Fig. 6, there are few overlapped areas between the saliency map (colored in blue) and human fixation map (marked in red). The correlation of the first row is 0.154, of the second row is -0.057, and the third row's correlation is 0.179.

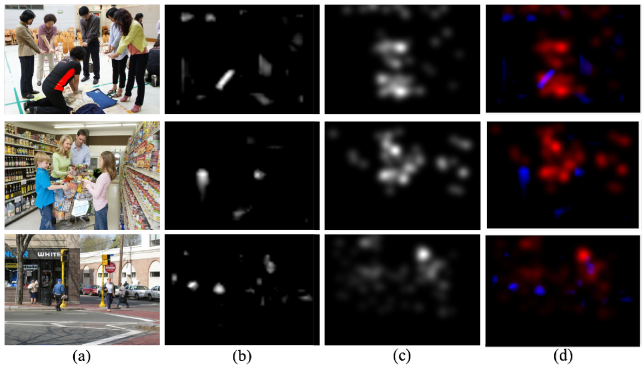

Fig. 6. Example of comparison of saliency map and human fixation map (a) original image (b) Itti's saliency map (c) human fixation map (d) corrleation map between saliency map and fixation map.

B. *Comparing context index within both salient and non-salient area*

For indicating the influence of top-down processing, weanalyze the contextindex from the first to third fixations within salient and non-salient areasseparately. Within the salient area, average context index for the first fixation was 3.53, for the second fixation was 3.40 and for the third fixation was 3.29 as shown in Fig. 7. Even though the fixation points stayed in salient area, the context index was comparatively high.

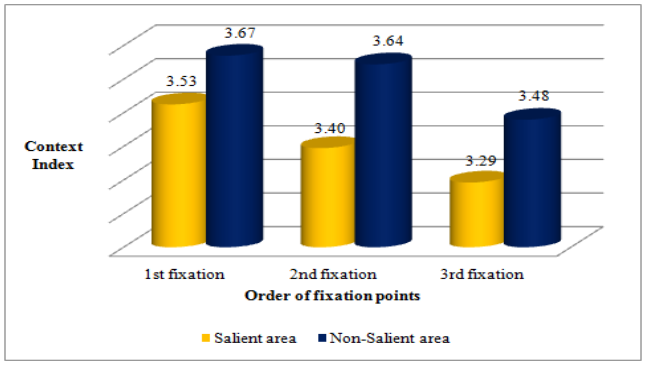

Fig. 7. Average of context index from the first to third fixations

Fig. 8. Examples of that eye-movement fixated within non-salient and contextually important objects (a) original image (b) Itti's saliency map and (c) scan path data printed on the original image

It suggests that eye-movements were derived from not only bottom-up saliency but also top-down process simultaneously. To investigate the actual scan path data on original image like Fig. 9, in the case of both the contextually important objects and visually salient regions are overlapped, the early fixation focused within that area – the example image shows early stage of eye-movements fixated on the face and monitor which are visually salient and concurrently, important in context of the image. There are about 10 images which the contextually important objects and bottom-up-based salient area are overlapped among 35 visual stimuli; for early fixations, approximately 32.22% of present data within 10 images show both saliency index and context index were coincidentally high in comparison to other images contained only 4.33% of the tendency. It indicates if contextually meaningful objects are also visually salient, they would robustly attract attention.

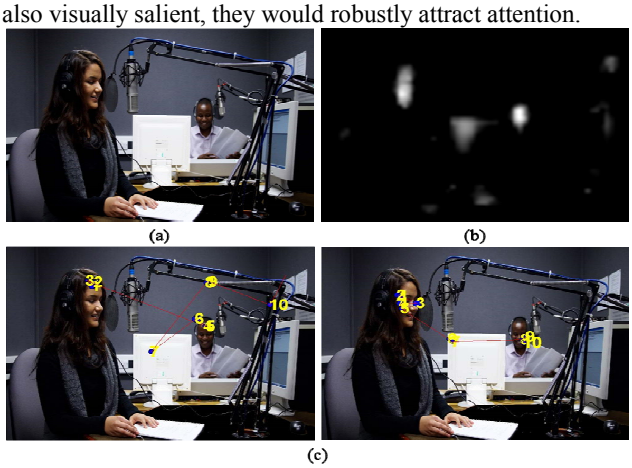

Fig. 9. Example of experimental stimuli contextual importance and salient area overlapped (a) original image (b) Itti's saliency map and (c) scan path on the original image

The context index within non-salient area was high: Within the non-salient area, average context index was 3.67 for the first fixation, 3.64 for the second fixation and 3.48 for the third fixation (Fig. 7). It demonstrates the eye-movement within non-salient area moved toward contextually important objects under the effect of top-down process. About 70% of fixation data tended to be stayed within non-salient area where the contextually important object was located and the example of the subjects' scan path is shown in Fig. 8.

C. *Analysis of the fixation data by fixation counts and order in sequence on contextually important objects*

Through analysis of the fixation points, we changed the direction of our analysis so that the fixation counts and order in sequence of focused objects showed a new possibility for deployed eye-movements.

We found that subjects fixated on the object considered contextually most important within the image the most, and the second most important object secondly most. We observed a significant tendency whereby the objects were most frequently fixated toward based on the order of importance. Fig. 10 shows the average number of fixations corresponding to objects accumulated from the first to the tenth fixations. The graph shows the average fixation counts and increasing fixation tendency by objects' importance. Subjects fixated on the least important object, index 1, an average of 6.4 times while the most important object, index 5 was fixated an average of 24.8 times. They then fixated on the second most important object, index 4, an average of 20 times. By the importance of objects, subjects fixated in increasing order 6.4, 9.1, 13.6, 20 and 24.8 times, respectively. In other words, the most important objects in terms of context were fixated more often.

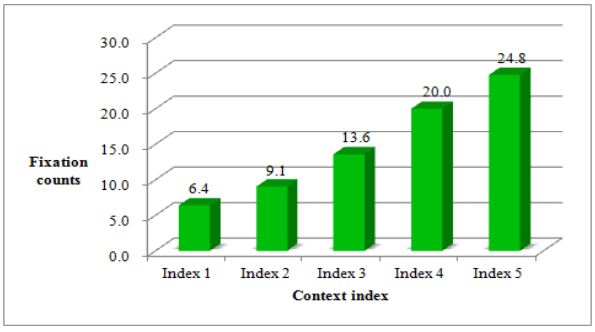

Fig. 10. Average number of fixations on the object corresponding to indices

The fixated order by objects in eye-movement sequence from the first to the tenth was checked for analysis. As can be seen in Fig. 11, the more important object was for the context of an image, the earlier subjects fixated on that object. The average order being fixated of eye-movement was within 2.09 for the most important object (index 5), within 3.20 for the second most important object, within 3.43 for index 3, within 3.83 for index 2 and within 4.46 in order for index 1: less important objects were fixated later in order.

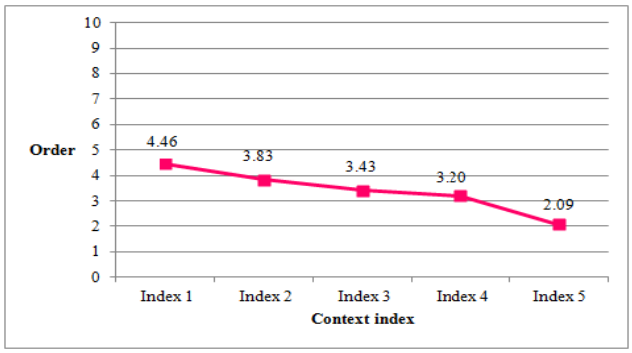

Fig. 11. Average fixation order of object corresponding to indices

The results show that there is a certain attentional tendency understood by examining contexts. First, the saliency model was found to be a poor predictor of attention for early visual processing and the correlation between the saliency map and human fixation map was also rather low. Second, within both salient and non-salient areas, the context index was relatively high. It indicates two implications that (i) what the context index was high within salient area means there was interrelation between top-down processing and bottom-up saliency affected on eye-movement and, that is, it guides fixations more powerfully (ii) context index was also high for non-salient area and it explains the fixations moved toward contextually important objects by top-down process. Third, the most important objects in terms of context were fixated on earlier and more often. In other words, the importance of the objects strongly influenced visual attention, suggesting that a top-down process influenced eye-movements to a greater degree in the context-related images.

IV. Discussion

We examined how eye-movements are attracted to certain areas within an image containing contextual information. For context-related images, contextual importance facilitated attention more than visual saliency and, that is, top-down processing is a strong attentional guide. This result could occur for a variety of reasons. For instance, the scenes depicted everyday life situations providing individuals with something familiar and this intimacy can drive the attention to contextually important objects first. Although our statistical data show certain commonalities in visual attention, we admit there are likely individual differences in how people attend: one subject mostly fixated on people during the experiment, for example. Overall, we examined the visual information processing by analyzing human attention. We observed that the "context" is an important and strong guide for visual attention. We would expect that these results could be applied for image segmentation, human and machine interactions and the general field of cognitive psychology. In future studies, we will demonstrate more clearly the relation between "context" and visual attention; we will continue to explore human visual processing systems.

Acknowledgment

This research was supported by WCU(World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education (R31-10062), by the KORUS-Tech program (kt-2010-SW-AP-FS0-0004) sponsored by the Korea Ministry of Science, ICT and Planning(MSIP), the National Research Foundation of Korea(NRF) Grant funded by the Ministry of Education (2012R1A1A3008188), by Basic Science Research Program through NRF of Korea, funded by MOE(NRF-2010-0020210), and by MSIP, Korea under ITRC NIPA-2013-(H0301-13-3001).

References

[1] Gibbs, John. Mise-en-Scčne: Film Style and Interpretation. vol. 10. Wallflower Press, 2002.

[2] Corbetta, Maurizio, Erbil Akbudak, Thomas E. Conturo, Abraham Z. Snyder, John M. Ollinger, Heather A. Drury, Martin R. Linenweber et al. "A Common Network of Functional Areas for Attention and Eye Movements," Neuron, vol. 21, 1998, pp.761-773.

[3] Keith Rayner, "Eye movements in reading and information processing: 20 years of research," Psychological Bulletin, vol.124 no.3, 1998, pp.372-422.

[4] Lowe, David G. "Distinctive image features from scale-invariant keypoints." International Journal of Computer Vision, vol.60, no. 2, 2004, pp.91-110.

[5] Oliva, Aude, and Antonio Torralba. "Scene-centered description from spatial envelope properties." Biologically Motivated Computer Vision Lecture Notes in Computer Science vol. 2525, 2002, pp 263-272.

[6] Laurent Itti and Christof Koch. "A saliency-based search mechanism for overt and covert shifts of visual attention." Vision research, vol. 40, no. 10-12, 2000, pp.1489-1506.

[7] Laurent Itti and Christof Koch, "Computational modeling of visual attention," Nature reviews neuroscience, vol.2, 2001, pp.194-203.

[8] Torralba, Antonio, Aude Oliva, Monica S. Castelhano, and John M. Henderson. "Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search." Psychological review 113, no. 4, 2006, pp.766-786.

[9] Elazary, Lior, and Laurent Itti. "Interesting objects are visually salient," Journal of Vision, vol. 8, no. 3, 2008, pp.1-15.

[10] Cerf, Moran, E. Paxon Frady, and Christof Koch. "Faces and text attract gaze independent of the task: Experimental data and computer model." Journal of vision, vol. 9, no. 12, 2009, pp.1-15.

[11] Chun, Marvin M. "Contextual cueing of visual attention." Trends in cognitive sciences, vol.4, no. 5, 2000, pp. 170-178.

[12] Spain, Merrielle, and Pietro Perona. "Some objects are more equal than others: Measuring and predicting importance." In proceeding of Computer Vision 2008, 2008, pp. 523-536.

[13] M. Nyström and K. Holmqvist. "An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data." Behavior Research Methods, vol.42, no.1, 2010, pp.188--204.

[14] Dooseok Kang, Sukhan Lee, and Yu-Bu Lee, "Human Visual Attention with Context-Specific Top-down Saliency," in proceeding of International Conference of Robotics and Biomimetics 2011, 2011, pp.2055-2060.

[15] Ouerhani, Nabil, Roman von Wartburg, Heinz Hugli, and Rene Muri. "Empirical validation of the saliency-based model of visual attention." Electronic letters on computer vision and image analysis, vol.3, no. 1, 2004, pp.13–24.