Comparing Computer-predicted Fixations to Human Gaze

Yanxiang Wu
School of Computing
Clemson University
yanxiaw@clemson.edu

Andrew T Duchowski
School of Computing
Clemson University
andrewd@cs.clemson.edu

ABSTRACT

This paper investigates the human gaze on natural photos and compares them with the computer-predicted fixations by *Super Gaussian Component analysis (SGC)* about their similarity in image areas.

Author Keywords
Human eye fixation, saccadic eye movement, visual saliency, statistical saliency predict

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: undetermined

General Terms
Verification

INTRODUCTION

Human eyes can move at about three times per second via very rapid eye movement to reorient the high-resolving power of the fovea. This movement is called *saccades*. But when it comes to detecting the pattern information, we have to maintain a relatively stable gaze, which is called a*fixation*, for a period of time. There is a lot of research about what factors determine where will be the fixation location.*Visual Saliency* is a human vision mechanism that distinct some items in the world stand out from its neighbors and immediately grab ones attention. This mechanism contributes to the high efficiency of human vision and perceptual system that can efficiently process information in a complex scene.

Visual Saliency hypothesis believe that fixation sites are selected based on image properties generated in a bottom-up manner. Therefore it is the visual properties that determines what part of the image should be the fixation location. Researches in visual saliency tried to draw this mechanism to computer system so that computer program can achieve a faster processing speed in computer vision system. Itti[4] proposed a Saliency-Based Visual Attention model that extract many components such as intensity, color and orientations from input image and combine the analysis results into a *Saliency Map* that would be the fixation location. This is a Bottom-Up, Neural Network approach. For the past decades, extensive researches have been done to generate eye-movements and estimate the visual saliency.

There is another hypothesis called *Cognitive Control hypothesis* claimed that the fixation sites are selected based on the needs of the cognitive system in relation to the task. As opposite to the *Visual Saliency hypothesis*, this hypothesis believe that the eye fixation is determined by cognitive information gathering needs rather than inherent visual salience [3].

It was widely accepted that both of stimulus-driven bottom-up factor and task-driven top-bottom factors affect the eye-movements of subjects. The bottom-up approach is utilizing the saliency map that is pre-computed by low-level features to estimate the gaze allocation. While the top-bottom approach is mainly driven by tasks and previous researches indicate that saliency model work much better than random models [5].

In this experiment, the hypothesis is: the Fixation points generated by *Super Gaussian Component* analysis method is similar in its location to the human gaze. And in this paper we assume that computer vision system can utilize the human gaze in 2D image to segment or track object in the real scene.

BACKGROUND

Henderson conducted a series of experiments to identify the relationship between fixation and saliency map. And the results showed that cognitive factors are a critical and likely dominant determinant of fixation locations in the active viewing of scene[3]. However, it did not show the same result in non-task situation.

Bruce [1] conducted experiments that collected fixation density map based on human eye fixation points. Sun[5] filtered out the subject-wise inconsistency of the result and proposed a hypothesis that 1) Saliency is very sparse 2) High saliency value tends to be located surrounding the region with abundant structural information. And they find that these characteristics of saliency share great similarity with super-Gaussianity, which is synonymous with sparse and structurized in statistic.

Therefore, Sun[5] proposed a *Super Gaussian Component* Analysis framework that tries to solve the question of "What

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

components in visual images draw fixations" instead of the traditional question of "What properties draw attention". The *Super Gaussian Component(SGC)* analysis framework divide the input image into a bunch of small image patches and then use Kurtosis Maximization to search for the *SGC* pursuits which will be used to filter the original image to get the instant *response map*. Once the *response map* generated, Winner-Takes-All **(WTA)** and Inhibition-of-Return **(IoR)** principles will applied to the *response map* to get the SG component what will be used to estimate saccadic dynamically. And the self-information of the *SGC* will be used to estimate the Visual Saliency of input image during the progress, and the more SGC are involved, the more details will appear in the saliency map[5].

According to their result, a stimulus which is conform to super Gaussian distribution is more likely to gather human gaze. Provided with this conclusion, it reaches the more natural underlying of visual saliency. And the conclusion can be applied to many hard-to-do it perfect vision problems such as accuracy tracking, registration etc. Therefore, the interesting part of Xiaoshuai's research paper in this works would be mainly focused on its capability of estimating the saliency map.

Object detection and tracking is a classic topic of research in Computer Vision community. It is mainly about using computer vision algorithms to segment object from background and recognizing them by machine learning algorithms. The big challenge of object detection and tracking is not only because of the high intra-class variations likes shape, pose, appearances, but also because of the occlusion, illumination changes. Even if there are considerable progresses done over past years, the problem is still challenging because the algorithm may not feasible for real-time system. For example, Angela [2] proposed a robust Hough Forest algorithm for object detection and tracking. However, the algorithm is too complex to use for real-time system. Also, one big problem of temporal object tracking is the *drift* problem that the tracked object position will be offset from the actual one with the time going. Therefore, people are looking for an approach that is able to detect and track object with reasonable complexity and have reasonable accuracy.

In this paper, the motivation is to validate the feasibility of using the computer-predicted to detect object(s) in natural scenes. Our hypotheses is the *SGC* based statistic model is an effective way to find the saliency area in a natural scene since its similarity to human gaze in freely looking at photos, therefore, it should be able to provide sufficient saliency value to segment the proto-object from natural background.

The original research have already conducted experiment on standard benchmark image database and proposed its conclusion that the algorithm is robust to noise, contrast changes, brightness changes. In this research we will conduct experiments with 2 parts: 1) Validate the similarity of *SGC* to human gaze of specific photos in the benchmark database. The photo chosen will be in concerning about the object as a whole instead of the details of the object. 2) Validate the similarity of human gaze to *SGC* predicted fixation area in a series of new photos we get from natural scene. The photos we get for this test set would focus on objects such as human and manufactured objects in complex scene.

METHODOLOGY

Apparatus

The experiment apparatus is Tobii 1750 Eye Tracker. It is an unobtrusive equipment. The user will be required to sit in front of it and conducting a calibrate process before experiment begin. The Tobii can tracking the user with a degree of accuracy of 0.5 degree, about 50 pixels. The sample rate is 50 HZ and the measurement may be slightly imprecise.

Experiment design

The experiment will conducted with 20 subjects whose randomly chose from university student, they will be required to freely look at 24 images that consisted with 10 images in benchmark databases which denoted as group *BenchmarkSet* and 14 images that is taken in natural scene which denoted as group *NaturalPhotoSet*. Subjects will be required to watch at the center point of the screen in between images so that reducing the interference from the previous photo.

Subjects will be told that there is no tasks to do with the content of any photo, therefore they can look at the photos passively.

In order to compare and reduce the cognitive impact of the fixation area result from the order of photo viewing.The subjects will be divided to 2 groups that each of group consisted with 10 participants. The image displayed to the subject in the same group will be the same order, but the order will be randomly shuffled for each of the subject in the other group. And the results will be compared for both of the same image between human gaze and between human gaze and computer generated saliency map.

The approach used in this paper to compensate the imprecise is used in the experiment that require the use look at certain point for a while and measure the offset location of the tracked result.

Then the mean offset vector V^i_{off}will be calculated for every subject. And the compensated fixation for each fixation point detected is:

$$P^{ij}_c = P^{ij} + V^i_{off} \tag{1}$$

where P^{ij}_c is the compensated fixation location and P^{ij} is the tracked fixation location which is collected from the experiment.

To comparing the result of human gaze with the *SGC* generated visual saliency area. The first fixation will be ignored. The fixation density map is normalized to [0, 1], and only the fixation density that is larger than 0.5 will be preserved. In accordance to the original result. We use *Area Under ROC Curve (AUC)* to compare the results quantitatively.

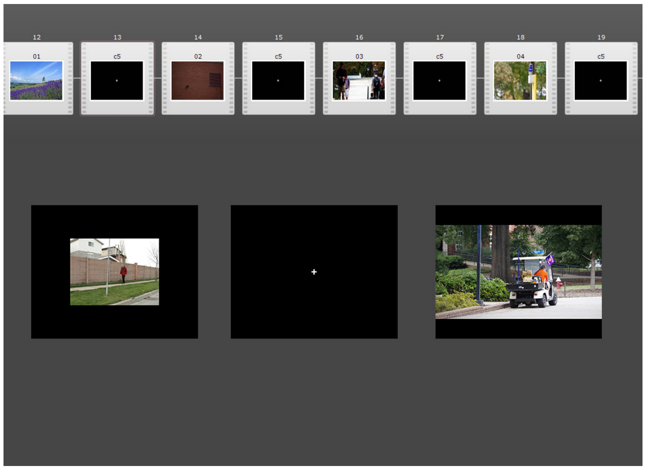

Figure 1. The workflow of testing image set in Tobii Studio

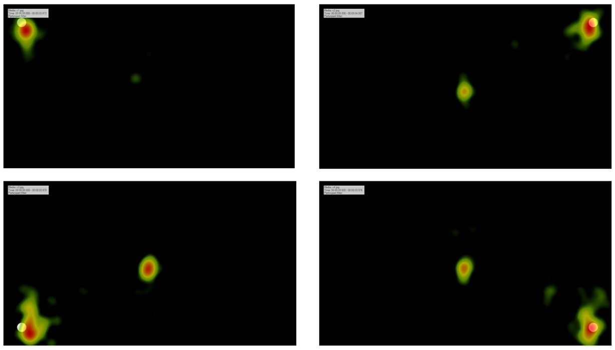

Figure 2. The overall fixation calibration result

RESULT AND ANALYSIS

We conducted the experiment with 21 participant, and 1 failed to complete the experiment. There are 20 valid participant whose sample rate in Tobii studio is larger than 83%. There are about 7463 valid fixation points left after eliminate the first fixation point for each image and normalized the fixation and remove the fixation value whose less than 0.5.

However, we find that the offset of tracking is unsuitable to compensate by the method we present in last section since the offset is neither linear or 3d surface. It is actually unbalanced due to the change of the user. Figure 3 and 4 demonstrated the situation. But We find the overall accuracy of the detected fixation is quite better than single participant (p ¡ 0.05). Figure 2 shows the result of the preset calibration image in four corners.

We have 2 groups of participants, one of them viewing the images in certain order and the other group viewing the images in random order. There is no affection from the order. (p ¡ 0.03).

According to Sun[5] , the AUC(SE score they get for the test set is 0.7903. And the AUC(SE score in our test is 0.6801 for the 27 generated the fixation points. And this value is

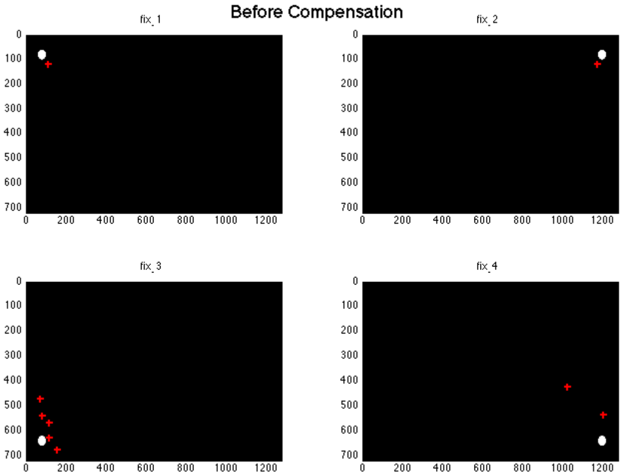

Figure 3. The fixation data used to generate compensation vector

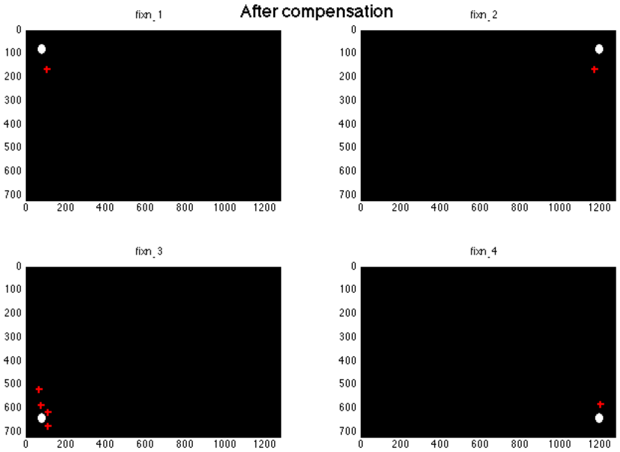

Figure 4. The fixation data used to generate compensation vector

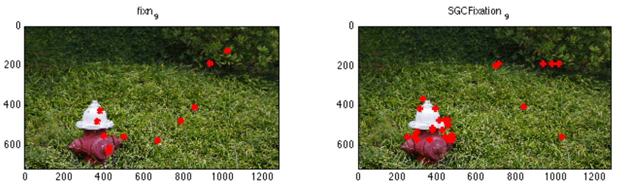

Figure 5. The result of human gaze (left) and the fixation generated by SGC

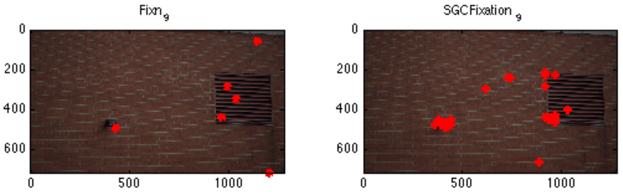

Figure 6. The result of human gaze (left) and the fixation generated by SGC

slightly changed since it is the fixation points is random variable in their model. Therefore, the fixation generated by this model is basically similar to the human gaze as it claimed in Sun's paper.

However, the story is slightly different if we exam the result in more detail. At first glance, the model do have very strong similarity to the human gaze like the result showed in Figure **??** and Figure **??**.

When it comes to the situation that did not contains any distinguishing objects. Figure **??** shows the participants tend to look at the bulb above the center and skim the table and bulb. Which is reasonable for us to understand what's in the scene. However, the fixation generated by SGC is quite unreasonable. It actually useless for object segmentation from the background.

And the situation is even more interesting when it comes to a complex scene with human presence. Figure 8 shows a scene with lots of vehicles. Human shows strong tendency to focus on the face while the SGC shows no preference for the face. Instead, it looks some area that have high contrast but not important in the image.

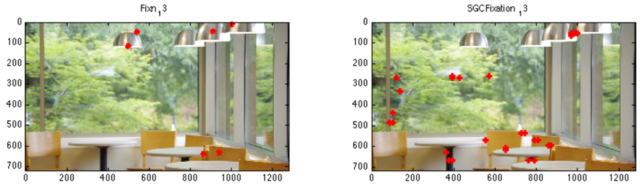

Figure 7. The result of human gaze (left) and the fixation generated by SGC in a scene that contains no distinguishing object. Human is tend to look bulb above the center. But the algorithm generate fixation in areas never saw by human

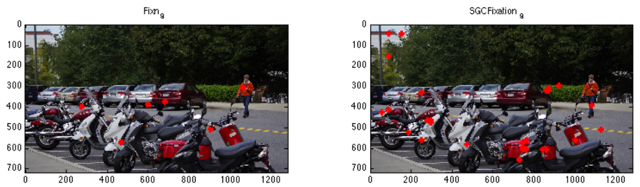

Figure 8. When the human presence, especially the face presence in the scene. There is different between the fixation

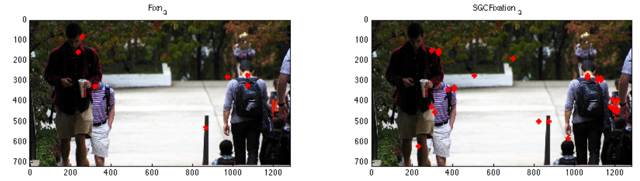

Figure 9. Most participants looks at the coke-cola, but SGC shows no bias for that.

Also Figure 9 shows another example for this fact. Most participants looked at the coke-cola and the guy who carry it. But SGC generated fixation basically focused on other areas.

Other discovery

When we draw the heat map which combines the fixation from all participants for each image, we discover a interesting situation. For the image group *BenchmarkSet*, the heat map clearly shows that the human gaze almost perfectly matches the potential proto-object in each images. However, this interesting result did not showed up in the *NaturalPhotoSet* . In fact, this discovery is opposite to our initial expectation since the images from *NaturalPhotoSet* have much more details (1280x720 down scaled from 16M pixel photo, while the photo from benchmark dataset is 1280x960 scaled up from 681x551) and better sense of beauty (according to the user's oral response after experiment, not in questionnaire).

And it is also showed more accuracy than the fixation data from the original benchmark set.

One possible explanation for this result maybe because we scaled up the image, and this result in certain level of blur. And the blur of detail may result in participant look for the details with more efforts.

CONCLUSION AND DISCUSSION

In this paper, we presented the result of an experiment to compare the human gaze to the fixation generated by *Super Gaussian Component* analysis. We exam their similarity by the AUC score. According to the result, *SGC* have good performance to generate fixation that similar to human gaze.

However, we also find that the situation is different when it comes to specific case of scene. If we classed the image to 3 different groups: 1) simple natural scene have no human or signs. 2) simple scene without principal object in the

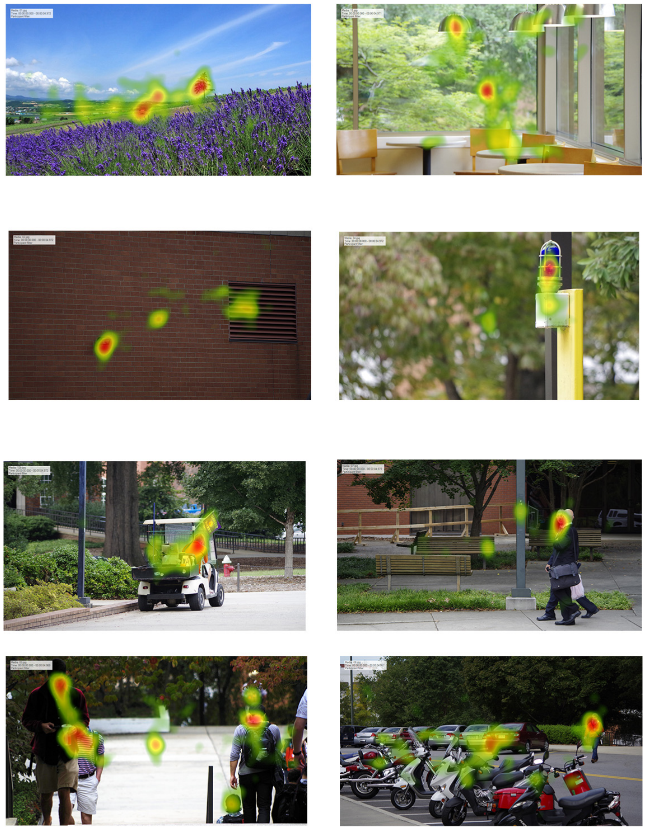

Figure 10. The human gaze in natural scene may affected by prior knowledge. Whenever human face presented, the fixation is largely focused on the human

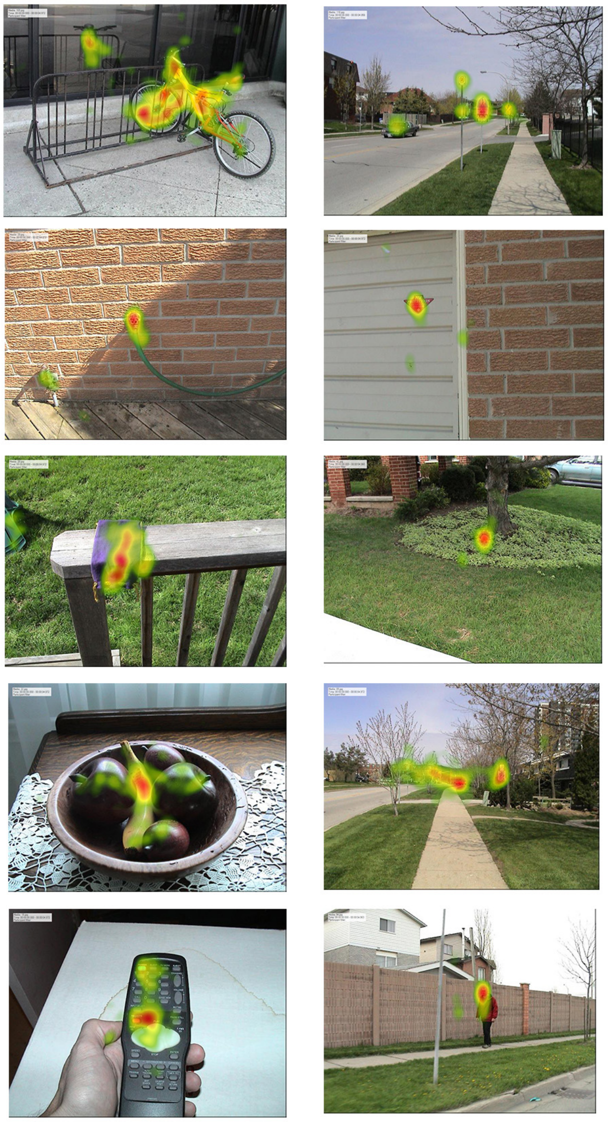

Figure 11. The human gaze of the images from *BenchmarkSet* sort of perfectly match the potential proto-object

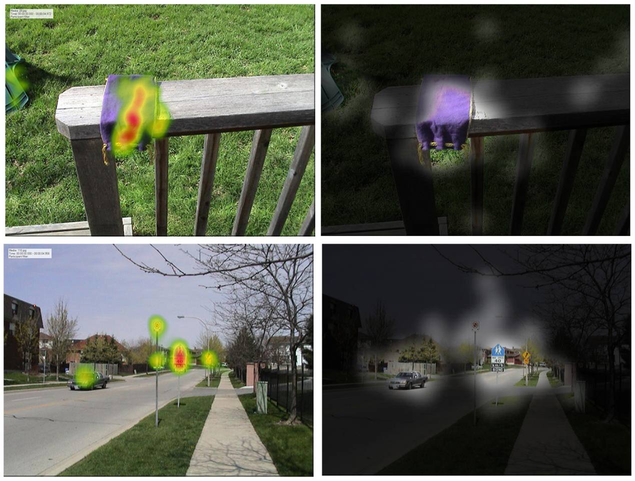

Figure 12. The human gaze heat map in our dataset (left), the human gaze heat map from the benchmark set raw data (right)

scene. 3) anything with human or signs. The *SGC* shows strong ability in 1). But a little bit less effective in 2). And is quite different from human gaze in 3). It is clear that we human understanding the world with our prior knowledge. And we look at the environment with the object regard our knowledge. Therefore, even *SGC* is incapable to track object as effective as our human being in terms of lack of the prior knowledge we have. However. *SGC* do provide a effective algorithm to generate fixations that may be useful for processing by later layers of the vision system.

In addition, we find that the participants' fixation in the stimulus from *BenchmarkSet* group seems match potential proto-object (AOI) more exactly than the result from the images from *NaturalPhotoSet* group. We have considered this situation in our experiment design. But this discovery indeed worthy to be researched.

REFERENCES

1. Bruce, N. D. B., and Tsotsos, J. K. Saliency based on information maximization. In *NIPS* (2005).

2. Gall, J., Yao, A., Razavi, N., Van Gool, L., and Lempitsky, V. Hough forests for object detection, tracking, and action recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 33*, 11 (nov. 2011), 2188 –2202.

3. Henderson, J., Brockmole, J., Castelhano, M., and Mack, M. Visual saliency does not account for eye movements during visual search in real-world scenes. *Eye movements: A window on mind and brain* (2007), 537–562.

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

5. Sun, X., Yao, H., and Ji, R. What are we looking for: Towards statistical modeling of saccadic eye movements and visual saliency. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (june 2012), 1552 –1559.