Expanding Saliency Models With Object Value: An Experimental Study

Mehrshad Tavana[1],Sajad Ahmadinabi[1],Mehran Khorshidi[1],Javad Khodadoust[1]

[1] *Department of Electrical Engineering ,Sharif University of Technology*

Abstract

When humans observe natural scenes without a specific task, their visual attention is naturally drawn towards areas or objects that exhibit salient features and stand out from the background. This phenomenon has been extensively studied using saliency models, which aim to predict eye movements during image or video viewing. These models provide insights into the fact that not all components of our visual environment are equally captivating. Certain objects have the ability to effortlessly capture our attention by "popping out" from their surroundings in a bottom-up fashion. The foundation of most bottom-up attention control models is the concept of a saliency map. A saliency map is a explicit two-dimensional map that encodes the conspicuity of objects in the visual environment. It highlights regions or objects that are visually distinct and likely to attract human attention. By analyzing these saliency maps, researchers gain a deeper understanding of the mechanisms behind attentional processes. In this paper, we obtained AUC-Judd, AUC-Borji, KL-divergence and NSS metrics for different features and observed that subband, Torralba and Horizon features were better compatible with human saccade patterns than other features. In addition, we conducted experiments to assess the robustness of these features against Gaussian and Salt & Pepper noises. Our findings indicate that all the features exhibited a high level of resistance. In order to explore the impact of object value on subjects' saccade patterns, we designed a task. Through this task, we discovered that subjects tend to direct their first saccades and subsequent fixations towards objects that possess either high or low value.

Keywords: Saliency Map; Fixation Map;Normalized Scanpath Saliency(NSS); Bottom-Up; Top-Down; Value

1. INTRODUCTION

The human brain is a remarkable organ, capable of processing vast amounts of information while selectively attending to stimuli relevant for perception and cognition. Computational models of saliency take images as input and generate a topographical map that represents the saliency levels of different regions in the image for a human observer. Most models of saliency computationally encode how different an area is from what surrounds it based on a model of a biologically plausible set of features that mimic early visual processing (Koch and Ullman, 1985). In saliency models, various low-level visual features such as intensity, color, orientation, texture, and motion are extracted from the image at different scales. These features are used to compute individual saliency maps for each feature. The saliency maps are then normalized and combined using linear or non-linear methods to create a master saliency map that represents the saliency of each pixel. The Itti and Koch Model (Itti and Koch, 2001) provides a theoretical framework that elucidates how visual attention operates within the human mind. A substantial body of evidence supports the two-component framework for the control of visual attention deployment (Treisman and Gelade, 1980). This framework suggests that attentional mechanisms consist of two distinct components (Theeuwes , 2010). The first component is known as the "top-down" control. It involves voluntary and goal-directed processes that guide attention based on internal factors such as task demands, expectations, and prior knowledge. The second component is referred to as the "bottom-up" control. This component

Images	Metric	Features						
		Subband	Itti	Color	Torralba	Horizon	Dist to center	All
Original	AUC-Judd	0.706 ± 0.020	0.472 ± 0.010	0.493 ± 0.008	0.722 ± 0.017	0.713 ± 0.017	0.207 ± 0.007	0.729 ± 0.018
	AUC-Borji	0.695 ± 0.020	0.457 ± 0.004	0.486 ± 0.005	0.712 ± 0.017	0.690 ± 0.019	0.204 ± 0.006	0.715 ± 0.018
	KL-divergence	9.032 ± 0.152	9.423 ± 0.042	9.366 ± 0.018	9.061 ± 0.069	9.456 ± 1.375	10.13 ± 0.1335	8.988 ± 0.152
	NSS	0.781 ± 0.516	-0.205 ± 0.239	-0.074 ± 0.150	0.801 ± 0.307	0.733 ± 0.367	-1.11 ± 0.127	0.928 ± 0.554
Gaussian Noise $\sigma = 0.01$	AUC-Judd	0.716 ± 0.020	0.470 ± 0.012	0.5058 ± 0.006	0.735 ± 0.018	0.738 ± 0.013	0.188 ± 0.006	0.744 ± 0.016
	AUC-Borji	0.706 ± 0.020	0.4518 ± 0.005	0.498 ± 0.005	0.725 ± 0.017	0.713 ± 0.016	0.185 ± 0.006	0.727 ± 0.016
	KL-divergence	9.015 ± 0.146	9.447 ± 0.047	9.362 ± 0.018	9.042 ± 0.072	9.202 ± 0.654	10.19 ± 0.135	8.973 ± 0.131
	NSS	0.806 ± 0.457	-0.257 ± 0.171	0.002 ± 0.099	0.848 ± 0.132	0.821 ± 0.351	-1.196 ± 0.122	0.950 ± 0.487
Gaussian Noise $\sigma = 0.1$	AUC-Judd	0.703 ± 0.025	0.451 ± 0.014	0.541 ± 0.009	0.720 ± 0.023	0.722 ± 0.019	0.184 ± 0.006	0.722 ± 0.023
	AUC-Borji	0.693 ± 0.024	0.437 ± 0.008	0.526 ± 0.007	0.711 ± 0.023	0.697 ± 0.022	0.182 ± 0.006	0.707 ± 0.022
	KL-divergence	8.960 ± 0.933	9.358 ± 0.933	9.245 ± 0.888	8.975 ± 0.883	9.159 ± 1.53	10.10 ± 1.17	8.935 ± 0.926
	NSS	0.767 ± 0.434	-0.283 ± 0.181	0.137 ± 0.074	0.815 ± 0.332	0.774 ± 0.362	-1.191 ± 0.131	0.890 ± 0.482
Salt & Pepper Noise $d = 0.05$	AUC-Judd	0.703 ± 0.025	0.455 ± 0.015	0.540 ± 0.009	0.721 ± 0.023	0.722 ± 0.019	0.184 ± 0.006	0.722 ± 0.023
	AUC-Borji	0.693 ± 0.024	0.440 ± 0.008	0.525 ± 0.007	0.711 ± 0.023	0.697 ± 0.022	0.181 ± 0.006	0.708 ± 0.021
	KL-divergence	8.957 ± 0.936	9.357 ± 0.933	9.246 ± 0.889	8.974 ± 0.882	9.158 ± 1.53	10.10 ± 1.17	8.932 ± 0.926
	NSS	0.769 ± 0.435	-0.279 ± 0.181	0.132 ± 0.074	0.817 ± 0.331	0.775 ± 0.365	-1.191 ± 0.131	0.890 ± 0.482

Table 1. Evaluation results of different metrics for different features

is automatic and reflexive, driven by salient stimuli in the environment that capture attention involuntarily. Saliency is independent of the nature of the particular task, operates very rapidly and is primarily driven in a bottom-up manner. Electrophysiological evidence suggests the presence of multiple neural maps in the pulvinar, superior colliculus, and intraparietal sulcus that are specifically involved in encoding the saliency of visual stimuli (Robinson and Petersen , 1992; Colby and Goldberg , 1999).

In this paper, we will calculate the AUC-Judd, AUC-Borji, KL-divergence, and NSS metrics for various features (Subband, Itti, Color, Torralba, Horizon, Distance to center). Subsequently, we will evaluate their robustness against Gaussian and Salt & Pepper noises. In the following, using the designed task (see materials and methods: 3.3), we will explore the influence of value on subjects' saccadic eye movements and fixation patterns.

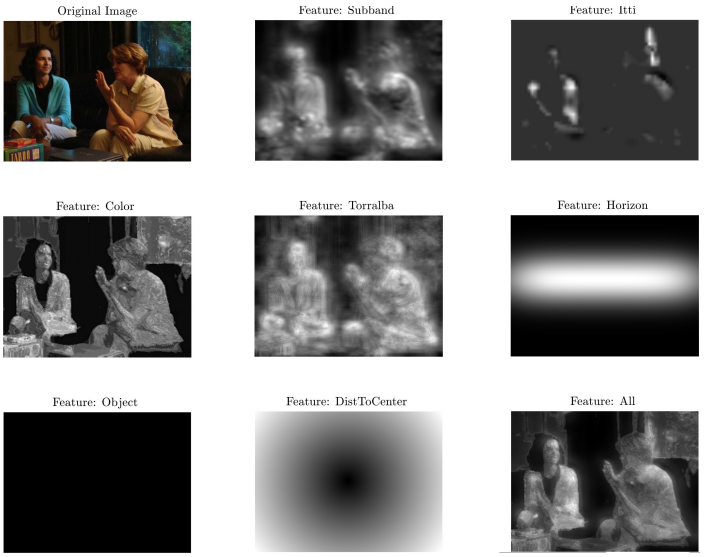

Figure 1: Different feature maps for a sample image.

2. RESULTS

2.1. Evaluate Saliency Models

In the first part of our study, we evaluated various features on the MIT dataset (see Materials and Methods) using two modes: without noise and with noise. For a sample image, these features are shown in figure 1. The performance of these features was measured using four metrics: AUC-Judd, AUC-Borji, KL-divergence, and NSS. The results are presented in table 1. The calculation of each metrics is explained in detail in the Materials and Methods section.

As depicted in table 1, the subband, torralba, and horizon features demonstrate a strong alignment with the saccade patterns observed in humans. This observation is supported by all four metrics utilized in our study. Also, in this table, we can observe that all the features exhibited resistance to gaussian and salt & pepper noises.

2.2. Effect of value on the subjects' saccade patterns

In the second part of the paper, we aimed to examine the impact of value on the subjects' saccade patterns through a designed task (see materials and methods). We can see some examples of heatmaps obtained for subjects in figure 2. The results of this task are presented in table 2.

The table provides information about the objects with the lowest and highest values for each subject. Additionally, it indicates the object towards which the

	Animals				Drinks				Foods				Instruments				Snacks			
Subjects	Min score	Max score	fist sacc	fixation	Min score	Max score	fist sacc	fixation	Min score	Max score	fist sacc	fixation	Min score	Max score	fist sacc	fixation	Min score	Max score	fist sacc	fixation
1	rhino	panda	rhino	panda	drink	red juice	drink	drink	-	all	pizza	pizza	trumpet	guitar	guitar	guitar	biscuit	seeds	cheetos	biscuit
2	rhino	panda	cat	panda	drink	orange juice	orange juice	orange juice	chicken	pizza	beef stew	chicken	trumpet	piano	trumpet	trumpet	cheetos	biscuit	cheetos	biscuit
3	rhino	panda	cat	rhino	drink	red juice	orange juice	orange juice	chicken	pizza	chicken	pizza	piano	trumpet	trumpet	trumpet	biscuit	seeds	seeds	seeds
4	rhino	cat	cat	panda	red juice	red juice	red juice	red juice	chicken	beef stew	pizza	pizza	trumpet	piano	trumpet	piano	biscuit	cheetos	cheetos	cheetos
5	rhino	panda	cat	panda	drink	red juicce	red juice	red juice	pizza	beef stew	pizaa	beef stew	trumpet	piano	guitar	piano	biscuit	cheetos	seeds	cheetos
6	rhino	cat	cat	cat	drink	red juice	drink	drink	pizza	beef stew	pizza	pizza	all	-	trumpet	trumpet	cheetos	seeds	seeds	seeds
7	-	all	cat	cat	drink	orange juice	orange juice	drink	pizza	chicken	chicken	chicken	-	all	trumpet	piano	cheetos	seeds	cheetos	cheetos
8	cat	panda	cat	rhino	drink	red juice	red juice	drink	pizza	chicken	pizza	beef stew	trumpet	piano	piano	piano	biscuit	cheetos	biscuit	biscuit
9	rhino	cat	cat	rhino	orange juice	drink	orange juice	orange juice	pizza	chicken	chicken	chicken	piano	trumpet	piano	piano	biscuit	cheetos	biscuit	biscuit
10	cat	rhino	rhino	panda	orange juice	drink	orange juice	red juice	pizza	chicken	chicken	pizza	trumpet	guitar	trumpet	trumpet	cheetos	seeds	seeds	cheetos
11	rhino	cat	cat	cat	drink	red juice	drink	red juice	all	-	beef stew	pizza	trumpet	piano	trumpet	piano	biscuit	cheetos	cheetos	biscuit
12	cat	rhino	cat	rhino	red juice	orange juice	orange juice	red juice	pizza	chicken	chicken	pizza	guitar	trumpet	trumpet	piano	cheetos	biscuit	cheetos	biscuit
13	rhino	panda	rhino	rhino	drink	red juice	drink	drink	beef stew	pizza	beef stew	chicken	trumpet	piano	trumpet	piano	biscuit	cheetos	biscuit	biscuit

Table 2: Results of behavioral task

Figure 2: Heatmaps of one subject

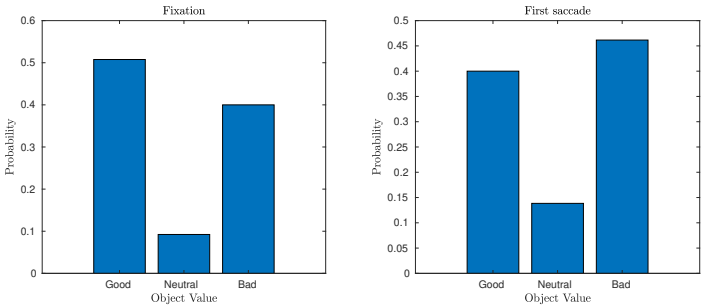

Figure 3: First saccade and fixation results.

subjects directed their first saccade and the object they subsequently fixated on.

By analyzing this table, we generated two bar plots in figure 3. These plots illustrate that the subjects' initial saccade and fixation tend to occur on objects with either high or low value, while the least frequent occurrence of first saccades and fixations was observed on neutral objects (objects with middle values). Furthermore, figure 3 reveals that the subjects predominantly directed their initial saccades towards objects with low values (figure 3,right), and subsequently exhibited a higher tendency to fixate on objects with high values (figure 3,left).

3. MATERIALS AND METHODS

3.1. Fixation Map

The fixation map provides a visual representation of the areas in an image or scene that draw the most attention from human observers. After recording the eye data of subjects, it is important to generate a fixation map that can be used to compare against models saliency map. To obtain the fixation map, we can utilize the fixation points recorded from subjects eye data. These fixation points indicate the specific locations where the subject's gaze was focused during the observation task.

Fixation points f^i for the ith observer are defined as

$$f^i(x) = \sum_{k=1}^{M} \delta(x - x_f(k)) \tag{1}$$

where x is the spatial coordinates vector, $x_f(k)$ is the spatial coordinates of the kth visual fixation, M is the number of visual fixations for the ith observer and $\delta(.)$ is the Kronecker symbol.

In the case of N observers, the final fixation map is obtained by calculating the average of their respective fixation points

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} f^i(x) \tag{2}$$

The fixation map S is obtained by convolving the fixation points f with an isotropic bidimensional gaussian function.

$$S(x) = f(x) * G_\sigma(x) \tag{3}$$

where σ is standard deviation of gaussian function. In general, σ is commonly considered equivalent to one degree of visual angle.

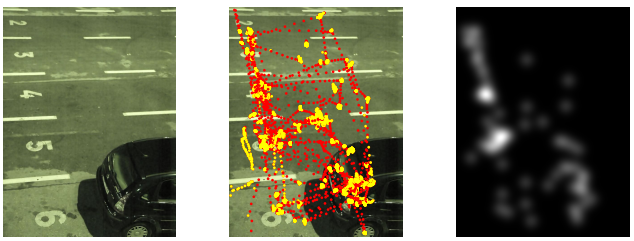

Figure 4: How to calculate fixation map. left: original image, middle: image and all subjects' fixations, right: fixation map

3.2. MIT Data-Set

The MIT dataset (Judd et al , 2009) comprises 1003 images that were used to record eye tracking data from fifteen users who freely viewed these images. The images had a longest dimension of 1024 pixels, and the other dimension varied between 405 and 1024 pixels, with the majority at 768 pixels. Out of the total images, 779 were in landscape orientation, while 228 were in portrait orientation. The viewers consisted of both males and

Figure 5: Sample stimulus images used in Our Task

females aged between 18 and 35. Two of the viewers were researchers involved in the project, while the rest were naïve viewers. All viewers were seated approximately two feet away from a 19-inch computer screen with a resolution of 1280x1024 pixels. The viewing environment was a dark room, and the viewers used a chin rest to stabilize their heads during the experiment. During each trial, the viewers' gaze paths were recorded by an eye tracker connected to a separate computer. The viewers were presented with each image at full resolution for a duration of 3 seconds, followed by 1 second of viewing a gray screen before the next image was shown.

3.3. Our Data-Set and Task

To investigate the effect of value on saccade patterns, we designed a task that involved five categories (Animals, Foods, Instruments, Snacks, Drinks) of objects (figure 5). Within each category, three specific objects were selected, and these objects were presented to the subjects as a single image. Prior to performing the task, subjects were required to determine the value of each of the 15 objects using MATLAB's Psychtoolbox. They assigned a value ranging from -10 to 10 for each object.Following the value assignment phase, subjects participated in the main task, which lasted for one minute. During this task, subjects viewed the five images in a free-viewing mode, with each image displayed for 12 seconds. The eye-tracking data of the subjects during this task was recorded using the website https://www.realeye.io. 13 subjects participated in this task.

3.4. Kullback-Leibler (KL) divergence

Kullback-Leibler (KL) divergence is a mathematical measure used in information theory to quantify the difference between two probability distributions. It provides a way to compare how one distribution, let's say P, differs from another distribution, Q. The KL divergence is defined as:

$$KL(P||Q) = \sum_x P(x) log \frac{P(x)}{Q(x)} \tag{4}$$

The KL divergence is always non-negative. The KL divergence with a value of zero indicating that the two distributions are identical. As the KL divergence increases, it signifies a greater dissimilarity between the distributions.

In the field of saliency research, the computation of Kullback-Leibler (KL) divergence can vary depending on how saliency predictions and ground truth fixations are interpreted as probability distributions. To prevent any potential confusion regarding the specific implementation of KL divergence, we referring to a particular variant as KL-Judd (Bylinskii et al , 2016). KL-Judd metric involves taking a saliency map P and a ground truth fixation map Q as input. The purpose of this metric is to assess the loss of information that occurs when using P as an approximation of Q:

$$KL(P||Q) = \sum_x Q(x) log(\epsilon + \frac{Q(x)}{\epsilon + P(x)}) \tag{5}$$

where ϵ is a regularization constant.

3.5. Area Under the Curve (AUC): AUC-Judd, AUC-Borji

In the context of predicting fixation locations on an image, a saliency map can be regarded as a classifier that determines which pixels are likely to be fixated and which ones are not. In signal detection theory, the Receiver Operating Characteristic (ROC) measures the tradeoff between true and false positives at various discrimination thresholds. The ROC curve is created by plotting a model's True-Positive Rate (TPR) versus its False-Positive Rate (FPR) at various classification thresholds. True Positive Rate (TPR) represents the probability that a positive sample is correctly predicted as positive. False Positive Rate (FPR) represents the probability that a negative sample is incorrectly predicted as positive. The Area Under the Receiver Operating Characteristic (ROC) Curve, commonly known as AUC, is a widely used metric for evaluating saliency maps. In the evaluation of saliency maps, the saliency map is conceptualized as a binary classifier that distinguishes fixated regions from non-fixated regions using different

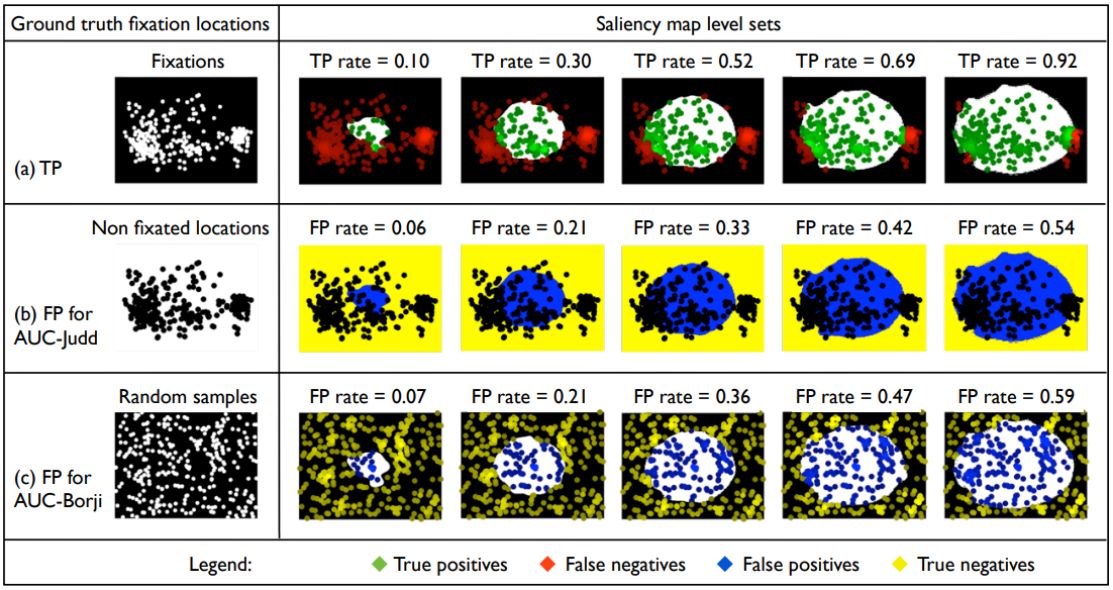

Figure 6: How true and false positives are calculated under different AUC metrics: AUC-Judd and AUC-Borji

threshold values or level sets (Bylinskii et al , 2016). The advantage of using AUC is that it allows us to compare different models without directly examining their ROC curves. There are different variants of AUC, two examples of which are: AUC-Judd (Judd et al , 2009) and AUC-Borji (Borji et al , 2013).

In AUC-Judd, the true positive rate (TP rate) for a given threshold is calculated as the ratio of true positives to the total number of fixations. True positives are determined by saliency map values above the threshold at fixated pixels. On the other hand, the false positive rate (FP rate) is the ratio of false positives to the total number of saliency map pixels at the given threshold. False positives are defined as saliency map values above the threshold at unfixated pixels.

AUC-Borji utilizes a uniform random sample of image pixels as negatives for evaluation purposes. It considers saliency map values above the threshold at these randomly sampled pixels as false positives. Figure 6 shows how to calculate AUC in Judd and Borji methods.

AUC ranges from 0 to 1, where higher values indicate better performance. A perfect classifier will have an AUC of 1, indicating that it can perfectly separate positive and negative samples. On the other hand, a completely random classifier will have an AUC of 0.5, indicating that its predictions are no better than chance.

3.6. Normalized Scanpath Saliency (NSS)

The normalized scanpath saliency (NSS) is a metric that involves a saliency map and a set of fixations (Le Meur and Baccino, 2013). The main idea behind this metric is to measure the saliency values at fixation locations along a subject's scanpath. To standardize the saliency values, you need to calculate the z-score of each value. The z-score represents the number of standard deviations a particular value is away from the mean. This process allows you to transform the saliency values into a distribution with a mean of zero and a standard deviation of one (figure 7).

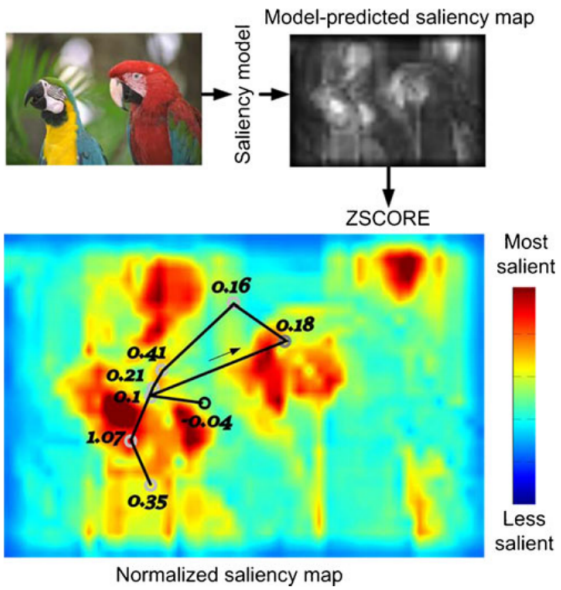

Figure 7: Example of normalized scanpath saliency computation.

The NSS value for a given fixation location is computed on a small neighborhood centered on that location. The NSS is the average of NSS for all fixations M of an observer (Algorithm 1). A higher NSS value indicates a better match between the saliency values and the fixation locations.

Algorithm 1 NSS algorithm

Step 1 : Standardize the saliency values in order to have a zero mean and unit standard deviation:

$$Z_{SM}(x) = \frac{SM(x) - \mu}{\sigma} \tag{6}$$

where Z_{SM} is the standardized saliency map and

$$\mu = \frac{1}{|I|} \sum_{t \in I} SM(x_t) \tag{7}$$

$$\sigma = \frac{1}{|I|} \sqrt{\frac{1}{|I|} \sum_{t \in I} (SM(x_t) - \mu)^2} \tag{8}$$

Step 2 : The NSS value for a given fixation location is computed on a small neighborhood centered on that location:

$$NSS(x_f(k)) = \sum_{j \in \pi} K_h(x_f(k) - x_j) Z_{SM}(x_j) \tag{9}$$

where K is a kernel with a bandwidth h and pi is neighborhood.

Step 3 : The NSS is the average of $NSS(x_f(k))$ for all fixations M of an observer. It is given by

$$NSS = \frac{1}{M} \sum_{k=1}^{M} NSS(x_f(k)) \tag{10}$$

4. DISCUSSION

In this paper, we conducted an analysis of various features and evaluated their performance based on the AUC-Judd, AUC-Borji, KL-divergence, and NSS metrics. Our findings revealed that the three features, namely subband, torralba, and horizon, exhibited the highest consistency with the saccade patterns observed in the subjects. Subsequently, we conducted an analysis to evaluate the robustness of these features against various types of noise, including gaussian and salt & pepper noise. Our observations revealed that all of the features demonstrated a remarkable resilience to these types of noise. In the subsequent section of the article, we aimed to investigate how the value assigned to objects influenced the saccade patterns exhibited by the subjects. To accomplish this, we designed a task in which the subjects were presented with five categories of images, each category containing three objects. Prior to engaging in the task, the subjects were asked to assign a value to each of the fifteen objects based on their individual judgments or predetermined criteria. This allowed us to examine the potential impact of object value on the subsequent patterns of saccadic eye movements displayed by the subjects. In this experiment, we made an intriguing observation: subjects exhibited a greater degree of attention towards objects that they had either rated with the highest score or the lowest score. This finding suggests that the subjects' attention was significantly influenced by the subjective value judgments they assigned to the objects. Objects with extreme values, whether highly positive or highly negative, appeared to attract more attention from the subjects during the task.

As a future work, the incorporation of value-based objects as an additional channel in salience models can be explored. This improvement in the model involves integrating information about the scores of objects assigned by subjects during the experiment and then fitting a score-based gaussian distribution to these objects.

REFERENCES

Koch C, Ullman S. Shifts in selective visual attention: towards the underlying neural circuitry. Hum Neurobiol. 1985;4(4):219-27.

Itti L, Koch C. Computational modelling of visual attention. Nat Rev Neurosci. 2001 Mar;2(3):194-203. doi: 10.1038/35058500. PMID: 11256080.

Le Meur O, Baccino T. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. Behav Res Methods. 2013 Mar;45(1):251-66. doi: 10.3758/s13428-012-0226-9. PMID: 22773434.

Treisman AM, Gelade G. A feature-integration theory of attention. Cogn Psychol. 1980 Jan;12(1):97-136. doi: 10.1016/0010-0285(80)90005-5. PMID: 7351125.

Theeuwes J. Top-down and bottom-up control of visual selection. Acta Psychol (Amst). 2010 Oct;135(2):77-99. doi: 10.1016/j.actpsy.2010.02.006. Epub 2010 May 26. PMID: 20507828.

Robinson, D.L. and Petersen, S.E. (1992) The Pulvinar and Visual Salience. Trends in Neurosciences, 15, 127-132. http://dx.doi.org/10.1016/0166-2236(92)90354-B

Colby CL, Goldberg ME. Space and attention in parietal cortex. Annu Rev Neurosci. 1999;22:319-49. doi: 10.1146/annurev.neuro.22.1.319. PMID: 10202542.

Itti, Laurent. Models of Bottom- Up and Top-Down Visual Attention. Dissertation (Ph.D.), California Institute of Technology. doi:10.7907/MD7V-NE41

T. Judd, K. Ehinger, F. Durand and A. Torralba, "Learning to predict where humans look," 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 2009, pp. 2106-2113, doi: 10.1109/ICCV.2009.5459462.

Z. Bylinskii, T. Judd, A. Oliva, A. Torralba and F. Durand, "What Do Different Evaluation Metrics Tell Us About Saliency Models?," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 3, pp. 740-757, 1 March 2019, doi: 10.1109/TPAMI.2018.2815601.

A. Borji, D. N. Sihite and L. Itti, "Quantitative Analysis of Human-Model Agreement in Visual Saliency Modeling: A Comparative Study," in IEEE Transactions on Image Processing, vol. 22, no. 1, pp. 55-69, Jan. 2013, doi: 10.1109/TIP.2012.2210727.