Evaluation of Saliency Models and Eye Tracking Data Using the Itti Model and Gazerecorder Tool

Ali Shahbazi[a]

[a]Student, EE. Department, Sharif University of Technology

This report presents an evaluation of the Itti saliency model's performance in predicting human attention patterns compared to ground truth eye-tracking data. Natural images from a provided dataset were processed using the Itti saliency model to generate saliency maps. The fixation points of selected subjects were overlaid on these maps to assess the model's accuracy. Additionally, the Gazerecorder tool was utilized to obtain human gaze locations on the images. Metrics such as Normalized Scanpath Saliency (NSS) and Receiver Operating Characteristic (ROC) curves were employed for quantitative analysis. The results revealed that while the Itti model showed good alignment with human gaze data for some images, discrepancies were observed for complex scenes and high-level features. The study highlights the importance of considering both low-level and high-level features in attention models and provides insights into the strengths and limitations of the Itti saliency model.

Itti saliency model | Visual attention | Gazerecorder tool | NSS | ROC | Fixation points

Introduction

In recent years, the field of visual attention and eye tracking has gained significant attention due to its applications in various domains, such as neuroscience, computer vision, and human-computer interaction. Understanding how humans perceive and selectively attend to visual stimuli can provide valuable insights into human cognition and behavior.

The purpose of this report is to evaluate the performance of the Itti saliency model, the provided dataset, and the data obtained from the Gazerecorder tool. The Itti model is a popular computational model that simulates human visual attention by generating saliency maps. The dataset used in this study consists of a collection of images, while the Gazerecorder tool enables the recording of eye movements and the generation of personal heat-maps.

By conducting this evaluation, we aim to gain a better understanding of the strengths and limitations of the Itti saliency model in capturing human visual attention. The findings from this study can contribute to advancements in visual attention research and aid in the development of more accurate and robust models for applications such as image recognition, object detection, and human-computer interaction.

Results

Itti Saliency Model. The Itti saliency model is a computational model that aims to simulate human visual attention by predicting the regions in an image that are most likely to capture human attention. It operates by analyzing low-level features of an image, such as color, intensity, and orientation, and combining them to create a saliency map (Figure 1).

The saliency map generated by the Itti model represents the spatial distribution of salient regions in the image. In other words, it assigns higher values to locations that are more likely to attract attention and lower values to less salient regions. This process mimics the way humans prioritize certain areas in their visual field based on the distinctiveness and relevance of visual features.

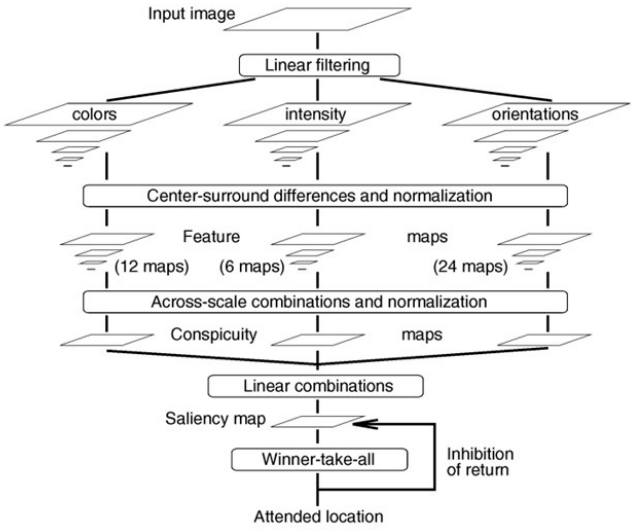

Fig. 1. General architecture and block diagram of Itti saliency model. Obtaining different saliency maps for each feature and then performing linear combination of the normalized maps.

The input images utilized in this study consist of natural images sourced from the data path, encompassing a total of four images with dimensions of 1024×768. To speed up the processing time, these images are resized to 640×480 (Figure 2). The saliency analysis is conducted using the Itti saliency-based model, which incorporates low-level features based on the methodology proposed by (1). By employing this model, distinct saliency maps are obtained, each corresponding to a specific low-level feature.

By integrating all the individual saliency maps derived from the low-level features, a comprehensive saliency map is generated, representing the combined intensity values for each pixel. This intensity map serves as an indicator of the probability of capturing human attention within the image (Figure 3).

The regions of the image that exhibit the highest saliency values can be leveraged to generate a subsequent map, which predicts the likely locations of gaze fixation. This new map is constructed by convolving the saliency map with a circular filter (Figure 4). The resulting map highlights the areas of the image that are more likely to capture and hold human gaze.

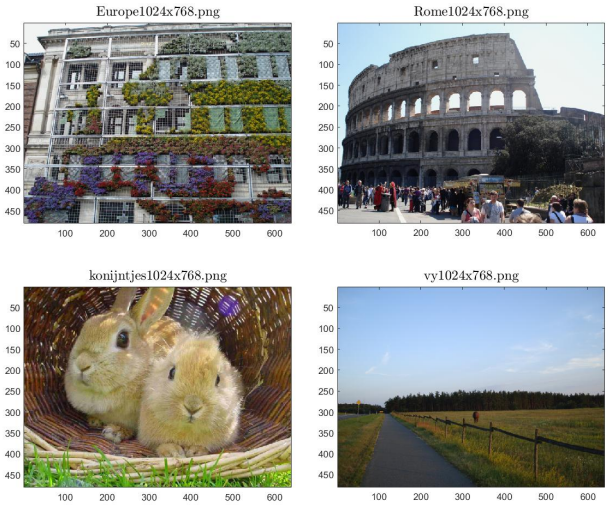

Fig. 2. Images as the input stimuli for the Itti saliency model, containing animals, nature, people, and buildings. Resized to lower dimensions to speed up the process.

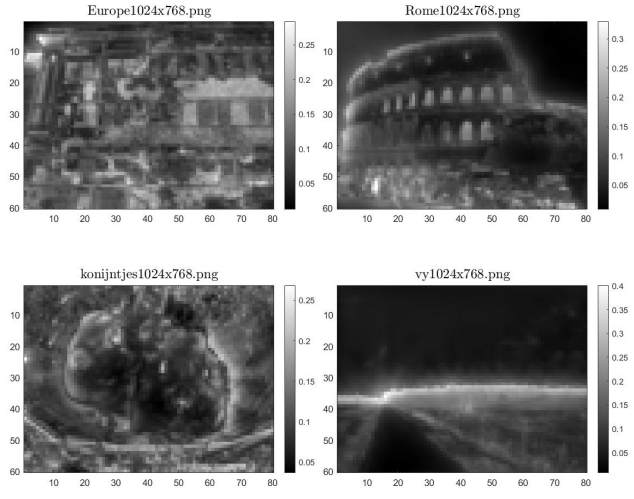

Fig. 3. Saliency maps for the input images. Brighter areas indicate higher chance to gain attention based on the low-level features.

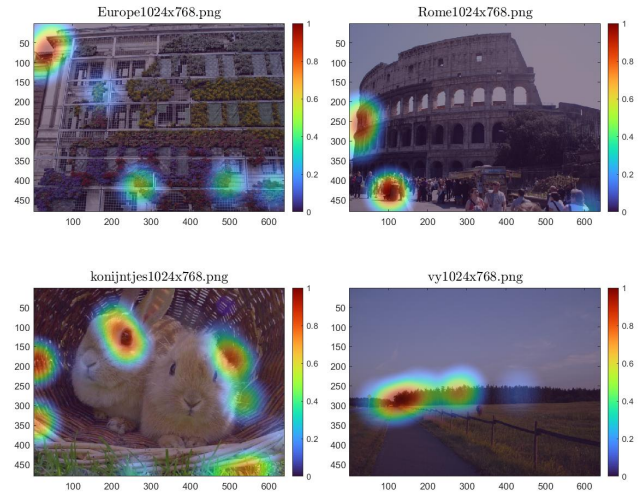

Fig. 4. Gaze map from the convolution of the saliency map and a circular filter. Regions with higher intensity are more likely to predict human gaze locations.

Eye Tracker Data. The eye tracker data contains information regarding the gaze location at each time bin, accompanied by corresponding labels indicating the type of eye movement (fixation, saccade, blink, etc.). Upon filtering out blink occurrences and extracting the fixation points, the resulting data representation is depicted in Figure 5. A noticeable observation is that the gaze map generated by the Itti model fails to capture numerous fixation points exhibited by the subject.

In the final image, there appears to be a relatively good alignment between the eye tracker data and the Itti map. However, certain scenarios, such as when presented with a face (as exemplified by the rabbit face in the third image), the Itti model encounters difficulties in accurately predicting the eye tracker data. This discrepancy can be attributed to the distinct types of attention being utilized. While the Itti model primarily focuses on extracting low-level features and bottom-up attention, human attention comprises a combination of both low-level and high-level features. For instance, attending to the face of an animal represents an example of high-level feature attention.

Gazerecorder Tool. By utilizing the Gazerecorder website, it is possible to create a new study and share the experiment link with others. This allows invited individuals to view the selected images and obtain their respective gaze locations. The experimental process involves a calibration step using a webcam, after which the actual experiment begins. In Figure 6, the gaze locations of a subject are visualized. However, it is important to note that access to download the data is restricted and requires payment.

As depicted in Figure 6, the fixation points exhibited by humans demonstrate a higher level of complexity compared to a simplistic low-level saliency map. Human attention is often directed towards high-level features such as faces, people, and actions. For instance, in the second image depicting Rome, the human subject's attention is primarily drawn to the individuals, their faces, and the ongoing actions, rather than focusing predominantly on the surrounding buildings. Similarly, in the third image featuring a rabbit, humans tend to exhibit a strong

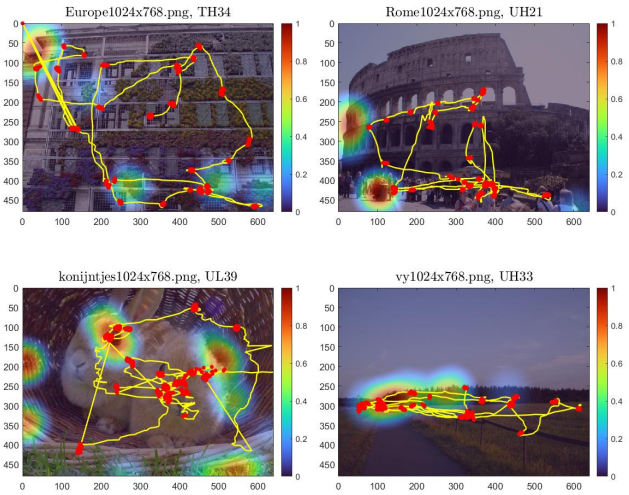

Fig. 5. The fixation points of the selected subject and gaze map generated from Itti model superimposed onto the input image.

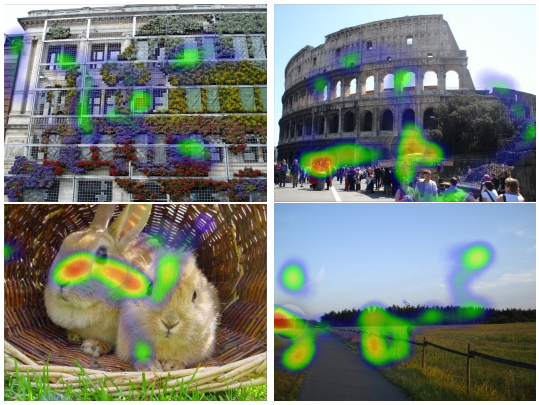

Fig. 6. Gaze locations of a subject recorded with the Gazerecorder tool.

inclination to attend to the animal's face and specifically its eyes. These observations highlight the significance of high-level features in influencing human attention patterns.

Normalized Scanpath Saliency and Receiver Operating Characteristics. Normalized Scanpath Saliency (NSS) is a metric used to assess the agreement between a saliency model and human eye-tracking data. It quantifies how well the saliency map generated by the model aligns with observed gaze patterns. NSS involves extracting saliency values from the model's map at the corresponding fixation locations in the eye-tracking data and normalizing them by subtracting the mean and dividing by the standard deviation. Positive NSS values indicate better alignment, suggesting that the model accurately predicts the regions attracting human attention. Negative NSS values indicate a mismatch between the model's predictions and the observed gaze patterns.

NSS provides researchers with a robust measure to evaluate and compare saliency models' performance in capturing human attention. It helps assess the ability of these models to identify salient regions that attract gaze. By using NSS, researchers can gain insights into the strengths and limitations of different saliency models, contributing to the development

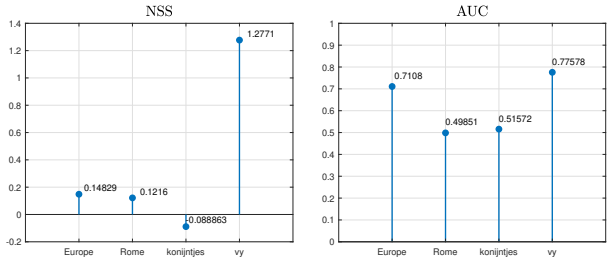

Fig. 7. The normalized scanpath saliency and receiver operating characteristic obtained from comparing the ground truth human subject data and Itti saliency map.

and improvement of visual attention modeling techniques.

The Receiver Operating Characteristic (ROC) curve is a useful tool for comparing saliency models to human eye-tracking data. It illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) at different discrimination thresholds.

To construct an ROC curve, saliency models assign saliency values to image locations and compare them to the ground truth eye-tracking data. Varying the threshold for saliency values generates different points on the curve, representing different trade-offs between sensitivity and specificity. A model with better performance will have a curve closer to the top-left corner, indicating a higher true positive rate and a lower false positive rate.

The area under the ROC curve (AUC) summarizes the overall performance of a saliency model. A higher AUC value signifies better discriminative ability between salient and non-salient areas. By analyzing the ROC curves and comparing AUC values, researchers can quantitatively assess and compare the performance of different saliency models in capturing human attention patterns observed in eye-tracking data.

Within the implemented code, we used the graph-based visual saliency (GBVS) method to calculate the Area Under the Curve (AUC). The obtained results are depicted in Figure 7. Notably, the AUC score indicates the level of alignment between the Itti saliency map and the human eye-tracking data.

In particular, the results reveal that the Itti saliency map and the human data are perfectly aligned for the last image (vy). However, for the second and third images (Rome and konijntjes), there is a significant disparity. While the AUC value for the first image (Europe) is relatively high at 0.71, the NSS measure remains relatively low at 0.15.

Discussion

The conducted study focused on evaluating the performance of the Itti saliency model in predicting human attention patterns compared to ground truth eye-tracking data. To achieve this, the study followed a systematic approach, starting with retrieving natural images from the provided dataset and feeding them into the Itti saliency model to obtain saliency maps. The fixation points of the selected subjects were then plotted on these maps to visualize the model's performance.

The results of this evaluation revealed some interesting findings. It was observed that the Itti saliency model generally showed good alignment with the human eye-tracking data, particularly for certain images. However, there were notable discrepancies in its performance across different stimuli. For

instance, in images with complex scenes, such as the second image (Rome) and the third image (konijntjes), the model struggled to accurately predict the eye-tracking data. This mismatch can be attributed to the fact that human attention is not solely based on low-level features, which the Itti model primarily focuses on, but is also influenced by high-level features, such as faces and actions.

To further evaluate the performance of the Itti saliency model, the study employed metrics such as Normalized Scanpath Saliency (NSS) and Receiver Operating Characteristic (ROC) curves. These metrics provided quantitative measures for comparing the model's predictions to the eye-tracking data. The NSS score revealed the extent of correspondence between the model's saliency map and human fixation points, while the AUC from ROC analysis indicated the overall discriminative ability of the model in distinguishing between salient and non-salient regions.

In conclusion, the study shed light on the strengths and limitations of the Itti saliency model in predicting human attention patterns. While the model demonstrated good alignment for certain images, its performance was affected by the complexity of scenes and the incorporation of high-level features in human attention. The findings underscore the importance of considering both low-level and high-level features in attention models.

Materials and Methods

Itti Saliency Model. The codes used in this study have been obtained from the mentioned repository. Specifically, the `itti_model_demo.m` script has been customized and utilized to generate the figures presented in this report. The modified versions of the codes can be accessed in the dedicated `Code` folder, provided alongside this report.

Eye Tracker Data. The eye tracking data used in this study is acquired from a high-speed system provided by SensoMotoric Instruments (SMI), featuring a sampling frequency of 500 Hz. A comprehensive description of the dataset can be found in the work by Larsson et al. (2). Additionally, this dataset has been utilized in subsequent studies by Larsson et al. (3) and Andersson et al. (4), further highlighting its significance in eye tracking research.

Gazerecorder Tool. The human gaze locations on the images were obtained using the web-based version of the Gazerecorder tool, accessible through the website https://app.gazerecorder.com/. It is important to note that, unfortunately, downloading the data from the tool is not available without purchasing the app.

GBVS. The GBVS method to compute AUC was obtained from this website that seems to be unavailable for now. Codes are available in the `Code/GBVS` folder.

NSS Calculation. The code for computing the NSS is retreived from this repository.

ACKNOWLEDGMENTS. We would like to sincerely acknowledge our supervisor, Dr. Ebrahimpour, for generously granting us additional time to complete this assignment. We also extend our thanks to the authors of the Itti saliency model and the creators of the Gazerecorder tool, whose resources and tools were indispensable for our research. The dataset made available in the mentioned repository served as a crucial component for our evaluation.

References

1. L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998. .
2. Linnéa Larsson, Marcus Nyström, and Martin Stridh. Detection of saccades and postsaccadic oscillations in the presence of smooth pursuit. *IEEE Transactions on biomedical engineering*, 60(9):2484–2493, 2013.
3. Linnéa Larsson, Marcus Nyström, Richard Andersson, and Martin Stridh. Detection of fixations and smooth pursuit movements in high-speed eye-tracking data. *Biomedical Signal Processing and Control*, 18:145–152, 2015.
4. Richard Andersson, Linnea Larsson, Kenneth Holmqvist, Martin Stridh, and Marcus Nyström. One algorithm to rule them all? an evaluation and discussion of ten eye movement event-detection algorithms. *Behavior research methods*, 49:616–637, 2017.