ASSIGNMENT 3

“Analyzing Salient Regions of Images: A Comparison of FOA Model Predictions and Human Fixation Patterns”

Vision Neuroscience
Prof. Ebrahimpour

Deniz Rezapur
denizrezapur@gmail.com

Abstract:

This report presents a visual attention system that is inspired by the behavior and neuronal architecture of the early primate visual system. The system combines multiscale image features into a single topographical saliency map and uses a dynamical neural network to select attended locations in order of decreasing saliency. By rapidly selecting conspicuous locations to be analyzed in detail, the system efficiently breaks down the complex problem of scene understanding. Its computational efficiency, accuracy, and ability to outperform other state-of-the-art visual attention models make it a promising tool for a variety of real-world applications such as robotics, autonomous vehicles, and image and video analysis.

Introduction:

Attention is a fundamental aspect of human perception and cognition, allowing us to selectively process and prioritize information from our environment. Models of attention in computational neuroscience have been developed to understand this process and to create artificial systems that can efficiently process sensory inputs. One such model is the visual attention system, which aims to replicate the behavior and neuronal architecture of the early primate visual system.

In recent years, there has been significant progress in developing visual attention systems that can efficiently extract salient features from complex scenes. One approach is the "feature integration theory," which explains human visual search strategies. This model decomposes visual input into a set of topographic feature maps, and different spatial locations then compete for saliency within each map, such that only locations which locally stand out from their surround can persist. All feature maps feed, in a purely bottom-up manner, into a master "saliency map," which topographically codes for local conspicuity over the entire visual scene.

In this report a visual attention system that is built upon the "feature integration theory" is reviewed. The model uses a saliency map that codes for local conspicuity over the entire visual scene and is endowed with internal dynamics that generate attentional shifts. Unlike other models of attention that require top-down guidance to shift attention, this model represents a complete account of bottom-up saliency. It provides a fast and massively parallel method for selecting a small number of interesting image locations to be analyzed by more complex and time-consuming object recognition processes.

Material & Methods:

The visual attention system presented in this report takes static color images as input, typically digitized at 640 x 480 resolution. The input image is then processed using dyadic Gaussian pyramids to generate nine spatial scales, each with different horizontal and vertical reduction factors ranging from 1:1 to 1:256 in eight octaves. Each feature is computed by a set of linear "center-surround" operations, which are similar to visual receptive fields. These operations are particularly well-suited to detecting locations that stand out from their surround and are a general computational principle in the retina, lateral geniculate nucleus, and primary visual cortex. Despite its simplified nature, the model reproduces many of the behaviors of Wang's original spiking neuron model.

In this model, center-surround is implemented as the difference between fine and coarse scales. This approach allows the system to efficiently detect salient features in the input image and to generate a topographical saliency map that highlights the most conspicuous locations in the scene.

The visual attention system described in the report uses a combination of feature maps for intensity, color, and orientation to create three "conspicuity maps." These maps are then used to compute a saliency map, which serves as the input to a winner-take-all neural network.

The saliency map is modeled as a layer of leaky integrate-and-fire neurons, which receive excitatory inputs from the conspicuity maps. The winner-take-all network ensures that only the most active location in the saliency map remains, while inhibiting other locations to prevent immediate return to previously attended locations.

In addition, the model incorporates a mechanism to bias the selection of subsequent salient locations to those spatially close to the current focus of attention. This is achieved by activating a small excitation in the saliency map in a near surround of the current focus of attention.

The visual attention system is based on biologically plausible mechanisms and has been shown to replicate several phenomena observed in human visual processing, including the "inhibition of return" effect. It provides a useful framework for understanding the neural basis of attention and has potential applications in areas such as object recognition and scene understanding.

The focus of attention (FOA) is modeled as a simple disk with a fixed radius of one sixth of the smaller input image dimension. The time constants, conductances, and firing thresholds of the simulated neurons are chosen to ensure that the FOA jumps from one salient location to the next in approximately 30-70 ms and that an attended area is inhibited for approximately 500-900 ms. This prevents cycling through a limited number of locations and ensures thorough scanning of the image. The system is stable over time for all images studied, and all parameters are fixed in the implementation.

Results:

1- **Results of model for the given pictures**

In this section of my report, I utilized the FOA (Fixation Over Areas) model code to predict the points of fixation on the given pictures and generate their corresponding saliency heat maps. (Fig.1, Fig.2, Fig.3, Fig.4, Fig.5).

For one of these pictures (Europe picture), the three maps of intensity, color, and orientation are mentioned as an example. (Fig.6).

It is important to note that the FOA (Fixation Over Areas) model is deterministic, meaning that when the same picture is inputted into

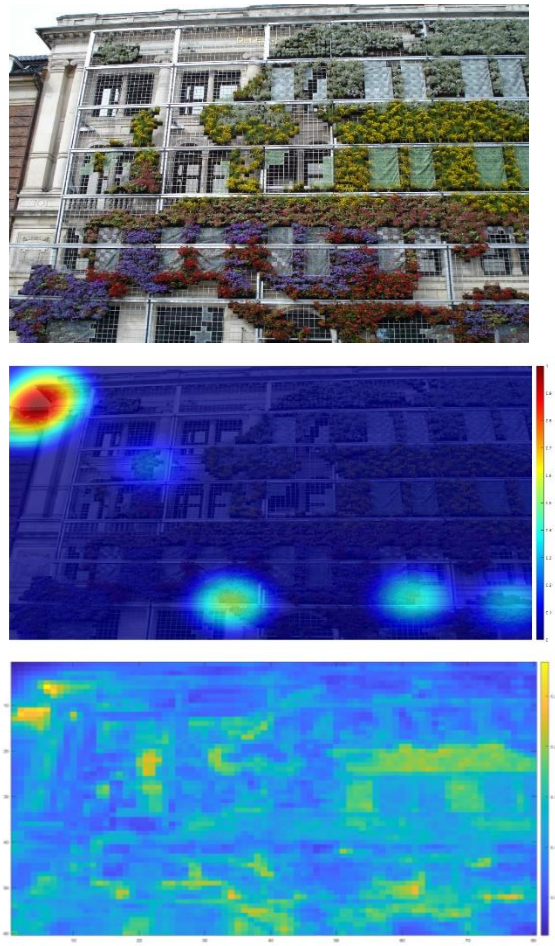

Figure 1- Europe Image: Up: Original image. Middle: Heatmap of FOA points. Down: Overall Salience Map

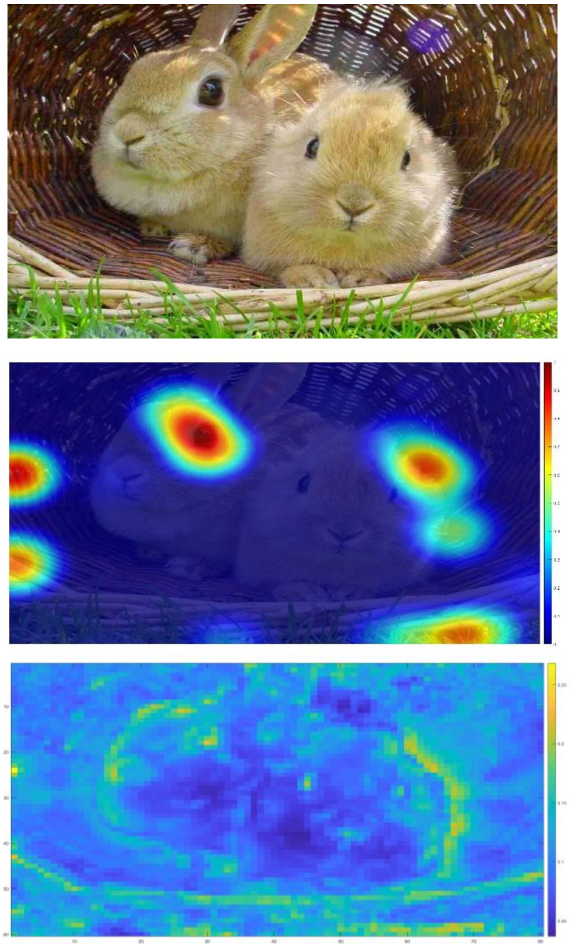

Figure 2- Rabbits Image: Up: Original image. Middle: Heatmap of FOA points. Down: Overall Salience Map

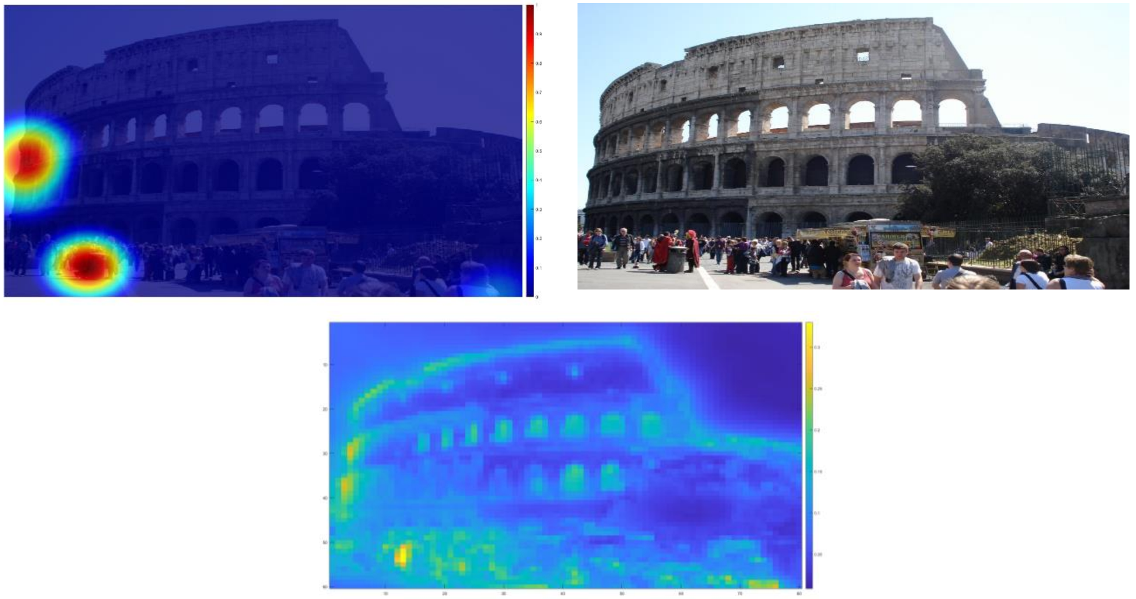

Figure 3- Rome Image: Up-Right: Original image. Up-Left: Heatmap of FOA points. Down: Overall Salience Map

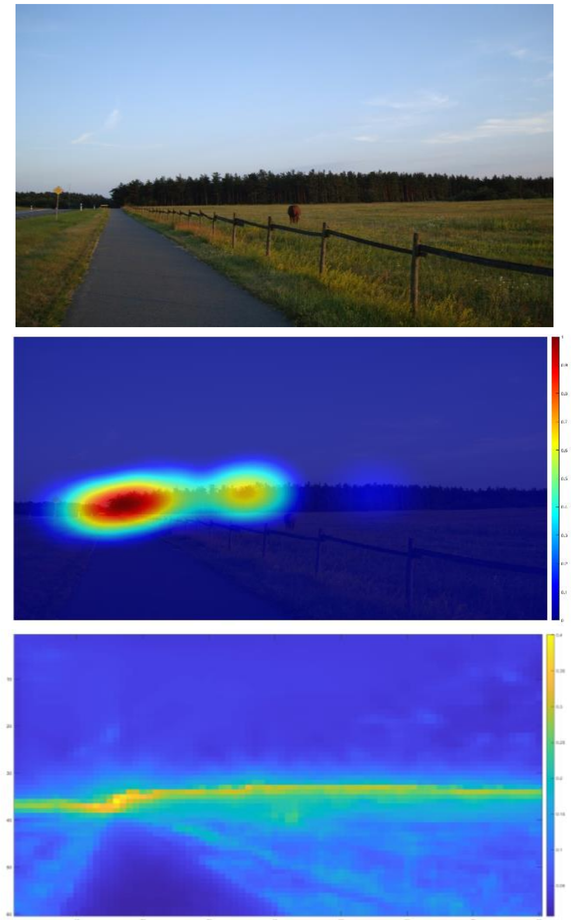

Figure 4- VYImage: Up: Original image. Middle: Heatmap of FOA points. Down: Overall Salience Map

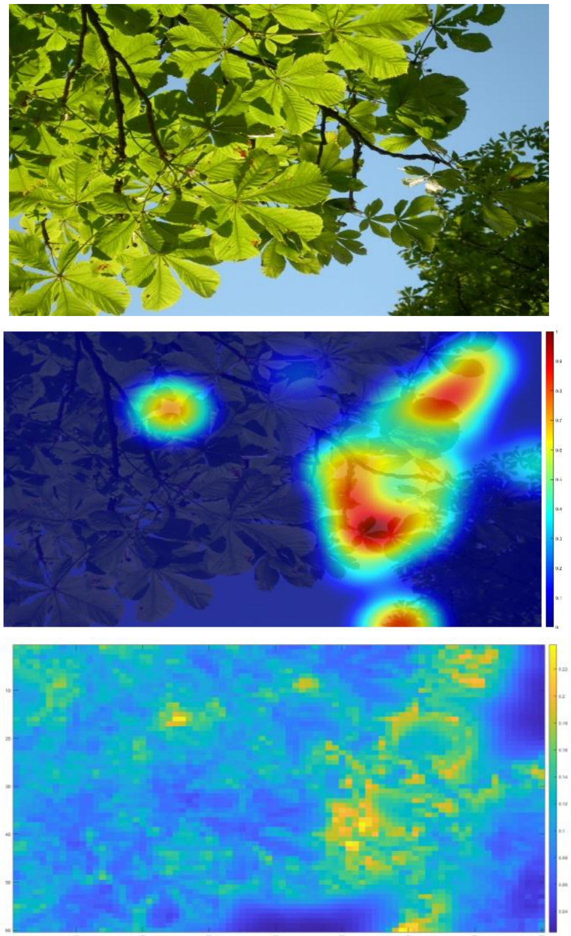

Figure 5- Blad Image: Up: Original image. Middle: Heatmap of FOA points. Down: Overall Salience Map

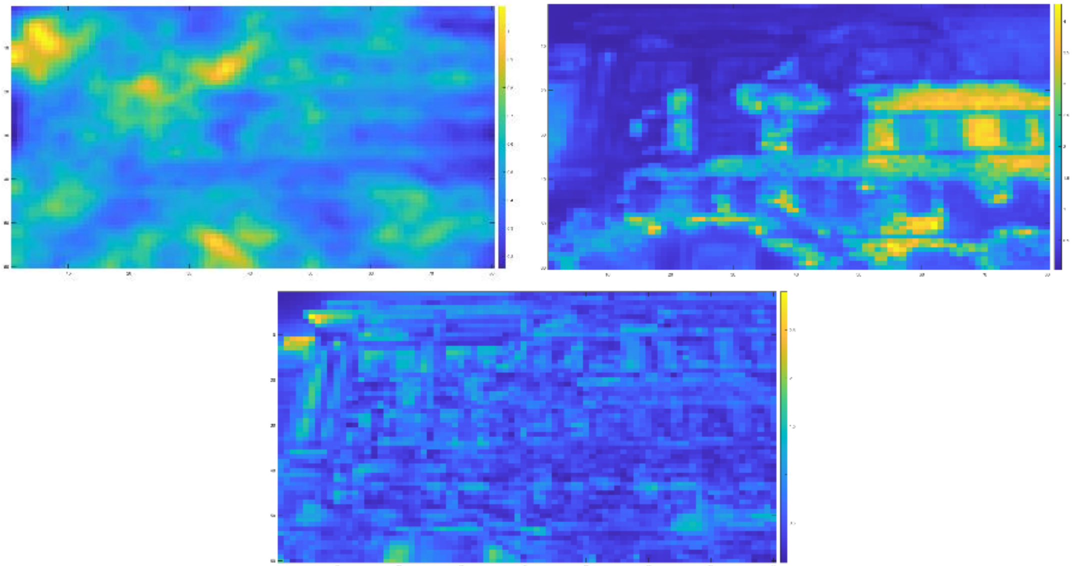

Figure 6- Europe Image: Up-Right: Color Overall Map. Up-Left: Orientation Overall Map. Down: Intensity Overall Map

the model multiple times, the resulting FOA points and saliency heat map will be identical each time. This characteristic of the model ensures consistency in its predictions and allows for reliable analysis of the salient regions of the images.

2- Results of the given dataset

In this particular dataset, only the first four pictures were utilized, and the Blad picture was omitted and not used in the analysis. To analyze the saliency of these four images, I utilized two manual functions to fit the given points to the pictures. For each of the four images, several data files were provided, and I applied the functions to all of them. By doing so, I obtained the following best results (Fig.7, Fig.8, Fig.9, and Fig.10)

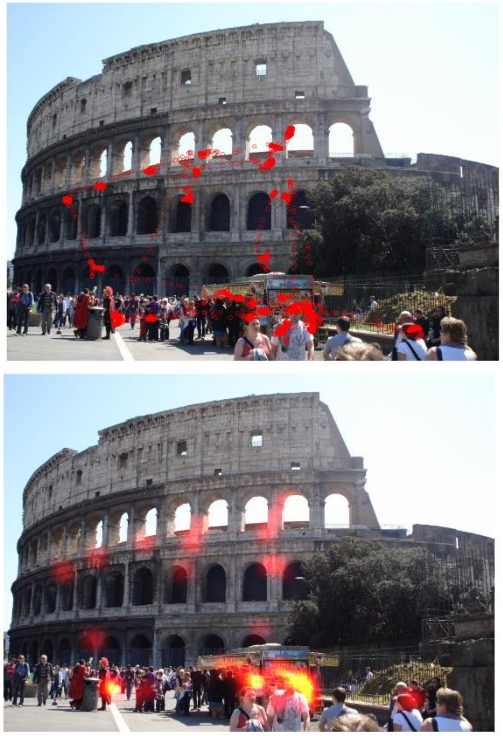

Figure 8- Rome Image. Up: eye data points Down: Heatmap

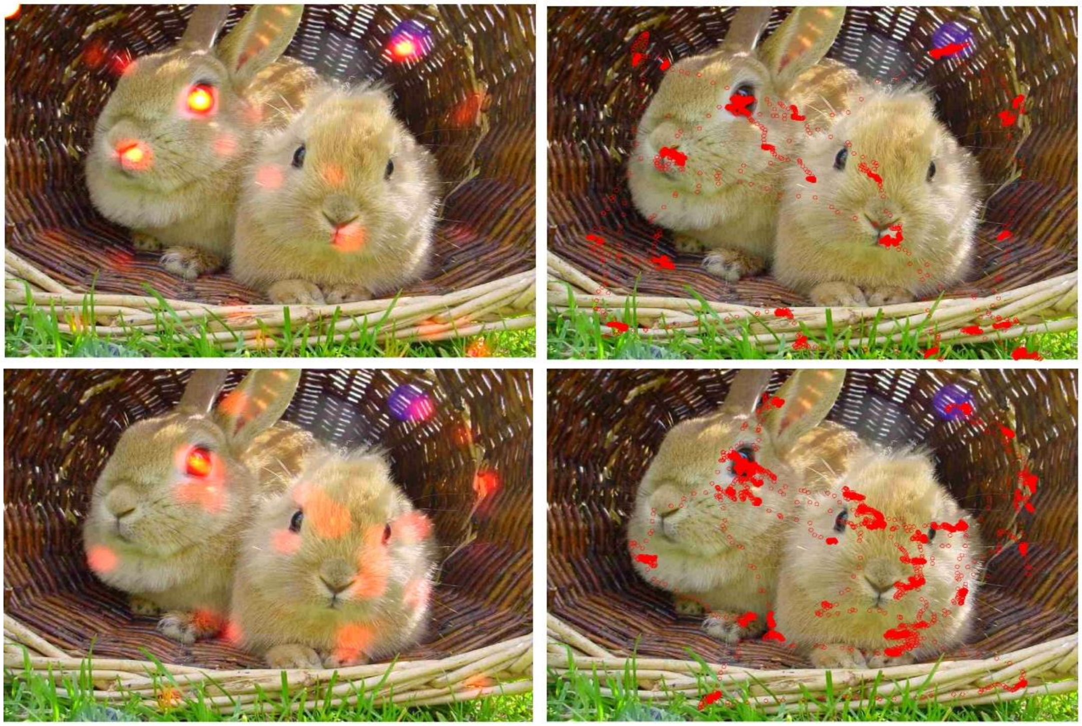

Figure 7- Rabbit Image. Up and down images are the best results for this picture. The right column pictures are points of eye data and the left column pictures are the heatmaps

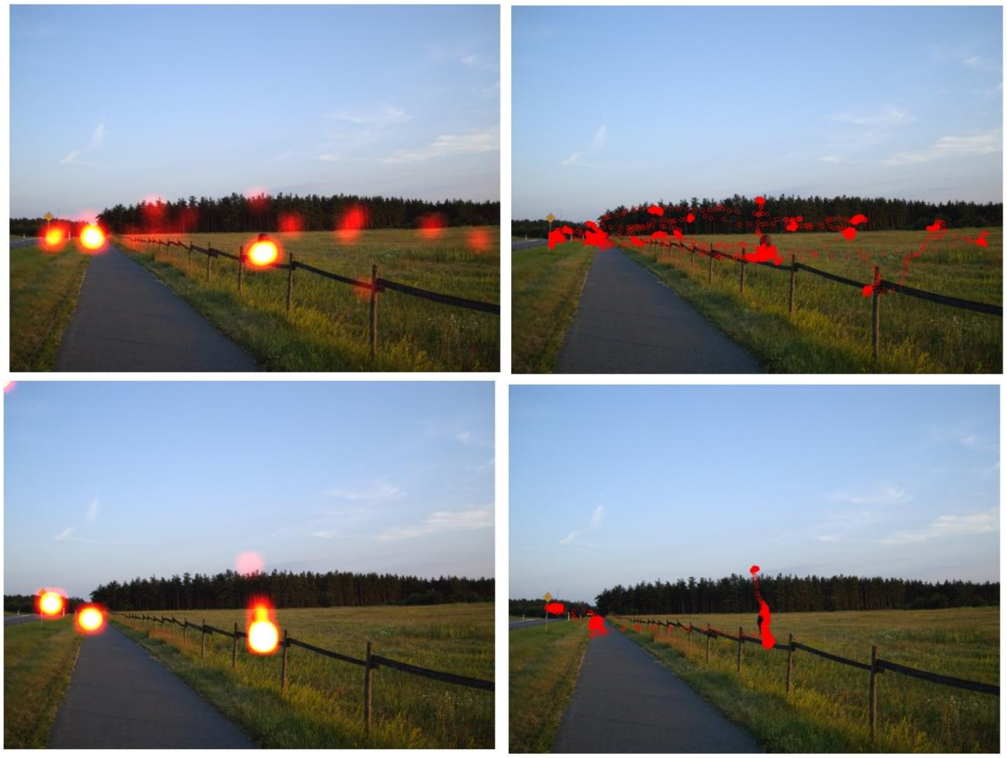

Figure 9- VY Image. Up and down images are the best results for this picture. The right column pictures are points of eye data and the left column pictures are the heatmaps

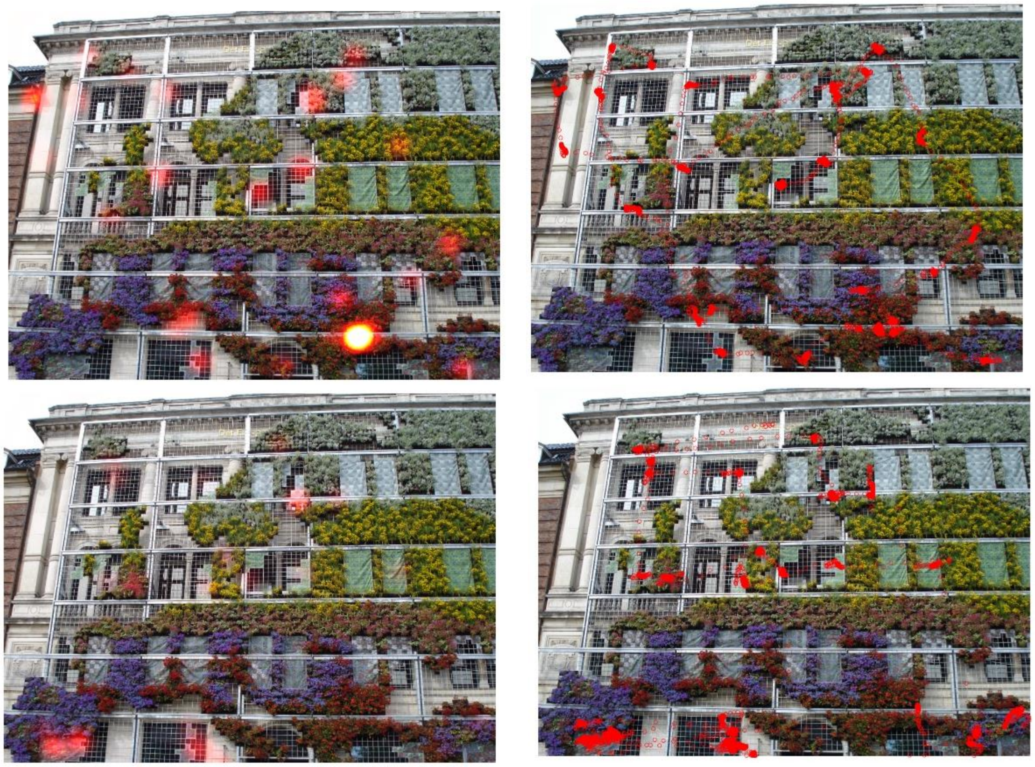

Figure 10- Europe Image. Up and down images are the best results for this picture. The right column pictures are points of eye data and the left column pictures are the heatmaps

3- Heatmaps of my own gaze recording data

In this section, I utilized an online gaze recorder service to record data on the fixation points of human subjects while viewing the four images for which subject data files were available. The service provided an Excel file containing the recorded fixation points, as well as a heat map of the salient regions of the images. These recorded fixation points will be used in the next section to calculate the ROC measure. The resulting heat maps generated by the gaze recorder service can be seen in Figure 11.

It is important to note that the heat maps generated using the online gaze recorder service were recorded prior to any analysis of the results from the FOA model or the subject data files. This was done to ensure that the recorded fixation points were not biased by any prior knowledge of the salient regions of the images. The recorded heat maps were used as a baseline for comparison with the results obtained from the FOA model and the subject data files in the subsequent sections of the report.

4- Comparison of the results of the model, the results of the subjects and my own results

In this section, I will compare the results obtained from the FOA model, the data recorded by myself using an online gaze recorder service, and the data obtained from the subject data files. It should be noted that the gaze recorder services that I found and was recommended in the assignment file did not provide the heat map separately from the pictures and Excel files containing the

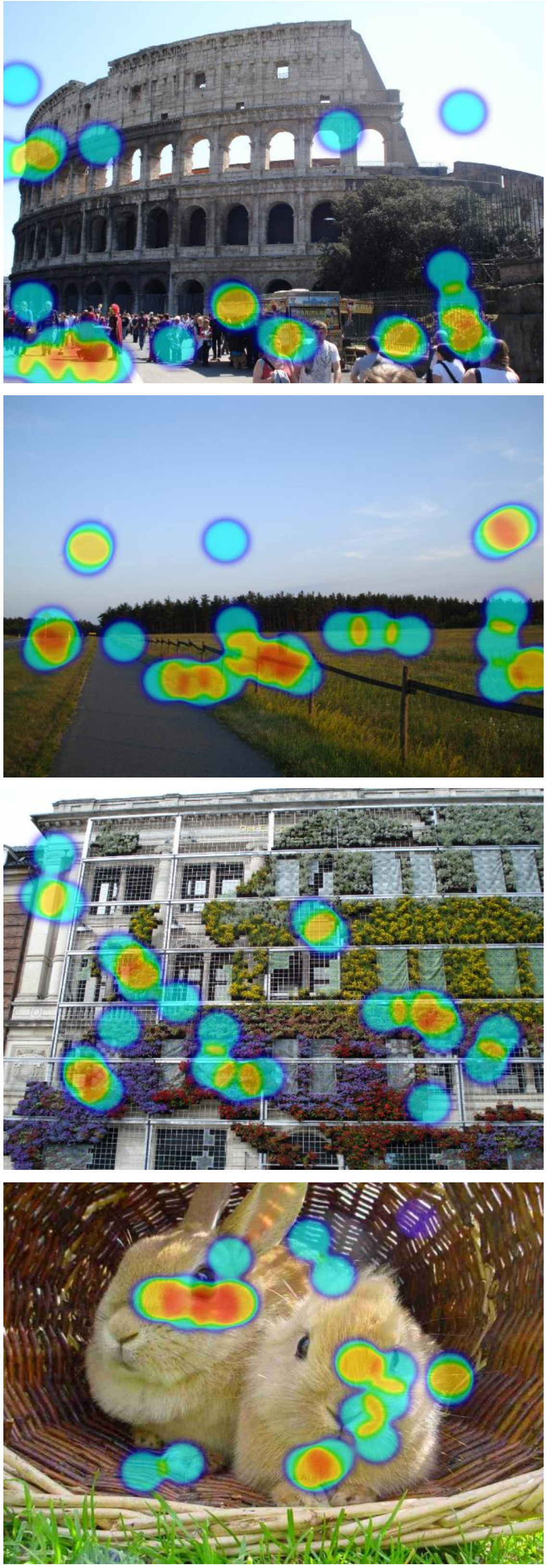

Figure 11- From top: Heatmaps of Rome, VY, Europe, and Rabbit images

recorded fixation points. Additionally, the function that I found for calculating ROC required both the positions and the heat map together, making it impossible to use those methods for comparison.

Therefore, to compare the results, I relied on the heat maps generated by the FOA model, the online gaze recorder service, and the subject data files. While the FOA model provided predictions of the main FOA points, the gathered data from myself and the subjects indicated that top-down attention also played a role in determining the salient regions of the images. As a result, there were some differences between the predictions of the FOA model and the recorded fixation points of the human subjects.

Overall, by comparing the heat maps generated by the different methods, I was able to gain insights into the salient regions of the images and the factors that influenced human attention.

Discussion:

In this report, I utilized various methods to analyze the salient regions of four images and gain insights into the factors that influence human attention.

Firstly, I used a FOA (Fixation Over Areas) model to predict the points of fixation on the images and generate saliency heat maps. The model's deterministic nature ensured consistency in its predictions, allowing for reliable analysis of the salient regions of the images. However, the model's predictions were limited to the main FOA points and did not account for the influence of top-down attention.

To address this limitation, I utilized an online gaze recorder service to record data on the fixation points of human subjects while viewing the images. The recorded fixation points were used to generate a heat map of the salient regions of the images. The resulting heat maps provided insights into the salient regions of the images and the factors that influenced human attention, including top-down attention.

Furthermore, to compare the results obtained from the FOA model, the gaze recorder service, and the subject data files, I relied on the heat maps generated by each method. While the FOA model provided predictions of the main FOA points, the gathered data from myself and the human subjects indicated that top-down attention also played a role in determining the salient regions of the images. As a result, there were some differences between the predictions of the FOA model and the recorded fixation points of the human subjects.

In conclusion, the methods utilized in this report provided valuable insights into the salient regions of the images and the factors that influence human attention. The results highlight the importance of considering both bottom-up and top-down attention in understanding how humans perceive and process visual information.

Reference:

- A Model of Saliency-Based Visual Attention for Rapid Scene Analysis
  Laurent Itti, Christof Koch, and Ernst Niebur