A Review on "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis"

Zahra Kavian[a]

[a]*Sharif University of Technology, Tehran,*

Abstract

The visual attention system could model with a saliency map, a kind of topography map that is a combination of multi-scale image features. This article introduces the Itti attention model and shows how the salience value of stimulus guide the location of the visual bottom-up selective attention. We see the correlation between the salience value and the fixation location is almost more than the chance level.

Keywords: Visual attention, scene analysis, feature extraction, visual search

1. Introduction

The primate visual system receives a enormous information each moment, and it is capable of responding fast as opposed to the limited speed of the neural hardware. A well-known hypothesis is the visual system only processes the amount of information precisely, and the other remains unprocessed. This selection is based on the particular region of the visual field. It is known as a "focus of attention" which scans the scene in a rapid, bottom-up, saliency-driven.

The model in this article is opposed by Koch and Ullman and is based on the "feature integration theory" which describes the human visual search strategy. This theory says the visual network input is decomposed into some feature maps. In other words, some low-level features are extracted in the early stage and then combined later. Different regions compete which other to be chosen as a saliency in each map. In the end, this map is combined and generates the saliency map 1.

First, we describe this model precisely and show the output for some stimuli. Second, we compare the saliency map of the model with the focus location points recorded by the eye tracker.

2. Model and Saliency Map

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

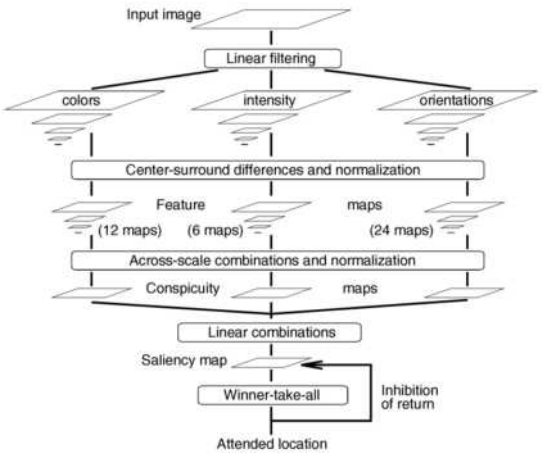

Figure 1: General architecture of the model.

The input image is low-filtered and subsampled based on dyadic Gaussian pyramids, and the horizontal and vertical images are reduced to eight octaves. Three low-level features (color, intensity, and orientation) are extracted by linear "center-surround" operations akin to visual receptive fields. The center is a pixel at three : $c \in \{2, 3, 4\}$ and the co-responded surround is defined as: $s = c+\theta$, with $\theta \in \{3, 4\}$ .The difference between each center and the surround is considered.

How do these features extract? Consider r, g, and b are the red, green, and blue channels of the input image:

- Intensity Map :

  intensity contrast, due to on-center and off-center neurons.

$$I = \frac{(r + g + b)}{3}$$
$$I(c, s) = |I(c) \ominus I(s)|$$

- Color Maps :

At first, four broadly-tuned color channels are created as:

$$R = r - (\frac{g+b}{2})$$
$$G = g - (\frac{r+b}{2})$$
$$B = b - (\frac{g+r}{2})$$
$$Y = (\frac{g+r}{2}) - (\frac{|g-r|}{2}) - b$$

Second, the "color double-opponent" system, a neuron in which the center of its receptive field inhibits with color and excites with another, is presented by two color maps as follows:

$$RG = |R(c) - G(c)| \ominus |R(s) - G(s)|$$

- Local Orientation Map

  This feature map encodes due to orientation-selective neurons in the primary visual cortex. " The local orientation information is obtained from I using oriented Gabor pyramids $O(s, q)$, where $s \in [0 \dots 8]$ represents the scale and $q \in \{0°, 45°, 90°, 135°\}$ is the preferred orientation:

$$O(c, s, \theta) = |O(c, \theta) - O(s, \theta)|$$

For each center and surround scale, the model finds these channels and linearly combines the normalized maps, and at the end find the saliency map (s) as follows:

$$\bar{I} = \oplus_{c=2}^{4} \oplus_{s=c+3}^{c=4} N(I(c, s))$$
$$\bar{C} = \oplus_{c=2}^{4} \oplus_{s=c+3}^{c=4} [N(RG(c, s) - N(BY(c, s))]$$
$$\bar{O} = \sum_{\theta \in \{0,45,90,135\}} N(\oplus_{c=2}^{4} \oplus_{s=c+3}^{c=4} N(O(c, s, \theta)))$$
$$S = \frac{1}{3}(N(\bar{I}) + N(\bar{C}) + N(\bar{O}))$$

3. Results

3.1. Salience Map of Stimulus

In this part, I have four stimulus images and want to obtain a saliency map of the Itti Model (Figure 2, and 7 to 10). For one stimulus, I show each stage of the model (Figure 2 to 6).

3.2. Eye Tracker data

In this part, the data of the eye tracker is shown on the saliency map. Both of them have an acceptable overlap (Figure 11 to 14).

	Rabbit	vy	Rome	Europe
NSS	-0.0667	0.0252	-0.0881	0.1923
AUC	0.4987	0.5053	0.4914	0.5085

Table 1: Performance model for a subject for each stimulus image.

3.3. Gazerecorde

In this part, we should use an online eye-tracker and record our gaze for the previous stimulus images. Unfortunately, I couldn't work on this site. I've used it several times, but it couldn't calibrate my system. So, I use a similar app, in which I record my focused zone. It was acceptable but not pierced. The points on the heat map are upper than their original position. You can see the result in Figure 15.

3.4. NSS & ROC

"Normalized Scanpath Saliency (NSS)" measures the correspondence between the saliency map and the grand truth. This measurement is introduced as an average normalized salience value at the fixation point. You see its function below:

```
s_map = saliency_map;
f_map = logical(fixation_map>0.5);

if size(s_map) ~= size(f_map)
    s_map = reshape(s_map,size(f_map,1),
        size(f_map,2));
end

s_map = zscore(s_map);
nss= mean(s_map(f_map));
```

Receiver operating characteristic (ROC) is another analysis to quantify the classifier's performance. To find AUC (Area Under Curve), I use the Matlab function; "perfcurve".

Now, we use these two methods to find how saliency could follow the human gaze position. You could compare these methods for each image in the table.

4. Summary and conclusions

Itti model has changed the view about human attention. It could explain how the visual system responds fast and accurately, but this model is not capable enough to describe the location of the focus points. On the one hand, up-down connections do not use in this model. On the other hand, additional factors have significant effects on human perception, such as intellectual or personality. The results also prove the previous claim.

References

Itti, Laurent and Koch, Christof and Niebur, Ernst. *LaTeX: A model of saliency-based visual attention for rapid scene analysis* 1998.

Parkhurst, Derrick and Law, Klinton and Niebur, Ernst. *LaTeX: Modeling the role of salience in the allocation of overt visual attention* 2002.

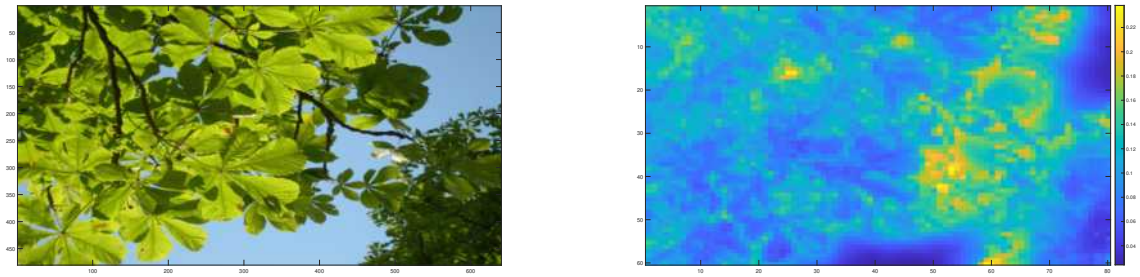

Figure 2: Right: original image. Left: salience map

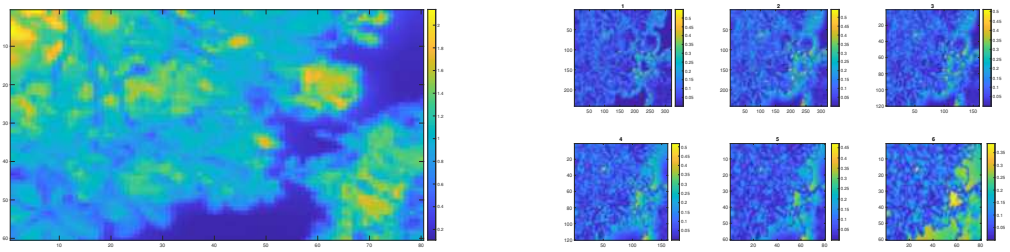

Figure 3: Right: single color channels. Left: center surround of the gaussian pyramid of the intensity

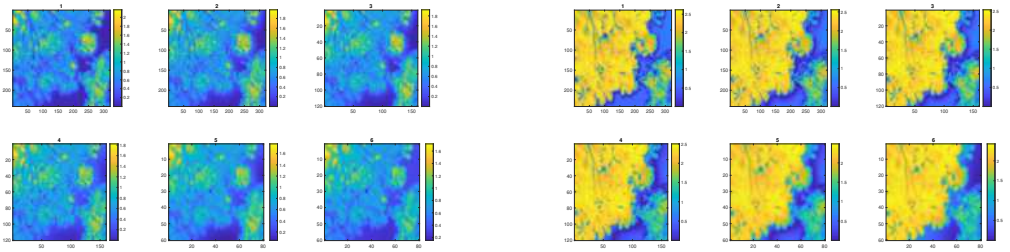

Figure 4: Right: RG center surround. Left: BY center surround

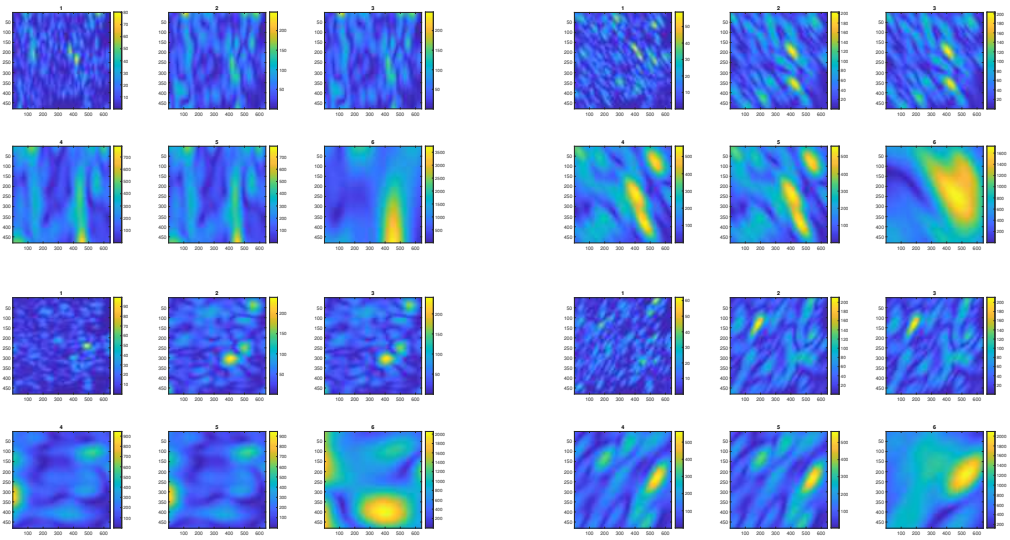

Figure 5: 1-4th direction center surround

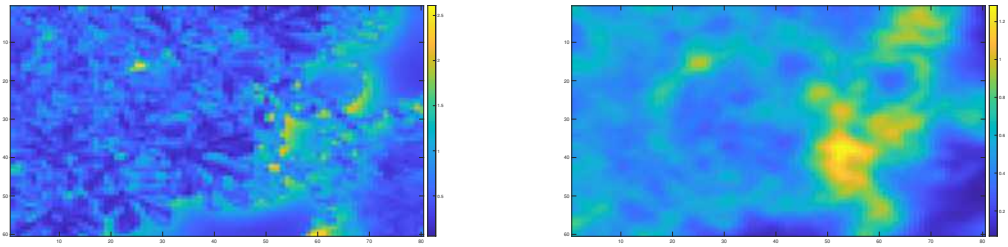

Figure 6: Right: conspiquity map of intensity. Left: conspiquity map of colors

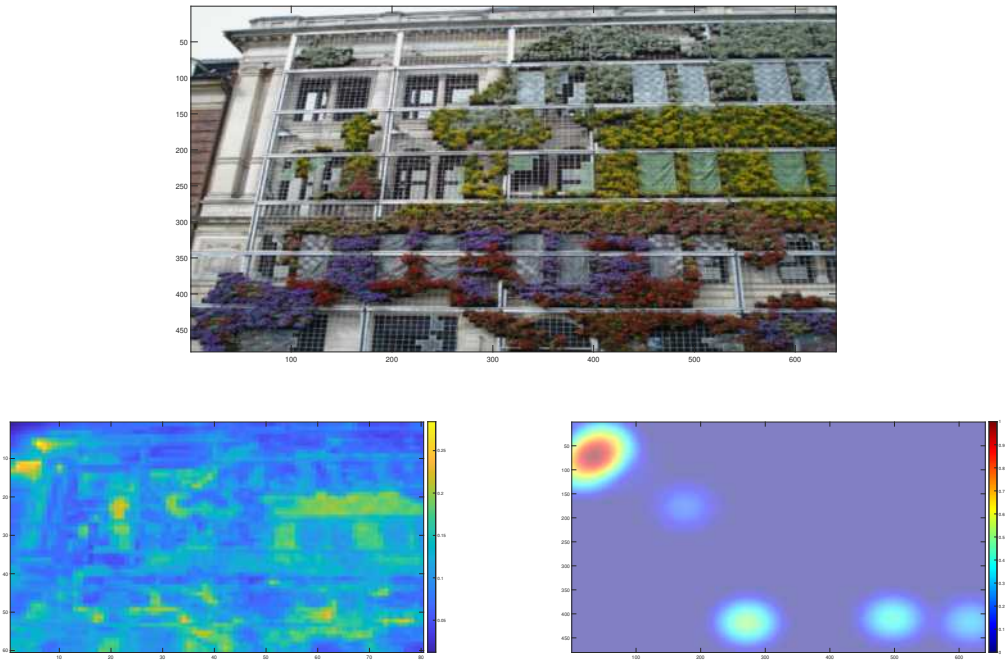

Figure 7: Top: original image. Right: Saliency Map. Left: Focus of Attention

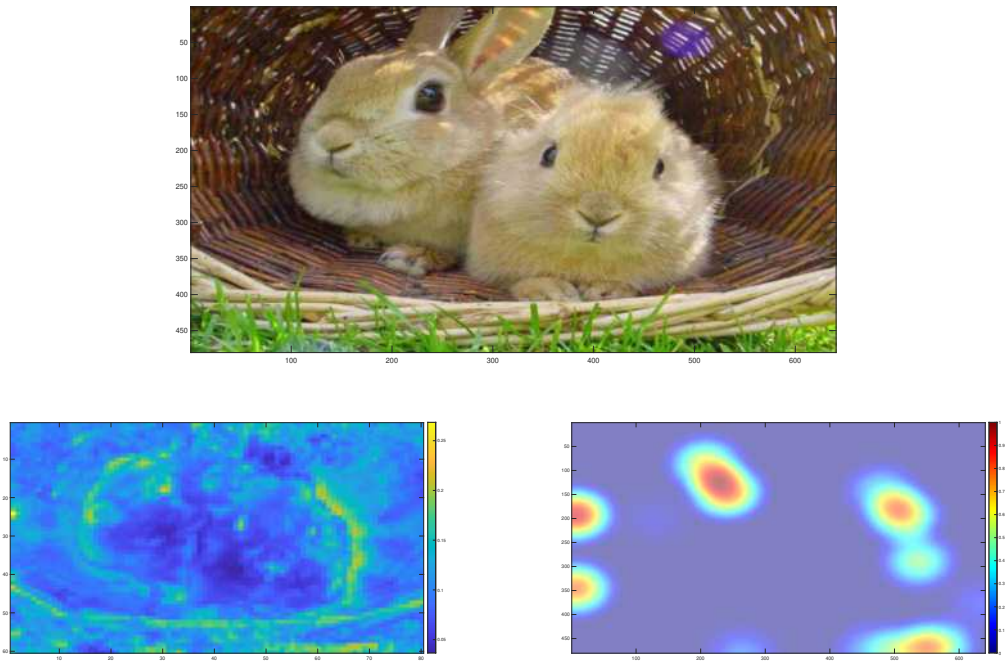

Figure 8: Top: original image. Right: Saliency Map. Left: Focus of Attention

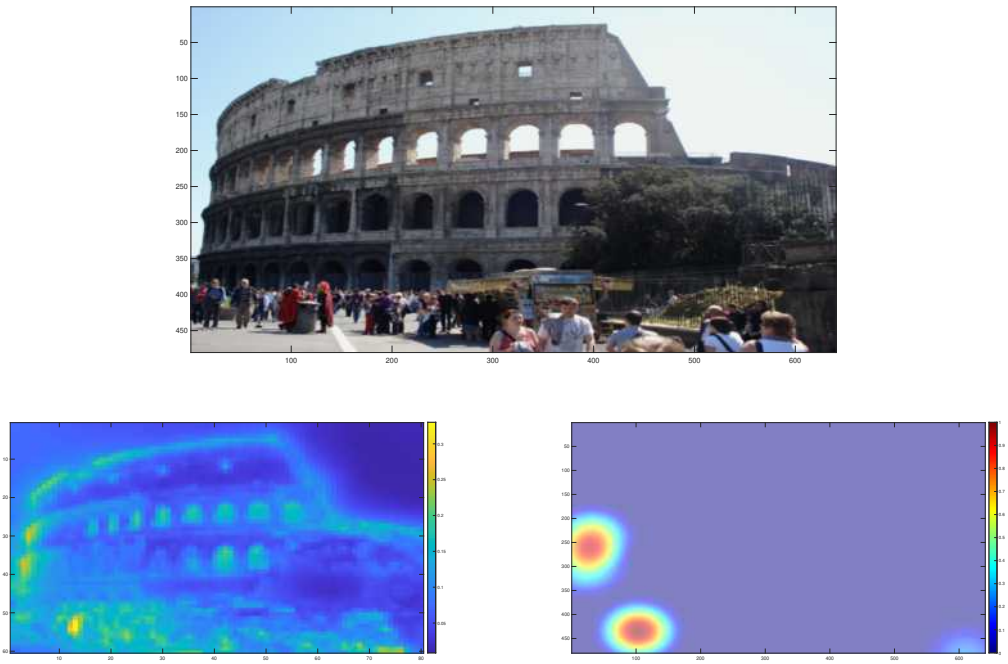

Figure 9: Top: original image. Right: Saliency Map. Left: Focus of Attention

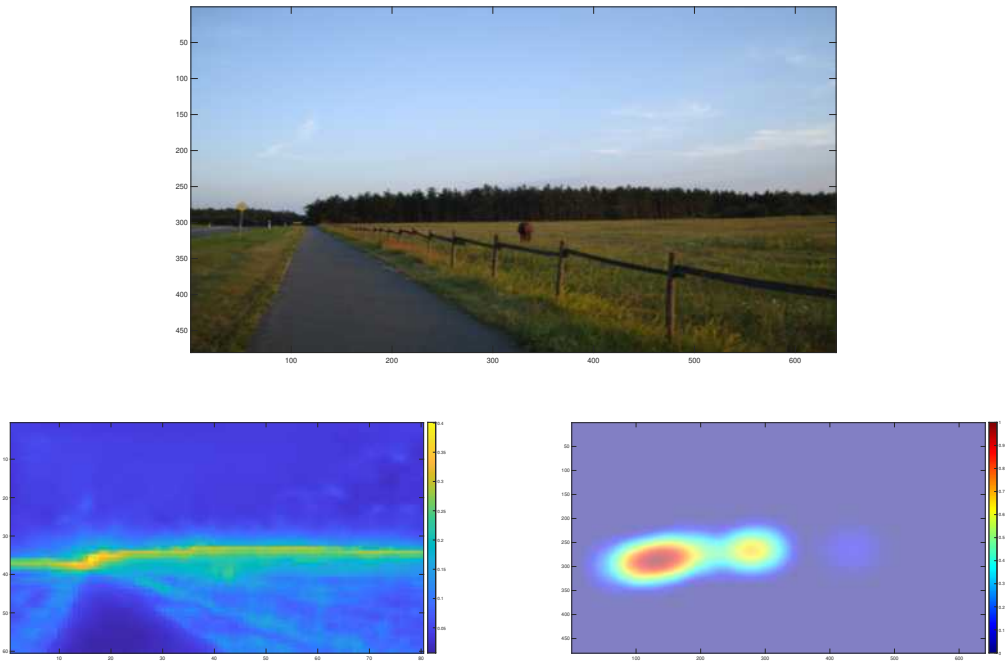

Figure 10: Top: original image. Right: Saliency Map. Left: Focus of Attention

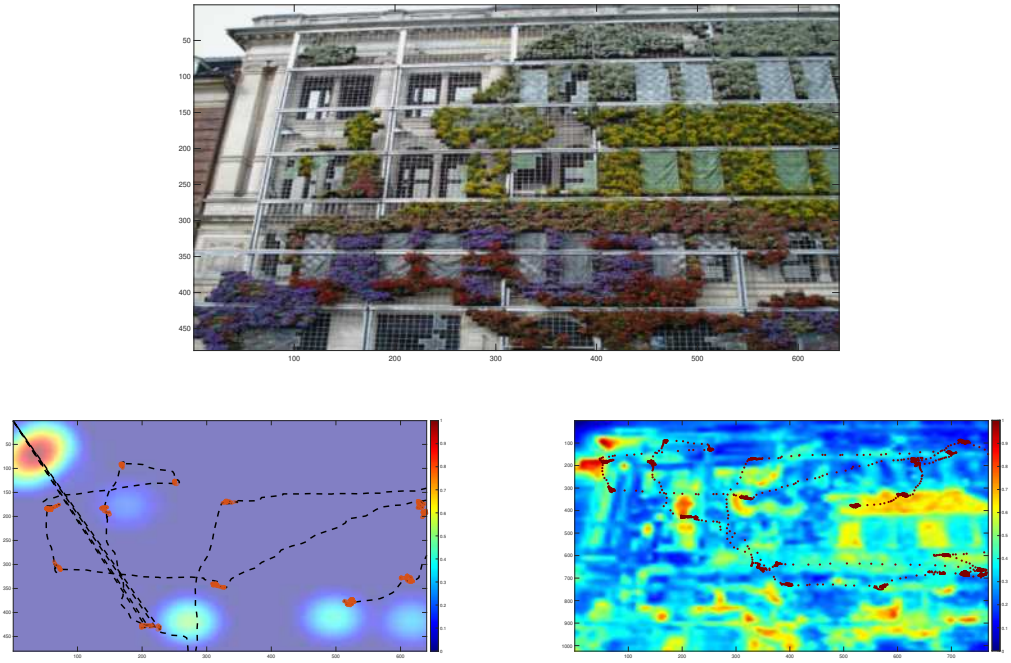

Figure 11: Right: Saliency Map. Left: Focus of Attention

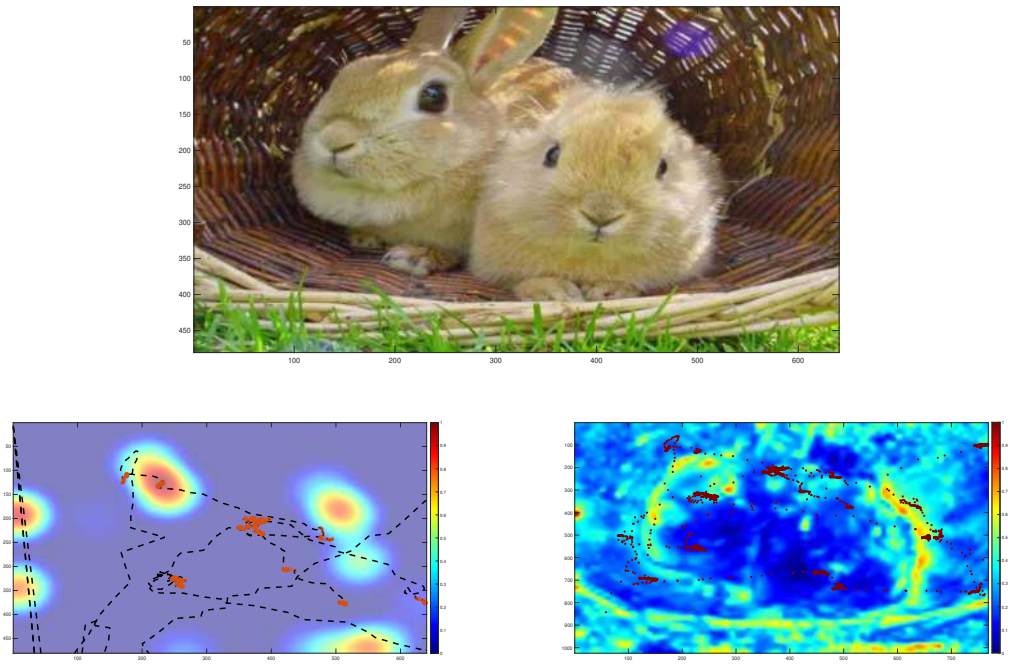

Figure 12: Right: Saliency Map. Left: Focus of Attention

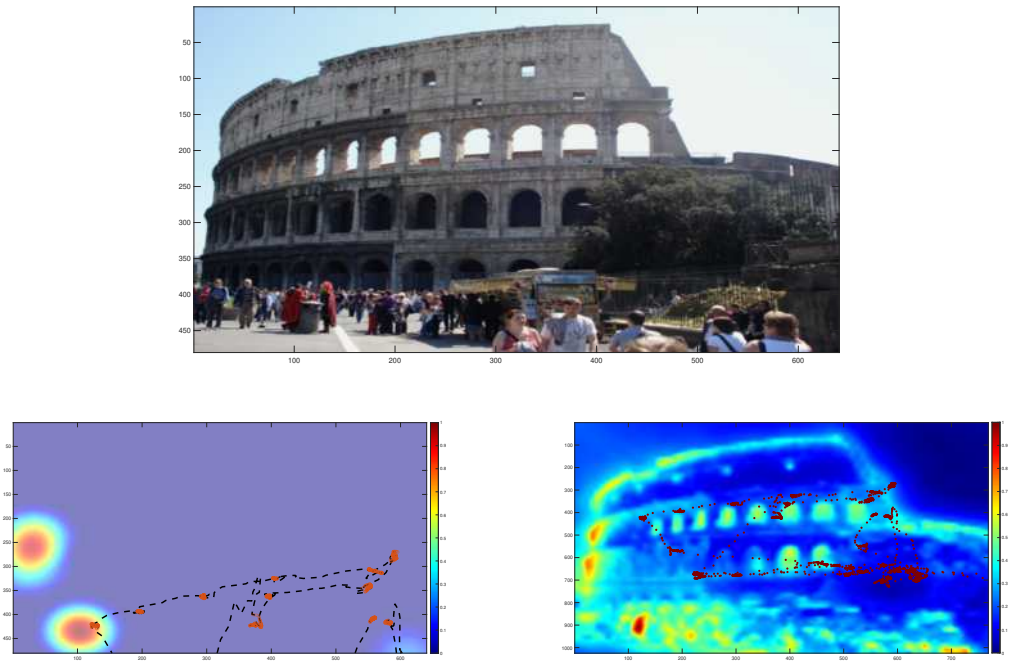

Figure 13: Right: Saliency Map. Left: Focus of Attention

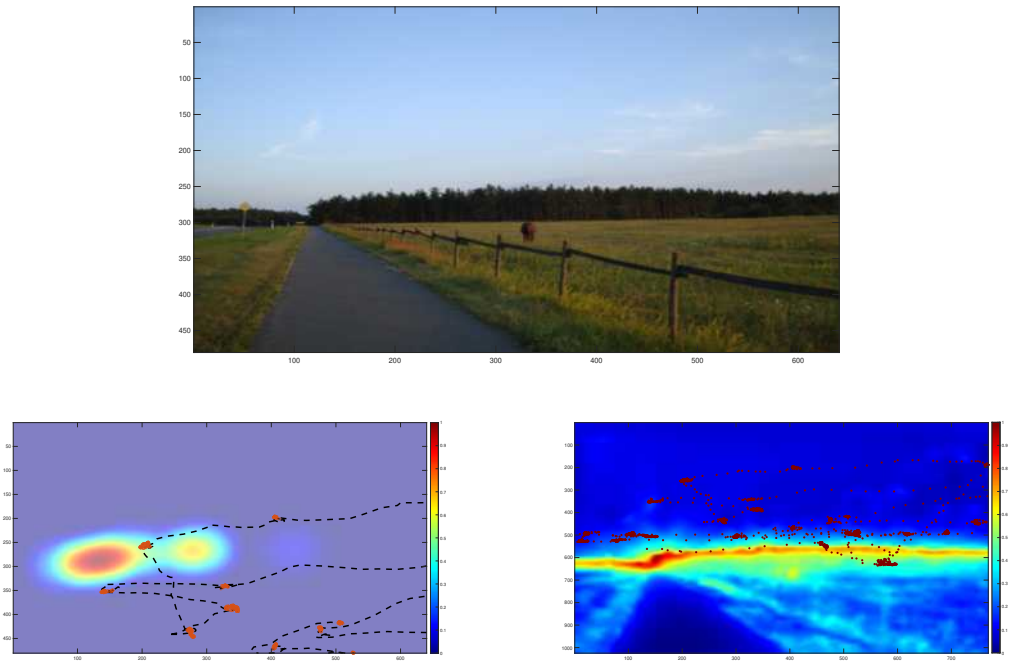

Figure 14: Right: Saliency Map. Left: Focus of Attention

5. Appendix

I run this model for some video frames that shows a carousel. We see a dynamic heatmap based on the person's location (Figure 16).

Figure 15: My Focus of Attention

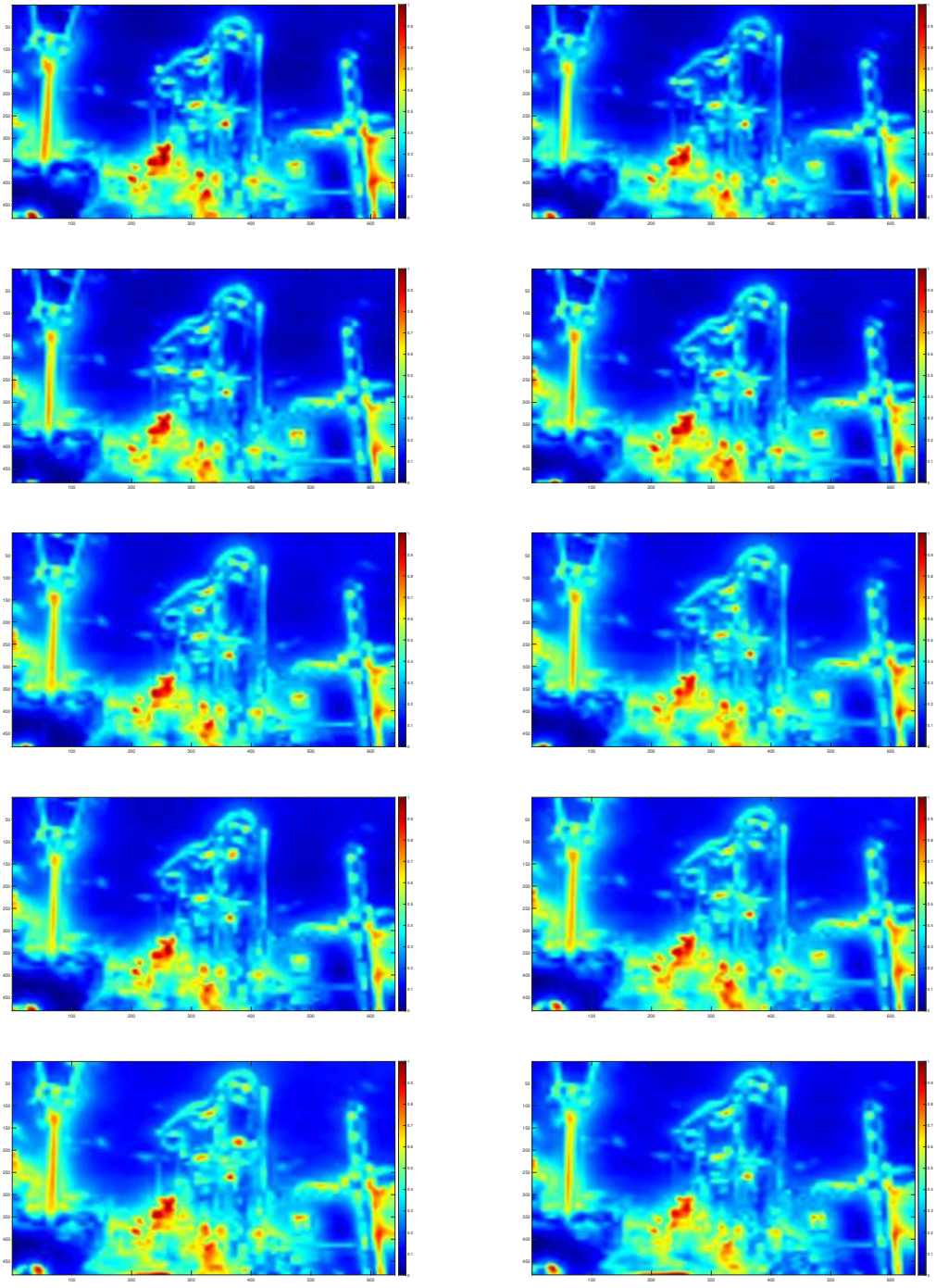

Figure 16: Saliency map of a video