CMPUT 616
Implementation of a Visual Attention Model

Nathan Funk

April 14, 2004

1 Introduction

This project implements the visual attention model described by Itti, Koch and Niebur in their 1998 paper "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis" [6]. The paper describes a model of the bottom-up driven shifting of the *focus of attention.* Much like a rapidly shifting spotlight, the focus of attention scans a scene by selectively choosing a portion of the visual field for detailed examination. This behavior is commonly observed in primates which direct their attention on small salient sections of a scene rather than scanning the entire scene in a fixed pattern. This enables interpretation of a relatively complex scene in real time [11].

The model outlined in the original paper is based on a biologically plausible architecture developed by Koch and Ullman [7]. It ignores the top-down, task-dependent influence on the focus of attention. It is also not concerned with object recognition, but rather focused on finding features of interest which may in a later stage be used for object identification.

This report provides an overview of the model, describes details of the implementation and also discusses the experiments performed with the implementation.

2 Objectives

The main objectives of this project are as follows:

- Implement the model.
- Discuss the differences between the implementation and the original model.
- Report the expected effects of these differences.
- Evaluate the results of a set of experiments.
- Compare experimental results to those presented in the paper.

3 Background

3.1 Overview

As the model overview diagram shows (Figure 1), the information flow is unidirectional except for the inhibition of return. The structure is parallel for the majority of the steps. Gaussian pyramids

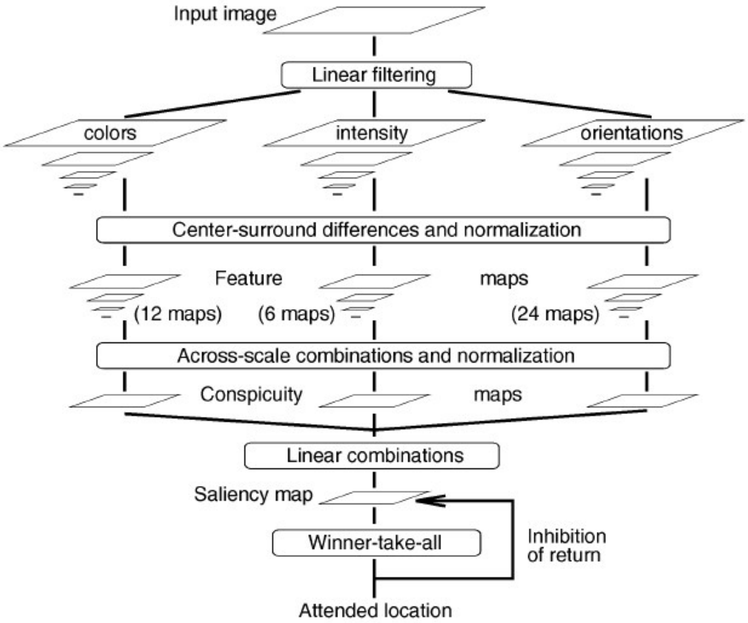

Figure 1: Overview of the model. Note the up-side-down pyramids representing multiple scale representations of the input channels and feature maps (from Itti, Koch and Niebur [6]).

are employed ranging from 1:1 (scale 0) to 1:256 (scale 8) [4] [1]. These are used for the computation of center-surround differences as outlined below.

By creating so-called *feature maps* from the image, the original color data in RGB format is decomposed into multiple channels. The three feature maps proposed in the paper are *intensity*, *color* and *orientation*, although it is also noted that other features such as motion could be added to the model. Each feature map is determined through a set of “center-surround” differences which detect spatial discontinuities for each feature. The same principle is found in the retina, lateral geniculate nucleus as well as the primary visual cortex. The center-surround differences are found by performing *across-scale subtraction* which will be denoted as “$\ominus$”. In particular, the difference between a pixel at scale $c \in \{2, 3, 4\}$ (the center) and its corresponding pixel at scale $s = c + \delta$ (the surround), where $\delta \in 3, 4$ is used. This results in six different scale combinations.

The goal of the individual feature maps is to determine salient features of different types separately. These maps can then be combined into the *saliency map* which is a single-scale image showing the calculated saliency at every point in the image. A simple approach for directing the focus of attention (FOA) would be to select the most active locations in the saliency map. A more neuronally plausible implementation is however provided.

The saliency map feeds into a 2D layer of *leaky integrate-and-fire* neurons which again feed into a second layer of neurons. This second layer is modeled as a “winner-take-all” (WTA) network. The winner of WTA layer directs the FOA and also causes local inhibition on the saliency map layer through the *inhibition of return*.

3.2 Feature Maps

3.2.1 Intensity

From the original image with red, green and blue color channels r, g, b respectively, the intensity I is calculated as $I = (r + g + b)/3$. As noted in Itti's PhD thesis [5], this assumption is a very rough approximation in comparison to typical weighted intensity calculation as described by Foley *et al.* [3].

A Gaussian pyramid $I(\sigma)$ is created with nine scales ($\sigma \in [0..8]$). This pyramid is used to determine center-surround differences similar to cells responding to a bright center and dark surround, and dark center bright surround [9]. A total set of six maps $\mathcal{I}(c, s)$ are calculated with:

$$\mathcal{I}(c, s) = |I(c) \ominus I(s)|. \tag{1}$$

3.2.2 Color

For the color feature map, the color information is first converted to a broadly-tuned channel system with four channels: Red $R = r - (g + b)/2$, green $G = g - (r + b)/2$, blue $B = b - (r + g)/2$, and yellow $Y = (r + g)/2 - |r - g|/2 - b$. A pyramid is constructed for each of these channels.

The feature map is based on the idea of chromatic opponency for red/green and blue/yellow such as in the human primary visual cortex [2]. Two maps are created from the new color channels:

$$\begin{aligned} \mathcal{RG}(c, s) &= |(R(c) - G(c)) \ominus (G(s) - R(s))|, \text{and} && (2) \\ \mathcal{BY}(c, s) &= |(B(c) - Y(c)) \ominus (Y(s) - B(s))|. && (3) \end{aligned}$$

3.2.3 Orientation

The orientation feature map is based on the intensity image I as calculated for the intensity step. Four oriented Gabor pyramids $O(\sigma, \theta)$ are constructed, where $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$. For each pyramid, the orientation contrast is determined with:

$$\mathcal{O}(c, s, \theta) = |O(c, \theta) \ominus O(s, \theta)|. \tag{4}$$

A total of 24 maps are created for the orientation (6 maps for each angle).

3.3 The Normalization Operator

For the subsequent steps, it is necessary to *normalize* various maps. The proposed normalization operator $\mathcal{N}(.)$ ensures that the different modalities resulting from the previous processing steps are made comparable. Furthermore, the operator promotes maps with few dominating peaks, and attenuates maps with many similar peaks as shown in Figure 2.

The normalization algorithm can be summarized as follows:

1. Normalize the map within a range $[0..M]$.
2. Determine $\bar{m}$, which is the average of all local maxima without the global maximum M.
3. Multiply the map by $(M - \bar{m})^2$.

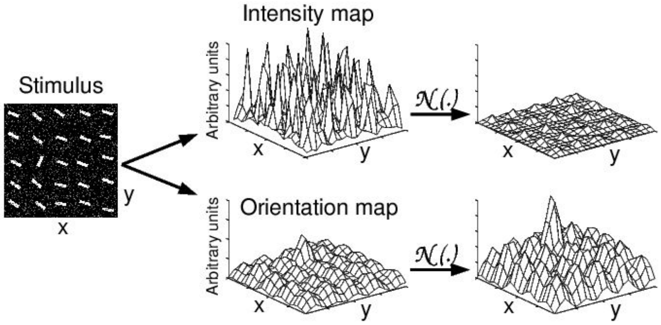

Figure 2: Example of using the normalization operator $\mathcal{N}(.)$ on two different feature maps. The noisy intensity map is attenuated, while the orientation map, featuring a single dominant peak, is promoted (from Itti, Koch and Niebur [6]).

3.4 Conspicuity Maps

Before combining the feature maps to the saliency map, they first need to be all brought into one scale. This is accomplished by normalization and across-scale addition, denoted by "$\oplus$", for each of the feature maps. The result are the three conspicuity maps $\bar{\mathcal{I}}$, $\bar{\mathcal{C}}$, and $\bar{\mathcal{O}}$:

$$\bar{\mathcal{I}} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{I}(c,s)), \tag{5}$$

$$\bar{\mathcal{C}} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} [\mathcal{N}(\mathcal{RG}(c,s)) + \mathcal{N}(\mathcal{BY}(c,s))] \tag{6}$$

$$\bar{\mathcal{O}} = \sum_{\theta\in\{0^\circ,45^\circ,90^\circ,135^\circ\}} \mathcal{N}(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{O}(c,s,\theta))), \tag{7}$$

3.5 Saliency Map

Finally, the final input $\mathcal{S}$ into the saliency map can be determined by averaging the three normalized conspicuity maps:

$$\mathcal{S} = \frac{1}{3}\left(\mathcal{N}(\bar{\mathcal{I}}) + \mathcal{N}(\bar{\mathcal{C}}) + \mathcal{N}(\bar{\mathcal{O}})\right). \tag{8}$$

The saliency map (SM) is the first neural layer as shown in Figure 3. It is excited by $\mathcal{S}$ and simultaneously inhibited by the inhibition of return (IOR). The winner-takes-all layer (WTA) is also modeled with leaky integrate and fire neurons. It should be noted that the SM neurons however never reach their firing threshold. When a WTA neuron reaches it's threshold potential, three mechanisms are activated:

- The FOA is shifted to the position of the winning neuron.
- All neurons in the WTA are reset to zero.

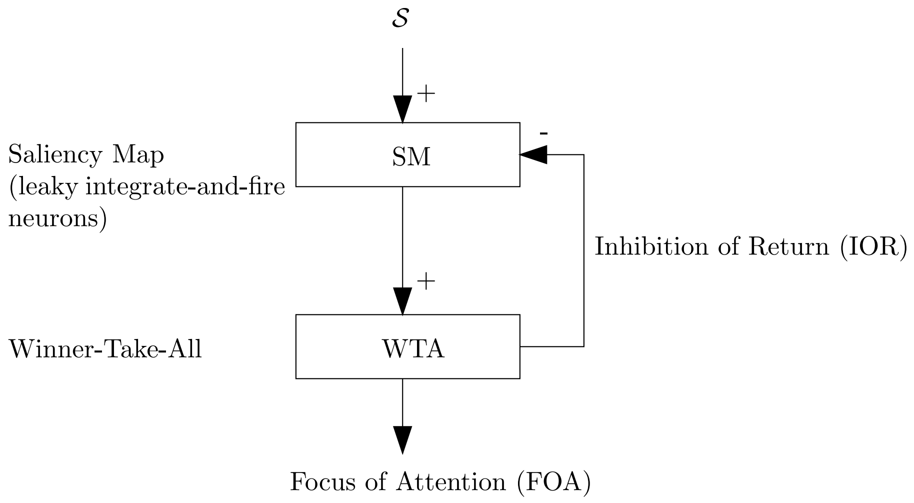

Figure 3: Detailed view of the neural layers.

- IOR is added to the local area of the winning neuron.

Adding IOR to the saliency map effectively prevents the FOA from immediately returning to a recently attended location. Although not explicitly stated in the paper, Itti's PhD thesis [5] mentions that a spacial difference of Gaussians (DoG) can be employed for the IOR. The excitatory lobes around the inhibitory center act as a *proximity preference*. This gives a slight preference to salient features that are nearer to the previously attended locations.

The radius of the FOA is set to be one sixth of the smaller of the image width or height. The IOR is modeled as a transient inhibition with a time of approximately 500-900 ms, which has been psychophysically observed in [10]. The parameters of the neural layers (time constants, conductances, and firing thresholds) were chosen to achieve FOA jumps every 30 to 70 ms in simulated time.

4 Implementation details

The entire implementation is performed as a set of MATLAB scripts. The generation of $\mathcal{S}$ is accomplished by calling `makeS(imageFile)`. This stores $\mathcal{S}$ in a separate file. By calling `timeSim(imageFile, dt, tMax)` the time simulation of the neural layers is started with a time step of `dt` and a total simulated time of `tMax`.

The majority of the implementation such as the construction of the feature maps and conspicuity maps is the same as described in Itti, Koch and Niebur's paper. The assumptions made in this implementation, as well as differences from the original implementation and their expected effects are discussed here.

4.1 Pyramid Construction

Initially, the pyramid construction was planned to be compared with using multiple Gaussian filters with increasing standard deviation. The pyramid construction is essentially repeated low pass filtering (LPF) with a lower frequency cut-off at each scale. Since the scaled images need to be up-sampled before performing across-scale subtraction and addition, they could instead all be kept at a single resolution yet filtered at increasing Gaussian standard deviations. Furthermore, since the up-sampling of course scaled images requires an interpolation method (typically bilinear), the results from a single LPF pass without scaling would provide a more accurate representation of the low-frequency components in the image.

It was soon discovered that there is a major disadvantage associated with this approach. For the smallest scales (e.g. 7 and 8), the equivalent Gaussian convolution filter size is very large (over 100x100). The filtering then becomes very computationally expensive, resulting in a significant slow-down of the model. Since experiments were expected to take on the order of 10 minutes to complete, this approach was not fully implemented. The quality improvement would likely be minimal, since the linear interpolation of low frequency images does not introduce much error.

If the initially planned approach was ever implemented, using a *separable* filter would be expected to improve the performance. However this might introduce similar errors to those caused by bilinear interpolation.

The method employed in the final implementation uses the `imresize()` function from the MATLAB Imaging Toolbox. It can perform a LPF with variable size before resizing the image. The size of this filter was noted to have an effect on the saliency map and is therefor examined further in the experiment section. The paper cited by Itti, Koch and Niebur uses a 5x5 separable filter [4]. For this reason a 5x5 filter was used in for all the experiments unless otherwise noted. The interpolation method for both up and down-sampling was chosen to be bilinear.

A slight error was made in the implementation, resulting in all scales being shifted by one index. The scale indexing in the paper starts at 0 for the 1:1 image. In this implementation, scale 1 is used for the 1:1 image. This causes all feature map calculations to be performed at double the resolution as in the paper. However, this is equivalent to starting with a higher resolution image and has no qualitative effect on the results.

4.2 Neuron Layers

The description of the neural layers in the paper does not include all the necessary details and parameter values necessary for an implementation. For this reason, all the assumptions made are summarized here.

Both the SM layer and the WTA layer are implemented with leaky integrate-and-fire neurons. For these neurons, a parallel resistor-capacitor model based on Lapicques paper [8] and Tuckwell's analysis [12] is employed. The change in potential for a given input current I (not to be confused with the image intensity) is modeled with:

$$C\frac{\mathrm{d}V}{\mathrm{d}t} = -\frac{V}{R} + I, \tag{9}$$

where C is the capacitance, R is the resistance, and V is the membrane potential. This differential equation can be discretized in order to be able to determine the membrane potential $V(t+\delta t)$ after a time step of δt given a previous potential $V(t)$ and an input current $I(t)$. The resulting equation is:

$$V(t+\delta t) = \left(1 - \frac{\delta t}{CR}\right)V(t) + \frac{\delta t}{C}I(t). \tag{10}$$

Since the neurons in the SM layer never reach their threshold, and the potentials in the WTA layer are reset to 0 immediately after firing, the action potential does not need to be modeled.

The values of the parameters were adjusted to give similar results as the paper. This is justified since the parameters in the paper were also adjusted to achieve the desired output. For both the

SM and WTA layers the resistance R is 100 MΩ and the time constant $\tau = CR$ is 10 ms. The input current I into the SM layer is modeled as 5 $10^{-3}(\mathcal{S} - IOR)$. The input current into the WTA is set to the value of the SM potential. Finally, the WTA layer threshold is 20 mV as described in Itti's thesis.

Due to the discrete time model, it is possible that multiple neurons in the WTA layer reach their threshold in the same time step. This was observed in initial experiments. The effect was minimized by reducing the time step size to 1 ms.

The IOR and proximity preference were modeled similar to Itti's description in his thesis [5]. A difference of Gaussians is employed, where IOR is updated with the equation

$$IOR(t + \delta t, x, y) = IOR(t, x, y) - 5e^{\frac{dx^2+dy^2}{\left(\frac{r_{FOA}}{2}\right)^2}} + e^{\frac{dx^2+dy^2}{r_{FOA}^2}}, \tag{11}$$

where x and y are the image coordinates, dx and dy are the distance of the image coordinates from the FOA center, and r_{FOA} is the radius of the FOA. The inhibitory center is weighted with 5, making it's effect much stronger than that of the proximity preference (the last term in the equation). Also note that the standard deviations are chosen as $\frac{r_{FOA}}{2}$ for the center and r_{FOA} for the surround.

The decay of the IOR is modeled by a 1% attenuation at every time step. For a 1 ms time step this simulates a decay to 1% of the original value within 500 ms, similar to the targeted inhibition time frame of 500-900 ms in the original paper.

5 Experiments

5.1 Basic operation

A number of experiments were performed to ensure the proper performance of the model. The following test was targeted at the orientation filter. It is similar to the example image given for the normalization operator in the paper (see Figure 4).

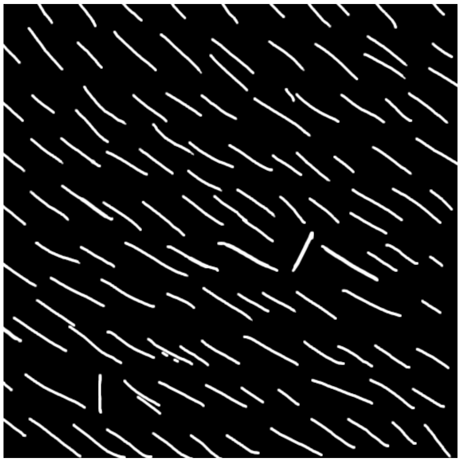

Figure 4: The orientation test image `o-test.png`.

Figure 5 shows the resulting feature maps, which appear as expected. Figures 6, 7, and 8 show the $\mathcal{S}$ map, SM neuron layers, and final FOA path for this test image respectively.

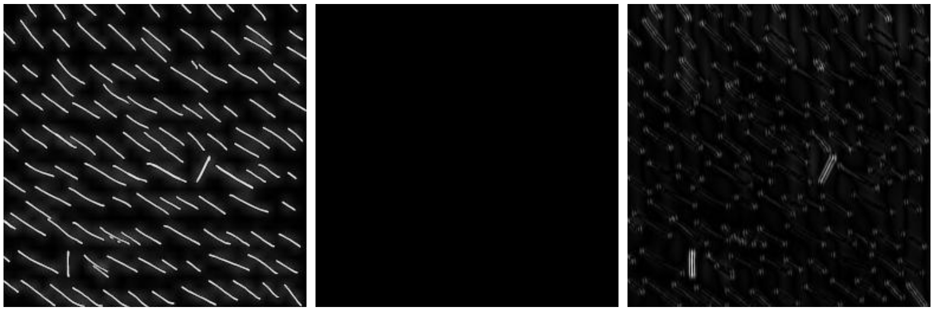

Figure 5: The resulting intensity, red-green, and orientation feature maps respectively. The angle of the orientation map is 90°. Since the image is gray-scale the response in the red-green feature map is 0. The orientation map correctly highlights the vertical line in the bottom left.

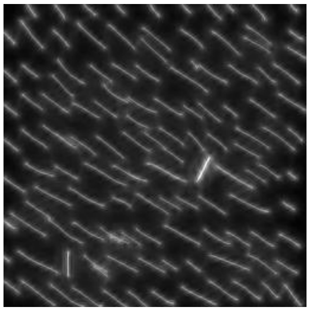

Figure 6: The $\mathcal{S}$ map resulting from the normalization and averaging of the conspicuity maps.

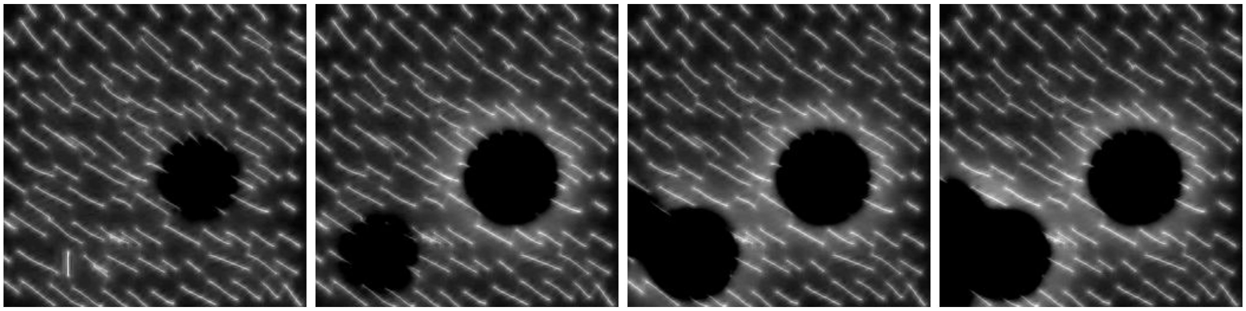

Figure 7: The SM neuron layer at 30 ms, 52 ms, 70 ms, and 86 ms respectively. The images clearly show the proximity preference around each inhibitory region. Although the proximity preference does not appear to affect the second position of the FOA, it has a strong effect on the subsequent locations since no other strong salient regions are present.

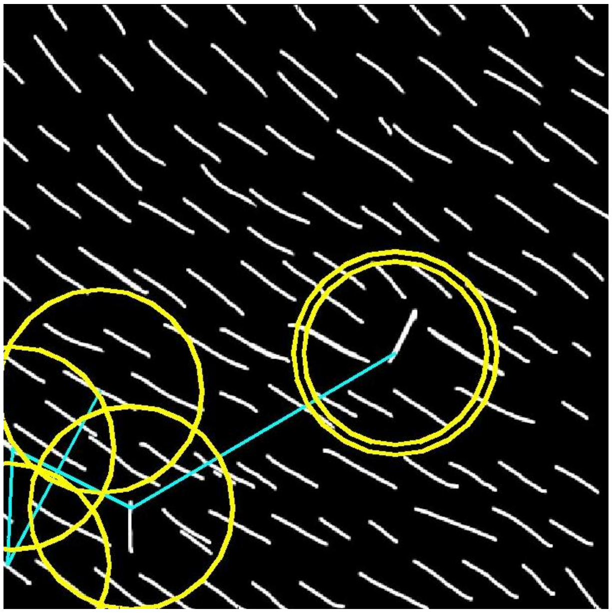

Figure 8: The resulting path of the FOA. The first FOA position is marked by the doubled circle. The circles outline the region of the FOA. This result was achieved with a time step of 2 ms.

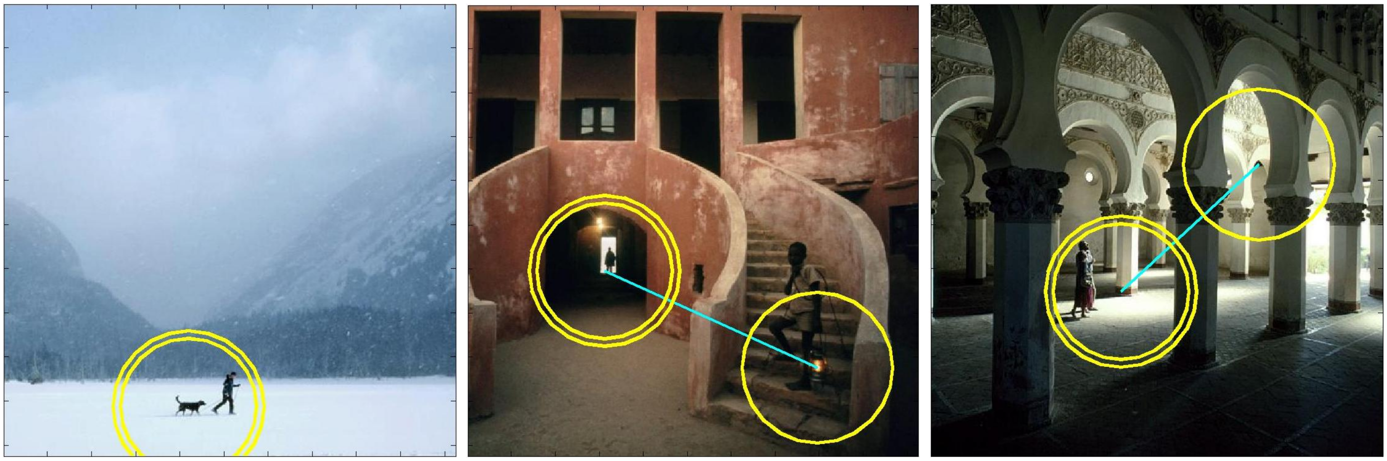

Figure 9: Example results on a set of images from `nationalgeographic.com`. All images were generated with a time step of 1ms, and restricted to a simulated time of 100ms.

5.2 Study of filtering effects

As mentioned in the implementation section, it was noticed during the implementation that the LPF filtering of the images prior to down-sampling in the pyramid construction appeared to affect the final results. In particular, it was noted that when applying a strong LPF, the FOA would be attracted to the edges of the image. After experimenting with different filter strengths, the cause was determined.

When filtering the image, a zero-padding around the edges is necessary for the convolution. This padding must be at least half the convolution kernel size. For large Gaussian convolution filters, the zero-padding causes a significant darkening of bright images around the edges. When the across-scale differences are calculated, these dark edges result in a brightening of the feature maps at the edges, and hence also affect the saliency map. This *false saliency* around the image edges finally causes the FOA to be attracted to the edge regions.

As Figure 10 shows, increasing the filter strength affects the edge and corner regions of the $\mathcal{S}$ map.

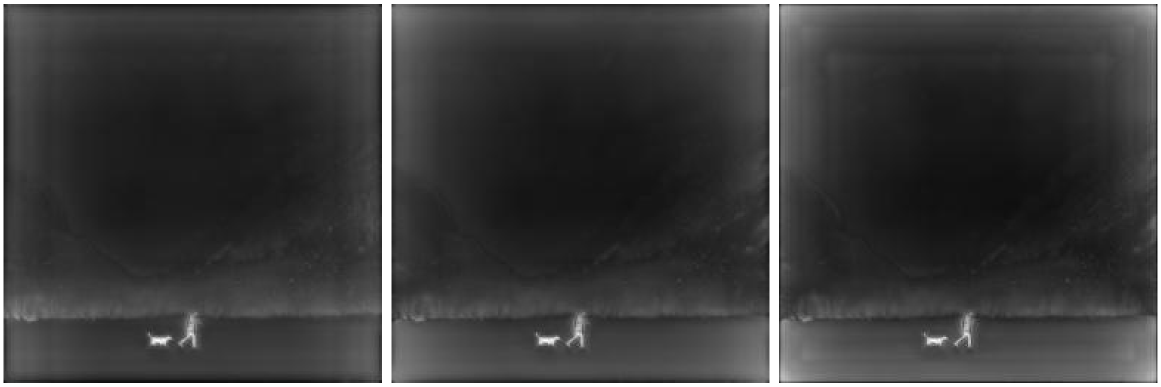

Figure 10: The effect on $\mathcal{S}$ of increasing the filter size from no filter, to 3x3, to 15x15 respectively. Note the false saliency caused around the edges and particularly in the corners.

5.3 Images showing poor results

Although the performance on most images was good, the path of the FOA on some images did not seem appropriate as Figure 11 shows.

The FOA locations are clustered rather than being spread out across the image. The likely cause for this effect is that the proximity preference is too strong. Reducing the weight in the IOR update (equation 11) from 1 to about 1/2 should produce better results. Unfortunately, due to time constraints, a full examination was not possible.

The third image in Figure 11 shows that the saliency due to the color feature maps (in this case red-green) is not sufficient to direct attention towards the 6 in the center. The saliency map (not shown) does not highlight the area of the 6. In fact, the area around the 6 has higher saliency than the 6 itself. This is likely due to the circle-texture.

6 Conclusions

The implementation was successful in the sense that the model was completely implemented and shows good results. Still, the evaluation of the results is quite subjective as also pointed out in the

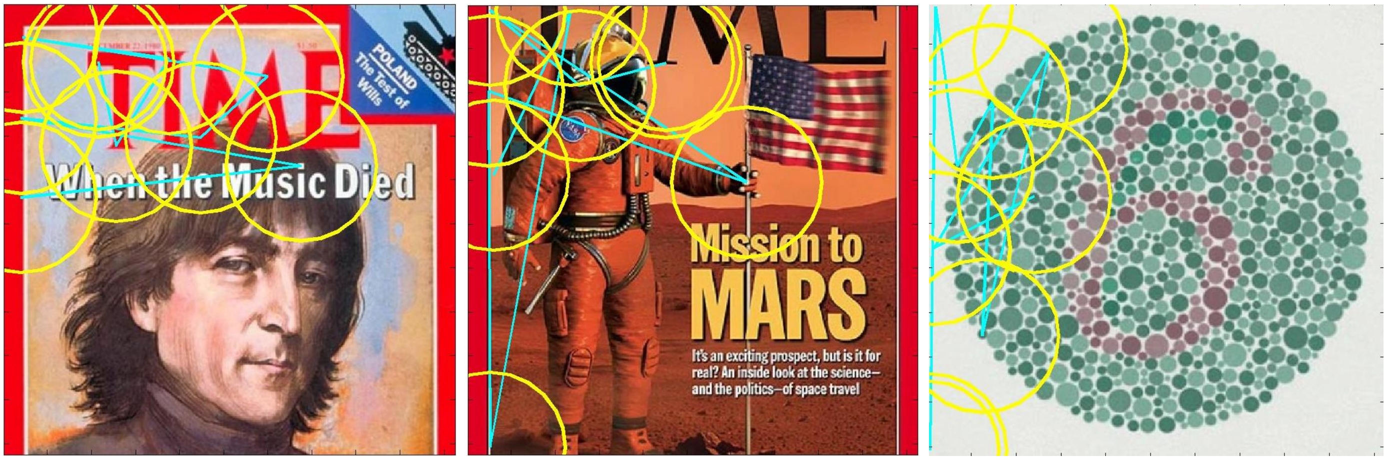
TIME
When the Music Died
POLAND
The Test of Wills
Mission to
MARS
It's an exciting prospect, but is it for real? An inside look at the science— and the politics—of space travel

Figure 11: Example of poor performance results. In all images, the FOA locations are strongly clustered in a restricted area, rather than being spread out over the image. By reducing the strength of the proximity preference, these results could probably be avoided. The first two images are from `time.com`, the third is a Ishihara red-green test image from `http://www.uni-mannheim.de/fakul/psycho/irtel/color/ishihara.html`.

original paper.

A number of observations were made during the implementation and are summarized here:

- Keeping the image at a single resolution and increasing the standard deviation of a Gaussian filter for each "scale" is expected to produce similar, if not better results. However the computational cost of large convolution filters makes it impractical.

- Using a leaky integrate-and-fire model for both the SM and WTA neural layers requires adjusting many parameters to achieve the desired results. The experiments showed the performance strongly depends on the proper configuration of these parameters.

- Using a difference of Gaussians to model the inhibition of return and proximity preference together is useful. However, the parameters such as the weight and standard deviation of each Gaussian need to be manually adjusted to achieve good results. Neither the paper nor Itti's thesis suggest a better way of obtaining these parameters.

- Since no ground truth measurements were available for comparison, the evaluation of the model is subjective.

References

[1] P. J. Burt and E. H. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, COM-31(4):532–540, 1983.

[2] S. Engel, X. Zhang, and B. Wandell. Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature*, 388(6,637):68–71, July 1997.

[3] J. D. Foley, A. van Dam, S. Feiner, and J. Hughes. *Computer Graphics, Principles and Practice (2nd ed.)*. Addison-Wesley, New York, NY, 1990.

[4] H. Greenspan, S. Belongie, R. Goodman, P. Perona, S. Rakshit, and C. H. Anderson. Overcomplete steerable pyramid filters and rotation invariance. In *Proc. IEEE Computer Vision and Pattern Recognition*, pages 222–228, June 1994.

[5] L. Itti. *Models of Top-Down and Bottom-Up Visual Attention.* PhD thesis, California Institute of Technology, Pasadena, California, 2000.

[6] 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.

[7] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.

[8] L. Lapicque. Recherches quantitatives sur l'excitation électrique des nerfs traitée comme une polarisation. *J. Physiol. Pathol. Gen.*, 9, 1907.

[9] A. G. Leventhal. *The Neural Basis of Visual Function: Vision and Visual Dysfunction*, volume 4. CRC Press, Boca Raton, Fla., 1991.

[10] M. I. Posner and Y. Cohen. Components of visual orienting. In H. Bouma and D. G. Bouwhuis, editors, *Attention and Performance*, volume 10, pages 531–556. Erlbaum, Hilldale, N.J., 1984.

[11] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. H. Lai, N. Davis, and F. Nuflo. Modelling visual attention via selective tuning. *Artificial Intelligence*, 78(1-2):507–545, October 1995.

[12] H. Tuckwell. *Introduction to Theoretical Neurobiology*, volume 1. Cambridge University Press, Cambridge, UK.