# A MULTI-RESOLUTION SALIENCY FRAMEWORK TO DRIVE FOVEATION

Siddharth Advani, John Sustersic, Kevin Irick and Vijaykrishnan Narayanan

The Pennsylvania State University, University Park, PA, USA

## ABSTRACT

The Human Visual System (HVS) exhibits multi-resolution characteristics, where the fovea is at the highest resolution while the resolution tapers off towards the periphery. Given enough activity at the periphery, the HVS is then capable to foveate to the next region of interest (ROI), to attend to it at full resolution. Saliency models in the past have focused on identifying features that can be used in a bottom-up manner to generate conspicuity maps, which are then combined together to provide regions of fixated interest. However, these models neglect to take into consideration the foveal relation of an object of interest. The model proposed in this work aims to compute saliency as a function of distance from a given fixation point, using a multi-resolution framework. Apart from computational benefits, significant motivation can be found from this work in areas such as visual search, robotics, communications etc.

***Index Terms***— Foveation, Saliency, Multi-resolution

## 1. INTRODUCTION

Visual search has become an important prerogative for new-age computer systems. Tasks such as recognizing scenes in a photograph, identifying models of vehicles, locating missing items in a retail stores, autonomously navigating through a departmental store etc. have great relevance to today's information technology (IT) demands. The Human Visual System (HVS) has been built in such a way that it becomes necessary to move the eyes in order to facilitate such tasks. Understanding the efficiency with which our eyes intelligently take in pertinent information to perform different tasks has significant impact to building the next-generations autonomous systems.

Visual attention has gained a lot of traction in computational neuroscience research over the past few years. Various computational models [1], [2], [3] have used low-level features to build information maps which are then fused together to form what is popularly called as a saliency map. Given an image to observe, this saliency map in essence provides a compact representation in terms of what is most important in the image.

There are a couple of handicaps when applying these computational models directly to the next-generation autonomous machines. These models assume that the human eye uses its full resolution across the entire field of view (FOV), which is not the case. The resolution drops off from the center of the fovea towards the periphery and the HVS is adept at foveating so as to investigate other areas in the periphery when attention is drawn in that direction. As shown in Figure 1, the distribution of cones is highly concentrated around the fovea and there is a steep fall-off in resolution beyond 10 degrees of the fovea [4]. Our eyes thus need to foveate, to allow points of interest to fall on the fovea – the region of highest resolution. It is only after this foveation process that we are capable of gathering complete information from the object of interest that drew our attention to it. It is due to this reason that humans tend to select nearby locations more frequently than distant targets and salience maps need to be computed taking this into account to improve the predictive power of the models [5], [6]. The second handicap of these models is that due to their computational complexity, it becomes extremely challenging to build real-time embedded systems that are capable of being driven by visual attention. Although significant success has been achieved by designing custom accelerators [7], [8] that take advantage of the streaming nature of the computations involved, achieving real-time frame rates for high-definition videos (approximately 2 megapixels/frame) is still a tall order. The framework proposed in this paper aims to tackle both these problems in the course of solving the bigger question of where to foveate next, when designing an autonomous robot. The rest of this paper is organized as follows. Section 2 describes the proposed multi-resolution methodology. Performance evaluation results and comparisons with the original model are presented in Section 3. A discussion on related work in this area is a part of Section 4 while further evolution of this framework and concluding remarks are a part of Section 5.

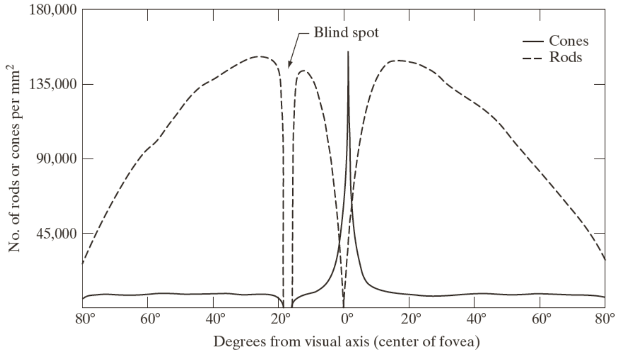


Figure 1: Distribution of rods and cones in the retina [4]

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

## 2. METHODOLOGY

We choose an information theoretic computational saliency model, AIM (Attention based on Information Maximization) as a building block for our foveation framework. AIM has been benchmarked against many other saliency models and it has proven to come significantly close to human fixations [9]. The model looks to compute visual content as a measure of surprise or rareness using Shannon's self-information theory. The algorithm is divided into three major sections. The first section involves creating a sparse representation of the image by projecting it on a set of basis learned using Independent Component Analysis (ICA). The next section involves a density estimation using a histogram back projection technique. Finally a log-likelihood is computed across the entire basis to give the final information map. For more detailed information on the algorithm and the theory behind it, one is pointed to [1], [10].

For our experiments, we use a ½" Format C-Mount Fisheye Lens having a focal length of 1.4 mm and a Field of View (FOV) of 185°. The images captured are 1920x1920 in size. The images have some inherent non-linearity as one moves away from the center, which is similar to the way the human eyes perceive the world around. In order to model the steep roll-off in resolution from fovea to periphery, we build a three level Gaussian pyramid from the original image. To do this, we first extract a 50% high-resolution center region from Level 1 as our fovea, as shown in Figure 2. After blurring and downsampling, a second region is cropped out from Level 2, representing the mid-resolution region. Another round of blurring and downsampling leaves us with the entire FOV but at a much lower resolution (Level 3). It should be noted that as the resolution drops off, the FOV is gradually increasing in our framework.

We run AIM on each of these three regions, which returns corresponding information maps. These information maps represent the salient regions at different resolutions as shown in Figure 3 (c). There are a number of ways in which to fuse these information maps to give a final multi-resolution saliency map. We believe that an adaptive weighting function on each of these maps will be a valuable parameter to tune in a dynamic environment. However for this work, which focuses on static images, we use weights of $w_1 = 1/3$, $w_{2\,=}$ 2/3 and $w_3 = 1$ for the high-resolution fovea, the mid-resolution region and the low-resolution region respectively. We use these weights since pixels in the fovea occur thrice across the pyramid while pixels in the mid-resolution region occur twice. These weights thus prevent the final saliency map from being overly center-biased. Since these maps are of different size, they are appropriately up-sampled and zero-padded before adding them up.

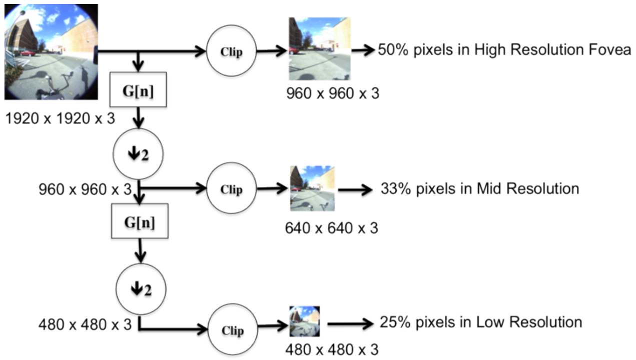
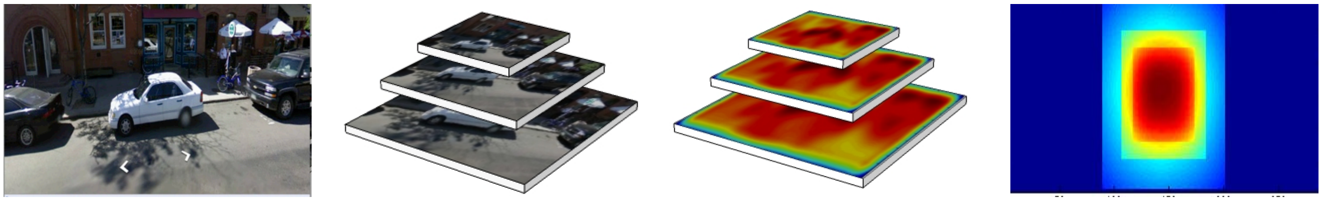


Figure 2: Image breakdown into a 3-level architecture which has a central high-resolution fovea, a mid resolution region and a low-resolution region. The lowest level covers the entire FOV

## 3. RESULTS

To validate our model (MR-AIM), we ran experiments on a series of patterns as shown in Figure 4. First we considered a series of spatially distributed red dots of same dimensions against a black background (Figures 4 (a) and 4 (b)). As can be seen in the saliency result (Figures 4 (e) and 4 (f)) there is a gradual decrease in saliency as one moves away from the fovea (Red corresponds to regions of higher saliency while Blue corresponds to regions of lower saliency). Onsets are considered to drive visual attention in a dynamic environment, so in Figure (c) we next considered the arrivals of new objects of interest within the fovea (red dot) and towards the periphery (yellow dot). Maximum response is obtained in the region around the yellow dot (Figure 4 (g)). Next we consider a movement of the yellow dot further away from the fovea (Figure 4 (d)). Again we notice a slight shift in saliency moving attention towards the center (Figure 4 (h)). These experiments give us valuable information on the dynamics of visual attention in a volatile situation.



(a) (b) (c) (d)

Figure 3: Methodology of the proposed framework from Left to Right (a) Input Image (b) Image Pyramids with increasing FOV (c) Visual Attention Saliency Maps (d) Multi-resolution Attention Map by fusing (c) with different weights

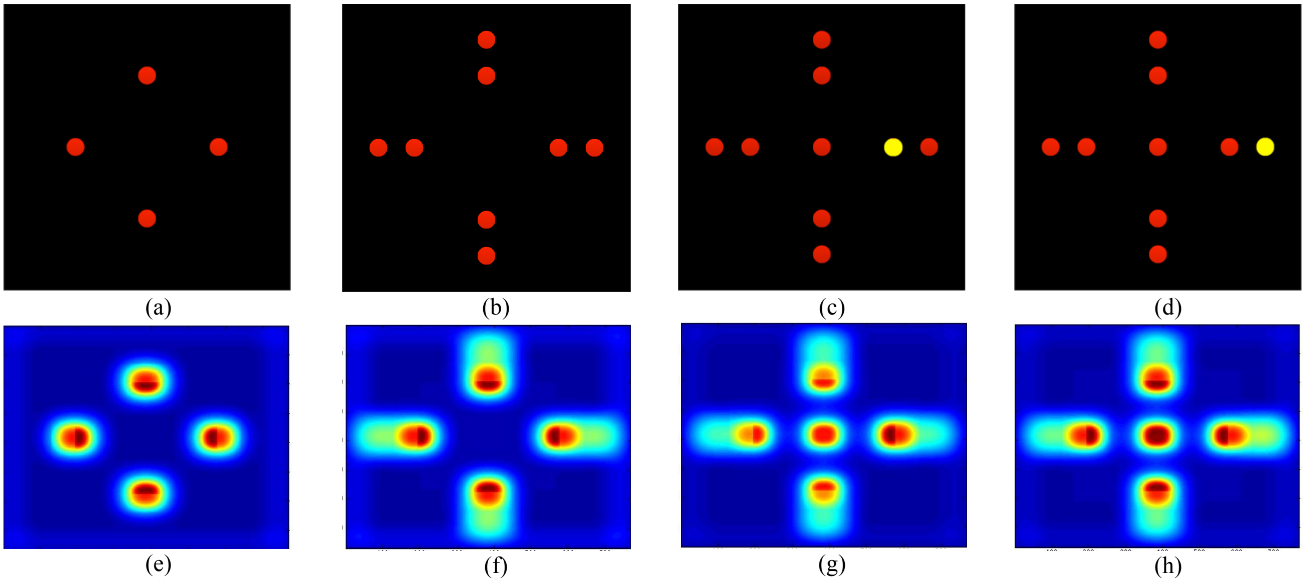


Figure 4: Saliency results with different spatial perturbations. (a)-(d): Original Images. (e)-(h): Saliency Results

Our next set of experiments was to compare the multi-resolution model with the original AIM model and evaluate the former, both, in terms of quality and performance. It should be noted here that the dataset provided in [9] has images of maximum size 1024x768, while the framework designed here is ideally targeted towards high resolution images that contain a lot of salient objects. Figure 6 (Column 1) shows an example of such an image with increasing size from top to bottom. Column 2 depicts results from the original AIM model. Column 3 shows the output of the MR-AIM. For smaller image sizes, AIM does a very good job in spotting the main ROIs. But as the image size starts to increase, it starts to pick edges as most salient. This is due to the limited size (21x21) of the basis kernels used. Increasing the size of the kernels is not a viable option for a real-time system, since that would in turn increase the computation time. MR-AIM has no such problem. Since it operates on smaller image sizes at different resolutions, it can detect objects at different scales. There is a bias towards objects in the center, but the weights do a significant job in capturing ROIs towards the periphery as well. It should be noted here that MR-AIM would not pick up objects that become extremely salient in the periphery, but with the addition of other channels of saliency, like motion, in the periphery, it will be more robust in a dynamic environment.

Another set of experiments run was on a series of video frames captured in an environment where there was sufficient activity in the periphery to activate attention as shown in Figure 7. We compare our model to other models rather than verifying against actual eye-tracking data since such data is not readily available for high-definition images. The top row shows Itti's results for frame numbers 17, 22 and 27. The middle row shows AIM's results for the respective frames while the bottom row shows MR-AIM's response. For a fair comparison we deactivated the inhibition of return in Itti's model. Both AIM and MR-AIM capture the onset of the bicyclist in frame 22 successively. These experiments offer us significant confidence about the qualitative performance of MR-AIM.

To evaluate the quantitative performance improvement we gained by this framework, we ran multiple iterations of AIM on different image sizes and compared the run times with MR-AIM. All experiments were run using MATLAB on an Intel Xeon 2.4 GHz processor. As shown in Figure 5, we achieve more than 2x speedup in software. Our experiments showed that with a six-stage multi-resolution framework, we could achieve up to 15x speedup, albeit at the cost of significant loss in resolution at the lower stages of the pyramid, thus resulting in noisier results. This framework then becomes an ideal candidate for a reconfigurable accelerator where different stages of the pyramid can be instantiated based on the environment one is operating in. If high throughput is a requirement, the accelerator can be dynamically reconfigured to instantiate more stages, while if a high priority navigation process is being carried out then three stages could be used.

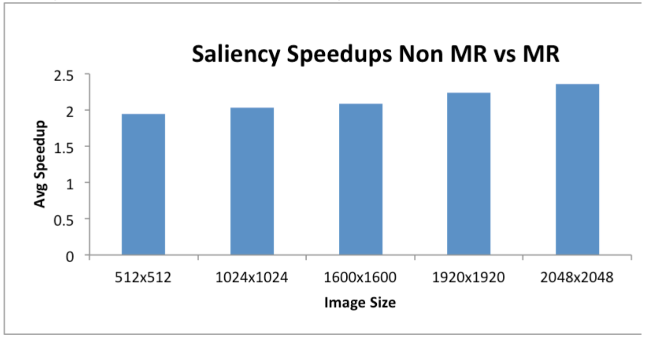


Figure 5: Quantitative Evaluation of AIM versus MR-AIM

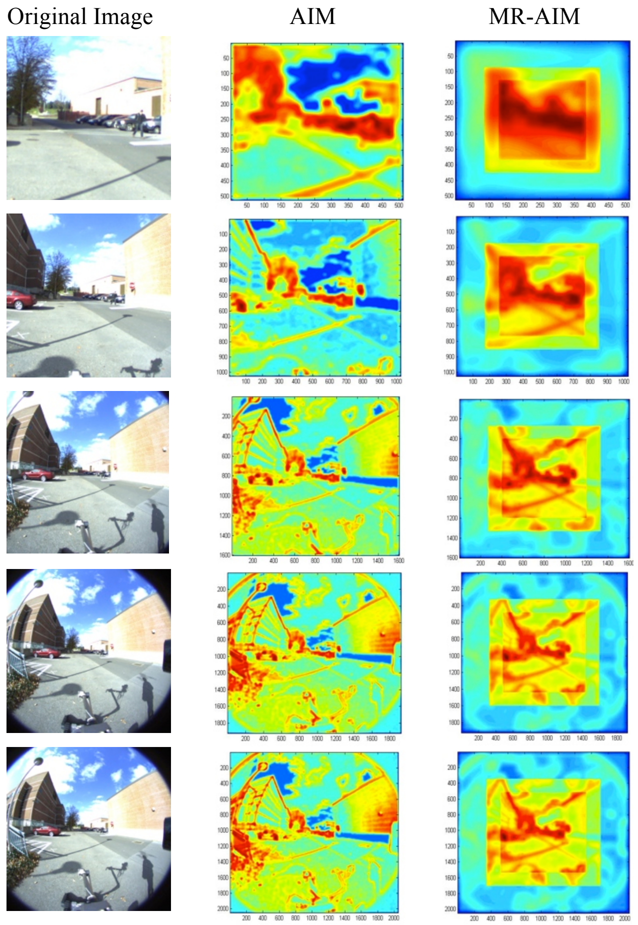


Figure 6: Qualitative Comparison of AIM versus MR-AIM. Column 1: From top to bottom the images are 512x512, 1024x1024, 1600x1600, 1920x1920 and 2048x2048* (* = extrapolated). Column 2: AIM. Column 3: MR-AIM

## 4. RELATED WORK

Early studies on foveation using the concept of multi-resolution pyramid was carried out in [11] for low-bandwidth video communication. Their work however looked at encoding the image itself in a more efficient way rather than look at it in terms of computing saliency. The Gaze Attention Fixation Finding Engine (GAFFE) model [12] builds on the foveation model of [11] to compute saliency in gray-scale images and arrives at various fixation points (not necessarily in the order carried out by the HVS). Another aspect to consider is that they run their model on the foveated (compressed) image rather then computing the fixations as an inherent part of the saliency model. A multi-resolution foveation model was proposed in the context of saliency by using the phase spectrum of quaternion Fourier transform (PQFT) [13]. This was also in context of image and video compression rather than attempting to predict fixation points in a dynamic environment. A recent foveation-saliency approach in [14] was used to improve attention prediction under a quality assessment task. While the state-of-the-art computational models [1], [2], [3] use bottom-up features to compute visual attention, the HVS uses a combination of top-down decision making and bottom-up feature extraction to digest information coming through the ventral stream [15]. Thus, designing a biologically plausible system capable of predicting where to look next is a challenging task. We believe that our work is different from the point of view of answering this very question and the multi-resolution framework modeling the front-end of the HVS is a stepping-stone in that direction.

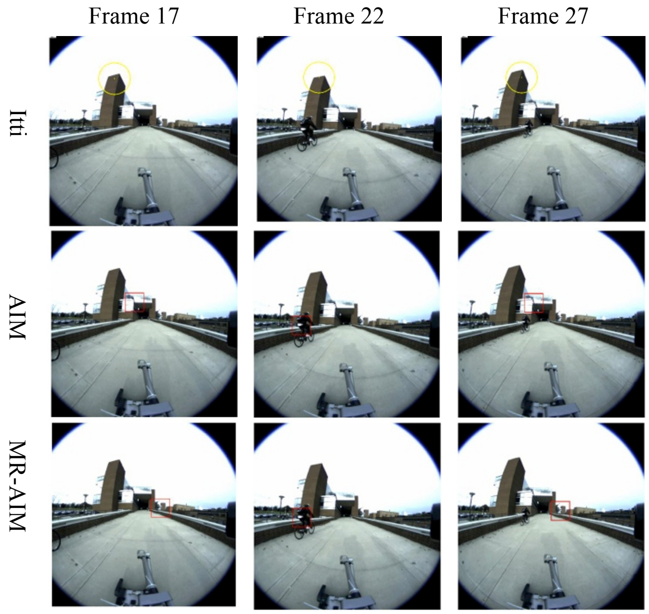


Figure 7: Qualitative comparison of Itti's model (yellow circle shows focus of attention, AIM (red box shows maximum saliency score) and MR-AIM (red box shows maximum saliency score) at three different time instances captured from a video.

## 5. CONCLUSIONS

A multi-resolution framework for visual saliency is presented. The way the HVS operates is modeled using the framework, where resolution rolls of as one would move away from the point of fixation. Qualitative and quantitative comparisons were made with other state-of-the-art models. Apart from biological plausibility, we show significant computational benefits that would enable the design of next-generation autonomous systems driven by visual attention. Future work includes extending this framework by adding additional channels of saliency in the periphery and introducing top-down bias to drive foveation.

## 6. ACKNOWLEDGEMENTS

The authors acknowledge discussions with Nandhini Chandramoorthy (The Pennsylvania State University). This material is based upon work supported in part by the Office of Naval Research under grant number N00014-11-1-0893. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the sponsor. This work was also supported in part by NSF Grants 1213052, 0916887 and Intel Science & Technology Center on Embedded Computing.

## 7. REFERENCES

[1] N. D. B. Bruce and J. K. Tsotsos, "Saliency, Attention, and Visual Search: An Information Theoretic Approach," *Journal of Vision*, vol. 9, no. 3, pp. 5.1–24, Jan. 2009.

[2] 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*, vol. 20, no. 11, pp. 1254–1259, 1998.

[3] L. Itti and C. Koch, "Computational Modelling of Visual Attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, Mar. 2001.

[4] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*. Prentice Hall, 2007.

[5] D. Parkhurst, K. Law, and E. Niebur, "Modeling the Role of Salience in the Allocation of Overt Visual Attention," *Vision Research*, vol. 42, no. 1, pp. 107–23, Jan. 2002.

[6] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard, "Eye Guidance in Natural Vision : Reinterpreting Salience," *Journal of Vision*, vol. 11, pp. 1–23, 2011.

[7] S. Bae, Y. C. P. Cho, S. Park, K. M. Irick, Y. Jin, and V. Narayanan, "An FPGA Implementation of Information Theoretic Visual-Saliency System and Its Optimization," *2011 IEEE 19th Annual International Symposium on Field-Programmable Custom Computing Machines*, pp. 41–48, May 2011.

[8] S. Kestur, D. Dantara, and V. Narayanan, "SHARC : A Streaming Model for FPGA Accelerators and its Application to Saliency," *Design, Automation, and Test in Europe*, 2011.

[9] T. Judd, F. Durand, and A. Torralba, "A Benchmark of Computational Models of Saliency to Predict Human Fixations," 2012.

[10] N. D. B. Bruce and J. K. Tsotsos, "Saliency Based on Information Maximization," *Advances in Neural Information Processing Systems*, vol. 18, pp. 155–162, 2006.

[11] W. S. Geisler and J. S. Perry, "A Real-Time Foveated Multiresolution System for Low-Bandwidth Video Communication," *Proceedings of the SPIE: The International Society for Optical Engineering*, vol. 3299, 1998.

[12] U. Rajashekar, I. van der Linde, a C. Bovik, and L. K. Cormack, "GAFFE: A Gaze-Attentive Fixation Finding Engine," *IEEE Transactions on Image Processing*, vol. 17, no. 4, pp. 564–73, Apr. 2008.

[13] C. Guo and L. Zhang, "A Novel Multiresolution Spatiotemporal Saliency Detection Model and its Applications in Image and Video Compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185–98, Jan. 2010.

[14] L. J. Gide, Milind S and Karam, "Improved Foveation and Saliency-based Visual Attention Prediction under a Quality Assessment Task," *Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 200–205, 2012.

[15] C. Lang, G. Liu, J. Yu, and S. Yan, "Saliency Detection by Multitask Sparsity Pursuit," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 1327–38, Mar. 2012.