Predicting Visual Attention using Gamma Kernels

Ryan Burt, Eder Santana, Jose C. Principe
Computational NeuroEngineering Laboratory
Department of Electrical and Computer Engineering
University of Florida
Gainesville, FL 32601

Nina Thigpen, Andreas Keil
Center for the Study of Emotion and Attention
Department of Psychology
University of Florida
Gainesville, FL 32601

***Abstract*—Saliency measures are a popular way to predict visual attention. However, saliency is normally tested on sets of single resolution images that are unlike what the human vision system sees. We propose a new saliency measure based on convolving images with 2D gamma kernels which function as a comparison between a center and a surrounding neighborhood. The two parameters in the gamma kernel provide an ideal way to change the size of both the center and the surrounding neighborhood, which makes finding saliency at different scales simple and fast. We test the new saliency measure on both the CAT2000 database and the Toronto database and compare the results with other simple saliency methods. In addition, we test the methods on a foveated version of the Toronto database to test whether these methods perform well in a fixation system similar to the human vision system. Gamma saliency is shown to both perform better and compute faster than the competing methods in both the standard databases and the foveated version.**

***Index Terms* - Saliency, Gamma kernel, Image processing, Foveation**

I. Introduction

Humans have the ability to view a scene and form an overall representation in a remarkably short length of time. However, due to the complexity of visual search, it is reasonable to assume that humans do not fixate on and process every small region in an image [11]. Instead, the entire image is quickly sent through a pyramidal processing mechanism that selects fixation regions for more attention [19]. By selecting only these small regions, the human vision system (HVS) is able to quickly process pieces of the scene to form an overall representation that is stored in the brain.

However, current techniques in image processing tend to process entire images by convolving them with learned filters. Here, we can take inspiration from the HVS and only process a number of subregions and form an overall representation of the entire scene. In the same way that convolving learned filters over an image is a step beyond scanning pixel-by-pixel, processing still images as videos of small frames composed of visually interesting regions could be a further step that simply discards large regions of the image that have little to no effect on classification.

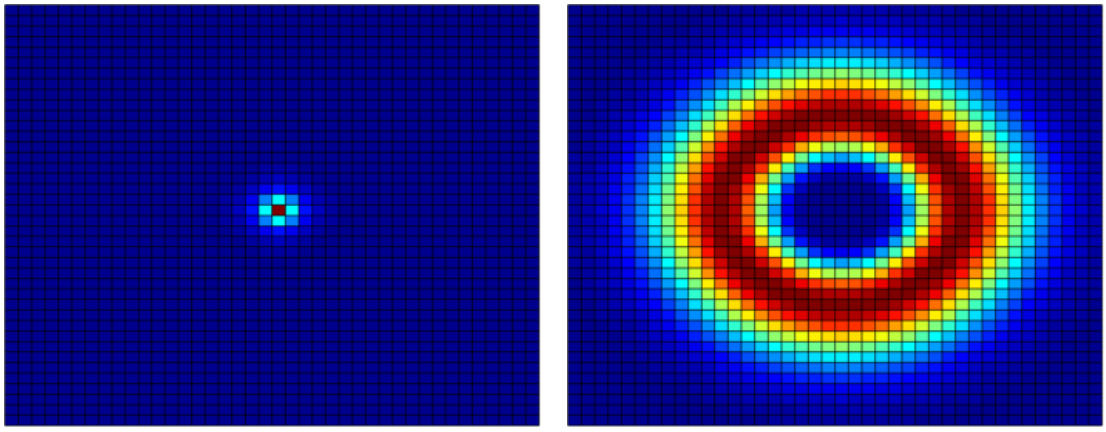
Saliency is defined as the state or quality by which an object stands out relative to its neighbors. An object tends to be more salient if it is brightly colored, flashy, and altogether different from its surroundings. By using saliency as a proxy for visual attention, it could be possible to create a system that quickly selects regions of interest for more computationally intensive processing, then combine these representations into an overall understanding of a complex scene in much the same way the HVS works.

Since the introduction of Itti's method in 1998 [11], saliency has become a popular way to predict visual attention in images and could therefore be used to segment out the interesting regions for faster processing. Most saliency measures work by combining a number of simple features such as color, intensity, and orientation to find distinct regions in images that could attract the human eye. Two competing views of saliency are the center-surround methods that compare a local center to a neighborhood such as Itti [11], Fast and Efficient Saliency [17], AIM [3], Graph Based Visual Saliency, [10], and Region Covariance Saliency [6]; and the global context methods that compare regions to other regions from any location in the image, such as Torralba [18], and RARE2012 [16].

However, there are fundamental differences between how these saliency measures are tested and how the human vision system uses saliency to direct attentive exploration of the surrounding scene. Human vision has full access to high resolution data only in a small region called the fovea where the focus of attention is centered, approximately 3 degrees of visual angle around the point at which gaze is directed during a given moment in time. Thus, in addition to bottom-up saliency, the human brain must infer/extrapolate or remember parafoveal and peripheral information, or use a combination of the two, to compute targets of interest for future fixation locations: As shown from empirical research on saccadic exploratory eye movements, these future fixations will target the blurry, low resolution regions in the visual periphery. In order for saliency metrics to properly mimic the human vision system, they must therefore be able to find regions of interest outside the initial focal area. However, saliency algorithms applied to digital images have per definition access to the full resolution across the field of view.

To address this crucial difference between the biological and computational study, a framework is needed to transform images from single resolution to multi-resolution. Using images with a clear field of focus and a blurred periphery is called foveated imaging. Foveated imaging has been used in other areas in image and video processing to this date, mainly for compression and faster processing [12], [8]. In addition, some

This work is supported by the Office of Naval Research (ONR) grant #N00014-14-1-0542.



(a) The center kernel (b) The surround kernel

Fig. 1. Visual representation of the center and surround kernels

saliency metrics have been tested in multi-resolution images in an attempt to speed computation and improve results [1], [9], but study in this area is still limited.

We propose that to build a vision system that adheres as closely as possible to the human standard, the saliency measure should be capable of predicting regions of interest outside the initial focal area. To this end, we have selected several current saliency metrics as well as created our own and will study them in a standard fixation database, but the images will be foveated before calculating the saliency maps.

Outside of saliency, gamma kernels have been used for target detection [14], [**?**]. The circular shape of the gamma kernels is ideal for comparing a center region to a local neighborhood, and the size of each can easily be controlled through the use of two parameters, which allows for easily changed scales. In addition, the gamma kernel has many properties such as the ability to be computed recursively and the smoothness of the neighborhood that make it well suited to signal processing methods.

In this paper we introduce a new saliency measure based on convolutional 2D gamma kernels. These kernels function as a quickly computed saliency measure since the main difference calculation can be done with a single convolution on each of the feature vectors. We first Gamma saliency in Section II, Section III contains the results of the experiments on both standard and foveated datasets, and Section IV concludes the paper.

II. Method

Similar to the Gamma CFAR, Itti method, and others (though the CFAR is not a saliency measure, but a specific target detector), Gamma saliency is based on the center surround principle: a region is salient if it is different from the neighborhood. In order to compute these local differences, we use a 2D gamma kernel that emphasizes a center while contrasting it with a local neighborhood through convolution:

$$g_{k,\mu}(n_1, n_2) = \frac{\mu^{k+1}}{2\pi k!}\sqrt{n_1^2 + n_2^2}^{k-1} e^{-\mu\sqrt{n_1^2+n_2^2}} \quad (1)$$

For this kernel, n_1 and n_2 are the local support grid, μ is the shape parameter, and k is the kernel order. Using μ and k, we can control the shape of the kernel: when $k = 1$ the kernel peak is centered around zero. For larger kernel orders, the peak is centered k/μ away from the center. In addition, smaller values of μ will increase the bandwidth of the peak.

With these parameters we can construct a 2D shape that compares a center region to a surrounding neighborhood by subtracting a kernel with order $k > 1$ from a kernel with order $k = 1$. The 1st order kernel functions as the center while the higher order kernel forms the surrounding neighborhood. By adjusting the shape parameter and order of the neighborhood kernel we can control the size and location of the neighborhood relative to the center, and as well as adjust the size of the center by using the shape parameter for the center kernel. Fig. 1 shows an example of a center kernel with parameters $\mu = 1$ and $k = 1$ along with the surround kernel with parameters $\mu = 1$ and $k = 10$.

For a multiscale saliency measure, we simply combine multiple kernels of different sizes before the convolution stage (2). A kernel with a larger center scale is subtracted by a surround kernel with a larger and further removed neighborhood, effectively searching for larger objects by comparing more overall area in the image. By summing all the kernels before the convolution stage, we create a system which is capable of computing saliency at different scales without adding extra computation beyond a simple summation. The kernel summation is described in (2), where all k for even m are 1 to create the center kernels. The number of different scales is $m/2$.

$$g_{total} = \sum_{m=0}^{M-1} = (-1^m) g_m(k_m, \mu_m) \quad (2)$$

In addition to the circular shape of the neighborhood, the gamma kernel has other useful properties that can be exploited. The shape of the neighborhoods is smooth, which is in contrast to other methods which choose neighborhoods that sample

Method \ Metric	ROC (Judd)	ROC (Borji)	Similarity	Correlation	NSS	Time (s)
Itti	.712	.597	.384	.275	.341	.280
AIM	.746	.632	.403	.363	.479	1.10
Torralba	.684	.600	.374	.292	.360	.78
GBVS	.848	.677	.488	.570	**.638**	1.03
FES	.847	.586	.520	.572	.446	**.21**
RARE2012	.785	.625	.477	.551	.489	1.39
RCS	.747	.609	.431	.414	.413	15.84
Gamma	**.862**	**.695**	**.588**	**.581**	.546	**.21**

TABLE I
ATTENTION PREDICTION RESULTS ON THE TORONTO DATABASE

Method \ Metric	ROC (Judd)	ROC (Borji)	Similarity	Correlation	NSS	Time (s)
Itti	.700	.570	.377	.206	.258	.25
AIM	.772	.628	.437	.335	.497	1.04
Torralba	.770	.619	.437	.324	.448	1.20
GBVS	.844	.642	.498	.486	.510	1.05
FES	.812	.576	.562	.628	.368	.29
RARE2012	.822	.643	.466	.408	**.511**	1.37
RCS	.763	.593	.431	.292	.352	14.91
Gamma	**.852**	**.676**	**.592**	**.633**	.468	**.21**

TABLE II
ATTENTION PREDICTION RESULTS ON THE CAT2000 DATABASE

Method \ Metric	ROC (Judd)	ROC (Borji)	Similarity	Correlation	NSS
Itti	.737	.597	.403	.314	.369
AIM	.794	.657	.433	.458	.561
Torralba	.784	.650	.433	.469	.539
GBVS	.839	.664	.502	.603	**.594**
FES	.846	.571	.487	.536	.403
RARE2012	.841	.656	.525	.632	.591
RCS	.819	.629	.517	.595	.517
Gamma	**.858**	**.684**	**.607**	**.649**	.483

TABLE III
ATTENTION PREDICTION RESULTS ON THE FOVEATED TORONTO DATABASE

at a fixed radius. Also, the gamma kernel can be computed recursively. Though we don't make use of the recursive computation here in favor or pre-computing the kernel for speed, the recursive property could be exploited to extend this method to work in a temporal structure such as video saliency.

With this local difference measure, the rest of the saliency measure is constructed similarly to the other center surround methods [13]: the image is broken into feature matrices, each matrix is convolved with the multiscale kernel, the matrices are combined and exponentiated to accentuate peaks, then postprocessing is performed to boost results using a Gaussian blur and a center bias.

The feature matrices are composed of the CIELab color space, which has three matrices - one luminence matrix and two color opponency matrices. In CIELab space, the distance between two colors can be calculated using simply the Euclidean distance, which is a useful property that we take advantage of in the convolution. Each of these matrices is convolved with the multiscale gamma kernel to get the saliency measure in each channel (3). In the following equations, $\bullet$ is the convolution operator.

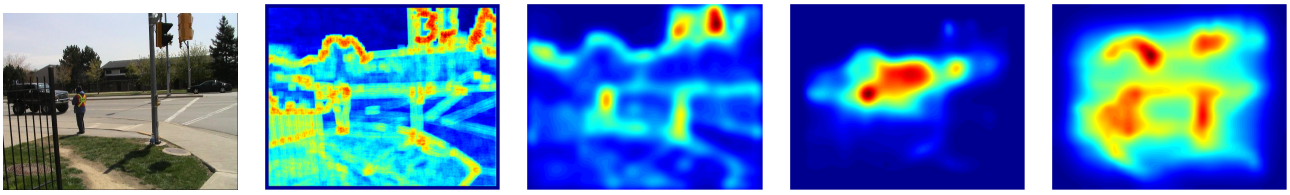
$$S = \frac{|g \bullet L| + |g \bullet a| + |g \bullet b|}{3} \tag{3}$$

Once we have the overall combined saliency map, there are a few common postprocessing mechanisms used to improve results. First, the main peaks in the measure are accentuated by raising the combined map to a power $\alpha > 1$. Next, it is well known that humans tend to fixate on the center of images, so a Gaussian weighting is applied to the center of the image where the variance of the Gaussian is dependent on the image size. Finally, to reduce the effects of noise and created a more streamlined map, the map is blurred using a small Gaussian kernel (4) as in [17].

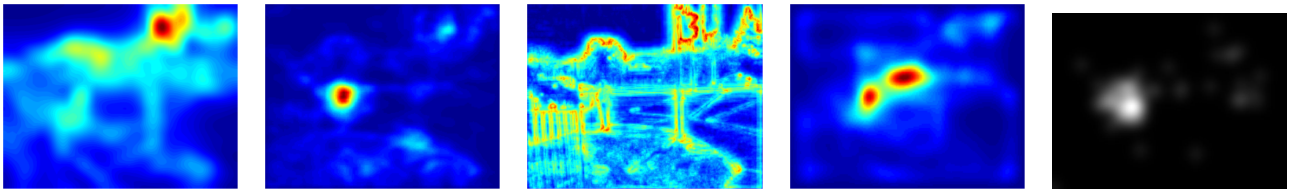
$$S = (S^{\alpha} G(\sigma^2)) \bullet G(.5) \tag{4}$$

III. RESULTS

Results were computed on the Toronto dataset [4] and the CAT2000 training database [2]. The Toronto database consists of 120 images shown to 20 students for four seconds of free-viewing. The CAT2000 database has 2000 images drawn from 20 different categories for a wide variety of image foregrounds and backgrounds, as well as the fixation data from 18 different observers. The observers were given the task of free-viewing each image for five seconds with one degree of visual angle corresponding to roughly 38 pixels in each image. Each set of saliency maps were computed with the default set of parameters recommended by the algorithms. For Gamma



(a) Original Image (b) AIM (c) RCS (d) FES (e) GBVS



(f) Itti (g) RARE2012 (h) Torralba (i) Gamma (j) Fixation Map

Fig. 2. A comparison between different saliency measures for an image in the Toronto Database

Saliency, the parameters used were $k = [1, 1, 1, 26, 25, 19]$, $\mu = [2, 1, .5, 2, 1, .5]$, and $\alpha = 5$.

The maps were then compared to the collected fixation data using these metrics: the area under ROC curve created by Judd, the area under ROC curve by Borji, the similarity measure, the correlation coefficient, and the normalized scanpath saliency. The area under ROC curve by Judd is measured as the proportion of saliency map values above a threshold at the fixation locations to the number of values below the threshold at the fixation locations. In contrast, Borji's version of the area under ROC curve measure the proportion of true postives to false positives, which are the values in the saliency map above a threshold that do not correspond to a fixation location. The similarity measure treats each map as a distribution and computes the histogram intersection. The correlation measure is Pearson's linear coefficient between the two maps. Lastly, the normalized scanpath saliency refers to the mean value of the normalized saliency map at fixation locations [5]. In each of the metrics, the higher number indicates a better result.

For calculating the computation time, each algorithm was set to produce a saliency map sized 128x171 to ensure that algorithms that downsample don't have an inherent advantage for computation time. All times were computed on PC running Matlab R2012a on an i5-2310 clocked at 2.9GHz.

Table 1 shows the results from comparing the saliency maps with the fixation maps in the Toronto database across five different metrics along with the mean time to create a saliency map from a single image in the database, with the best results for each metric in bold. Gamma saliency performs the best in four of five metrics, with the closest competitor being GBVS. Gamma saliency is also the fastest since it is based on a convolutional filter. Table 2 shows the results for the CAT2000 database. Once again Gamma saliency performs the best in 4 of 5 metrics and computes the saliency maps in the fastest times. Fig. 2 shows the resulting saliency maps from a single image in the Toronto Database for a qualitative analysis.

To create the foveated dataset, the present study created images that are increasingly blurred around a small high-resolution area (artifical fovea). To create these images, we used the fast method developed by Geisler and Perry for images and videos in 2002 [7]. This method creates arbitrary visual fields in displays that allows for relatively high frame rates so that the visual field in the displays can be controlled in real time. We used this method to blur each image in the Toronto database around the center point, creating an artificial fovea that corresponds to the approximate size of the original center fixation.

Table 3 shows the results for each saliency measure on the foveated Toronto database. Gamma saliency still performs the best across most of the metrics, which shows that it could be used in a fixation system that approximates the HVS by using foveated inputs. Interestingly, the foveation actually improves the results obtained by most saliency measures, possibly by natually adding a blur and center bias that has been shown to improve results in previous studies.

IV. Conclusion

In this paper we prosed a new saliency method based on a 2D gamma kernel that functions as a convolution filter to estimate local saliency. Using 2D gamma kernels gives us an efficient method that also lends itself easily to the multiscale architecture preferred in most saliency algorithms. We show that the results are better than other comparably simple methods and that the computation time is extremely fast.

Also, we showed that foveating images before processing not only better approximates the working of the HVS, but it also improves the results. This could be due ot the natural center bias and blurring involved in foveation.

Future work could include adapting the shape parameter to find scales that fit the input data, which would eliminate the need to either fix or scan the parameters. Another step would be to extend this measure to work in videos, possibly using the recursive calculation. Finally, we will also include this saliency work in an architecture that saves computation in computer vision systems by only processing salient regions.

REFERENCES

[1] S. Advani, J. Susteric, K. Irick, and V. Narayanan, "A multi-resolution saliency framework to drive foveation" *International Conference on Acoustics, Speech, and Signal Processing*, pp. 2596–2600, 2013.

[2] A. Borji and L. Itti, "CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research" *arXiv:1505.03581*, 2015

[3] N. D. B. Bruce and J. K. Tsotsos, "Saliency Based on Information Maximization" *Advances in Neural Information Processing Systems*, pp.155–162, 2005.

[4] N. D. B. Bruce and J. K. Tsotsos, "Attention Based on Information Maximization" *Journal of Vision*, vol. 7, no. 9, pp.950, 2007.

[5] Z. Bylinski, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "MIT Saliency Benchmark" http://saliency.mit.edu/

[6] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances" *Journal of Vision*, vol. 13, no. 4, 2013.

[7] W. Geisler and J. Perry, "Real-time Simulation of Arbitrary Visual Fields" *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*, pp.83–87, 2002.

[8] W. Geisler, J. Perry, and J. Najemnik, "Visual serach: The role of peripheral information measured using gaze-contingent displays" *Journal of Vision*, vol. 6, no. 9, pp.858–873, 2006.

[9] C. Guo and L. Zhang, "A novel mulitresolution spatiotemporal saliency detection model and its applications in image and video compression" *IEEE Transactions on Signal Processing*, vol. 10, no. 1, pp.185–198, 2009.

[10] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency" *Advances in Neural Information Processing Systems*, pp.545–552, 2006.

[11] 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 , 1998.

[12] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention" *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp.1304–1318, 2004.

[13] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations" 2012.

[14] J.C. Principe, A. Radisavljevic, M. Kim, J. Fisher, M. Hyett and L. Novak, "Target pre-screening based on 2D gamma kernels", *Proceedings of SPIE*, pp.251–258, April 1995.

[15] J.C. Principe, M. Kim, and J. Fisher, "Target discrimination in synthetic aperture radar (SAR) using artificial neural networks", *Trans. on Image Processing*, pp.1136–1149, vol. 7, no. 8, August 1998.

[16] N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, and T. Dutoit, "RARE2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis" *Signal Processing: Image Communication*, issn:0923-5965, 2012.

[17] H. R. Tavakoli, E. Rahtu, and J. Heikkilä, "Fast and Efficient Saliency Detection Using Sparse Sampling and Kernel Density Estimation" *Scandanavian Conference on Image Analysis*, pp. 666–537675, May 2011.

[18] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, "Contextual Guidance of eye movements and attention in real-world scenes: the role of global features in object search" *Psychological Review*, vol. 113, no. 4, 2006.

[19] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning" *Artificial Intelligence*, vol. 78, no. 1, pp.507-545, 1995.