HIGH RESOLUTION BIOLOGICALLY INSPIRED SALIENT REGION DETECTION

Yusuf Saber[1] , Matthew Kyan[2]

Ryerson University
Department of Electrical and Computer Engineering
350 Victoria Street, Toronto, Ontario, Canada M5B 2K3

[1]ysaber@ee.ryerson.ca, [2]mkyan@ee.ryerson.ca

ABSTRACT

This paper presents a novel method for salient region detection. With the same biologically inspired logic as Itti et al., we use wavelet decomposition to obtain a ’centre’ scale, and filtered versions of the original image for the ’surround’ scale. Our method yields saliency map that is of the same resolution of the investigated image, with a detailed outline and highlight of the salient object. Previous methods are discussed in detail and comparisons are made as a validation process for our method.

Index Terms— Saliency, visual attention, human visual system, wavelets.

1. INTRODUCTION

According to William James, the father of American psychology, a two component framework is implemented in the human visual system (HVS) for attentional deployment[1]. This framework suggests that the HVS directs its attention to an object based on bottom-up cues as well as top-down cues. Bottom-up cues are those that unintentionally grab one’s attention, an example of which is face detection. Top-down cues are those that are intentionally looked for, an example of which is face recognition. Hence, the top-down attention model requires prior information, while the bottom-up model does not. Because of this, much research has been explored into modeling the bottom-up concept of attention. To display points of attention, a two dimensional function is developed which maximizes where the attention exists the most. This function is known as a saliency map. In the digital realm, a saliency map is used to determine points of attention in a scene. In terms of a human’s gaze, the location on the saliency map that yields the highest energy is where the gaze would unintentionally be directed.

Although various methods exist to extract saliency, they all have the same fundamental backbone: to compare local information with global information. The difference between the local and global is directly correlated to the local region’s saliency, where pixels that are largely different than the global representation yield a higher saliency. Some of the more basic methods try to extract salient pixels by computing some representation of the overall image and subtracting each pixel in the image from that value [2]. Equation (1) gives this formula, where α is the representation of the entire image (the average pixel value of the image for example), s_p is the saliency of pixel p, and i_p is the intensity of pixel p.

$$s_p = |i_p - \alpha| \qquad (1)$$

One of the most prominent methods of detecting saliency is proposed by Itti et al. [3] which is a biologically inspired one that traverses the image and spatially compares a small local region to a larger local region; doing so with respect to various features of the image such as intensity, color, orientation, hue, and/or texture[3]. Although this method is quite robust, it requires a lot of processing time and yields a very low resolution map.

Although low resolution saliency maps have their applications, such as robotic direction, Achanta et al. [4] proposed a set of definitions for the generation of saliency maps that serves a wider range of applications. These definitions, discussed in section 2.2, propose that the saliency map yield high resolution maps (the same resolution as the investigated image) that outlines and includes the entire salient object, without picking up noisy elements in the image. Such a saliency map has potential for an array of applications including image segmentation, image compression, and object tracking in video.

Although Achanta et al. succeeded in fulfilling their requirements, their method implements the center-surround function on a purely pixel-to-global basis, not considering pixels to their direct surrounding, but to the image as a whole. The proposed method in this paper uses Itti et al.’s original center-surround function and applies in a high resolution setting, using wavelets to allow for lossless processing.

The rest of this paper is organized as follows: Section 2 discusses, in detail, some of the relevant previous work in the literature of saliency map detection. Section 3 discusses the proposed method as well as the basis of its construction.

978-1-4577-1303-3/11/$26.00 ©2011 Crown

Section 4 shows some results and comparisons to previous methods, showing the significance of the proposed method. Section 5 gives some concluding remarks.

2. PAST APPROACHES

To get a better understanding of how the proposed method was compiled, a short study of the appropriate past methods for saliency detection are discussed.

2.1. Itti's Method

Itti et al.'s method [3] stands out amongst the rest because it is the first to put the feature integration theory [5] into practise. The feature integration theory states that the HVS scans various features for attentive points, and then combines all the points to provide a perception of attention.

Itti first extracts features from the image (intensity, color, and orientation, among others) and then performs a centre-surround operation on each of the features [3]. The centre-surround operation is the basis for all bottom-up attention models. It is used to compare local regions to larger, comparably global, regions. The centre-surround operation, when applied to a feature map, yields a conspicuity map of the attentive regions in the image. A conspicuity map is a saliency map of a single feature. Itti proposes that many scales be analyzed to obtain a true comparison between the local and global information. He uses Gaussian pyramids to downsample the image, which results in lossy images, which ultimately provide a very low resolution saliency map when resized to the size of the original image. Finally, all of the conspicuity maps for all of the scales must be normalized and combined to form a final saliency map.

The normalization operator proposed requires some discussion because of its importance. Rather than simply limiting the pixel values to a certain range, Itti's normalization operator is an iterative process that allows high energy points in the saliency map to further stand out, while suppressing low energy points. This is a useful operator because different maps usually contain different dynamic ranges. This means that if one map contains a large number of high-energy points, none will stand out, while a map with many low-energy points but only one or two high-energy points will require those points to stand out. This function operates as such:

1. Find the global maximum, M
2. Find the average of all other local maxima, m
3. Globally multiply the entire map by $(M - m)^2$

This operation allows for a fair comparison between the conspicuity maps and an even, balanced summation of them.

2.2. Frequency Tuned Method

As mentioned in the introduction of this paper, Achanta et al. [4] proposed a set of definitions to yield a more versatile saliency map. Those definitions are:

1. Emphasize the largest salient objects
2. Uniformly highlight whole salient regions
3. Establish well-defined boundaries of salient objects
4. Disregard high frequencies arising from texture, noise, and blocking artifacts
5. Efficiently output full resolution saliency maps

To meet all of the aforementioned requirements, Achanta et al define their saliency map function as:

$$S(x, y) = |I_\mu - I'(x, y)| \tag{2}$$

where I_μ is the mean image value and $I'(x, y)$ is the corresponding image pixel value in the Gaussian blurred version of the original image. Blurring the image using a Gaussian kernel allows high frequencies to be disregarded (requirement 4) and hence produces a salient regions map. Since no downsampling occurs, the yielded saliency map is of the same resolution as the original image.

Ngau et al. [6] later proposed that instead of blurring the image to remove high-frequency noise and artifacts, one could perform wavelet decomposition to yield a downsampled approximation of the image. This would allow the image to be reconstructed without really losing any information, and without resorting to blurring the image which, in a sense, is a de-resolution operation and results in loss of salient objects that are very small in size.

Ngau et al. applied the same formula as Achanta et al. (Equation 2) except that they applied it to the approximation of the wavelet decomposition (using only one level). They then reconstructed the image back to its original size.

Our method uses wavelets to allow for lossless processing, while incorporating the center-surround operation for a true comparison of the local and global information.

3. THE PROPOSED METHOD

There are two main drawbacks with Itti et al.'s method in that it requires a lot of processing time (so it cannot be applied in real time), and it is of very low resolution (so an exact outline of the object cannot be extracted). Our goal was to overcome these drawbacks while still including the feature integration theory to allow for a biologically inspired process.

Wavelet theory has recently taken on a large role in image processing due to its lossless resizing capabilities. Since the one factor eluding to loss of resolution in Itti's method is the downsampling using Gaussian pyramids, we chose to use wavelet decomposition and take the one-level-down approximation as the downsample to the original image. Thus, performing the centre-surround function using the original image and said approximation, we have the capability to resize the resulting conspicuity map back to the size of the original image without any loss in resolution.

Figure 1 shows the block diagram of the proposed method. It should be noted that Figure 1 only shows the grayscale component as well as the red-green color opponency, and not the blue-yellow opponency. This was done to save space. The blue-yellow opponency block diagram is identical to the red-green one except it uses the blue and yellow color maps rather than the red and green maps.

The centre-surround used in the proposed method is the same used by Itti et al. Taking the wavelet approximation as the centre scale, c(θ), and the original map as the surround scale, s(θ), the centre-surround function is implemented by downsizing the surround scale to the size of the centre-scale using bicubic interpolation, and performing absolute point-by-point subtraction. It is now clear why wavelets are necessary, since the output of the centre-surround function is not the same size as the original image.

In Figure 1, the input image is first put through a set of linear filters to extract grayscale information (conversion as in NTSC) as well as red, blue, green, and yellow information (conversion as in [3]). Each of these information maps are then decomposed using a wavelet filter to extract an approximation as well as horizontal, vertical, and diagonal detail maps. With regards to the grayscale map, the wavelet decomposition components are labeled GrA (approximation), GrH (horizontal), GrV (vertical), and GrD (diagonal). The same applies to the red and green maps, except the prefix Gr is replaced with R and G, respectively. The centre-surround operation (shown as the θ operator) is performed using the approximation and original information map. Such is the case for the grayscale component. However, to implement color opponency, the centre-surround operation is applied to the red information map with the green wavelet approximation, and vice versa, and similarly with the blue-yellow opponency (not shown). The outputs of the centre-surround operation are then used to reconstruct the image using the wavelet decomposition components. All of the resulting conspicuity maps are then normalized (using Itti's normalization function [3]) and summed to form the final saliency map.

In our method, we did not implement orientation saliency as Itti did [3]. This is because at such a high scale (the original and one scale down), an orientation map looks similar to an edge map. This obstructs the final saliency map as the edges in the orientation map may not lay exactly at the boundaries of the regions yielded by the intensity and color maps and hence some objects appear to be surrounded by a ringing/halo effect. Also, the orientation results did not provide much new information anyway, and ignoring them also improves processing time.

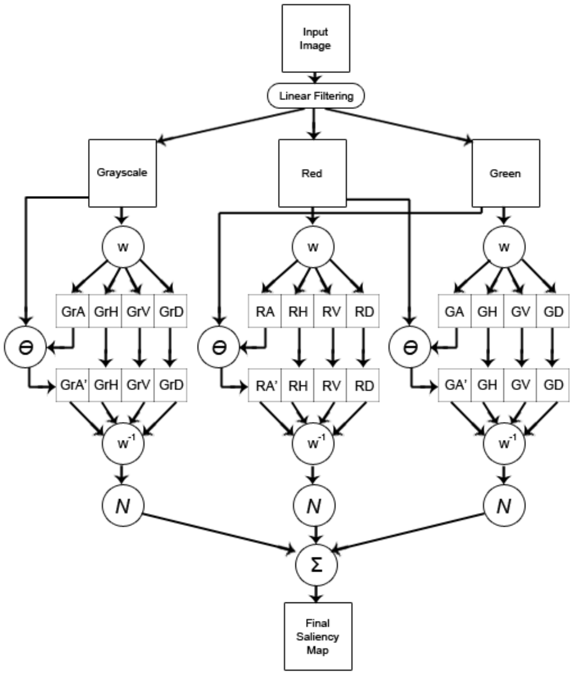

Fig. 1: Block diagram of the proposed method.

4. RESULTS AND ANALYSIS

To test the validity of the proposed method, a few test images were taken from the MSRA Salient Object Database[7]. Both Achanta's method and our method were run on the sample images. Other methods could have been used for comparison, however, Achanta's method is the only one that extracts high resolution maps, therefore other methods such as Itti's [3], and the spectral residual [8], were not shown in the comparison. Recently, Achanta created a method that uses maximum symmetric surround (MSSS) to detect salient regions [9] and the results are an improvement over the frequency-tuned method, but not by much (see [9]). Because the improvement in the results is not drastic, this method was also not used for comparison.

The proposed method provides many benefits, one of which is that it considers information from multiple scales. This is of particular importance when it comes to color opponency because of the varying representations of information each color map provides. With intensity, a local-global comparison can be fairly done within the same scale on the same map. However, a local-global comparison on the same scale across two different color maps can cause some issues with respect to the dynamic ranges of the maps. Using multiple scales as well as reciprocated evaluations (i.e. green centre vs red surround and red centre vs green surround) overcomes any issues of dynamic ranges while maintaining the integrity of the centre-surround operation.

The results in Table 1 show that our method clearly yields better results in terms of completeness of the salient object. Achanta's method highlights various parts of the image, but the salient object is not distinct. Our method also highlights small non-salient areas of the image, however, the salient object is still easily distinguishable.

One might notice that in our maps, smooth regions tend to be blocky. There are two possible reasons why this occurs. On one hand, the center-surround operator may not be recalling much of a difference within that block, so it returns a constant value. On the other hand, the wavelet operator may not be decomposing these regions, which in turn would affect the centre-surround operator as previously mentioned. To overcome this issue, a thresholding function could be applied to separate the object from the background entirely.

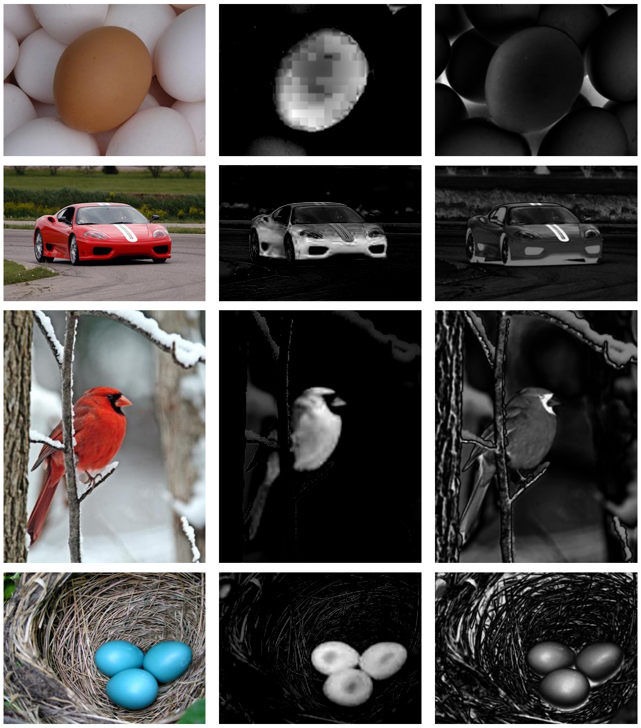

Table 1: Left column: Original images from [7]. Middle column: Our results. Right column: Achanta et al's results (using code from [4]).

5. CONCLUSIONS

The benefit of wavelets in image processing is clearly demonstrated in this paper. Biologically inspired methods have an upper hand over methods that focus on computational efficiency, such as [4] and [8], in that they provide a plausible representation of how attention is sought. However, they have long been associated with computational inefficiency. The introduction of wavelets overcomes this problem, while keeping the benefits of using the feature integration theory intact.

Color opponency can be further explored using the LUV color circle to promote more opponencies. Using only red-green and blue-yellow opponencies limits the potential of other present opponencies and perhaps steals the spotlight from potentially more appropriate opponencies such as yellow-violet and orange-blue, which are direct opposites on the LUV color circle, whereas blue-yellow are not direct opposites. The LUV space is an appropriate model to apply opponency to because it provides perceptual uniformity.

6. REFERENCES

[1] W. James, "The principles of psychology," *Harvard University Press, Cambridge, Massachussetts*, 1980/81.

[2] H. X. Q. Zhang, "Extracting regions of interest in biomedical images," *2008 International Seminar on Future BioMedical Information Engineering*, 2008.

[3] E. N. L. Itti, C. Koch, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions On Pattern Analysis and Machine Intelligence, Vol. 20, No. 11.*, 1998.

[4] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[5] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology, vol. 12, no. 1*, 1980.

[6] C. Ngau, L. Ang, and K. Seng, "Bottom-up visual saliency map using wavelet transform domain," *3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT)*, 2010.

[7] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum, "Learning to detect a salient object," *Proc. IEEE Cont. on Computer Vision and pattern Recognition (CVPR), Minneapolis, Minnesota*, 2007.

[8] L. Z. X. Hou, "Saliency detection: A spectral residual approach," *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07.*, 2007.

[9] R. Achanta and S. Susstrunk, "Saliency detection using maximum symmetric surround," *Proceedings of 2010 IEEE 17th International Conference on Image Processing*, 2010.