A Modified Approach to Biologically Motivated Saliency Mapping

Shane Grant
Department of Computer Science
University of California, San Diego
La Jolla, CA 92093
wgrant@ucsd.edu

Kevin A Heins
Department of Mathematics
University of California, San Diego
La Jolla, CA 92093
kheins@ucsd.edu

Abstract

The human eye is trained to find objects of greater importance in its field of vision. Hence visual attention systems have been proposed to replicate this ability for computer vision. In this paper, a leading method to mirror mammalian saliency detection will be evaluated and expanded upon to more accurately model the most salient features of an image. We compare our approach against traditional methods via a simple survey of human participants.

1. Introduction

When observing a crowded scene, the human eye can quickly filter out noise and focus on only regions of interest. Modeling this natural behavior proves to be quite difficult because human vision is not well understood. However, we aim to emulate what is thought to be subconscious preprocessing of the visual field of view.

The human eye seems to be guided by both object recognition through discriminant analysis and stimulus-driven signals that announce that a location is sufficiently different from its surroundings to be worthy of attention. These two theories are known as top-down and bottom-up, respectively.

Our approach focuses on building a bottom-up saliency map to highlight regions worthy of further analysis. Previous approaches to this problem include a biologically plausible approach first made by Itti *et al*. [3], and a supervised approach developed by Liu *et al*. [6]. Our approach most closely follows that proposed by Itti *et al*.

2. Background

The eyes rapidly focus on new objects of interest and process information. Modelling such a system could provide an efficient and viable active vision system.

2.1. Biological Saliency

In 1985 Koch and Ullman proposed a theory to describe the underlying neural mechanisms of vision and bottom-up saliency [4]. They posited that the eye selects several features that pertain to a stimulus in the visual field and combines these features into a single topographical 'saliency map.' Thus an area of an image that differs greatly by these features (color, intensity, orientation, etc.) could be classified as more salient than its surroundings.

In the retina, photoreceptors, horizontal, and bipolar cells are the processing elements for edge extraction. After visual input is passed through a series of these cells edge information is delivered to the visual cortex. In addition, a neural circuit in the retina creates opponent cells which recieve inhibitory and excitory responses from various cones in the eye. These systems combine with further processing in the lateral geniculate nucleus (LGN), which plays a role in detecting shape and pattern information such as symmetry, as a preprocessor for the visual cortex to find a saliency region [8].

2.2. Computational Saliency

The Koch and Ullman model was purely theoretical until it was directly applied to computer vision by Niebur and Koch in 1996 [7], and refined by Itti *et al*. [3, 2]. They combined features such as intensity, color, orientation, and motion to create the saliency map. These saliency maps have been used for numerous applications, including predicting eye movement [9] and prioritizing selection [10].

3. The Model

Input is provided in the form of a color image in any resolution. This image is split into several visual feature maps based upon different aspects of human vision. A linear combination of these maps forms the final saliency map, which can be used for further processing as needed.

3.1. Color Space Feature Maps

The standard way of representing an image on a computer involves using the RGB color space, which represents image pixels as ordered triples containing the intensity of red, green, and blue color in a pixel. It has been an industry hardware standard and a computationally simple model for color.

Our approach uses the CIE L*a*b* color space, which was designed to closely mimic how human vision is believed to perceive color [1]. It describes colors along three axes in 3D space - the L* axis corresponds to the luminosity of a color - a value of 0 is pure black, while a value of 100 is pure white. This component closely matches with human perception of lightness.

The other two axes, a* and b*, theoretically have no upper or lower bounds. Negative a* corresponds to a green color, positive to a red color. Negative b* to a blue color, positive to a yellow color. In reality these are often bounded between -128 and 127, depending on the number of bits used to represent the color.

The a* and b* maps replicate the color double-opponent system wherein in the center of the receptive field, neurons are excited by one color and inhibited by another, while the converse is true in the surrounding area. This can be computationally represented via a center-surround difference, which highlight regions of contrast within the map by computing the absolute difference of an image at different scales [3]:

$$M(c,s) = |M(c) \ominus M(s)| \tag{1}$$

$$c \in \{2,3,4\}, \delta \in \{3,4\}, s = c + \delta$$

where $\ominus$ represents interpolated pixel-wise differencing.

To generate these different scales, a pyramid of eight dyadic scales is calculated on the raw color map channels with a Gaussian kernel applied at each decimation. We use the kernel proposed by Walther to reduce shifting of the image during this process [12].

3.2. Orientation Feature Maps

The orientation of the edges provide another cue to finding the most salient objects within an image. Salient objects may have edge orientations that differ from the surrounding area [3].

We determine edge orientations using a set of Gabor filters at different scales and orientations. In total, four orientations, $\{0, \pi/4, \pi/2, 3\pi/4\}$, each at two different scales, are computed. These filters are designed to have a strong response to regions in the image that match their orientation [11].

A Gabor filter is obtained by modulating a sinusoid with a Gaussian:

$$g(x,y,\theta,\phi) = e^{-\frac{x^2+y^2}{\sigma^2}} e^{2\pi\theta\mathbf{i}(x\cos\phi + y\sin\phi)} \tag{2}$$

Each scale/orientation pair is run through the same dyadic pyramid as the color feature maps to produce a total of 24 orientation feature maps after the center-surround differences are taken.

3.3. Symmetry Feature Maps

The last cue we consider for our saliency map are local bilateral differences. Symmetry tends to be an important mechanism to identify the structure of objects. Objects likely to be salient, such as man-made objects, plants, and animals, tend to have pronounced symmetry; hence it is believed to be an important aspect of bottom-up detection [8].

To identify symmetry within an image, the local frequencies are analyzed to determine local symmetries and asymmetries. Gabor filters with two different periodic functions are used: a cosine wave to identify local symmetries and a sine wave to identify local asymmetries. Specifically, log Gabor filters are used, which use a Gaussian only when viewed from a logarithmic frequency scale. The differences of these even and odd filters are taken over several orientations, to provide a penultimate symmetry map [5]:

$$Sym(x) = \frac{\sum_n \lfloor |e_n(x)| - |o_n(x)| - T \rfloor}{\sum_n A_n(x) + \epsilon} \tag{3}$$

where e_n is the even cosine function, o_n is the odd sine function, A_n is the magnitude of the filter response vector, ϵ is a term to prevent division by zero, and T is a noise compensation term.

Once the symmetry map has been calculated, center surround differences are taken once again to provide a final symmetry map which can then be used with the aforementioned maps.

3.4. Final Saliency Map

The final saliency map is created with a linear combination of so called ”conspicuity maps,” defined as follows:

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

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

$$\bar{O} = \sum_{\theta \in \{0,\pi/4,\pi/2,3\pi/4\}} \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(O(c,s,\theta)) \tag{6}$$

$$\bar{S} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(S(c,s)) \tag{7}$$

The $\bigoplus$ operator consists of reduction of each map to scale $1 : 16$ and pixel-wise addition. L, A, and B refer

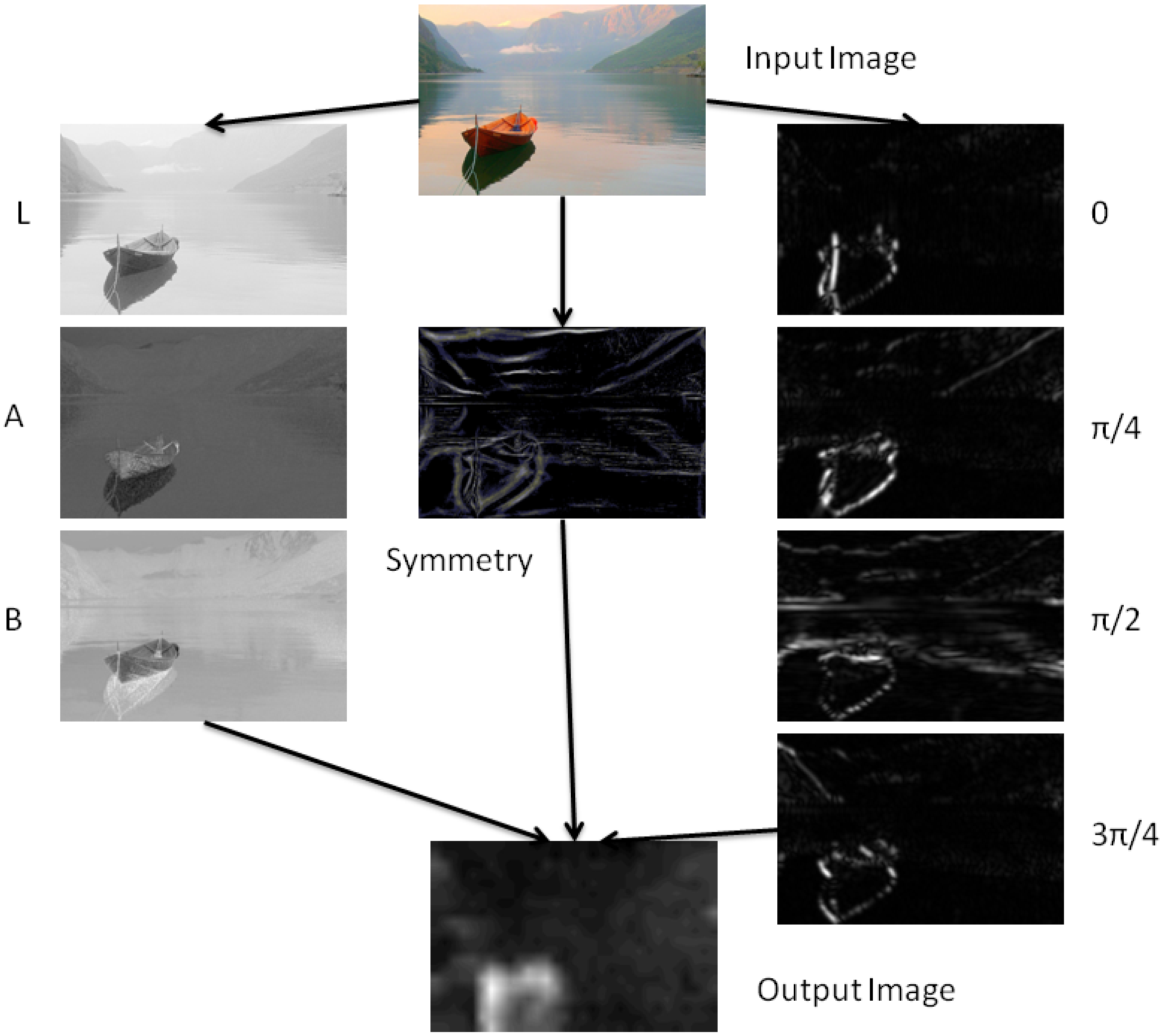


Figure 1. The original input image is divided into several feature channels: lightness (L), red-green opponency (A), blue-yellow opponency (B), symmetry, and four orientation channels. Center-surround differences are then taken to create feature maps, and then combined to created the final output image.

to the raw luminosity, a* (red-green), and b* (blue-yellow) color maps. O refers to the orientation maps while S refers to the symmetry maps. The normalization operator $N(.)$ is performed at each step of this process.

As seen in biological systems, inhibition appears weakest at the center of attention and weaker at the extremities [2]. We utilize a two dimensional difference-of-Gaussians (DoG) filter to emulate this.

In our implementation, this is accomplished by using two Gaussian kernels and taking their difference. They are defined as follows:

$$DoG\,(x,y) = \frac{c_{ex}^2}{2\pi\sigma_{ex}^2}\, e^{-\frac{x^2+y^2}{2\sigma_{ex}^2}} - \frac{c_{inh}^2}{2\pi\sigma_{inh}^2}\, e^{-\frac{x^2+y^2}{2\sigma_{inh}^2}} \tag{8}$$

Where $\sigma_{ex} = 2\%$ and $\sigma_{inh} = 25\%$ of the input image width, $c_{ex} = 0.5$ and $c_{inh} = 1.5$. The normalization operation iteratively applies this DoG filter to the image a set number of times:

$$M \leftarrow |M + M * DoG - C_{inh}|_{\geq 0} \tag{9}$$

where C_{inh} a constant inhibitory term and all negative values are replaced with zero.

The result of this normalization is that isolated salient regions are excited while images with numerous similarly salient regions are suppressed.

The ultimate saliency map is created using a simple linear combination of every conspicuity map. The weights are distributed evenly amongst the different categories of feature maps - image intensity, color opponency, orientation,

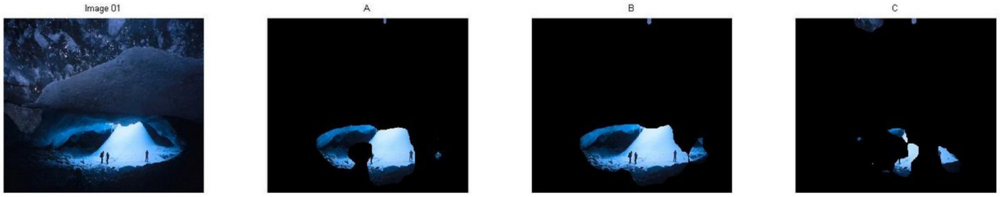


Figure 2. An example of an image shown to survey participants. The original image followed by three randomly ordered masked variants. In this case, A is LAB, B is SYM, and C is RGB.

and symmetry. In our implementation, the weights applied to each conspicuity map are equal.

$$\begin{aligned} Saliency &= w_L N\left(\bar{L}\right) + w_C N\left(\bar{C}\right) \\ &+ w_O N\left(\bar{O}\right) + w_S N\left(\bar{S}\right) \end{aligned}$$

$$\sum_{i \in \{L,C,O,S\}} w_i = 1$$

$$Saliency_{final} = N_{final}\left(Saliency\right) \tag{10}$$

where $N_{final}\left(.\right)$ is a final normalization operation based upon maximal suppression. Values are first normalized to a fixed range $[0..M]$ before the global maximum M and the average $\bar{m}$ of all other local maxima are found. Finally the entire map is multiplied by $\left(M - \bar{m}\right)^2$ to finish the normalization procedure.

4. Results and Evaluation

To evaluate our model of saliency, we set up three differing models: 1) Itti's algorithm (RGB), 2) Ittis algorithm with our approximation for human color vision (LAB), and 3) Itti's algorithm including human color and symmetry detection (SYM). Saliency maps were computed for the same image on all three models, and all non salient regions were masked to provide a final image. Due to inherent subjectivity of saliency, several individuals performed evaluation by ordering the models best to worst for a collection of 24 images with one or more prominent subjects. Best was defined as the image that best captured whatever the individual found most important in the original image.

To ensure validity of the evaluation, none of the images were used in the testing of the model and have no predetermined biases towards any of the models. The order in which the three separate models display was random, and each model was simply labeled ”A”, ”B”, or ”C”.

4.1. Results

For each image, the model with the highest ranking was given a value of 3, followed by a value of 2, and a value of 1 for the worst. Ties were allowed, and the point values were averaged accordingly. A total of 13 respondents participated in the survey. Overall, the SYM model masking had the highest average value of 2.3686, followed by 1.9359 for LAB, and 1.6955 for RGB. In our survey the model with the two new features, symmetry and the LAB color space, was favored.

First, a two sample t-test was run to test whether the scores were significant enough to claim one is greater than another. Testing with the standard deviation of the 24 images averaged scores, the following test was implemented:

$$H_0 : Score_1 = Score_2$$

$$H_1 : Score_1 > Score_2$$

The corresponding t- and p- values are as follows:

Models	t-score	p-value
LAB $>$ RGB	1.5214	0.0675
SYM $>$ LAB	3.4031	$6.9468 * 10^{-4}$
SYM $>$ RGB	4.5370	$2.0399 * 10^{-5}$

It can be concluded that the inclusion of symmetry clearly helps the salient regions, but the null hypothesis cannot be outright rejected for LABs improvement over RGB.

To test whether LAB was superior to RGB, the SYM results were discarded, and instead we ran a test to see whether LAB is superior to RGB over 50% of the time. The test was set up as follows:

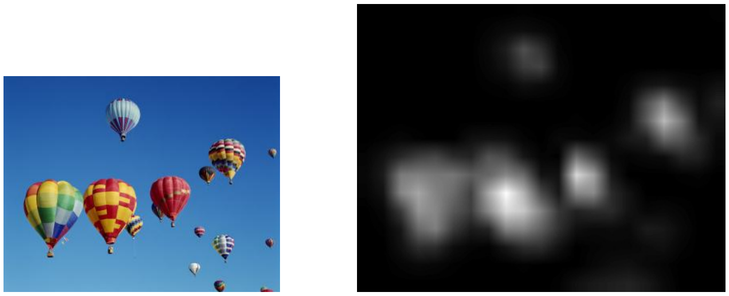
$$H_0 : p_0 = 0.5$$

$$H_1 : p_0 > 0.5$$

A z-value of 2.6042 and a p-value of 0.004604 were reported. Thus there is some further evidence that LAB is superior to RGB for significance levels above 0.01.

5. Conclusion

There seems to be evidence that using a color model closer to that of human vision results in superior results compared with traditional RGB based models. Furthermore, there is evidence that symmetry is a relevant feature in attentional selectivity.



(a) Original Image

(b) Saliency Map w/ LAB + Symmetry

Figure 3. An example of the input/output of the algorithm.

However, occlusion presents a problem for local symmetry filters, which have no way to detect this. When a symmetrical image is covered or distorted, the symmetry can dissapear and the filter will have little response.

Scale is important in determining which objects are salient in a scene. Depending on the area of search, objects which were previously not important can suddenly become salient. Our algorithm does not address this issue.

Finally, it seems as if learning is crucial to the advancement of saliency detection. The features we utilize in our algorithm are extracted without any knowledge of the input image or contextual clues. Work by Liu *et al*. has begun to address this issue and incorporate learning with traditional saliency detection algorithms [6].

6. Acknowledgements

Special thanks to Serge Belongie and the class of CSE190a for help and feedback throughout the project, and to all the participants of our survey.

References

[1] G. Hoffmann. Cielab color space. Technical report, University of Applied Sciences in Emden, 2003. 2

[2] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40:1489–1506, 2000. 1, 3

[3] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence*, 20:1254–1259, 1998. 1, 2

[4] C. Koch and S. Ullman. Shifts in selective visual attention: Towards an underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985. 1

[5] P. Kovesi. Symmetry and asymmetry from local phase. *Tenth Australian Joint Conference on Artificial Intelligence*, pages 185–190, 1997. 2

[6] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *CVPR*, 2007. 1, 5

[7] E. Niebur and C. Koch. Control of selective visual attention: Modeling the 'where' pathway. *Neural Information Processing Systems*, 8:802–808, 1996. 1

[8] S.-J. Park, J.-K. Shin, and M. Lee. Biologically inspired saliency map model for bottom-up visual attention. *2nd Workshop on Biologically Motivated Computer Vision*, pages 418–426, 2002. 1, 2

[9] D. Parkhurst, K. Law, and E. Niebur. Modelling the role of salience in the allocation of overt visual attention. *Vision Research*, 42:107–123, 2002. 1

[10] D. Parkhurst and E. Niebur. Variable resolution displays: A theoretical, practical and behavioral evaluation. *Human Factors*, 44:611–629, 2002. 1

[11] V. S. N. Prasad and J. Domke. Gabor filter visualization. Technical report, University of Maryland, 2005. 2

[12] D. Walther. *Interactions of Visual Attention and Object Recognition: Computational Modeling, Algorithms, and Psychophysics*. PhD thesis, California Institue of Technology, 2006. 2