# Proto-Object Based Visual Saliency Model with a Motion-Sensitive Channel

Jamal Lottier Molin*, Alexander F. Russell, Stefan Mihalas†, Ernst Niebur‡, Ralph Etienne-Cummings§
*‡§Johns Hopkins University, Baltimore, MD, 21218
†Allen Institute for Brain Science, Seattle, WA, 98103
[ *jmolin1 | §retienne | ‡niebur ] @jhu.edu,

***Abstract*— The human visual system has the inherent capability of using selective attention to rapidly process visual information across visual scenes. Early models of visual saliency are purely feature-based and compute visual attention for static scenes. However, to model the human visual system, it is important to also consider temporal change that may exist within the scene when computing visual saliency. We present a biologically-plausible model of dynamic visual attention that computes saliency as a function of proto-objects modulated by an independent motion-sensitive channel. This motion-sensitive channel extracts motion information via biologically plausible temporal filters modeling simple cell receptive fields. By using KL divergence measurements, we show that this model performs significantly better than chance in predicting eye fixations. Furthermore, in our experiments, this model outperforms the Itti, 2005 dynamic saliency model and insignificantly differs from the graph-based visual dynamic saliency model in performance.**

## I. Introduction

The ability of the human and animal visual system to segment visual scenes into conspicuous regions is essential for higher-level processing. The retina rapidly outputs large amounts of visual information estimated up to 100Mbps per optic nerve, thus, parallel processing of the entire visual field is unlikely [1], [2]. The visual system must use selective attention and filter out insignificant information in order to quickly minimize the amount of information forwarded for further processing. The brain uses two mechanisms that work together to compute a saliency map for a given visual stimulus: bottom-up attention and top-down attention. Bottom-up attention is solely a function of information within the visual scene. Top-down attention, rather, is biased attention that is a function of one's internal state and goals. In this work, however, we seek to model a biologically-plausible model for computing bottom-up, dynamic visual attention. External top-down components can easily be integrated with this model for modulating the computed bottom-up-based saliency map in an application-specific manner [3].

Early models of the visual system compute visual saliency by segmenting information in the visual field into sub-units, or feature maps, and then performing a sequence of computational tasks on the individual feature channels, simultaneously. These models then combine the computed attention maps from the individual feature channels into a final saliency map that can then be used for further processing [7]. The saliency map is a single-valued map representing saliency within the visual field such that pixels/objects of higher saliency are awarded higher values. We present an extension to a previously designed object-based visual saliency model, [5], by implementing a temporal component for computing a dynamic saliency map for visual scenes exhibiting motion.

### A. Feature-based Saliency Models

Arguably, the most common feature-based saliency model is the model proposed by Itti, Koch, and Niebur in 1998 [7]. This model utilizes ideas of Koch and Ullman's work [8], which supported the feature integration theory [16] and described biological evidence that attention is processed from individual feature maps that code for conspicuity throughout the entire visual field. The model proposed by Itti, Koch, and Niebur is a biologically-plausible purely feed-forward, bottom-up model of visual saliency for static scenes. Many recent models of visual saliency have used this model as a framework for their computational mechanisms. However, these early models computed saliency on static scenes, neglecting temporal effects when it exists within the visual field.

Selective attention is not only important for static visual stimuli, but also when motion exists within the visual receptive field. In this work we consider only first-order (opposed to second-order) motion, more specifically non-direction-specific temporal change. First-order motion is a function of spatio-temporal properties of image intensity over multiple frames [4]. Currently, there is no concrete understanding of the distinct location where first-order motion and second-order motion are processed [4]. There does exist increasing amounts scientific support that first-order motion is processed along the dorsal pathway, starting with simple cells in V1 (primary visual cortex) that encode information pertaining to locations of temporal change within the visual stimulus [4]. We utilize a model of these simple cells' receptive fields in this work.

Within the past decade, more research has been conducted towards developing saliency models that consider temporal information for computing saliency maps in response to dynamic visual stimulus. Studies have shown that models that utilize temporal information to compute saliency maps perform much better at predicting eye fixations than those that utilize only static features [9]. Itti [9] extended the model implemented by Itti, Koch, and Niebur [7] to include a motion and flicker (on/off) motion channel. In this model, motion was considered as being another "feature" which was included in the final summation of individual feature channels' attention maps to

978-1-4799-1471-5/13/$31.00 ©2013 IEEE

then form a master saliency map. Other models use the Itti, Koch, and Niebur model as a framework for their dynamic saliency models. Zhang et. al. designed a machine learning-based dynamic saliency model that uses natural statistics, which is learned from prior experience [10]. Itti and Baldi designed a Bayesian surprise-based model that uses Bayesian probabilistic principles and defines saliency as a deviation from expectation based on internal models of the visual world [11]. A graph-based visual saliency model was implemented by Harel et. al. and uses a similar approach to Itti, 2005 [9] but modifies the normalization process at the individual feature channels [12]. However, it is important to note these models discussed thus far are purely feature-based.

### *B. Object-based Saliency Models*

There has been an increasing amount of neurophysiological evidence that supports the idea that attention depends on structural organization of a scene into tentative objects [3]. In contrast to the feature integration theory, object-based attention supports Gestalt psychology, which states that whole objects are perceived prior to individual features. Previous visual saliency models have utilized an object-based approach [14],[15]. However, these models do not use the idea that saliency is a function of perceptual organization of a scene into tentative proto-objects. In the remaining of this paper, we present an object-based approach for computing saliency for dynamic scenes. We utilize the proto-object based model proposed by [5] and introduce an extension to this model by incorporating a motion-sensitive channel modeling temporal activity encoded in simple cells in V1. This is not considered an additional feature channel, but rather is an independent pathway that computes temporal information which modulates border ownership activity. In other words, computed proto-objects within the visual scene exhibiting motion are awarded with higher saliency prior to any normalization or linear summations of individual feature channels. This model is able to compute a dynamic saliency map capable of predicting eye fixations far better than chance.

## II. Proto-object Based Model without Motion Component

As previously stated, this model for computing dynamic visual saliency is an extension of the proto-object based saliency model implemented by [5]. This original model is a biologically-plausible, purely feed-forward proto-object based model for visual saliency of static stimuli. It utilizes the concept of grouping cells and border-ownership neurons that code for figure-ground relationships of the visual scene. This provides an estimate of the location and size of the proto-objects that exist within the visual scene. The grouping mechanism works by first extracting local edge information using simple and complex edge cells. ON-center and OFF-center center-surround filters are then used to extract locations of bright objects on dark background and dark objects on light background. The local edge information is then combined with the OFF-center and ON-center center-surround responses to compute border ownership responses. It is important to note that the ON-center and OFF-center center-surround responses equally impact border-ownership responses [5]. This step also includes a normalization operator that awards higher saliency to high center-surround activity and lower saliency to low center-surround activity. This normalization step accounts for the fact that biological responses of border-ownership cells are independent of figure-ground contrast polarity. Gestalt principles of continuity and proximity are then applied to the border-ownership responses to activate grouping cells. This is accomplished by using a max operator at each pixel for selecting the most active border ownership cell from a pool of all possible orientations. The grouping cell activity is computed by integrating the most active border ownership cell activity in a circular manner. The activity of the grouping cells provides an estimate of the location of proto-objects within the receptive field. Image pyramids are used throughout this process so that the computations are scale invariant [5].

Although this grouping mechanism provides an estimate for proto-object location, it only provides saliency information in regards to figure-ground relationships. By itself, it cannot award saliency to proto-objects unique in their feature composition. In order to account for this, concepts used by Itti et. al. [7] are used after grouping has been processed. Three different feature channels are used: intensity, color, and orientation. Within the color channel there are four color-opponency sub-channels (red-green, green-red, blue-yellow, yellow-blue). Within the orientation channel there are four orientation sub-channels ($0 - 3\pi/4$ in increments of $\pi/4$). The responses of each channel are then fed into the grouping mechanism to compute a grouping pyramid coding for saliency activity of proto-objects within the visual scene. The grouping pyramids are then fed into the framework seen in [7] where they are first normalized and merged into a single saliency map at each individual feature channel. They are then normalized across feature channels and linearly summed to form a final saliency map. This encodes saliency not only for figure-ground relationships but also saliency in regards to feature composition as a function of the proto-objects.

## III. Integration of Motion Component

We introduce a motion-sensitive channel integrated with the previously described model [5] for modulating border-ownership activity at the grouping level. The complete model can be seen in Figure 1. There is neurophysiological evidence that non-direction selective simple cells in the primary visual cortex encode temporal information of dynamic visual stimuli [13]. There are two types of these simple cells, strongly phasic and weakly phasic. Strongly phasic consists of a filter that has a strong excitatory component (strong positive lobe) and strong inhibitory component (strong negative lobe). In this model we utilize strongly phasic cells that exist within the magnocelluar pathway due to their high temporal resolution. Parkhurst [6] describes a biologically-plausible mathematical model for approximating the temporal profile $r(t)$ of such strongly phasic simple cell receptive fields. See Equation 1

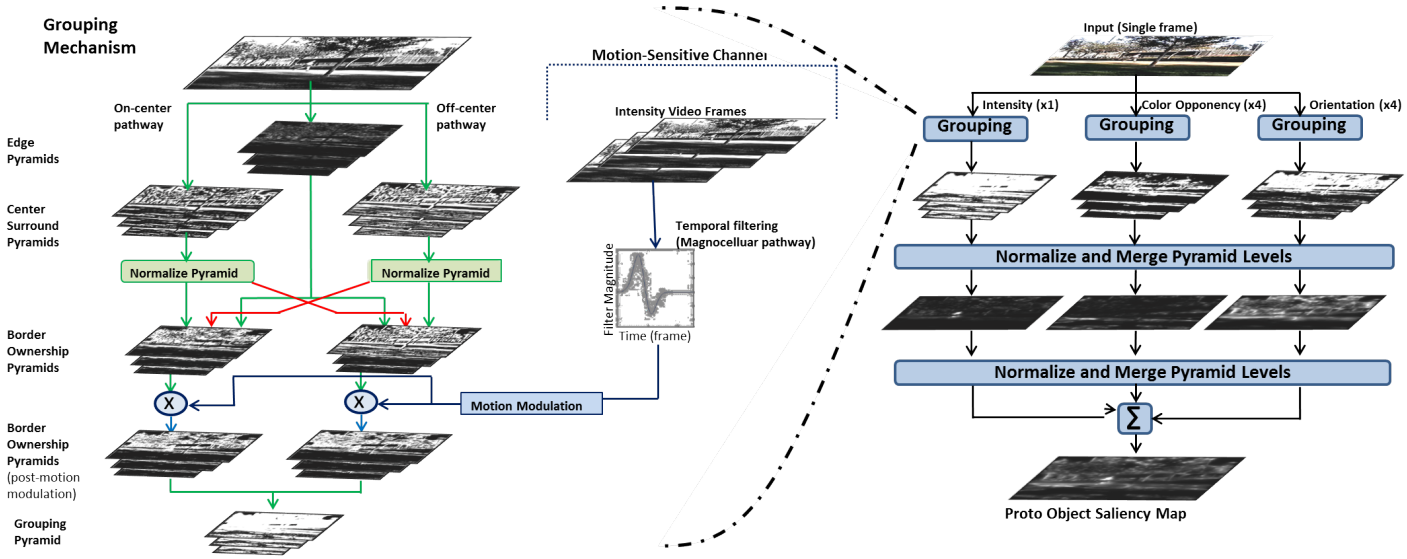


Fig. 1. Proto-object Based Model with Motion-Sensitive Channel (Green - Excitatory connections, Red - Inhibitory connections, Blue - Modulation)

below.

$$r(t) = \alpha(t - \tau - \delta)e^{\beta(t-\tau)^2}, \quad (1)$$

Here, $\alpha$ and $\beta$ are response amplitude parameters, $\tau$ is a time shift parameter and $\delta$ determines the degree to which the response is strongly phasic in time. These parameters were fit to the responses of strongly phasic simple cell characteristics observed by [13]. The resulting temporal filter, $r(i)$, was applied over the current frame and eight previous frames of the grayscaled video sequences and can be seen in Equation 2.

$$\begin{aligned} result(i) = &-0.0028 * frame(i) + 0.0759 * frame(i-1) \\ &+ 0.1681 * frame(i-2) - 0.0910 * frame(i-3) \\ &- 0.1073 * frame(i-4) - 0.0179 * frame(i-5) \\ &- 0.0084 * frame(i-6) - 0.0083 * frame(i-7) \\ &- 0083 * frame(i-8) \quad (2) \end{aligned}$$

In this equation, $i$ represents the current frame and $frame()$ is the grayscaled frame of the video sequence at a specific point in time. And the $result()$ (of same size as a single frame) at the current frame $i$ is the result of convolving $r$ along the temporal dimension of the frames.

The motion-sensitive channel operates only on the intensity channel. The temporal filter previously discussed is applied to the current frame and a fixed number of previous frames, at each scale/level of the intensity pyramid. The temporal responses are then normalized to a scalar value between 0 and 1, where 0 represents absence of any temporal activity and 1 represents highest possible temporal response. These temporal responses are then used to multiplicatively modulate the border-ownership activity prior to the final grouping step (see Figure 1). This affects the grouping mechanism by awarding a higher saliency (in regards to figure-ground relationship) to proto-objects exhibiting higher temporal activity, i.e. motion. There exists a motion weight parameter for the motion-sensitive channel. This parameter functions such that increasing the weight parameter proportionally increases the amount of modulation of the border-ownership activity induced by temporal activity from the motion-sensitive channel.

## IV. Experimental Results

To validate this model we compared it against eye fixation data from 4 arbitrary videos used in Itti et. al. [9] experiments. In these experiments, 8 subjects' eye fixation data was recorded for various video sequences. For this work we used 4 different subjects' fixation data for each video. The videos were displayed at a resolution of 640 x 480 with a frame rate of 30.1341 Hz. Here we compare our model to two state-of-the-art dynamic visual saliency models: Graph Based Saliency model [12] and the Itti, 2005 model [9]. A single frame of each model's computed dynamic saliency map on one of the videos used can be seen in Figure 2.

To evaluate the saliency models' ability to predict eye fixations, we use the Kullback-Leibler divergence (KLD) as a metric. To compute KLD, first, the 3x3 matrix (circular aperture of $5.6°$) of computed saliency values around the target location of a fixation point is obtained. A max operator over these nine saliency values is applied. We call the result of the max operator, $S_{sac}$. We then divide $S_{sac}$ by the max saliency value across the saliency map at the time of the fixation ($S_{max}$). This results in a ratio of the saliency value at the region of the eye fixation, to the max value of the saliency map at the time of the fixation ($S_{sac}/S_{max}$). We compute $S_{sac}/S_{max}$ for each fixation throughout a video giving us a distribution of saliency values for actual fixations. These same computations are then computed for uniformly chosen random fixations giving us a distribution of $S_{rand}/S_{max}$ values ($S_{max}$ values remain the same for both actual and random fixations). We then histogram the resulting values


Fig. 2. Original video overlaid with fixation data ('colored squares') and single frame of the computed dynamic saliency map for each model (Video: 'tvad04', Frame: 310/312)

for $S_{sac}/S_{max}$ and $S_{rand}/S_{max}$, giving us actual fixation distributions and random fixation distributions, respectively [9]. The KL divergence of the actual fixation distributions from the random fixation distributions in computed. As a baseline, we can then denote "chance" as a KL divergence value equal to 0. Henceforth, the greater the KL divergence value, the better.

Using each model's computed saliency map, for each video, each model's KLD was measured for each subject using the method previously described. Table 1 shows the results of the KLD measurements (scores). These results reflect the average KLD over all subjects' eye fixation data and further averaged over all 4 videos. These subjects were focused on the center of the screen prior to starting the video, causing the eye fixation data to have strong center bias at the initiation of the video. Top-down factors are also not considered in this model which also effect the subjects' eye movements. Henceforth, these KLD values may underestimate the models' true performance for predicting bottom-up attention. Nonetheless, this proto-object based model outperforms the Itti, 2005 model (+0.1009) and insignificantly differs from the GBVS model (+0.0165) in performance with respect to the computed KLD in this experiment. Moreover, this model performs significantly better than chance (chance being a KLD score of 0) in predicting eye movements. Significance was calculated using a paired t-test between the results of the proto-object model and the Itti, 2005 model, as well as between the proto-object model and the GBVS model. The $p$-values seen in Table I indicate that the results between the proto-object model and Itti, 2005 model are highly significant while the results between the GBVS model are much less significant. Such $p$-values are expected considering only 4 videos were used from the dataset. Nonetheless, considering only average KLD scores, these results are promising and by comparing the models against fixation data for the entire video dataset used in [9], we can further validate and compare the proto-object based model's performance.

TABLE I

AVERAGE KL DIVERGENCE AND SIGNIFIANCE OF PROTO-OBJECT KLD RESULTS IN COMPARISON TO OTHER MODELS (P-VALUE)

| ***Metric*** | **Proto** | **Itti,2005** | **GBVS** |
|---|---|---|---|
| *KLD* | 0.41902 | 0.31811 | 0.40257 |
| *p-value* | - | 0.0676 | 0.7884 |

## V. CONCLUSION

We have proposed a proto-object based model for computing dynamic visual saliency. It utilizes a motion-sensitive channel that computes temporal information within a dynamic visual scene. This temporal information is computed using a biologically-plausible temporal filter modeling properties of simple cells in the primary visual cortex. This temporal activity is then used to modulate the notion of proto-objects within the visual field. From our experiments, this model insignificantly differs in performance from the state-of-the-art feature-based GBVS dynamic saliency model [12] and outperforms the Itti, 2005 model [9]. Continuing work includes comparing our results of this model against all eye fixation data from the complete dataset of videos used in [9] and further, comparing our results to other dynamic visual saliency models including [10], [11], [14], [15]. Finally, we seek to implement this model in FPGA hardware to allow for real-time processing and a more ideal neuromorphic implementation.

## REFERENCES

[1] K. Koch, J. McLean, M. Berry, P. Sterling, V. Balasubramanian, and M. A. Freed, "Efficiency of information transmission by retinal ganglion cells," *Curr. Biol.*, vol.14, no.17, pp. 1523 - 1530, 2004.

[2] S. P. Strong, R. Koberle, R. R. de Ruyter van, Steveninck, and W. Bialek, "Entropy and information in neural spike trains," *Phys. Rev. Lett.*, vol.80, pp.197-200, Jan 1998.

[3] S. Mihalas, Y. Dong, R. von der Heydt, E. Neibur, "Mechanisms of perceptual organization provide auto-zoom and auto-localization for attention to objects," *PNAS*, vol.108, no.18, pp. 7583–7588, 2011.

[4] L. M. Vaina, S. Soloviev, "First-order and second-order motion: neurological evidence for neuroanatomically distinct systems," *Progress in Brain Research*, vol. 144, pp. 197–212, 2004.

[5] A. F. Russell, S. Mihalas, E. Niebur, and R. Etienne- Cummings, "Proto-object based visual saliency," In review, 2012.

[6] D. Parkhurst, "Selective Attention in Natural Vision: Using Computational Models to Quantify Stimulus-driven Attentional Location," Ph.D. dissertation, Dept. of Neuroscience, Johns Hopkins University, Baltimore, MD, 2002.

[7] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE PAMI*, vol.20, no.11, pp.1254–1259, 1998.

[8] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, vol.4, pp.219-227, 1985.

[9] L. Itti, "Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes," *Visual Cognition*, vol.12, no.6, pp.1093–1123, 2005.

[10] L. Zhang, M. H. Tong, G. W. Cottrell, "SUNDAy: Saliency Using Natural Statistics for Dynamic Analysis of Scenes,"emphThirty-first Annual Cognitive Science Society Conference, 2009.

[11] L. Itti, P.F. Baldi, "Bayesian Suprise attracts human attention," *Vision Research*, vol.49, pp.1295–1306, 2008.

[12] J. Harel, C. Koch, P. Perona, "Graph-Based Visual Saliency," *Advances in Neural Information Processing Systems*, vol. 19, pp.545–552, 2007.

[13] R. L. De Valois, N. P. Cottaris, L. E. Mahon, S.D. Elfar, J.A. Wilson, "Spatial and temporal receptive fields of geniculate and cortical cells and directioanl selectivity," *Vision Researh*, vol.20, pp.3685–3702, 2000.

[14] D. Walther, C. Koch, "Modeling attention to salient proto-obejcts," *Neural Networks*, vol.19, no.9, pp.1395–1407, 2006.

[15] M. Wischnewski, A. Belardinelli, W.X. Schneider and J.J. Steil, "Where to Look Next? Combining Static and Dynamic Proto-objects in a TVA-based Model of Visual Attention," *Cognitive Computation*, vol.2, no.4, pp.326–343, 2010.

[16] A.M. Treisman, G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol.12, no.1, pp.97–136, 1980.