Towards a Unified Multiresolution Vision Model for Autonomous Ground Robots

John Sustersic^{*1}, Brad Wyble², Siddharth Advani³, Vijaykrishnan Narayanan³

¹The Applied Research Laboratory
The Pennsylvania State University
University Park, PA 16804 USA
814.863.3015
jps263@psu.edu
* corresponding author

²Department of Psychology
The Pennsylvania State University
University Park, PA 16804 USA
bpw10@psu.edu

³Department of Computer Science and Engineering
The Pennsylvania State University
University Park, PA 16804 USA
{ska130, vijay}@cse.psu.edu

ABSTRACT

While remotely operated unmanned vehicles are increasingly a part of everyday life, truly autonomous robots capable of independent operation in dynamic environments have yet to be realized – particularly in the case of ground robots required to interact with humans and their environment. We present a unified multiresolution vision model for this application designed to provide the wide field of view required to maintain situational awareness and sufficient visual acuity to recognize elements of the environment while permitting feasible implementations in real-time vision applications. The model features a kind of color-constant processing through single-opponent color channels and contrast invariant oriented edge detection using a novel implementation of the Combination of Receptive Fields model. The model provides color and edge-based salience assessment, as well as a compressed color image representation suitable for subsequent object identification. We show that bottom-up visual saliency computed using this model is competitive with the current state-of-the-art while allowing computation in a compressed domain and mimicking the human visual system with nearly half (45%) of computational effort focused within the fovea. This method reduces storage requirement of the image pyramid to less than 5% of the full image, and computation in this domain reduces model complexity in terms of both computational costs and memory requirements accordingly. We also quantitatively evaluate the model for its application domain by using it with a camera/lens system with a 185 degree field of view capturing 3.5M pixel color images by using a tuned salience model to predict human fixations.

KEYWORDS

computer vision; Gaussian pyramid; visual salience; parallel algorithms;

1 INTRODUCTION

Models of visual saliency, as a surrogate measure of attention to indicate visually interesting image features, has been an active field of research in both computational neuroscience as well as in computer vision; sometimes symbiotically feeding on the other's findings, and sometimes moving ahead independently [1–5]. These saliency models have used low-level features to build information maps, which are then fused together to form what is popularly called a saliency map. A compilation of how these models performed, when compared against human fixations, show that they do remarkably well [6]. 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. In [7], the authors looked at adapting a model [4] to develop a biologically inspired multi-resolution framework for salience, which showed both quantitative and qualitative advantages. However the current approach offers a major advantage relative to [7], which is that the processing of the visual input through a series of maps can be precisely controlled. This top-down control permits parametric manipulation of the computation that can be

modulated based on the task at hand. The result is a computational algorithm that is both tunable, and computationally efficient, thus making it ideal for real-time, on-demand processing in an autonomous system.

The proposed algorithm is based on simple cells in the primary visual cortex that are believed to extract local contour information from a visual scene. This information serves as the building block of early vision, and is important for such tasks as determining salience, contour processing, object recognition, and scene gist determination. The CORF (Combination of Receptive Fields) model [8] considers a computational model of a simple cell as an alternative to the 2D Gabor function (GF) model [9]. The Scale Invariant Feature Transform (SIFT) is a well-known method used for object recognition using a Difference of Gaussian (DoG) approach to produce translation invariant feature vectors [10]. These models process visual input in a similar fashion as the earliest stages of other models that describe hierarchical processing of objects, such as HMAX [11], which simulates the visual ventral pathway and HOP [12], which recognizes complex objects by combining simpler parts into more complex representations. In [13], the authors show that recognition rates improve when incorporating attention in cluttered and crowded conditions. In [14], the authors discuss various state-of-the-art models with regards to a rapid visual categorization task and postulate that attention is important in un-trained environments.

The model described here incorporates a novel formulation of the CORF algorithm that is neuromorphic, tunable, and can support both attention and object recognition. We describe how this model may be used to determine high-salience regions of visual images as a form of attention, which is important in object recognition and scene understanding in humans.

2 THEORY

2.1 Single Opponent Color Difference of Gaussian Model

We begin by describing a formulation for the computation of Difference of Gaussian (DoG), defining center ($G_{c,\sigma}$) and surround ($G_{s,\sigma}$) Gaussians conventionally and constraining the ratios of inner and outer Gaussians to be 2 as found in electrophysical studies in mammalian LGN cells [8]:

$$G_{c,\sigma}(x, y) \stackrel{\text{def}}{=} \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (1)$$

$$G_{s,\sigma}(x, y) \stackrel{\text{def}}{=} \frac{1}{2\pi(2\sigma)^2} e^{-\frac{x^2+y^2}{2(2\sigma)^2}} \quad (2)$$

where σ is the standard deviation of the inner Gaussian. Let $P \in \mathbb{R}_+$, $P \leq P_{max}$ and let the input image $\mathcal{I}: \mathbb{N}^2 \rightarrow \mathbb{P}^3$. As in [15] we consider opponent color channels Red-Green (R-G), Red-Cyan (R-C), Yellow-Blue (Y-B), and White-Black (W-K), so we must project from RGB color space into the higher-dimension RGBCYW color space to compute these opponent color channels. Let $\Omega(r, g, b) = \langle r, g, b, 0.5g + 0.5b, 0.5g + 0.5r, 0.299r + 0.587g + 0.114b \rangle$ where $r, g, b \in P$. Thus $\Omega: \mathbb{P}^3 \rightarrow \mathbb{P}^6$ and describes how yellow, cyan and white (intensity) channels are derived from an RGB pixel (note that this differs from [15]). Thus $\Omega(\mathcal{I})$ denotes the image $\mathcal{I}$ in RGBCYW color space. We then convolve $\Omega(\mathcal{I})$ with both center and surround Gaussians to yield center and surround responses $G_{c,\sigma} * \Omega(\mathcal{I})$ and $G_{s,\sigma} * \Omega(\mathcal{I})$, respectively. Thus the input image $\mathcal{I}$ may be thought of as a discrete function mapping from pixel coordinate space ($\mathbb{N}^2$) to RGB, and $G * \Omega(\mathcal{I})$ is a function mapping pixel coordinate space to RGBCYW color space. Mathematically, $G * \Omega(\mathcal{I}): \mathbb{N}^2 \rightarrow \mathbb{P}^6$.

Let $[\cdot]$ be the index operator that selects a color channel by name (e.g. $\langle 255, 0, 0, 0, 0, 0 \rangle[R] = 255$). The center-on or positive DoG_σ^+ may then be defined as:

$$DoG_\sigma^+ \stackrel{\text{def}}{=} \left\langle \begin{array}{l} |G_{c,\sigma} * \Omega(\mathcal{I})[R] - G_{s,\sigma} * \Omega(\mathcal{I})[G]|^+ \\ |G_{c,\sigma} * \Omega(\mathcal{I})[R] - G_{s,\sigma} * \Omega(\mathcal{I})[C]|^+ \\ |G_{c,\sigma} * \Omega(\mathcal{I})[Y] - G_{s,\sigma} * \Omega(\mathcal{I})[B]|^+ \\ |G_{c,\sigma} * \Omega(\mathcal{I})[W] - G_{s,\sigma} * \Omega(\mathcal{I})[W]|^+ \end{array} \right\rangle \quad (3)$$

where $|\cdot|^+$ designates half-wave rectification ($|x|^+ = x \forall x \geq 0, 0 \text{ otherwise}$). This defines center-on DoG_σ^+ as a four-dimension color space consisting of red channel center minus the green channel surround, red center minus cyan surround, etc. Thus the DoG_σ^+ function may be thought of as a 3D convolution that transforms an image from RGB color space into a single-opponent color space $\langle R-G, R-C, Y-B, W-K \rangle$.

Unlike previous work [8], $DoG_\sigma^+ \neq -DoG_\sigma^+$ and is defined:

$$DoG_{\sigma}^{-} \stackrel{\text{def}}{=} \left\langle \begin{array}{l} |G_{c,\sigma} * \Omega(\mathcal{I})[G] - G_{s,\sigma} * \Omega(\mathcal{I})[R]|^{+} \\ |G_{c,\sigma} * \Omega(\mathcal{I})[C] - G_{s,\sigma} * \Omega(\mathcal{I})[R]|^{+} \\ |G_{c,\sigma} * \Omega(\mathcal{I})[B] - G_{s,\sigma} * \Omega(\mathcal{I})[Y]|^{+} \\ |G_{s,\sigma} * \Omega(\mathcal{I})[W] - G_{c,\sigma} * \Omega(\mathcal{I})[W]|^{+} \end{array} \right\rangle \quad (4)$$

Note that the definition of the white channel is conceptually the same as that used in the CORF model [8]; therefore one may think of this single-opponent color model as a generalization of CORF and we call it CCORF (Color CORF). To illustrate what is generated from this model, we present DoG responses for the test image as illustrated in Figure 1.

Figure 1 - Test image and single opponent color responses. From left to right on the top row: test image, Red-Green DoG, Green-Red DoG, Red-Cyan DoG, Cyan-Red DoG, Yellow-Blue DoG, Blue-Yellow DoG. The second row is the White-Black DoG and the Black-White DoG (Note minor level adjustments have been made to make image readable in print.)

2.2 Cropped Gaussian Pyramid

Figure 2 illustrates the concept of transforming an image into a ‘Cropped Gaussian Pyramid’ (CGP) in which pyramid levels are progressively cropped such that full resolution is preserved only in a narrow field of view fovea, and there is a progressive loss of acuity with increasing eccentricity due to increasingly sparse representations in larger levels of the pyramid. This yields a multi-resolution representation of the image in which the number of pixels in the CGP is much less than the input image. Conversely, traditional Gaussian Pyramids [16] (GPs) have more pixels than the image. For example, an imaging system producing 1920x1920 (3.51Mpix) images can be represented in a 6-level CGP in only 166.78Kpix, with 45% of those in the fovea¹, while a 6-level GP model would require 4.91Mpix.

Traditionally, GPs are constructed by successive (sequential) Gaussian blur and downsampling operations. In the downsampling step for imagery data, fully $\frac{3}{4}$ of the computational work done in blurring is discarded in current practice, but in a CGP an even larger percent of computational work will be discarded in the progressive cropping. We therefore separate computations of the levels and combine this with previously described (DoG) computation to limit total computational work to only required computations while enhancing parallelizability – significantly reducing the time complexity of CGP pyramid construction at the tradeoff of progressively reduced visual acuity in peripheral regions of the image. Thus the CGP may be thought of as a piece-wise linear approximation of biological foveated vision.

Each level of the pyramid is computed independently, directly from the source image. The 3D convolution of the source image with the previously described center and surround Gaussian kernels of this image in a level’s region of interest produces center-on and center-off single-opponent color channels, which are then half-wave rectified and compressed logarithmically. Divisive normalization completes the process for a level, but as we are in the log domain this is computed as simple subtraction. Repeating the process for each pyramid level completes the construction of the CGP and processing illustrated in Figure 2.

2.3 Oriented Edge Detection using MR-CCORF

In the original CORF model, two sets of collinear subunits taking inputs from the center-on (DoG⁺) and center-off (DoG⁻) units were used to define a CORF cell, with the locations of the subunits corresponding to the intersection of concentric circles and maximal responses of DoG⁺ and DoG⁻ units in the presence of the preferred stimulus [8]. One line of subunits operates on DoG⁺ and the other operates on DoG⁻, the preferred stimulus orientation and location is collinear with these lines and equidistant to them, and the weights on the subunit were distributed normally about the CORF cell center. The authors use two concentric circles but note that the model can be generalized to any number; we note that one may generalize this model further to a two-parameter Gaussian where the 2D Gaussian kernel is ‘split’ along the centerline of the CORF cell. Let θ be the cell’s preferred orientation. Let $\sigma_{\parallel}$ be the standard deviation of a Gaussian weighting function in the direction parallel to the cell’s preferred orientation and $\sigma_{\perp}$ that perpendicular. We define the excitatory CCORF kernel:

$$G_{\theta, \sigma_{\parallel}, \sigma_{\perp}}^{+}(x, y) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{2\pi\sigma_{\parallel}\sigma_{\perp}} e^{-\left(\frac{(x\cos\theta - y\sin\theta)^2}{2\sigma_{\parallel}^2} + \frac{(x\sin\theta + y\cos\theta)^2}{2\sigma_{\perp}^2}\right)} & \text{if } x\sin\theta + y\cos\theta < 0 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

The inhibitory CCORF kernel is then defined:

¹ In humans, ~50% of cortical resources processing retinal information are allocated to the fovea.

Figure 2 - Schematic diagram of preprocessing in the multi-resolution color CORF pyramid. The model is designed to approximate tradeoffs found in biological vision systems between visual acuity and field of view and approximates processing in retina and LGN. We use half-wave rectification after [8] which performed ~16% better w.r.t saliency than half-square rectification. The 3.51Mpix image compresses to a representation of only 166.78Kpix.

$$G_{\theta, \sigma_{\parallel}, \sigma_{\perp}}^{-}(x, y) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } x \sin \theta + y \cos \theta < 0 \\ \frac{1}{2\pi\sigma_{\parallel}\sigma_{\perp}} e^{-\left(\frac{(x \cos \theta - y \sin \theta)^2}{2\sigma_{\parallel}^2} + \frac{(x \sin \theta + y \cos \theta)^2}{2\sigma_{\perp}^2}\right)} & \text{otherwise} \end{cases} \quad (6)$$

Following [8], we compute CCORF unit response as the weighted geometric mean of its subunits; however we take the DoG^+ and DoG^- directly as excitatory and inhibitory subunits, respectively. Recall that a weighted geometric mean is equivalent to an arithmetic mean in the log domain:

$$\left(\prod_{i=1}^n a_i^{\omega_i}\right)^{1/\sum \omega_i} = e^{\left[\frac{1}{\sum \omega_i} \sum_{i=1}^n \omega_i \ln(a_i)\right]} \quad (7)$$

Furthermore, note that this corresponds to taking dot products of the excitatory and inhibitory kernels with the log of the corresponding DoG. Therefore, computing the CCORF responses as a weighted geometric mean for the image can be expressed as the convolution of excitatory and inhibitory kernels with the logarithms of DoG^+ and DoG^- , respectively, and noting that $\int \int \left(G_{\theta, \sigma_{\parallel}, \sigma_{\perp}}^+(x, y) + G_{\theta, \sigma_{\parallel}, \sigma_{\perp}}^-(x, y)\right) dx dy \stackrel{\text{def}}{=} 1$, we define the CCORF operator:

$$\text{CCORF}_{\theta, \sigma_{\parallel}, \sigma_{\perp}} = e^{\left[G_{\theta, \sigma_{\parallel}, \sigma_{\perp}}^+ * \ln(\text{DoG}_{\sigma}^+) + G_{\theta, \sigma_{\parallel}, \sigma_{\perp}}^- * \ln(\text{DoG}_{\sigma}^-)\right]} \quad (8)$$

It is significant to notice that the CCORF operator like the original CORF operator is an oriented edge detector, not an oriented line detector – thus the response for any θ is not equal to the response for $\theta \pm \pi$.

Figure 3a illustrates how these single-opponent color channels are used in the computation of oriented edges using this novel CORF operator, the computation of color and oriented edge conspicuity, and bottom-up saliency. As before, each level of the pyramid may be computed independently. At each level, the 3D convolution of the oriented CCORF kernels with the single opponent color channels produces the response for each kernel, generating a set of log-domain oriented edge responses for the input image. The net response is computed by simply pooling across all orientations:

$$CCorf_{\sigma_{\parallel}, \sigma_{\perp}} \stackrel{\text{def}}{=} \sum_{\theta} CCorf_{\theta, \sigma_{\parallel}, \sigma_{\perp}} \quad (9)$$

Since oriented edge responses are in the log domain, exponentiation of responses for individual orientations is performed before taking the sum. Repeating these computations for each pyramid level completes the computation of oriented edge responses for MR-CCORF. An example of a pair of partitioned Gaussians is shown in Figure 3b.

Figure 3 includes an example of the various intermediates and results of this computation. Examining the third pyramid level from the input image in 3a, figures 3c and 3d show DoG^+ and DoG^- responses for W-K as weakly oriented edge detectors, and figures 3e, 3f, and 3g show the $CCorf_{\theta, \sigma_{\parallel}, \sigma_{\perp}}$ response for $\theta = 0, \pi/4, \pi/2$, respectively, and $\sigma_{\parallel} = 4, \sigma_{\perp} = 0.6$. Kernel size is 4×4 , and the net response is pooled over 16 orientations $\Theta = \{0, \pi/8, \pi/4, \dots, 15\pi/8\}$. Figures 3l and 3m show DoG^+ and DoG^- responses for the R-G channels with the RGB region of interest shown for comparison in 3i along with the red-only channel in 3j and green-only channel in 3k. Note that DoG provides significant noise immunity in addition to color constancy. Figure 3n is the net $CCorf$ R-G response. Representative CCORF responses for the test image of Figure 1 are shown in Figure 4 for the four cardinal orientations.

To illustrate the ability of this formulation of CCORF to localize oriented edges, we evaluated CCORF responses to a set of white-black test stimuli with oriented edges of 0, 14, 25.9, and 45 degrees using 72 CCORF units spanning 360 degrees (i.e. 5 degree increments; test stimuli show in 5c). We plotted the CCORF response for each stimulus at $x=128$ over a range of y near $y=128$ to demonstration translational localization in addition to responses over the set of CCORF units to demonstration rotational localization. Highlights of these results are illustrated in Figure 5, and show the CCORF reliably shows peak response near the actual applied stimulus orientation and precise translational localization. Recalling that in pixel space edges with orientations of 0, 14, 25.9 and 45 correspond to slopes of 0:4, 1:4, 2:4, and 4:4, respectively, a kernel of size 4×4 's precise response to orientation will vary depending upon the precise local structure of the stimulus. It should therefore not surprise that for the shallow 14 degree orientation the 4×4 kernel found the orientation to be zero degrees. However the 6×6 kernel (response shown in Figure 5) does correctly resolve the applied stimulus orientation. Interestingly, the 25.9 degree stimulus response peaked at 35 degrees. We repeated this test for kernel sizes up to 16×16 in addition to varying the ratio of semi-major to semi-minor axis on the CCORF kernel, and obtained the same results. These results suggest that the CCORF formulation is approaching the information-theoretic limits in resolving oriented edges in pixel imagery. By horizontal and vertical symmetries, these results from 0 to 45 degrees sufficiently describe the CCORF performance over the full range of 360 degrees. Furthermore, these results show that 4×4 kernel sizes are sufficient for oriented edge detection at the intervals we use in our model – 16 oriented edges over 360 degrees, though the results suggest a local pooling mechanism might outperform winner-take-all classification of the angle of applied stimulus. Finally, the results suggest that still better oriented edge localization resolution may be achieved using larger kernel sizes, particularly if combined with local pooling.

levels, even the larger 16x16 kernel cannot correctly discern the applied stimulus (figures 5e and 5f). The intermediate blurring done in the original CORF model would likely help resolve the lower frequency signal from the higher frequency noise, but we note that in a multiresolution application such as this, these blurred, lower-resolution images would naturally be computed at subsequent levels of the pyramid – this was the original motivation for omitting the blurring layer from this model. This suggests that cross-level excitatory inputs to a pooling, orientation classification layer that uses local excitatory and inhibitory inputs would resolve oriented edges in even noisier stimuli. Antiphase inhibition as demonstrated in subsequent work on the original CORF model [20] would be simple to integrate to this model and as shown in the literature is expected to improve edge detection in high signal-to-noise ratio applications.

2.4 Multi-resolution Color Conspicuity

Multi-resolution color conspicuity is computed first by pooling across the single-opponent color channels (i.e. DoG^+ and DoG^- responses), and then pooling across pyramid levels. Let $A \stackrel{\text{def}}{=} \langle \alpha_0, \dots, \alpha_7 \rangle$ where $\alpha_i \in \mathbb{R}, 0 \leq \alpha_i \leq 1$ and $\sum \alpha_i = 1$. Let single-resolution color conspicuity be defined as $s_{c,A}(x, y) \stackrel{\text{def}}{=} [\text{DoG}^+(x, y) \text{ DoG}^-(x, y)] \cdot A$ where $[\text{DoG}^+ \text{ DoG}^-]$ represents the concatenation of the two P^4 vectors into P^8 . Thus $s_{c,A}: P^4 \times P^4 \rightarrow P$.

Pooling across the pyramid levels is complicated by the fact that the levels do not represent the same spatial regions – that is, while all levels have pixels that map to the fovea, only the lowest resolution level have pixels that map to the extreme periphery. A simple pooling method for this application is to sum across the levels and normalize by the number of levels that participate in the summation. We note, however, that human fixations tend to be center-weighted – so much so that a naïve salience model consisting of a centered Gaussian is used as a benchmark and often performs quantitatively better than sophisticated salience models [5-6]. The question of whether or not this center-biasing is an artifact of eye-tracking experiments, a deep statistical property of the structure of natural images, or some combination thereof, we do not need address; we merely observe the well-established property that human fixations of images exhibit this characteristic, thus salience algorithms should have this characteristic as well. We observe that a pooling function that simply sums across the levels and normalizes by the total number of levels would be advantageous for at least some applications such as bottom-up salience as the CGP naturally has more pixel representation in the fovea: all levels have pixels to contribute to the fovea, where only the lowest resolution level has any pixel representation of the periphery.. Without knowing a priori how much foveal weighting is best for a particular application, we define a single-parameter linear weight model. Let $L \in \mathbb{N}$ be the number of levels in the pyramid and $l \in \{0, 1, \dots, L-1\}$ be the levels where $l=0$ corresponds to the fovea. We define the weight of a level as a function of $m \in \mathbb{R}_+$:

$$f(m, l) = \frac{ml+1}{\sum_{l=0}^{L-1} ml+1} = \frac{ml+1}{L+0.5mL(L-1)} \quad (10)$$

Conspicuity may then be computed as the weighted sum across the levels, using traditional upsampling of the lower resolution levels. For color conspicuity: $\mathcal{C}_{c,m_c}(x, y) \stackrel{\text{def}}{=} \sum_{l=0}^{L-1} f(m_c, l) \cdot s_{c,A}(x, y)$, where $s_{c,A}(x, y) \stackrel{\text{def}}{=} 0$ outside a level's region of interest (ROI). Thus when $m=0$ the weighting across levels effectively becomes the aforementioned normalization by the number of levels; as m grows, the weighting shifts progressively towards the lower resolution levels. Note however that since the fovea has representation in all levels, even for large m the fovea still is emphasized though conspicuity resolution at the fovea drops as the lower resolution levels dominate the conspicuity expression.

2.5 Multi-resolution Oriented Edge Conspicuity and Saliency

Oriented edge conspicuity is computed similarly: $\mathbb{C}_{oe,m_{oe}}(x,y) \stackrel{\text{def}}{=} \sum_{l=0}^{L-1} f(m_{oe}, l) \cdot \text{CCor}f_{\sigma_{\parallel}, \sigma_{\perp}}(x,y)$, where $\text{CCor}f_{\sigma_{\parallel}, \sigma_{\perp}}(x,y) \stackrel{\text{def}}{=} 0$ outside a level's ROI. Finally, saliency may be computed as a weighted linear combination of color and oriented edge conspicuity: $\mathcal{S}(x,y) = \alpha_c \mathbb{C}_{c,m_c}(x,y) + \alpha_{oe} \mathbb{C}_{oe,m_{oe}}(x,y)$. Thus saliency may be computed using a multi-resolution CGP and may be thought of as a kind of computation performed in a compressed domain. Note that unlike other models such as Itti-Koch [1-2], intensity is not explicitly used in the computation of saliency. We found that when averaged over all images in the MIT training data, the 2D covariance between intensity and fixation map is approximately zero. We therefore omit the intensity channel from our model.

3 METHODOLOGY

We evaluate the MR-CCORF model by quantifiably comparing it against the current state-of-the-art for saliency algorithms using the MIT saliency benchmark [6]. First, we must determine which parameters to use in the model. We define the parameters for the CGP by fitting a model to the 1920×1920 images produced by our fisheye-lens 185° FOV camera system. There are several ways to do this, and we consider two methods. In the first method, we allow the field of view subtended by a pyramid level to grow linearly (constant rate) as the lineal resolution of the levels is downsampled by 2. Beginning with a fovea radius of 120 pixels, the 1920^2 image fits an 8-level model, however the lowest resolution levels become very small (on the orders of tens of pixels square), so we stop at 6 levels as illustrated in figures 1 and 2. In the second method, we allow the field of view to grow exponentially such that the size of the pyramid levels is fixed at a radius of 120 pixels. This yields a 4-level model for 1920^2 images. We select the 6-level model based on the quantitative advantages of fewer number of pixels in the CGP (166.78Kpix in 6-level vs. 225Kpix in 4-level models), fast computational time ($\propto$ to number of pixels in CGP), and better parallelizability.

To maximize the detail of the DoG operators, we choose $\sigma_0 = 0.25$ (effectively a 2D Dirac delta kernel). Subsequent CGP levels are then bound by the relation $\sigma_i = 2\sigma_{i-1}$ such that the outer Gaussian at one level is identical to the inner Gaussian at the next level. For the CCORF operator parameters, 4×4 kernels are used with $\sigma_{\parallel} = 4, \sigma_{\perp} = 0.6$, and we consider 16 orientations $\Theta = \{0, \pi/8, \pi/4, \dots, 15\pi/8\}$. Other configurations were evaluated and we found that these parameters qualitatively superior in response and quantifiably better in computational performance. 8 kernels are shown in Figure 4 with DoG^+ and DoG^- components shown as positive and negative values, respectively.

3.1 Metrics

We evaluate the MR-CCORF model by quantifiably comparing it against the current state-of-the-art for saliency algorithms using the MIT saliency benchmark [6]. At the time of writing this paper, the scoring metrics used in this benchmark included Receiver Operating Characteristics (ROC - also commonly known as Area Under Curve), Similarity and Earth Mover's Distance (EMD).

The ROC has been a standard metric in the past to evaluate the performance of saliency models by plotting true positives versus false positives. However, [17] provide a poignant example of how the ROC for two models can be very similar even though their false alarms may be significantly different. Thus, the ROC curves fail to penalize models as a function of the spatial deviation between the predicted saliency map and the actual fixation map.

To address this, the Similarity metric takes into account spatial discrepancies when evaluating different models. Mathematically, the similarity S between two maps A and B is $S = \sum_{i,j} \min(P_{i,j}, Q_{i,j})$, where $\sum_{i,j} P_{i,j} = \sum_{i,j} Q_{i,j} = 1$. Here $0 \leq S \leq 1$ where $S = 0$ implies that the two maps are completely different while $S = 1$ indicates that the maps are identical.

The EMD metric was first introduced as a metric for image retrieval in [18]. It relates to the minimal cost that must be paid to transform one distribution to another and can be modeled as a transportation problem solved using linear optimization, but is computationally expensive. In [19], the authors introduce $\widehat{\text{EMD}}$, a linear time variant of the EMD. Mathematically, $\widehat{\text{EMD}}$ is represented as

$$\widehat{\text{EMD}}(P, Q) = \left(\min_{\{f_{i,j}\}} \sum_{i,j} f_{i,j} d_{i,j} \right) + \left| \sum_i P_i - \sum_j Q_j \right| \times \max_{i,j} \{d_{i,j}\}$$

$$\text{s.t. } f_{i,j} \geq 0, \sum_j f_{i,j} \leq P_i, \sum_i f_{i,j} \leq Q_j, \sum_{i,j} f_{i,j} = \min(\sum_i P_i, \sum_j Q_j)$$

where each $f_{i,j}$ denotes the amount transported from the i th supply to the j th demand. $d_{i,j}$ is the ground distance between bin i and bin j in the distribution. In the context of saliency, a larger EMD indicates a larger difference between the fixation map

and the predicted map while an EMD of zero indicates that the maps are identical.

(a)

(b)

Figure 5 – Localization capability of oriented edge detection using CCORF. Figure 5(a) illustrates CCORF responses for 72 oriented edges (5 degree increments) to an applied white-black contrast area oriented at 14 degrees, with edge passing through $x=128, y=128$. Linear-domain CCORF response shown in main graph with peak response at orientation of 15 degrees, precisely at $y=128$. The inset in figure 5(a) shows the responses in log domain to illustrate the discrimination of oriented edges in presence of aliasing effects using small kernels.

Figure 5(b) illustrates normalized peak responses to different stimuli with orientations of 0, 14, 26, and 45 degrees (shown in 5c, top 4 images). Bottom 2 images in 5c show noisy 45 degree stimuli, 5d-f show CCORF responses. Note that the in the inset to 5a for which there is no noise, only aliasing artifacts are observed which is in stark contrast to the noisy responses in 5e-f. See text for discussion.

Overall, a good saliency model would have a high ROC score, a high Similarity score and a low EMD score. While we do use these three metrics to evaluate MR-CCORF, we consider each of these metrics equally important in weight when it comes to ranking one model higher than the other. Thus, a saliency model would be ranked higher than another one if it were better on any two of the three metrics considered here.

4 EXPERIMENTAL RESULTS

To determine good parameters to use for salience applications, we performed a sensitivity analysis of MR-CCORF against the MIT saliency training dataset which consists of 1003 images of indoor and outdoor scenes and human fixation data from 15 subjects [5], thus training the MR-CCORF model on the MIT saliency training set. Observing that the W-K and K-W channels carry weakly oriented edge information much the same as the net *CCorf* response, we define the single-opponent pooling weight, A , to be a uniform distribution over the six channels carrying color information. Then beginning with $\alpha_c = \alpha_{oe} = 0.5$ we evaluated the salience model against the MIT training data for $m_c = \{0, 0.25, 1, 10, 1000\}$ and $m_{oe} = \{0, 1, 10, 1000\}$. The results are illustrated in Figure 4 and show that $m_c = 0$ is the argmax for similarity and area under the ROC curve metrics and the argmin for the Earth Mover’s Distance (EMD) metric, while the optimal value of m_{oe} varies somewhat across the three metrics with $m_{oe} = 0$ is argmax of the similarity metric, $m_{oe} = 1$ is argmax of ROC metric and $m_{oe} = 0.1$ is the argmax of EMD. Next we took $m_c = m_{oe} = 0$ and explored the parameter space over a range of distributions between α_c, α_{oe} and found that the optimal distribution for all three metrics is $\alpha_c = 0.19, \alpha_{oe} = 0.81$ for $m_c = m_{oe} = 0$ as shown in Figure 6. Finally we set $m_c = 0, \alpha_c = 0.19, \alpha_{oe} = 0.81$ and explore the parameter space between $m_{oe} = 0$ and 1 and find that $m_{oe} = 0, m_c = 0, \alpha_c = 0.19, \alpha_{oe} = 0.81$ appears to be an optimal point in the parameter space searched with respect to the MIT saliency training dataset for the three metrics of interest. Figure 6 illustrates quantitative model performance.

4.1 Model Performance at Predicting Human Fixations

Figure 7a illustrates model performances for MR-CCORF, MR-AIM [7] and AIM [4], and shows that MR-CCORF scores better on the EMD and Similarity metrics and is competitive with MR-AIM on the ROC metric for this data set. We then evaluate the trained MR-CCORF model against the MIT saliency benchmark and compare performance of MR-CORF with AIM, MR-AIM, and Itti-Koch [1-2] models. The results are shown in figure 7b and show that when trained against the MIT training data, the MR-CCORF set outperforms the others on the EMD metric on two of the three datasets and is competitive on the other metrics.

We further evaluate the model for its application domain by quantitatively comparing its performance with that of the AIM [4] and MR-AIM [7] saliency models using a dataset of 175 images of indoor and outdoor scenes collected with the 185° FOV lens and GigE camera system on our robotic testbed and human fixation data collected on these images. Fixation data were collected from 21 human participants, aged 18-35 using an Eyelink 1000 eye tracker that was calibrated to an average error of under 1 degree using 13 points. Subjects were placed in a chinrest, 60 cm from a 24” display. Images were displayed at 1152 x 1152 pixels, subtending 30 degrees. Eye data were collected for 3.5 seconds per image. To encourage vigilance, participants were instructed that they would periodically be asked to describe the image they just had just seen in 3-5 words. This occurred randomly on 30% of trials. Images were randomly ordered for each participant. Each data point collected during the presentation period, sampled at 1000hz, was used as the center of a Gaussian distribution with a standard deviation of 60. These distributions were overlaid to compute salience heat maps for model verification. Comparative model performance for MR-CCORF, MR-AIM and AIM are shown in figure 7c and indicated that MR-CCORF outperforms both models on EMD and Similarity metrics and is competitive with MR-AIM on the ROC metric.

4.2 Computational Complexity

Having quantitatively compared the performance of MR-CCORF with state-of-the-art models in the task of predicting human fixations, we complete the comparison by analyzing the computational costs of the models to show that MR-CCORF provides competitive performance while requiring orders of magnitude less storage and computation. First we consider the computational cost of constructing a Gaussian image pyramid from 2560x1920 images. We assume the default $(2\sigma) 5 \times 5$ Gaussian kernels are convolved with the preceding level’s image, and downsampling by two operations as used in common image processing packages, and we compute the number of operations (multiplications and additions) required to compute each pyramid level. 2D convolution is well known to be $\mathcal{O}(J_w J_h k_w k_h)$ where J_w and J_h are the image height (in this analysis, 1920) and image width (2560), respectively, and k_w, k_h are the height and width of the kernel. It is also well known that the Gaussian kernel is separable with complexity $\mathcal{O}(J_w J_h k_w + J_w J_h k_h)$, so we consider the separable method while noting that MR-CCORF can take greater advantage of the separability of the Gaussian kernels as the model can require large kernels. CGP construction is considered as described herein for the same input image size (2560x1920), and we consider both unseparated and separated computation. The results are illustrated in figure 8.

Figure 6 - Quantitative MR-CCORF model performance over a range of parameters on the MIT saliency training dataset. For Earth Mover's Distance metric, smaller is better; for Similarity and Area under ROC Curve metrics, larger is better. See text for discussion.

Figure 8a and 8b show total operations per level for GP and CGP computations considering separable convolution, respectively. While the computational costs of the GP decrease exponentially through the levels, computational costs increase exponentially for the CGP. Though the number of pixels in each level decrease through the levels in the CGP, the kernel size grows exponentially and dominated computational costs. The case is more interesting for the separated formulation as shown in figure 8b. Recalling that inseparable 2D convolution is quadratic in the size of kernel while separable 2D convolution is linear in kernel size, separating the Gaussian kernel has a greater impact in the CGP, significantly mitigating the costs of the larger Gaussian kernels and creating interesting dynamics in the competing trends of reducing pixels per level and increasing kernel sizes through the levels. Finally, figure 8f shows the GP and CGP storage requirements normalized by the source image size. Since the GP contains the full image as its first level, it is clear that the number of pixels in the GP must be greater than that of the source image; however the CGP contains only a small fraction of the full-resolution input image as part of its representation, and the total number of pixels in the CGP is an order of magnitude less than the source image.

The total number of operations is illustrated in figure 8e and show that CGP computation requires 2/3 the number of operations as GP computation when non-separated convolution is used; when separated kernels are considered, CGP computation is only 11.78% that of GP.

Next we consider the computational costs of edge detection in Itti-Koch style GP models and MR-CCorF. We assume 4x4 Gabor kernels for the Itti-Koch style GP model and consider the 4x4 CCorF kernels described in this work. Since both kernels are inseparable and have the same size, it follows that computational cost will be proportional to the number of pixels in the corresponding image pyramid. Figure 9a shows the computational costs of edge detection both in total and for each level, and as one might expect the total computational cost is an order of magnitude less for the CGP. Note that there is no requirement to use CCorF kernels on the CGP – one may choose to use Gabor kernels on a CGP or CCorF kernels on the GP; the computational cost of the Gabor on a CGP would be the same as using CCorF on a CGP and vice-versa. Thus the real computational advantage of the CGP is the order reduction in the number of pixels in the model.

Finally, we consider the computational costs of pyramid generation and feature extraction for Itti-Koch, MR-AIM and MR-CCorf saliency models. For Itti-Koch and MR-CCorf models, the feature extraction is the edge detectors (Gabor and CCorf kernels, respectively); for MR-AIM the feature is an (inseparable) 21x21 template, and following [7] we consider 25 templates. For pyramid construction in all models we assume separable convolutions are used, and as noted that all models have inseparable feature kernels. Figure 9b summarizes computational and memory (storage) complexity of these three models normalized by the Itti-Koch model complexity. While MR-AIM has a slightly reduced storage and pyramid generation complexity over Itti-Koch, it has a significantly higher computational cost of feature extraction due to the large, inseparable feature kernels used. MR-CCorf not only has significantly reduced storage complexity (3.48% that of the GP used in Itti-Koch) but also a total computational complexity only 6.64% that of Itti-Koch and only 1.46% that of MR-AIM. Thus MR-CCORF provides quantitatively comparable results at predicting human fixations in 3 different image datasets at a small fraction of the computational and storage complexity of other saliency models.

4.3 Parallelizability

While convolutional algorithms have great potential for fine-grained parallelization as each dot product of a kernel with some data can be computed concurrently with others, the MR-CCORF algorithm described here has significant advantages for additional, coarse-grained parallelization. Traditional GPs are computed using sequential ‘blur and downsample’ computations; necessarily this constitutes a true data dependence between pyramid levels. However in the CGP described here, each level is dependent only on the input image itself and thus each level may be computed completely in parallel. Recalling the data flow described in figures 2 and 3, this corresponds to horizontal data flows and means that from input image through CGP level computation and edge detection, there are no inter-level dependencies. Once all the levels have been computed, then summary information may be computed across levels (vertical data flow in figures 2 and 3), however these summary computations do not depend on each other and may also be computed in parallel.

Finally, salience is computed as a summary of summaries (another horizontal data flow), and this computation may be

Figure 7 - MR-CCORF, MR-AIM, and AIM saliency model performance in predicting human fixations over 3 different data sets. (a) MIT training data set. (b) MIT saliency benchmark. (c) Penn State set: 21 subjects, 175 images with 185 FOV in indoor and outdoor scenes

Figure 8 - Computational cost in image pyramid construction. (a) 6-level GP pyramid computation per level (b) CGP pyramid computation per level. Inset is a magnified view of the same. (c) GP and CGP computation cost comparison. (d) Comparing image pyramid storage requirements normalized by image size. Y-axis in (a-c) are in units of number of operations.

pipelined with the preceding horizontal dataflow. Thus the architecture of the MR-CCorf model provides both fine and coarse-grained parallelizability that is easily realizable using multithreading programming methods and classic barrier-based thread synchronization, and this model promises to yield even greater performance potential when parallelized on GPU or FPGA for low-power embedded systems like vision-guided robots.

5 CONCLUSIONS AND FUTURE WORK

We present a unified multiresolution vision model featuring color-constant processing through single-opponent color channels and contrast invariant oriented edge detection using a novel implementation of the Combination of Receptive Fields model. The model provides a color and edge-based salience assessment, as well as a compressed color image representation suitable for subsequent object identification and other later-vision processing. We find that visual salience can be computed from a cropped Gaussian pyramid with significant reductions both in computational and memory complexity, and that the quantitative performance of the algorithm is competitive with the current state-of-the-art. Moreover, the structure of this algorithm naturally exposes both fine grain and coarse grain parallel computation. Using this biomimetic model of early cognitive vision, we will next explore novel conspicuities (e.g. stereo disparity), task-specific forms of conspicuity and later cognitive vision processing (e.g. object detection) for resource-constrained embedded systems such as autonomous ground vehicles.

While the performance of this model shows it to be a viable, computable model of early vision for autonomous vehicles there are many opportunities to extend this work. While MR-CCORF provides competitive performance with MR-AIM, outperforming on the EMD metric for 2 of 3 data sets, we believe that it is the relative shape of the spatial prior implicit in the published MR-AIM model (a very coarse 2D spatial Gaussian) relative to that used in this formulation of MR-CCORF

Figure 9 - (a) Comparing computational costs of edge detection for 4x4 Gabor kernels on a GP and 4x4 CCORF kernels on a CGP. (b) Computational and storage complexities of Itti-Koch style (GP-based), MR-AIM, and MR-CCORF early cognitive vision models.

(coarse 2D cone) vice the methodology that slightly favors MR-AIM in some cases. We will test this hypothesis in future work. Antiphase inhibitory effects as shown in the original CORF model [20] would be simple to implement and explore, offering improved performance in high signal to noise ratio applications; combined with more complex excitatory and inhibitory local pooling models it would be expected to not only improve edge detection in the presence of noise but also more reliably detect and classify edge stimuli. Such improved performance in early cognitive vision models is reasonably expected to yield improvement in downstream visual tasks such as object recognition, particular in real vision applications required in robotics applications.

6 ACKNOWLEDGMENT

The authors wish to thank Brian Smith for his insightful review comments, Garrett Swan, Tilke Judd and Michael Moore for their help in model evaluation, John Koch, Kevin Fisher, Amar Rasheed and Dylan Levan for their work in collecting datasets. 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.

7 VITAE

Dr. John Sustersic is a Research Associate with the Applied Research Laboratories Autonomous Control and Intelligent Systems Division at The Pennsylvania State University and the Director of its Cognitive Perception and Computational Intelligence Laboratory. John earned his degree in Electrical Engineering from the Cleveland State University before continuing at the Pennsylvania State University, earning his Ph.D. in Computer Science and Engineering in 2009. John studies real-time cognitive perception and decision-making systems for autonomous robots in dynamic, unconstrained environments.

Dr. Brad Wyble is Assistant Professor of Psychology at the Pennsylvania State University. Brad earned his B.A. in Computer Science at Brandeis University in 1995 before his Masters (1999) and later his Ph.D. in Psychology from Harvard in 2003. Brad studies the interaction of visual attention and memory systems within the brain, using a mixture of computational modeling, behavioral experiments, and electrophysiology.

Mr. Siddharth Advani is currently pursuing his Ph.D. in Computer Science & Engineering at the Pennsylvania State University. He is a Research Assistant in the Microsystems Design Lab studying hardware-software co-design for high performance embedded systems. Siddharth earned his B.S. in Electronics Engineering at the Pune University in 2005 and his M.S. in Electrical Engineering at The Pennsylvania State University in 2009.

Dr. Vijaykrishnan Narayan is Professor of Computer Science & Engineering and Electrical Engineering at the Pennsylvania State University. Vijay received his Bachelors in Computer Science & Engineering from University of Madras, India in 1993 and his Ph.D. in Computer Science & Engineering from the University of South Florida, USA, in 1998. He is the co-director of the Microsystems Design Lab, which has more than 50 graduate students.

8 REFERENCES

- [1] L. Itti, C. Koch, and E. Niebur. (1998) "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.
- [2] L. Itti and C. Koch. (2001) "Computational Modelling of Visual Attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203.
- [3] J. Harel, C. Koch, and P. Perona. (2006) "Graph-Based Visual Saliency," *Advances in Neural Information Processing Systems*.
- [4] N. D. B. Bruce and J. K. Tsotsos. (2009) "Saliency, Attention, and Visual Search: An Information Theoretic Approach," *Journal of Vision*, vol. 9, no. 3, pp. 5.1–24.
- [5] T. Judd, K. Ehinger, F. Durand, and A. Torralba (2009) "Learning to Predict Where Humans Look," *IEEE Proceedings of The 12th International Conference on Computer Vision*, pp. 2106–2113.
- [6] T. Judd, F. Durand, and A. Torralba. (2012) "A Benchmark of Computational Models of Saliency to Predict Human Fixations," MIT Technical Report.
- [7] S. Advani, J. Sustersic, K. Irick, and V. Narayanan. (2013) "A Multi-Resolution Saliency Framework To Drive Foveation," *IEEE Proceedings of The 38th International Conference on Acoustics, Speech, and Signal Processing*.
- [8] G. Azzopardi and N. Petkov. (2012) "A CORF Computational Model of a Simple Cell that Relies on LGN Input Outperforms the Gabor Function Model," *Biological Cybernetics*, vol. 106, no. 3, pp. 177–89.
- [9] J. G. Daugman. (1985) "Uncertainty Relation for Resolution in Space, Spatial Frequency, and Orientation Optimized by Two-Dimensional Visual Cortical Filters," *Journal of the Optical Society of America*, vol. 2, no. 7, pp. 1160–1169.
- [10] D. G. Lowe. (2004) "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110.
- [11] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust Object Recognition with Cortex-Like Mechanisms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 411–426, 2007.
- [12] I. Kokkinos and A. Yuille. (2009) "HOP $\square$: Hierarchical Object Parsing," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pp. 802–9.
- [13] S. Chikkerur, T. Serre, and T. Poggio. (2009) "Attentive Processing Improves Object Recognition".
- [14] S. M. Crouzet and T. Serre. (2011) "What are the Visual Features Underlying Rapid Object Recognition?," *Frontiers in Psychology*, vol. 2.
- [15] Zhang, J., Barhomi, Y., Serre, T. (2012) A new biologically inspired color image descriptor. In *Proceedings of the IEEE Conference on European Conference on Computer Vision (ECCV)*, Florence, Italy.
- [16] Burt, P. J. (May 1981). "Fast filter transform for image processing". *Computer Graphics and Image Processing* 16: 20–51. doi:10.1016/0146-664X(81)90092-7
- [17] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes.," *J. Vis.*, vol. 11, no. 3, pp. 1–15, Jan. 2011.
- [18] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *J. Comput. Vis.*, pp. 1–20, 2000.
- [19] O. Pele and M. Werman, "A Linear Time Histogram Metric for Improved SIFT Matching," in *European Conference on Computer Vision*, 2008.
- [20] G. Azzopardi, A. Rodriguez-Sanchez, J. Piater, N. Petkov, "A Push-Pull CORF Model of a Simple Cell with Antiphase Inhibition Improves SNR and Contour Detection," *PLOS One*, July 24, 2014, DOI: 10.1371/journal.pone.0098424