Saliency Estimation Using a Non-Parametric Low-Level Vision Model

Naila Murray, Maria Vanrell, Xavier Otazu, and C. Alejandro Parraga
Computer Vision Center, Computer Science Department
Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain
nmurray@cvc.uab.es, xotazu@cvc.uab.es, maria.vanrell@uab.es, aparraga@cvc.uab.es

Abstract

Many successful models for predicting attention in a scene involve three main steps: convolution with a set of filters, a center-surround mechanism and spatial pooling to construct a saliency map. However, integrating spatial information and justifying the choice of various parameter values remain open problems. In this paper we show that an efficient model of color appearance in human vision, which contains a principled selection of parameters as well as an innate spatial pooling mechanism, can be generalized to obtain a saliency model that outperforms state-of-the-art models.

Scale integration is achieved by an inverse wavelet transform over the set of scale-weighted center-surround responses. The scale-weighting function (termed $ECSF$) has been optimized to better replicate psychophysical data on color appearance, and the appropriate sizes of the center-surround inhibition windows have been determined by training a Gaussian Mixture Model on eye-fixation data, thus avoiding ad-hoc parameter selection. Additionally, we conclude that the extension of a color appearance model to saliency estimation adds to the evidence for a common low-level visual front-end for different visual tasks.

1. Introduction

Saccadic eye movements are perhaps one of the most defining characteristics of the human visual system, allowing us to rapidly sample images by changing the point of fixation. Although many factors may determine what image features are selected or discarded by our attentional processes, it has been useful to separate these into two categories of factors: bottom-up and top-down. The former comprises automatically-driven (instantaneous) processes while the later comprises processes that are dependent of the organism's internal state (such as the visual task at hand or the subject's background). While the difficulties of understanding internal states are usually dealt with by machine learning techniques trained on general prior knowledge, image-driven processes are usually tackled by building saliency maps, getting inspiration from low-level biological processes which are better known than the more elusive top-down mechanisms. Saliency maps are topographical maps of the visually salient parts of scenes (saliency at a given location is in turn determined by how different this location is from its surround in color, orientation, motion, depth, etc. [10]). Computing these maps is still an open problem whose interest is growing in computer vision [6, 5, 3, 18, 9, 8].

Several computational models have already been proposed to predict human gaze fixation, some of which are inspired by biological mechanisms (usually well known low-level processes) while others are based on learning techniques that directly train from human fixation data.

Among the biologically-inspired models of saliency, the model of Itti *et al.* [7] is one of the most influential, summing the scale-space center-surround excitation responses of feature maps at different spatial frequencies and orientations and feeding the result into a neural network, the output of which measures saliency. Gao *et al.* [4] approached saliency at a location as the discriminatory power of a set of features describing that location to distinguish between the region and its surround. Bruce & Tsotsos [3] considered saliency at a location to be quantified by the self-information of the location with respect to its surrounding context - either the entire image, or more localized pixel regions. Zhang *et al.* [18] also proposed a method based on self-information, but using a spatial pyramid to produce local features (with the contextual statistics being generated from a collection of natural images rather than a local neighborhood of pixels or a single image). Seo & Milanfar [14, 15] uses a self-resemblance mechanism to compute saliency, where a region with dissimilar curvature compared to its surroundings was designated as being highly salient. In a typical learning-based approach [9, 8], salient features are learned and combined using eye-tracking data, with the learning techniques serving to reduce the number of model parameters that must be tuned.

In the most common bottom-up modelling framework,

(a) (b) (c)

Figure 1: (a) Two salient features of a scene outlined in green and red. In (b) and (c) we show the spatial scale and orientations at which each object is most prominent. Because these scales and orientation are different for the two features, integrating information contained in the spatial pyramid is critical.

attention in a scene involves a scale-space decomposition of the input image using a set of linear filters, a center-surround operation over the decomposition, and some kind of spatial pooling to build the final saliency map. However, two main questions at the core of this approach remain unresolved: (a) how to integrate the information derived from the multiple scales of the decomposition, and (b) how to adjust the various parameters in order to obtain a general mechanism. Integrating scale information is of particular importance as salient features in a scene and in different scenes may occupy different spatial frequencies, as shown in Figure 1. Therefore a mechanism to locate salient features at different levels of the spatial pyramid and combine these features into a final map is critical.

In this paper, we propose a computational model of saliency that follows the typical three-step architecture described above, while trying to answer the above questions through a combination of simple, neurally-plausible mechanisms that remove nearly all arbitrary variables. Our proposal in this paper generalizes a particular low-level model developed to predict color appearance [13] and has three main levels:

In the first stage, the visual stimuli are processed in a manner consistent with what is known about the early human visual pathway (color-opponent and luminance channels, followed by a multi-scale decomposition). The bank of filters used (Gabor-like wavelets) and the range of spatial scales (in octaves) are biologically justified [1, 16, 17] and commonly used in low-level vision modelling.

The second stage of our model consists of a simulation of the inhibition mechanisms present in cells of the visual cortex, which effectively normalize their response to stimulus contrast. The sizes of the central and normalizing surround windows were learned by training a Gaussian Mixture Model (GMM) on eye-fixation data.

The third stage of our model integrates information at multiple scales by performing an inverse wavelet transform directly on weights computed from the non-linearization of the cortical outputs. This non-linear integration is done through a weighting function similar to that proposed by Otazu *et al.* [13] and named *Extended Contrast Sensitivity Function* ($ECSF$), but optimized to fit psychophysical color matching data at different spatial scales.

Our fitted $ECSF$ is at the core of our proposal and represents its most novel component. It had been previously adjusted by fitting the same low-level model to predict matching of color inductive patterns by human observers. The fact that this function can also model saliency provides support for the hypothesis of a unique underlying low-level mechanism for different visual tasks. This mechanism can be modelled either to predict color appearance (by applying the inverse wavelet transform onto the decomposed coefficients modulated by the $ECSF$ weights) or visual salience (by applying the transform to the weights themselves instead). In addition, we introduce a novel approach to selecting the size of the normalization window, which reduces the number of parameters that must be set in an ad-hoc manner.

Our two main contributions can be summarized as follows:

1. A framework for integrating scale through a simple inverse wavelet transform over the set of weighted center-surround outputs.

2. A reduction of ad-hoc parameters. This was done by introducing training steps on both color appearance and eye-fixation psychophysical data.

The rest of this paper is organized as follows. In section 2 we present the low-level color vision model and our fitted ECSF. In section 3, we use the resulting weights of the model to compute saliency while in section 3.1 we evaluate the model's performance. Section 3.2 summarizes the results and section 4 discusses further work.

2. A low level vision model

The saliency estimation method we propose in this work is an extension of a low level visual representation derived from the unified color induction model developed by Otazu *et al.* [12, 13]. In these works the authors propose a multiresolution model that predicts brightness and color appearance, respectively. Color perception is the result of several adaptation mechanisms which cause the same patch to be perceived differently depending on its surround. Areas A and B of both images in Figure 2 are perceived as having different brightness (in panel a) and/or different color (in panel c) respectively, although in both cases they are physically identical (intensity and RGB color channel profiles are plotted as solid lines in the corresponding panels (b) and (d)). These illusions [1] are predicted by the color model of Otazu *et al.*[13], shown in dashed lines in Figure 2 (panels

[1] the Checkershadow and Beau-lotto illusions were created by E.H. Adelson and Beau Lotto respectively.

(b) and (d)). For example, area A is darker in graphic (b) and area B is more orange-ish in graphic (d).

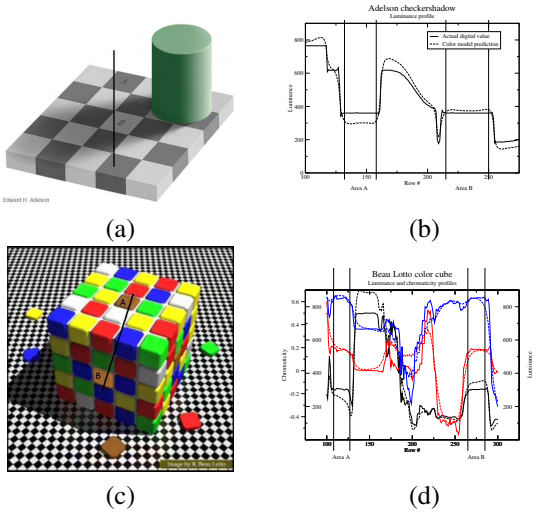


Figure 2: Brightness and color visual illusions with their corresponding image profiles (continuous lines, panels b and d) and model predictions profiles (broken lines, in panels b and d).

In the first stage of Otazu *et al.*'s model, an image is convolved with a bank of filters using a multi-resolution wavelet transform. The resulting spatial pyramid contains wavelet planes oriented either horizontally (h), vertically (v) or diagonally (d). The coefficients of the spatial pyramid obtained using the wavelet transform can be considered an estimation of the local oriented contrast. For a given image I, the wavelet transform is denoted as

$$WT(I_c) = \{w_{s,o}\}_{s=1,2,\ldots,n\,;\,o=h,v,d} \tag{1}$$

where $w_{s,o}$ is the wavelet plane at spatial scale s and orientation o and I_c represents one of the opponent channels $O1$, $O2$ and $O3$ of image I. Each opponent channel is decomposed into a spatial pyramid using the wavelet transform, WT. This transform contains Gabor-like basis functions and the number of scales used in the decomposition is given by $n = log_2 D$ for an image whose largest dimension is size D.

In the second stage, the contrast energy $a_{x,y}$ around a wavelet coefficient $\omega_{x,y}$ centered at position x, y is estimated by convolving the local region with a binary filter h. The shape of the filter varies with the orientation of the wavelet plane on which it operates, as shown in Figure 5. For example, for a horizontal wavelet plane, $a_{x,y}$ is computed by

$$a_{x,y} = \sum_j \omega_{x-j,y} 2h_j \tag{2}$$

where h_j is the j-th coefficient of the one-dimensional filter h. The contrast energy is computed for coefficients at all spatial locations and spatial scales. Filter h_j defines a region around the central wavelet coefficient $\omega_{x,y}$ where the activity $a_{x,y}$ is calculated. The interaction between this central region and surrounding regions produces a center-surround effect. In order to model this center-surround effect, the energies of the central region $a^{cen}_{x,y}$ and the surround region $a^{sur}_{x,y}$ are compared using

$$r_{x,y} = (a^{cen}_{x,y})^2/(a^{sur}_{x,y})^2. \tag{3}$$

The energy of the surrounding regions, $a^{sur}_{x,y}$, is computed in an analogous manner to $a^{cen}_{x,y}$, with the only difference being the definition of the filter h, also shown in Figure 5. A non-linear scaling of $r_{x,y}$ is performed to produce the final center-surround energy measure $z_{x,y}$:

$$z_{x,y} = r^2_{x,y}/(1 + r^2_{x,y}) \tag{4}$$

so that $z_{x,y} \in [0, 1]$. When $z_{x,y} \to 0$, central activity $a^{cen}_{x,y}$ is much lower than surround activity $a^{sur}_{x,y}$. Similarly, when $z_{x,y} \to 1$, central activity is much higher than surround activity. Therefore, $r_{x,y}$ may be interpreted as a saturated approximation to the relative central activity $a^{cen}_{x,y}$. The size of central and surround regions are used to define the size of the corresponding h_j filters.

It is well-known that color appearance is dependent on spatial frequency. Mullen [11] described human sensitivity to local contrast in color opponent channels with a generalized Contrast Sensitivity Function (CSF), which is a function of spatial frequency. Adopting this idea, Otazu *et al.* define an extended contrast sensitivity function ($ECSF$) which is parametrized by spatial scale s and center-surround contrast energy. Spatial scale is inversely proportional to spatial frequency ν such that $s = log_2(1/\nu) = log_2(T)$, where T is the period and thus denotes one frequency cycle measured in pixels. The function $ECSF$ is defined as

$$ECSF(z, s) = z \cdot g(s) + k(s) \tag{5}$$

where the function $g(s)$ is defined as

$$g(s) = \begin{cases} \beta e^{-\frac{s^2}{2\sigma_1^2}} & s \le s^g_0 \\ \beta e^{-\frac{s^2}{2\sigma_2^2}} & otherwise \end{cases} \tag{6}$$

Here s represents the spatial scale of the wavelet plane being processed, β is a scaling constant, and σ_1 and σ_2 define the spread of the spatial sensitivity of $g(s)$. The s^g_0 parameter defines the peak spatial scale sensitivity of $g(s)$. In Equation 5, the center-surround activity z of wavelet coefficients are modulated by $g(s)$. An additional function, $k(s)$, was introduced to ensure a non-zero lower bound on

$ECSF(z,s)$:

$$k(s) = \begin{cases} e^{-\frac{s^2}{2\sigma_3^2}} & s \leq s_0^k \\ 1 & otherwise \end{cases} \quad (7)$$

Here, σ_3 defines the spread of the spatial sensitivity of $k(s)$ and s_0^k defines the peak spatial scale sensitivity of $k(s)$.

The function $ECSF$ is used to weight the center-surround contrast energy $z_{x,y}$ at a location, producing the final response $\alpha_{x,y}$:

$$\alpha_{x,y} = ECSF(z_{x,y}, s_{x,y}). \quad (8)$$

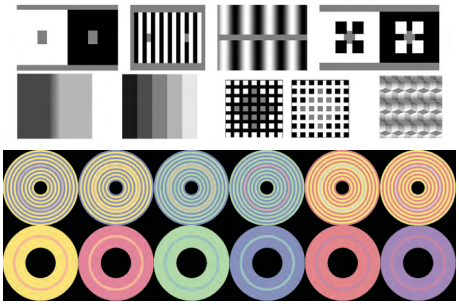
$\alpha_{x,y}$ is the weight that modulates the wavelet coefficient $\omega_{x,y}$. The perceived image channel $I_c^{perceived}$ that contains the color appearance illusions are obtained by performing an inverse wavelet transform on the wavelet coefficients $\omega_{x,y}$ at each location, scale and orientation, after the coefficients have been weighted by the $\alpha_{x,y}$ response at that location:

$$I_c^{perceived}(x,y) = \sum_s \sum_o \alpha_{x,y,s,o} \cdot \omega_{x,y,s,o} + C_r \quad (9)$$

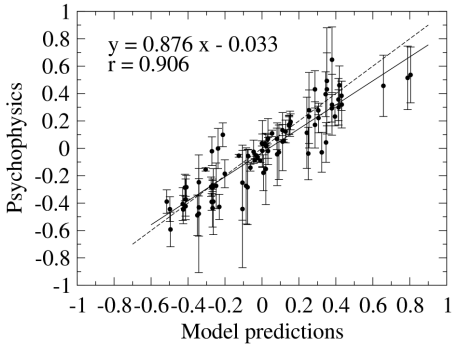
Here o represents the orientation of the wavelet plane of $\omega_{x,y,s,o}$ and C_r represents the residual image plane obtained from WT.

The model of Otazu *et al.* was capable of replicating the psychophysical data obtained from two separate experiments. In the first experiment, by Blakeslee *et al.* [2], observers performed asymmetric brightness matching tasks in order to match the illusions present in regions of the stimuli. Some example brightness stimuli are shown in Figure 3(a). The second experiment was performed by Otazu *et al.* [13] in an analogous fashion, but with observers performing asymmetric color matching tasks rather than tasks involving brightness. Some example color stimuli used in these experiments are shown in Figure 3(a).

Our saliency estimation model is based on the previous stages we have just described. However, to obtain parameters for the intensity and color $ECSF(z,s)$ functions, we used the psychophysical data, which was provided to us by the authors of [2] and [13], to perform a least squares regression in order to select the parameters of the functions. Our results are given in table 1. Both fitted $ECSF(z,s)$ functions maintain a high correlation rate ($r = 0.9$) with the color and lightness psychophysical data, as shown in Figure 3(b). Note that both chromaticity channels share the same $ECSF(z,s)$ function. The profiles of the resulting optimized $ECSF(x,s)$ functions for brightness and chromaticity channels are shown in Figure 4. The functions enhance contrast energy responses in a narrow passband and suppresses contrast energy for low spatial scales (high spatial frequencies). The magnitude of the enhancement or suppression increases with the magnitude of the center-surround contrast energy, z.



(a)



(b)

Figure 3: (a) Examples of images used in psychophysical experiments. (b) Correlation between model prediction and psychophysical data. The solid line represents the model linear regression fit and the dashed line is the ideal fit. Since measurements involve dimensionless measures and physical units, they were arbitrarily normalized to show the correlation.

Param.	σ_1	σ_2	σ_3	β	s_0^g	s_0^k
Intensity	1.021	1.048	0.212	4.982	4.000	4.531
Color	1.361	0.796	0.349	3.612	4.724	5.059

Table 1: Parameters for $ECSF(z,s)$ obtained using least square regression.

3. Building saliency maps

In the previous section we described a low-level visual representation that predicts color appearance phenomena. This model concluded with equation 9 which can be reformulated as

$$I_c^{perceived}(x,y) = WT^{-1}\{\alpha_{x,y,s,o} \cdot \omega_{x,y,s,o}\} \quad (10)$$

where $I_c^{perceived}$ is a new version of the original channel in which image locations may have been modified by the α weight, either by a blurring or an enhancing effect. The colors of modified locations have either been assimilated (averaged) to be more similar to the surrounding color or contrasted (sharpened) to be less similar to the surround.

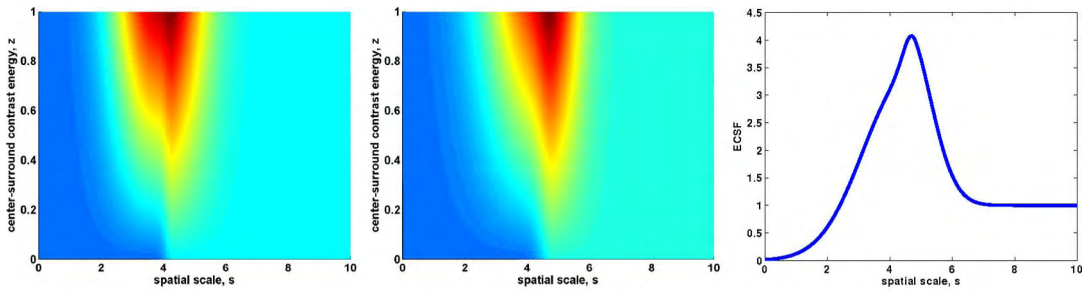


Figure 4: Weighting functions for (a) intensity and (b) chromaticity channels: Bluer colors represent lower values while redder colors indicate higher values. (c) shows a slice of $ECSF(z, s)$ for chromaticity channels, for $z = 0.9$. For a wavelet coefficient corresponding to a scale between approximately 3 and 6, its center-surround contrast energy is boosted. Coefficients outside this passband are either suppressed (for low spatial scales) or remain unchanged (for high spatial scales).

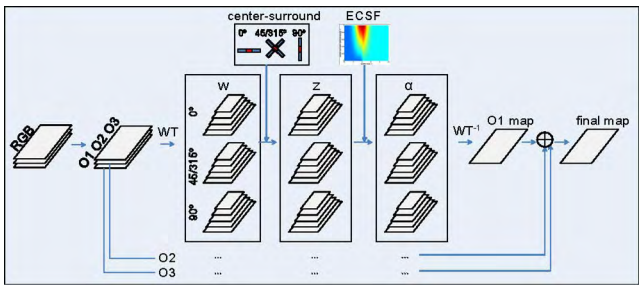


Figure 5: Schematic of our saliency approach. Red sections of the center-surround filters correspond to the central filters while blue sections correspond to the surround filters.

To obtain predictions of saliency using this color representation, we hypothesize that image locations that undergo enhancement are salient, while locations that undergo blurring are non-salient. In this sense we can define the saliency map of an specific image channel by the inverse wavelet transform of the α weight. Thus the saliency map, S_c, of the image channel I_c at the location x, y can be easily estimated as

$$S_c(x, y) = WT^{-1}\{\alpha_{x,y,s,o}\}. \tag{11}$$

By removing the wavelet coefficients $\omega_{x,y,s,\bullet}$ and performing the inverse transform solely on the weights computed at each image location we provide an elegant and direct method for estimating image saliency from a generalized low level visual representation.

To combine the maps for each channel into the final saliency map, S, we compute the Euclidean norm $S = \sqrt{S_{O1}^2 + S_{O2}^2 + S_{O3}^2}$. The steps of the saliency model are illustrated in Figure 5.

This process generalizes the color appearance model to one which estimates saliency. The main advantage of our method is the integration of multi-scale information by the inverse wavelet transform, and the use of the $ECSF$, whose parameters are biologically justifiable. However, there are still some parameters to be set. The strength of the center-surround inhibition defined by Equation 4 is highly dependent on the size of the binary h_j filters, which define the extent of the local central and surround regions of a wavelet coefficient $\omega_{x,y}$. We posit that the central region around a feature ought to span the responses to that feature in the wavelet plane. However, the extent of the wavelet response to a feature differs with each spatial scale. Therefore we designed the size of the central region to be just large enough to span wavelet responses for a feature at the most salient spatial scale. The most salient spatial scale was taken to be the scale to which both $ECSF(z, s)$ functions are most sensitive, approximately $s = 4$.

We consider the central region encompassing a feature to be a Region Of Interest (ROI). Therefore, we estimate the required size of the central region by determining the typical size of a ROI. To determine this size we first created a Gaussian Mixture Model (GMM) of the locations of the eye fixations for a subset of 70 images from the Bruce & Tsotsos dataset [3], one of the datasets we will use to evaluate our model. The GMMs contained 5 components, each of which clusters the locations of a set of eye-fixations and thus represents an ROI. The standard deviation of each component is therefore interpreted to be the radius of an ROI. Across the 70 images, the average radius of a Gaussian component was 27 pixels. At $s = 4$, the radius would now be $2(27/2^{s-1}) = 2(27/8) = 6.8$. Therefore we set the size of the local central region to be twice this radius, or 13 pixels. The radius of the surround region was set to be 26 pixels, twice the size of the central region, so that the resulting center-surround region spanned 65 pixels. The sizes of the central and surround regions were used when performing the evaluation on both datasets introduced in the

upcoming section describing our experimental results.

An important point to consider is whether the peak spatial scales of the $ECSF$ functions, approximately $s = 4$, are consistent with the peak spatial frequencies of the human CSFs for chromatic and achromatic channels, which have been estimated to be around 2 and 4 cycles per degree (cpd) respectively [11]. That is, are the spatial scales being enhanced by the $ECSF$ functions consistent with the spatial frequencies to which humans are most attuned? If we assume that the ROI spans a feature with a spatial frequency between 2 and 4 cpd, 106 pixels contain 2-4 spatial periods. The spatial scale that corresponds to 2 cpd is $s = log_2(T) = log_2(106/2) = 5.7$, while 4 cpd corresponds to $s = 4.7$. These spatial scales are indeed consistent with the peak $ECSF$ spatial scales obtained by least squares regression.

3.1. Experimental results

We evaluated our model's performance with respect to predicting human eye fixation data from two image datasets. To assess the accuracy of our model we used both the well-known receiver operating characteristic (ROC) and Kullback-Leibler (KL) divergence as quantitative metrics. The ROC curve indicates how well the saliency map discriminates between fixated and non-fixated locations for different binary saliency thresholds while the KL divergence indicates how well the method distinguishes between the histograms of saliency values at fixated and non-fixated locations in the image. For both of these metrics, a higher value indicates better performance.

Zhang *et al.* noted that several saliency methods have image border effects which artificially improve the ROC results [18]. To avoid this issue and ensure a fair comparison of saliency methods we adopt the evaluation framework described by Zhang *et al.* [18], which involves modified metrics for both the area under the ROC curve (AROC) and KL divergence. For each image in the dataset, true positive fixations are fixations for that image, while false positive fixations are fixations for a *different* image from the dataset, chosen randomly. This avoids the true positive fixations having a center bias with respect to the false positive fixations. Because the false fixations for an image are randomly chosen, a new calculation of the metrics is likely to produce a different value. Therefore we computed the metrics 100 times in order to compute the standard error. The saliency maps are shuffled 100 times. On each occasion, the KL-divergence is computed between the histograms of saliency values at unshuffled fixation points and shuffled fixation points. When calculating the area under the ROC curve, we also used 100 random permutations of the fixation points.

The first dataset we use was provided by Bruce & Tsotsos in [3]. This popular dataset is commonly used as the benchmark dataset for comparing visual saliency predictions between methods. The dataset contains 120 color images of indoor and outdoor scenes, along with eye-fixation data for 20 different subjects. The mean and the standard error of each metric are reported in Table 2. We performed this evaluation on five state-of-the-art methods as well as our proposed method and as Table 2 shows, our method exceeds the state-of-the-art performance as measured by both metrics.

Model	KL (SE)	AROC (SE)
Itti *et al.*[7]	0.1130 (0.0011)	0.6146 (0.0008)
Bruce & Tsotsos[3]	0.2029 (0.0017)	0.6727 (0.0008)
Gao *et al.*[4]	0.1535 (0.0016)	0.6395 (0.0007)
Zhang *et al.*[18]	0.2097 (0.0016)	0.6570 (0.0008)
Seo & Milanfar[15]	0.3432 (0.0029)	0.6769 (0.0008)
Our method	0.4265 (0.0030)	0.7013 (0.0008)

Table 2: Performance in predicting human eye fixations from the Bruce & Tsotsos dataset.

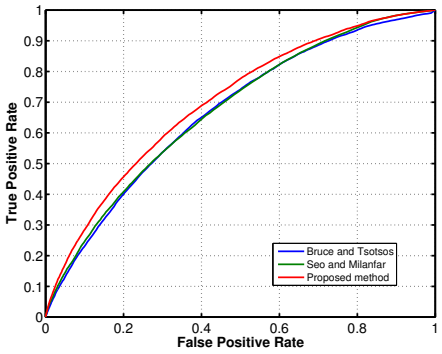


Figure 6: ROC curves for Bruce & Tsotsos, Seo & Milanfar, and the proposed method.

The second dataset we used was introduced by Judd *et al.* in [8]. This dataset contains 1,003 images of varying dimensions, along with eye fixation data for 15 subjects. In order to be able to compare fixations across images, only those images whose dimensions were 768x1024 pixels were used, reducing the number of images examined to 463. This dataset is more challenging than the first as its images contain more semantic objects which are not modeled by bottom-up saliency, such as people, faces and text. Therefore, as would be expected, the AROC and KL divergence metrics are lower for all bottom-up visual attention models. The results, obtained using the same evaluation method described previously, are shown in Table 3 and indicate that once again our method exceeds state-of-the-art performance.

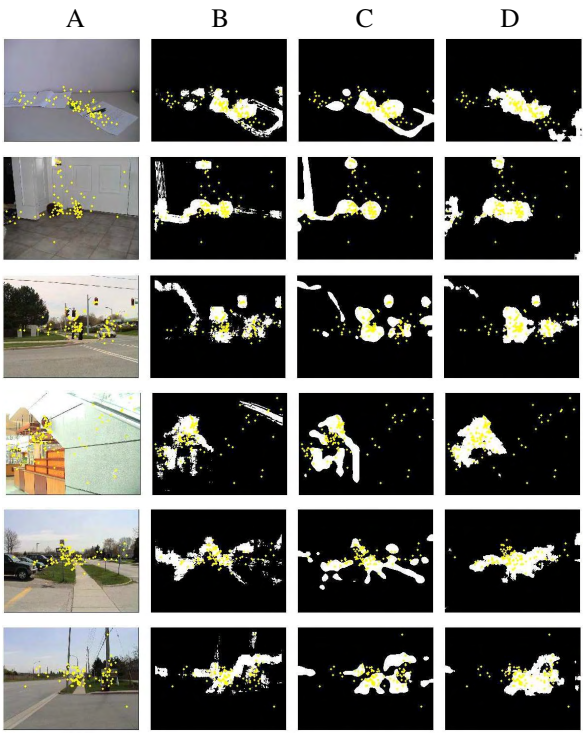


Figure 7: Qualitative analysis of results for Bruce & Tsotsos dataset: Column A contains original image. Columns B, C, and D contain thresholded saliency maps obtained from Bruce & Tsotsos, Seo & Milanfar and our method, respectively. The saliency maps have each been thresholded to their top 10% most salient locations. Yellow markers indicate eye fixations. Our method is seen to be less sensitive to low-frequency edges such as street curbs and skylights, which is in line with human eye fixations.

Model	KL (SE)	AROC (SE)
Bruce & Tsotsos [3]	0.2629 (0.0025)	0.6501 (0.0008)
Seo & Milanfar [15]	0.2700 (0.0025)	0.6462 (0.0007)
Our method	0.2788 (0.0021)	0.6640 (0.0006)

Table 3: Performance in predicting human eye fixations from the Judd *et al.* dataset.

3.2. Discussion

Figure 7 illustrates the benefit of our method when compared to Bruce & Tsotsos [3] and Seo & Milanfar [15]. The saliency maps have each been thresholded to their top 10% most salient locations and show that the most salient regions of our saliency map better correspond to the fixations of human observers. In addition, the ROC curves for the three methods in Figure 6 show that our method has fewer false positives at higher thresholds, indicating that the proposed method is better able to detect the most salient regions of the image.

Figure 8 shows qualitative results for the second dataset, provided by Judd *et al.* [8]. Here there is also a higher correlation between the most salient regions of our saliency map, and human eye fixations, when compared with Bruce & Tsotsos and Seo & Milanfar.

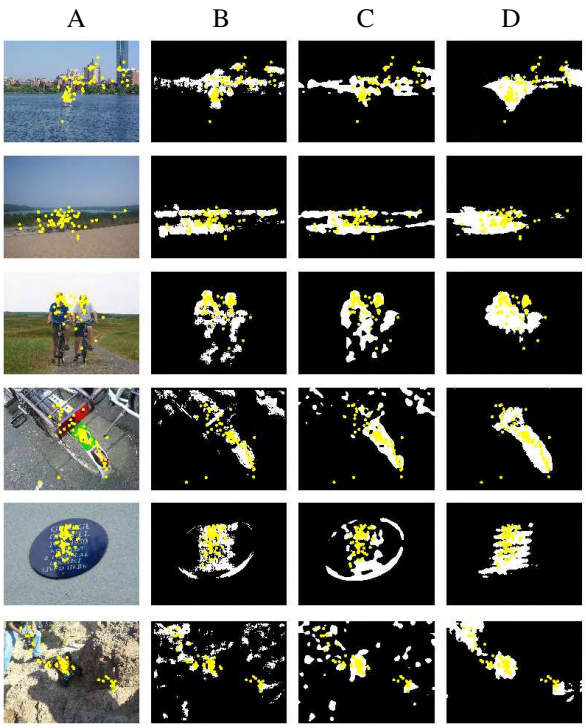


Figure 8: Qualitative analysis of results for Judd *et al.* dataset: Column A contains original image. Columns B, C, and D contain saliency maps obtained from Bruce & Tsotsos, Seo & Milanfar and our method, respectively. The saliency maps have each been thresholded to their top 10% most salient locations.

We attribute our model's success to the fact that it is less sensitive to low-frequency edges in the images, such as skylines and road curbs. In addition, we avoid excessive sensitivity to textured regions by suppressing high-frequency information using the weighting functions $ECSF(z, s)$. As Figure 4 shows, the weighting function is more sensitive to mid-range frequencies. The previous methods included in Table 2 either select information at one scale or combine scale information from subband pyramids by an unweighted linear combination while in our method, $ECSF(z, s)$ acts as a bandpass filter in the image's spatial frequency domain, and provides a biologically plausible mechanism for combining spatial information.

Finally, we also investigated how the performance of our saliency model changed depending on the peak spatial scale of $ECSF(z, s)$ for the intensity channel, which is the channel which contains the majority of the saliency information. Figure 9 shows that, as expected from psychophysical data, when low or high frequencies are enhanced and mid-range

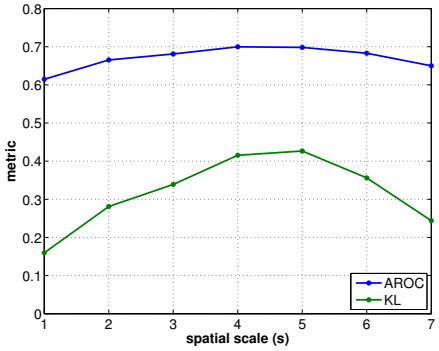


Figure 9: Change in AROC and KL metrics with change in peak frequency of $ECSF(z, s)$ for intensity: The best s for both these metrics ($s = 4$ for AROC and $s = 5$ for KL) are in line with the value determined using psychophysical experiments.

frequencies are inhibited the performance of the model suffers. The model performs best when mid-range frequencies are enhanced and low or high frequencies are inhibited. In addition, the best scale range for these metrics, between 4 and 5, correspond to the value determined using psychophysical experiments.

4. Conclusions and further work

The proposed saliency model can be summarized by the following pipeline:

$$I_c \xrightarrow{WT} \{\omega_{s,o}\} \xrightarrow{CS} \{z_{s,o}\} \xrightarrow{ECSF} \{\alpha_{s,o}\} \xrightarrow{WT^{-1}} S_c$$

where CS represents the center-surround mechanism and $ECSF$ is the extended contrast sensitivity function. The main advantage of our formulation is the use of a scale-weighting function that is less sensitive to non-salient edges and provides a biologically plausible mechanism for integrating scale information contained in the spatial pyramid. Additionally, we reduced ad-hoc parameters by learning the appropriate size of local central regions. At the moment, the size of this region is optimized for the most salient spatial scale and held constant for other frequencies. Further work will include modulating the size of this region with respect to spatial scale, as well as exploring less ad-hoc means of representing the suppressive surround region.

5. Acknowledgments

We would like to thank Hae Jong Seo for sharing his evaluation code. This work has been supported by Projects TIN2007-64577, TIN2010-21771-C02-1 and Consolider-Ingenio 2010-CSD2007-00018 from the Spanish Ministry of Science. C. Alejandro Parraga was funded by grant RYC-2007-00484.

References

[1] C. Blakemore and F. W. Campbell. On the existence of neurons in the human visual system selectively sensitive to the orientation and size of retinal images. *The Journal of Physiology*, 203(1):237260, 1969.

[2] B. Blakeslee and M. E. McCourt. Similar mechanisms underlie simultaneous brightness contrast and grating induction. *Vision Research*, 37(20):2849–2869, 1997.

[3] N. D. Bruce and J. K. Tsotsos. Saliency based on information maximization. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 155–162, MIT Press, 2006. MIT Press.

[4] D. Gao, V. Mahadevan, and N. Vasconcelos. On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, 8(7:13):1–18, 2008.

[5] D. Gao and N. Vasconcelos. Bottom-up saliency is a discriminant process. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–6, 2007.

[6] L. Itti and C. Koch. Computational modelling of visual attention. *Nature reviews. Neuroscience*, 2(3):194–203, March 2001.

[7] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 1998.

[8] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[9] W. Kienzle, F. Wichmann, B. Schalkopf, and M. O. Franz. A nonparametric approach to bottom-up visual saliency. In *In Proceedings of NIPS 19*. MIT Press, 2007.

[10] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, 4(4):219–227, 1985.

[11] K. Mullen. The contrast sensitivity of human color-vision to red green and blue yellow chromatic gratings. *Journal of Physiology*, pages 381–400, 1985.

[12] X. Otazu, A. Parraga, and M. Vanrell. Multiresolution wavelet framework models brightness induction effects. *Vision Research*, 48:733–751, 2008.

[13] X. Otazu, C. A. Parraga, and M. Vanrell. Toward a unified chromatic induction model. *Journal of Vision*, 10(12), 2010.

[14] H. J. Seo and P. Milanfar. Nonparametric bottom-up saliency detection by self-resemblance. In *IEEE CVPR, 1st International Workshop on Visual Scene Understanding*, pages 45 –52, 2009.

[15] H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12):15.1–27, 2009.

[16] A. Werner. The spatial tuning of chromatic adaptation. *Vision Research*, 43(15):1611 – 1623, 2003.

[17] C. Yu, S. Klein, and D. Levi. Facilitation of contrast detection by cross-oriented surround stimuli and its psychophysical mechanisms. *Journal of Vision*, 2(3):243255, 2002.

[18] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):–, 2008.