Low-level Spatio-Chromatic Grouping for Saliency Estimation

Naila Murray, *Student Member, IEEE,* Maria Vanrell, *Member, IEEE,* Xavier Otazu, and C. Alejandro Parraga

Abstract—We propose a saliency model termed SIM (Saliency by Induction Mechanisms) which is inspired by a low-level spatio-chromatic model that has successfully predicted color perception phenomena. In so doing, we hypothesize that these phenomena, known collectively as color induction, are directly related to saliency. SIM uses geometrical grouplets to enhance complex low-level visual features such as corners, and suppress relatively simpler features such as edges. As SIM's parameters have been fitted on psychophysical data, it is largely non-parametric. SIM outperforms state-of-the-art methods in predicting eye-fixations on two datasets and using two metrics.

Index Terms—Computational models of vision, color, hierarchical image representation

1 INTRODUCTION

The ability to predict the attentional gaze of observers viewing a scene has wide applications, from object recognition and visual aesthetics to marketing and user interface development. As a result, a great deal of research effort has been devoted to developing models of human visual attention. Visual attention is thought to comprise bottom-up and top-down components. This paper focuses on bottom-up attention or saliency, which relates to cues such as local contrast, color and motion.

There is a wide spectrum of methods for modeling attention[1], from biologically-inspired models to learning-based approaches. A typical learning-based method is that of Keinzle et al. [2], who sampled small image patches at eye-fixation locations and learned which of these patches classify fixation locations well. Following in this vein, the model of Judd et al. [3] combined the information contained in different saliency methods to produce a single saliency map, by using a support vector machine.

Among the more bio-inspired models of saliency, the model of Itti *et al.* [4] is one of the most influential. It uses a neural network to output a saliency map after training the network with center-surround excitation responses of feature maps obtained after a single layer of linear filters are applied to the input image. Each feature map contains information from one of three cues: orientation, color or scale. Gao *et al.* [5] considered the saliency of a local region to be quantified by the discriminatory power of a set of features describing that region to distinguish the region from its surrounding context. Bruce & Tsotsos [6] approached local saliency as the self-information of local patches with respect to its surrounding patches, where the surround could be considered a localized surround region or the remainder of the entire image. In [6], an ICA basis set of filters was learnt from RGB patches extracted from images and used to represent the local patches. As was also found by Hou & Zhang [7] in a similar approach, the basis set consisted mainly of oriented Gabor-like patches with opponent color properties. Zhang *et al.* [8] also proposed a method which uses self-information, but in this case a spatial pyramid was used to produce local features and a database of natural images, rather than a local neighborhood of pixels or a single image, provided contextual statistics. In addition, Zhang *et al.* extracted features from a spatial pyramid of each of the three opponent color channels. Seo & Milanfar [9] used kernel regression-based self-resemblance to compute saliency, and considered a region to be salient when its curvature was different to that of its surround.

In these bio-inspired approaches, there remain several major challenges, including:

- generating the optimal feature maps for estimating saliency [2];
- holistically combining saliency information from these feature maps, which are extracted from multiple scales, orientations and color channels [10]; and
- selecting the many model parameters (such as the number, type and orientation of filters, and coefficients for non-linear normalisations and activation functions) present in such models [11].

In this work, we propose a saliency model which addresses the above issues by making two main contributions:

- *The authors are with the Computer Vision Center, Universitat Autònoma de Barcelona in Bellaterra, Spain.*
E-mail: {nmurray, xotazu, maria,Alejandro.Parraga}@cvc.uab.es

- We adapt a low-level color induction model in order to predict saliency. The resultant saliency model inherits an extended Contrast Sensitivity Function (termed the $ECSF$), which provides a biologically-inspired manner of integrating scale, orientation and color. The $ECSF$ has been fitted to psychophysical data and as a result requires no parameter tuning. As such, it may be considered as prior knowledge included in SIM.
- We use "geometrical grouplets" [12] to produce a sparse and efficiently-computed image representation that enhances features known to guide attention and suppresses non-salient features.

The proposed model exceeds the performance of state-of-the-art saliency estimation methods in predicting eye-fixations for two datasets and using two metrics.

The remainder of this article is organized as follows: in section 2 we describe the color induction principles that underlie our saliency model. In section 3 we describe our sparse image representation based on geometrical grouplets. Our entire saliency estimation framework is detailed in section 4. In section 5 we discuss quantitive and qualitative experimental results and we draw several conclusions in section 6. A preliminary version of this work appeared in [13].

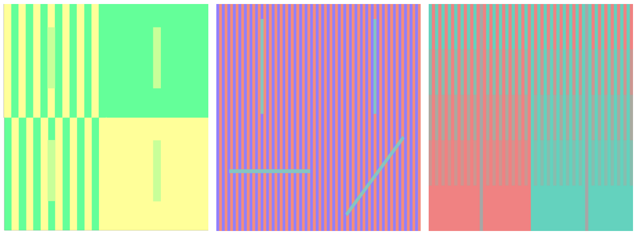
2 MODELING LOW-LEVEL COLOR VISION

Two decades ago, a modular paradigm arose in biological vision stating that color perception occurs in the visual system in a specific cortical area, V4 [14]. This modular paradigm was adopted by Itti *et al.* for saliency [4]. In the intervening years however, a large body of evidence has emerged which supports the view of a more interlinked processing of color and form in the human visual cortex [15].

In this work we adapt a computational model of color perception [16] to the problem of saliency estimation. The model is based on a non-modular approach to combining color, scale and orientation and has been designed to predict well-known color induction phenomena. Color induction refers to perceived changes in the color appearance of a stimulus due to surround influence, and may be demonstrated using common visual illusions. Our adaptation of the color perception model of [16] is motivated by our hypothesis that *factors related to color induction phenomena also inform on local saliency*.

The model of [16] captures the effect of three key properties on the perceived color of stimuli. In the following paragraphs we describe these effects and how they have been incorporated into our saliency model.

First, the perceived color of a stimulus is influenced by the *surround spatial frequency*. Fig. 1(a) shows how



(a) (b) (c)

Fig. 1. Perceived color of the stimulus depends on the (a) color and frequency of the surround; (b) relative orientation of the stimuli to the surround; (c) self-contrast of the surround.

surround spatial frequency affects the perceived colors of 4 identical stimuli. In a high-frequency background the color of the stimulus approaches that of the surround (top left stimulus becomes more greenish while the bottom left becomes yellowish). In a low-frequency background the stimulus's perceived color moves away from the surround color (top right stimulus becomes more yellowish when surrounded by green; bottom right more greenish when surrounded by yellow). These induction effects are termed assimilation and contrast respectively.

Second, *orientation* also influences color appearance. In Fig. 1(b) we can observe that the relative orientation between the stimulus and the surround provokes a perceptual change. While the top left and right stimuli clearly undergo assimilation (a greenish perception when surrounded by pink, and a bluish perception when surrounded by blue), the stimuli at bottom appear closer to their true cyan color. This is because assimilation is greatest when the stimulus and background have the same orientation.

These two effects are incorporated into our saliency model by representing images using a wavelet decomposition, which jointly encodes the spatial frequency and orientation of image stimuli. Given an image I, the wavelet decomposition of one of its channels I_c is

$$WT(I_c) = \{w_{s,o}\}_{1 \leq s \leq S, o=\{h,v,d\}} \quad (1)$$

where $w_{s,o}$ is the wavelet plane at spatial scale s and orientation o. For an image whose largest dimension is size D, the decomposition produces $S = log_2 D$ scales. The wavelet transform WT uses Gabor-like basis functions, as Gabor functions resemble the receptive fields of neurons in the cortex.

Third, *surround contrast* also plays a crucial role in how color is perceived. As shown in Fig. 1(c), chromatic assimilation is reduced and chromatic contrast is increased when the surround contrast decreases. Therefore the amount of induction at an image location is modulated by the surround contrast at that location. The surround contrast of a stimulus at position x, y can be modeled as a divisive normalization, which we term the normalized center contrast, $z_{x,y}$, around a wavelet coefficient $w_{x,y}$. It is estimated as a normalization of the variance of the coefficients of the central region $a^{cen}_{x,y}$ normalized by the variance of the coefficients of the surround region $a^{sur}_{x,y}$:

$$z_{x,y} = \frac{(a^{cen}_{x,y})^2}{(a^{cen}_{x,y})^2 + (a^{sur}_{x,y})^2}. \tag{2}$$

Divisive normalization has been shown by Simoncelli and Schwartz [17] to remove statistical dependencies present in wavelet decompositions of natural scenes and, in this instance, may be viewed as a center-surround contrast mechanism.

The three effects mentioned above are integrated using an extended Contrast Sensitvity Function ($ECSF$). The $ECSF$ determines the type of induction depending on the orientation at a specific spatial frequency, and the amount of induction depending on the surround contrast. This function is inspired by the well-known CSF that was measured in [18] for luminance and colour contrast.

The $ECSF$ we use is a function of spatial scale s and normalized center contrast z. 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 $ECSF$ function is defined as $ECSF(z, s) = z \cdot g(s) + k(s)$. Here z is modulating the function $g(s)$, which is an aproximation to the psychophysical CSF and is itself introducing assimilation or contrast depending on the spatial frequency s. Function $g(s)$ is defined as

$$g(s) = \begin{cases} \beta e^{-\frac{(s-s_0^g)^2}{2\sigma_1^2}} & s \le s_0^g \\ \beta e^{-\frac{(s-s_0^g)^2}{2\sigma_2^2}} & otherwise \end{cases} \tag{3}$$

Here β is a scaling constant, and σ_1 and σ_2 define the spread of the spatial sensitivity of $g(s)$. The s_0^g parameter defines the peak scale sensitivity of $g(s)$. An additional function, $k(s)$, was introduced to ensure a non-zero lower bound on $ECSF(z, s)$:

$$k(s) = \begin{cases} e^{-\frac{(s-s_0^k)^2}{2\sigma_3^2}} & s \le s_0^k \\ 1 & otherwise \end{cases} \tag{4}$$

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

In the induction model of [16], the output of the $ECSF$ was used to weight wavelet coefficients, after which an inverse wavelet transform was performed, producing a new "perceived" image. This reconstructed image replicates color induction phenomena perceived by human observers. For our saliency model, we use these *induction weights* output by the $ECSF$ as a measure of the *saliency* of a feature given its orientation, spatial frequency and center-surround contrast properties.

We have fitted all the parameters of the $ECSF$ in order to predict psychophysical data from two experiments, one involving brightness and the other involving color induction. In the first experiment, by Blakeslee *et al.* [19], observers viewed two stimuli with the same luminance but different perceived brightness. They were then asked to modify the brightness of one of the stimuli to match the perceived brightness of the other stimulus. The second experiment was conducted by Otazu *et al.* [16] in an analogous fashion, but with observers performing asymmetric color and brightness matching tasks rather than tasks involving only brightness. In these matching experiments, the difference between the original physical (color or brightness) values of the stimulus and the modified physical values was recorded as a measure of induction. Least squares regression was used to select the parameters of the functions that best reproduce this data (given in Table 1) in the perceived image output by the induction model. Examples of stimuli used in these experiments are shown in Figure 2.

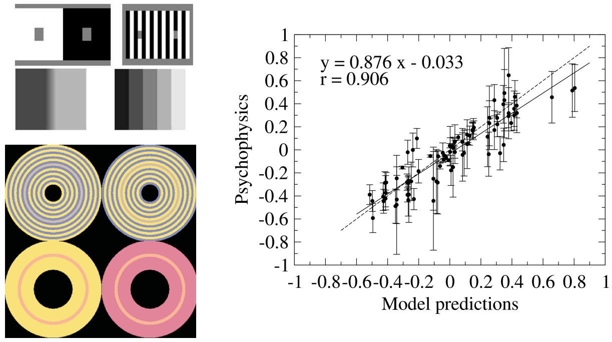


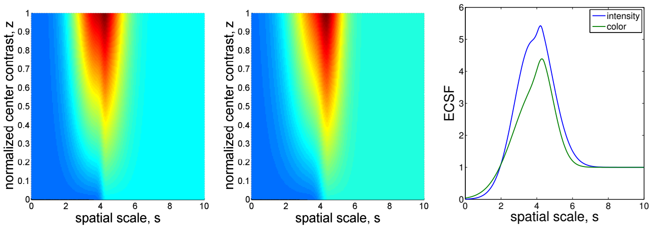
Fig. 2. (a) Examples of experimental stimuli. (b) Correlation between model prediction and psychophysical data. The solid line represents the model linear regression fit and the dashed line is the ideal fit. Since measurements involve dimensionless measures and physical units, they were arbitrarily normalized to show the correlation.

TABLE 1
Fitted Parameters for $ECSF(z, s)$ functions.

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

As the human visual system has different contrast sensitivities for color and luminance, two different $ECSF$ functions were fitted using these data, one for intensity channels ($ECSF_I$) and another for chromatic channels ($ECSF_C$). Both fitted $ECSF(z, s)$ functions maintained a high correlation rate ($r = 0.9$) with the color and brightness psychophysical data (see Figure 2). Their profiles are shown in Fig. 3. The functions enhance normalized center contrast in a narrow passband and suppress this contrast for low spatial scales. The magnitude of the enhancement increases with the magnitude of the normalized center contrast, z, as observed in Figs. 3(a) and 3(b). These $ECSF$s have peak spatial scales in the wavelet decomposition that correspond to peak spatial frequencies between 2-5 cpd, which agree with previous psychophysical estimations [18].

As stated above, we use a wavelet transform with Gabor-like basis functions as an image representation.



(a) (b) (c)

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

This representation agrees with a long-standing view of the early human sensory system as an efficient information processing system [20], [21]. In this view, an objective of early sensory coding is to transform the visual signal into a sparse, statistically independent representation where *redundancy has been removed*. Wavelet decompositions are highly sensitive to edges, in addition to more complex features resulting from super-imposed orientations, such as corners and terminations. However, in comparision with edges, complex features are preferentially fixated on when humans free-view natural images, [22], [23]. Therefore, to estimate saliency, an image representation with higher responses for complex features, relative to the responses for simple features, is desirable. To construct such an improved image representation we will employ the Grouplet Transform (GT).

3 GROUPLET TRANSFORM FOR IMAGE REPRESENTATION

The GT [12] is an additional stage of the image representation that renders it more responsive to complex features. The GT is applied to each wavelet plane $w_{s,o}$ using a modified Haar transform, computed using a lifting scheme.

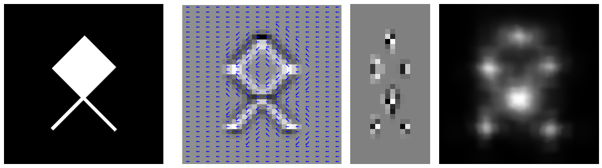
The GT as a modified Haar Transform

The Haar transform (HT) decomposes a signal into a residual (lower-frequency) component and a detail (higher-frequency) component. When the signal is a wavelet plane $w_{s,o}$, its residual data $r_{s,j,o}$ is initialized as $r_{s,1,o} = w_{s,o}$. The grouplet scale j increases from 1 to J, where J is the number of scales. For a horizontal wavelet support, the HT groups consecutive residual coefficients $r_{s,j,o}(2x-1, y)$ and $r_{s,j,o}(2x, y)$ at scale j to compute the residual at the subsequent scale $j+1$:

$$r_{s,j+1,o}(x, y) = \frac{r_{s,j,o}(2x-1, y) + r_{s,j,o}(2x, y)}{2}. \quad (5)$$

The detail data is computed as a normalized difference of the consecutive residual coefficients:

$$d_{s,j+1,o}(x, y) = \frac{r_{s,j,o}(2x, y) - r_{s,j,o}(2x-1, y)}{2^j}. \quad (6)$$



(a) (b) (c) (d)

Fig. 4. Grouping associated wavelet coefficients: (a) shows the input image; (b) shows the association field at j = 1 over a vertically orientated wavelet plane (dark coefficients in the wavelet plane are negative, bright coefficients are positive and gray coefficients are close to zero). The association field (arrows) groups coefficients. The resultant grouplet detail plane in (c) is more sparse than the wavelet plane, preserving only the variations occurring at the corners and terminations; (d) shows the final saliency map (see section 4).

A GT is a Haar transform in which the residual and detail coefficients are computed between pairs of elements which are not necessarily consecutive, but are *paired along the contour to which they both belong*. To ascertain the contour along which coefficients should be paired, an "association field" is defined using a block matching algorithm. In this field, associations occur between points and their neighbors in the direction of maximum regularity. In this way, the association field encodes the anisotropic regularities present in the image. The regularities in $r_{s,j,o}$ are suppressed in $d_{s,j+1,o}$ by equation 6. Therefore, the GT is in essence a differencing operator applied to neighboring wavelet responses along a contour. Neighbors with similar values produce low responses in $d_{s,j+1,o}$ while those with differing values or singularities produce high responses, as illustrated in Fig. 4. By computing $d_{s,j,o} \forall j = 1, ..., J$, points are grouped across increasingly long distances. Each resultant grouplet plane is a sparser representation that contains comparatively higher coefficients for complex geometrical features, whilst simple features are suppressed.

In our saliency model, we apply the GT to wavelet coefficients in order to obtain this improved representation in which salient features are more prominent. It has been suggested that the hierarchical application of the GT to wavelet coefficients may mimic long-range horizontal connections between simple cells in area V1 [12].

4 SALIENCY ESTIMATION

In previous sections we made two hypotheses on what constitutes a salient visual stimulus. First, we claimed that a region is salient if its color is enhanced by the surround. We have shown that enhancement can depend on frequency, orientation and contrast of the surround. We proposed adapting a color induction model based on wavelets to indicate color contrasted regions. Second, we claimed that complex image features such as corners, terminations or crossings emerging from contours are salient. We proposed that a grouplet transform be used to enhance these complex features in the image representation.

Considering both hypotheses, here we propose a 6-stage model that estimates saliency by enhancing image locations with certain local spatio-chromatic properties and/or contour singularities. Our model contains the main stages of a color induction model [16], which uses a wavelet decomposition and a function that modulates wavelet coefficients according to their local properties. We introduce a grouplet transform that enables the grouping of simple features whilst maintaining singularities. Below, we describe the stages of our saliency model.

Stage (I): Color representation Three opponent color channels are obtained from image I by converting each (RGB) value, after γ correction, to the opponent space so that $O1 = \frac{R-G}{R+G+B}$, $O2 = \frac{R+G-2B}{R+G+B}$ and $O3 = R+G+B$.

Stage (II): Spatial decomposition Each channel is decomposed in two successive steps. The first one uses the wavelet transform in equation 1, obtaining $\{w_{s,o}\}$. Subsequently, on each wavelet plane the grouplet transform in equation 6 is applied:

$$I_c \xrightarrow{WT} \{\omega_{s,o}\} \xrightarrow{GT} \{d_{s,j,o}\} \tag{7}$$

where $d_{s,j,o}$ denotes the detail plane at scale j. For a wavelet plane whose largest dimension is size D, $J = log_2 D$. To group features, the association field for a wavelet plane is initialized perpendicularly to its orientation o. Thus for a horizontal wavelet plane, the Haar differencing in equation 6 is conducted column-wise. A diagonal wavelet plane captures high frequency information in both horizontal and vertical orientations. Therefore the grouplet transform is applied to such planes in both horizontal and vertical orientations separately, leading to two sets of grouplet planes for each diagonal wavelet plane.

Stage (III): Normalized Center Contrast (NCC) We compute the NCC, $z_{s,j,o}(x,y)$, for every grouplet coefficient $d_{s,j,o}(x,y)$ using equation 2.

Stage (IV): Induction weights ($ECSF$) The $ECSF$ function is used to compute induction weights $\alpha_{s,j,o}(x,y)$ for every grouplet coefficient $d_{s,j,o}(x,y)$:

$$\alpha_{s,j,o}(x,y) = ECSF(z_{s,j,o}(x,y), s). \tag{8}$$

The $ECSF_C$ function is used for channels $O1$ and $O2$, while $ECSF_I$ is used for channel $O3$. The $\alpha_{s,j,o}(x,y)$ weight gives a measure of saliency for location (x,y) in $d_{s,j,o}$. The $ECSF$ acts so that $z_{s,j,o}$ values with scales s in the passband of the $ECSF$ are enhanced, while those with scales outside of this passband are suppressed.

Each $\alpha_{s,j,o}$ plane is resized to the size of its corresponding wavelet plane $w_{s,o}$ using bicubic interpolation, and then summed to produce $\alpha_{s,o}$ for that wavelet plane:

$$\alpha_{s,o}(x,y) = \sum_j \varphi(\alpha_{s,j,o}(x,y)) \tag{9}$$

where $\varphi(\cdot)$ denotes bicubic interpolation.

Stages (V)-(VI): Saliency Map Recovery Finally, an inverse wavelet transform is performed on the spatial pyramid of $\alpha_{s,o}$ planes to produce the final saliency map S_c for an image channel. At this point the pipeline of the model may be summarized as

$$I_c \xrightarrow{WT} \{\omega_{s,o}\} \xrightarrow{GT} \{d_{s,j,o}\} \xrightarrow{NCC} \{z_{s,j,o}\} \xrightarrow{ECSF} \{\alpha_{s,j,o}\} \xrightarrow{\varphi} \{\alpha_{s,o}\} \xrightarrow{WT^{-1}} S_c$$

The saliency maps for all three image channels are combined to form the final saliency map S using the Euclidean norm $S = \sqrt{S_{O1}^2 + S_{O2}^2 + S_{O3}^2}$. The method, termed SIM for Saliency by Induction Mechanisms, is summarized schematically in Fig. 5.

Designing the center and surround regions

In stage III of the method, normalized center contrast is measured. The number of pixels spanning the center region and the extended region, comprising both the center and surround regions, were chosen so as to resemble the receptive and extra-receptive fields of V1 cortical cells respectively, in a similar fashion to Gao *et al.* [5]. Various studies [24], [25] estimate the central region of the receptive field in V1 cells to correspond on average to a visual angle, β, of approximately $1°$. The size of a feature, l, that subtends this visual angle when shown on a screen is computed as $l = d \cdot \tan\beta$, where d is the distance from the observer to the screen. Therefore, the number of pixels P_c that correspond to feature l is $P_c = (d \cdot \tan\beta)/(\frac{mon}{res})$, where mon is the size of the monitor and res is the average of the horizontal and vertical resolution of the displayed image. We used this P_c value as the diameter of the central region. The diameter of the extra-receptive field has been estimated to be at least 2 to 5 times that of the receptive field [26], [27]. We experimented with diameters in this range and found a size of 5.5 times that of the central region to perform well. These diameters were held constant throughout the image decomposition so that the effective sizes increase with spatial scale.

5 EXPERIMENTS

To evaluate SIM, we applied it to the problem of predicting eye-fixations in two image datasets. The accuracy of the predictions were quantitatively assessed using both the Kullback-Leibler (KL) divergence and the receiver operating characteristic (ROC) metrics. The KL divergence measures how well the method distinguishes between the histograms of saliency values at fixated and non-fixated locations in the image.

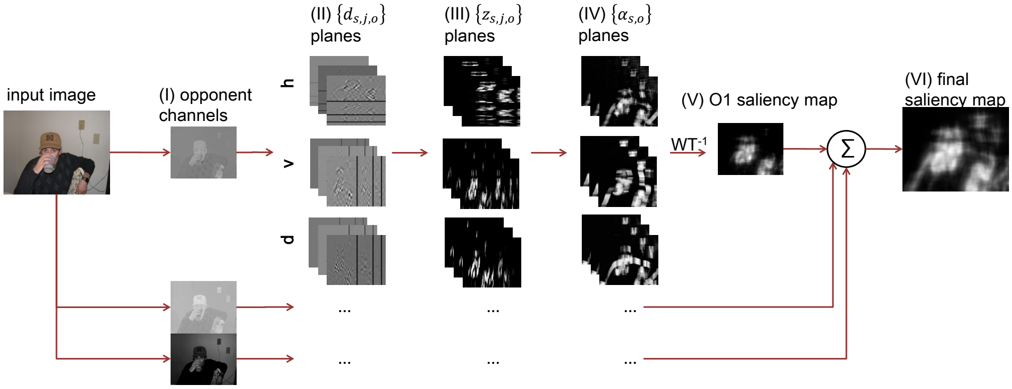


Fig. 5. Schematic of SIM: (I) The image is converted to the opponent space. (II) Each channel is decomposed into wavelet planes, and each wavelet plane is decomposed into grouplet planes (demarcated with black lines). (III) Contrast responses from grouplet planes are calculated and combined to produce contrast response planes. (IV) The $ECSF$ produces induction weights planes $\alpha_{s,o}$. (V) The $\alpha_{s,o}$ planes are combined by an inverse wavelet transform to produce the final channel map. (VI) The 3 channels maps are then combined.

The ROC curve measures how well the saliency map discriminates between fixated and non-fixated locations for different binary saliency thresholds. For both metrics, a higher value indicates better performance.

As noted by Zhang *et al.*, image border effects in several saliency methods result in artificial improvements in the ROC measure [8]. Therefore we adopt the evaluation framework described in [8] in order to avoid this issue and ensure a fair comparison of methods. This evaluation framework comprises modified metrics for the area under the ROC curve (AROC) and KL divergence metrics. For each image in the dataset, fixations for that image are denoted true positives, while the fixations for a randomly chosen different image in the dataset are denoted false positives for that image. With this formulation, any center bias of the true positive fixations with respect to the false positive fixations is avoided. The random selection of false positive fixations means that a new calculation of the metrics is likely to produce a different value. Therefore, in order to compute the standard error (SE), the metrics were computed 100 times. In addition, we shuffled the saliency maps 100 times. With each shuffling, the KL-divergence between the histograms of saliency values at unshuffled fixation points and shuffled fixation points was computed. We also performed 100 random permutations of the fixation points when calculating the AROC.

The first eye-fixation dataset used [6] is a popular benchmark dataset for comparing eye-fixation predictions between saliency models. It contains 120 color images, with 511x681 resolution, of indoor and outdoor scenes, along with the recorded eye-fixations of 20 subjects. The evaluation was performed on 7 state-of-the-art methods as well as SIM, both with and without the GT. The results are reported in Table 2. We see that, with or without the GT, SIM exceeds the state-of-the-art performance as measured by both metrics. Further, the addition of the GT improves upon SIM's performance.

TABLE 2
Performance on Bruce & Tsotsos dataset.

Model	KL (SE)	AROC (SE)
Itti [4]	0.1913 (0.0019)	0.6214 (0.0007)
AIM [6]	0.3228 (0.0023)	0.6711 (0.0006)
SUN [8]	0.2118 (0.0019)	0.6377 (0.0007)
GBVS [28]	0.1909 (0.0015)	0.6324 (0.0006)
Seo [9]	0.3558 (0.0027)	0.6783 (0.0007)
DVA [7]	0.3227 (0.0024)	0.6795 (0.0007)
SIGS [29]	0.3679 (0.0025)	0.6868 (0.0007)
SIM with GT, w/o ECSF	0.3786 (0.0029)	0.6877 (0.0008)
SIM w/o GT, with ECSF	0.4456 (0.0031)	0.7077 (0.0007)
SIM with GT, with ECSF	**0.4925** (0.0034)	**0.7136** (0.0007)

The second eye-fixation dataset used was provided by Judd *et al.* in [3]. This dataset contains 1,003 color images of varying dimensions, along with the recorded eye-fixations of 15 subjects. Because fixations must be compared across images, only those images whose dimensions were 768x1024 pixels were used, reducing the number of images examined to 463. The images in this dataset contain a greater number of semantic objects which are not modeled by bottom-up saliency, such as people, faces and text, and as such is more challenging than the first. Therefore, the AROC and KL divergence metrics are lower for all the saliency models, as one would expect. The results, shown in Table 3 indicate that once again SIM exceeds state-of-the-art performance.

Implementation Details

The Bruce & Tsotsos dataset was collected on a 21 inch monitor with $d = 29.5$ inches. For images with 511x681 resolution, the diameter of the central region, P_c, $= 18$ pixels. The Judd *et al.* dataset was collected on a 19 inch monitor with $d = 24$ inches. For images with 768x1024 resolution, $P_c = 24$ pixels. For a MATLAB implementation running on an Intel Core 2 Duo CPU at 3.00 GHz with 2GB RAM, typical run times for color images of sizes 128x128, 256x256 and 512x512 pixels are 0.6, 1.2 and 3.2 seconds respectively.

TABLE 3
Performance on Judd *et al.* dataset.

Model	KL (SE)	AROC (SE)
Itti [4]	0.2073 (0.0014)	0.6285 (0.0005)
AIM [6]	0.2647 (0.0016)	0.6506 (0.0004)
SUN [8]	0.1832 (0.0012)	0.6244 (0.0004)
GBVS [28]	0.1207 (0.0008)	0.5880 (0.0003)
Seo [9]	0.2749 (0.0015)	0.6479 (0.0004)
DVA [7]	0.2924 (0.0016)	0.6565 (0.0005)
SIGS [29]	0.2953 (0.0014)	0.6555 (0.0004)
SIM with GT, w/o ECSF	0.2885 (0.0016)	0.6618 (0.0005)
SIM w/o GT, with ECSF	0.3021 (0.0017)	0.6695 (0.0005)
SIM with GT, with ECSF	**0.3678** (0.0020)	**0.6788** (0.0005)

5.1 Discussion

Qualitative comparisons between two state-of-the-art methods [6], [9] and SIM are displayed in Figs. 6 and 9. One can see that for the proposed method (column (d)), the most salient regions correspond better to eye-fixations and highly salient features are located at a variety of spatial frequencies.

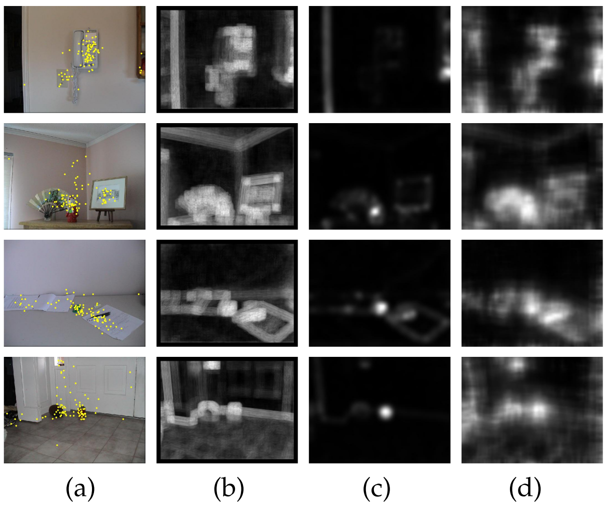


Fig. 6. Qualitative results for Bruce & Tsotsos dataset: Column (a) contains the original image. Columns (b), (c), and (d) contain saliency maps obtained from [6], [9] and SIM respectively. Yellow markers indicate eye fixations. Our method is seen to more clearly distinguish salient regions from background regions and to better estimate the extent of salient regions.

In addition, the model is less sensitive to low-frequency edges such as skylines and road curbs while avoiding excessive sensitivity to high-frequency textured regions. The weighting function $ECSF(z,s)$ is critical to these effects as it is more sensitive to mid-range frequencies, as Fig. 3 shows. As a result, it acts as a bandpass filter in the image's spatial frequency domain, and provides a biologically-inspired mechanism for combining spatial information at different scales. The importance of this combination is evidenced by the fact that SIM's performance decreases significantly (though it is still state-of-the-art) when the $ECSF$ is removed (see Tables 2 and 3). The GT further lowers sensitivity to low-frequency edges.

Scale selection and combination are inherent challenges for saliency estimation methods. The best-performing state-of-the-art methods ([**?**],[7],[29]) perform scale selection by choosing an image resolution that gives best performance as measured on saliency data. However, even when using data from the test domain, namely saliency estimation, the performance of these methods are lower than SIM's, which uses a scale combination method fitted using experimental data from a different problem domain, namely color perception prediction. In addition, Seo *et al.* reported no improvement when combining saliency maps computed at different scales [9]. Therefore the inclusion of an effective scale combination mechanism is one important way in which our method differs from previous ones.

One can also see in the figures that regions of high saliency are more clearly distinguished from background regions. Other methods may provide good localization for salient regions at few spatial scales [9] or may detect poorly localized regions at many spatial scales [6]. Our method strikes a good balance between localization of salient regions and detection of salient regions at different spatial scales. This is reflected in the large improvements in KL divergence achieved for both datasets. The increased discriminative power is due to the fact that the background features present in the wavelet planes are attenuated by the grouplet transform, as illustrated in Fig. 7. These background features tend to be small, isolated features which, while present in wavelet planes, do not persist beyond the first few grouplet planes.

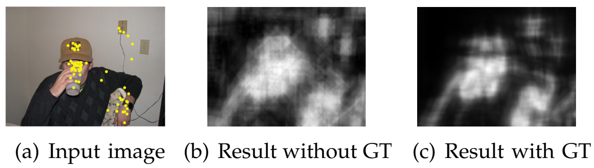


Fig. 7. The GT attenuates spatially isolated features.

The grouplet transform itself may be considered a center-surround mechanism, as it measures the difference in amplitude between a coefficient and its neighbor. Consequently, regions of the wavelet plane with similar amplitudes, and therefore low contrast, are attenuated in their grouplet planes, while regions of the wavelet plane with large differentials between their amplitudes are enhanced. Therefore the grouplet transform acts to further distill the information present in the wavelet transform, preserving only features which are spatially extensive and strongly contrasting with their surroundings.

Our model required parameters to be set for the $ECSF$ and the center-surround regions. The $ECSF$ parameters were set using psychophysical data and are dataset-independent. Therefore our only free parameters are the center-surround region sizes. As mentioned in section 4, the center region's size was set to correspond to 1° of visual angle, and the surround size was set to be 5.5 times the size of the center region. We found results to be very stable for surround-to-center region ratios from 3-6 and for center sizes of $1^\circ \pm 0.2$. As such, our model is robust to uncertainty in the choice of free parameters.

We also investigated the effect of varying the spatial scale for which the $ECSF(z, s)$ gives the highest response, denoted by s_0. We varied s_0 for the $ECSF$ of the intensity channel, the channel containing the majority of the saliency information. Fig. 8 shows that SIM performs best when mid-range frequencies are enhanced and low or high frequencies are inhibited. Furthermore, the best scale range for these metrics, between 4 and 6, is consistent with the value determined using psychophysical data, $s_0 = 4.2$ (see Fig. 3(a)).

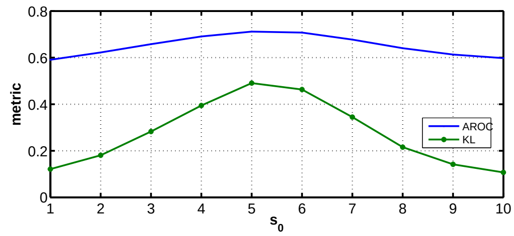
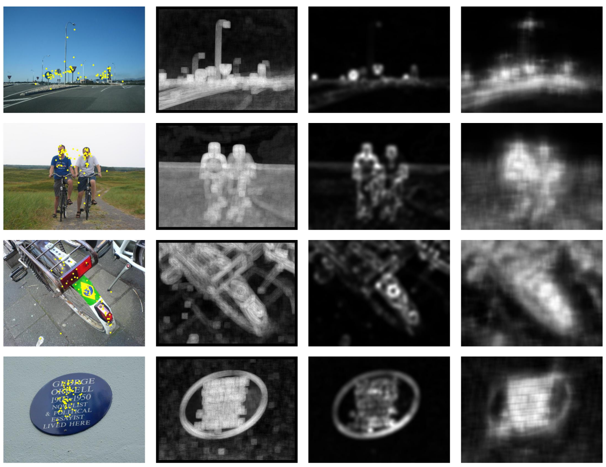


Fig. 8. Change in AROC and KL metrics with change in s_0 for intensity $ECSF(z, s)$, for the Bruce & Tsotsos dataset, using SIM with GT. The best s_0 for both these metrics are in line with the value determined using psychophysical experiments.



(a) (b) (c) (d)

Fig. 9. Qualitative results for Judd *et al.* dataset: Column (a) contains the original image. Columns (b), (c), and (d) contain saliency maps obtained from [6], [9] and SIM respectively. Yellow markers indicate eye fixations.

6 CONCLUSIONS

We proposed a saliency model, SIM, based on a biologically-inspired low-level spatio-chromatic representation. SIM measures saliency using the result of the perceptual integration of color, orientation, local spatial frequency and surround contrast. The parameters of our integration mechanisms have been fitted to psychophysical data. In addition, we have shown that saliency estimation is improved if we insert a grouping stage that suppresses simple edges, thereby avoiding strong saliency responses for such features. We demonstrate that SIM exceeds state-of-the-art performance in predicting eye-fixations on two datasets and using two metrics. Its success raises an intriguing topic for future research, namely, whether there exists a similar architecture for both the low-level visual saliency machinery and the colour perception machinery in humans.

ACKNOWLEDGMENTS

This work has been supported by Projects 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] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 99, no. PrePrints, 2012.

[2] W. Kienzle, F. Wichmann, B. Schalkopf, and M. O. Franz, "A nonparametric approach to bottom-up visual saliency," in *Advances in neural information processing systems 19*. MIT Press, 2007.

[3] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE Int'l Conf. Computer Vision*, 2009.

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

[5] D. Gao, V. Mahadevan, and N. Vasconcelos., "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *Journal of Vision*, vol. 8, no. 7:13, pp. 1–18, 2008.

[6] N. D. Bruce and J. K. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. MIT Press: MIT Press, 2006, pp. 155–162.

[7] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," *Advances in neural information processing systems*, vol. 21, pp. 681–688, 2008.

[8] 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*, vol. 8, no. 7, pp. –, 2008.

[9] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance." *Journal of Vision*, vol. 9, no. 12, pp. 15.1–27, 2009.

[10] Q. Zhao and C. Koch, "Learning visual saliency by combining feature maps in a nonlinear manner using adaboost," *Journal of Vision*, vol. 12, no. 6, 2012.

[11] N. Pinto, D. Doukhan, J. DiCarlo, and D. Cox, "A high-throughput screening approach to discovering good forms of biologically inspired visual representation," *PLoS computational biology*, vol. 5, no. 11, p. e1000579, 2009.

[12] S. Mallat, "Geometrical grouplets," *Applied and Computational Harmonic Analysis*, vol. 26, no. 2, pp. 161 – 180, 2009.

[13] N. Murray, M. Vanrell, X. Otazu, and C. Parraga, "Saliency estimation using a non-parametric low-level vision model," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 433–440.

[14] S. Zeki, J. Watson, C. Lueck, K. Friston, C. Kennard, and R. Frackowiak, "A direct demonstration of functional specialization in human visual cortex," *The Journal of Neuroscience*, vol. 11, no. 3, pp. 641–649, 1991.

[15] R. Shapley and M. J. Hawken, "Color in the cortex: single- and double-opponent cells," *Vision Research*, vol. 51, no. 7, pp. 701 – 717, 2011, vision Research 50th Anniversary Special Issue.

[16] X. Otazu, C. A. Parraga, and M. Vanrell, "Toward a unified chromatic induction model," *Journal of Vision*, vol. 10, no. 12, 2010.

[17] E. P. Simoncelli and O. Schwartz, "Modeling surround suppression in v1 neurons with a statistically-derived normalization model," in *Advances in neural information processing systems 2*. MIT Press, 1999, pp. 153–159.

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

[19] B. Blakeslee and M. E. McCourt, "Similar mechanisms underlie simultaneous brightness contrast and grating induction," *Vision Research*, vol. 37, no. 20, pp. 2849–2869, 1997.

[20] F. Attneave, "Some informational aspects of visual perception," *Psychol Rev*, vol. 61, no. 3, pp. 183–93, 1954.

[21] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, Mar 2001.

[22] C. Zetzsche, K. Schill, H. Deubel, G. Krieger, E. Umkehrer, and S. Beinlich, "Investigation of a sensorimotor system for saccadic scene analysis: an integrated approach," in *Proceedings of the fifth international conference on simulation of adaptive behavior on From animals to animats 5*. Cambridge, MA, USA: MIT Press, 1998, pp. 120–126.

[23] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision Research*, vol. 45, no. 8, pp. 2397–2416, Aug 2005.

[24] J. R. Cavanaugh, W. Bair, J. A. Movshon, J. R, W. Bair, and J. A. Movshon, "Nature and interaction of signals from the receptive field center and surround in macaque v1 neurons," *J Neurophysiol*, pp. 2530–2546, 2002.

[25] A. Smith, K. Singh, A. Williams, and M. Greenlee, "Estimating Receptive Field Size from fMRI Data in Human Striate and Extrastriate Visual Cortex," *Cerebral Cortex*, vol. 11, no. 12, pp. 1182–1190, 2001.

[26] L. Chao-Yi and L. Wu, "Extensive integration field beyond the classical receptive field of cat's striate cortical neurons–classification and tuning properties," *Vision Research*, vol. 34, no. 18, pp. 2337 – 2355, 1994.

[27] G. Walker, I. Ohzawa, R. Freeman *et al.*, "Suppression outside the classical cortical receptive field," *Visual neuroscience*, vol. 17, no. 3, pp. 369–379, 2000.

[28] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Advances in neural information processing systems*, vol. 19, p. 545, 2007.

[29] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, p. 194, 2012.