Low-Level Spatiochromatic 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 based on a low-level spatiochromatic model that has successfully predicted chromatic induction phenomena. In so doing, we hypothesize that the low-level visual mechanisms that enhance or suppress image detail are also responsible for making some image regions more salient. Moreover, SIM adds geometrical grouplets to enhance complex low-level features such as corners, and suppress relatively simpler features such as edges. Since our model has been fitted on psychophysical chromatic induction data, it is largely nonparametric. 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 saliency [1], from biologically inspired models to learning-based approaches. Among the more bioinspired models, Itti et al.'s [2] 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 is applied to the input image. Each feature map contains information from one of three cues: orientation, color, or scale. Gao et al. [3] 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 and Tsotsos [4] 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 [4], an ICA basis set of filters was learned from RGB patches extracted from images and used to represent the local patches. As was also found by Hou and Zhang [5] in a similar approach, the basis set consisted mainly of oriented Gabor-like patches with opponent color properties. Zhang et al. [6] also proposed a method that 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 and Milanfar [7] used kernel regression-based self-resemblance to compute saliency and considered a region to be salient when its curvature was different from that of its surround.

In these bioinspired approaches, there remain several major challenges, including:

- generating the optimal feature maps for estimating saliency [8],
- holistically combining saliency information from these feature maps, which are extracted from multiple scales, orientations and color channels [9], and
- selecting the many model parameters (such as the number, type, and orientation of filters, and coefficients for nonlinear normalizations and activation functions) present in such models [10].

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

- We adapt a low-level color induction model in order to predict saliency. The resultant saliency model inherits an extended contrast sensitivity function (*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 saliency by induction mechanisms (SIM).
- We use "geometrical grouplets" [11] to produce a sparse and efficiently computed image representation that enhances features known to guide attention and suppresses nonsalient 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 paper 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 quantitative and qualitative experimental results, and we draw several conclusions in Section 6. A preliminary version of this work appeared in [12].

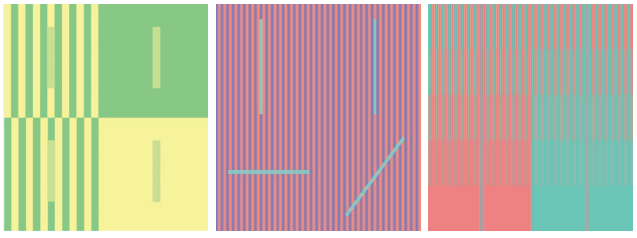
2 Modeling Low-Level Color Vision

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 [13]. This modular paradigm was adopted by Itti et al. for saliency [2]. In the intervening years, however, a large body of evidence has emerged that supports the view of a more interlinked processing of color and form in the human visual cortex [14].

In this work, we adapt a computational model of color perception [15] to the problem of saliency estimation. The model is based on a nonmodular 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 Otazu et al. [15] is motivated by our hypothesis that *factors related to color induction phenomena also inform on local saliency*.

- *N. Murray is with the Xerox Research Centre Europe, 6 Chemin de Maupertuis, Meylan 38240, France. E-mail: nmurray@cvc.uab.es.*
- *M. Vanrell, X. Otazu, and C.A. Parraga are with the Centre de Visió per Computador, Edifici O Campús UAB, Cerdanyola del Vallès (Bellaterra), Barcelona 08193, Spain.*
  E-mail: {maria, xotazu, Alejandro.Parraga}@cvc.uab.es.

Manuscript received 9 Aug. 2012; revised 30 Jan. 2013; accepted 14 May 2013; published online 10 June 2013.
Recommended for acceptance by Y. Matsushita.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2012-08-0613.
Digital Object Identifier no. 10.1109/TPAMI.2013.108.



(a) (b) (c)

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

The model of Otazu et al. [15] 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. 1a shows how surround spatial frequency affects the perceived colors of four 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 stimulus 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 stimulus more greenish when surrounded by yellow). These induction effects are termed assimilation and contrast, respectively.

Second, *orientation* also influences color appearance. In Fig. 1b, we can observe that the relative orientation between the stimulus and the surround provokes a perceptual change. While the top-left and top-right stimuli clearly undergo assimilation (a greenish perception when surrounded by pink and a bluish perception when surrounded by blue), the stimuli at the 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

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

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. Note that we cannot use an exact Gabor transform as it does not have a complete inverse transform, a property which will be required in a later stage of our method.

Third, *surround contrast* also plays a crucial role in how color is perceived. As shown in Fig. 1c, 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 (NCC), $z_{x,y}$, 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 $a^{sur}_{x,y}$:

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

Divisive normalization has been shown by Simoncelli and Schwartz [16] to remove statistical dependences 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 *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 [17] for luminance and color contrast.

The *ECSF* we use is a function of spatial scale s and NCC 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 approximation to the psychophysical CSF and is itself introducing assimilation or contrast depending on the spatial frequency s. Function $g(s)$ is defined as

$$g(s) = \begin{cases} \beta e^{-\frac{(s-s^g_0)^2}{2\sigma_1^2}}, & s \le s^g_0, \\ \beta e^{-\frac{(s-s^g_0)^2}{2\sigma_2^2}}, & \text{otherwise.} \end{cases} \tag{3}$$

Here, β is a scaling constant, and σ_1 and σ_2 define the spread of the spatial sensitivity of $g(s)$. The s^g_0 parameter defines the peak scale sensitivity of $g(s)$. An additional function, $k(s)$, was introduced to ensure a nonzero lower bound on $ECSF(z, s)$:

$$k(s) = \begin{cases} e^{-\frac{(s-s^k_0)^2}{2\sigma_3^2}}, & s \le s^k_0, \\ 1, & \text{otherwise.} \end{cases} \tag{4}$$

Here, σ_3 defines the spread of the spatial sensitivity of $k(s)$ and s^k_0 defines the peak scale sensitivity of $k(s)$.

In the induction model of Otazu et al. [15], 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 and McCourt [18], 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. [15] 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 these data (given in Table 1) in the perceived image output by the induction model. Examples of stimuli used in these experiments are shown in Fig. 2.

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 Fig. 2). Their profiles are

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

shown in Fig. 3. The functions enhance NCC in a narrow passband and suppress this contrast for low spatial scales. The magnitude of the enhancement increases with the magnitude of the NCC, z, as observed in Figs. 3a and 3b. These *ECSF*s have peak spatial scales in the wavelet decomposition that correspond to peak spatial frequencies between 2 and 5 cpd, which agree with previous psychophysical estimations [17].

As stated above, we use a wavelet transform with Gabor-like basis functions as an image representation. This representation agrees with a long-standing view of the early human sensory system as an efficient information processing system [19], [20]. 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 superimposed orientations, such as corners and terminations. However, in comparison with edges, complex features are preferentially fixated on when humans free-view natural images, [21], [22]. 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 GT for Image Representation

The GT [11] 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 (HT), computed using a lifting scheme.

3.1 The GT as a Modified HT

The 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 $r_{s,j,o}$ are 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

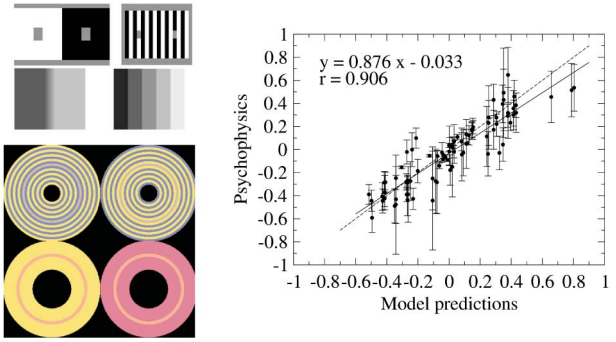
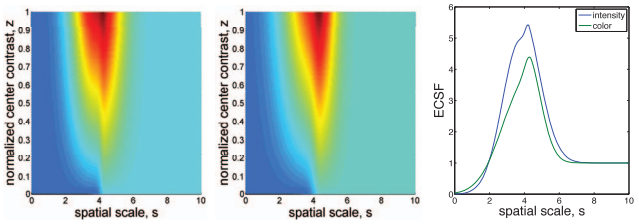


Fig. 2. (a) Examples of experimental stimuli. (b) Correlation between model prediction and psychophysical data. The solid line represents the model linear regression fit and the dashed line represents the ideal fit, i.e., perfect correlation. Since measurements involve dimensionless measures and physical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



(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) 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).

$r_{s,j,o}(2x,y)$ at scale j to compute the residual at the subsequent scale $j+1$:

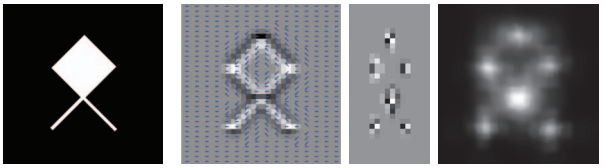
$$r_{s,j+1,o}(x,y) = \frac{r_{s,j,o}(2x-1,y) + r_{s,j,o}(2x,y)}{2}. \quad (5)$$

The detail data are 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 GT is an HT in which the residual and detail coefficients are computed between pairs of elements that 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 (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 $d_{s,j,o}\ \forall j = 1,\ldots,J$, points are grouped across increasingly long distances. Each resultant grouplet plane is a sparser representation that contains comparatively higher coefficients for complex geometrical features, while 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 [11].



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

Fig. 4. Grouping-associated wavelet coefficients. (a) The input image. (b) 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) The final saliency map (see Section 4).

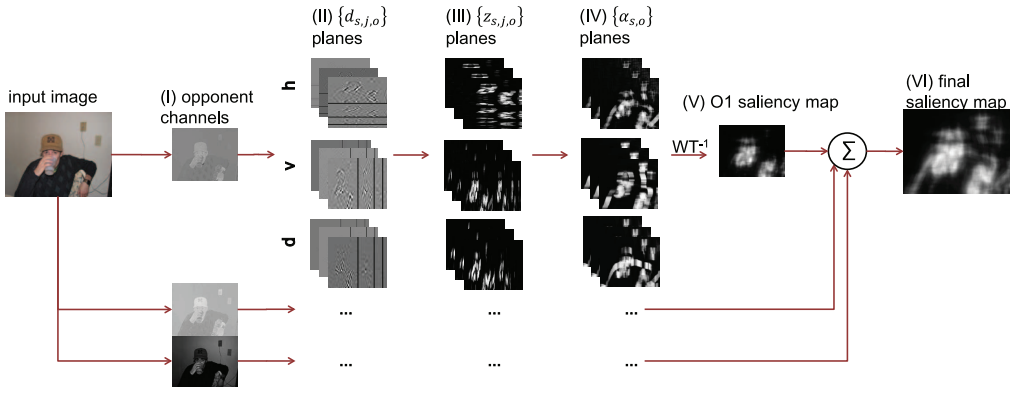


Fig. 5. Schematic of SIM. (I) The image is converted to the opponent space. (II) Each channel is decomposed into wavelet planes, and each wavelet plane is decomposed into grouplet planes (demarcated with black lines). (III) Contrast responses from grouplet planes are calculated and combined to produce 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 three-channel maps are then combined.

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 contrast regions. Second, we claimed that complex image features such as corners, terminations, or crossings emerging from contours are salient. We proposed that a GT be used to enhance these complex features in the image representation.

Considering both hypotheses, here we propose a six-stage model that estimates saliency by enhancing image locations with certain local spatiochromatic properties and/or contour singularities. Our model contains the main stages of a color induction model [15] which uses a wavelet decomposition and a function that modulates wavelet coefficients according to their local properties. We introduce a GT that enables the grouping of simple features while 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 into two successive steps. The first one uses the wavelet transform in (1), obtaining $\{w_{s,o}\}$. Subsequently, on each wavelet plane, the GT in (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 (6) is conducted column-wise. A diagonal wavelet plane captures high-frequency information in both horizontal and vertical orientations. Therefore, the GT 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): NCC. We compute the NCC, $z_{s,j,o}(x, y)$, for every grouplet coefficient $d_{s,j,o}(x, y)$ using (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, is summarized schematically in Fig. 5.

4.1 Designing the Center and Surround Regions

In stage III of the method, NCC is measured. The number of pixels spanning the center region and the extended region, comprising both the center and surround regions, was chosen so as to resemble the receptive and extra-receptive fields of V1 cortical cells, respectively, in a similar fashion to Gao et al. [3]. Various studies [23], [24] 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, P_{e-r}, has been estimated to be at least two to five times that of the receptive field [25], [26]. 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

TABLE 2
Performance on the Bruce and Tsotsos Dataset

Model	KL (SE)	AROC (SE)
Itti [2]	0.1913 (0.0019)	0.6214 (0.0007)
AIM [4]	0.3228 (0.0023)	0.6711 (0.0006)
SUN [6]	0.2118 (0.0019)	0.6377 (0.0007)
GBVS [27]	0.1909 (0.0015)	0.6324 (0.0006)
Seo [7]	0.3558 (0.0027)	0.6783 (0.0007)
DVA [5]	0.3227 (0.0024)	0.6795 (0.0007)
SIGS [28]	0.3679 (0.0025)	0.6868 (0.0007)
SIM w/o GT	0.4456 (0.0031)	0.7077 (0.0007)
SIM	**0.4925** (0.0034)	**0.7136** (0.0007)
SIM w/o $ECSF$	0.3786 (0.0029)	0.6877 (0.0008)
SIM with tuned P_c, P_{e-r}	0.4920 (0.0034)	0.7138 (0.0007)

TABLE 3
Performance on the Judd et al. Dataset

Model	KL (SE)	AROC (SE)
Itti [2]	0.2073 (0.0014)	0.6285 (0.0005)
AIM [4]	0.2647 (0.0016)	0.6506 (0.0004)
SUN [6]	0.1832 (0.0012)	0.6244 (0.0004)
GBVS [27]	0.1207 (0.0008)	0.5880 (0.0003)
Seo [7]	0.2749 (0.0015)	0.6479 (0.0004)
DVA [5]	0.2924 (0.0016)	0.6565 (0.0005)
SIGS [28]	0.2953 (0.0014)	0.6555 (0.0004)
SIM w/o GT	0.3021 (0.0017)	0.6695 (0.0005)
SIM	**0.3678** (0.0020)	**0.6788** (0.0005)
SIM w/o $ECSF$	0.2885 (0.0016)	0.6618 (0.0005)
SIM with tuned P_c, P_{e-r}	0.3663 (0.0020)	0.6774 (0.0005)

constant throughout the image decomposition so that the effective sizes increase with the 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 was 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 nonfixated locations in the image. The ROC curve measures how well the saliency map discriminates between fixated and nonfixated 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 [6]. Therefore, we adopt the evaluation framework described in [6] 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), both metrics were computed 100 times, each time using a different random permutation of the fixation points as false positives. The KL divergence between the histograms of saliency values at true-positive fixation points and false-positive fixation points was computed.

The first eye-fixation dataset used [4] is a popular benchmark dataset for comparing eye-fixation predictions between saliency models. It contains 120 color images, with 511×681 resolution, of indoor and outdoor scenes, along with the recorded eye fixations of 20 subjects, to whom the images were presented for 4 seconds. The evaluation was performed on seven state-of-the-art methods as well as SIM. The results are reported in Table 2. We see that, even 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.

The second eye-fixation dataset used was provided by Judd et al. [29]. This dataset contains 1,003 color images of varying dimensions, along with the recorded eye fixations of 15 subjects, to whom the images were presented for 3 seconds. Because fixations must be compared across images, only those images whose dimensions were $768 \times 1{,}024$ 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 and Tsotsos dataset was collected on a 21-inch monitor with $d = 29.5$ inches. For images with 511×681 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 $768 \times 1{,}024$ resolution, $P_c = 24$ pixels. For a MATLAB implementation running on an Intel Core 2 Duo CPU at 3.00 GHz with 2 GB RAM, typical run times for color images of sizes 128×128, 256×256, and 512×512 pixels are 0.6, 1.2, and 3.2 seconds, respectively.

5.1 Discussion

Qualitative comparisons between two state-of-the-art methods [4], [7] 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.

In addition, the model is less sensitive to low-frequency edges such as skylines and road curbs, while avoiding excessive

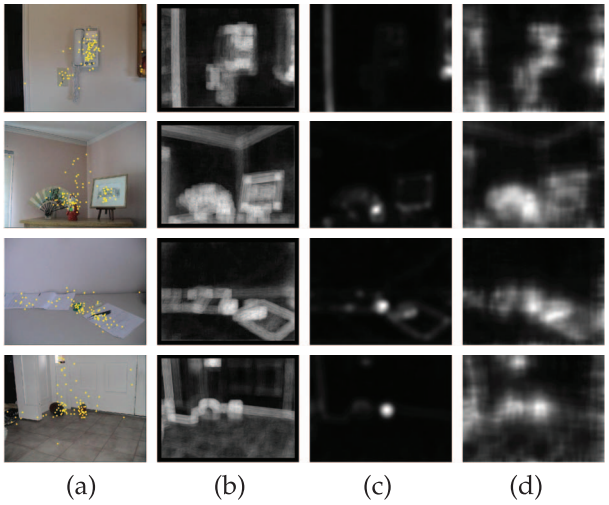
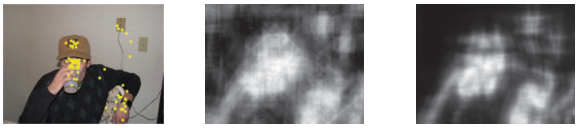


Fig. 6. Qualitative results for the Bruce and Tsotsos dataset. Column (a) contains the original image. Columns (b), (c), and (d) contain saliency maps obtained from [4], [7], and SIM, respectively. Yellow markers indicate eye fixations. Our method is seen to more clearly distinguish salient regions from background regions and to better estimate the extent of salient regions.



(a) Input image (b) Result without GT (c) Result with GT

Fig. 7. The GT attenuates spatially isolated features.

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 shown in Fig. 3. 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, SIM w/o *ECSF*). The GT further lowers sensitivity to low-frequency edges.

Scale selection and combination are required for all saliency estimation methods and have proven challenging. Most state-of-the-art methods (e.g., [7], [5], [28]) perform scale selection by simply choosing an image resolution that gives best performance as measured on eye-fixation data test sets. However, even when using data from the test domain, the performances 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 induction prediction. In addition, Seo and Milanfar reported no improvement when combining saliency maps computed at different scales [7]. 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 [7] or may detect poorly localized regions at many spatial scales [4]. 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 GT, 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.

The GT 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 GT acts to further distill the information present in the wavelet transform, preserving only features that 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. However, when the viewing conditions of the images are unknown, P_c and P_{e-r} cannot be determined in this manner. In such a case, these values may be fitted as hyperparameters of the model. We found that for $P_c = 17$ and $P_{e-r} = 91$, SIM maintains its performance for both metrics and both datasets (see Tables 2 and 3, SIM with tuned P_c, P_{e-r}). Moreover, the performance is quite stable for a wide range of

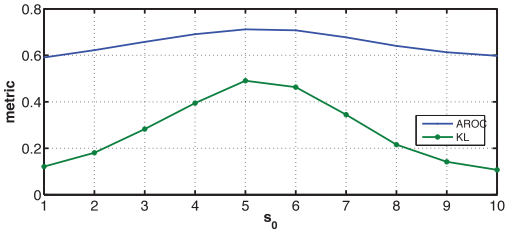


Fig. 8. Change in AROC and KL metrics with change in s_0 for intensity $ECSF(z, s)$, for the Bruce and Tsotsos dataset, using SIM with GT. The best s_0 for both these metrics is in line with the value determined using psychophysical experiments.

values of P_c and P_{e-r} (see Section 1 of the supplemental material, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2013.108, for related experimental results). 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. 3a).

6 Conclusions

We proposed a saliency model, SIM, based on a biologically inspired low-level spatiochromatic 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 question for further research, namely, whether the model designed to predict color perception and adapted to saliency estimation can be used to model other low-level visual tasks.

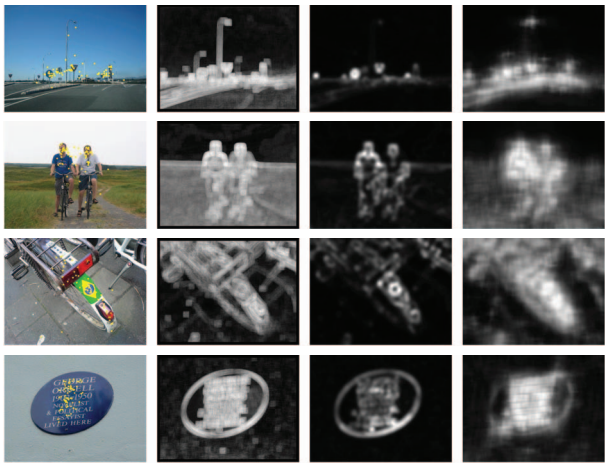


Fig. 9. Qualitative results for the Judd et al. dataset. Column (a) contains the original image. Columns (b), (c), and (d) contain saliency maps obtained from [4], [7], and SIM, respectively. Yellow markers indicate eye fixations.

ACKNOWLEDGMENTS

This work was supported by Projects TIN2010-21771-C02-1, 2009-SGR-669, 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. 35, no. 1, pp. 185-207, Jan. 2013.
[2] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 11, pp. 1254-1259, Nov. 1998.
[3] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the Plausibility of the Discriminant Center-Surround Hypothesis for Visual Saliency," *J. Vision,* vol. 8, no. 7, pp. 1-18, 2008.
[4] N.D. Bruce and J.K. Tsotsos, "Saliency Based on Information Maximization," *Advances in Neural Information Processing Systems 18,* Y. Weiss, B. Schölkopf, and J. Platt, eds., pp. 155-162, MIT Press, 2006.
[5] X. Hou and L. Zhang, "Dynamic Visual Attention: Searching for Coding Length Increments," *Proc. Conf. Advances in Neural Information Processing Systems,* vol. 21, pp. 681-688, 2008.
[6] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, and G.W. Cottrell, "SUN: A Bayesian Framework for Saliency Using Natural Statistics," *J. Vision,* vol. 8, no. 7, 2008.
[7] H.J. Seo and P. Milanfar, "Static and Space-Time Visual Saliency Detection by Self-Resemblance," *J. Vision,* vol. 9, no. 12, pp. 1-27, 2009.
[8] W. Kienzle, F. Wichmann, B. Schalkopf, and M.O. Franz, "A Nonparametric Approach to Bottom-Up Visual Saliency," *Advances in Neural Information Processing Systems 19,* MIT Press, 2007.
[9] Q. Zhao and C. Koch, "Learning Visual Saliency by Combining Feature Maps in a Nonlinear Manner Using Adaboost," *J. Vision,* vol. 12, no. 6, 2012.
[10] 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.
[11] S. Mallat, "Geometrical Grouplets," *Applied and Computational Harmonic Analysis,* vol. 26, no. 2, pp. 161-180, 2009.
[12] N. Murray, M. Vanrell, X. Otazu, and C. Parraga, "Saliency Estimation Using a Non-Parametric Low-Level Vision Model," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 433-440, 2011.
[13] S. Zeki, J. Watson, C. Lueck, K. Friston, C. Kennard, and R. Frackowiak, "A Direct Demonstration of Functional Specialization in Human Visual Cortex," *The J. Neuroscience,* vol. 11, no. 3, pp. 641-649, 1991.
[14] 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.
[15] X. Otazu, C.A. Parraga, and M. Vanrell, "Toward a Unified Chromatic Induction Model," *J. Vision,* vol. 10, no. 12, 2010.
[16] E.P. Simoncelli and O. Schwartz, "Modeling Surround Suppression in V1 Neurons with a Statistically-Derived Normalization Model," *Advances in Neural Information Processing Systems 2,* pp. 153-159, MIT Press, 1999.
[17] K. Mullen, "The Contrast Sensitivity of Human Color-Vision to Red Green and Blue Yellow Chromatic Gratings," *J. Physiology,* vol. 359, pp. 381-400, 1985.
[18] 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.
[19] F. Attneave, "Some Informational Aspects of Visual Perception," *Psychological Rev.,* vol. 61, no. 3, pp. 183-193, 1954.
[20] L. Itti and C. Koch, "Computational Modelling of Visual Attention," *Nature Rev. Neuroscience,* vol. 2, no. 3, pp. 194-203, Mar. 2001.
[21] 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," *Proc. Fifth Int'l Conf. Simulation of Adaptive Behavior: From Animals to Animats 5,* pp. 120-126, 1998.
[22] 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.
[23] J.R. Cavanaugh, W. Bair, and J.A. Movshon, "Nature and Interaction of Signals from the Receptive Field Center and Surround in Macaque v1 Neurons," *J. Neurophysiology,* vol. 88, pp. 2530-2546, 2002.
[24] 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.
[25] 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.
[26] G. Walker et al., "Suppression Outside the Classical Cortical Receptive Field," *Visual Neuroscience,* vol. 17, no. 3, pp. 369-379, 2000.
[27] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency," *Proc. Conf. Advances in Neural Information Processing Systems,* vol. 19, p. 545, 2007.
[28] X. Hou, J. Harel, and C. Koch, "Image Signature: Highlighting Sparse Salient Regions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 34, no. 1, pp. 194-201, Jan. 2012.
[29] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to Predict Where Humans Look," *Proc. IEEE Int'l Conf. Computer Vision,* 2009.