Exploiting Local and Global Patch Rarities for Saliency Detection

Ali Borji
USC
borji@usc.edu

Laurent Itti
USC
itti@usc.edu

Abstract

We introduce a saliency model based on two key ideas. The first one is considering local and global image patch rarities as two complementary processes. The second one is based on our observation that for different images, one of the RGB and Lab color spaces outperforms the other in saliency detection. We propose a framework that measures patch rarities in each color space and combines them in a final map. For each color channel, first, the input image is partitioned into non-overlapping patches and then each patch is represented by a vector of coefficients that linearly reconstruct it from a learned dictionary of patches from natural scenes. Next, two measures of saliency (Local and Global) are calculated and fused to indicate saliency of each patch. Local saliency is distinctiveness of a patch from its surrounding patches. Global saliency is the inverse of a patch's probability of happening over the entire image. The final saliency map is built by normalizing and fusing local and global saliency maps of all channels from both color systems. Extensive evaluation over four benchmark eye-tracking datasets shows the significant advantage of our approach over 10 state-of-the-art saliency models.

1. Introduction

The human visual system has to process an enormous amount of incoming information ($\sim 10^8$ bit/s) from the retina. Similarly, in computer vision, many systems suffer from the high computational complexity of inputs, especially when these systems are supposed to work in real time. Visual saliency is a concept that offers efficient solutions for both biological and artificial vision systems. It is basically a process that detects scene regions different from their surroundings (often referred as bottom-up saliency). Then, higher cognitive and usually more complex operations are focused only on the selected areas.

Recently, modeling visual saliency has raised much interest in theory and applications (see [47] for a review). For example in computer vision, it has been used for image and video compression [49], image segmentation, and object recognition [52]. In computer graphics, detecting salient regions has been employed for content-aware image cropping, photo collage [50], and stylization of images [53]. Saliency computation has also applications in other areas such as advertisement design [51] and visual prosthetics [48]. Our focus in this paper is proposing a new and more predictive (with respect to human eye tracking data) model of bottom-up visual saliency by integrating local and global saliency detection in both RGB and Lab color spaces (see Fig. 1).

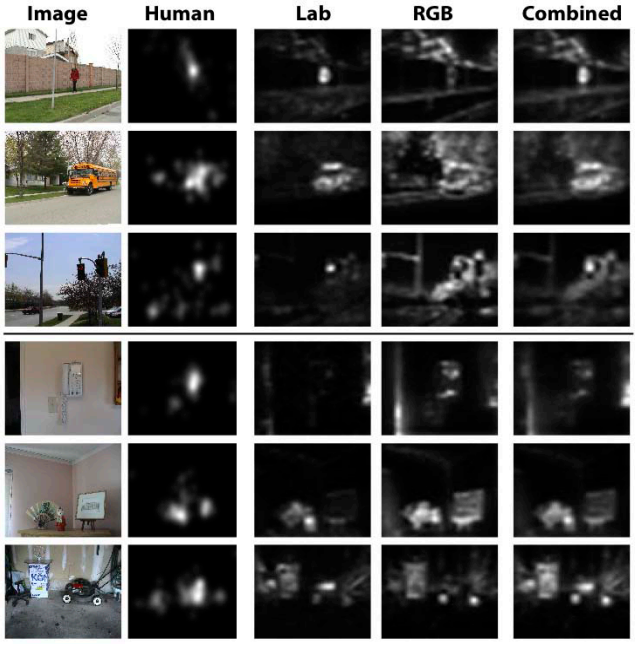

Figure 1. **One color system does not work for all images.** Top (Bottom): Sample images where our model is able to detect the outliers in CIE Lab (RGB) color space. For some images both color spaces work equally well. Last column shows combined maps from both color spaces. Images are taken from the TORONTO dataset [14].

Related works on saliency modeling. A majority of computational models of attention follow the structure adapted from the Feature Integration Theory (FIT) [15] and the Guided Search model [1]. Koch and Ullman [19] proposed a computational architecture for this theory and Itti *et al.* [4] were among the first ones to fully implement and maintain it. The main idea here is to compute saliency in

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

each of several features (e.g., color, intensity, orientation; saliency is then the relative difference between a region and its surrounding) in parallel, and to fuse them in a scalar map called the "saliency map". Le Meur *et al.* [18] adapted the Koch-Ullman's model to include features of contrast sensitivity, perceptual decomposition, visual masking, and center-surround interactions. Some models have added features such as symmetry [20], texture contrast [36], curvedness [21], or motion [41] to the basic structure.

In addition to the mentioned cognitive models, several probabilistic models of visual saliency have been developed over the past years. In these models, a set of statistics or probability distributions are computed from either the current scene, or from a set of natural scenes over space or time or both. Itti and Baldi [10] defined surprising stimuli as those which significantly change beliefs of an observer, measured as the Kullback-Leibler (KL) distance between posterior and prior beliefs. Harel *et al.* [7] used graph algorithms and a measure of dissimilarity to achieve efficient saliency computation with their Graph Based Visual Saliency (GBVS) model. Torralba *et al.*'s contextual guidance model [26] consolidates low-level salience and scene context when guiding search. Areas of high salience within a selected contextual region are given higher weights on an activation map than those that fall outside the selected contextual region. Some Bayesian models formulate visual search and derive a measure of bottom-up saliency as a by-product. For example, Zhang *et al.*'s model [12], Saliency Using Natural statistics (SUN), combines top-down and bottom-up information to guide eye movements during real-world object search tasks. However, unlike Torralba *et al.*'s model, SUN implements target features as the top-down component. Gao and Vasconcelos [23] define saliency as maximizing classification accuracy. They utilize the KL distance to measure mutual information between features at a scene location and class labels. The higher mutual information between a region and class of interest, the higher the saliency of that region. Seo and Milanfar [11] using local regression kernels build a "self-resemblance" map, which measures the similarity of a feature matrix at a pixel of interest to its neighboring feature matrices.

Bruce and Tsotsos [14] proposed the Attention based on Information Maximization (AIM) model by employing the first principles of information theory. They model bottom-up saliency as the maximum information sampled from an image. More specifically, saliency is computed as Shannon's self-information $-log\ p(f)$, where f is a local visual feature. Hou and Zhang [9] introduced the Incremental Coding Length (ICL) approach to measure the respective entropy gain of each feature. The goal is to maximize the entropy of the sampled visual features.

Some models measure saliency in the frequency domain. Hou and Zhang [8] propose a method based on relating extracted spectral residual features of an image in the spectral domain to the spatial domain. Guo *et al.* [24] show that incorporating the Phase spectrum of the Quaternion Fourier Transform (PQFT) instead of the amplitude transform leads to better saliency predictions in the spatio-temporal domain.

Some models learn saliency. Kienzle *et al.* [2] utilize support vector machines (SVM) to learn saliency of each image patch directly from human eye tracking data. Similarly, Judd *et al.* [3] train a linear SVM from human eye movement data, using a set of low, mid, and high-level image features to define salient locations. Feature vectors from highly fixated locations are assigned class label $+1$ while less fixated locations are assigned label -1. Zhao and Koch [22] used least-squares regression to learn the weights associated with a set of feature maps from subjects freely fixating natural scenes drawn from four different eye-tracking data sets. They find that the weights can be quite different for different data sets, but their face-detection and orientation channels are usually more important than color and intensity channels. Navalpakkam and Itti [29] define visual saliency in terms of signal to noise ratio (SNR) of a target object versus background and learn parameters of a linear combination of low-level features that cause the highest expected SNR for detecting a target from distractors.

Contributions. The models reviewed above fall into two general categories: 1) models that calculate saliency by implementing local center-surround operations (e.g., Itti *et al.* [4], Surprise [10], Judd *et al.* [3], GBVS [7], and Rahtu *et al.* [39]), 2) models that find salient regions globally by calculating rarity of features over the entire scene (e.g., AIM [14], SUN [12], Torralba [26], SRM [8], ICL [9], and Rarity model [25]). Our first contribution is to propose a unified model that benefits from the advantages of both approaches, which thus far have been treated independently. Note that the ideas of local and global context have been (separately) considered in the past [44][17] by salient object detection/segmentation approaches, but those have not yet been tested with human fixation prediction, which is the goal of most models (including ours).

Almost all saliency approaches utilize a color channel. Some have used RGB (e.g., [4][3][14][12]) while others have employed Lab (e.g., [42][18][39]), inspired by the finding that it better approximates human color perception. In particular, Lab aspires to perceptual uniformity, and its L component closely matches human perception of lightness, while the a and b channels approximate the human chromatic opponent system. RGB, on the other hand, is often the default choice for scene representation and storage. We argue that employing just one color system does not always lead to successful outlier detection. In Fig.1, we show that interesting objects in some images are more salient in Lab color space, while, for some others, saliency detection works better in RGB. Hence, a yet unexplored strat-

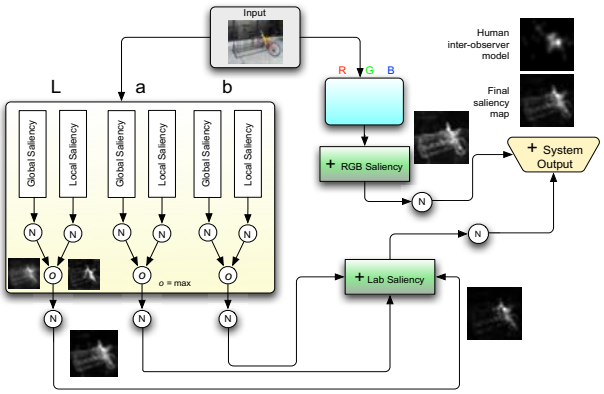

Figure 2. **Diagram of our proposed model.** First, the input image is transformed into Lab and RGB formats. Then, in each channel of a color space, a global saliency map based on rarity of an image patch in the entire scene, and a local saliency map, the dissimilarity between a patch and its surrounding window, are computed, normalized, and combined. Outputs of color channels (i.e., L, a, or b, similarly for RGB) are normalized and combined once more to form the output of a color system. The final map is the summation of the normalized maps in two color spaces.

egy, which is our second contribution, is combining saliency maps from both color spaces.

We compare accuracy of our model and its subcomponents with the mainstream models over four benchmark eye tracking datasets. These are top-ranked models that previous studies have shown to be significantly predictive of eye fixations in free viewing of natural scenes.

2. Proposed Saliency Model

Our proposed framework is presented in Fig. 2. An input image in two formats (Lab and RGB) undergoes the same saliency detection and the resultant maps in each color system are normalized and summed. In each color format, two local and global saliency operations are applied to each color sub-channel separately. While the first operation detects outliers in a local surrounding, the latter calculates the rarity of a feature or a region over the entire scene. Then, local and global rarities are combined to generate the output of each channel. Channel output maps are then normalized and summed once more to generate the saliency map. The whole process can be performed over several scales. There is no need to directly calculate the orientation channel in our model (since some patches from the chosen ensemble will emulate it; see below).

There is a large body of behavioral support for both local and global operations from the cognitive science literature. While early studies favored the thesis that local contrast attracts attention [15][19][4], recent work has shifted toward understanding top-down conceptual factors which seem to operate at the object level (see Figs. 1 and 7 for some examples). Such factors in free-viewing include human body, signs, cars, faces and text [3][45]. Particularly, Einhäuser *et al.* [30], showed that objects predict human fixations better than low-level saliency. Also it has been shown that interesting objects are more salient within a scene, providing support in favor of object-based attention [43]. As rare features are more likely to belong to a single object and since objects are rare compared to background in natural scenes, we believe that global saliency can help detecting top-down object-level concepts. Thus, instead of leaning on only one component (local or global), an effective strategy is integrating both of these complementary processes.

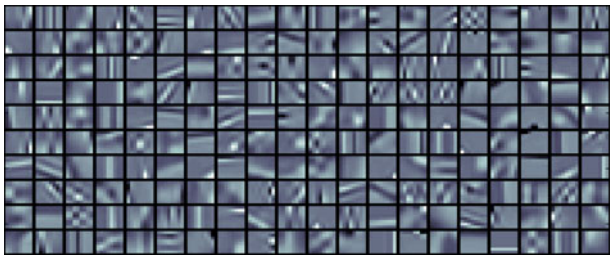

Figure 3. **A dictionary of 200 basis functions** learned from a large repository of natural images for the L channel of the Lab color space. Image size and patch size (w) were 512×512 and 8×8, respectively.

We estimate saliency on a patch-by-patch basis: each image patch is projected into the space of a dictionary of image patches (basis functions) learned from a repository of natural scenes. Each patch of an image is then represented by a vector of basis coefficients that can linearly reconstruct it.

2.1. Image representation

It is well known that natural images can be sparsely represented by a set of localized and oriented filters [27][28]. Also, recent progress in computer vision has demonstrated that sparse coding is an effective tool for image representation for several applications such as image classification [31][32], face recognition [33], image denoising [34], as well as saliency detection [14][12][9]. The underlying idea behind sparse coding is that a vision system should be adapted based on statistics of the visual environment where it is supposed to operate. As a supporting evidence for this theory, it has been shown that receptive fields (RF) of some neurons in V1 cortex resemble those RFs that are learned by sparse coding algorithms [27].

Mathematically, given a set of n $m-$dimensional basis signals (dictionary) $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \dots \mathbf{d}_n] \in \mathbb{R}^{m \times n}$, the sparse coding of an input signal $\mathbf{x} \in \mathbb{R}^m$ can be found by solving an "l_1-norm minimization problem":

$$\boldsymbol{\alpha}^*(\mathbf{x}, \mathbf{D}) = \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \frac{1}{2} ||\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}||_2^2 + \lambda_1 ||\boldsymbol{\alpha}||_1 \qquad (1)$$

where $||.||_1$ denotes the l_1-norm and λ_1 is a regularization parameter. Thus, $\mathbf{x} \sim \mathbf{x}^{'} = \mathbf{D}\alpha^*$ where $\mathbf{x}^{'}$ is the estimation of $\mathbf{x}$. To learn the dictionary $\mathbf{D}$, considering a training set of q data samples $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots \mathbf{y}_q]$ in $\mathbb{R}^{m \times q}$ an empirical cost function $g_q(\mathbf{D}) = \frac{1}{q} \sum_{i=1}^{q} l_u(\mathbf{y}_i, \mathbf{D})$ is minimized,

where $l_u(\mathbf{y}, \mathbf{D})$ is:

$$l_u(\mathbf{y}_i, \mathbf{D}) = min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \frac{1}{2}||\mathbf{y}_i - \mathbf{D}\boldsymbol{\alpha}||_2^2 + \lambda_1||\boldsymbol{\alpha}||_1 \quad (2)$$

We represent an image patch by a linear combination of some basis functions which correspond or act as feature detectors in early visual areas of the brain (neuron receptive fields or transfer functions). Given an input image, it is first resized to $2^w \times 2^w$ pixels where patch size w is selected in a way that 2^w is divisible to w. Let $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \dots \mathbf{p}_n\}$ represent the set of linearized image patches from top-left to bottom-right with no overlap. Then using Eq. 1, coefficients that reconstruct each patch are calculated and are used to represent that patch. By reshaping reconstructed patches and aligning them, the original image can be reproduced.

To learn a dictionary of patch bases (i.e., minimizing $g_q(\mathbf{D})$), we extracted $500,000$ 8×8 image patches (for each sub channel of RGB or Lab) from 1500 randomly selected color images from natural scenes. Each basis function in the dictionary is a $8 \times 8 = 64$D vector. A sample learned dictionary of size 200 is shown in Fig. 3. We experimented with different dictionary sizes (10, 50, 100, 200, 400, and 1000) and realized that fixation prediction results did not change much. The sparse codes $\boldsymbol{\alpha}_i$ are computed with the above basis using the LARS algorithm [5] implemented in the SPAMS toolbox[1].

2.2. Measuring visual saliency

Our model is based on two saliency operations. The first one, local contrast, considers the rarity of image regions with respect to (small) local neighborhoods (guided by the well-established computational architecture of Koch and Ullman [19] and Itti *et al.* [4]). The second operation, global contrast, evaluates saliency of an image patch using its contrast with respect to the patch statistics over the entire image. Finally, local and global contrast maps are consolidated. We repeat the process for each channel of both RGB and Lab color systems and fuse saliency maps of each sub channel of a color space to generate a saliency map for each color system. At each stage, maps are normalized before integration (See Fig. 2).

Local saliency. Local saliency (S_l) in our model is the average weighted dissimilarity between a center patch i (blue rectangle in Fig. 4) and its L patches in a rectangular neighborhood (red rectangle in Fig. 4):

$$S_l^c(\mathbf{p}_i) = \frac{1}{L} \sum_{j=1}^{L} W_{ij}^{-1} D_{ij}^c \quad (3)$$

where W_{ij} is the Euclidean distance between the center patch i and the surround patch j. Thus, those patches further away from the center patch will have less influence on

[1] http://www.di.ens.fr/willow/SPAMS/index.html

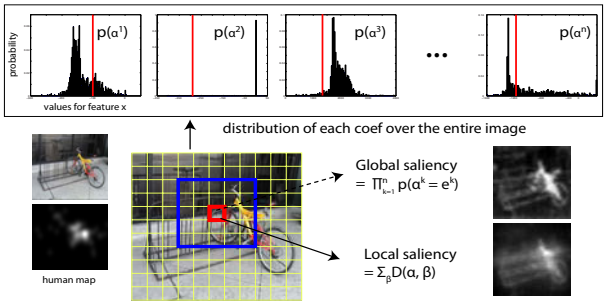

Figure 4. **Illustration of global and local saliency for an image patch.** Global saliency measures the rarity of patch in the entire scene while local rarity measures the difference between a patch and its surrounding context.

the saliency of the center patch. D_{ij} denotes the Euclidean distance between patch i and patch j in the feature space between $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_j$, vectors of coefficients for patches i and j, respectively derived from sparse coding (Sec. 2.1). While here we use the Euclidean distance (l_2 distance), the KL distance [23][38], l_1 distance [17], or correlation coefficient have also been used in the past to calculate patch similarity. Superscript c denotes color sub channels ($L, a, or\ b$ in Lab or $R, G, or\ B$ in RGB).

Global saliency. It often happens that a local patch is similar to its neighbors but the whole region (i.e., local + surrounding) is still in global rarity in the entire scene. Using only the local saliency may suppress areas within a homogeneous region resulting in blank holes, which sometimes impedes object-based attention (e.g., a uniformly textured object would only be salient at its borders). To remedy such shortcoming, we build our global saliency operator guided by the information-theoretic saliency measure of Bruce and Tsotsos [14]. Instead of each pixel, here we calculate the probability of each patch $P(\mathbf{p}_i)$ over the entire scene and use its inverse as the global saliency:

$$\begin{aligned} S_g^c(\mathbf{p}_i) &= P(\mathbf{p}_i)^{-1} = \left(\Pi_{j=1}^{n} P(\alpha_{ij})\right)^{-1} \\ log\left(S_g^c(\mathbf{p}_i)\right) &= -log(P(\mathbf{p}_i)) = -\sum_{j=1}^{n} log(P(\alpha_{ij})) \\ S_g^c(\mathbf{p}_i) &\propto -\sum_{j=1}^{n} log(P(\alpha_{ij})) \end{aligned} \quad (4)$$

To calculate $P(\mathbf{p}_i)$, we assume that coefficients $\boldsymbol{\alpha}$ are conditionally independent from each other. This is to some extent guaranteed by the sparse coding algorithm [5]. For each coefficient of the patch representation vector (i.e., α_{ij}), first a binned histogram (100 bins here) is calculated from all of the patches in the scene and is then converted to a pdf ($P(\alpha_{ij})$) by dividing to its sum. If a patch is rare in one of the features, the above product will get a small value leading to high global saliency for that patch overall. Fig. 4 illustrates the process of calculating global saliency.

Combined saliency. Local and global saliency maps are

then normalized and combined:

$$S^c_{lg}(\mathbf{p}_i) = \mathcal{N}(S^c_l(\mathbf{p}_i)) \circ \mathcal{N}(S^c_g(\mathbf{p}_i)) \quad (5)$$

where $\circ$ is an integration scheme (i.e., $\{+, *,$ max, *or* min$\}$). Through the experiments, we found that max in this stage leads to slightly higher accuracy than others. Then, saliency values of a patch in all channels are normalized and summed again to generate the saliency of a patch in each color system. For Lab color system, we have:

$$S^{Lab}_{lg}(\mathbf{p}_i) = \sum_{c \in L,a,b} \mathcal{N}(S^c_{lg}(\mathbf{p}_i)) \quad (6)$$

The same operation applies to the RGB color space. Final saliency for a patch is then summation of normalized saliency maps in both color systems:

$$S_{lg}(\mathbf{p}_i) = \mathcal{N}(S^{Lab}_{lg}(\mathbf{p}_i)) + \mathcal{N}(S^{RGB}_{lg}(\mathbf{p}_i)) \quad (7)$$

Normalization ($\mathcal{N}$). Similar to [4] first, the average of all local maxima (defined as greater than 4 neighboring points) with intensity above a threshold is calculated (M_l). Then a map is multiplied by $p = (M_g - M_l)^2$ where M_g is the global maximum in the map (known as maxnorm).

Extension to the scale space. Since objects appear at different sizes and depths, it is necessary to perform saliency detection at several spatial scales. To make our approach multi-scale, we calculate the saliency of downsampled images (divisions by 2) from the original image and then take the average after normalization:

$$S(\mathbf{x}) = \frac{1}{M} \sum_{i=1}^{M} \mathcal{N}(S^i_{lg}(\mathbf{x})) \quad (8)$$

where M is the number of scales and $S^i_{lg}(\mathbf{x})$ is the saliency of pixel $\mathbf{x}$ derived from the saliency map created by Eq. 7. Finally, we smooth the resultant map by convolving it with a small Gaussian kernel for better visualization.

Handling center-bias. A tedious and challenging factor in saliency modeling is handling center-bias in eye tracking data, which is the tendency of human subjects to preferentially look near the image center [13]. This generates a high central peak in the overall 2D histogram of fixations, resulting in high scores for a trivial saliency model whose map is just a Gaussian blob at the image center. To account for center-bias, some models intrinsically (e.g., GBVS [7], E-Saliency [37]) or extrinsically (e.g., Judd *et al.* [3] and Yang *et al.* [16]) add a center prior to their algorithms[2]. Here, instead of adding center bias to our model, we use a scoring metric that discounts center-bias in a non-parametric manner (See next section) when evaluating our and other saliency models against eye-tracking data.

[2]Some models add center-bias by either fitting a 2D Gaussian to fixation data or simply by just using the average fixation map (i.e., 2D histogram).

3. Experimental Setup

To validate our proposed method, we carried out several experiments on four benchmark datasets using the "shuffled AUC" score described below[3]. The main reason behind employing several datasets is that current datasets have different image and feature statistics, stimulus variety, biases (e.g., center-bias), and eye tracking parameters. Hence, it is necessary to employ several datasets as models leverage different features that their distribution varies across datasets.

Evaluation metric. The most widely used score for saliency model evaluation is the **AUC** [14]. In AUC, human fixations for an image are considered as the positive set and some points from the image are randomly chosen (uniformly) as the negative set. The saliency map is then treated as a binary classifier to separate the positive samples from the negatives. By thresholding over the saliency map and plotting true positive rate vs. false positive rate, an ROC curve is achieved and its underneath area is calculated. A problem with AUC is that it generates a large value for a central Gaussian model and is thus affected by center-bias [13]. To tackle center bias, Zhang *et al.* [12] introduced **shuffled AUC** score, with the only difference that instead of selecting negative points randomly from a uniform distribution, all human fixations (except the positive set) are used as the negative set. Shuffled AUC score generates a value of 0.5 for both a central Gaussian and a completely uniform map. Please note that in addition to shuffled AUC, there are also some other scores that have been often used in the past, for example Normalized Scanpath Saliency (NSS) [35], KL distance [10], and Correlation Coefficient [46]. But here we avoid using them as they are all affected by center-bias. Instead, we adopt the shuffled AUC score which is becoming a standard for saliency model evaluation [40][12].

Utilized fixation datasets are briefly described below.

TORONTO[4] [14]. This is the most widely used dataset for model comparison. It contains 120 color images with resolution of 511×681 pixels from indoor and outdoor environments. Images are presented at random to 20 subjects for 3 seconds with 2 seconds of gray mask in between.

MIT[5] [3]. This is the largest dataset containing 1003 images (resolution from 405×1024 to 1024×1024 pixels) collected from Flicker and LabelMe datasets. There are 779 landscape and 228 portrait images. Fifteen subjects freely viewed images for 3 sec. with 1 sec. delay in between.

KOOTSTRA[6] [20]. This dataset contains 101 images from 5 different categories: 12 animals, 12 automan, 16 buildings, 20 flowers, and 41 natural scenes. Images are ob-

[3]Our software for score calculation and saliency maps over 4 datasets are available at: https://sites.google.com/site/saliencyevaluation/.

[4]Available at: http://www-sop.inria.fr/members/Neil.Bruce

[5]This dataset is available at: http://people.csail.mit.edu/tjudd/

[6]This dataset is available at: http://www.csc.kth.se/~kootstra/

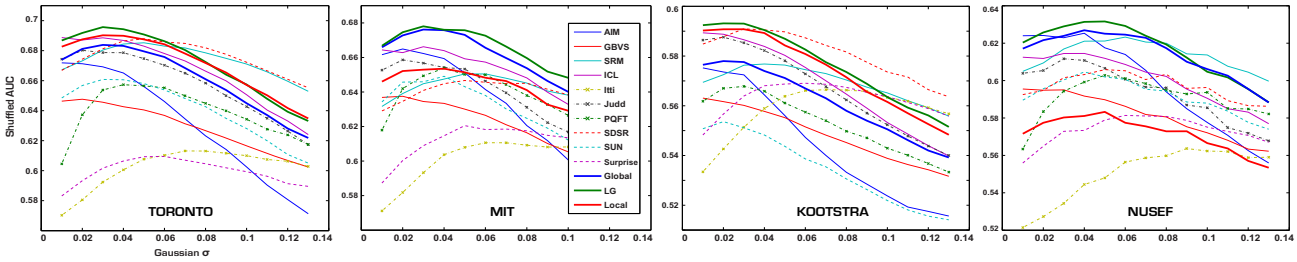

Figure 5. **Model comparison.** Fixation prediction accuracy of our saliency operations (Local, Global, LG (Local + Global)) along with 10 state-of-the-art models over 4 benchmark datasets. X-axis indicates the σ of the Gaussian kernel (in image width) by which maps are smoothed. NUSEF dataset contains some images with copyright which are not easily accessible and we don't use. Only 412 images are used here.

Dataset	AIM [14]	GBVS [7]	SRM [8]	ICL [9]	Itti [4]	Judd [3]	PQFT [24]	SDSR [11]	SUN [12]	Surprise [10]	Local S_l	Global S_g	LG S_{lg}	Gauss	IO
TORONTO [14]	0.67	0.647	0.685	**0.691**	0.61	0.68	0.657	0.687	0.66	0.605	**0.691**	0.69	**0.696**	0.50	**0.73**
Optimal σ	0.01	0.02	0.05	0.01	0.07	0.03	0.04	0.05	0.03	0.06	0.04	0.03	0.03	-	-
MIT [3]	0.664	0.637	0.65	**0.666**	0.61	0.658	0.65	0.646	0.649	0.62	0.653	**0.676**	**0.678**	0.50	**0.75**
Optimal σ	0.02	0.02	0.05	0.03	0.06	0.02	0.04	0.05	0.04	0.05	0.04	0.04	0.03	-	-
KOOTSTRA [20]	0.575	0.563	0.576	0.589	0.57	0.587	0.57	**0.59**	0.55	0.566	**0.591**	0.578	**0.593**	0.50	**0.62**
Optimal σ	0.01	0.01	0.04	0.01	0.07	0.02	0.03	0.03	0.02	0.07	0.03	0.02	0.03	-	-
NUSEF [6]	**0.623**	0.595	0.62	0.614	0.56	0.61	0.60	0.60	0.60	0.58	0.583	**0.627**	**0.632**	0.49	**0.66**
Optimal σ	0.04	0.01	0.06	0.03	0.09	0.03	0.05	0.04	0.04	0.06	0.05	0.04	0.05	-	-

Table 1. **Maximum performance of models shown in Fig. 5**. Numbers in second rows are the sigma values where models take their maximum performance. Parameter settings: Surround window size = 1; number of scales = 1 (256×256). Accuracies of three best models over each dataset are shown in bold face font. LG is the Local+Global model and IO stands for the human inter-observer model.

served by 31 subjects in the age range of 17 to 32 for 5 seconds. Image resolution is 768×1024 pixels. This dataset is specially challenging because there are not explicit objects or salient regions within many of the images.

NUSEF[7] **[6].** This dataset includes 758 images containing emotionally affective scenes/objects such as expressive faces, nudes, unpleasant concepts, and interactive actions. In total, 75 subjects free-viewed part of the image set for 5 seconds each (on average 25 subjects per image).

4. Performance Evaluation

Here, along with the evaluation of our model, we also compare 10 state-of-the-art bottom-up saliency models. Softwares for these models are publicly available[8]. Additionally, we implemented two simple models, to serve as baseline: Gaussian Blob (Gauss) and Human inter-observer (IO). Gaussian blob is simply a 2D Gaussian shape drawn at the center of the image; it is expected to predict human gaze well if such gaze is strongly clustered around the center [13]. The human model outputs, for a given stimulus, a map built by integrating fixations from other subjects than the one under test while they watched that stimulus. The human map is usually smoothed by convolving with a small Gaussian kernel. This model provides an upper-bound on prediction accuracy of saliency models to the degree that, different humans may be the best predictors of each other. Model maps were resized to the size of the original image, onto which eye data have been recorded.

An important parameter in model comparison is smoothness (blurring) of saliency maps [40]. Here, we smoothed the saliency map of each model by convolving it with a variable size Gaussian kernel. Fig. 5 presents the shuffled AUC score of models over the range of standard deviations σ of the Gaussian kernel in image width (from 0.01 to 0.13 in steps of 0.01). The maximum score value over this range for each model is shown in Table 1. Smoothing the saliency map dramatically affects the accuracy of some models (e.g., Itti, Surprise, PQFT, and SUN). Our combined saliency model (local + global and RGB + Lab) outperforms other models over 4 datasets with a larger margin over the MIT and NUSEF datasets. Our local and global saliency operators have less accuracy than the combined model but are still above several models. Results show that global saliency works better than local saliency operator over large datasets (MIT and NUSEF) while they are close to each other over TORONTO dataset. Models were more successful over the TORONTO and MIT datasets and less over KOOTSTRA and NUSEF, possibly because of the higher complexity of stimuli in these datasets. The NUSEF dataset contains many affective and emotional stimuli while KOOTSTRA dataset contains images without well-defined interesting and salient objects (e.g., nature scenes, trees, and flowers).

[7] Available at: http://mmas.comp.nus.edu.sg/NUSEF.html

[8] **AIM**: http://www-sop.inria.fr/members/Neil.Bruce/
GBVS: http://www.klab.caltech.edu/~harel/
SRM & ICL: http://www.klab.caltech.edu/~xhou/
Itti & Surprise: http://ilab.usc.edu/toolkit/
Judd: http://people.csail.mit.edu/tjudd/
PQFT: http://visual-attention-processing.googlecode.com/
SDSR: http://alumni.soe.ucsc.edu/~rokaf/
SUN: http://cseweb.ucsd.edu/~l6zhang/

Dataset	RGB			Lab			RGB + Lab		
	S_l	S_g	S_{lg}	S_l	S_g	S_{lg}	S_l	S_g	S_{lg}
TORONTO	0.646	0.647	0.653	0.670	0.660	0.660	0.678	0.668	**0.683**
MIT	0.627	0.639	0.640	0.646	0.644	0.651	0.658	0.663	**0.667**
KOOTSTRA	0.574	0.572	0.578	0.572	0.555	0.570	0.589	0.573	**0.591**
NUSEF	0.599	0.610	0.610	0.556	0.596	0.592	0.569	0.614	**0.616**

Table 2. **RGB vs. Lab for saliency detection.** S_l: Local; S_g: Global; S_{lg}: Local + Global. Parameter settings: scales (M) =1 (256×256); Window size = 1. Results are over original saliency maps without smoothing.

Among compared models, ICL [12], AIM [14], SDSR [11], and Judd *et al.* [3] performed higher than the rest. Itti *et al.* [4] and Surprise [10] models are ranked at the bottom. As we expected, a trivial Gaussian blob located at the image center scores around 0.5 over all datasets and human model scores the best providing a gold standard for visual saliency models. Humans are less correlated over KOOTSTRA and NUSEF datasets.

Lab vs. RGB for saliency detection. To assess the power of Lab and RGB color spaces for saliency detection, we performed an experiment using each of the two color systems. Results over all four datasets are shown in Table 2. According to our results, it is not possible to tell which color system is the best. The Lab color space leads to higher accuracies over TORONTO and MIT datasets while RGB works better over KOOTSTRA and NUSEF datasets. Integrating both color systems leads to higher accuracy than each one taken separately, consistently over all four datasets. This indicated the importance of saliency integration over both color systems. Also note that over each component (local or global), combination of RGB and Lab leads to higher performance than each of the color systems.

Influence of surrounding window size and number of scales. Here we analyze how the size of the surrounding window (and hence number of neighbors) and number of spatial scales affect performance of our models. As the left diagram of Fig. 6 shows, increasing the number of neighbors reduces the accuracy of the local saliency operator. Correspondingly, this reduces the accuracy of the combined model. Note that the global operator is not affected by this parameter. Shown in the right panel of Fig. 6, increasing the number of scales enhances the results to a certain point (here using 3 scales [256, 128, and 64]) and then drops (when using 4 scales [512 256 128 64]).

Runtime aspects. It takes approximately 5 seconds for our model to process a 256×256 image in both RGB and Lab color spaces using a personal computer running Linux Ubuntu with 5.8 GB RAM and 12 Core Intel i7 3.2 GHz CPU. Our model is faster than AIM (16 sec), Judd (without object detectors)(4.7 sec), close to SDSR (2.4) and GBVS (2), and slower than PQFT (1 sec), SUN (1), Itti (0.28) (using [52]), and ICL (0.1) models. Our global saliency operator is appx. 3 times faster than our local saliency operator.

For qualitative assessment, we show in Fig. 7 saliency maps of our combined saliency model and compared models for sample images from TORONTO and MIT datasets.

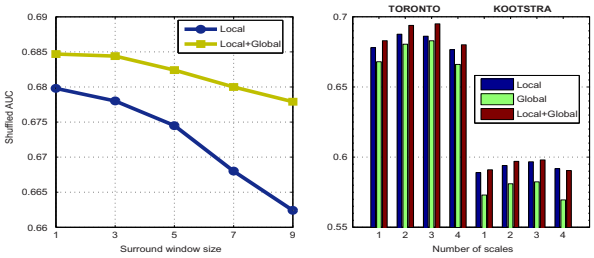

Figure 6. **Parameter analysis**. Left: Effect of the surround window size on accuracy over TORONTO dataset using 256×256 images ($M = 1$). Right: Influence of scale on results over TORONTO and KOOTSTRA datasets (window size =1). First three bars are 256,128,64 and fourth one represents four scales 512,256,128,64.

5. Conclusions and Future Works

We enhance the state-of-the-art in saliency modeling by proposing an accurate and easy-to-implement model that utilizes image representations in both RGB and Lab color spaces. Furthermore, we introduce one local and one global saliency operator each representing a class of previous models to some extent. We conclude that integration of local and global saliency operators works better than just using either one, which encourages more research in this direction. Similarly, combining both color systems strongly benefits saliency detection and eye fixation prediction.

There are two areas that we would like to improve upon. The first one is incorporating top-down factors for fixation prediction. The large gap between models and the human inter-observer model (see Table 1) is mainly due to role of top-down concepts (e.g., faces, text [45], people, and cars [3], affective and emotional stimuli or actions within scenes [6]) when freely viewing scenes. While some of these factors have been utilized for saliency detection in the past [3], adding more top-down features (e.g., by reliable detection of text on natural scenes) can scale up accuracy of current models. The second area is extending our model for saliency detection in spatio-temporal domain (videos).

Supported by the National Science Foundation (grant number BCS-0827764), and the Army Research Office (W911NF-08-1-0360 and W911NF-11-1-0046), and U.S. Army (W81XWH-10-2-0076). The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.

References

[1] J. M. Wolfe and T. S. Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nat. Rev. Neurosci.*, 5:1-7, 2004. 1

[2] W., Kienzle, A. F., Wichmann, B., Scholkopf, and M. O. Franz. A nonparametric approach to bottom-up visual saliency. *NIPS*, 2007. 2

[3] T. Judd, K. Ehinger, F. Durand and, A. Torralba. Learning to predict where humans look, *ICCV*, 2009. 2, 3, 5, 6, 7

[4] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE PAMI*, 1998. 1, 2, 3, 4, 5, 6, 7

[5] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *J. of Machine Learning*, 2010. 4

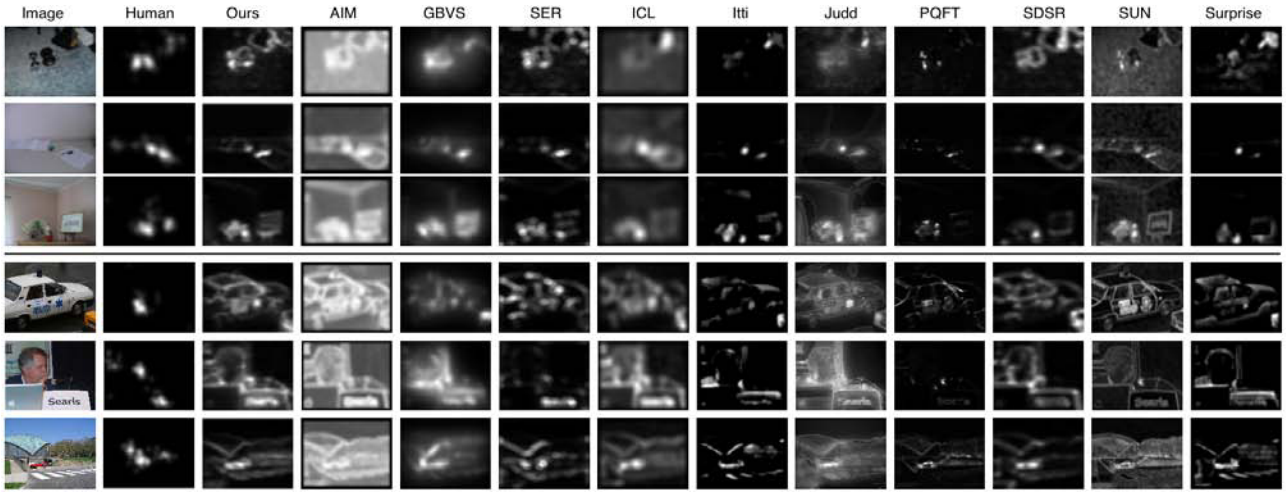

Figure 7. **Visual comparison of our combined saliency model and 10 state-of-the-art models over samples from TORONTO (top) and MIT datasets.**

[6] R. Subramanian, H. Katti, N. Sebe, M. Kankanhalli, and T.S. Chua. An eye fixation database for saliency detection in images. *ECCV*, 2010. 6, 7

[7] J. Harel, C. Koch, P. Perona. Graph-based visual saliency. *NIPS*, 2006. 2, 5, 6

[8] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. *CVPR*, 2007. 2, 6

[9] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. *NIPS*, 2008. 2, 3, 6

[10] L. Itti and P. Baldi. Bayesian surprise attracts human visual attention. *NIPS*, 2005. 2, 5, 6, 7

[11] H.J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9, 2009. 2, 6, 7

[12] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, SUN: A Bayesian framework for saliency using natural statistics. *J. of Vision*, 8(32):1-20, 2008. 2, 3, 5, 6, 7

[13] B.W. Tatler. *J. Vision*, 14(7):1-17, 2007. 5, 6

[14] N.D.B. Bruce and J.K. Tsotsos. Saliency based on information maximization. *NIPS*, 2005. 1, 2, 3, 4, 5, 6, 7

[15] A.M. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psych.*, 12:97-136, 1980. 1, 3

[16] Y. Yang, M. Song, N. Li, J. Bu, and C. Chen. What is the chance of happening:A new way to predict where people look. ECCV,2010. 5

[17] L. Duan, C. Wu, J. Miao, L. Qing, and Y. Fu. Visual saliency detection by spatially weighted dissimilarity. CVPR 2011. 2, 4

[18] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model bottom-up visual attention. *PAMI*, 2006. 2

[19] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, 1985. 1, 3, 4

[20] G. Kootstra, A. Nederveen, and B. de Boer. Paying attention to symmetry. *BMVC*, 2008. 2, 5, 6

[21] R. Valenti, N. Sebe, and T. Gevers. Image saliency by isocentric curvedness and color. . *ICCV*, 2009. 2

[22] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, 11(3), 2011. 2

[23] D. Gao, V. Mahadevan, and N. Vasconcelos. The discriminant center-surround hypothesis for bottom-up saliency. *NIPS*, 2007. 2, 4

[24] C. Guo and L. Zhang. A novel multiresolution spatiotemporal saliency detection model and Its applications in image and video compression. *IEEE Trans. on Image Processing*, 2010. 2, 6

[25] M. Mancas. Computational attention: Modelisation and application to audio and image processing. PhD. thesis, 2007. 2

[26] A. Torralba, A. Oliva, M. Castelhano and J.M. Henderson. Contextual guidance of attention in natural scenes: The role of Global features on object search. *Psychological Review*, 2006. 2

[27] B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 1996. 3

[28] E. Simoncelli and B. Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24, 2001. 3

[29] V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. *CVPR*, 2006. 2

[30] W. Einhäuser, M. Spain, and P. Perona. Objects predict xations better than early saliency. *Journal of Vision*, 2008. 3

[31] C. Kanan and G. Cottrell. Robust classification of objects, faces, and flowers using national image. *CVPR*, 2010. 3

[32] F. Bach, J. Mairal, J. Ponce, and G. Sparo. Sparse coding and dictionary learning for image analysis. *CVPR*, 2010. 3

[33] A. Yang, A. Ganesh, Z. Zhou, S. Sastry, and Y. Ma. A review of fast l1-minimization algorithms for robust face recognition. *http://arxiv.org*, 2010. 3

[34] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3336-3745, 2006. 3

[35] R. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Res.*, 45, 2005. 5

[36] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Res.*, 2002. 2

[37] T. Avraham, M. Lindenbaum. Esaliency (Extended Saliency): Meaningful attention using stochastic image modeling. *PAMI*, 2010. 5

[38] D.A. Klein and S. Frintrop. Center-surround divergence of feature statistics for salient object detection. *ICCV*, 2011. 4

[39] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilá. Segmenting salient object from images and videos. *ECCV*, 2010. 2

[40] X. Hou, J. Harel, and Christof Koch. Image Signature: Highlighting sparse salient regions. *IEEE PAMI*, In press. 5, 6

[41] L. Itti, N. Dhavale, and F. Pighin. Realistic avatar eye and head animation using a neurobiological model of visual attention. *SPIE*, 2003. 2

[42] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil. Decorrelation and distinctiveness provide with human-like saliency. *ACIVS*, 5807, 2009. 2

[43] L. Elazary and L. Itti. Interesting objects are visually salient. *J. Vision*, 2008. 3

[44] M.M Cheng, G.X Zhang, N.J. Mitra, and X. Huang, and S.M. Hu. Global Contrast based Salient Region Detection. *CVPR*, 2011. 2

[45] M. Cerf, J. Harel, W. Einhäuser, and C. Koch. Predicting gaze using low-level saliency combined with face detection. *NIPS*, 2007. 3, 7

[46] N. Ouerhani, R. von Wartburg, H. Hugli, and R.M. Muri. Empirical validation of saliency-based model of visual attention. *Electronic Letters on Computer Vision and Image Analysis*, 2003. 5

[47] A. Toet. Computational versus psychophysical image saliency: A comparative evaluation study. *IEEE trans. PAMI*, 2011. 1

[48] N. Parikh, L. Itti, and J. Weiland. Saliency-based image processing for retinal prostheses. *J. Neural Eng.* 7, 2010. 1

[49] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Trans. Image Process.*, 2004. 1

[50] J. Wang, J. Sun, L. Quan, X. Tang, and H.Y Shum. Picture collage. *CVPR*, 1:347-354, 2006. 1

[51] R. Rosenholtz, A. Dorai, and R. Freeman. Do predictions of visual perception aid design? *ACM Transactions on Applied Perception (TAP)*, 2011. 1

[52] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 2006. 1, 7

[53] D. DeCarlo and A. Santella. Stylization and abstraction of photographs. *ACM Trans. on Graphics*, 2002. 1