Learning a saliency map using fixated locations in natural scenes

Qi Zhao

Computation and Neural Systems, California Institute of Technology, Pasadena, CA, USA

Christof Koch

Computation and Neural Systems, California Institute of Technology, Pasadena, CA, USA, & Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea

Inspired by the primate visual system, computational saliency models decompose visual input into a set of feature maps across spatial scales in a number of pre-specified channels. The outputs of these feature maps are summed to yield the final saliency map. Here we use a least square technique to learn the weights associated with these maps from subjects freely fixating natural scenes drawn from four recent eye-tracking data sets. Depending on the data set, the weights can be quite different, with the face and orientation channels usually more important than color and intensity channels. Inter-subject differences are negligible. We also model a bias toward fixating at the center of images and consider both time-varying and constant factors that contribute to this bias. To compensate for the inadequacy of the standard method to judge performance (area under the ROC curve), we use two other metrics to comprehensively assess performance. Although our model retains the basic structure of the standard saliency model, it outperforms several state-of-the-art saliency algorithms. Furthermore, the simple structure makes the results applicable to numerous studies in psychophysics and physiology and leads to an extremely easy implementation for real-world applications.

Keywords: computational saliency model, feature combination, center bias, inter-subject variability, metric


Introduction

Humans and other primates move their eyes to select visual information from any one visual scene. This allows them to bring the high-resolution part of their retina, the fovea, onto relevant parts of the image, thereby deploying processing recourses to the most relevant visual information and interpret complex scenes in real time. Besides understanding the mechanism that drives this selection of interesting parts in the image, predicting interesting locations as well as locations where people are likely to look has many real-world applications. Computational models can be applied to various computer vision tasks such as navigational assistance, robot control, surveillance systems, object detection and recognition, and scene understanding. Such predictions also find applications in other areas including advertising design, image and video compression, pictorial database querying, and gaze animation.

Starting from the Feature Integration Theory of Treisman and Gelade (1980) and the proposal by Koch and Ullman (1985) for a map in the primate visual system that encodes the extent to which any location in the field of view is conspicuous or salient, based on bottom-up, task-independent factors, a series of ever refined algorithms has been designed to predict where subjects will fixate in synthetic or natural scenes (Einhauser, Spain, & Perona, 2008; Foulsham & Underwood, 2008; Itti, Koch, & Niebur, 1998; Oliva, Torralba, Castelhano, & Henderson, 2003; Parkhurst, Law, & Niebur, 2002; Walther, Serre, Poggio, & Koch, 2005). In these models (Itti & Koch, 2000; Itti et al., 1998; Parkhurst et al., 2002), low-level attributes such as color, intensity, and orientation combined to yield maps through center–surround filtering at numerous spatial scales. Subsequently, Einhäuser et al. (2006) and Krieger, Rentschler, Hauske, Schill, and Zetzsche (2000) suggested incorporating higher order statistics to fill some of the gaps between the predictive powers of current saliency map models. One way of doing this is by adding more semantic feature channels such as faces or text into the saliency map. This significantly improves the accuracy of prediction (Cerf, Frady, & Koch, 2009; Einhäuser et al., 2008). The extent to which such bottom-up, task-independent saliency maps predict human fixational eye movements under free-viewing conditions remains under active investigation (Donk & Zoest, 2008; Foulsham & Underwood, 2008; Masciocchi, Mihalas, Parkhurst, & Niebur, 2009). Bottom-up saliency has also been adopted (Chikkur, Serre, Tan, & Poggio, 2010; Navalpakkam & Itti, 2005; Rutishauser & Koch, 2007) to mimic top-down searches. However, we here only...
consider task-independent scrutiny of images as they might occur when people are gazing at a scene without looking for anything in particular.

Within each feature channel, various normalization methods have been proposed (Itti & Koch, 1999) to integrate the multi-scale feature maps into a final one. The focus of these methods are on unifying the maps across different dynamic ranges and extraction mechanisms so that salient objects appearing strongly in a few maps are less likely to be masked by others.

Despite advances in image features and normalization methods, *linear summation* of feature channels into the final saliency map remains the norm (Cerf et al., 2009; Harel, Koch, & Perona, 2007; Itti & Baldi, 2006; Itti et al., 1998). Linear summation has some psychophysical support (Nothdurft, 2000) and is simple to apply. However, (Koene & Zhaoping, 2007; Li, 2002) have raised psychophysical arguments against linear summation strategies. In addition, prior work (Itti, 2005; Peters, Iyer, Itti, & Koch, 2005) has been aware of the different strengths contributed by different features to perceptual salience. We here investigate the importance of different bottom-up features in driving gaze allocation, including inter-subject variability, by learning an optimal set of feature weights using the constraint linear least square algorithm and perform quantitative analysis on four recent eye movement data sets (Bruce & Tsotsos, 2009; Cerf et al., 2009; Judd, Ehinger, Durand, & Torralba, 2009; Subramanian, Katti, Sebe, Kankanhalli, & Chua, 2010).

Under standard testing conditions, a strong central bias is seen, that is, subjects tend to look at the center of the image. Several explanations for this phenomenon have been suggested. Some attributed the center bias to the drop in visual system sensitivity in the periphery (Parkhurst et al., 2002; Peters et al., 2005) and to a motor bias in the saccadic system that favors small amplitude saccades over large amplitude ones (Bahill, Adler, & Stark, 1975; Gajewski, Pearson, Mack, Bartlett, & Henderson, 2005; Pelz & Canosa, 2001). These two factors combined with the fact that scene viewing experiments typically start in the center result in a central fixation bias. The experimental setup (users are placed centrally in front of the screen; Judd et al., 2009; Tatler, 2007; Zhang, Tong, & Cottrell, 2009; Zhang, Tong, Marks, Shan, & Cottrell, 2008) and the bias toward centering the eyeball within its orbit reinforce the tendency to look toward the center (Fuller, 1996; Pare & Munoz, 2001; Tatler, 2007; Zambarbarieri, Beltrami, & Versino, 1995). However, Vitu, Kapoula, Lancelin, and Lavigne (2004) demonstrated that it is the screen center rather than the straight-head position—the orbital center—that produces the central fixation tendency. Many (Einhäuser et al., 2008; Judd et al., 2009; Parkhurst et al., 2002; Reinagel & Zador, 1999; Tatler, Baddeley, & Gilchrist, 2005) assumed that the bias arises from image feature distributions. As human photographers place objects of interest in the center, it is not surprising that subjects will look at such centrally placed objects. Lastly, Le Meur, Le Callet, Barba, and Thoreau (2006) and Tatler (2007) pointed out that the center of the scene offers strategic advantages—it is an optimal location for extracting information from the scene and a convenient location for the efficient exploration of the scene.

Previous work (Judd et al., 2009; Parkhurst et al., 2002; Peters et al., 2005) described this bias via a single Gaussian or exponential spatial filter. The Gaussian/exponential type prior is effective but not readily justified. We here consider both time-varying and constant factors that give rise to this effect. That is, we consider the possibility that the center bias may be stronger early on and then diminish over time (or vice versa). We show that the saccade sequence follows a Gaussian process and that the distribution of fixations is a mixture of Gaussians. Furthermore, by proving the convergence of the Gaussian covariance matrix, we justify approximating this time-varying process via a single kernel.

**Methods**

**Data set**

This study analyzes eye movements from four recent data sets (Bruce & Tsotsos, 2009; Cerf et al., 2009; Judd et al., 2009; Subramanian et al., 2010; Figure 1).

In the FIFA data set (Cerf et al., 2009), fixation data were collected from 8 subjects performing a 2-s-long free-viewing task on 180 color natural images (28° × 21°). They were asked to rate, on a scale of 1 through 10, how interesting each image was. Scenes were indoor and outdoor still images in color. Images include faces in various skin colors, age groups, gender, positions, and sizes.

The second data set from Bruce and Tsotsos (2009; referred here as the Toronto database) contains data from 11 subjects viewing 120 color images of outdoor and indoor scenes. Participants were given no particular instructions except to observe the images (32° × 24°), 4 s each. One distinction between this data set and that of the FIFA (Cerf et al., 2009) is that a large portion of images here do not contain particular regions of interest, while in the FIFA data set most contain very salient regions (e.g., faces or salient nonface objects).

The eye-tracking data set from Judd et al. (2009; referred to as MIT database) is the largest one in the community. It includes 1003 images collected from Flickr and LabelMe. Eye movement data were recorded from 15 users who free-viewed these images (36° × 27°) for 3 s. A memory test motivated subjects to pay attention to the images: they looked at 100 images and needed to indicate which ones they had seen before.

The NUS database recently published by Subramanian et al. (2010) includes 758 images containing semantically...
affective objects/scenes such as expressive faces, nudes, unpleasant concepts, and interactive actions. Images are from Flickr, Photo.net, Google, and emotion-evoking IAPS (Lang, Bradley, & Cuthbert, 2008). In total, 75 subjects free-viewed (26° × 19°) part of the image set for 5 s each (each image was viewed by an average of 25 subjects).

Fixation maps

From the recorded eye movement data, psychophysical fixation maps are constructed to globally represent the successive fixations of subjects viewing the images. Formally, for each subject $i$ viewing image $j$, assuming that each fixation gives rise to a Gaussian-distributed activity, all gaze data are represented as the recorded fixations convolved with an isotropic Gaussian kernel $K_G$ as

$$H_i^j(x) = a \sum_{k=2}^{f} K_G \left( \frac{x - x_k}{h} \right),$$

where $x$ denotes the 2d image coordinates. $x_k$ represents the image coordinates of the $k$th fixation, and $f$ is the number of fixations. The bandwidth of the kernel, $h$, is set to approximate the size of fovea, and $a$ normalizes the map. An example of a fixation map is shown in Figure 2b. Note that the first fixation of each image is not used as it is always the center of the image.

Bottom-up saliency model

The eye movement data are used to analyze and validate the predictions of attentional allocation by computational models. We use the biologically inspired bottom-up driven saliency model recently developed by Cerf et al. (2009), which adds a face map to the standard Itti–Koch saliency model (Itti et al., 1998). Briefly, the Itti–Koch model includes two color channels (blue/yellow and red/green), one intensity channel, and four orientation channels ($0^\circ$, $45^\circ$, $90^\circ$, $135^\circ$). Raw maps of nine spatial scales ($0–8$) are created using dyadic Gaussian pyramids. Six center–surround difference maps are then constructed as point-wise differences across pyramid scales to capture local contrasts ($\text{center level } c = \{2, 3, 4\}$, $\text{surround level } s = c + \delta$, where $\delta = \{2, 3\}$). A single conspicuity map for each of the color, intensity, and orientation feature channels is built through across-scale addition of the center–surround difference maps and is represented at scale 4 (Figure 2c). For the face channel, the conspicuity map is generated by running the Viola and Jones (2001) face detector. Although different from early visual features such as color, intensity, and orientation, face attracts attention strongly and rapidly, independent of task; therefore, it is also considered part of the bottom-up saliency pathway (Cerf et al., 2009).

Conspicuity maps are used to construct the feature vectors for learning. As shown in Figure 2, for an image location $x$, the values of the color, intensity, orientation, and face conspicuity maps at this particular location are extracted and stacked to form the sample vector $v(x) = [C(x) I(x) O(x) F(x)]^T$.

The fixation maps (Equation 1, Figure 2b) are represented at the same scale as the conspicuity maps, and the real number from the fixation map is stored with the feature vector.

Learning optimal weights

To quantify the relevance of different features in deciding where to look, we use linear, least square regression with constraints to learn the weights from eye movement data.

Formally, let $C, I, O,$ and $F$ be the stacked vectors of the color, intensity, orientation, and face values at all image locations and let us denote $V = [C I O F]$, $M_{\text{fix}}$ as vectorized fixation map that is represented as the recorded fixations convolved with an isotropic Gaussian kernel, and $w = [w_C w_I w_O w_F]^T$ as the weights of the feature channels. The objective function is

$$\arg\min_w \|V \times w - M_{\text{fix}}\|^2,$$  \hspace{1cm} (2)

subject to

$$w \geq 0.$$  \hspace{1cm} (3)

The problem is solved using an active set method similar to that described in Gill, Murray, and Wright (1981).

Figure 3 provides an illustration of how feature weights affect the final saliency maps. The weight of a feature indicates the importance of that particular feature in deciding where to look at.

To investigate the level of inter-subject variability, we learn optimal weights for each individual as well as for the
entire population of subjects. The only difference is the fixation maps we feed the algorithm (Equation 2).

**Modeling the center bias**

We consider two types of center biases—time-varying and constant ones (Figure 4a, black circles)—and model the eye movement as a Gaussian process.

1. We model any time-dependent center bias using a 2D Gaussian filter centered at the current fixation as $\mathcal{N}(\mathbf{c}_t, \Sigma_f)$. Here $\mathbf{c}_t$ is the location of the current fixation that changes with time, and $\Sigma_f = \begin{pmatrix} \sigma_f^2 & 0 \\ 0 & \sigma_f^2 \end{pmatrix}$, where $\sigma_f$ denotes a space constant and is fixed a priori. Note that although the mean of the distributions changes with time, the standard deviation reflects inherent biological properties and we set it as a constant during the viewing process (see the two small black circles in Figure 4a).

2. We model any time-independent center bias (due, for instance, to the straight-ahead position, the tendency to center the eyeball within its orbit, and the tendency to look at the screen center due to strategic advantages) via a 2D Gaussian centered at the screen center as $\mathcal{N}(\mathbf{0}, \Sigma_h)$ (see the large black circle in Figure 4a). Since the multiplication of Gaussian functions is still Gaussian functions, a single Gaussian here is equivalent to modeling each factor using a Gaussian and then multiplying them for the compound effect. As before, $\Sigma_h = \begin{pmatrix} \sigma_h^2 & 0 \\ 0 & \sigma_h^2 \end{pmatrix}$, where $\sigma_h$ is set a priori.

In the following, we first discuss the saccade (a 2D vector from the current fixation to the next fixation) distribution, followed by the fixation distribution.
Saccade distribution

In this section, we use symbols with \( \hat{\cdot} \) to represent distributions and those with \( \tilde{\cdot} \) to denote instances of variables.

Given the current (time \( t \)) fixation position \( \tilde{c}_t \), the two Gaussian factors just described multiply to produce the center bias effect; therefore, the distribution of the \( (t + 1) \)th saccade (from the \( t \)th location to the \( (t + 1) \)th location) is

\[
\hat{x}_{t+1} \sim \mathcal{N}(0, \Sigma_h) \cdot \mathcal{N}(\tilde{c}_t, \Sigma_f),
\]

(4)

where \( \tilde{c}_0 = 0 \) since the eye movement starts at the center of the screen. In this and the next subsection, the subscript \( 0 \) denotes the initial fixation, which is generally not used for analysis as it is the center of the screen. The subscript \( t \) refers to the \( t \)th fixation starting from the fixation following the initial one.

Since the multiplication of two Gaussian functions is another Gaussian function, according to Equation 4, the saccade sequence, that is the 2D vector from the current to the next fixation, \( \{\tilde{x}_t\}_{t=1,2,...} \), follows a Gaussian process.

Fixation distribution

Denoting the fixation distribution at time \( t \) as \( \hat{X}_t \), the fixation distribution at time \( t + 1 \), \( \hat{X}_{t+1} \), can be written as the integral of saccade distributions over all possible locations \( \tilde{c}_t \), weighted by the probability of generating each location \( \tilde{c}_t \) from \( \hat{X}_t \).

We derive that \( \hat{X}_{t+1} \sim \mathcal{N}(0, \Sigma_h) \cdot \mathcal{N}(\hat{X}_t, \Sigma_f) \) (see Appendix A for the derivation) and \( \hat{X}_1 = \tilde{x}_1 \). Since the convolution of two Gaussian functions is another Gaussian function, as is the multiplication of two Gaussian functions, we have

\[
\hat{X}_{t+1} \sim \mathcal{N}(\hat{c}_{t+1}, \Sigma_{t+1}),
\]

(5a)

where

\[
\Sigma_{t+1} = (\Sigma_h^{-1} + (\Sigma_t + \Sigma_f)^{-1})^{-1}, \quad \Sigma_t = (\Sigma_h^{-1} + \Sigma_f^{-1})^{-1},
\]

(5b)

and

\[
c_{t+1} = \Sigma_{t+1} \Sigma_h^{-1} 0 + \Sigma_{t+1} (\Sigma_t + \Sigma_f)^{-1} (c_t + 0), \quad \hat{c}_1 = 0.
\]

(5c)

It is obvious that the mean of the fixation distribution (Equation 5c) is \( 0 \). Further, we prove that their covariance matrix (Equation 5b) converges. Formally, we denote \( \{t\}_{t=1,2,...} \) as the sequence of successive fixations. The covariance matrix of the distribution at these fixations are \( \{\Sigma_t\}_{t=1,2,...} \), where \( \Sigma_t \) is defined in Equation 5b. We prove (see Appendix A) the following.

**Theorem 2.1.** The sequence of \( \{\Sigma_t\}_{t=1,2,...} \) is convergent.

The same mean and the convergence of the covariance matrix suggest that after a certain amount of viewing time (empirically after making 3–5 fixations), the fixation distribution does not vary much, as illustrated in Figure 4. This convergence property empirically justifies the use of a single Gaussian filter instead of a mixture of Gaussians to model the central bias as a function of the fixation number.

Similarity measures

There are several similarity measures to quantitatively evaluate the performance of saliency models. These measures include the Receiver Operating Characteristics (ROC; Green & Swets, 1966), the Normalized Scanpath Saliency (NSS; Peters et al., 2005), correlation-based measures (Jost, Ouerhani, von Wartburg, Mäuri, & Häggl, 2005; Rajashekhar, van der Linde, Bovik, & Cormack, 2008), the least square index (Henderson, Brockmole, Castelhan, & Mack, 2007; Mannan, Ruddock, & Wooding, 1997), and the “string-edit” distance (Brandt & Stark, 1997; Choi, Mosley, & Stark, 1995; Hacisalihzade, Allen, & Stark, 1992). Among them, ROC is the most popular method and most widely used in the community. The inherent limitation of ROC, however, is that it only depends on the ordering of the fixations (ordinality) and does not capture the metric amplitude differences. In practice, as long as the hit rates are high, the area under the ROC curve (AUC) is always high regardless of the false alarm rate (Figure 5). Therefore, an ROC analysis, while very useful, is by itself insufficient to describe the deviation of predicted fixation patterns from the actual fixation map. To conduct a more comprehensive evaluation, we also employ the NSS (Peters et al., 2005) and the Earth Mover’s Distance (EMD; Rubner, Tomasi, & Guibas, 2000) that measure the real difference rather than only ordering of the values. By definition, NSS (Peters et al., 2005) evaluates saliency values at fixated locations. It works by first linearly normalizing the saliency map to have zero mean and unit standard deviation. Next, it extracts from each point corresponding to the fixation locations along a subject’s scanpath its computed saliency and averages these values to compute the NSS that is compared against the saliency distribution of the entire image (which is, by definition, zero mean). The NSS is the average distance between the fixation saliency and zero. A larger NSS implies a greater correspondence between fixation locations and the saliency predictions. A value of zero indicates no such
correspondence. Unlike the NSS that focuses on the saliency values of the scanpath, EMD (Rubner et al., 2000) captures the global discrepancy of two distributions. Intuitively, given two distributions, EMD measures the least amount of work needed to move one distribution to map onto the other one. It is computed through linear programming and accommodates distribution alignments well. A larger EMD indicates a larger overall discrepancy between the two distributions.

Given the extant variability among different subjects looking at the same image, no saliency algorithm can perform better (on average) than the area under the ROC curve dictated by inter-subject variability. We compute an ideal AUC by measuring how well the fixations of one subject can be predicted by those of the other \( n - 1 \) subjects, iterating over all \( n \) subjects and averaging the result. These AUC values are 78.6% for the FIFA data set, 87.8% for the Toronto data set, 90.8% for the MIT data set, and 85.7% for the NUS data set. In general, we express the performance of saliency algorithms in terms of normalized AUC (nAUC) values, which are the AUC values using the saliency algorithm normalized by the ideal AUC. A strong saliency model should have an nAUC value close to 1, a large NSS, and a small EMD value.

### Results and discussions

We test our algorithms on four data sets: the FIFA (Cerf et al., 2009), Toronto (Bruce & Tsotsos, 2009), MIT (Judd et al., 2009), and the NUS (Subramanian et al., 2010) data sets.

#### Experiment 1—using the FIFA data set

We first compare the models with equal weights and learned weights on the FIFA data set (Cerf et al., 2009). The data set of 180 images is divided into 130 training images and 50 testing ones. For models with a center bias, a center Gaussian function is learned from the training data. The optimal weights for the general model are

\[
\begin{bmatrix}
0.027 & 0.024 & 0.222 & 0.727
\end{bmatrix}^T.
\]

The optimal weights for subject-specific models vary slightly, while the ranking of the four features—from faces and orientation to color and intensity—remains constant. “CBM” stands for Center Bias Modeling. The NSSs of the linear models with optimal weights are noticeably larger than those with equal weights, and the NSSs of models with CBM are larger than those without CBM, suggesting a greater correspondence between fixation locations and the salient points predicted by the models. The EMD is considerably smaller using optimal weights or/and with CBM, reflecting superior global consistency between saliency and fixation maps.

<table>
<thead>
<tr>
<th>Equal weights (Cerf et al., 2009; Itti et al., 1998)</th>
<th>Without CBM</th>
<th>With CBM</th>
</tr>
</thead>
<tbody>
<tr>
<td>nAUC</td>
<td>0.924</td>
<td>0.952</td>
</tr>
<tr>
<td>NSS</td>
<td>0.845</td>
<td>1.50</td>
</tr>
<tr>
<td>EMD</td>
<td>5.26</td>
<td>3.90</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Optimal weights</th>
<th>Without CBM</th>
<th>With CBM</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Subject-specific</td>
<td>General</td>
</tr>
<tr>
<td>nAUC</td>
<td>0.944</td>
<td>0.945</td>
</tr>
<tr>
<td>NSS</td>
<td>1.32</td>
<td>1.35</td>
</tr>
<tr>
<td>EMD</td>
<td>4.41</td>
<td>4.33</td>
</tr>
</tbody>
</table>

Table 1. Quantitative comparisons of 6 models on the FIFA data set. The optimal weights for the general model are \([w_C \ w_l \ w_o \ w_f]_{\text{opt}}^T = [0.027 \ 0.024 \ 0.222 \ 0.727]^T\). The optimal weights for subject-specific models vary slightly, while the ranking of the four features—from faces and orientation to color and intensity—remains constant. “CBM” stands for Center Bias Modeling. The NSSs of the linear models with optimal weights are noticeably larger than those with equal weights, and the NSSs of models with CBM are larger than those without CBM, suggesting a greater correspondence between fixation locations and the salient points predicted by the models. The EMD is considerably smaller using optimal weights or/and with CBM, reflecting superior global consistency between saliency and fixation maps.
data and multiplied with the spatial-information-free saliency maps.

From Table 1, we observe that: (1) by setting proper weights to different feature channels from constrained linear least squares, the model improves significantly. This suggests that we do rely on certain features more than others in deciding where to look at and such features should be emphasized in the final saliency map. (2) Using both equal and optimal weights, models with center bias modeling perform consistently better than those without such a spatial prior. This confirms the previous discussions that when looking at an image, we tend to look at the center of the image. In laboratory setting where all the data were collected, this center bias is stronger than scenarios where subjects can freely move their heads. Computational saliency models that account for such a bias is shown to have better predictability power. (3) Compared with nAUC, the performance difference is indeed better reflected by NSS and EMD. (4) There is no significant improvement using subject-specific models over the general model. This suggests that—at least when averaging over 130 different images—subjects accord the same weights to faces, orientation, color, and intensity channels.

### Experiment 2—using the Toronto data set

Compared with the FIFA data set (Cerf et al., 2009), the Toronto data set (Bruce & Tsotsos, 2009) contains many fewer faces and other distinct large objects in an image and is therefore considered a more difficult data set. We divide the 120 images into 80 training images and 40 testing images. As there are fewer fixation data than the FIFA data set, we build only general models in this experiment and focus on the comparisons on models with equal weights and optimal weights and models with and without a center bias.

Table 2 presents the normalized AUC, NSS, and EMD values against the fixation data. Comparing the 4th column of Table 2 to the 4th column of Table 1, where both are results from models with learned weights, the results on the FIFA data set are better, consistent with the aforementioned fact that the many faces and other objects in the FIFA images consistently attract people’s gaze. Comparing the 2nd and 4th columns of Table 2 with their counterparts with center bias modeling (the 3rd and 5th columns), we see a significant performance improvement where the center bias is modeled.

Table 3 summarizes a performance comparison of four popular saliency algorithms against our model.

### Experiment 3—using the MIT data set

Following Judd et al. (2009), we divide the MIT data set into 903 training images and 100 testing images for experiment.

Earlier for the Toronto data set, we used the aggregate data from all 11 subjects and all fixations made in the viewing period to achieve high statistical confidence. Since there is a larger number of images and fixation data in this MIT data set, we also fit the model to individual subjects and measure the variance of model parameters for individual subjects. We conduct one additional experiment that tests how the optimal weights and model performance change as a function of the fixation.

Figure 6 plots the optimal weights as a function of the fixation. When the training data include only the first fixations on all images with all subjects, the learned weight of the face channel is the largest; it decreases monotonically as more fixations per image per subject are used. Both the orientation and color channels display a clear opposite trend. These results demonstrate that compared with other bottom-up channels, face attracts attention not only strongly but fast, consistent with findings reported in Cerf et al. (2009).

Figure 7 illustrates—using NSS—model performance with respect to the number of fixations made (different metrics were used for evaluation and suggested consistent conclusions here; therefore, only results in NSS are displayed to avoid redundancy). Results of 5 computational models are shown—(i) a center Gaussian model

---

### Table 2

Quantitative comparisons of 4 models on the Toronto data set. The optimal weights are \( \begin{bmatrix} w_C & w_I & w_O & w_F \end{bmatrix}^T = [0.403 \ 0.067 \ 0.530 \ 0]^T \). The weight for the face channel is 0 as there are few frontal faces in this data set. “CBM” stands for Center Bias Modeling.

<table>
<thead>
<tr>
<th>Models</th>
<th>Equal weights Without CBM</th>
<th>Equal weights With CBM</th>
<th>Optimal weights Without CBM</th>
<th>Optimal weights With CBM</th>
</tr>
</thead>
<tbody>
<tr>
<td>nAUC</td>
<td>0.828</td>
<td>0.943</td>
<td>0.834</td>
<td>0.948</td>
</tr>
<tr>
<td>NSS</td>
<td>0.872</td>
<td>1.49</td>
<td>0.920</td>
<td>1.54</td>
</tr>
<tr>
<td>EMD</td>
<td>4.85</td>
<td>3.09</td>
<td>4.50</td>
<td>2.90</td>
</tr>
</tbody>
</table>

### Table 3

Normalized AUC values for different saliency models on the Toronto data set. Our model is based on optimized weights and a center bias.

<table>
<thead>
<tr>
<th>Models</th>
<th>nAUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Itti et al. (Cerf et al., 2009; Itti et al., 1998)</td>
<td>0.828</td>
</tr>
<tr>
<td>Gao et al. (2007)</td>
<td>0.880</td>
</tr>
<tr>
<td>Bruce and Tsotsos (2009)</td>
<td>0.890</td>
</tr>
<tr>
<td>Hou and Zhang (2008)</td>
<td>0.903</td>
</tr>
<tr>
<td>Our model</td>
<td>0.948</td>
</tr>
</tbody>
</table>
where the variance of the Gaussian kernel is learned from
the 903 training images but no other biasing for faces,
orientation, color, and intensity takes place, (ii) the
standard model with equal weights, (iii) the standard
model with optimal weights, and (iv, v) two models with
center bias where the center Gaussian is multiplied with
the standard models to compensate for the center bias.

Figure 7 demonstrates that (1) models with optimal
weights outperform those with equal weights. (2) A center
Gaussian that models a spatial prior together with the
usual features consistently shows improved performance.
(3) Saliency decreases with time, consistent with the

Figure 6. Optimal weights (learned from all 15 subjects) with respect to viewing time for the MIT data set. The weight of face decreases while the weights for other channels increase, indicating that face attracts attention faster than the other channels.

Figure 7. Illustration of model performance with respect to viewing time for the MIT data set. The performance of all these bottom-up saliency models degrade with viewing time, as more top-down factors come into play.
findings (Mannan, Kennard, & Husain, 2009) that initial fixations are more driven by bottom-up features compared to later ones.

We train subject-specific optimal weights from fixation data of one particular subject and evaluate the resulting subject-specific model using nAUC, NSS, and EMD. For all 15 subjects, the face channel is weighted the most heavily, followed by orientation, color, and intensity. The mean of the optimal weights are \( \frac{w_C}{C_0}, \frac{w_I}{I_0}, \frac{w_O}{O_0}, \frac{w_F}{F_0} \) = \[0.109 0.072 0.278 0.541\]^T, and the standard deviation is \( \sigma_C, \sigma_I, \sigma_O, \sigma_F \) = \[0.028 0.022 0.039 0.054\]^T. We use the trained weights to build subject-specific saliency models, and the model performance is reported in the 6th and 8th columns of Table 4. Again, the improvement compared to the model trained on the population data is marginal. For a performance summary of 7 models (the aforementioned 5 models (Figure 7) and 2 subject-specific ones (the 6th and 8th columns)), see Table 4.

**General discussions**

In the visual search literature, it is well known that some features can be used more efficiently for the deployment of top-down attention than others (Burgess & Ghandeharian, 1984; Motter & Belky, 1998; Rajashekar, Bovik, & Cormack, 2006; Rao, Zelinsky, Hayhoe, & Ballard, 2002). In particular, color is a strongly guiding feature (Motter & Belky, 1998; Williams & Reingold, 2001; Williams, 1966), whereas at least one reports the opposite (Zelinsky, 1996). Recently, Ehinger, Hidalgo-Sotelo, Torralba, and Oliva (2009) report that among all global and local features considered the context model (Torralba, Oliva, Castelhano, & Henderson, 2006) is the best predictor in a search-people task. As suggested by Rutishauser and Koch (2007), such differences in predictability can occur as a result of which features define the target.

While these discussions focus on what features of the target are used preferentially to bias the search, the weighting of different features in bottom-up, free-viewing tasks is less investigated. We here learn a set of optimal feature weights using linear regression with constraints. Training data are taken from four recent eye-tracking data sets (Bruce & Tsotsos, 2009; Cerf et al., 2009; Judd et al., 2009; Subramanian et al., 2010), where 500 images were used for training and the remaining for testing. 

**Experiment 4—using the NUS data set**

Lastly, we conduct experiments on the NUS data set (Subramanian et al., 2010), where 500 images were used for training and the remaining for testing. 

Table 5 summarizes the performance of models with equal and optimal weights, with and without center bias modeling. Despite the considerably richer semantic contents in this data set, the weights of the four bottom-up channels are consistent with the other three data sets: face and orientation are more important than color and intensity. The performance of bottom-up saliency model is significantly improved after addressing difference strengths of features for perceptual saliency and with a center bias modeling.

<table>
<thead>
<tr>
<th>Equal weights</th>
<th>Optimal weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centered Gaussian</td>
<td>Without CBM</td>
</tr>
<tr>
<td>nAUC</td>
<td>0.904</td>
</tr>
<tr>
<td>NSS</td>
<td>1.06</td>
</tr>
<tr>
<td>EMD</td>
<td>3.20</td>
</tr>
</tbody>
</table>

Table 5. Quantitative comparisons of 5 models on the NUS data set. The optimal weights for the general model are \( \frac{w_C}{w_I}, \frac{w_O}{w_F} \) = \[0.054 0.049 0.256 0.641\]^T. “CBM” stands for Center Bias Modeling.
computational saliency models. We explicitly point out that performance should be judged by a combination of metrics, in our case, ROC, NSS, and EMD.

Appendix A

Derivation of fixation distribution

\[X_{t+1} \sim N(0, \Sigma_h) \cdot (X_t * N(0, \Sigma_f))\]

Similar as before, in the following derivations, symbols with ^ represent distributions and those with ~ denote instances of variables.

Recall that at the current fixation location \(\tilde{c}_t\), the \((t+1)\)th saccade distribution is \(\tilde{x}_{t+1} \sim N'(0, \Sigma_h) \cdot N(\tilde{c}_t, \Sigma_f)\) (Equation 4). Thus, given \(\tilde{c}_t\), the probability of the next fixation at \(\tilde{c}_{t+1}\) is given by \([N'(0, \Sigma_h) \cdot N(\tilde{c}_t, \Sigma_f)](\tilde{c}_{t+1})\) (in this derivation, we use \([\cdot]\) to denote a distribution and \([\cdot](\cdot)\) as a value of the distribution at a specific point).

Since the current fixation follows the distribution \(\tilde{x}_t\), the probability of fixating at a particular location \(\tilde{c}_t\) is \(\tilde{x}_t(\tilde{c}_t)\). Integrating the saccade distributions over all possible locations \(\tilde{c}_t\) yields

\[
\tilde{x}_{t+1}(\tilde{c}_{t+1}) = \int_{\tilde{c}_t} \tilde{x}_t(\tilde{c}_t) \cdot [N'(0, \Sigma_h) \cdot N(\tilde{c}_t, \Sigma_f)](\tilde{c}_{t+1}) d\tilde{c}_t
\]

\[
= \int_{\tilde{c}_t} \tilde{x}_t(\tilde{c}_t) \cdot [N'(0, \Sigma_h)](\tilde{c}_{t+1}) \cdot [N(\tilde{c}_t, \Sigma_f)](\tilde{c}_{t+1}) d\tilde{c}_t
\]

\[
= [N'(0, \Sigma_h)](\tilde{c}_{t+1}) \cdot \int_{\tilde{c}_t} \tilde{x}_t(\tilde{c}_t) \cdot [N(\tilde{c}_t, \Sigma_f)](\tilde{c}_{t+1}) d\tilde{c}_t
\]

\[
= [N'(0, \Sigma_h)](\tilde{c}_{t+1}) \cdot (\tilde{x}_t * [N(0, \Sigma_f)](\tilde{c}_{t+1})).
\]

(A1)

From Equation A1, we have

\[\tilde{x}_{t+1} \sim N'(0, \Sigma_h) \cdot (\tilde{x}_t * N(0, \Sigma_f)).\]  

(A2)

This completes the derivation.

Proof of Theorem 2.1

**Proof.** Since the \(t\)th covariance matrix \(\Sigma_t\) is given by

\[
\begin{pmatrix}
\sigma_t^2 & 0 \\
0 & \sigma_t^2
\end{pmatrix}
\]

the convergence of \(\{\Sigma_t\}_{t=1,2,\ldots}\) is equivalent to the convergence of \(\{\sigma_t^2\}_{t=1,2,\ldots}\). To prove the convergence of this series, it suffices to show that it is both
upper bounded and strictly monotonic increasing. First, recall that
(1) the convolution of two Gaussian functions is another Gaussian function, in particular,
\[
N(\mu_a, \sigma_a^2) * N(\mu_b, \sigma_b^2) \propto N(\mu_a + \mu_b, \sigma_a^2 + \sigma_b^2).
\]  
(A3)

(2) The multiplication of two Gaussian functions is also another Gaussian function, particularly
\[
N(\mu_a, \sigma_a^2) \cdot N(\mu_b, \sigma_b^2) \propto N(\mu_c, \sigma_c^2),
\]  
(A4a)
where
\[
\sigma_c^2 = ((\sigma_a^2)^{-1} + (\sigma_b^2)^{-1})^{-1}.
\]  
(A4b)

Since \( \hat{X}_{t+1} \sim N(0, \Sigma_c) \cdot \hat{X}_t * N(0, \Sigma_b) \), using Equations A3, A4a, and A4b, we obtain
\[
\sigma_{t+1}^2 = ((\sigma_t^2 + \sigma_f^2)^{-1} + (\sigma_h^2)^{-1})^{-1},
\]  
(A5a)
and
\[
\sigma_t^2 = ((\sigma_f^2)^{-1} + (\sigma_h^2)^{-1})^{-1}.
\]  
(A5b)

From Equation A5a, we see that \( \sigma_t^2 < \sigma_f^2 \) for \( t = 1, 2, \ldots \), therefore \( \{\sigma_t^2\}_{t=1,2,\ldots} \) is upper bounded by \( \sigma_f^2 \).
To prove monotonicity, let \( Q(t) \) be the statement that \( \sigma_{t+1}^2 > \sigma_t^2 \). This is equivalent to
\[
f_{\text{diff}}(t) = \sigma_{t+1}^2 - \sigma_t^2 = \frac{\sigma_t^4 + \sigma_f^2 \cdot \sigma_f^2 - \sigma_f^2 \cdot \sigma_h^2}{\sigma_t^2 + \sigma_f^2 + \sigma_h^2} > 0.
\]  
(A6)

Basic step. To prove \( Q(1) \), simply substitute Equations A5a and A5b into Equation A6:
\[
f_{\text{diff}}(1) = \frac{\sigma_f^2 \cdot \sigma_h^2}{(\sigma_f^2 + \sigma_h^2)(\sigma_f^2 + \sigma_f^2 + \sigma_h^2)} > 0,
\]  
(A7)

therefore, \( \sigma_2^2 > \sigma_1^2 \).

Inductive step. This time we assume \( Q(t - 1) \), i.e., \( \sigma_t > \sigma_{t-1} \), and prove \( Q(t) \).

Combining Equation A5a and \( \sigma_t^2 > \sigma_{t-1}^2 \) results
\[
\sigma_{t+1}^2 = ((\sigma_t^2 + \sigma_f^2)^{-1} + (\sigma_h^2)^{-1})^{-1} > ((\sigma_{t-1}^2 + \sigma_f^2)^{-1} + (\sigma_h^2)^{-1})^{-1} = \sigma_t^2.
\]  
(A8)

This proves \( Q(t) \) and completes the proof.

**Acknowledgments**

This research was supported by the NeoVision Program at DARPA, the ONR, the Mathers Foundation, and the WCU (World Class University) Program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (R31-10008).

Commercial relationships: none.
Corresponding author: Qi Zhao.
Email: qzhao@klab.caltech.edu.
Address: BBB, 216-76, Caltech, 1200 E. California Blvd., Pasadena, CA 91125, USA.

**References**


