Learning visual saliency by combining feature maps in a nonlinear manner using AdaBoost

Qi Zhao

Computation and Neural Systems, California Institute of Technology, Pasadena, CA, USA
Now at Department of Electrical and Computer Engineering, National University of Singapore, Singapore

Christof Koch

Computation and Neural Systems, California Institute of Technology, Pasadena, CA, USA
Allen Institute for Brain Science, Seattle, WA, USA

To predict where subjects look under natural viewing conditions, biologically inspired saliency models decompose visual input into a set of feature maps across spatial scales. The output of these feature maps are summed to yield the final saliency map. We studied the integration of bottom-up feature maps across multiple spatial scales by using eye movement data from four recent eye tracking datasets. We use AdaBoost as the central computational module that takes into account feature selection, thresholding, weight assignment, and integration in a principled and nonlinear learning framework. By combining the output of feature maps via a series of nonlinear classifiers, the new model consistently predicts eye movements better than any of its competitors.

Keywords: AdaBoost, computational saliency model, feature integration

Citation: Zhao, Q., & Koch, C. (2012). Learning visual saliency by combining feature maps in a nonlinear manner using AdaBoost. *Journal of Vision, 12*(6):22, 1–15, http://www.journalofvision.org/content/12/6/22, doi:10.1167/12.6.22.

Introduction

Humans and other primates shift their gaze to allocate processing resources to a subset of the visual input. Understanding and emulating the way that humans free-view natural scenes has both scientific and economic impact. Thus, a computational model predicting where humans look has general applicability in a wide range of tasks relating to surveillance, in-home healthcare, advertising, and entertainment.

In the past decade, a large body of computational models (Itti, Koch, & Niebur, 1998; Parkhurst, Law, and Niebur, 2002; Oliva, Torralba, Castelhano, & Henderson, 2003; Walther, Serre, Poggio, & Koch, 2005; Foulsham & Underwood, 2008; Einhäuser, Spain, & Perona, 2008; Masciocchi, Mihalas, Parkhurst, & Niebur, 2009; Chikkerur, Serre, Tan, & Poggio, 2010) have been proposed to predict gaze allocation, many of which were inspired by neural mechanisms (Koch & Ullman, 1985). A common approach is to (a) extract visual features such as color, intensity, and orientation, (b) compute individual feature maps using biologically plausible filters such as Gabor or Difference of Gaussian filters, and (c) linearly integrate these maps to generate a final saliency map.

Our study focused on the last step—irrespective of the method used for extracting and mapping features, we investigated general principles of how different features are integrated to form the saliency map. In the literature, there are physiological and psychophysical studies in support of "linear summation" (i.e., linear integration with equal weights) (Treisman & Gelade, 1980; Nothdurft, 2000; Engmann et al., 2009), "max" (Li, 2002; Koene & Zhaoping, 2007), or "MAP" (Vincent, Baddeley, Troscianko, & Gilchrist, 2009) type of integration, in which linear summation has been commonly employed in computational modeling (Itti et al., 1998; Cerf, Frady, & Koch, 2009; Navalpakkam & Itti, 2005). Later, under the linear assumption, (Itti & Koch, 1999) suggested various ways to normalize the feature maps. Hu, Xie, Ma, Chia, and Rajan (2004) computed feature contributions to saliency using "composite saliency indicator," a measure based on spatial compactness and saliency density. Zhao and Koch (2011) analyzed the image features at fixated and nonfixated locations directly from eye tracking data and quantified the strengths of different visual features in saliency using constrained linear regression.

Lacking sufficient neurophysiological data concerning the neural instantiation of feature integration for saliency, we did not make any assumption regarding the integration stage. Instead, we took a data-driven

doi: 10.1167/12.6.22 Received May 25, 2011; published June 15, 2012 ISSN 1534-7362 © 2012 ARVO

approach and studied feature integration strategies based on human eye movement data. In particular, the new model aims to automatically and simultaneously address the following issues: (a) select a set of features from a feature pool in the absence of assuming which feature type is needed, (b) find an optimal threshold for each feature, and (c) estimate optimal feature weights according to their contributions to perceptual saliency.

Our approach

We used AdaBoost as a unifying framework to address these issues. The new model learns the integration of features at multiple scales in a principled manner and with the following key advantages. First, it selects from a feature pool the most informative features that have significant variability. This framework can easily incorporate any additional features and select the best ones in a greedy manner. Second, it finds the optimal threshold for each feature (Li, 2002). Third, it makes no assumption of linear superposition or equal weights of features. Indeed, we explicitly demonstrated that certain types of nonlinear combination significantly and consistently outperform linear combinations. This raises the question of the extent to which the primate brain takes advantages of such nonlinear integration strategies. Future psychophysical and neurophysiological research are needed to answer this question.

Related work

Different criteria to quantify saliency exist. Itti and Baldi (2006) hypothesized that the information-theoretical concept of spatio-temporal surprise is central to saliency. Raj, Geisler, Frazor, and Bovik (2005) derived an entropy minimization algorithm to select fixations. Seo and Milanfer (2009) computed saliency using a “self-resemblance” measure, in which each pixel of the saliency map indicates the statistical likelihood of saliency of a feature matrix given its surrounding feature matrices. Bruce and Tsotsos (2009) presented a model based on “self-information” after Independent Component Analysis (ICA) decomposition (Hyvarinen & Oja, 2000) that is in line with the sparseness of the response of cortical cells to visual input (Field, 1994). Wang, Wang, Huang, and Gao (2010) defined the Site Entropy Rate as a saliency measure, also after ICA decomposition. In most of the saliency models, features are predefined. Some commonly used features include contrast (Reinagel & Zador, 1999), edge content (Baddeley & Tatler, 2006), intensity bispectra (Krieger, Rentschler, Hauske, Schill, & Zetzsche, 2000), color (Jost, Ouerhani, von Wartburg, Müri, & Hügli, 2005), and symmetry (Privitera & Stark, 2000), as well as more semantic ones such as faces and text (Cerf et al., 2009). On the other hand, various inference algorithms were designed for saliency estimation. For example, Avraham and Lindenbaum (2009) used a stochastic model to estimate the probability that an image part is of interest. In Harel, Koch, and Perona (2007), an activation map within each feature channel was generated based on graph computations. In Carbone and Pirri (2010), a Bernouli mixture model is proposed to capture context dependency.

In contrast to these models that were built upon particular image features, inference algorithms, or objective functions, another category of saliency models that learn from eye movement data, have become popular. Kienzle, Wichmann, Scholkopf, and Franz (2006) directly learnt from raw image patches whether or not these were fixated. Determining the size and resolution of the patches was not straightforward, and compromises had to be made between computational tractability and generality. In addition, the high dimensional feature vector of an image patch required a large number of training samples. Subsequently, Judd, Ehinger, Durand, and Torralba (2009) constructed a large eye tracking database and learnt a saliency model based on low, middle, and high-level image features. Both Kienzle et al. (2006) and Judd et al. (2009) applied support vector machine (SVM) to learn the saliency models. Different from these approaches, we used an AdaBoost-based model that combines a series of base classifiers to model the complex input data. With an ensemble of sequentially learned models, each base model covered a different aspect of the data set (Jin, Liu, Si, Carbonell, & Hauptmann, 2003). In addition, unlike Kienzle et al. (2006) who used raw image data, and Judd et al. (2009) who included a set of hand-crafted features, the features in the current framework are simple, biologically plausible ones (Itti et al., 1998; Parkhurst et al., 2002; Cerf et al., 2009).

Alternative approaches employ learning based saliency models based on objects rather than pixels (Khuwuthyakorn, Robles-Kelly, & Zhou, 2010; Liu et al., 2011). To make the object detection step robust and consistent, pixel neighborhood information is considered. Thus, Khuwuthyakorn et al. (2010) extended generic image descriptors of Itti et al. (1998) and Liu et al. (2011) to a neighborhood-based descriptor setting by considering the interaction of image pixels with neighboring pixels. In other efforts, Conditional Random Field (CRF) (Lafferty, McCallum, & Pereira, 2001) that encodes interaction of neighboring pixels effectively detects salient objects in images (Liu et al., 2011) and videos (Liu, Zheng, Ding, & Yuan, 2008), although CRF learning and inference are quite slow.

The section Features and bottom-up saliency model introduces the architecture of the standard saliency model and the bottom-up features used. Learning

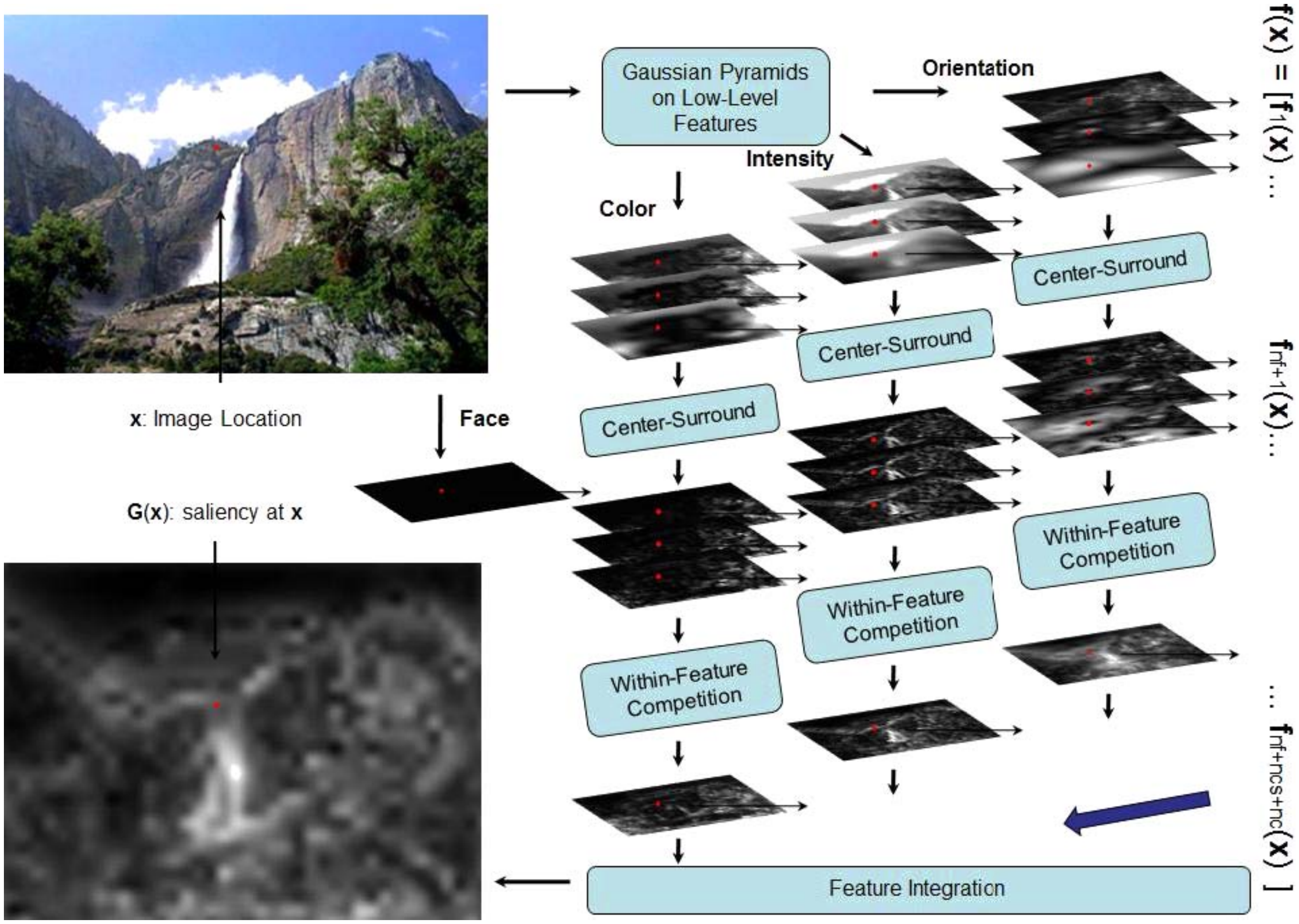

Figure 1. Illustration of the bottom-up saliency model. A sample vector for learning from one particular location x (marked in red) is shown on the right.

nonlinear feature integration using AdaBoost presents the new architecture. Experimental results demonstrates comparative and quantitative results, and General discussions and future work concludes the paper and discusses future work.

Features and bottom-up saliency model

To focus on the comparisons of feature integration mechanisms for saliency, we used a simple set of biologically plausible features (Itti et al., 1998; Cerf et al., 2009) that included two color channels (blue/yellow and red/green), one intensity channel, and four orientation channels (0°, 45°, 90°, 135°). For each of these channels, (a) six raw maps of spatial scales (2–7) were created using dyadic Gaussian pyramids, (b) six center-surround (c-s) difference maps were then constructed as point-wise differences across pyramid scales to capture local contrasts (center level $c = \{2, 3, 4\}$, surround level $s = c + \delta$, where $\delta = \{2, 3\}$), and (c) a single conspicuity map for each of the seven features was built through across-scale addition and within-feature competition (represented at scale 4). We included a single face channel, generated by running the Viola and Jones face detector (Viola & Jones, 2001). Although different from more primitive features such as color, intensity, and orientation, faces attract gaze in an automatic and task-independent manner (Cerf et al., 2009). The architecture of our model is illustrated in Figure 1.

Previous models (Itti et al., 1998; Cerf et al., 2009) generated a saliency map based on summation of the c-s maps. We here used information directly from all $nf = 42$ raw feature maps, $ncs = 42$ c-s maps, and $nc = 4$ conspicuity maps to construct the feature vectors for learning. As shown in Figure 1, for an image location $\mathbf{x}$, the values of all the relevant maps at this particular location were extracted and stacked to form the sample vector $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x})\ f_2(\mathbf{x}) \cdots f_{nf+1}(\mathbf{x}) \cdots f_{nf+ncs+1}(\mathbf{x}) \cdots f_{nf+ncs+nc}(\mathbf{x})]^T$, where T is the transpose of the vector.

Each feature vector has a label of either 1 or −1, indicating whether the location is fixated or not. Fixation maps were constructed by convolving recorded fixations with an isotropic Gaussian kernel (Zhao &

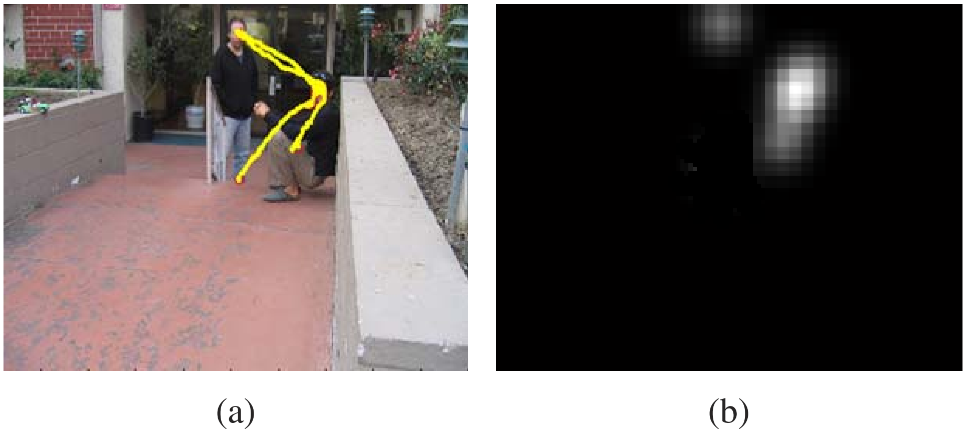

Figure 2. Fixation map illustration. (a) Original image with eye movements of one subject. (b) Fixation map of the same subject free-viewing the image shown in (a) (the first fixation is the center of the image and not included in the fixation map).

Koch, 2011); an example is shown in Figure 2b. 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

$$H_i^j(\mathbf{x}) = \alpha \sum_{k=2}^{f} K_G\left(\frac{\mathbf{x} - \mathbf{x}_k}{h}\right), \tag{1}$$

where $\mathbf{x}$ denotes the two-dimensional image coordinates. $\mathbf{x}_k$ represents the image coordinates of the kth fixation, and f is the number of fixations. The bandwidth of the kernel, h, is set to approximate the size of fovea, and α normalizes the map. The fixation maps were represented at the same scale as the conspicuity maps (to avoid too large maps, we also limited the largest spatial dimension to 40 [Harel et al., 2007]). We set $h = 2$ to approximate the size of fovea.

Note that the first fixation in each image is not used as it is—by design—always the center of the image. To assign a label to each sample vector, the continuous fixation map is converted into binary labels by using a sampling technique: locations of positive examples are sampled from the maps (i.e., an image location with a larger value in the fixation map has a higher probability of being sampled as a positive sample), and locations of negative examples are sampled uniformly from nonactivated areas (i.e., with values smaller than a threshold $th = 0.001$ in our implementation) of the fixation maps.

Learning nonlinear feature integration using AdaBoost

To quantify the relevance of different features across multiple scales in deciding where to look, we learned—using AdaBoost—nonlinear integration of features $G(\mathbf{f}) : R^d \rightarrow R$, where d is the dimensionality of the feature space. The AdaBoost algorithm (Freund & Schapire, 1996; Friedman, Hastie, & Tibshirani, 1998; Schapire & Singer, 1999; Vezhnevets & Vezhnevets, 2005) is one of the most effective methods for object detection (Viola & Jones, 2001; Chen & Yuille, 2004). As a special case of boosting, the final strong classifier is a weighted combination of weak classifiers that are iteratively built. Subsequent weak classifiers are tweaked in favor of the misclassified instances.

Formally,

$$G(\mathbf{f}) = \sum_{t=1}^{T} \alpha_t g_t(\mathbf{f}), \tag{2}$$

where $g_t(\cdot)$ denotes the weak learner and $G(\cdot)$ the final classifier, here an estimate of saliency. α_t is the weight of $g_t(\cdot)$, as would be described in Algorithm 1 below. T is the number of weak classifiers. Instead of taking the sign of the AdaBoost output, as conventionally in classification, we use the real value $G(\mathbf{f})$ to form a saliency map. Details are described in Algorithm 1 and in Figure 3.

Algorithm 1 Learning A Nonlinear Feature Integration Using AdaBoost

Input: Training dataset with N images and eye movement data from M subjects. A testing image Im.

Output: Saliency map associated with Im.

Training stage:

1. For all locations in N images, sample $\{\mathbf{x}_s\}_{s=1}^S$ with labels $\{y_s\}_{s=1}^S$. See Features and bottom-up saliency model for details of sampling. Compute $\{\mathbf{f}_s\}_{s=1}^S = \{\mathbf{f}(\mathbf{x}_s)\}_{s=1}^S$ as a stack of features for the sample at location $\mathbf{x}_s$.
2. Initialize weights to be $\{w_s = \frac{1}{S}\}_{s=1}^S$, where S is the number of samples.

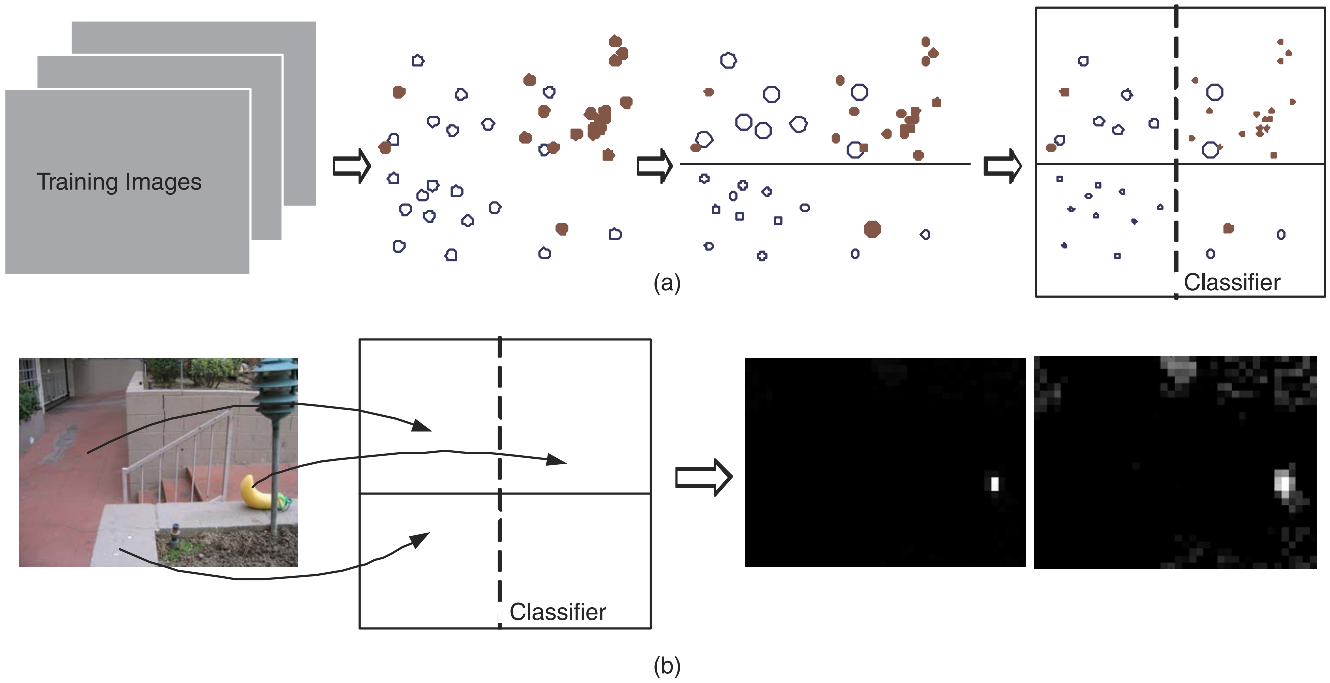

Figure 3. Illustration of the AdaBoost-based saliency model. (a) Training stage: using samples from training images, weak classifiers are trained through iterations and combined to form a strong classifier. We use a *d*-dimensional (a typical value of *d* in this work is 88) space but plot a two-dimensional one here for illustration. (b) Testing stage: for a new image, the feature vectors of image locations are calculated and input to the strong classifier to obtain saliency scores. For the two output maps shown on the right, the left map is the output of the strong classifier, where brighter regions denote more salient areas; and the right map is the result of the left map after a sigmoid transformation for illustration purposes.

3. For $t = 1, \ldots, T$ (where T is the number of weak classifiers)
   a. Train a weak classifier g_t: $R^d \rightarrow \{-1, 1\}$, which minimizes the weighted error function $g_t = \arg\min_{g_u \in G}$, where

$$\varepsilon_u = \sum_{s=1}^{S} w_t(s)[y_s \neq g_u(\mathbf{f}_s)].$$

   b. Set the weight of g_t as

$$\alpha_t = \frac{1}{2} ln \frac{1 - \varepsilon_t}{\varepsilon_t}.$$

   c. Update sample weights

$$w_{t+1}(s) = \frac{w_t(s) exp[-\alpha_t \cdot y_s \cdot g_t(\mathbf{f}_s)]}{Z_t},$$

   where Z is a normalization factor.
4. Saliency is defined as

$$G(\mathbf{f}) = \sum_{t=1}^{T} \alpha_t g_t(\mathbf{f}).$$

Testing stage (for a new image *Im*): for each location $\mathbf{x}$ in *Im*, compute the feature vector $\mathbf{f}(\mathbf{x})$, then apply the strong classifier $G(\mathbf{f}) : R^d \rightarrow R$ (Equation 2) to obtain the saliency value of $\mathbf{f}(\mathbf{x})$.

The algorithm has two stages. During the initial training stage, weak classifiers are built iteratively and combined to form the final strong classifier. In the second, testing stage, new images are applied with the strong classifier, which outputs saliency scores.

It is worth noting that the AdaBoost algorithm can be interpreted as a greedy feature selection process that selects from a feature set good ones that nevertheless have significant variety (Viola & Jones, 2001). We restrict each weak learner to depend on a single channel. As shown in Step 3a of Algorithm 1, the algorithm goes over all feature dimensions and picks the feature channel that best separates the positive and negative examples (while testing each of the channels, thresholds are tested in an exhaustive manner, and the threshold with the smallest classification error is used for that particular channel). Iteratively, AdaBoost automatically selects features from a feature pool and finds the optimal threshold and the optimal weight for each feature channel it selects. The total number of weak classifiers, T in Step 3, can correspond to the number of features in a feature pool, or to a much smaller value.

In Step 3c, samples are reweighted such that those misclassified by previously trained weak classifiers are

		Linear integration with optimal weights			Nonlinear integration		
		Subject-specific			Subject-specific		
	Linear summation	Mean	SD	General	Mean	SD	General
nAUC	0.924	0.945	0.0136	0.944	0.959	0.0161	0.953
NSS	0.845	1.35	0.0720	1.32	1.47	0.0711	1.42
EMD	5.26	4.33	0.236	4.41	2.68	0.150	2.87

Table 1. Quantitative comparison when integrating the four channels in a linear summation, optimal linear, or nonlinear manner on the FIFA dataset. The nonlinear model outperforms the linear ones on all three measures.

weighted more in future iterations. Thus, subsequently trained weak classifiers will place emphasis on these previously misclassified samples. Intuitively, a variety of feature channels is obtained by such a reweighting mechanism since subsequently selected weak learners that best separate the reweighted samples are usually very different from those channels already selected and that misclassified the samples.

Experimental results

Experimental paradigm

Datasets

This study analyzed eye movements from four recent datasets (Cerf et al., 2009; Bruce & Tsotsos, 2009; Judd et al., 2009; Subramanian, Katti, Sebe, Kankanhalli, & Chua, 2010).

In the FIFA dataset (Cerf et al., 2009), fixation data were collected from eight subjects performing a 2-sec-long "free-viewing" task on 180 color natural images (28° × 21°). Participants were asked to rate, on a scale of 1 through 10, how interesting each image was. Scenes were colored indoors and outdoors still images. Images included faces in various skin colors, age groups, gender, positions, and sizes.

The second dataset from Bruce and Tsotsos (2009), referred to 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 sec each. One distinction between this dataset and the FIFA one (Cerf et al., 2009) is that a large portion of images here do not contain particularly regions of interest, while in the FIFA dataset most contain very salient regions (e.g., faces, or salient nonface objects).

The eye tracking dataset published by Judd et al. (2009) (referred to as the MIT database) is the largest one in the community. It includes 1,003 images collected from *Flickr creative commons* and *LabelMe*. Eye movement data were recorded from 15 users who free-viewed these images (36° × 27°) for 3 sec each. A memory test was provided at the end to motivate the 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 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 part of the image set for 5 sec each (each image of size 26° × 19° was viewed by an average of 25 subjects).

Similarity measures

Unlike most saliency papers that used solely the area under the ROC curve (AUC) to quantify model performance, Zhao and Koch (2011) showed that in practice, as long as hit rates are high, the AUC is always high regardless of the false alarm rate. Therefore, an ROC analysis is, by itself, insufficient to describe the deviation of predicted fixation patterns

			Nonlinear integration			
				Raw/c-s feature level		
	Linear summation	Linear integration with optimal weights	Conspicuity level 4 channels	Top 10 of 88 channels	88 channels	88 channels with CBM
nAUC	0.828	0.834	0.836	0.912	0.916	0.982
NSS	0.872	0.920	0.913	1.35	1.37	1.88
EMD	4.85	4.50	3.66	3.28	3.20	2.11

Table 2. Quantitative comparisons of linear and nonlinear integrations on the Toronto dataset. "CBM" stands for Center Bias Modeling.

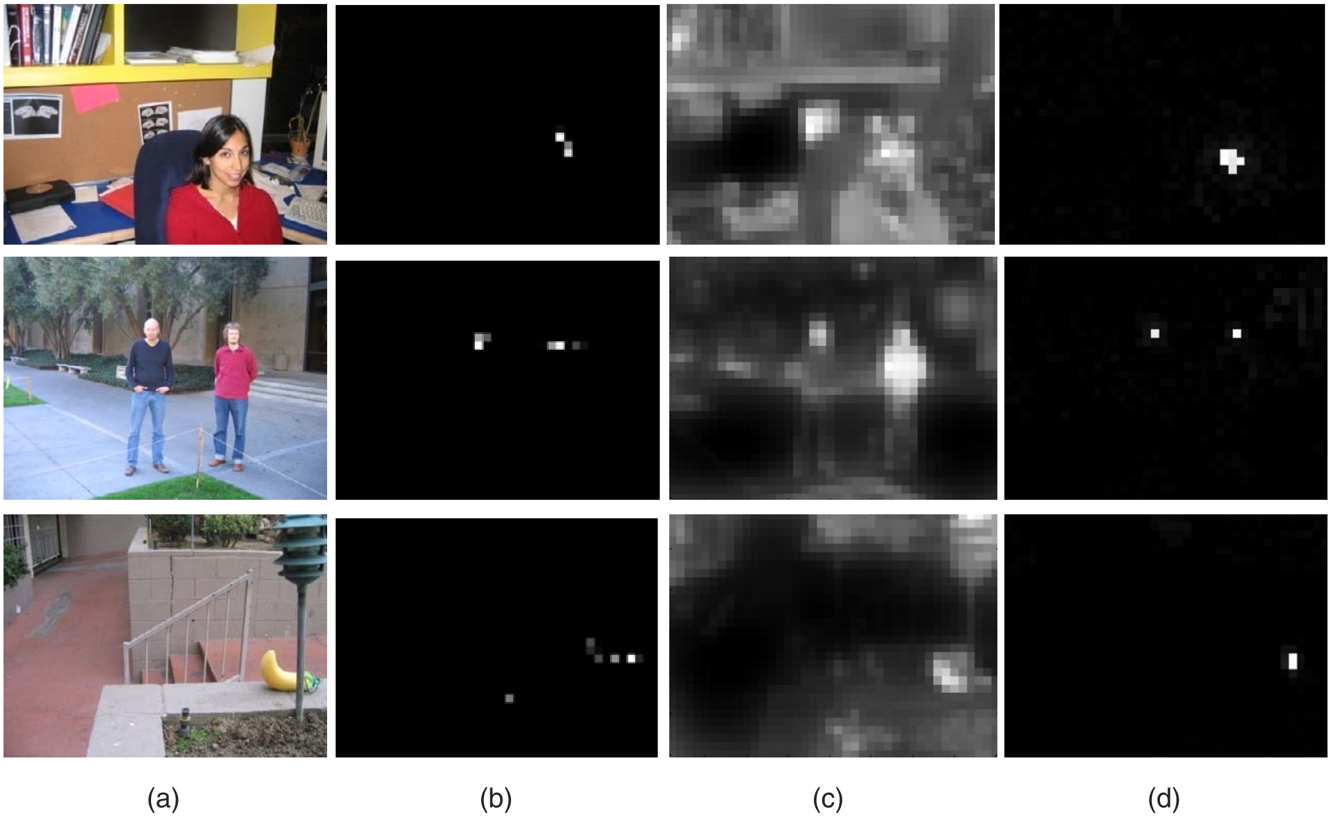

Figure 4. Comparisons between the linear and nonlinear algorithms. (a) Sample images from the FIFA dataset. (b) Fixation map for one subject. (c) Saliency map from the standard linear summation model. (d) Saliency map from AdaBoost learning using subject-specific data.

from the actual fixation map. We use three complementary similarity measures (Zhao & Koch, 2011) for a more comprehensive assessment—AUC in addition to the Normalized Scanpath Saliency (NSS) (Parkhurst et al., 2002; Peters, Iyer, Itti, & Koch, 2005) and the Earth Mover's Distance (EMD) (Rubner, Tomasi, & Guibas, 2000) that measure differences in value. Both AUC and NSS compare maps primarily at the exact locations of fixation, while EMD accommodates shifts in location and reflects the overall discrepancy between two maps on a more global scale.

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 intersubject variability. The ideal AUC is computed 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 ideal AUC values were 78.6% for the FIFA dataset, 87.8% for the Toronto dataset, 90.8% for the MIT dataset, and 85.7% for the NUS dataset (Zhao & Koch, 2011). We express the performance of saliency algorithms in terms of normalized AUC (nAUC) values, which is the AUC 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 an EMD value close to 0.

Performance

For our comparisons, we used both linear models with equal weights (Itti et al., 1998; Cerf et al., 2009) and linear models with optimal weights (Zhao & Koch, 2011), for which the weights of the four conspicuity maps (i.e., color, intensity, orientation, and face maps) were optimized by a linear regression with constraints.

We divided each dataset into a training set and a testing set, and sampled 10 positive samples (i.e., fixated locations) and the same number of negative samples (i.e., locations that were not fixated) from each image for training and testing.

FIFA dataset

In the first experiment, we compared linear summation and nonlinear integration on the FIFA dataset. The dataset of 180 images was divided into 130 training and 50 testing images. We trained subject-specific models using eye movement data from one observer,

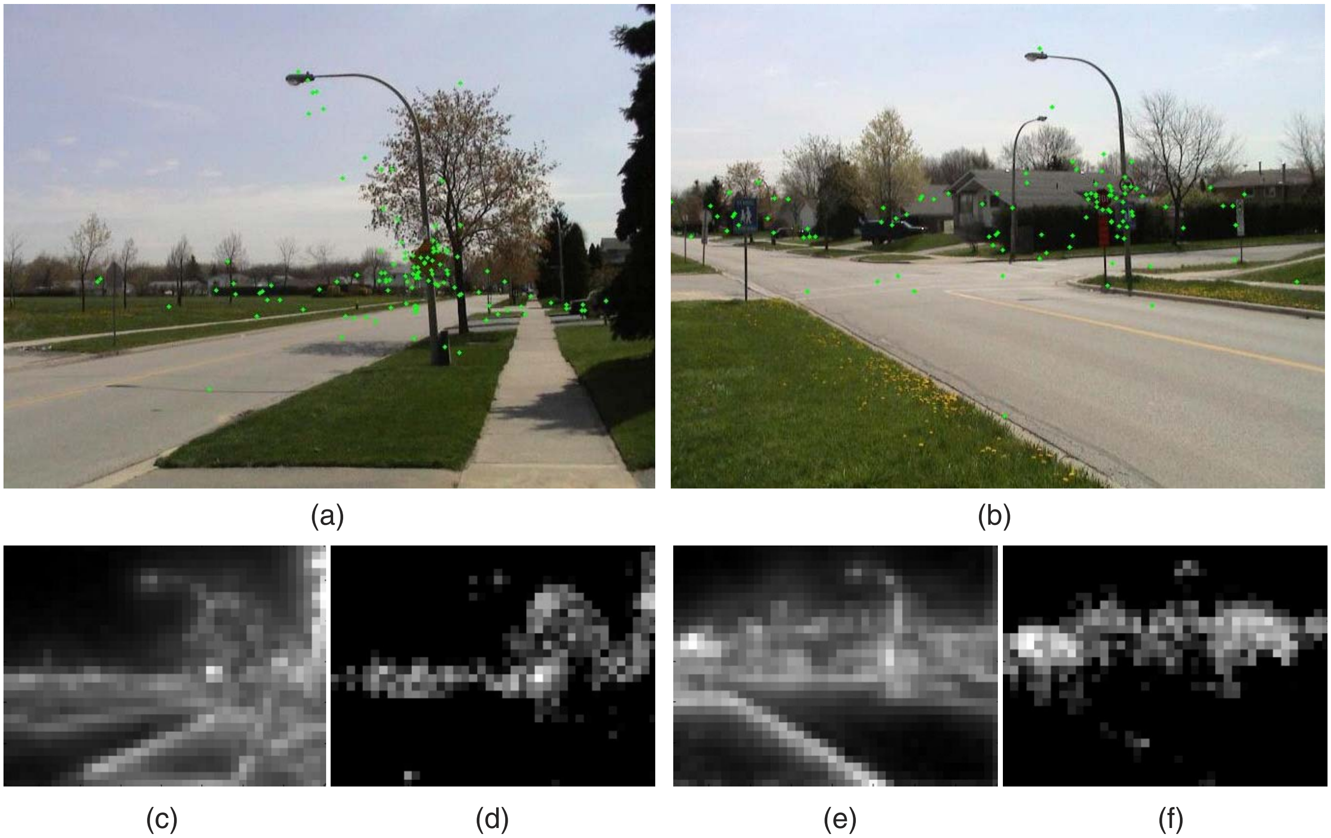

Figure 5. Comparisons between the linear and nonlinear algorithms. (a and b) Two sample images from the Toronto dataset with fixations (green dots) from all subjects. (c and e) Saliency maps from the standard linear summation model. (d and f) Saliency maps from AdaBoost learning.

and a subject-independent, general model using data from all eight subjects. We limited the nonlinear integration to the level of conspicuity maps and included three conspicuity maps and a face map in the feature pool (that is, $d = 4$).

As illustrated in Figure 4 and Table 1, AdaBoost learning outperformed linear summation.

Using all three similarity measures, the AdaBoost-based nonlinear integration outperformed the linear integration with optimal weights (Zhao & Koch, 2011), for which the weights for the four conspicuity maps were optimized, although the performance difference was better reflected by NSS and EMD, compared with nAUC. The performance differences are not significant between the average and the subject-specific data, suggesting unremarkable intersubject variability in terms of feature preference.

Toronto dataset

We divided the image set of 120 images into 80 training images and 40 testing ones. Because there were fewer fixations in the Toronto dataset than in the FIFA dataset, we built only general models in this experiment. When the conspicuity maps and the face map were used as candidate features ($d = 4$), nonlinearity was performed at a coarser level because conspicuity maps were constructed by linear summations of the c-s maps. In contrast, when all feature channels described in Features and bottom-up saliency model were included ($d = 88$), we were exploiting nonlinearity at a deeper level.

Figure 5 illustrates comparative results with linear summation and nonlinear integration. Subjects tended to fixate on regions of trees, houses, and poles, rather than the roads, grasses, or their boundaries that evoked a strong response in the linear summated saliency map (Figure 5c and e). Here, linear summation did not work well. Using a learning based approach, however, the model could learn from the fixation data which featured combinations that are more likely to attract attention.

A quantitative summary of linear integrations and nonlinear integrations with different levels is shown in Table 2. Comparing the fourth column of Table 2 to the last column of Table 1, in which both use four candidate features, the results on the FIFA dataset are better, consistent with the aforementioned fact that the FIFA dataset was relatively easier due to the presence of large faces and objects in most of the images.

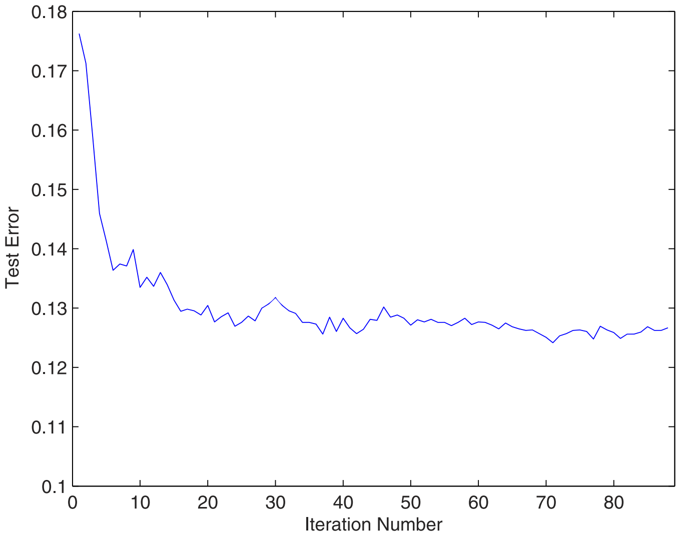

Figure 6. Test error as AdaBoost adds up to 88 weak classifiers, tested on the Toronto dataset. Performance does not increase much after 10 iterations.

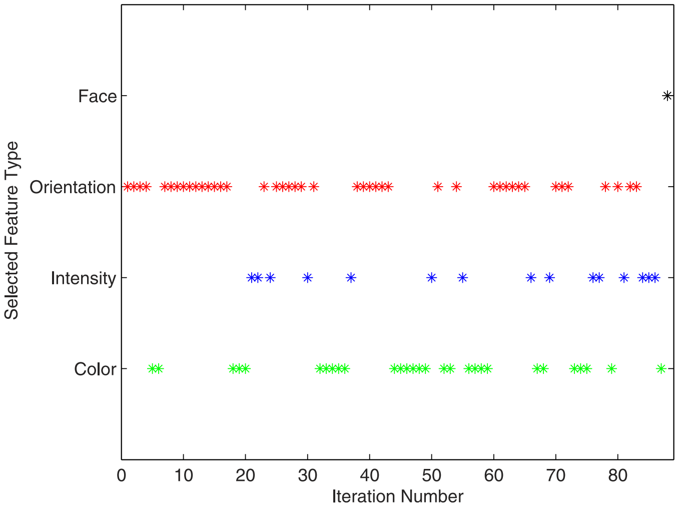

Figure 7. Selected feature type of AdaBoost as a function of the number of weak classifiers (on the Toronto dataset). Each iteration selects one feature in the four categories of color, intensity, orientation, and face (denoted by * in different colors). The features selected by the first 10 classifiers (shown to be the most important from Figure 6) are mostly orientation channels.

Figure 6 illustrates the test error as a function of the number of weak classifiers. It suggests that the performance does not increase noticeably after around 10 classifiers are included. To further show selected features by AdaBoost, Figures 7 and 8 illustrate the feature selection results as a function of iteration number. Particularly, in Figure 7, we divided all 88 feature maps into four categories of color, intensity, orientation, and face, and results indicated that 8 out of 10 top features were orientation channels (the remaining two were color channels). We broke down all channels but the face channel into their constitutive scale-dependent maps: raw feature maps of six scales and center-surround maps of six combinations, each with two color, one intensity, and four orientation channels, as introduced in Features and bottom-up saliency model. Figure 8 illustrates a relatively even distribution of the top 10 features over different scales, showing the necessity of using a multiscale approach to capture saliency.

We built saliency maps using the top 10 features selected by AdaBoost from the feature pool, as well as saliency maps using all features. The fifth and sixth columns in Table 2 indicate that after selecting the most discriminative features, the rest did not improve performance much. This, along with Figure 6, demonstrates the feature selection capability of AdaBoost: the algorithm selects the best features without efforts from domain experts. On this dataset, the most informative features selected by AdaBoost were orientation maps at various scales. On the FIFA, MIT, and NUS datasets, the most important feature was the face channel, followed by orientation, color, and intensity. The Toronto dataset ranked the face channel low since the dataset included few frontal faces.

AdaBoost can act in the same way as the lasso prior (Friedman, Hastie, Rosset, Tibshirani, & Zhu, 2004), and its cost function is different from the linear model. To rule out the probability that the improved performance is due to these two factors, we performed lasso regression and lasso logistic regression, using the same data and features. The lasso type regression was particularly helpful in some cases due to its tendency to prefer solutions with fewer nonzero coefficients, reducing the number of variables. We used it with 88 feature channels rather than 4. We obtained AUC: 0.881; NSS: 1.16; EMD: 3.91 with lasso regression and AUC: 0.886; NSS: 1.20; EMD: 3.85 with lasso logistic regression. The performance was better than the models using four conspicuity maps as features, in which each sub-feature-map belonging to the four broad feature categories (i.e., color, intensity, orientation, face) were inherently linearly added with equal weights to produce

					Our model	
Models	Itti et al. (1998)	Gao et al. (2007)	Bruce & Tsotsos (2009)	Hou & Zhang (2008)	Without CBM	With CBM
nAUC	0.828	0.880	0.890	0.903	0.916	0.982

Table 3. Normalized AUC for different saliency models on the Toronto dataset. “CBM” stands for Center Bias Modeling.

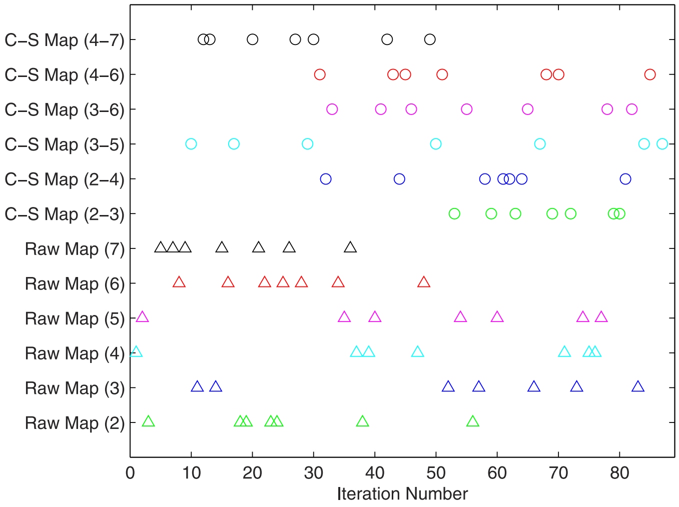

Figure 8. Selected scale as AdaBoost adds more and more classifiers (on the Toronto dataset). "7" is the spatial scale containing the highest spatial frequencies.

one conspicuity map for each feature category. On the other hand, with the same number of feature channels (i.e., 88), the models with various regression methods still underperform the one with AdaBoost, showing that the performance increase of AdaBoost was not due to the lasso prior or the different form of the cost function.

Under standard testing conditions, a strong center bias is seen (Tatler, 2007; Zhang, Tong, Marks, Shan, & Cottrell, 2008; Zhang, Tong, & Cottrell, 2009; Judd et al., 2009). In Zhao and Koch (2011), both time-varying (Bahill, Adler, & Stark, 1975; Pelz & Canosa, 2001; Parkhurst et al., 2002; Peters et al., 2005; Gajewski, Pearson, Mack, Bartlett, & Henderson, 2005) and constant (Zambarbieri, Beltrami, & Versino, 1995; Fuller, 1996; Vitu, Kapoula, Lancelin, & Lavigne, 2004; Le Meur, Le Callet, Barba, & Thoreau, 2006; Tatler, 2007; Zhang et al., 2008; Zhang et al., 2009; Judd et al., 2009) factors that contribute to this bias were considered. The center bias was modeled as a dynamic process that can be well approximated using a single kernel (Zhao & Koch, 2011). We built a center model that is a Gaussian function learned from the training data and multiplied it by the saliency maps to compensate for the center bias. Performance was boosted by considering such a spatial prior term (last column of Table 2).

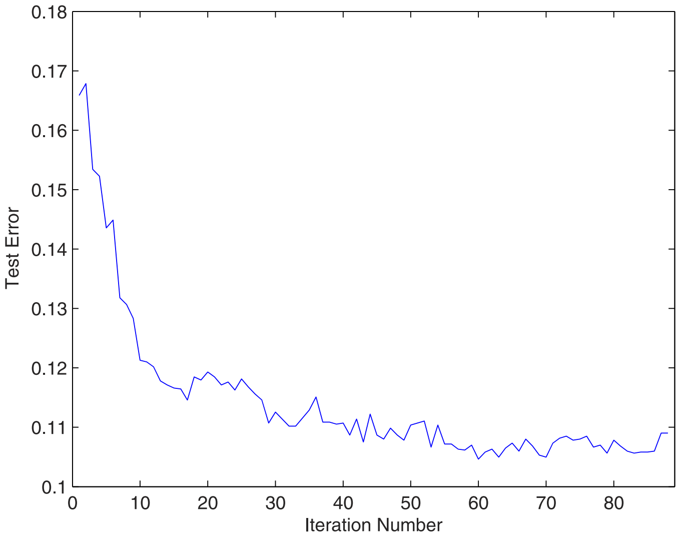

Figure 9. Test error as AdaBoost adds in more weak classifiers (on the MIT dataset). Considering more than 10–20 classifiers does not increase performance substantially.

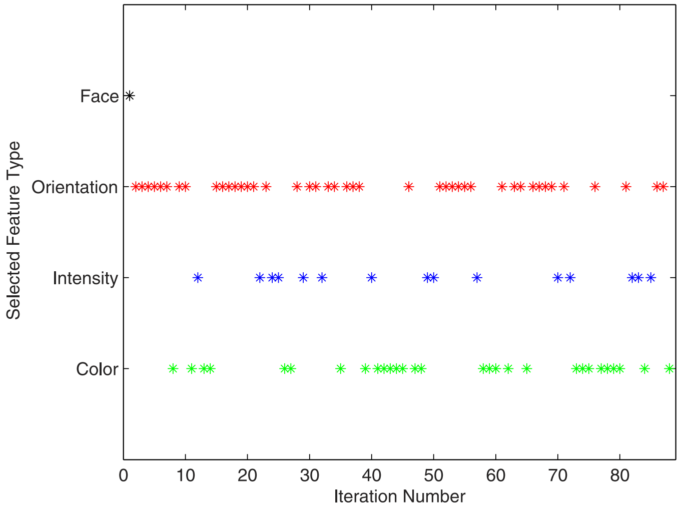

Figure 10. Selected feature type as AdaBoost iterates (on the MIT dataset). Different from the Toronto dataset, the most informative feature is the face channel, followed by orientation, color, and intensity.

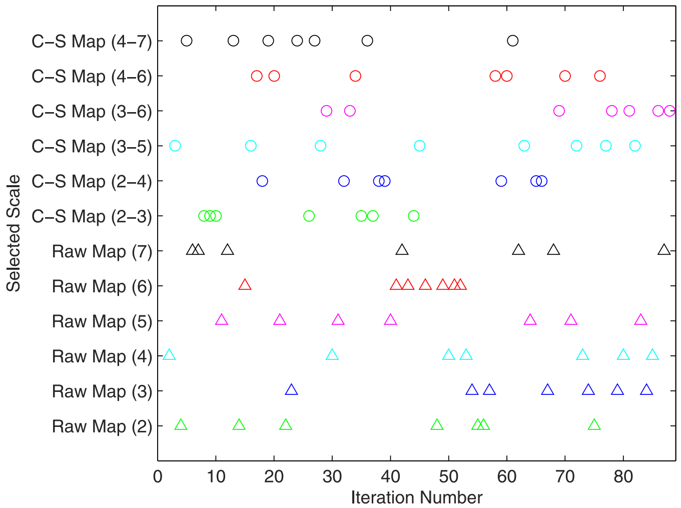

Figure 11. Selected scale as AdaBoost iterates (on the MIT dataset). Selected scales are quite evenly distributed across various scales.

	Centered Gaussian	Linear summation		Linear integration with optimal weights		Nonlinear integration	
		Without CBM	With CBM	Without CBM	With CBM	Without CBM	With CBM
nAUC	0.869	0.776	0.899	0.792	0.910	0.876	0.977
NSS	1.07	0.635	1.19	0.725	1.24	1.17	1.83
EMD	3.56	4.73	3.04	4.53	2.88	3.56	2.24

Table 4. Performance comparisons of the linear and nonlinear integrations with and without spatial channels on the MIT dataset. Scores of the centered Gaussian model (a pure spatial model) are presented as a reference. “CBM” stands for Center Bias Modeling.

The Toronto dataset has been used as a benchmark in several recent publications. Table 3 compares the performance of these models.

MIT dataset

Similar to Judd et al. (2009), we divided the MIT dataset into 903 training images and 100 testing image. Nonlinear integration is learned using $d = 88$ channels.

Since the Toronto dataset included few frontal faces, to provide a more complete view of informative features selected by AdaBoost, on this MIT dataset we again visualized the test error as a function of iteration numbers and the feature selection process, i.e., we split the maps into groups based on their feature types and scales, and illustrated the feature type/scale selected as AdaBoost iterates (Figures 10 and 11). Different from the Toronto dataset, the most informative feature was the face channel, followed by orientation, color, and intensity. This was consistent with findings in Cerf et al. (2009) and Zhao and Koch (2011), which demonstrated that faces strongly and rapidly attract gaze, independent of any task. The discrepancy with the Toronto dataset arose from the fact that the Toronto dataset contained few frontal faces, therefore the learning algorithm could not reliably learn face-related information from the limited data. When the training data contained sufficient frontal faces (such as in the MIT, the FIFA, and the NUS datasets), the face channel was always the most important, compared with color, intensity, and orientation. Particularly, the numbers selected in the first 10 iterations were 1, 0, and 8 for the color, intensity, and orientation channels (the remaining one is the face channel). We performed the same experiments on the FIFA and NUS datasets and obtained similar results.

Quantitative results are reported in Table 4. Though several earlier works had been aware of center bias (Tatler, 2007; Zhang et al., 2008; Judd et al., 2009) for the first time explicitly added a spatial prior and reported improved model performance. The nAUC for our AdaBoost-based models were 0.977 and 0.876, with and without spatial information modeling. In comparison, in Judd et al. (2009), the nAUCs for all features with and without the spatial prior were 0.923 and 0.859, respectively.

NUS dataset

Lastly, we conducted experiments on the NUS dataset, using 500 images for training and the remaining for testing. We again included $d = 88$ channels for learning nonlinear integration.

Despite the considerably richer semantic contents in this dataset, the conclusions from the experiments were consistent with those from the previous three datasets (Table 5): the performance of the saliency model was consistently improved by a nonlinear feature integration and a center bias model.

General discussions and future work

Itti and Koch (1999) pointed out that one difficulty in combining different feature maps into a unique scalar saliency map is that these maps are not directly comparable, with different dynamic ranges and extraction mechanisms. For example, salient objects appearing strongly in one orientation map risk being masked by noise or less salient objects in other maps either with larger dynamic ranges or with a larger number of such

	Centered Gaussian	Linear summation		Linear integration with optimal weights		Nonlinear integration	
		Without CBM	With CBM	Without CBM	With CBM	Without CBM	With CBM
nAUC	0.904	0.793	0.922	0.829	0.938	0.842	0.947
NSS	1.06	0.706	1.15	0.858	1.28	0.891	1.33
EMD	3.20	4.85	3.04	4.55	2.97	4.10	2.68

Table 5. Quantitative comparisons of seven models on the NUS dataset. “CBM” stands for Center Bias Modeling.

maps. Several normalization schemes (Itti & Koch, 1999; Le Meur et al., 2006; Ouerhani, Bur, & Hugli, 2006; Onat, Libertus, & König, 2007) were introduced to alleviate this problem. We exploited the strategy that if signals in certain channels are not sufficiently strong, such channels should not contribute to the final saliency map. A principled framework to automatically find the thresholds and weights for different feature channels is the AdaBoost-based algorithm. Our computational experiments demonstrated the superior performance of the nonlinear compared with linear combination schemes.

There is considerable psychophysical evidence in favor of certain features (e.g., color and luminance) contributing linearly to saliency (Treisman & Gelade, 1980; Nothdurft, 2000; Engmann et al., 2009). This raises the question of the extent to which nonlinear integration is pursued by the human visual system.

Our work focused on early visual saliency, but the framework could incorporate other features as well. For example, text (Cerf et al., 2009), other interesting objects (Einhäuser et al., 2008), and contextual cues (Torralba, Oliva, Castelhano, & Henderson, 2006) could be added into the framework to see the correlation (Baddeley & Tatler, 2006) and relevance of different features or different categories of features in static images. Furthermore, the current studies can be generalized to dynamic scenes to develop a model to predict "spatiotemporal" deployment of gaze.

Though not the focus of this paper, we showed, using three complementary measures, the existence of a strong center bias in all datasets. This was largely due to the experimental setup (Tatler, 2007; Zhang et al., 2008, 2009; Judd et al., 2009) and the feature distributions of the image sets (Reinagel & Zador, 1999; Parkhurst et al., 2002; Tatler, Baddeley, & Gilchrist, 2005; Einhäuser, Spain, & Perona, 2008; Judd et al., 2009). To avoid this bias, several pioneering studies (Hayhoe & Ballard, 2005; Land, 2006; Pelz et al., 2000; Schumann et al., 2008) track eye movements when observers move in the real world, avoiding many of the limitations of viewing photos on monitors. A subject ripe for further investigation is to apply the current framework using data from the unrestrained eye-tracking experiments with full-field-of-view (e.g., while subjects are walking).

Acknowledgments

The authors would like to thank Jonathan Harel for helpful discussions. This research was supported by the NeoVision program at DARPA, by the ONR, by the G. Harold & Leila Y. Mathers Charitable Foundation, and by the WCU (World Class University) program funded by the Ministry of Education, Science and Technology through the National Research Foundation of Korea (R31-10008).

Commercial relationships: none.

Corresponding author: Qi Zhao.
E-mail: qzhao@klab.caltech.edu.
Address: California Institute of Technology, Pasadena, California, USA.

References

Avraham, T., & Lindenbaum, M. (2009). Esaliency (Extended Saliency): Meaningful attention using stochastic image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 99*(1), 693–708.

Baddeley, R., & Tatler, B. (2006). High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision Research, 46*(18), 2824–2833.

Bahill, A., Adler, D., & Stark, L. (1975). Most naturally occurring human saccades have magnitudes of 15 degrees or less. *Investigative Ophthalmology & Visual Science, 14*(6), 468–469, http://www.iovs.org/content/14/6/468. [PubMed] [Article].

Bruce, N., & Tsotsos, J. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, *9*(3):5, 1–24, http://www.journalofvision.org/content/9/3/5, doi:10.1167/9.3.5. [PubMed] [Article].

Carbone, A., & Pirri, F. (2010). Learning saliency. An ICA based model using Bernoulli mixtures. In *Proceedings of Brain Inspired Cognitive Systems.*

Cerf, M., Frady, E., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, *9*(12): 10, 1–15, http://www.journalofvision.org/content/9/12/10, doi:10.1167/9.12.10. [PubMed] [Article].

Chen, X., & Yuille, A. (2004). Detecting and reading text in natural scenes. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 366–373).

Chikkerur, S. S., Serre, T., Tan, C., & Poggio, T. (2010). What and where: A Bayesian inference theory of attention. *Vision Research, 50*(22), 2233–2247.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, *8*(14):18, 1–26, http://www.journalofvision.

org/content/8/14/18, doi:10.1167/8.14.18. [PubMed] [Article].

Engmann, S., ‘t Hart, B. M., Sieren, T., Onat, S., König, P., & Einhäuser, W. (2009). Saliency on a natural scene background: Effects of color and luminance contrast add linearly. *Attention, Perception, & Psychophysics, 71*(6), 1337–1352.

Field, D. (1994). What is the goal of sensory coding? *Neural Computation, 6*:559–601.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, *8*(2):6, 601–617, http://www.journalofvision.org/content/8/2/6, doi:10.1167/8.2.6. [PubMed] [Article].

Freund, Y., & Schapire, R. (1996). Game theory, on-line prediction and boosting. In *Conference on Computational Learning Theory* (pp. 325–332). New York: ACM.

Friedman, J., Hastie, T., Rosset, S., Tibshirani, R., & Zhu, J. (2004). Discussion of three boosting papers. *Annals of Statistics*, *32*(1), 102–107.

Friedman, J., Hastie, T., & Tibshirani, R. (1998). Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, *38*(2), 337–374.

Fuller, J. (1996). Eye position and target amplitude effects on human visual saccadic latencies. *Experimental Brain Research*, *109*(3), 457–466.

Gajewski, D.A., Pearson, A.M., Mack, M.L., Bartlett, F.N., & Henderson, J.M. (2005). Human gaze control in real world search. In L. Paletta, J. Tsotsos, E. Rome, & G. Humphreys (Eds.), *Attention and Performance in Computational Vision* (Vol. 3368, pp. 83–99). New York: Springer-Verlag.

Gao, D., Mahadevan, V., & Vasconcelos, N. (2007). The discriminant center-surround hypothesis for bottom-up saliency. In *Advances in Neural Information Processing Systems* (pp. 497–504). Cambridge, MA: MIT Press.

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. In *Advances in Neural Information Processing Systems* (pp. 545–552).

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194. Cambridge MA: MIT Press.

Hou, X., & Zhang, L. (2008). Dynamic visual attention: searching for coding length increments. In *Advances in Neural Information Processing Systems* (pp. 681–688). Cambridge, MA: MIT Press.

Hu, Y., Xie, X., Ma, W., Chia, L., & Rajan, D. (2004). Salient region detection using weighted feature maps based on the human visual attention model. In *IEEE Pacific-Rim Conference on Multimedia* (pp. 993–1000). Berlin, Heidelberg: Springer-Verlag.

Hyvarinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, *13*(4–5), 411–430.

Itti, L., & Baldi, P. (2006). Bayesian surprise attracts human attention. In *Advances in Neural Information Processing Systems* (pp. 547–554).

Itti, L., & Koch, C. (1999). Comparison of feature combination strategies for saliency-based visual attention systems. In *Proc. SPIE Human Vision and Electronic Imaging* (pp. *3644*:473–82). Bellingham, WA: SPIE.

Itti, L., Koch, C., & Niebur, E. (1998). A model for saliency based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.

Jin, R., Liu, Y., Si, L., Carbonell, J., & Hauptmann, A. G. (2003). A new boosting algorithm using input-dependent regularizer. In *International Conference on Machine Learning*. Palo Alto, CA: AAAI Press.

Jost, T., Ouerhani, N., von Wartburg, R., Müri, R., & Hügli, H. (2005). Assessing the contribution of color in visual attention. *Computer Vision and Image Understanding*, *100*(1–2), 107–123.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *IEEE International Conference on Computer Vision*. Washington, DC: IEEE Computer Society.

Khuwuthyakorn, P., Robles-Kelly, A., & Zhou, J. (2010). Object of interest detection by saliency learning. In *European Conference on Computer Vision* (pp. 636–649). Berlin, Heidelberg: Springer-Verlag.

Kienzle, W., Wichmann, F., Scholkopf, B., & Franz, M. (2006). A nonparametric approach to bottom-up visual saliency. In *Advances in Neural Information Processing Systems* (pp. 689–696). Cambridge, MA: MIT Press.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, *4*(4), 219–227.

Koene, A., & Zhaoping, L. (2007). Feature-specific interactions in salience from combined feature contrasts: Evidence for a bottom-up saliency map in V1. *Journal of Vision*, *7*(7):6, 1–14, http://www.journalofvision.org/content/7/7/6, doi:10.1167/7.7.6. [PubMed] [Article].

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye movements: An investigation with higher-order statistics. *Spatial Vision*, *13*(2–3), 201–214.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning* (pp. 282–289). San Francisco: Morgan Kaufmann Publishers, Inc.

Land, M. (2006). Eye movements and the control of actions in everyday life. *Progress in Retinal and Eye Research*, *25*(3), 296–324.

Lang, P., Bradley, M., & Cuthbert, B. (2008). (IAPS): Affective ratings of pictures and instruction manual. In *Technical report, University of Florida*.

Le Meur, O., Le Callet, P., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model the bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(5), 802–817.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, *6*(1), 9–16.

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., et al. (2011). Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(2), 353–367.

Liu, T., Zheng, N., Ding, W., & Yuan, Z. (2008). Video attention: Learning to detect a salient object sequence. In *IEEE Conference on Pattern Recognition* (pp. 1–4) Washington, DC: IEEE Computer Society.

Masciocchi, C., Mihalas, S., Parkhurst, D., & Niebur, E. (2009). Everyone knows what is interesting: Salient locations which should be fixated. *Journal of Vision*, *9*(11):25, 1–22, http://www.journalofvision.org/content/9/11/25, doi:10.1167/9.11.25. [PubMed] [Article].

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, *45*(2), 205–231.

Nothdurft, H. (2000). Salience from feature contrast: Additivity across dimensions. *Vision Research*, *40*: 1183–1201.

Oliva, A., Torralba, A., Castelhano, M., & Henderson, J. (2003). Top-down control of visual attention in object detection. In *IEEE International Conference on Image Processing* (pp. *I*:253–256). Washington, DC: IEEE Computer Society.

Onat, S., Libertus, K., & König, P. (2007). Integrating audiovisual information for the control of overt attention. *Journal of Vision*, *7*(10):11, 1–16, http://www.journalofvision.org/content/7/10/11, doi:10.1167/7.10.11. [PubMed] [Article].

Ouerhani, N., Bur, A., & Hugli, H. (2006). Linear vs. nonlinear feature combination for saliency computation: A comparison with human vision. In *Lecture Notes in Computer Science* (pp. *4174*:314–323). Berlin, Heidelberg: Springer-Verlag.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107–123.

Pelz, J. B., & Canosa, R. (2001). Oculomotor behavior and perceptual strategies in complex tasks. *Vision Research*, *41*(25–26), 3587–3596.

Pelz, J. B., Canosa, R., Kucharczyk, D., Babcock, J., Silver, A., & Konno, D. (2000). Portable eye tracking: A study of natural eye movements. In *Proceedings of Human Vision and Electronic Imaging V* (pp. *3959*:566–582). Bellingham, WA: SPIE.

Peters, R., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, *45*(18), 2397–2416.

Privitera, C., & Stark, L. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(9), 970–982.

Raj, R., Geisler, W., Frazor, R., & Bovik, A. (2005). Contrast statistics for foveated visual systems: Fixation selection by minimizing contrast entropy. *Journal of the Optical Society of America A*, *22*(10), 2039–2049.

Reinagel, P., & Zador, A. (1999). Natural scene statistics at the centre of gaze. *Network*, *10*(4), 341–350.

Rubner, Y., Tomasi, C., & Guibas, L. (2000). The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision*, *40*(2), 99–121.

Schapire, R., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, *37*(3), 297–336.

Schumann, F., Einhäuser, W., Vockeroth, J., Bartl, K., Schneider, E., & König, P. (2008). Salient features in gaze- aligned recordings of human visual input during free exploration of natural environments. *Journal of Vision*, *8*(14):12, 1–17, http://www.journalofvision.org/content/8/14/12, doi:10.1167/8.14.12. [PubMed] [Article].

Seo, H., & Milanfar, P. (2009). Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, *9*(12):15, 1–27, http://www.journalofvision.org/content/9/12/15, doi:10.1167/9.12.15. [PubMed] [Article].

Subramanian, R., Katti, H., Sebe, N., Kankanhalli, M., & Chua, T. S. (2010). An eye fixation database for saliency detection in images. In *European Conference on Computer Vision* (pp. *6314*:30–43). Berlin, Heidelberg: Springer-Verlag.

Tatler, B. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14):4, 1–17, http://www.journalofvision.org/content/7/14/4, doi:10.1167/7.14.4. [PubMed] [Article].

Tatler, B., Baddeley, R., & Gilchrist, I. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, *45*(5), 643–659.

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*:766–786.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Vezhnevets, A., & Vezhnevets, V. (2005). Modest AdaBoost—Teaching AdaBoost to generalize better. *In Graphicon*.

Vincent, B. T., Baddeley, R. J., Troscianko, T., & Gilchrist, I. D. (2009). Optimal feature integration in visual search. *Journal of Vision*, *9*(5):15, 1–11, http://www.journalofvision.org/content/9/5/15, doi:10.1167/9.5.15. [PubMed] [Article].

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. *I*:511–518). Washington, DC: IEEE Computer Society.

Vitu, F., Kapoula, Z., Lancelin, D., & Lavigne, F. (2004). Eye movements in reading isolated words: Evidence for strong biases towards the center of the screen. *Vision Research*, *44*(3), 321–338.

Walther, D., Serre, T., Poggio, T., & Koch, C. (2005). Modeling feature sharing between object detection and topdown attention. *Journal of Vision*, *5*(8): 1041–1041, http://www.journalofvision.org/content/5/8/1041, doi:10.1167/5.8.1041. [Abstract].

Wang, W., Wang, Y., Huang, Q., & Gao, W. (2010). Measuring visual saliency by site entropy rate. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2368–2375). Washington, DC: IEEE Computer Society.

Zambarbieri, D., Beltrami, G., & Versino, M. (1995). Saccade latency toward auditory targets depends on the relative position of the sound source with respect to the eyes. *Vision Research*, *35*(23–24), 3305–3312.

Zhang, L., Tong, M., & Cottrell, G. (2009). SUNDay: Saliency using natural statistics for dynamic analysis of scenes. In *Proceedings of the 31st Annual Cognitive Science Conference* (pp. 2944–2949).

Zhang, L., Tong, M., Marks, T., Shan, H., & Cottrell, G. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, *8*(7):32, 1–20, http://www.journalofvision.org/content/8/7/32, doi:10.1167/8.7.32. [PubMed] [Article].

Zhao, Q., & Koch, C. (2011). Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, *11*(3):9, 1–15, http://www.journalofvision.org/content/11/3/9, doi:10.1167/11.3.9. [PubMed] [Article].