Boosting Bottom-up and Top-down Visual Features for Saliency Estimation

Ali Borji
Department of Computer Science
University of Southern California, Los Angeles, CA 90089
http://ilab.usc.edu/~borji

Abstract

Despite significant recent progress, the best available visual saliency models still lag behind human performance in predicting eye fixations in free-viewing of natural scenes. Majority of models are based on low-level visual features and the importance of top-down factors has not yet been fully explored or modeled. Here, we combine low-level features such as orientation, color, intensity, saliency maps of previous best bottom-up models with top-down cognitive visual features (e.g., faces, humans, cars, etc.) and learn a direct mapping from those features to eye fixations using Regression, SVM, and AdaBoost classifiers. By extensive experimenting over three benchmark eye-tracking datasets using three popular evaluation scores, we show that our boosting model outperforms 27 state-of-the-art models and is so far the closest model to the accuracy of human model for fixation prediction. Furthermore, our model successfully detects the most salient object in a scene without sophisticated image processings such as region segmentation.

1. Introduction

Visual attention is a cognitive process that helps humans and primates rapidly select the highly relevant information from a scene. This information is then processed finer by high-level visual processes such as scene understanding and object recognition. The notion of relevance is determined by two factors. The first one, often referred as bottom-up visual saliency, is a task-independent component based on only low-level and image-based outliers and conspicuities. The second component is based on volitionally-controlled mechanisms that determine the importance of scene regions in daily-life tasks such as driving.

The process of visual attention has been the subject of numerous studies in psychology, neurosciences, and computer vision. Correspondingly, several computational models of attention have been proposed in machine learning, computer vision, and robotics. Several applications have also been proposed and have further raised interest in this field including: image thumb-nailing [7], automatic collage creation [5], foveated image/video compression [6][9], non-photorealistic rendering [8], and advertisement design [10].

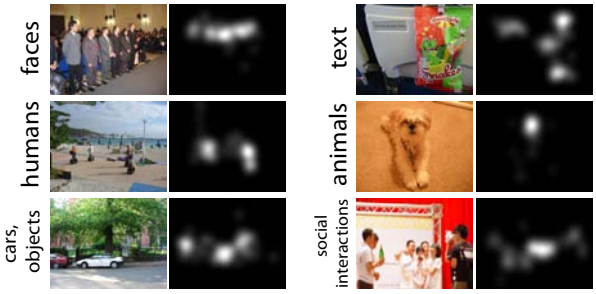

Figure 1. Human fixation map for sample images from the MIT [1] dataset. Top-down concepts including people, social interactions, animals, cars, signs, faces, and text attract human attention.

Models of bottom-up saliency have often been evaluated against predicting human fixations in free-viewing task. Today, many saliency models based on variety of techniques with compelling performance exist and still each year new models are introduced. Yet, there is a large gap between models and the human Inter-Observer (IO) model for predicting eye fixations. The IO "model" outputs, for a given stimulus, a map built by integrating eye fixations from other subjects than the one under test while they watched that stimulus. This model is expected to provide an upper bound on prediction accuracy of models to the extent that, different humans may be the best predictors of each other. The mentioned gap between models and human is largely due to the role of top-down factors (See Fig. 1).

It is believed that at early stages of free viewing (first few hundred milliseconds), mainly image-based conspicuities guide attention and later on, high-level factors (e.g., actions and events) direct eye movements [53][39]. These high-level factors may not necessarily translate to bottom-up saliency (e.g., based on color, intensity, or orientation) and should be taken into account separately. For instance, a human's head may not stand out from the rest of the scene but may attract attention. Thus, combining high-level concepts and low-level features seems inevitable to scale up current models and reach the human performance.

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

Some top-down factors in free-viewing are already known although active investigation still continues to discover more semantic factors. For instance, Einhäuser *et al.* [11] proposed that objects are better predictors of fixations than bottom-up saliency. Cerf *et al.* [14] showed that faces and text attract human gaze. Elazary and Itti [12] showed that interesting objects (annotations from LabelMe dataset [46]) are more salient. Subramanian *et al.* [13], by recording eye fixations over a large affective image dataset, observed that fixations are directed toward emotional and action stimuli and duration of fixations are longer on such stimuli. Similarly, Judd *et al.* [1], by plotting image regions at the top salient locations of the human saliency map (made of fixations), observed that humans, faces, cars, text, and animals attract human gaze probably because they convey more information in a scene. Alongside, some personal characteristics such as experience, age, and culture change the way humans look at images [54].

Inspired by [1], we propose three contributions to saliency learning. First, we combine the best of the two worlds: bottom-up and top-down factors. By comparing 29 saliency models, we consolidate features that the best bottom-up models have found predictive of human fixations with top-down factors such as faces, humans, cars, etc. We train several linear and non-linear classifiers from these features to fixations. Second, we emphasize more on internal parts of attention grabbing objects (e.g., top parts of humans) for more accurate saliency detection. Through extensive experiments, we show that our combined approach, outperforms previous saliency learning methods ([1][48]) as well as other state-of-the-art works over 3 datasets using 3 evaluation scores. Third, we show that our model is able to detect the most salient object in a scene close to performances of the mainstream salient region detection schemes.

Related work. Saliency models in general can be categorized as cognitive (biological) or computational (mathematical) while some happen in between. Several models are based on the bottom-up saliency model by Itti *et al.* [4]. This model is the first implementation of the Koch and Ullman's computational architecture [15] based on the Feature Integration Theory [16]. In this theory, an image is decomposed into low-level attributes such as color, intensity, and orientation across several spatial scales which are then linearly or non-linearly normalized and combined to form a master saliency map. An important element of this theory is the idea of center-surround that defines saliency as distinctiveness of an image region to its immediate surroundings. This model also proposes a suitable framework for adaptation of visual search theories and object detection models (e.g., [18]). Based on the idea of decorrelation of neural responses, Diaz *et al.* [29] proposed an effective model of saliency known as Adaptive Whitening Saliency (AWS). Le Meur *et al.* [33], Marat *et al.* [36], and Kootstra *et al.* [17] are other models guided by cognitive findings.

Another class of models are based on probabilistic formulation. Torralba [32] proposed a Bayesian framework for visual search which is also applicable for saliency detection. Bottom-up saliency is derived from their formulation as: $\frac{1}{p(F|G)}$ where F represents a global feature that summarizes the probability density of presence of the target object in the scene, based on analysis of the scene gist (G). Similarly, Zhang *et al.* [38] proposed SUN (Saliency Using Natural statistics) model in which bottom-up saliency emerges naturally as the self-information of visual features. Mancas [25] proposed local (small local neighborhood) and global (entire scene) rarities as saliency measures. Itti and Baldi [22] defined surprising stimuli as those that significantly change beliefs of an observer by computing the KL distance between posterior and prior beliefs. Graph based Visual Saliency (GBVS) [20] and E-saliency [26] are two other methods based on Bayesian and graphical models.

Decision theoretic interpretation of saliency states that attention is driven optimality with respect to the end task. Gao and Vasconcelos [35] argued that for recognition, salient features are those that best distinguish a class of interest from all other classes. Given some set of features $X = \{X_1, \cdots, X_d\}$, a location l and a class label Y with $Y_l = 0$ corresponding to samples drawn from the surround ($Y_l = 1$ for center region at l), saliency is then a measure of mutual information (usually KL divergence), computed as $I(X, Y) = \sum_{i=1}^{d} I(X_i, Y)$.

Frequency domain models are another class. Hou and Zhang [23] proposed Spectral Residual Model (SRM) by relating spectral residual features in spectral domain to the spatial domain. In [27], Phase spectrum of Quaternion Fourier Transform (PQFT) is utilized for saliency computation which is applicable for both static and dynamic stimuli.

Our proposed approaches are related to those models that learn mappings from image features to eye fixations using machine learning techniques. Kienzle *et al.* [2], Judd *et al.* [1], and Peters and Itti [47], used image patches, a vector of several features at each pixel, and scene gist, respectively for learning saliency. Zhao and Koch [48][49] learned optimal weights for saliency channel combination separately for each eye-tracking dataset. While they show tunning weights for each dataset results in high accuracies, learned weights sometimes do not agree over datasets. It is also unclear how this approach generalizes to unseen images. Here, we exploit more informative features and assess the capability of stronger classifiers for eye fixation prediction.

In addition to above models, some other models address salient region detection (e.g., Achanta *et al.* [44] and Cheng *et al.* [41]). The main goal of these models is to find and segment the most salient object or region in a scene. In principle, saliency detection and estimation (for fixation prediction) techniques are applicable interchangeably.

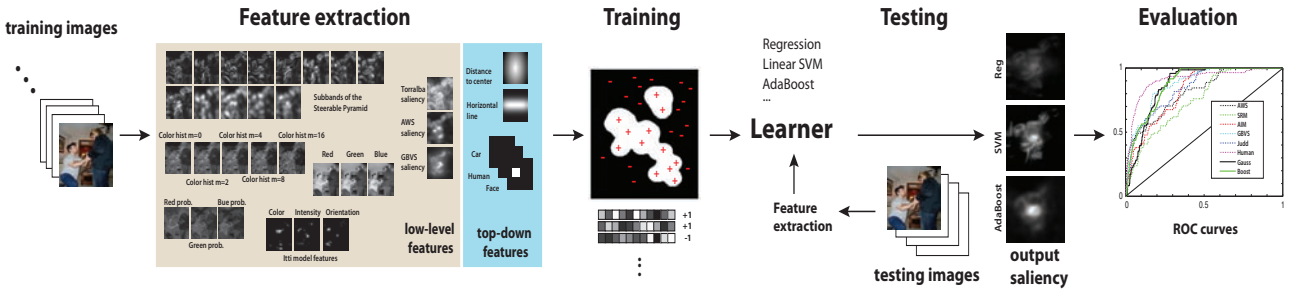

Figure 2. Illustration of our learning approach. A set of low-level and high-level visual features are extracted from some training images. Feature vectors corresponding to the top 20% (bottom 40%) of the human heatmap are assigned +1(-1) labels. Then a classifier is trained from these features and is used for predicting fixations on a test image. Finally, model performance is evaluated using an evaluation metric (see text).

2. Learning a model of visual saliency

In contrast to manually designed measures of saliency, we follow a learning approach by training classifiers directly from human eye tracking data. The basic idea is that a weighted combination of features, where weights are learned from a large repository of eye movements over natural images, can enhance saliency detection compared with unadjusted combination of feature maps. A learning approach has also the benefit of being easily applicable to visual search by enhancing feature weights of a target object.

In the following, we propose a Naive Bayesian formulation for saliency estimation. Let s be a binary variable denoting saliency of an image pixel at location $\mathbf{x} = (x, y)$ with feature vector $\mathbf{f}$, where "s equal 1" indicates that this pixel is salient (i.e., it can attract human eye) and zero, otherwise. The probability of pixel $\mathbf{x}$ being salient can be written as:

$$p(s|\mathbf{f},\mathbf{x}) = \frac{p(\mathbf{f}|s)p(s|\mathbf{x})}{p(\mathbf{f}|\mathbf{x})} = \frac{p(\mathbf{f}|s)p(s|\mathbf{x})}{p(\mathbf{f})} \\ = \frac{p(s|\mathbf{f})p(s|\mathbf{x})}{p(s)} \propto p(s|\mathbf{f})p(s|\mathbf{x}) \quad (1)$$

Above formula is based on the assumption that features can appear in all spatial locations (i.e., $\mathbf{x}$ and $\mathbf{f}$ are independent from each other thus $p(\mathbf{f}|\mathbf{x}) = p(\mathbf{f})$). We further assume that prior probabilities over s (i.e., a location being salient or not) are equal. The first term on the right side of Eq. 1 measures saliency due to features at an image pixel, while the second term measures saliency only based on the spatial location of a pixel. We learn $p(s|\mathbf{f})$ in a discriminative approach using classifiers from annotated data (fixated locations). We estimate $p(s|\mathbf{x})$ by:

$$p(s|\mathbf{x}) \propto 1 - d(\mathbf{x}, \mathbf{x}_o) \quad (2)$$

where $d(\mathbf{x}, \mathbf{x}_o)$ is the normalized distance of the pixel $\mathbf{x}$ from the center pixel $\mathbf{x}_o$. This resembles a Gaussian pdf that has been shown to explain fixations in free-viewing well [39].

2.1. Visual features

Low-level (bottom-up) features. Traditionally, intensity, orientation, and color have been used for saliency derivation over static images. Over dynamic scenes (videos), flicker and motion features have been added [55]. Several other low-level features have also been employed (e.g., size, depth, and optical flow) [56]. Here, we first resize each image to 200 × 200 pixels and then extract a set of features for every pixel. Similar to [1], we use low-level features as they have already been shown to correlate with visual attention and have underlying biological plausibility [16][15]. Low-level features are listed below:

- 13 local energy of the steerable pyramid filters in 4 orientations and 3 scales.
- 3 intensity, orientation, and color (Red/Green and Blue/Yellow) contrast channels as calculated by Itti and Koch's saliency method [4].
- 3 values of the red, green, and blue color channels as well as 3 features corresponding to probabilities of each of these color channels.
- 5 probabilities of above color channels as computed from 3D color histograms of the image filtered with a median filter at 6 different scales.
- 3 saliency maps of Torralba [32], AWS [29], and GBVS [20] bottom-up saliency models.

This results in 30 low-level features. Note that, center-surround operation is directly applied over maps of some features (e.g., Itti feature maps). Although in practice, it is possible to use any bottom-up model as a feature, here we utilize Torralba [32], AWS [29], and GBVS [20] models because these models have high fixation prediction power[1], employ radically different saliency mechanisms, are fast, and can be calculated from the other low-level features. Experimenting with other models did not help our results but we don't completely rule out such possibility. AWS model uses the Lab color space and decorrelates the feature maps while GBVS employs a measure of dissimilarity over image pixels to calculate saliency over a graph. Here, we exploit linear features. Our framework allows addition of other

[1] GBVS model performs higher than other models over AUC, CC, and NSS scores later shown in Fig. 5, and Fig. **??**. AWS model has the best shuffled AUC score which is a variant of AUC designed to tackle center bias by emphasizing more off-center fixations [38]. Please see supplement.

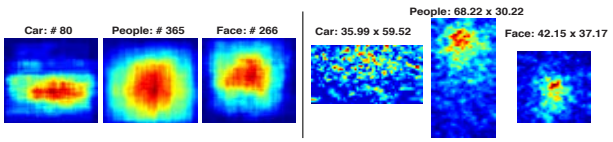

Figure 3. Left: mean position of object locations, Right: mean fixation locations over objects (Data from MIT dataset [1]).

non-linear features such as corners, to account for outliers due to changes in texture (famous egg in the nest or birthday candle images [38]). Extracted features for a sample image are illustrated in Fig. 2.

High-level (top-down) features. High-level features such as faces and text [14], people and cars [1], symmetry [17], and signs have been suggested to direct attention. These are prior knowledge that are learned through a human's life time. One challenge is detecting affective (emotional) features and semantic (high-level knowledge) scene properties, such as causality and action-influence, which are believed to be important in guiding attention. These factors influence both eye fixation locations and their durations [13]. High-level features that we include into our feature set are as follows:

- The horizontal line due to tendency of photographers to frame images and objects horizontally.
- Person and car detectors implemented by Felzenszwalb's Deformable Part Model (DPM) [50].
- Face detector using the Viola and Jone's code [51].

From annotated data, we noticed that certain regions in objects attract more attention, for example top-part of humans (head area) and face components (eyes, nose, and mouth)(see Fig. 3). To enhance saliency of those regions, we multiplied an object's detected region with the learned mean saliency map for that object (learned from training data). In general, performance of saliency detection by adding object detectors is highly dependent on the false positive rates of the employed detectors. For instance, if a face detector generates many false alarms for an image with no face then it dramatically reduces the evaluation scores. Unfortunately, despite high importance of text features in guiding gaze, to date, there is no reliable approach that can detect text in natural scenes.

Another important feature is the center prior based on the finding that majority of fixations happen near the center of the image (i.e., center-bias [39]). For fair comparison of classifiers with baseline approaches (AWS and GBVS models), here we treat the center feature separately. According to Eq. 2, we multiply saliency map of each model with $p(s|\mathbf{x})$, the distance of each pixel from the center.

Eventually, all features are augmented in a 34D (30 bottom-up + 4 top-down) vector (excluding center) and are fed to classifiers explained in the next section. Each feature map is resized into a 200×200 map which is then linearized into a 1×4000 vector (similarly for class labels).

2.2. Classifiers

We investigate the ability of linear and non-linear classifiers for fixation prediction. Linear classifiers are usually fast to compute via matrix operations and learned weights are easier to interpret. On the other hand, non-linear models are usually slower but more powerful. Fig. 2 shows a schematic illustration of our saliency learning method. We compile a large training set by sampling images at fixations. Each sample contains features at one point along with a $+1/-1$ label. Positive samples are taken from the top p percent salient pixels of the human fixation map (smoothed by convolving with a small Gaussian filter) and negative samples are taken from the bottom q percent. We chose samples from the top 20% and bottom 40% in order to have samples that were strongly positive and strongly negative. Training feature vectors were normalized to have zero mean and unit standard deviation and the same parameters were used to normalize test data. To evaluate our models, we followed a cross-validation approach. The whole dataset was divided into K parts, each with M images. Each time we trained the model from $K-1$ parts and tested it over the remaining part. Results are then averaged over all partitions.

Regression. Assuming a linear relationship between feature vector $\mathbf{f}$ and saliency s, we solve the equation $F \times W = S$ where F and S are matrices of $\mathbf{f}$ and s over training data. The solution is: $W = F^{+} \times S$, where F^{+} is the least-square pseudo-inverse of matrix F through SVD decomposition. To avoid numerical instability, those eigenvectors whose eigenvalues were less than half of the largest eigenvalue were discarded during computation of the pseudo-inverse. For a test image, features are first extracted and then the learned mapping was applied to generate a vector which is then resized to a 200×200 saliency map.

SVM. Using liblinear support vector machine[2]., a publicly available Matlab version of SVM, we also trained a SVM classifier. We adopted linear kernels as they are faster and perform as well as non-linear polynomial and RBF kernels for fixation prediction [1]. In testing, similar to Regression, instead of predicted labels (i.e., $+1/-1$), we use the value of $W^T\mathbf{f} + b$ where W and b are learned parameters.

Boosting. To investigate a non-linear mapping of features to saliency, we used AdaBoost algorithm [52], which has many appealing theoretical properties with applications in scene classification and object recognition. Given N labeled training examples (u_i, v_i) with $v_i \epsilon \{-1, +1\}$ and $u_i \epsilon U$, AdaBoost combines a number of weak classifiers h_t to learn a strong classifier $H(u) = sign(f(u)); f(u) = \sum_{t=1}^{T} \alpha_t h_t(u)$ where α_t is the weight of the t-th classifier. Here, we set the number of weak classifiers, T, to 10 which resulted in high accuracy and reasonable speed. Instead of the class label, we consider the real value of $H(u)$ to create

[2]Libsvm: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

a saliency map (i.e., $f(u)$). Final map is first smoothed by convolving with a small Gaussian kernel and is then passed through an exponential function for better illustration. We used the publicly available software for Gentle AdaBoost[3].

3. Experimental setup

A thorough evaluation of classifiers and features is presented in this section. Here, along with evaluation of our model, we also compare several models for future references. We were able to run 27 saliency models[4]. In addition, we also implemented two other simple yet powerful models: Gaussian Blob (Gauss) and Human inter-observer model. Gaussian blob is simply a 2D Gaussian shape drawn at the center of the image; it is expected to predict human gaze well if such gaze is strongly clustered around the image center. The inter-observer model outputs, for a given stimulus, a map built by integrating eye fixations from other subjects than the one under test while they watched that stimulus. Models maps were resized to the size of the original images onto which eye movements have been recorded.

3.1. Eye movement datasets

Since available eye movement datasets have different statistics, types of stimuli, and numbers of subjects, here we exploit 3 benchmark datasets for fair model comparison.

The first dataset, *MIT* [1][5], contains 1003 images collected from Flicker and LabelMe [46] datasets. The longest dimension of images is 1024 with other dimension ranging from 405 to 1024. There are 779 landscape and 228 portrait images. Fifteen human subjects viewed images. Images are shown for 3 seconds with 1 second gray screen between each two. The second dataset, *Toronto* [21][6], is the most widely-used dataset for saliency model evaluation. It contains 120 color images from indoor and outdoor scenes. Images are presented at random for 4 seconds with 2 seconds gray mask in between, to 20 subjects. *NUSEF*[7] is a recently introduced dataset with 758 images containing affective scenes/objects such as expressive faces, nudes, unpleasant concepts, and semantic concepts (action/cause). In total, 75 subjects freely viewed part of the image set for 5 second each (on average 25 subjects per image).

3.2. Evaluation metrics

Since there is no consensus over a unique scores for saliency model evaluation, we report results over three. A model that performs well should be good over all scores.

- *Area Under the ROC Curve (AUC)*: Using this score, the model's saliency map is treated as a binary classifier on every pixel in the image; pixels with larger saliency values than a threshold are classified as fixated while the rest of pixels are classified as non-fixated [21]. Human fixations are used as ground truth. By varying the threshold, the ROC curve is drawn as the *false positive rate* vs. *true positive rate*, and the area under this curve indicates how well the saliency map predicts actual human eye fixations.

- *Normalized Scanpath Saliency (NSS)* [47]: NSS is the response value at the human eye position, (x_h, y_h), in a model's predicted gaze map that has been normalized to have zero mean and unit standard deviation: $NSS = \frac{1}{\sigma_S}(S(x_h, y_h) - \mu_S)$. For an image, NSS is computed once for each saccade, and subsequently the mean and standard error are computed across the set of NSS scores.

- *Linear Correlation Coefficient (CC)*: The linear correlation coefficient measures the strength of a linear relationship between human fixation map (h) and saliency map (s) as: $CC(h,s) = \frac{cov(h,s)}{\sigma_h \sigma_s}$ where μ and σ are the mean and the standard deviation of a map, respectively.

	MIT [1]		Toronto [21]		NUSEF [13]	
Model	- C	+ C	- C	+ C	- C	+ C
Reg	0.775	0.820	0.787	0.826	0.754	0.784
SVM	0.773	0.835	0.789	0.837	0.758	0.806
Boost	0.806	0.836	0.805	0.838	0.781	0.815
AWS	0.75	0.824	0.770	0.832	0.731	0.790
GBVS	0.815	0.836	0.827	0.834	0.801	0.816
Reg (All)	0.805	0.832	0.815	0.834	0.779	0.796
SVM (All)	0.820	0.848	0.829	0.839	0.797	0.812
Boost (All)	**0.835**	**0.852**	**0.834**	**0.847**	**0.812**	**0.821**
Gaussian	**0.810**		**0.798**		**0.795**	
Human	**0.912**		**0.887**		**0.868**	

Table 1. AUC scores of models using 32 features over three datasets (average over 10 runs for MIT dataset). - C : without center prior; +C : with center. All: means 34 basic features are used (with AWS and GBVS features). Maximum among models in each column is shown in bold and black. Boosting outperforms linear SVM, regression classifiers as well as AWS and GBVS models when all features are added.

4. Model comparison and results

We validate our model by applying it to two problems: 1) eye movement prediction and 2) segmentation of the most salient object/region in a scene.

4.1. Fixation prediction

We trained and tested classifiers over the MIT dataset following cross-validation in Sec. 2.2 ($K = 10$, $M = 100$ except the last one containing 103). A trained model over all images from the MIT dataset was then applied to other datasets. Table 1 shows AUC scores of models.

Using 32 features (except AWS and GBVS features), boosting outperformed the other two classifiers. Results were enhanced with multiplying center-bias feature. Scores were higher than AWS model and slightly below GBVS. When adding center-bias, however, the difference between boosting and GBVS is smaller, sometimes boosting wins.

[3] http://graphics.cs.msu.ru/en/science/research/machinelearning/adaboosttoolbox

[4] Some models were available online. Some authors, either sent us codes/executables or saliency maps. Compared models include: Variance and Entropy, Itti *et al.* [4] (CIO channels), Surprise [22], VOCUS [19], Torralba [32], AIM [21], Saliency Toolbox (STB) [37], Le Meur *et al.* [33], GBVS [20], SRM [23], Marat *et al.* [36], Local and Global Rarity models [25], ICL [24], Kootstra *et al.* [17], SUN [38], PQFT [27], Yin Li *et al.* [30], SDSR [34], Judd *et al.* [1], Bian *et al.* [28], E-Saliency [26], Yan *et al.* [31], Li *et al.* [3], Tavakoli [40], and AWS [29].

[5] http://people.csail.mit.edu/tjudd/WherePeopleLook/index.html

[6] www-sop.inria.fr/members/Neil.Bruce

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

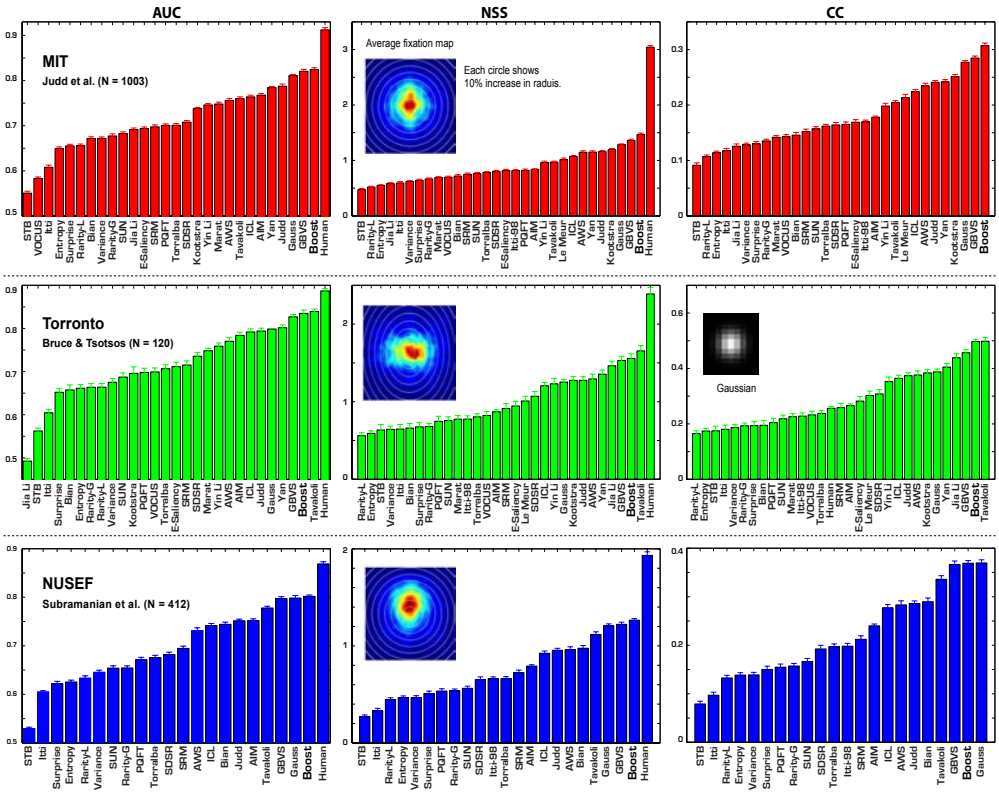

Figure 4. **Model Comparison** of 29 saliency models over 3 datasets and 3 scores. Only 412 images from NUSEF dataset were used due to copyright concerns. Note that all models are compared without adding center-bias. Error bars indicate standard error of the mean: $\frac{\sigma}{\sqrt{N}}$ where σ is the standard devision. Judd model is the result of SVM with no center feature (2nd row in Table 1). Over the MIT dataset, our boosting model scores the best using 3 evaluation scores followed by GBVS and Gaussian models. Over the Toronto dataset, Tavakoli *et al.* [40], achieves the highest scores and is followed by our model and GBVS. Over the NUSEF dataset, boosting outperforms all models over three scores (after Gauss using CC).

The interpretation behind this is that GBVS is a good model but has intrinsic center-bias in a way that multiplying center-bias does not change its performance much (compared when multiplying center-bias with other models).

When we added AWS and GBVS as bottom-up features to our feature set, performances of our boosting and SVM classifiers, outperformed GBVS consistently in all cases with and without center-bias (except SVM over NUSEF).

Results over NSS and CC scores without multiplying center-bias are shown in Fig. **??**. Boosting (no center but with AWS and GBVS as features) wins over GBVS and AWS in almost all cases. Overall, this figure shows that while many models score less than Gaussian model, our boosting model stands on top of Gaussian and is the best in majority of cases over 3 datasets and 3 scores. There is larger gap between models and IO model over NUSEF and MIT datasets because there are more stimuli with conceptual and top-down factors in these datasets. Tavakoli *et al.* [40] performed the best over Toronto dataset where lack of much top-down factors on images of this dataset, ranks boosting the second. Removing emphasize on internal parts of objects reduced AUC of Boosting (32D case in Table 1) over MIT dataset from 0.806 to 0.792.

Fig. 5 shows ROC curves of models in Table. 1 and the learned weight vector W of Regression and SVM classifiers over the MIT dataset. The most important features include:

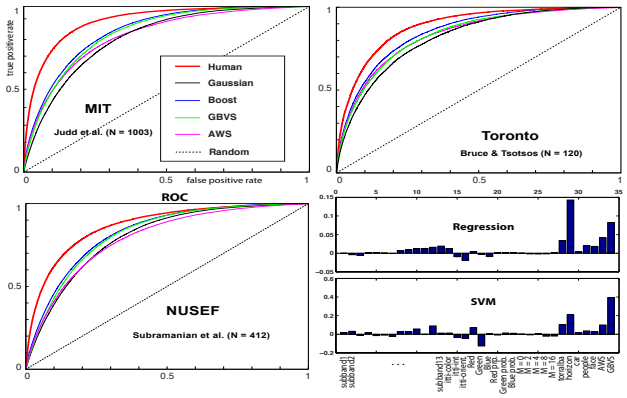

Figure 5. Comparison of our models (All features) with AWS and GBVS models. ROC curves are for the "+center" feature case.

horizontal line, GBVS, AWS, and Torralba saliency maps, as well as face and human detectors.

For our models, it takes 21.5 seconds to extract all necessary features ([0.25 0.54 10.2 0.19 0.31 4.4 3.6 2] seconds in order for Subband, Itti, Color, Torralba, Horizon, Objects, AWS, and GBVS maps) and 0.4 seconds to calculate saliency for a 200×200 image. A personal computer running Linux Ubuntu with 5.8 GB RAM and 12 Core 3.2 GHz Intel i7 CPU was used. The most expensive channel is color but it can be removed to make our model faster as it not a very important channel (See Fig. 5). Sample saliency maps for our classifiers and 5 best models are shown in Fig. 6.

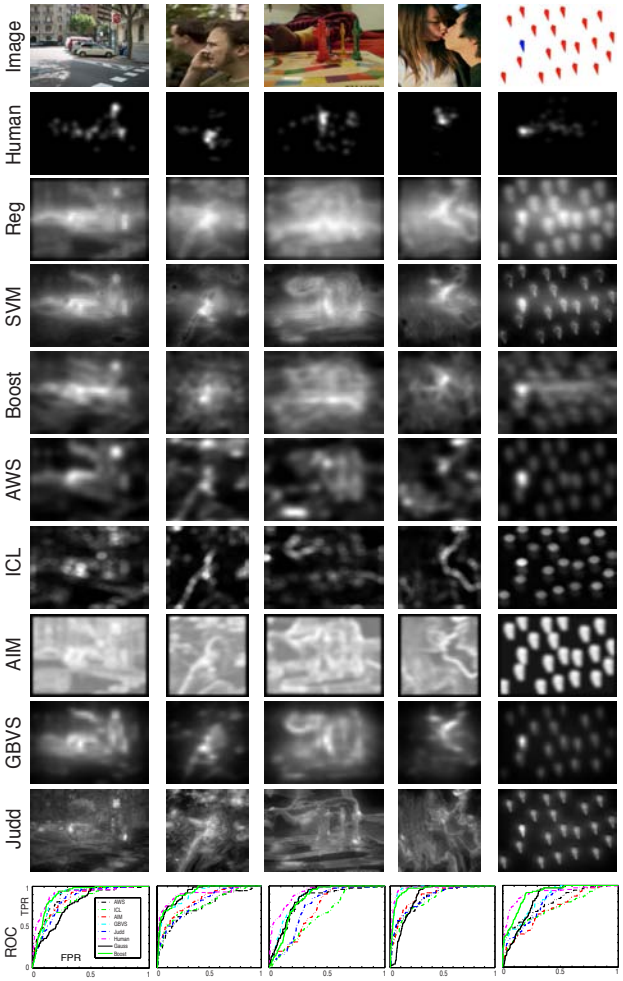

Figure 6. Visual comparison of saliency maps for sample images from MIT dataset [1] along with predictions of several models using AUC.

4.2. Application to salient object detection

Almost all salient region detection approaches utilize a saliency operator, where from there they start to segment the most salient object. Here, we show that our approach could provide a good such starting point. Promising results here is another evidence toward effectiveness of our model. We evaluated the results of our approach on the publicly available dataset, known as ASD, provided by Achanta *et al.* [44] which contains 1000 images of manually annotated objects. We compared our Boosting approach with 11 state-of-the-art salient object detection methods: IT [4], SR [23], GB [20] (GBVS), AC [42], FT [44], CA [43], MZ [45], HC, LC and RC [41], and G [39] which is a Gaussian at the center of the image. All bottom-up and top-down features including center prior were used.

We calculate precision and recall curves by binarizing the saliency map using every possible fixed threshold, similar to the fixed thresholding experiment in [44]. As seen from the comparison (Fig. 7), our saliency model outperforms several models while competing with the state-of-the-art models tailored for this task [44][41]. As there are many objects at the center, a trivial Gaussian model works better than several models.

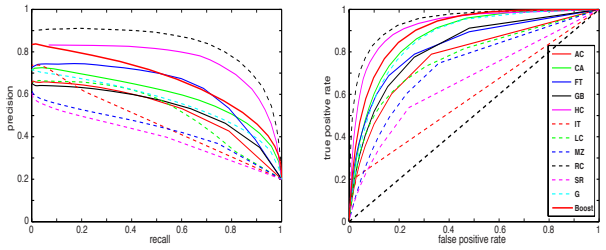

Figure 7. Precision-recall curves for naive thresholding of saliency maps using 1000 publicly available benchmark images [44].

Fig. 8 shows examples with human annotations and predictions of our model. As it can be seen our boosting model is able to successfully detect the most salient object, even in the condition that the salient object is not near the center.

Figure 8. Sample images from ASD dataset [44] with human annotations and unnormalized saliency map of our model.

5. Discussions and conclusions

We learned several models of visual saliency by integrating bottom-up and top-down features and compared their accuracy over the same data and scores. Our approach allows adding more features such as saliency maps of other bottom-up models or other top-down features. Among classifiers, AdaBoost has the best prediction accuracy followed by SVM and Regression. It outperforms majority of existing models and is thus far the *closest* model to human performance which can enhance performance of several approaches in computer vision. It has also competing performance for detecting the most salient object in a scene. An advantage of our approach is its generalization in a way that a classifier trained on one dataset performs well on the other datasets as opposed to training and testing for each dataset separately (as opposed to [48]). Our exhaustive comparison of the state-of-the-art models shows that although model rankings differ across datasets and scores, some models (GBVS [20], Judd *et al.* [1], Yan *et al.* [31], AWS [29], ICL [24], and Tavakoli *et al.*[40]) are better than the others.

One application of our method is conducting behavioral studies by comparing model parameters (W) across populations of human subjects for their differences in attention, for instance young vs. elderly or male vs. female. Although developing more effective bottom-up models based on purely low-level features is always welcomed, it is very important to discover and add more top-down factors for build-

ing more predictive models. As models are based on different saliency mechanisms, combining them may enhance the results thus helping to close the gap between humans and models in free-viewing of natural scenes.

References

[1] T. Judd, K. Ehinger, F. Durand and, A. Torralba. Learning to predict where humans look, *ICCV*, 2009. 1, 2, 3, 4, 5, 7

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

[3] J. Li, Y. Tian, T. Huang, and W. Gao, Probabilistic multi-task learning for visual saliency estimation in video, *IJCV*, 2010. 5

[4] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions PAMI*, 20(11), 1998. 2, 3, 5, 7

[5] J. Wang, J. Sun, L. Quan, X. Tang, and H.Y Shum. Picture collage. *CVPR*, 2006. 1

[6] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE T. Image Proc*, 2004. 1

[7] L. Marchesotti, C. Cifarelli, and G. Csurka. *ICCV*, 2009. 1

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

[9] S. Marat, M. Guironnet et, and D. Pellerin. Video summary using a visual attention model. *EUSIPCO*, 2007. 1

[10] R. Rosenholtz, A. Dorai, and R. Freeman. Do predictions of visual perception aid design? *ACM Trans on Applied Perception*, 2011. 1

[11] W. Einhäuser, M. Spain, and P. Perona. Objects predict xations better than early saliency. *Journal of Vision*, 2008. 2

[12] L. Elazary and L. Itti. Interesting objects are visually salient. *J. Vision*, 2008. 2

[13] R. Subramanian, H. Katti, N. Sebe, M. Kankanhalli, and T.S. Chua. An eye fixation database for saliency detection in images. *ECCV*, 2010. 2, 4, 5

[14] M. Cerf, J. Harel, W. Einhäuser, and C. Koch. *NIPS*, 2007. 2, 4

[15] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 1985. 2, 3

[16] A.M. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psych.*, 12:97-136, 1980. 2, 3

[17] G. Kootstra, A. Nederveen, and B. de Boer. Paying attention to symmetry. *BMVC*, 2008. 2, 4, 5

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

[19] S. Frintrop. VOCUS: A visual attention system for object detection and goal-directed search. Springer 2006. 5

[20] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. *NIPS*, 2006. 2, 3, 5, 7

[21] N.B Bruce and J.K Tsotsos. Saliency based on information maximization. *NIPS*, 2005. 5

[22] L. Itti and P. Baldi. Bayesian surprise attracts human attention.*NIPS*,2005. 2, 5

[23] X. Hou and L. Zhang.Saliency detection:A spectral residual approach.*CVPR*,2007. 2, 5, 7

[24] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. *NIPS*, 2008. 5, 7

[25] M. Mancas. Computational attention: Modelisation and application to audio and image processing. PhD. thesis, 2007. 2, 5

[26] T. Avraham, M. Lindenbaum. Esaliency (Extended Saliency): Meaningful attention using stochastic image modeling. *PAMI*, 32(4):693-708, 2010. 2, 5

[27] C. Guo and L. Zhang. A novel multiresolution spatiotemporal saliency detection model and Its applications in image and video compression. *IEEE Trans. on Image Processing*, 19(1), 2010. 2, 5

[28] P. Bian and L. Zhang. Biological plausibility of spectral domain approach for spatiotemporal visual saliency. *LNCS*, 2009. 5

[29] A.G. Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil. Decorrelation and distinctiveness provide with human-like saliency. *ACIVS*, 2009. 2, 3, 5, 7

[30] Y. Li, Y. Zhou, J. Yan, and J. Yang. Visual saliency based on conditional entropy. *ACCV*, 2009. 5

[31] J. Yan, J. Liu, Y. Li, and Y. Liu. Visual saliency via sparsity rank decomposition. *ICIP*, 2010. 5, 7

[32] A. Torralba. Modeling global scene factors in attention. *Journal of Optical Society of America*, 20(7), 2003. 2, 3, 5

[33] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE PAMI*, 2006. 2, 5

[34] H.J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9:1-27, 2009. 5

[35] D. Gao, S. Han, and N. Vasconcelos. Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE PAMI*, 2009. 2

[36] S. Marat, T. Ho-Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué. *IJCV*, 2009. 2, 5

[37] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395-1407, 2006. 5

[38] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, SUN: A Bayesian framework for saliency using natural statistics. *JOV*, 2008. 2, 3, 4, 5

[39] B.W. Tatler, The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor bases and image feature distributions. *J. Vision*, 14(7):1-17, 2007. 1, 3, 4, 7

[40] H.R. Tavakoli, E. Rahtu, and J. Heikkilä. *SCIA*, 2011. 5, 6, 7

[41] M.M. Cheng, G.X. Zhang, N.J. Mitra, Xiaolei Huang, and S.M. Hu. Global contrast based salient region detection. *CVPR*, 2011. 2, 7

[42] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk. Salient region detection and segmentation. *ICVS*, 2008. 7

[43] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *CVPR*, 2010. 7

[44] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk. Frequency-tuned salient region detection. *CVPR*, 2009. 2, 7

[45] Y.F. Ma and H.J. Zhang. Contrast-based image attention analysis by using fuzzy growing. *ACM Multimedia*, 2003 7

[46] B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman. LabelMe: A database and web-based tool for image annotation. *IJCV*, 2008. 2, 5

[47] R.J. Peters and L. Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. *CVPR*, 2007. 2, 5

[48] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, 11(3), 2011. 2, 7

[49] Q. Zhao and C. Koch. Learning visual saliency. *Information Sciences and Systems Conference*, 2011. 2

[50] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9), 2010. 4

[51] P. Viola and M. Jones. Robust real-time object detection. *IJCV*, 2001. 4

[52] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. of Computer and System Sciences*, 1997. 4

[53] R. Carmi and L. Itti. Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research*, 46(26), 2006. 1

[54] H. F. Chua, J.E. Boland, and R.E. Nisbett. Cultural variation in eye movements during scene perception. *PNAS*, 2005. 2

[55] L. Itti, N. Dhavale, and F. Pighin. *SPIE*, 2003. 3

[56] A. Toet. Computational versus psychophysical image saliency: A comparative evaluation study. *IEEE trans. PAMI*, 2011. 3