Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2012-001 January 13, 2012

A Benchmark of Computational Models of Saliency to Predict Human Fixations

Tilke Judd, FrØdo Durand, and Antonio Torralba

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

A Benchmark of Computational Models of Saliency to Predict Human Fixations

Tilke Judd, Frédo Durand, and Antonio Torralba

Abstract—Many computational models of visual attention have been created from a wide variety of different approaches to predict where people look in images. Each model is usually introduced by demonstrating performances on new images, and it is hard to make immediate comparisons between models. To alleviate this problem, we propose a benchmark data set containing 300 natural images with eye tracking data from 39 observers to compare model performances. We calculate the performance of 10 models at predicting ground truth fixations using three different metrics. We provide a way for people to submit new models for evaluation online. We find that the Judd et al. and Graph-based visual saliency models perform best. In general, models with blurrier maps and models that include a center bias perform well. We add and optimize a blur and center bias for each model and show improvements. We compare performances to baseline models of chance, center and human performance. We show that human performance increases with the number of humans to a limit. We analyze the similarity of different models using multidimensional scaling and explore the relationship between model performance and fixation consistency. Finally, we offer observations about how to improve saliency models in the future.

Index Terms—Saliency models, fixations, benchmark, center bias, fixation maps, metrics

1 Introduction

A Number of computational models of visual attention have been developed and the state of the art is rapidly improving. When a new model is introduced, it is often compared to a couple other models, but there is no clear way to quantitatively compare all the models against each other. To address this problem, we propose a benchmark data set, containing 300 natural images with eye tracking data from 39 observers to compare the performance of available models. Because models were initially developed with the aim to predict the human visual system, we measure the performance of a model by how well it *predicts where people look* in images in a free-viewing condition.

Where people look in an image is affected both by bottom-up and top-down mechanisms of visual attention. Many models only account for bottom-up mechanisms because it is hard to model the top-down mechanisms involved in all possible states of the observer (memories, culture, age, gender, experiences) and possible tasks (searching, browsing, recognizing). Our goal is to try and model attention for an *average* observer with a *free viewing* task. By free viewing we mean to imitate situations in which observers are viewing their world without a specific goal.

The difference between the performance of the best models today and the performance of humans shows that there is still room for improvement.

• *Tilke Judd, Frédo Durand and Antonio Torralba are with the Department of Electrical and Computer Engineering, Massachusetts Institute of Technology, Cambridge, MA, 02139. E-mail: {tjudd, fredo, torralba}@csail.mit.edu.*

1.1 Contributions

We make the following contributions:

- We provide an extensive review of the important computational models of saliency in Section 2.
- We introduce an image and eye tracking benchmark data set of 39 viewers on 300 images in Section 3.
- We show the performance of ten modern saliency models and three baseline models to predict human fixations on this benchmark data set under three metrics in Section 4. We provide a way to add new models to the comparison online.
- We show how human consistency, or the ability of a human fixation map to predict fixations from other humans, goes up as the number of observers increases in Section 4.3. The fixation map from 1 observer predicts fixations about as well as the best saliency models do (0.89 AuROC) and the limit of the performance of human appears to be 0.91 (AuROC). We extrapolate from this that our fixation map created from 39 observers is about 98% the way to a ground truth fixation map.
- We demonstrate that models that output blurrier maps and models that are biased to the center tend to perform better in Section 4.4. We optimize models for blurriness and centeredness and show that many models improve significantly.
- We see which models are similar and which are outliers by plotting them with multidimensional

scaling in Section 4.5.
- We show that model performance depends moderately on the human fixation consistency of an image in Section 4.6.
- We analyze what kinds of images models perform poorly on in Section 5 and suggest ways of improving existing saliency models.

2 Computational models of visual attention

In computer vision, robotics, human-computer interaction, computer graphics and design, there is a strong interest in having models, inspired by the human visual system, that selects the most relevant parts within a large amount of visual data. The objective of these models is to improve artificial vision systems by computing a numerical value of the likelihood of attending to, or the saliency of, every location in an image. The approaches vary in detail but have a similar structure. First we describe the structure and then group important models by type and describe them in rough chronological order.

2.1 General structure of computational models

Most computational models of attention have a structure that is adapted from feature integration theory [72] and the guided search model [85] and appears first in the algorithmic model of attention by Itti and Koch [40]. The main idea is to compute several features in parallel and to fuse their values in a representation which is usually called a *saliency map*.

The models generally include the following steps: First, the model computes one or several image pyramids from the input image to enable the computation of features at different scales. Then, image features are computed. Commonly used features include intensity, color, and orientation. Each feature channel is subdivided into several feature types (for example, r, g, b maps for color). Center-surround mechanisms or differences of Gaussians are used to collect within-map contrast into *feature maps*. The operation compares the average value of a center region to the average value of a surrounding region. The feature maps are summed up to feature dependent maps called *conspicuity maps*. Finally, the conspicuity maps are normalized, weighted and combined together to form the saliency map. The saliency map is usually visualized as gray-scale image in which the brightness of a pixel is proportional to its saliency.

A saliency map is often the final output of the model. However, some applications require the trajectory of image regions – mimicking human fixations and saccades. The selected image regions are local maxima in the saliency map. They might be determined by a *winner-take-all* approach and implemented with a notion of *inhibition of return* that ensures that all maxima are examined and prevents the focus of attention from staying at the most saliency region [58].

The way the different maps are fused is an important aspect of attentional systems. It is not clear how mapping and fusing happens in the brain, and computational systems use different approaches. Usually, a weighting function is applied to each map before summing up the maps. The weighting function determines the importance of features.

Before weighted maps are summed, they are usually normalized. This is done to weed out differences between a priori not comparable modalities with different extraction mechanisms. Additionally, it prevents channels that have more feature maps to be weighted higher than others.

After weighting and normalizing, the maps are summed to create the saliency map. Linear summation of feature channels into the final saliency map remains the norm.

The structure described so far is purely bottom up. Despite the well-known significance of top-down cues, most models consider only bottom-up computations because they are easier to model. Including other knowledge in a top-down matter is inspired by the Guided Search model and the theory of Biased Competition [15]. This is typically done by modulating the weights of the conspicuity maps before they are combined based on some top-down information about the scene or the task. Other ways of adding top-down information include adding context information, or faces, text and object detectors.

We explore specific examples of computational models of visual attention in the next section.

2.2 A selection of computational systems

The first computational model of visual attention was introduced by Koch and Ullman [40]. When first published, the model was not yet implemented but provided the algorithmic reasoning for later implementations. The winner-take-all (WTA) approach is an important contribution of their work.

Clark and Ferrier [14] were among the first to implement an attention system based on the Koch-Ullman model. It contains feature maps, which are weighted and summed up to a saliency map. Another early model was introduced by Milanese [47]. This work introduced concepts like conspicuity maps and feature computations based on center-surround mechanisms that are still used in models today.

The C++ *Neuromophic Vision Toolkit (NVT)* is another derivative of the Koch-Ullman model and is implemented and kept up to date by Itti and colleagues [35], [34], [32], [48]. This toolkit introduces image pyramids for the feature computations, which enables efficient processing.

Many others have tested this toolkit and suggested improvements: Parkhurst et al. [53] modified the basic

model to account for falloff in visual sensitivity. They noticed that the drop in visual sensitivity as a function of eccentricity on stimulus salience was an important determiner of attention and incorporated it in their model. Draper and Lionelle [16] introduced SAFE (selective attention as a front end) which modified the original approach such that it is more stable with respect to geometric transformations like translations, rotations, and reflections. Walther and Koch [83] extended this NVT model to attend to proto-object regions and created SaliencyToolBox (STB) Harel et al. [28] exploit the power, structure and parallel nature of graph algorithms to achieve efficient saliency computations of their Graph Based Visual Saliency model, which is based on the use of a *dissimilarity* metric. Le Meur et al. [41] [46] adapted the Koch-Ullman model to include the features of contrast sensitivity functions, perceptual decomposition, visual masking, and center-surround interactions. Others have updated the Koch-Ullman model by adding features such as symmetry [59] or curvedness [80].

Tsotsos' selective tuning model of visual attention [74] [75] [76] consists of a pyramidal architecture with an inhibitory beam. The model has been implemented for several features including luminance, orientation, color componency [76], motion [78] and depth from stereo vision [7]. Originally the selective tuning model processed one feature dimension only, but later it was extended to perform feature binding [62] [79].

The above approaches are based on biologically motivated feature selection, followed by center-surround operations, which highlight local gradients. Recently, some have hypothesized that fundamental quantities such as "surprise" and "self-information" and "signal to noise ratio" are at the heart of saliency and attention. Itti and Baldi [31] introduced a Bayesian model of surprise that aims to predict eye movements. Bruce and Tsotsos [6] [8] present a novel model for visual saliency computation built on a first-principles information-theoretic formulation dubbed Attention based on Information Maximization (AIM). They model bottom-up saliency as the maximum information sampled from an image.

Navalpakkam and Itti [49], [48], [50] define visual saliency in terms of signal to noise ratio (SNR). The model learns the parameters of a linear combination of low level features that cause the highest expected SNR for discriminating a target from distractors.

2.2.1 Models that add top-down components

The majority of the models described so far are bottom-up. However, it is well known that task is a strong influencer on our attention [86], especially in the context of search. In fact Henderson et al. [29] provide evidence that top-down information dominates real-world image search processes, such that the influence of low-level salience information on search guidance is minimal and others show that context is important [71] [51]. In order to correctly mimic the attention of humans, we have to successfully merge both bottom-up and top-down influences. Context of the scene is useful for speeding up search and recognition (we tend to look at the street rather than the sky when searching for our car) and can be added to models.

For example, Torralba et al.'s contextual guidance model [71] combines low-level salience and scene context when guiding search. Areas of high salience within a selected global region are given higher weights on an activation map than those that fall outside of the selected global region. The contextual guidance model outperformed a purely salience-driven model in predicting human fixation locations in a search task. This research has since been updated by Ehinger et al. [17].

Similar to the contextual guidance model, Zhang et al. [87] and Kanan et al. [38]'s Saliency Using Natural statistics (SUN) model combines top-down and bottom-up information to predict eye movements during real-world image search tasks. However, unlike the contextual guidance model, SUN implements target features as the top-down component. SUN outperformed a salience-driven model in predicting human fixation positions during real-world image search.

Both the contextual guidance model and the SUN model found that combining two sources of guidance significantly improved their abilities to predict human fixation locations, suggesting that humans similarly combine information types to guide search.

Goferman et al. [25] present context-aware saliency which aims at detecting the image regions that represent the scene and not just the most salient object. In addition to including low-level features such as contrast and color, they also consider global effects which suppress frequently occurring objects, they add a notion that visual forms may possess several centers of gravity, and they include detectors of human faces.

A second way to add top-down component to a model is to modulate the weights of the feature maps depending on the task at hand as originally explored by Wolfe et al. [85]. For example, if searching for a vertical green bottle, the model would increase the weights of the green and vertical orientation feature maps to allow those features to be attributed more saliency. In the salience map thus formed, all scene locations whose features are similar to the target become more salient and are more likely to draw attention. Navalpakkam and Itti [49], Elazary and Itti [20] and Gao et al. [22] use this approach.

Elazary and Itti [20] propose a model called SalBayes, which denotes the marriage between saliency and Bayesian modeling. At its core, the model learns the probability of an object's visual appearance having a range of values within a particular feature map.

In a search task, the model influences the various feature maps by computing the probability of a given target object for each detector within a feature map. As a result, locations in the maps with the highest probability are searched first.

Marchesotti et al. [45] use context by proposing a model for saliency detection based on the principle that images sharing global visual appearances are likely to share similar salience. Assuming that a large annotated image database is available, they retrieve the images most similar to the target image, build a simple classifier and use it to generate saliency maps. Their main application is image thumbnailing.

Similarly, Gao and Vasconcelos and colleagues [22], [23] and [24] propose a unified model for top-down and bottom-up saliency as a classification problem. They first applied this model to object detection [24] in which a set of features are selected such that a class of interest is best discriminated from other classes, and saliency is defined as the weighted sum of features that are salient for that class. In [22], they defined bottom-up saliency using the idea that pixel locations are salient if they are distinguished from their surroundings. They used difference of Gaussians (DoG) filters and Gabor filters, measuring the saliency of a point as the Kullback-Leibler (KL) divergence between the histogram of filter responses at the point and the histogram of filter responses in the surrounding region.

A third way to add top-down guidance to models is to incorporate the use of object detectors. The work of Cerf et al. [11][9][12][10] confirmed that faces and text strongly attract attention and showed that they were difficult to ignore even when doing so imposes a cost. They refined the Itti and Koch [35] model by adding a conspicuity map indicating the location of faces and text and demonstrate that this significantly improves the ability to predict eye fixations in natural images. They provide a working model which combines the saliency map computation of Itti and Koch model with the locations of faces based on the Viola Jones [82] face detection algorithm.

Additionally, Einhäuser et al. [18] showed that objects predict fixations better than early saliency. They add a human defined object-map to Itti and Koch model and show that fixations are predicted better by objects than by early saliency.

2.2.2 Fourier based models

Hou and Zhang [30] proposed a spectral residual approach based on the Fourier transform. The spectral residual approach does not rely on parameters and detects saliency rapidly. The difference between the log spectrum of an image and its smoothed version is the spectral residual of the image.

Wang and Li [84] build on Hou and Zhang's approach by combining spectral residual for bottom-up analysis with features capturing similarity and continuity based on Gestalt principles.

Guo and Zhang [27] later point out that the phase spectrum, not the amplitude spectrum, of an image's Fourier transform that is key to calculating the location of salient areas. They propose a novel multiresolution spatiotemporal saliency detection model called "phase spectrum of quaternion Fourier transform" (PQFT) to calculate the spatiotemporal saliency map of an image by its quaternion representation.

2.2.3 Region-based models

Another fundamental difference between saliency models is whether they are feature-based or incorporate some local grouping and are then called region-based models. Region-based models are well suited for object segmentation tasks [2][1][43][4]. Achanta and colleagues [2] and [1] present a method for salient region detection that exploits features of color and luminance and outputs full resolution saliency maps with well-defined boundaries of salient objects.

Liu et al. [43] suggest that saliency can be learned from manually labeled examples. They formulate salient object detection as an image segmentation problem, where they separate the salient object from the image background. They use features of multiscale contrast, center-surround histogram and color spatial-distribution.

Avraham and Lindenbaum [4] and [3] propose extended saliency (or ESaliency) that uses a validated stochastic model to estimate the probability that an image part is of interest. They use a region-based method by starting with a rough grouping of image regions, and then select regions that are unique with respect to the whole global scene rather than having local contrast.

2.2.4 Models that learn parameters

Most of the methods require many parameters such as the number of filters, type of filters, choice of the non-linearities, and a proper normalization scheme. These methods tend to emphasize textured areas as being salient regardless of their context. A good alternative is to use non-parametric approaches or learn the free parameters using machine learning.

Kienzle et al. [39] proposed to learn a visual saliency model directly from human eyetracking data using a support vector machine (SVM). They use ground truth data of eye tracking fixations on a small database of grey-scale image of natural scenes that have no particular salient object.

Seo and Milanfar [67] [66] use local regression kernels as features which fundamentally differ from conventional filter responses. They use a nonparametric kernel density estimation for these features, that results in a saliency map constructed from a local self-resemblance measure, indicating likelihood of saliency.

Judd et al. [37] use a linear SVM to learn weights of color, intensity and orientation features as well as feature channels from object, face and horizon detectors and a center gaussian. The weights are learned from ground truth training data of 15 subjects on 300 images.

Zhao and Koch [88] use a least square technique to learn the weights associated with a set of feature maps from subjects freely fixating natural scenes drawn from four different eye-tracking data sets. They find that the weights can be quite different for different data sets, but face and orientation channels are usually more important than color and intensity channels.

2.2.5 Models that include a center bias.

Researchers have shown several times that eye fixations tend to be biased towards the center of an image [54], [69] [70] [73], [41], [8]. Despite this, only Parkhurst and Niebur [54] and Zhao and Koch [88] and Judd et al. [37] have implemented a model that incorporates this bias.

2.2.6 Further reviews

If the reader is interested in further information, we refer the reader to a nice summary in Tsotsos et al. [77], a review by Rothenstein and Tsotsos [63] which presents a classification of models with details on the functional elements each includes, an overview by Shipp [68] that compares different models along the dimension of how they map onto system level circuits in the brain and a nice survey by Frintrop et al. [21] which covers models up to about 2005. However, the state of the art has changed dramatically in the last several years, warranting a new overview of the field.

2.3 Previous comparisons

Several authors have shown that regions of interest found by computational model of visual attention correlate with eye fixations [59] [53] [19] [29] [52] [6] [33] [57] and the reported area under receiver operating characteristic (ROC) curve, which measures how well the two correlate, has increased over time as models get better. However, these numbers cannot be compared directly as they come from different experiments using different images under different conditions.

To facilitate comparisons, several authors have produced openly available databases of images with fixations from eye tracking experiments [41] [42] [10] [6] [37] [61] and executable code of models that others can use. We provide links to these through our website.

Given the accessibility of image databases and code, some authors have recently made substantial comparisons. Zhang et al. [87] use the area under the ROC curve and KL-divergence to measure performance and compare against methods by [35], [6] and [22]. Seo and Milanfar [66] use the same measurement protocol and compare their model against the same three above and the SUN model [87]. Zhao and Koch [88] learn ideal weights for the Itti and Koch model using four different datasets. They use three different metrics, ROC curve, Normalized Scanpath Saliency (NSS) and Earth Mover's Distance, to evaluate the performance of the models.

Our work builds off of these comparisons. We differ in that we compare many more models (10 different ones) and we compare them on a new data set of images with a large amount of observers (39 observers). We measure performance of models using the area under the receiver operator characteristic (ROC), Earth Mover's Distance (EMD) and a similarity metric. This helps us make objective statements about how well the models predict fixations but also how they are similar and different from each other. Finally, we provide an online website to submit saliency maps from new models to be reviewed for performance. This means new models can easily be compared to many previous models at once using the same benchmark image data set and eye tracking data.

3 Experimental design

3.1 Benchmark data set

Images We collected 300 images from Flickr Creative Commons and personal image collections and recorded eye tracking data from 39 users who free-viewed these images. The longest dimension of each image was 1024 pixels and the second dimension ranged from 457 to 1024 pixels with the majority at 768 pixels. There were 223 landscape images and 77 portrait images.

Observers 39 observers (age range 18-50 years) participated in our eye tracking study. Each reported normal or corrected-to-normal vision. They all signed a consent form and were paid $15 for their time.

Method For the eye tracking experiment, we used a table-mounted, video-based ETL 400 ISCAN eye tracker which recorded observers' gaze paths at 240Hz as they viewed the series of images. Observers sat ~ 24 inches from a 19-inch computer screen of resolution 1280x1024px in a semi-dark room and used a chin rest to stabilize their head.

We used a five point calibration system, during which the coordinates of the pupil and corneal reflection were recorded for positions in the center and each corner of the screen. We checked camera calibration every 100 images and recalibrated if necessary. The average calibration error was less than one degree of visual angle (~ 35pixels). During the experiment, position data was transmitted from the eye tracking computer to the presentation computer so as to ensure that the observer fixated on a cross in the center of a gray screen for 500ms prior to the presentation of the next image. We displayed each image for 3 seconds.

Fig. 1. Five images from our benchmark data set (top), the fixation locations from our 39 observers (middle), and the corresponding fixation maps (bottom).

The instructions for this test were "You will see a series of 300 images. Look closely at each image. After viewing the images you will have a memory test: you will be asked to identify whether or not you have seen particular images before." We stated that there was a memory test to motivate users to pay attention however, we did not run a memory test at the end.

The raw data from the eye tracker consisted of time and position values for each sample. We used the method from [71] to define saccades by a combination of velocity and distance criteria. Eye movements smaller than the predetermined criteria were considered drift within a fixation. Individual fixation durations were computed as elapsed time between saccades and the position of each fixation was computed from the average position of each data point within the fixation. We discarded the first fixation from each scanpath to avoid the trivial information from the initial fixation in the center.

We obtain a continuous *fixation map* for an image from the eye tracking data by convolving a Gaussian filter across fixation locations of all observers. This is similar to the "landscape map" of [81]. We choose the size of the Gaussian to have a cutoff frequency of 8 cycles per image or about 1 degree of visual angle [18] to match with the area that an observer sees at high focus around the point of fixation. Figure 1 shows 5 images from our dataset, the fixations of 39 observers on the images, and the corresponding fixation maps. The fixation map is normalized to range between zero and one.

3.2 Saliency models

We compare ten computational models of attention in our benchmark. These models are major models of attention that have been introduced in the last five years (except for the Itti and Koch model which was originally implemented in 1998) and they offer executable code. In essence, these are the models that are *available and useable* to someone who might like to use one to build further applications.

We list the models we used here in the order they were introduced. The **Itti and Koch model** is based on the theoretical bottom-up feature-based model of Koch and Ullman [40], and was later implemented and improved [35], [34], [32], [83]. We used two implementations of this model: one implemented in Walther's Saliency Toolbox[1] and another which comes in the GBVS package[2]. The **Graph Based Visual Saliency (GBVS) model**[3] [28] is a graph-based implementation of the Itti and Koch model that uses a dissimilarity metric. **Torralba et al. model**'s [4] [71] incorporates the context of the scene. **Hou and Zhang's model**[5] [30] is based on the spectral residual of an image in the Fourier domain. Zhang et al. proposed the **SUN saliency model**[6][87] using natural statistics based on a Bayesian framework to estimate the probability of a target at every location. **Achanta et al.**[7] [1] provides a simple model which aims to cleanly extract objects from their background. The target application of their work is different from the current benchmark task, but it is still interesting to explore how it performs on our task. **Bruce and Tsotsos' AIM model**[8] [8] is a model for visual saliency computation built on a first principles information theoretic formulation dubbed Attention based on Information Maximization (AIM). **Judd et al.'s model**[9] [37] incorporates bottom-up and top-down image features and learns the appropriate weights for fusing feature channels together. The **Context-Aware Saliency**[10] [25] aims at detecting the image regions that represent the scene and not just the most salient object. Their goal is to find image regions that "tell the story of the image". Figure 2 shows saliency maps produced by each of the models.

The models vary greatly in the amount of salient pixels they return. For example, notice the difference in white or salient pixels between the Itti and Koch model and the Bruce and Tsotsos AIM model. In order to make comparisons across them, we match the histograms of the saliency maps to the histogram of the human fixation map for each image. In essence, this equalizes the amount of salient pixels allowed for a given image and asks each model to place the mass of salient pixels on the areas it finds most salient. Saliency maps with matched histograms are shown in Figure 3. We use these histogram-matched saliency maps for all our performance calculations. This affects the performance of some metrics but does not affect the performance for the ROC metric, which

1. http://www.saliencytoolbox.net/
2. http://www.klab.caltech.edu/~harel/share/gbvs.php
3. http://www.klab.caltech.edu/~harel/share/gbvs.php
4. Code provided by Antonio Torralba.
5. http://www.its.caltech.edu/~xhou/projects/spectralResidual/spectralresidual.html
6. http://cseweb.ucsd.edu/~l6zhang/
7. http://ivrgwww.epfl.ch/supplementary_material/RK_CVPR09
8. http://www-sop.inria.fr/members/Neil.Bruce/
9. http://people.csail.mit.edu/tjudd/WherePeopleLook
10. http://webee.technion.ac.il/labs/cgm/Computer-Graphics-Multimedia/Software/Saliency/Saliency.html

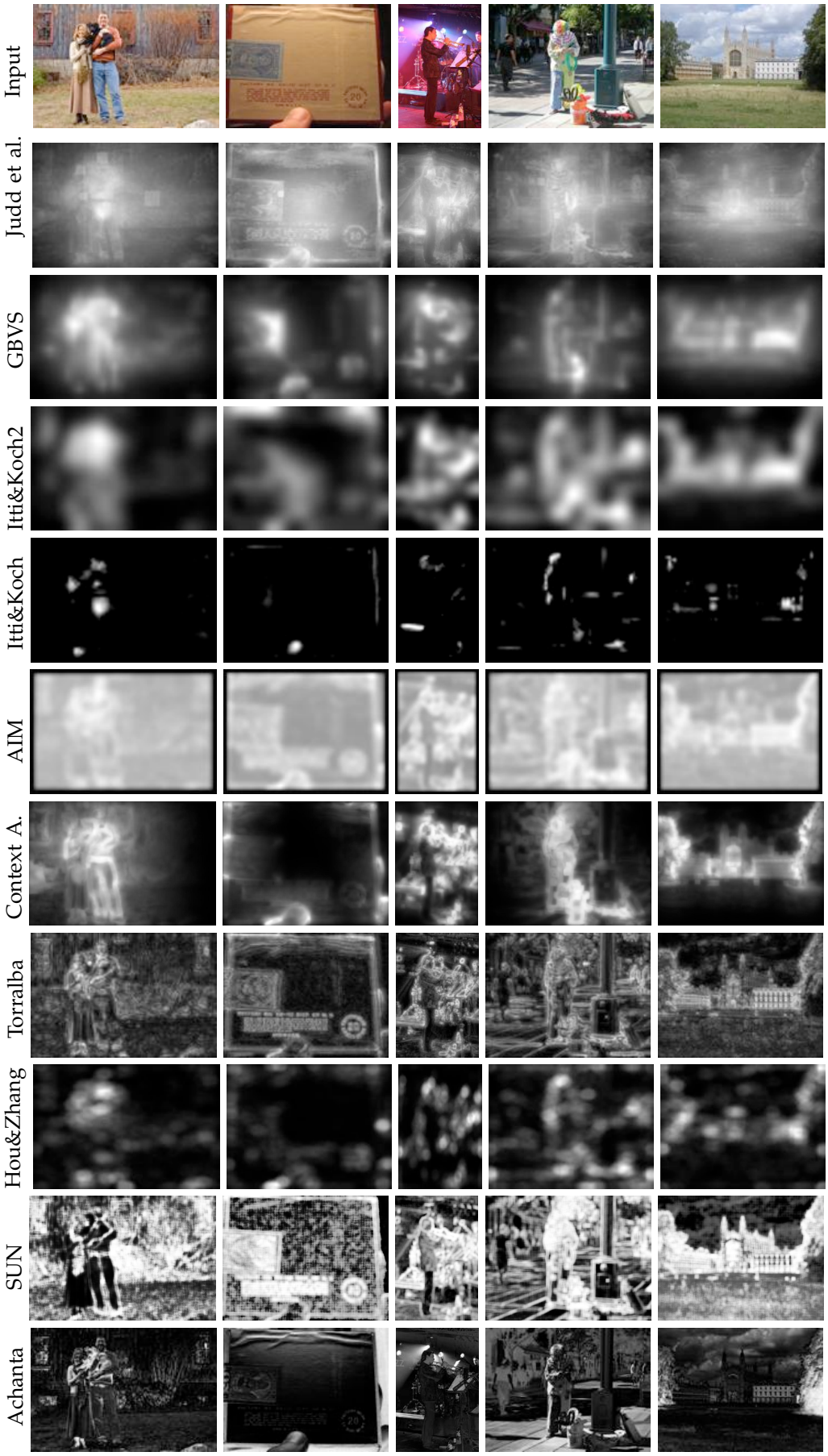

Fig. 2. Saliency maps from 10 different models.

Fig. 3. Histogram matched saliency maps. Matching the histograms between maps allows us to better understand which locations each model finds most salient

only depends on pixel ranking.

3.3 Baselines

In addition to comparing saliency models to each other, we compare them to three baselines shown in Figure 4. They have been histogram-matched to the fixation maps so that all baseline maps per image show the same amount of salient pixels.

Chance This model randomly selects pixels as salient. It is likely to be a very poor performing model.

Center This model predicts that the center of the image is the most salient. As the distance from the center of the image to the pixel increases, its saliency decreases. The model is created by stretching a symmetric Gaussian to fit the aspect ratio of a given image. This means that if the image is much longer than it is high, the Gaussian will have a longer horizontal axis. This stretched Gaussian performs slightly better than an isotropic Gaussian because it accounts for the tendency of objects of interest to be spread along the longer axis. Ideally, any good model of saliency should outperform the center model.

Human performance Humans should be the best predictors of where other humans will look. However they do not predict each other perfectly because of the variability of many human factors and the complexity of a given image. To calculate human performance we use the fixation map from n humans to predict fixations from other observers. In general, human performance goes up as you increase n. We explore this further in section 4.3.

3.4 Scoring metrics

Several metrics can be used to quantitatively evaluate the performance of saliency models. These measures include the Receiver Operating Characteristics (ROC) [26], the Normalized Scanpath Saliency (NSS) [57], correlation-based measures [36] [60], the least square index [29] [44], and the "string-edit" distance [5] [13] [65] among others.

Among these metrics, **ROC** is the most widely used in the community. According to [88] the inherent limitation of ROC, however, is that it only depends on the ordering of the fixations. In practice, as long as the hit rates are high, the area under the ROC curve (AUC) is always high regardless of the false alarm rate. While an ROC analysis is useful, it is insufficient to describe the spatial deviation of predicted saliency map from the actual fixation map. If a predicted salient location is misplaced, but misplaced close to or far away from the actual salient location, the performance should be different. To conduct a more comprehensive evaluation, we also use a measure of similarity and the Earth Mover's Distance (EMD) [64] that indicate spatial difference rather than only the ordering of the values. Though the ROC performance is not affected by our histogram matching, the similarity and EMD performance are.

The **similarity score (S)** is a measure of how similar two distributions are. After each distribution is scaled to sum to one, the similarity is the sum of the minimum values at each point in the distributions. Mathematically, the similarity S between two maps A and B is

$$S = \sum_{i,j} \min(P_{i,j}, Q_{i,j}) \text{ where } \sum_{i,j} P_{i,j} = \sum_{i,j} Q_{i,j} = 1.$$

A similarity score of one indicates the distributions are the same. A similarity score of zero indicates that they do not overlap at all and are completely different.

Earth Mover's Distance (EMD) [64] [55] is a measure of the distance between two probability distributions over a region. Informally, if the distributions are interpreted as two different ways of piling up a certain amount of dirt over the region, the EMD is the minimum cost of turning one pile into the other; where the cost is assumed to be amount of dirt moved times the distance by which it is moved. More formally, Pele and Werman [55] defined $\widehat{EMD}$ as:

$$\widehat{EMD}(P,Q) = (\min_{\{f_{ij}\}} \sum_{i,j} f_{ij} d_{ij}) + |\sum_i P_i - \sum_j Q_j| \max_{i,j} d_{ij}$$

$$s.t.\ f_{ij} \geq 0 \quad \sum_j f_{ij} \leq P_i \quad \sum_i f_{ij} \leq Q_j \quad \sum_{i,j} f_{ij} = \min(\sum_i P_i, \sum_j Q_j)$$

where each f_{ij} represents the amount transported from the ith supply to the jth demand. d_{ij} is the *ground distance* between bin i and bin j in the distribution. A larger EMD indicates a larger overall difference between the two distributions. An EMD of zero indicates that two distributions are the same. We use a fast implementation of EMD provided by Pele and Werman [55] [56][11] but without a threshold.

Overall, a good saliency model would have a high area under the ROC score, a high Similarity score and a low EMD score.

4 Experimental Results

In this section we show how well saliency maps predict human fixations under 3 metrics and examine which models have similar performance. We compare these to the performance of the baseline center and chance models. We also explore measures of human consistency under the three metrics.

We provide instructions on how to upload new models of saliency to be included in this comparison online.

We show that blurring saliency maps and adding a bias towards the center increases the performance of many models. To compensate for this, we optimize these parameters for each model and compare their increased performance.

11. Code at http://www.cs.huji.ac.il/~ofirpele/FastEMD/

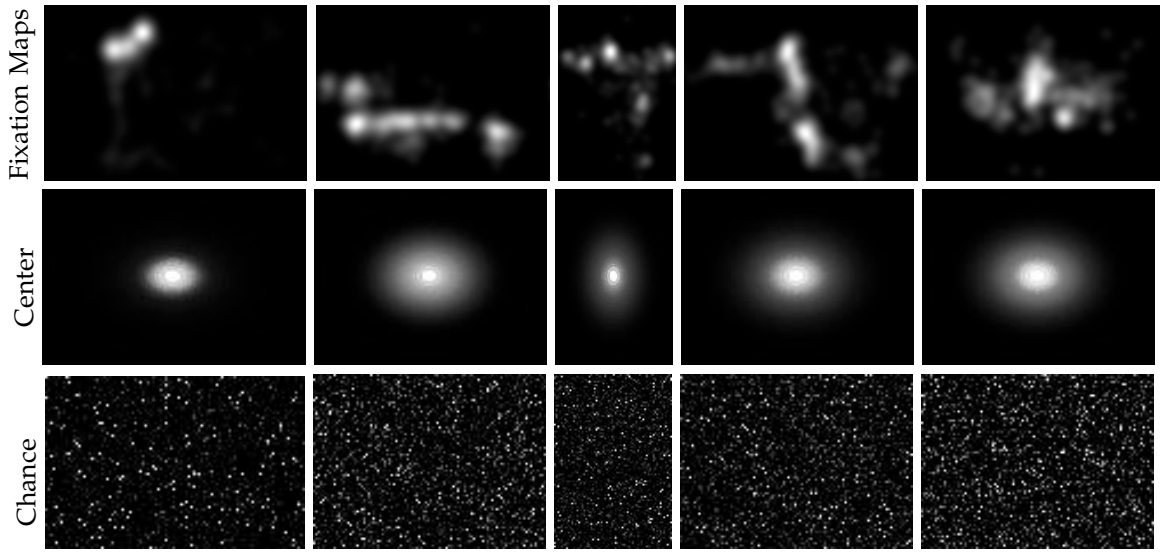

Fig. 4. Baseline models for predicting where people look include human fixation maps (top), the center model (middle), and the chance model (bottom).

In addition to assessing models' performance at predicting where people look, we measure how similar the models are to each other. We create a similarity matrix to view relationships between models and plot the models in a higher dimensional space using multidimensional scaling. This helps visualize how related the saliency models are.

Finally we describe the complexity of our images based on how similar each observer's fixations are to the average fixation map per image, and show how this image complexity affects saliency model performance results.

4.1 Model performances

Figure 5 shows the performance of saliency models using the three different metrics ROC, similarity and EMD. We measured performances using the histogram matched saliency maps from each model.

ROC The top chart of Figure 5 indicates how well saliency maps from each model predict ground truth fixations and shows the area under the ROC curve (AUR). For this metric, higher scores are better. We see that human performance is the highest and that all models perform better than chance. The center baseline outperforms many models of saliency. This is because of photographic bias of images and viewing-strategy of observers [73].

The Judd et al. and GBVS are the highest performing models and the only models to outperform the center. This is most likely because they are the only models to incorporate a global center bias.

The Context Aware, AIM, and Itti and Koch2 models have about the same performance. Looking at the saliency maps of these models shows that they share similar visual properties and select similar salient locations. This is interesting since the models are fundamentally quite different: Bruce and Tsotsos' AIM model aims to maximize information sampled from a scene and is derived from mathematical principles. On the other hand, the Itti and Koch model is a feature-based model with several parameters and a biologically inspired design. The context-aware model is based on the feature-based Itti and Koch model though it also includes global considerations, Gestalt principles and a face detector. Despite their differences, the models have similar performances. These three models also differ in how blurry their saliency maps are, with Context Aware having the least blurry maps and Itti and Koch 2 having the most blurry maps.

Blurrier saliency maps give better overall results as seen by the higher performance of the blurrier models (GBVS, Itti and Koch2, Bruce and Tsotsos AIM, Context Aware) over the more detailed, higher frequency models (SUN, Torralba, and Achanta). Instead of delineating salient regions from non salient regions in a binary fashion, having a natural decay is preferable. The nearby regions are likely to be salient and should be ranked higher than regions far from very salient objects.

The Itti and Koch and Achanta models perform the most poorly. Itti and Koch may perform poorly because its map is not blurry. The Achanta model selects objects with clean boundaries, but they are often not the salient objects. In general the Achanta model does well on simple images with one main salient object which the model selects well from the background. This is what the model was originally designed for, as it was intended to be a preprocess for image segmentation. When the images are complex scenes as in this benchmark, the model does not perform as well. For this data set, the Achanta model

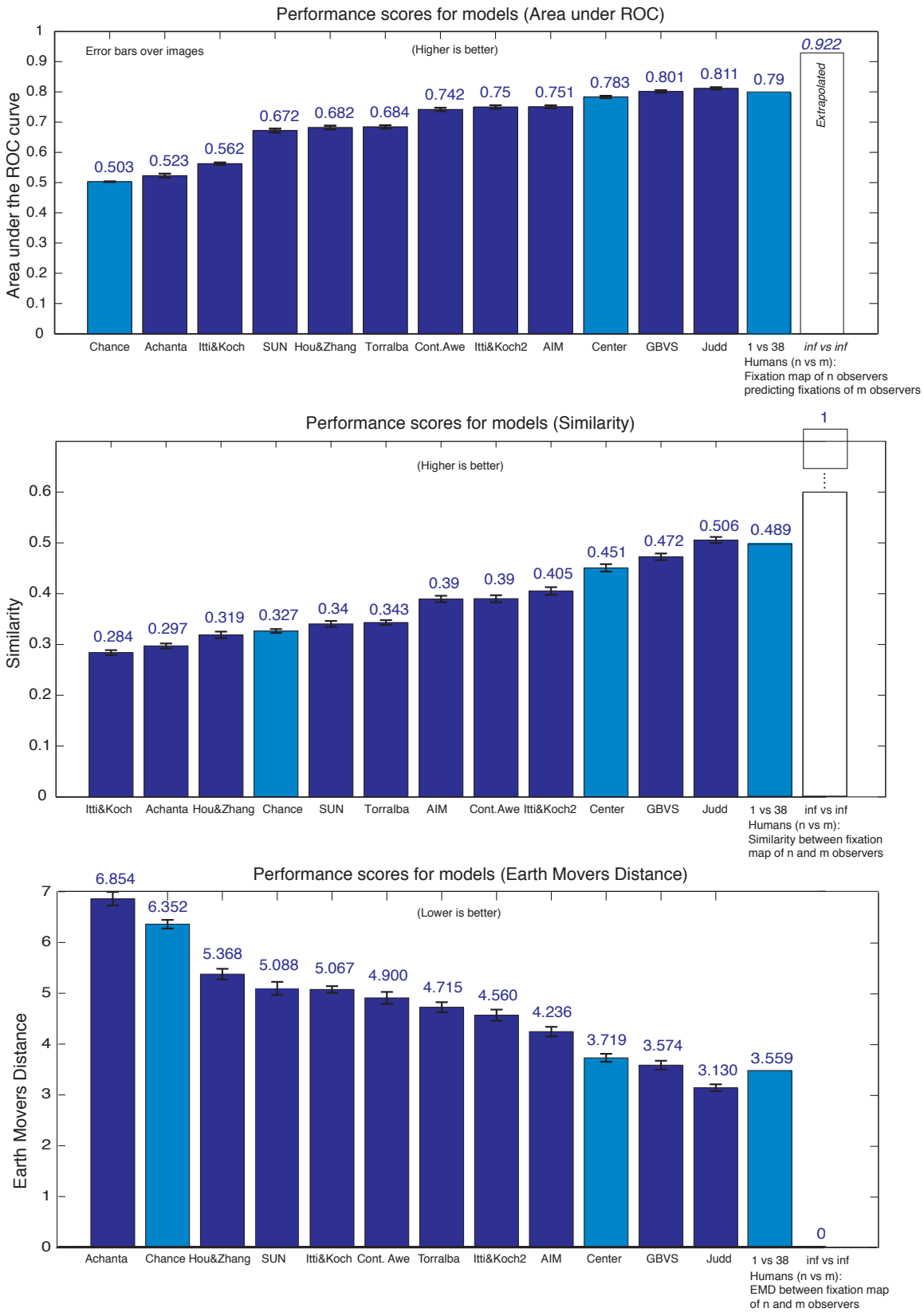

Fig. 5. Performance of all models using the ROC, similarity, and EMD metric. For the first two graphs, higher values are better. Lower values are better for EMD performance. Error bars show the uncertainty of the mean over 300 images. The Judd et al. and GBVS models perform the highest independent of the metric used.

performs about as well as chance.

The two implementations of the Itti and Koch map perform very differently. This is mostly due to implementation details of the two methods used with their out-of-the-box parameters and demonstrates that selecting good parameters can have a very large effect on performance. In this particular case, Itti and Koch 2, which was implemented in the GBVS toolbox, has a larger level of blur applied to the master map. We explore the effect of blurriness on the performance of a model in section 4.4.

Similarity The middle chart of Figure 5 represents how similar the saliency maps from different models are to the human fixation map. Higher scores indicate better performance and a perfect score of 1 would indicate that two maps are exactly the same. Overall the ordering of performances based on similarity is similar to the ordering from the ROC performances. The one major difference is that the chance model now performs better than 3 other models (IttiKoch, Achanta and HouZhang). Under the similarity metric, a conservative guess that salient locations are spread across the image will outperform a focussed guess that is incorrect.

Earth Mover's Distance The bottom chart of Figure 5 measures the earth mover's distance between the saliency map of a given model and the corresponding human fixation map. Lower values indicate better performance as less "earth" needs to be moved. Under this metric we see similar orderings as compared to the ROC metric, though some neighbor models are swapped in ranking.

4.2 Online benchmark

In addition to the analysis provided here, we provide performance data of existing models and instructions for submitting new models for evaluation online at http://people.csail.mit.edu/tjudd/SaliencyBenchmark/.

The site includes a short description of each model, a link to the code for the model, and the performance with each metric. To allow others to understand the type of images and data in the data set, we include fixations and fixation maps for 10 of the images. We also include saliency maps from each model for direct visual comparison.

The website also includes instructions for submitting new models for evaluation. The instructions are summarized here: 1) Download all 300 images in this data set. 2) Calculate saliency maps with the new model. The model should create maps that rank image locations by their likelihood to attract fixations of human observers as it will be measured for its ability to predict where people look. 3) Submit the saliency maps via email to tjudd@csail.mit.edu or online using the submission interface. 4) Wait for your AUR, Similarity and EMD scores. As the actual fixations on the 300 images are hidden from the world to avoid training and over fitting models, we run the scoring calculations and send you the results. 5) Your scores and a description of your model will be included on the website.

If the reader is interested in training new models based on images with fixation data sets, we provide a link to our previous MIT dataset of 1003 images with 15 observers introduced in [37] gathered in conditions similar to that of the new test data. In addition we link to several other smaller data sets that are publicly available.

4.3 Human consistency

Humans do not predict each other perfectly as they do not fixate at exactly the same locations in an image. Despite these variations, there is typically locations in an image that *most* observers look at. A fixation map created from an infinite number of observers would be the best predictor. We can measure human performance from a finite number of observers and approximate the limit with infinite viewers. Figure 5 includes two measures: how well 1 human predicts an average fixation map of 38 viewers (1 vs 38), and the extrapolated limit of human performance (infinity vs infinity).

4.3.1 1 vs 38

We plot how well a fixation map from 1 human viewer predicts the fixations (ROC metric) or the fixation map (Similarity and EMD metric) of the 38 other observers, averaged over all images and five variations of observers per image. This measure indicates how well a *single human* predicts an average of humans. This value is about as high as the highest performing saliency models under all metrics. Automatic saliency models have been touted as an inexpensive alternative to time and cost intensive eye tracking studies, and are good when eye tracking is not possible. For those who can do eye tracking, having *just one human observer* is about as accurate as the best saliency model. This accuracy goes up with every additional viewer. For situations where it is not possible to run an eye tracking experiment, automatic saliency models are a good alternative, though it is clear that saliency models still need to be improved.

4.3.2 Infinity vs Infinity

We would like to know how well a fixation map of infinite observers predicts fixations or a fixation map from a different set of infinite observers. As this cannot happen in practice, we find the value as a limit. As the number of observers goes to infinity, the two fixation maps converge to be exactly the same. Similarity becomes 1, and Earth Mover's Distance becomes 0. The area under the ROC curve converges to a value less than, but close to 1.

By plotting performance as the number of humans increases we can see both what the limit of human performance is and how quickly it converges to that limit. Figure 6(a) shows performance as the number of observers used for the reference fixation map and the number of observers whose fixations are being predicted increases from 1 to 19. (Because we only have 39 independent observers, we cannot calculate human performance beyond 19 vs 19). We fit these points to a power function of the form $f(x) = a*x^b+c$ and find the coefficients with their 95% confidence bounds in parentheses to be:

a = -0.07034 (-0.0810, -0.0596)
b = -0.3054 (-0.382, -0.229)
c = 0.9221 (0.911, 0.934)

The value $c = 0.922$ is the extrapolated limit for human performance under the ROC metric given an infinity of observers. When we plot $0.922 - f(x)$ in the log domain, we get Figure 6(b) which approaches 0 with a slope of -0.305. We repeat this method with the other metrics (see supplemental material for graphs): The similarity performance increases to 1 with a slope in the log scale of 0.34. EMD performance decreases to 0 with a slope in the log scale of -0.48.

4.3.3 How many observers is good enough?

A perfect ground truth fixation map would come from an infinity observers. A legitimate question is then, "how many observers is good enough?".

Our curve suggests that fixation maps from increasing number of observers quickly converge to fairly accurate ground truth fixation map as measured by the ability of the fixation map to predict new fixations under the ROC metric. 10 observers create a fixation map with an AUR performance of 0.887 which is $\sim 92\%$ of the possible performance increase from 0.5 to the limit of 0.922. Extrapolating from our fitted curves, we see that the fixation maps we create from 39 viewers has a performance of 0.899 which captures $\sim 95\%$ of ground truth, and are an accurate approximation.

TABLE 1
Fixation map accuracy depends on num observers.

num observers	AuROC perf	percentage of perf. limit
2	0.865	86.5%
5	0.879	90%
10	0.887	91.7%
20	0.894	93.3%
40	0.899	94.6%
1000	0.914	98%

The performance of a fixation map depends on the metric that is used. The ROC metric measures how well the fixation map locates the most fixated regions first. The similarity and EMD metric measure how well the fixation map represents both the fixated and non-fixated regions; they are required to get the salient areas as well as the non-salient areas right. Because of this, fixation maps converge to perfect ground truth more slowly under the Similarity and EMD metric. The extrapolated performance of our 39 observers under the similarity metric is ~ 0.78 (extrapolating from the best fit linear function) or ~ 0.80 (extrapolating from the best fit quadratic function) which is 78% or 80% of the distance between 0 and 1. The extrapolated performance of 39 observers under the EMD metric is ~ 0.81 which is 87% of the distance between 6.3 (Chance performance) and 0 (ideal performance).

4.4 Optimizing blurriness and centeredness of models

We notice that blurrier saliency maps tend to perform better at predicting fixations than saliency maps with sharp edges, and maps that include some bias toward the center perform better than maps without (see Figure 7). Since we are interested in knowing which models use features that best model human fixations independent of the implemented level of blurriness or bias towards the center, we optimize these parameters for each model and the recalculate performances (see Figure 8). We optimize the parameters on a training set of 100 images of the Judd et al. 2009 ICCV data set [37] by varying the parameters and choosing the value that maximizes performance.

4.4.1 Optimizing blur

We found the optimal Gaussian blur levels for the models to be Achanta sigma=40px, BruceTsotsos 100, ContextAware 30, GBVS 30, HouZhang 30, IttiKoch 80, IttiKoch2 60, Judd 20, SUNsaliency 40, Torralba 40. Figure 7 shows that in general, blurring the model enhances performance under the ROC metric (green bars in Figure), but by less than 10% for all models except IttiKoch. IttiKoch performance increases from 0.56 AUR to 0.65 AUR, for a 16% improvement. The original IttiKoch saliency maps have very sharp boundaries that put it at a disadvantage under the ROC metric because non-salient pixels near salient pixels are not given priority over non-salient pixels far from salient pixels. When the maps are blurred, this problem is eliminated as pixels close to salient regions are ranked higher than those farther away, and this leads to the jump in performance. BruceTsotsos AIM model has the second largest gain at 7% improvement. This is mostly a side effect of the non-salient boundaries of its saliency maps that get blurred and create a center-like bias.

4.4.2 Optimizing center bias

To optimize the bias towards the center for each model, we linearly combine the models' saliency map

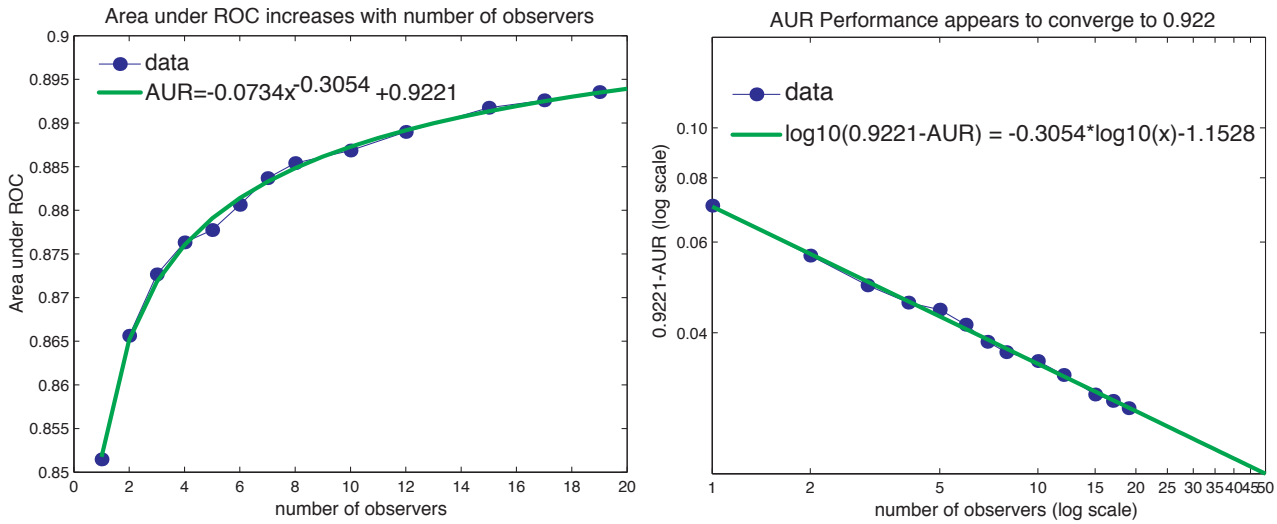

Fig. 6. Performance of x observers to predict fixations from x observers under the ROC metric. As number of human observers increases, prediction performance increases. This shows that human performance depends on the number of humans used. Based on extrapolation of the power curve that fits the data, it appears that limit of the human performance is 0.922 under the ROC metric. The plot on the right indicates the rate at which performance reaches that limit.

with the center map as follows:

$$\text{newMap} = w * \text{centerMap} + (1 - w) * \text{saliencyMap}$$

where w is the weight of the center map. We tried weights ranging from 0-1 and found optimal weights for the models to be Achanta w=1.0, BruceTsotsos w=0.5, Chance w=1.0, ContextAware w=0.7, GBVS w=0.6, HouZhang w=0.7, IttiKoch w=0.7, IttiKoch2 w=0.8, Judd w=0.3, SUNsaliency w=0.8, Torralba w=0.7. When w=1, the model is equivalent to the center model; the original model does not contribute to the final saliency map. Optimizing the center bias increases the performance of all models to be at or above the performance of the center (red bars in Figure). The Context-Aware model and Bruce and Tsotsos' AIM model improve when combined with the center map such that they perform as well as or better than the Judd et al. and GBVS models which where originally the highest performing.

Optimizing blur and the weight of the center map does not help close the gap between the highest performing models and human performance. Closing this gap is likely to depend on incorporating features that capture top-down influences on saliency, such as image semantics, object interaction, and better face, person, text and horizon detectors.

4.5 Multidimensional analysis

To understand which models are similar to each other, we calculate the similarity between the saliency maps of each model on each image and average the similarities across all images. This gives us a similarity between each of the models which we plot as a similarity matrix in Figure 9(a). The diagonals of the similarity matrix are 1 because a saliency map of a model is always the same as itself. The yellower the squares the more similar the maps of two models; the bluer the square, the more different the maps are. The center model is most similar to the Judd et al. model, GBVS and human fixations yet is dissimilar to the rest of the models. These are the only three models include a bias to the center. SUN saliency and the Torralba model are similar. Bruce and Tsotsos AIM is similar to Hou and Zhang, Torralba, Context Aware and Itti and Koch2. Itti and Koch, Achanta and Chance have the lowest similarity with all other models and have the bluest lines.

If we take 1 minus the similarity values, we get the dissimilarity values, or distances between models, that we can use to plot the models using multidimensional scaling as seen in Figure 9 (b) and (c). However, using 2 dimensions only accounts for 41% of the variance of the models; 3 dimensions accounts for 58% of the variance. These plots make it apparent that the chance model, Itti and Koch model and Achanta model are very different from the rest–they are outliers in the graph.

Because adding a center bias significantly changes the performance of each model, we redo the above analysis using the models with optimized center bias and show results in the bottom row of Figure 9. With optimized center bias, the models become much more similar to each other, as seen by the increased values and warmer colors of the similarity matrix. In multidimensional space, 2 dimensions accounts for 63% of the variance, 3 dimensions accounts for 73% of the variance. Chance, Achanta and Center models are now identically similar because their center weight is 1. Though many models have a higher performance,

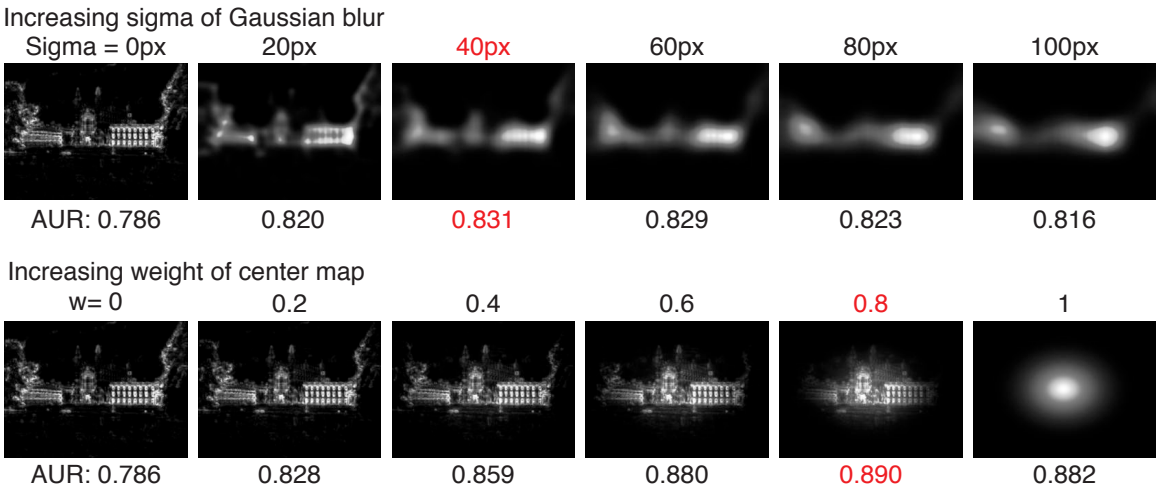

Fig. 7. Saliency maps with increasing blur and weight of center map. Red text indicates optimal parameters and maximum AUR performance for this example image.

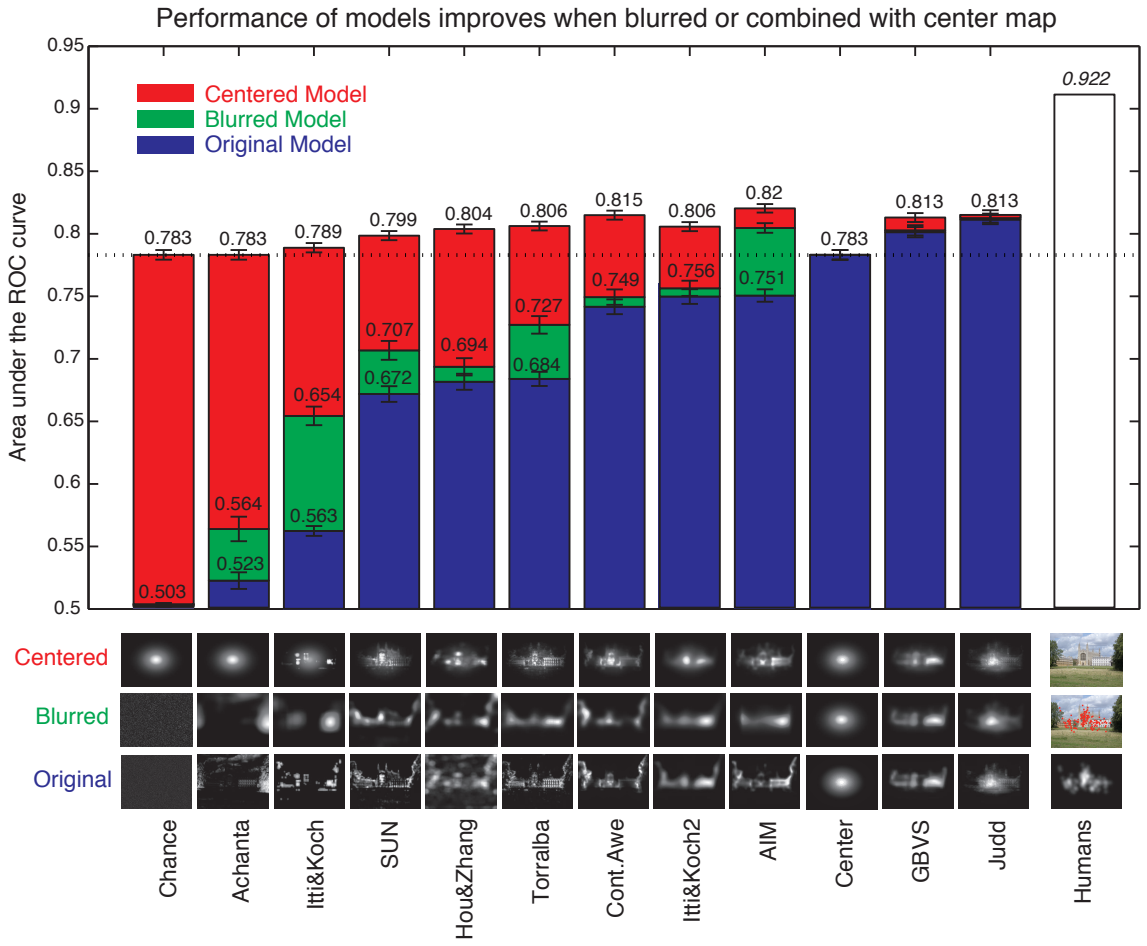

Fig. 8. Performance of the original, blurred, and centered models. When the models are optimized for blur or used in combination with the center map, model performance increases. Error bars show standard error over images.

the highest performing models do not become more similar to the fixation map. Other techniques beyond weighting the center must be used to increase the performance of saliency models.

4.6 Images ranked by fixation consistency

We find the fixation consistency for each image by measuring the similarity between one observer's fixation map to the average fixation map of all observers. Images where all observers look at the same locations have very high similarity scores; images where observers' fixations are very different from each other have low similarity scores. We rank images according to their consistency. The extremes of the ranking are shown in Figure 10. Images with people's faces, strong text, or one major object have have consistent fixations and high similarity scores. Complex outdoor scenes, cluttered data, or the notoriously detailed Where's Waldo images do not have consistent fixations and have low similarity scores. In these cases it may be harder for saliency models to accurately predict where people look.

To assess whether saliency model performance is affected by image fixation consistency, we used the image ranking and divided the ranked images into three consecutive bins corresponding to high, medium and low fixation consistency. Figure 11 shows the performance of 5 models under all three metrics for images in each bin.

As fixations consistency goes down, ROC scores decrease slightly (get worse), similarity scores increase slightly (get better), and earth mover's distances decrease slightly (get better). This shows that performance depends on both fixation consistency and on the metric used. For complex images with low fixation consistency, saliency maps have a hard time guessing what will be salient and often produce a spread out guess as to what is salient. This is advantageous to the Similarity and EMD metric but not to the area under the ROC curve metric.

We also binned images with respect to the total number of fixations per image (under the assumption images with higher number of fixations would select more complex images) and found the same trends.

5 Discussion

In this paper we have analyzed and compared current models of saliency. Though they have substantial power to predict human fixations, improvement is still needed. In this section we discuss what is necessary to improve current models.

5.1 When do models perform poorly or well?

We examine which images models have an easy or difficult time predicting fixations for and compare this to images that have high or low fixation consistency

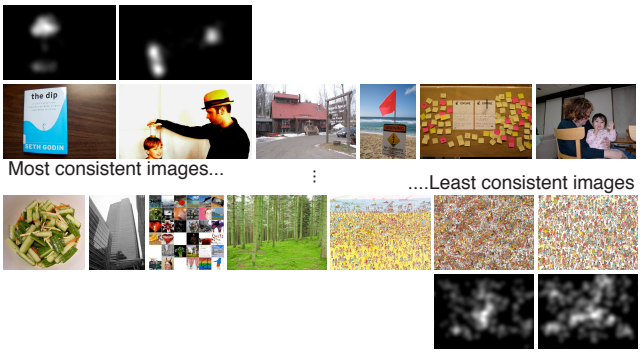

Fig. 10. Images ranked by similarity of human fixations. On the top are the images where humans are most consistent. The bottom row shows the fixation maps corresponding images where human fixations were most spread out.

in Figure 12. This compares how well models versus humans are at predicting fixations. To identify images for each quadrant we use image fixation consistency c defined before, and the average Similarity performance p over 3 top performing models (Judd et al., GBVS and Bruce and Tsotsos AIM models). We also use the descending ranked ordering of consistency cIX, and the descending ranked ordering of average model performance pIX.

Images in the top left quadrant have high fixation consistency and high model performance and are identified by their low summed rank $cIX + pIX$. These are images that saliency models perform well on. The images often have a strong salient object on a plain or simple background.

Images in the bottom left quadrant have high fixation consistency but low model performance and are identified by their low performance:consistency (low $p : c$) ratio. Though fixation consistency is high, models miss the important salient location. This typically happens when saliency model miss faces or text, or when the model fires on unusual colors or bright spots in the background that is not important to humans. For example, on the beach sign image, observers read the text and dismiss the orange colors while saliency models to the opposite. These images offer the largest potential area of improvement for saliency models. Humans are consistent about what is interesting, so models should be able to be to identify the same locations. We explore what should be added in the next subsection.

Images in the top right have low fixation consistency and high model performance and are identified by their high performance:consistency ratio (high $p : c$). On these images, fixations are spread out, and so are the saliency maps. The situation leads to high similarity scores but does not lead to high ROC scores.

Images in the bottom right have low fixation consistency and low model performance and are identified by their high summed rank $cIX + pIX$. Both humans and models do not predict fixations well. These are

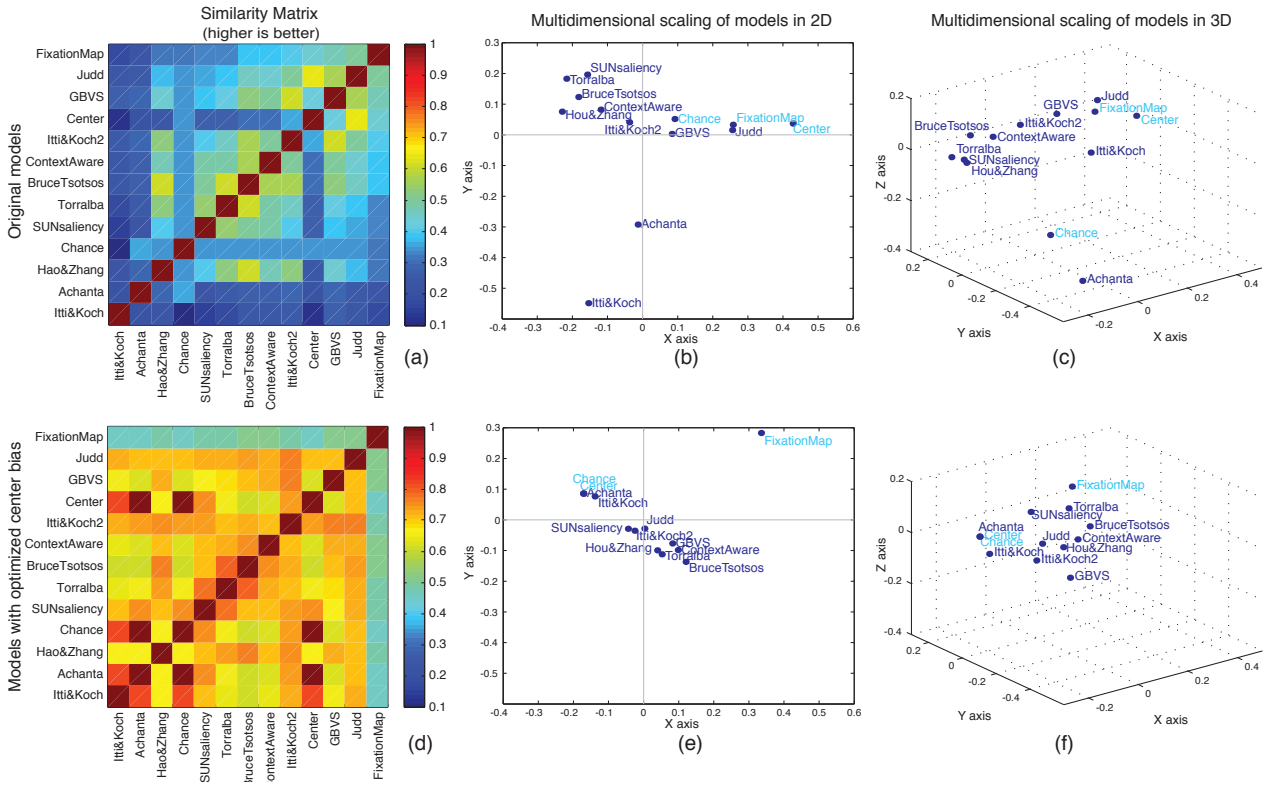

Fig. 9. The similarity matrices (a) and (d) show how similar each model is to all other models. They are plotted in the order of the first dimension of the multidimensional scaling. In (b), (e) and (c), (f), models are plotted in 2D and 3D space respectively using their 1-similarity value. The distance between the models' points in space gives an approximation of how different the models are from each other in higher dimensional space, though the plots only account for 41% (b), 58% (c), 63% (e), 73% (f) of the variance of the models. The top row shows the analysis between original models; the bottom line shows analysis between models with optimized center bias. After center bias optimization, the models are much more similar to each other as seen by the higher values in the similarity matrix (d).

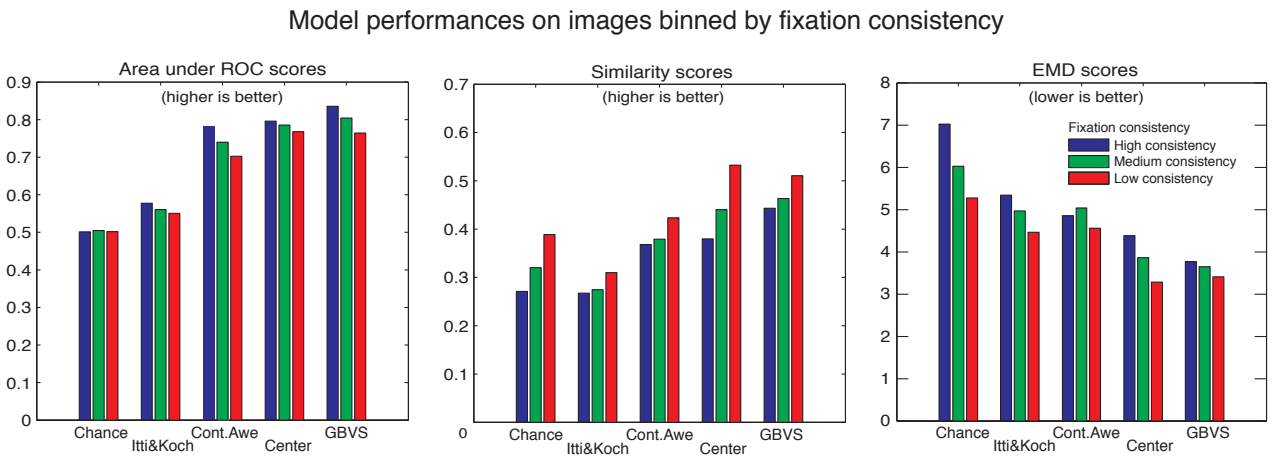

Fig. 11. **Top row:** Performance of 5 models for all three metrics on images binned by fixation consistency. As images become less consistent, performance scores get slightly worse under the ROC metric, but gets slightly better under the Similarity and EMD metrics. On images with low consistency, saliency maps often predict low saliency everywhere which produces higher Similarity and EMD scores but lower ROC scores. We found the same trend when images were binned by number of fixations per image.

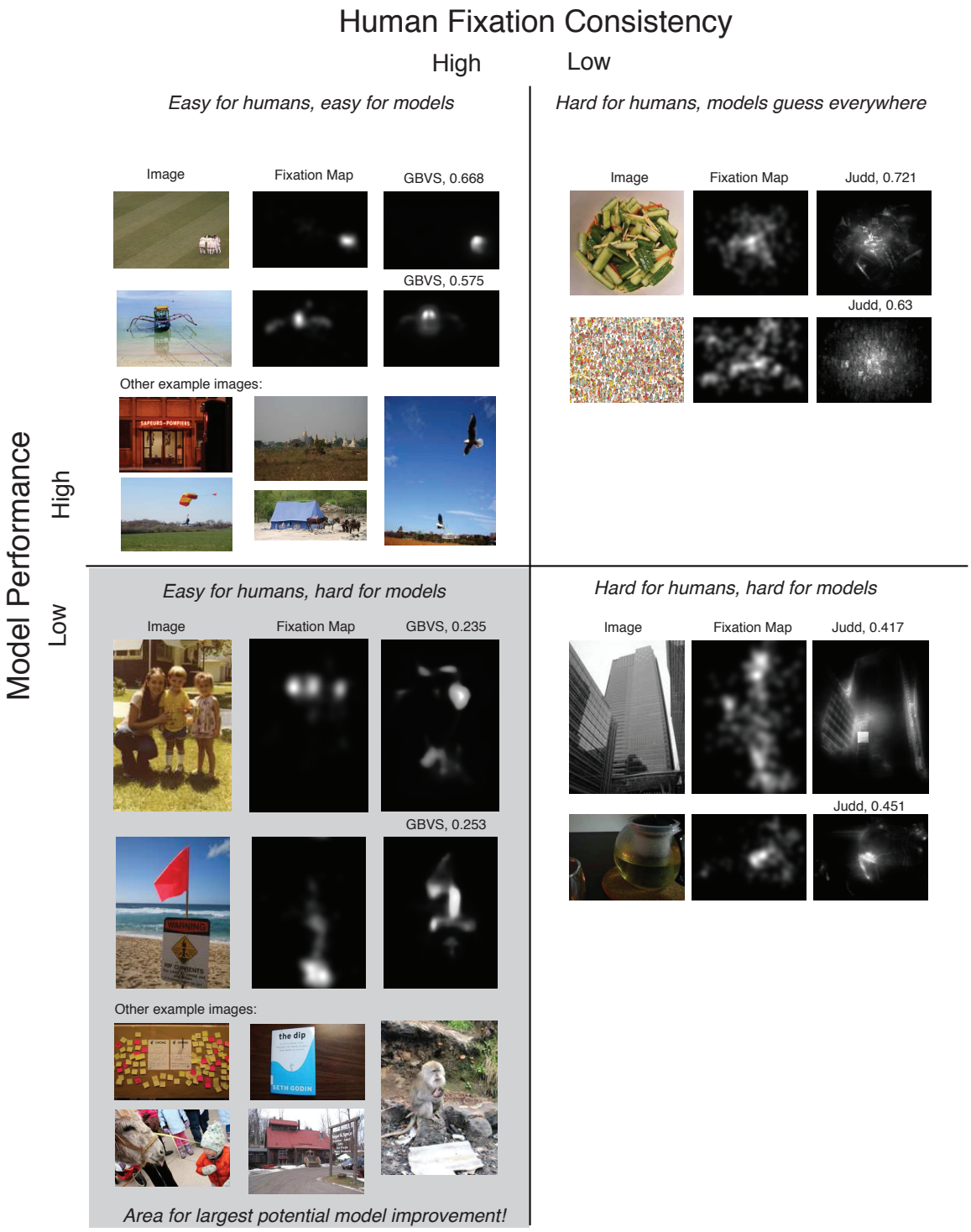

Fig. 12. Model performance versus fixation consistency. For images in the top left quadrant fixations are consistent and saliency models perform well. For images in the bottom left quadrant, fixations are consistent but saliency models miss the salient locations that humans fixate consistently on. These images provide the area of the largest potential improvement for saliency models. Images with non-consistent fixations sometimes get high scores in the similarity and EMD metric as the saliency models predict low saliency spread out across the image and this reflects non-consistent images well. Under the ROC metric, all non-consistent images tend to get low scores.

the hardest images to create saliency maps for.

5.2 How to improve models

Figure 13, which shows sample images from the benchmark data set and their respective fixation maps, demonstrates that observers are attracted to specific higher-level features in images. Our ability to improve the performance of models to predict human fixations lies in our ability to identify these higher-level features:

- Faces. Faces, whether they are human (d) or animal (g), real or cartoon, frontal or sideways (f), masked or not, in focus or blurry (a) or hidden, are all very salient to human observers. Current face detectors are adequate at finding non-blurry, human frontal faces, but miss many others.
- Components of faces. When a face is large, observers tend to be attracted to the eyes, nose and mouth with independent fixations as in image (b).
- Humans. Observers look at humans in images even if they are very small, as in image (c).
- Text. Humans read text when they see it, independent of whether it is clear typed text, painted cursive on a wall (d), small or large, scratched into a rock, or written in sand. Some but not all of these examples could be identified by current text detectors.
- Interactions between objects. When objects in an image interact, humans look at the players of the interaction. For example, observers look at the mailbox, the letter and the face of the woman putting a letter in the mailbox in image (f).
- Horizon. Snapshots taken by human photographers often include a horizon. Many interesting objects lie on the horizon as in image (e).
- Elements of surprise. Observers fixate on objects in unexpected locations as in image (g).

Why do bottom up features alone not capture these? Some could argue that text is salient because of its strong lines or color differential from its background and that this should be captured by low-level features. However, in the merry-go-round image (d), both the oval "0" lights on the attraction and the "Fun Forrest" text both have strong lines and are a different color from their surroundings and could both be considered letters, yet humans focus mostly on the Fun Forrest text. The difference between the two lies more in higher level context of the image.

6 Conclusion

We compare 10 recent models of saliency whose implementations are freely available and measure how well they predict fixations of 39 observers free-viewing 300 images. We report performance under three different metrics. We also provide a way for future models to be added to this comparison through an online benchmarking site.

We find that the Judd et al. [37] and the Graph Based Visual Saliency models [28] consistently outperform other models of saliency and the center model baseline. Other models have some predictive power as they perform better than chance but they often do not outperform the center model. Optimizing the models' blurriness and bias towards the center to models improves many models' performance, but does not improve the performance of the highest performing models.

We plot the models in multidimensional space and find that many models are similar in nature but that the Achanta and Saliency Toolbox implementation of the Itti and Koch model are different and appear as outliers.

We explore the notion of human performance: how well a fixation map from humans predicts fixations from other humans. The fixation map of just one observer has almost as much power as the highest performing saliency maps. As the number of observers for the fixation map increases, predictive performance increases much beyond the performance of models. It increases to a limit: with infinite observers, Similarity would be 1, EMD would be 0 and area under the ROC would be 0.922.

We explore which images models perform poorly on and they tend to end up in two categories: 1) images where human fixations are not consistent in which case even human performance at predicting other's fixations is low, 2) images where human fixations are consistent (observers agree on what is interesting in the image), but saliency models miss it. The second case typically happens when fixations are directed by top-down semantic understanding of the image. Models could be improved if they can identify specific objects of interest to humans: faces, components of faces, people, text, horizons, objects that interact with each other, and elements of surprise. We believe this provides the largest potential improvement for saliency models but is no small goal.

6.1 Main take-aways and recommendations

When creating a saliency model

- The blurriness and centeredness of a model should be optimized. One can use online data sets with public eye tracking data such as the MIT ICCV data set [37] to optimize these parameters.
- Consider that performance is different under different metrics. ROC performance captures whether a map accurately predicts the most fixated locations first. Similarity and EMD performance capture how well a map represents both fixated and non fixated regions. We feel that the ROC metric is adequate as it best captures our goal of predicting fixations.
- Aim to improve model performance on images where human fixations are consistent (observers

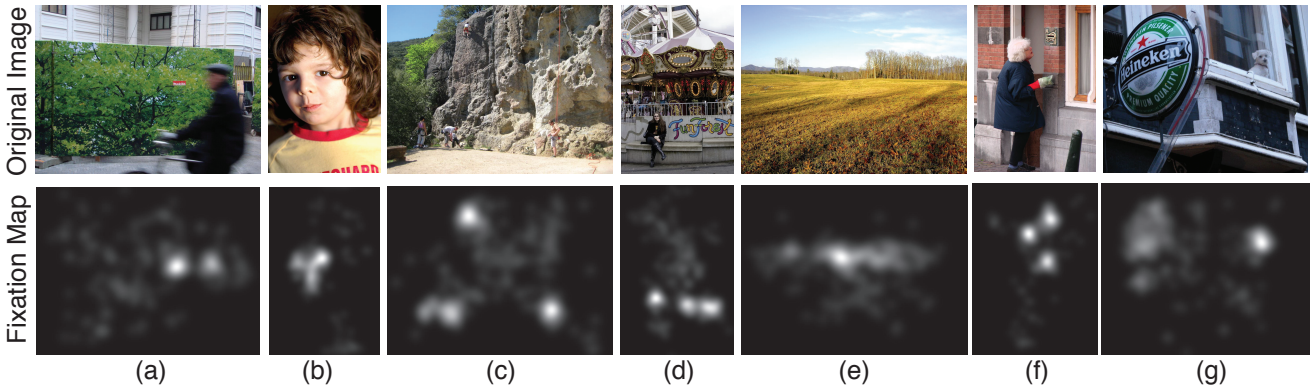

Fig. 13. Sample images from the benchmark data set with their respective fixation map from 39 observers. The fixations maps demonstrate that observers look at faces, components of faces, humans, text, horizons, interactions between objects, and elements of surprise.

agree on what is interesting in the image), as opposed to images where human consistency is low. This provides the largest potential improvement for saliency models.
- To compare a new model with many previous models, upload the saliency maps to our benchmark website. The model's performance will be directly comparable to others' performance.

When fixation information is needed for applications

- Consider again whether it is possible to incorporate eye tracking tests. With just 2 observers, one can get a more accurate prediction of where future observers will look than with the highest performing models right now.
- If constraints do not allow for eye tracking, use the Judd et al., GBVS or Bruce and Tsotsos AIM models to predict the most likely fixated areas. These models perform well and are available online though they vary in computation time. The best performing most recent models should be available on our website.

When running eye tracking tests to obtain ground truth fixation maps

- The more observers used, the more accurate the fixation map. However, using infinite observers is not necessary. 10 observers get $\sim 92\%$ the way to an ideal fixation map under the ROC metric. Our fixation maps with 39 observers reach $\sim 95\%$ of the way to ideal fixation map.
- The accuracy of a fixation map depends on the metric used to measure performance. When using a Similarity or EMD metric, more observers are needed to make an accurate map.

We have chosen to benchmark models by their performance to predict human fixations as saliency maps are used in applications as a proxy for what is important and interesting to humans. However, there are other measures of saliency model performance. One alternative is to measure their usefulness in an application: if the system performance increases in time or quality due to the model then it performs well. Exact correspondence with eye movements is not necessarily important.

ACKNOWLEDGMENTS

This research was supported by NSF CAREER awards 0447561 and IIS 0747120. Frédo Durand acknowledges Royal Dutch Shell and Quanta T-Party. We thank Aude Oliva for the use of her eye tracker.

REFERENCES

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1597 –1604, june 2009.

[2] Radhakrishna Achanta, Francisco Estrada, Patricia Wils, and Sabine Susstrunk. Salient region detection and segmentation. In Antonios Gasteratos, Markus Vincze, and John Tsotsos, editors, *Computer Vision Systems*, volume 5008 of *Lecture Notes in Computer Science*, pages 66–75. Springer Berlin / Heidelberg, 2008.

[3] Tamar Avraham and Michael Lindenbaum. Attention-based dynamic visual search using inner-scene similarity: Algorithms and bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:251–264, 2006.

[4] Tamar Avraham and Michael Lindenbaum. Esaliency (extended saliency): Meaningful attention using stochastic image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:693–708, 2010.

[5] Stephan A. Brandt and Lawrence W. Stark. Spontaneous eye movements during visual imagery reflect the content of the visual scene. *J. Cognitive Neuroscience*, 9:27–38, January 1997.

[6] Neil Bruce and John Tsotsos. Saliency based on information maximization. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 155–162. MIT Press, Cambridge, MA, 2006.

[7] Neil D. B. Bruce and John K. Tsotsos. An attentional framework for stereo vision. In *CRV'05*, pages 88–95, 2005.

[8] Neil D. B. Bruce and John K. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3), 2009.

[9] Moran Cerf, E. Paxon Frady, and Christof Koch. Using semantic content as cues for better scanpath prediction. In *Proceedings of the 2008 symposium on Eye tracking research & applications*, ETRA '08, pages 143–146, New York, NY, USA, 2008. ACM.

[10] Moran Cerf, E. Paxon Frady, and Christof Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, 9(12), 2009.

[11] Moran Cerf, Jonathan Harel, Wolfgang EinhŁuser, and Christof Koch. Predicting human gaze using low-level saliency combined with face detection. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *NIPS*. MIT Press, 2007.

[12] Moran Cerf, Jonathan Harel, Alex Huth, Wolfgang Einhaeuser, and Christof Koch. Decoding what people see from where they look: predicting visual stimuli from scanpaths. *International Workshop on Attention and Performance in Computational Vision*, 2008.

[13] Yun S. Choi, Anthony D. Mosley, and Lawrence W. Stark. String editing analysis of human visual search. *Optometry & Vision Science*, 72(7), 1995.

[14] J. J. Clark and N. J. Ferrier. Modal control of an attentive vision system. In *IEEE International Conference on Computer Vision*, 1988.

[15] R Desimone and J Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1):193–222, 1995.

[16] Bruce A. Draper and Albert Lionelle. Evaluation of selective attention under similarity transformations. *Comput. Vis. Image Underst.*, 100:152–171, October 2005.

[17] Krista Ehinger, Barbara Hidalgo-Sotelo, Antonio Torralba, and Aude Oliva. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, 17:945–978(34), 2009.

[18] Wolfgang Einhäuser, Merrielle Spain, and Pietro Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14), 2008.

[19] Lior Elazary and Laurent Itti. Interesting objects are visually salient. *J. Vis.*, 8(3):1–15, 3 2008.

[20] Lior Elazary and Laurent Itti. A bayesian model for efficient visual search and recognition. *Vision Research*, 50(14):1338 – 1352, 2010. Visual Search and Selective Attention.

[21] Simone Frintrop, Ro Erich, and Henrik I. Christensen. Computational visual attention systems and their cognitive foundations: A survey. *ACM Trans. Appl. Percept.*, 7:6:1–6:39, January 2010.

[22] Dashan Gao, Vijay Mahadevan, and Nuno Vasconcelos. On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, 8(7), 2008.

[23] Dashan Gao and Nuno Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. In *In Proc. NIPS*, pages 481–488, 2004.

[24] Dashan Gao and Nuno Vasconcelos. Integrated learning of saliency, complex features, and object detectors from cluttered scenes. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02*, CVPR '05, pages 282–287, Washington, DC, USA, 2005. IEEE Computer Society.

[25] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. In *CVPR'10*, pages 2376–2383, 2010.

[26] D. M Green and J. A Swets. *Signal detection theory and psychophysics*. John Wiley, 1966.

[27] Chenlei Guo and Liming Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *Trans. Img. Proc.*, 19:185–198, January 2010.

[28] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems 19*, pages 545–552. MIT Press, 2007.

[29] J M Henderson, J R Brockmole, M S Castelhano, and M Mack. Visual saliency does not account for eye movements during visual search in real-world scenes. *Eye Movement Research: Insights into Mind and Brain*, 2007.

[30] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–8, 2007.

[31] L. Itti and P. F. Baldi. Bayesian surprise attracts human attention. In *Advances in Neural Information Processing Systems, Vol. 19 (NIPS*2005)*, pages 547–554, Cambridge, MA, 2006. MIT Press.

[32] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2:194–203, 2001.

[33] Laurent Itti. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12:1093–1123, 2005.

[34] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40:1489–1506, 2000.

[35] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[36] Timothe Jost, Nabil Ouerhani, Roman von Wartburg, and Rene M˙ Assessing the contribution of color in visual attention. *Computer Vision and Image Understanding*, 100(1-2):107 – 123, 2005. Special Issue on Attention and Performance in Computer Vision.

[37] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[38] Christopher Kanan, Mathew H. Tong, Lingyun Zhang, and Garrison W. Cottrell. Sun: Top-down saliency using natural statistics, 2009.

[39] Wolf Kienzle, Felix A. Wichmann, Bernhard Schölkopf, and Matthias O. Franz. A nonparametric approach to bottom-up visual saliency. In Bernhard Schölkopf, John C. Platt, and Thomas Hoffman, editors, *NIPS*, pages 689–696. MIT Press, 2006.

[40] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurbiology*, 4:219–227, 1985.

[41] Olivier Le Meur, Patrick Le Callet, Dominique Barba, and Dominique Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28:802–817, May 2006.

[42] Ian Van Der Linde, Umesh Rajashekar, Alanc. Bovik, and Lawrence K. Cormack. Doves: a database of visual eye movements, 2008.

[43] Tie Liu, Jian Sun, Nan-Ning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1 –8, june 2007.

[44] S K Mannan, K H Ruddock, and D S Wooding. Fixation patterns made during brief examination of two-dimensional images. *Perception*, 26(8):1059–1072, 1997.

[45] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2232 –2239, 29 2009-oct. 2 2009.

[46] Olivier Le Meur, Patrick Le Callet, and Dominique Barba. Predicting visual fixations on video based on low-level visual features. *Vision Research*, 47(19):2483 – 2498, 2007.

[47] R Milanese. *Detecting salient regions in an image: from biological evidence to computer implementation*. PhD thesis, 1993.

[48] V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimal object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2049–2056, New York, NY, Jun 2006.

[49] Vidhya Navalpakkam and Laurent Itti. Modeling the influence of task on attention. *Vision Research*, 45(2):205 – 231, 2005.

[50] Vidhya Navalpakkam and Laurent Itti. Search goal tunes visual features optimally. *Neuron*, 53:605–617, 2007.

[51] A. Oliva, A. Torralba, M.S. Castelhano, and J.M. Henderson. Top-down control of visual attention in object detection. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 1, pages I – 253–6 vol.1, sept 2003.

[52] Nabil Ouerhani. Visual attention: From bio-inspired modeling to real-time implementation. ph.d. thesis, institut de microtechnique universit de neuchtel, switzerland, 2003.

[53] Derrick Parkhurst, Klinton Law, and Ernst Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107 – 123, 2002.

[54] D.J. Parkhurst and E. Niebur. Scene content selected by active vision. *Spatial Vision*, 16:125–154(30), 2003.

[55] Ofir Pele and Michael Werman. A linear time histogram metric for improved sift matching. In *ECCV*, 2008.

[56] Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *ICCV*, 2009.

[57] Robert J. Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397 – 2416, 2005.

[58] M. Posner and Y. Cohen. Components of visual orienting. *In: Attention and Performance X*, pages 531–556, 1984.

[59] Claudio M. Privitera and Lawrence W. Stark. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:970–982, September 2000.

[60] U. Rajashekar, I. van der Linde, A.C. Bovik, and L.K. Cormack. Gaffe: A gaze-attentive fixation finding engine. *Image Processing, IEEE Transactions on*, 17(4):564 –573, april 2008.

[61] Subramanian Ramanathan, Harish Katti, Nicu Sebe, Mohan Kankanhalli, and Tat-Seng Chua. An eye fixation database for saliency detection in images. In *ECCV 2010*, Crete, Greece, 2010.

[62] Albert L. Rothenstein and John K. Tsotsos. Selective tuning: Feature binding through selective attention. In *International Conference on Articial Neural Networks*, pages 548–557, 2006.

[63] Albert L. Rothenstein and John K. Tsotsos. Attention links sensing to recognition. *Image and Vision Computing*, 26(1):114 – 126, 2008. Cognitive Vision-Special Issue.

[64] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth movers distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:2000, 2000.

[65] Hacisalihzade S. S., Allen J.S., and Stark L. Visual perception and sequences of eye movement fixations: A stochastic modelling approach. *IEEE Transactions on Systems, Man and Cybernetics*, 22:474–481, 1992.

[66] Hae Jong Seo and P. Milanfar. Nonparametric bottom-up saliency detection by self-resemblance. *Computer Vision and Pattern Recognition Workshop*, 0:45–52, 2009.

[67] Hae Jong Seo and Peyman Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12), 2009.

[68] Stewart Shipp. The brain circuitry of attention. *Trends in Cognitive Sciences*, 8(5):223 – 230, 2004.

[69] Benjamin W. Tatler. 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), 2007.

[70] Benjamin W. Tatler and B.T. Vincent. The prominence of behavioural biases in eye guidance. *Visual Cognition*, 17(6-7):1029–1054, 2009.

[71] Antonio Torralba, Aude Oliva, Monica S. Castelhano, and John M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766–786, October 2006.

[72] Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97 – 136, 1980.

[73] Po-He Tseng, Ran Carmi, Ian G. M. Cameron, Douglas P. Munoz, and Laurent Itti. Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, 9(7), 2009.

[74] J.K. Tsotsos. Analyzing vision at the complexity level. *Behavioral and Brain Sciences*, 13:423–469, 1990.

[75] John K. Tsotsos. An inhibitory beam for attentional selection. In *Proceedings of the 1991 York conference on Spacial vision in humans and robots*, pages 313–331, New York, NY, USA, 1993. Cambridge University Press.

[76] John K. Tsotsos, Scan M. Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1-2):507 – 545, 1995. Special Volume on Computer Vision.

[77] John K. Tsotsos, Laurent Itti, and G Rees. A brief and selective history of attention. *Neurobiology of Attention*, 2005.

[78] John K. Tsotsos, Yueju Liu, Julio C. Martinez-Trujillo, Marc Pomplun, Evgueni Simine, and Kunhao Zhou. Attending to visual motion. *Computer Vision and Image Understanding*.

[79] John K. Tsotsos, Antonio Jose Rodriguez-Sanchez, Albert L. Rothenstein, and Eugene Simine. Different binding strategies for the different stages ofvisual recognition. In *Proceedings of the 2nd international conference on Advances in brain, vision and artificial intelligence*, BVAI'07, pages 150–160, Berlin, Heidelberg, 2007. Springer-Verlag.

[80] R. Valenti, N. Sebe, and T. Gevers. Image saliency by isocentric curvedness and color. In *IEEE International Conference on Computer Vision*, 2009.

[81] B. M. Velichkovsky, Marc Pomplun, Johannes Rieser, and Helge J. Ritter. *Attention and Communication: Eye-Movement-Based Research Paradigms*. Visual Attention and Cognition. Elsevier Science B.V., Amsterdam, 1996.

[82] Paul Viola and Michael Jones. Robust real-time object detection. In *International Journal of Computer Vision*, 2001.

[83] Dirk Walther and Christof Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395 – 1407, 2006. Brain and Attention, Brain and Attention.

[84] Zheshen Wang and Baoxin Li. A two-stage approach to saliency detection in images. *IEEE International Conference on Acoustics Speech and Signal Processing ICASSP*, pages 965–968, 2008.

[85] Jeremy M. Wolfe, Kyle R. Cave, and Susan L. Franzel. Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, Vol 15(3):419–433, 1989.

[86] A. L. Yarbus. Eye movements and vision. 1967.

[87] Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 2008.

[88] Qi Zhao and Christof Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, 11(3), 2011.

Tilke Judd received a B.S. degree in Mathematics from Massachusetts Institute of Technology (MIT) followed by an M.S. degree in Computer Science and a Ph.D. in Computer Science from MIT in 2007 and 2011 respectively, supervised by Frédo Durand and Antonio Torralba. During the summers of 2007 and 2009 she was an intern with Google and Industrial Light and Magic respectively. She was awarded a National Science Foundation Fellowship for 2005-2008 and a Xerox Graduate Fellowship in 2008.

Frédo Durand is an associate professor in Electrical Engineering and Computer Science at the Massachusetts Institute of Technology, and a member of the Computer Science and Artificial Intelligence Laboratory (CSAIL). He received his PhD from Grenoble University, France, in 1999, supervised by Claude Puech and George Drettakis. From 1999 till 2002, he was a post-doc in the MIT Computer Graphics Group with Julie Dorsey.

Antonio Torralba is an associate professor of Electrical Engineering and Computer Science at the Computer Science and Articial Intelligence Laboratory (CSAIL) at MIT. Following his degree in telecommunications engineering, obtained at the Universidad Politecnica de Cataluña, Spain, he was awarded a Ph.D. in Signal, Image, and Speech processing from the Institut National Polytechnique de Grenoble, France. Thereafter, he spent post-doctoral training at the Brain and Cognitive Science Department and the Computer Science and Articial Intelligence Laboratory at MIT.