Analysis of scores, datasets, and models in visual saliency prediction

Ali Borji[†] Hamed R. Tavakoli[+] Dicky N. Sihite[†] Laurent Itti[†]
[†] Department of Computer Science, University of Southern California, Los Angeles
[+] Center for Machine Vision Research, University of Oulu, Finland

Abstract

Significant recent progress has been made in developing high-quality saliency models. However, less effort has been undertaken on fair assessment of these models, over large standardized datasets and correctly addressing confounding factors. In this study, we pursue a critical and quantitative look at challenges (e.g., center-bias, map smoothing) in saliency modeling and the way they affect model accuracy. We quantitatively compare 32 state-of-the-art models (using the shuffled AUC score to discount center-bias) on 4 benchmark eye movement datasets, for prediction of human fixation locations and scanpath sequence. We also account for the role of map smoothing. We find that, although model rankings vary, some (e.g., AWS, LG, AIM, and HouNIPS) consistently outperform other models over all datasets. Some models work well for prediction of both fixation locations and scanpath sequence (e.g., Judd, GBVS). Our results show low prediction accuracy for models over emotional stimuli from the NUSEF dataset. Our last benchmark, for the first time, gauges the ability of models to decode the stimulus category from statistics of fixations, saccades, and model saliency values at fixated locations. In this test, ITTI and AIM models win over other models. Our benchmark provides a comprehensive high-level picture of the strengths and weaknesses of many popular models, and suggests future research directions in saliency modeling.

1. Introduction

A large number of models has been proposed for predicting where people look in scenes [1]. But due to a lack of an exhaustive coherent benchmarking system, to address several issues such as evaluation measures (e.g., at least 4 types of AUC measures have been used; supplement), center-bias, map characteristics (e.g., smoothing), and dataset bias, a lot of inconsistencies still exist. The discrepancy of results in previous works calls for a unified approach for gauging the progress in this field and for fair comparison of models.

Importance. Modeling visual saliency broadens our understanding of a highly complex cognitive behavior, which may lead to subsequent findings in other areas (object and scene recognition, visual search, etc.) [27][5]. It also benefits many engineering applications (e.g., object detection and segmentation, content-aware image re-targeting, image in-painting, visual tracking, image and video compression, crowds analysis and social gaming [2][6][24][37][30], determining the importance of objects in a scene [48, 44], memorability of image regions [49], and object recall [50].

Our contributions. We offer 3 main contributions: (1) discussing current challenges and directions in saliency modeling (evaluation metrics, dataset bias, model parameters, etc.) and proposing solutions, (2) comparing 32 models and their pros and cons in a unified quantitative framework over 4 widely-used datasets for fixation prediction (on classic and emotional stimuli) as well as scanpath prediction, and (3) stimuli/task decoding using saliency and fixation statistics. Hopefully, our study will open new directions and conversations and help better organize the saliency literature.

Previous benchmarks. Few attempts has been made for saliency model comparison, but their shortcomings have driven us to conduct a new benchmark considering the latest progress. Borji *et al.* [30] compared 36 models over 3 datasets. Judd *et al.* [41] compared 9 models over only 300 images. Some works have compared salient object detection and region-of-interest algorithms [42]. Some other benchmarks have compared models over applications such as image quality assessment [52]. While being very effective, previous comparisons have not correctly addressed all challenging parameters in model accuracy. For example, map smoothing which influences fixation prediction accuracy of models [43], or center-bias (tendency of humans to look towards the center of an image, such that a trivial model that predicts salience near the center may surpass other saliency models) [28, 25], have not been addressed in previous benchmarks. Here, we thoroughly investigate these shortcomings with additional comparison of models over scanpath sequences. We provide the latest update on saliency modeling, with the most comprehensive set of models, challenges/parameters, datasets, and measures.

2. Basic concepts and definitions

Here, we lay out the ground for the rest of the paper and explain some basic concepts of visual attention.

There is often confusion between saliency and atten-

tion. Saliency is a property of the perceived visual stimulus (bottom-up (BU)) or, at most, of the features that the visual system extracts from the stimulus (which can be manipulated by top-down (TD) cues). Attention is a much more general concept that depends on many cognitive factors of very high-level such as strategy for image search and interactions between saliency and search strategy, as well as subjective factors such as age and experience.

A major distinction between mechanisms of visual attention is the bottom-up vs. top-down dissociation. Bottom-up attention is reflexive, fast, likely feed-forward, and mainly deployed by stimulus saliency (e.g., pop-out). On the other hand, top-down attention is deliberative, slow, and powerful with variable selection criteria depending on the task.

Previous studies of attention differ mainly based on the type of stimuli they have employed and the tasks they have addressed. Visual stimuli used in neurophysiological and modeling works include: static (synthetic search arrays involving pop-out and conjunction search arrays, cartoons, or photographs) and over spatio-temporal dynamic stimuli (movies and interactive video games). These stimuli have been exploited for studying visual attention over three types of tasks: (1) free viewing, (2) visual search, and (3) interactive tasks (games or real-world tasks [54]).

What is the unit of attention? Do we attend to spatial locations, objects, or features? [5][27] A great deal of neurophysiological and behavioral evidence exists for all three. Space-based theories claim that humans deliberately attend to spatial locations where a target may appear. Similar observations indicate that visual attention is essentially guided by recognized objects, with low-level saliency contributing only indirectly [50]. Features are well explored by neural studies showing that neurons accommodate their response properties to render an object of interest more salient. Note that these are not exclusive concepts.

A closely related field to saliency modeling is salient region detection. While the goal of the former is to predict locations that grab attention, the latter attempts to segment the most salient object or region in a scene. Evaluation is often done by measuring precision-recall of saliency maps of a model against ground truth data (explicit saliency judgments of subjects by annotating salient objects or clicking on locations). Some models in two categories have compared themselves against each other, without being aware of the distinction and different goals of the models.

The majority of models described by the above concepts have focused on bottom-up, space-based, and static attention for explaining eye movements in free-viewing tasks.

3. Analysis of challenges and open problems

We discuss challenges that have emerged as more models have been proposed. These are open issues not only for research but also for performing fair model comparison.

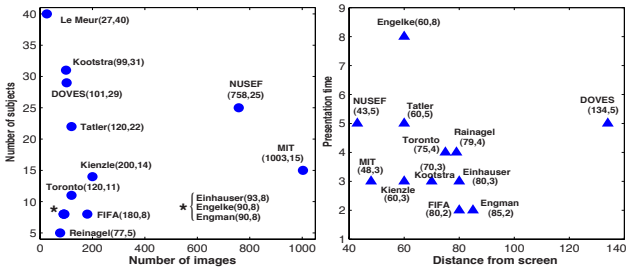

Figure 1. Parameters of 13 major eye movement datasets. The MIT [2] dataset is the largest one with 1003 images. It has a high degree of photographer bias and few number of eye tracking subjects. LeMeur [21] has only 27 images with the highest number of eye-tracking subjects (40). The Toronto dataset [10] has 120 images mainly indoor and in-city scenes. The 5 common objects in Judd dataset are human, face, text, animals, and cars.

Dataset bias. Available eye movement datasets vary on several parameters, for instance: number of images, number of viewers, viewing time per image, subject's distance from the screen, and stimulus variety [55]. Fig. 1 shows some popular eye tracking data sets, among them some are publicly available. Due to small size of the Toronto dataset and small number of subjects, its sole usage is less encouraged. Perhaps the best options so far are NUSEF [4] and Judd [2] datasets. NUSEF dataset (758 images and 25 subjects) contains a large number of affective stimuli making it more suitable for studying semantic attentional cues. As Fig. 1 shows larger fixation datasets with many images and eye-tracking subjects are needed. Because of the specialty of datasets (different optimal weights for features over different datasets [36]), a fair evaluation is to compare models over several datasets as presented in Sec. 4. Further, it has been shown that fixation density maps from different laboratories differ significantly due to inter-laboratory differences and experimental conditions [56].

A difficult challenge in fixation datasets which has affected fair model comparison is "Center-Bias (CB)", whereby humans often appear to preferentially look near an image's center [28]. Two important causes for CB are: (1) Viewing strategy where subjects start looking from the image center and (2) A perhaps stronger, photographer bias, which is the tendency of photographers to frame interesting objects at the center. Annoyingly, due to CB in data, a trivial saliency model that just consists of a Gaussian blob at the center of the image, often scores higher than almost all saliency models [2]. This can be verified from the average eye fixation maps of 3 popular datasets (See supplement). We observed higher central fixation densities for images with objects at the center compared with those with objects off the center. Another problem that is in essence similar to CB is handling invalid filter responses at image borders ("border effect", e.g., AIM model [10]; See [25]).

Some models explicitly (e.g., Judd) or implicitly (e.g., GBVS) have added center-bias (location prior) making fair comparison challenging. Three possible remedies are: (1)

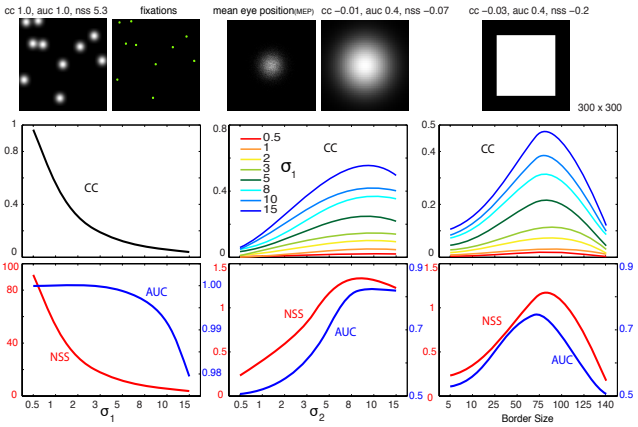

Figure 2. Score analysis. 1^{st} column: scores of a saliency map made by placing a variable Gaussian (σ_1) at fixated locations, 2^{nd} column: scores of the central Gaussian blob (σ_2), and the 3^{rd} column: scores of the image with variable border size. Results are averaged over 1000 runs with 10 randomly generated fixations from a Gaussian distribution to mimic center-bias [28] in data similar to heatmap in Fig. 3. Image size: [300 300].

Every model adds a central Gaussian. This adds Gaussian size and its weight as two additional parameters, (2) Collecting datasets with no CB. This is difficult since, even if we have an approach to uniformly distribute image content, viewing strategy still exists, and (3) Designing suitable metrics which we consider as the most reasonable approach.

Evaluation metrics. Traditionally, saliency models have been evaluated against eye movement datasets. In some cases, accuracy is whether one can predict what changes people will notice, or what they will remember or annotate [1]. We use three popular metrics for saliency evaluation: (1) Correlation Coefficient (CC) between a model (s) and human (h) saliency maps: $CC(s,h) = \frac{cov(s,h)}{\sigma_s \sigma_h}$, (2) Normalized Scanpath Saliency (NSS): the average of saliency values at n fixations in a normalized map ($NSS = \frac{1}{n}\sum_{i=1}^{n} \frac{s(x_h^i, y_h^i) - \mu_s}{\sigma_s}$)[51], and (3) Area Under the ROC Curve (AUC) where human fixations are considered as the positive set and some points from the image are uniformly chosen as the negative set. The saliency map is then treated as a binary classifier to separate the positive samples from negatives. By thresholding over this map and plotting true positive rate vs. false positive rate an ROC curve is achieved and its underneath area is calculated. Please see supplement for a subtle discussion of variations of AUC metrics. KL [9] and earth mover distance (EMD) [36] measures have also been used for model evaluation. Some studies have evaluated the sequence of fixations in scanpath [32, 31].

Fig. 2 shows analysis of how the above scores are affected by smoothness of the saliency map and possible center bias in the reference data. We generated some random eye fixations (sampled from a Gaussian distribution) and made a saliency map by convolving it with a Gaussian filter with variable sigma σ_1. Shown in the 1st column, increasing σ_1, reduces all 3 scores. Over AUC, however, the drop is moderate and the range is very small meaning as long as the hit rates are high, the AUC is high regardless of the false alarm rate [36][1]. Shown in the 2nd column, we placed a Gaussian at the center of the image and calculated the score again by varying the σ_2 of the central Gaussian as well as σ_1 of the Gaussian convolved with fixations (only for CC since for NSS and AUC, fixation positions are used). Increasing σ_2, raises 3 scores up to the maximum match between Gaussian and MEP map and then drops or saturates. CC scores are raised by increasing σ_1. Third column in Fig. 2 shows that by increasing the border size, scores reach a maximum and then drop, a similar effect to center-bias. These analyses show that smoothing the saliency maps and the size of the central Gaussian affect scores and should be accounted for fair model comparison. NSS score is more sensitive to smoothing. All of these scores suffer from center-bias.

Two other issues regarding scores are sensitivity to map normalization (a.k.a re-parameterization) and having well-defined bounds (and chance level). Some scores are invariant to continuous monotonic nonlinearity (e.g., KL) while some others are not (CC, NSS, and AUC). All scores are invariant to saliency map shifting and scaling. Some scores have well defined bounds (CC and AUC have lower and upper bounds) while some do not (KL and NSS; KL has lower bound and NSS has upper bound and chance level of 0).

A proper score for tackling CB is shuffled AUC (sAUC) [25] with the only difference to AUC being that instead of selecting negative points randomly, all fixations over other images are used as the negative set. This score is not affected by σ_2 and border size in Fig. 2. sAUC value for a central Gaussian and a white map is near 0.5 (i.e, fixations from other images as the negative set [28]). When using the method in [25] (i.e., saliency from other images but at fixations of the current image), this type of AUC leads to the exact value of 0.5 for the central Gaussian (See Supp.).

Features for saliency detection. Traditionally, intensity, orientation, and color (in LAB and RGB spaces) have been used for saliency derivation over static images. For dynamic scenes, flicker and motion features have been added. Furthermore, several other low-level features have been used to estimate saliency (size, depth, optical flow, etc.). High-level features (prior knowledge) such as faces [2], people [2], cars [2], symmetry [8], signs, and text [35] have been also incorporated. One challenge is detecting affective (emotional) features and semantic (high-level knowledge) scene properties (e.g., causality, action-influence) which have been suggested to be important in guiding attention (location and fixation duration) [4]. Models usually use all channels for all sorts of stimuli which makes them highly dependent on the false positive rates of employed feature

[1] Note that in [36] and [2], AUC is calculated by thresholding the saliency map and then measuring hit rate which is different from what we (and also [25, 28, 43]) do by spreading random points on the image.

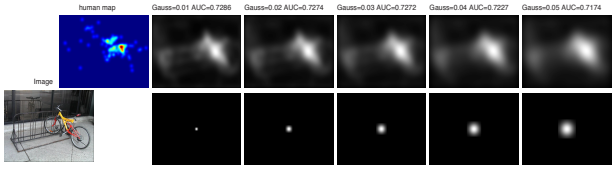

Figure 3. A sample saliency map smoothed by convolving with a variable-size Gaussian kernel (for the AWS model over an image of the Toronto dataset).

detectors (e.g., face or car detector). Since existing models use linear features, they render highly textured regions more salient. Non-linear features (e.g., famous egg in the nest or birthday candle images [25]) has been proposed but has not been fully implemented.

Parameters. Models often have several design parameters such as the number and type of filters, choice of non-linearities, within-feature and across-scale normalization schemes, smoothing, and center-bias. Properly tuning these parameters is important in fair model comparison and is perhaps best left for a model developer to optimize himself.

4. Saliency benchmark

We chose four widely-used datasets for model comparison: Toronto [10], NUSEF [4], MIT [2], and Kootstra [8]. Table 1 shows 30 models compared here. Additionally, we implemented two simple yet powerful models, to serve as baselines: **Gaussian Blob** (Gauss) and **Human inter-observer** (IO). Gaussian blob is simply a 2D Gaussian shape drawn at the center of the image; it is expected to predict human gaze well if such gaze is strongly clustered around the center. For a given stimulus, the human model outputs a map built by integrating fixations from other subjects than the one under test while they watched that same stimulus. The map is usually smoothed by convolving with a Gaussian filter. This inter-observer model is expected to provide an upper bound on prediction accuracy of computational models, to the extent that different humans may be the best predictors of each other. We resized saliency maps to the size of the original images onto which eye movements have been recorded. Please note that, besides models compared here, some other models may exist that might perform well (e.g., [37]), but are not publicly available or easily accessible. We leave them for future investigations.

We first measure how well a model performs at predicting where people look over static image eye movement datasets. We report results using sAUC score as it has several advantages over others. Results over other scores are shown in supplement. Note however, sAUC score alone is not the only criterion for our conclusions as it gives more credit to the off center information and favors true positives more. Next, we compare models for their ability of predicting the saccade sequence. Our conclusions are based on the premise that if a model is good it should perform well over all configurations (i.e., score, datasets, and parameters).

Predicting fixation locations: Model scores and average ranks using sAUC over four datasets are shown in Table 1. We smoothed saliency map of each model by convolving it with a Gaussian kernel (Fig. 3). We then plotted the sAUC of each model over the range of standard deviations of the Gaussian kernel in image width (from 0.01 to 0.13 in steps of 0.01) and calculated the maximum value over this range for each model. Compared with our rankings with original maps (supplement), now some models get a better score. Although the ranking order is not the same over four datasets, some general patterns are noticeable. The Gaussian model is the worst (not significantly better than chance) over all datasets as we expected. There is a significant difference between models and IO model. This difference is more profound over NUSEF and MIT datasets as they contain many stimuli with complex high-level concepts.

AWS model is significantly better than all other models followed by LG, AIM, Global Rarity, Torralba, HouCVPR, HouNIPS, SDSR, and Judd. Over the largest dataset (i.e, MIT), AWS, LG, AIM, Torralba performed better than other models. Over the NUSEF dataset, AIM, LG, Torralba, HouCVPR, HouNIPS models did the best. Kootstra, STB, ITTI (due to different normalization and map sparseness than ITTI98), and Marat ranked at the bottom. Interestingly, AWS model on the Kootstra dataset performs as good as the human IO. Our analyses show that CC, NSS, and AUC produce very high scores for the Gaussian; almost better than all models thanks to its center-preference (see supplement). Therefore, we do not recommend using them for saliency model comparison. Considering rankings sAUC, CC, and NSS scores, we noticed that models that performed well over sAUC are also ranked on top using other scores.

Fig. 4 shows model performance over stimulus categories of the NUSEF dataset for each model and average over all models. There is no significant difference over different categories of stimuli averaged over all models (Inset; See also supplement) although it seems that models perform better over face stimuli and the worst over portrait and nude (this pattern is more clear considering only

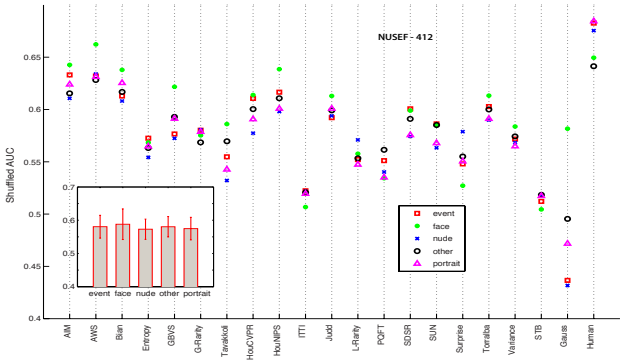

Figure 4. Model performance over categories of the NUSEF dataset. Gauss and Human IO are excluded from the average (i.e., inset). Number of images: Event: 36, Face: 52, Nude: 20, Other: 181, Portrait: 123.

Model	Gaussian-Blob	Inter-observer (IO)	Variance	Entropy	Itti *et al.* (ITTI98)	Itti *et al.* (ITTI)	Torralba	Vocus (Frintrop)	Surprise (Itti & Baldi)	AIM (Bruce & Tsotsos)	Saliency Toolbox (STB)	GBVS (Harel *et al.*)	Le Meur *et al.*	HouCVPR (Hou & Zhang)	Local Rarity (Mancas)	Global Rarity (Mancas)	HouNIPS (Hou & Zhang)	Kootstra *et al.*	SUN (Zhang *et al.*)	Marat *et al.*	PQFT (Guo *et al.*)	Yin Li *et al.*	SDSR (Seo & Milanfar)	Judd *et al.*	Bian *et al.*	ESaliency (Avraham *et al.*)	Yan *et al.*	AWS (Diaz *et al.*)	Jia Li *et al.*	Tavakoli *et al.*	Murray *et al.*	LG (Borji & Itti)	Avg. score over models
Ref.	[28]	-	-	[32]	[3]	[33]	[20]	[6]	[9]	[10]	[24]	[7]	[21]	[11]	[13]	[13]	[12]	[8]	[25]	[23]	[15]	[18]	[22]	[2]	[16]	[14]	[19]	[17]	[26]	[34]	[47]	[38]	-
Year	-	-	-	-	98	00	03	05	05	05	06	06	07	07	07	08	08	08	09	09	09	09	09	09	10	10	10	10	10	11	11	12	-
Code	M	M	C	C	C	C	C	C	M	M	M	S	M	M	M	M	E	M	S	M	M	M	M	M	S	M	E	E	M	11	11	11	-
Category	O	O	I	I	C	C	B	C	B/I	I	C	G	C	S	I	I	I	C	B	C	S	S	I	P	S	G	I	C	B	B	C	I	-
Torronto	.50	**.73**	.66	.65	.63	.62	.69	.66	.63	.69	.62	.65	.66	.69	.65	.69	.69	.61	.67	.64	.68	.69	.69	.68	.61	.65	.68	.72	.67	.64	.64	.70	.66
NUSEF	.49	**.66**	.62	.61	.57	.56	.63	-	.59	.64	.56	.59	-	.63	.60	.62	.63	-	.61	-	.61	-	.61	.61	.63	-	-	.64	-	.56	.57	.63	.60
MIT	.50	**.75**	.65	.64	.62	.61	.67	.65	.63	.68	.58	.64	.57	.65	.63	.67	.65	.60	.65	.62	.66	.65	.65	.66	.61	.62	.64	.69	-	.65	.65	.68	.64
Kootstra	.50	**.62**	.58	.57	.58	.57	.59	.60	.58	.59	.57	.56	.57	.59	.58	.61	.59	.56	.56	.54	.58	.59	.60	.59	.57	.56	.58	.62	.56	-	-	.59	.58
Avg Rank	-	-	4.8	5.8	7.3	8.3	3	4.7	6.8	2.5	8.8	6.5	8	3.5	6	2.8	3.5	9.3	5.3	8	4.3	4	3.8	4	7	5.8	5	1	6.7	6.5	6.7	2.5	-

Table 1. Compared visual saliency models. Abbreviations are: M: Matlab, C: C/C++, E: Executables, S: Sent saliency maps. Note that STB [24] and VOCUS [6] are two implementations of the Itti *et al.* [3] model. Numbers are maximum shuffled AUC scores of models by optimizing the saliency map smoothness (Fig. 3). See supplement for optimal sigma values of Gaussian kernel σ where models take their maximums (σ from $0.01 : 0.01 : 0.13$ in image width). Model category belongs to one of these categories [30]: Cognitive (C), Bayesian (B), Decision-theoretic (D), Information-theoretic (I), Graphical (G), Spectral-analysis (S), Pattern-classification (P), Others (O). We observe lower performance for the STB model compared with either ITTI or ITTI98 models over free-viewing datasets. Thus, it's use instead on the ITTI model (e.g., for model-based behavioral studies) is not encouraged. We employ two different versions of the Itti *et al.* model: ITTI98 and ITTI, which correspond to different normalization schemes. In ITTI98, each feature map's contribution to the saliency map is weighted by the squared difference between the globally most active location and the average activity of all other local maxima in the feature map [3]. This gives rise to smooth saliency maps, which tend to correlate better with noisy human eye movement data. In the ITTI model [33], the spatial competition for saliency is much stronger, and is implemented in each feature map as 10 rounds of convolution by a large difference-of-Gaussians followed by half-wave rectification. This gives rise to much sparser saliency maps, which are more useful than the ITTI98 maps when trying to decide on the single next location to look at (e.g., in machine vision and robotics applications). Kootstra dataset is the hardest one for humans (low IO agreement) and models. Next hardest in the NUSEF dataset. Note that symmetry model of Kootstra can not compete with the other models over the Kootstra dataset although there are many images with symmetric objects in this dataset. Numbers are rounded to their closest value (See supplement for more accurate values). Borji *et al.* [30], results are on original images while here we report optimized results over smoothed images. In our experiments here we used the myGauss=fspecial('gaussian',50,10) which was then normalized to [0 1]. In principle, smoothing Gaussian should be about $1^\circ - 2^\circ$ of the visual field.

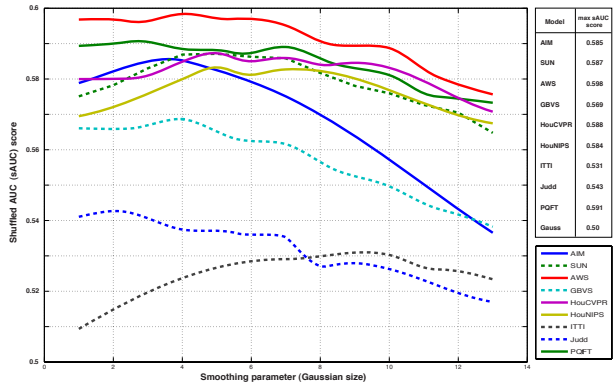

Figure 5. sAUC scores over emotional images of the NUSEF dataset. Results on the right-hand table are the maxima over smoothing range.

top-performing models). The AWS model did the best over all categories. HouNIPS, Judd, SDSR, Yan, and AIM also ranked at the top. Faces are often located at the center while nude and event stimuli are mostly off-center. Humans are more correlated for portrait, event, and nude stimuli. A separate analysis over the Kootstra dataset showed that models have difficulty in saliency detection over nature stimuli where there are less distinctive and salient objects (See supplement). This means that much progress remains to be done in saliency detection over stimuli containing conceptual stimuli (e.g, images containing interacting objects, actions such as grasping, living vs. non-living, object regions inside a bigger object i.e., faces, body parts, etc).

Behavioral studies have shown that affective stimuli influence the way we look at images. Humphrey *et al.* [45] showed that initial fixations were more likely to be on emotional objects than more visually salient neutral ones. Here we take a close look at model differences over emotional (affective) stimuli, using 287 images from the NUSEF belonging to the IAPS dataset [46]. Fig. 5 shows sAUC scores of 10 models over affective stimuli. These values are smaller than the ones for NUSEF (non-affective) shown in Table 1. Our results (using shuffled AUC and with smoothing similar to Table 1; see supplement) suggest that only a fraction of fixations landed on emotional image regions, possibly due to bottom-up saliency (interaction between saliency and emotion; AWS on emotional = 0.59, non-emotional = 0.69). Models AWS, PQFT, and HouNIPS outperform others over these stimuli. These models also performed well on non-emotional stimuli. Fig. 6 shows saliency maps of some emotional images.

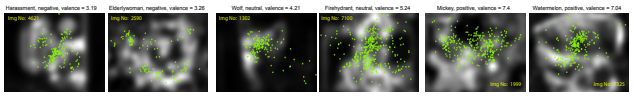

Figure 6. Sample emotional images with positive, negative, and neutral emotional valence from NUSEF dataset along with saliency maps of the AWS model. Note that in some cases saliency misses fixations.

Predicting scanpath: Not only humans are correlated in terms of the locations they fixate, but they also agree somewhat in the order of their fixations [31, 32]. In the context of saliency modeling, few models have aimed to predict scanpath sequence, partly due to difficulty in measuring

and quantizing scanpaths. Here, we first define a measure to quantify scanpath and then compare models in terms of their ability to generate saccades similar to human scanpath. In some applications it is desired to predict gaze one or multiple steps ahead (e.g., in advertising).

Algorithm 1 Scanpath evaluation

- **PHASE 1: generating human scanpath and clusters**
  Input: subjects' fixations.
  Output: fixation clusters & human scanpath strings.
  $BW \leftarrow$ List of all possible band-width values
  for `each band-width` $\in BW$ **do**
    Compute a clustering using the mean-shift algorithm
    Compute interaction between clusters
  end for
  Chose the band-width with the highest interaction rate
  Repeat the clustering for the chosen band-width and store clusters
  Assign a unique character to each cluster
  for `each subject` **do**
    Generate a string for scanpath by finding the closest cluster centers
    Store the string sequence
  end for
  Repeat the above process for all the images in the database

- **PHASE 2: Model evaluation on a given test image**
  Input: model generated saliency maps.
  Output: model's overall score.
  for `each image` **do**
    Take a model saliency map
    Load the sequence strings and clusters corresponding to that image
    for `each subject` **do**
      Generate an equal length sequence running IOR
      Generate a string sequence by finding the closest cluster
      Assign a matching score to the two strings
    end for
    Save the average score over all subjects
  end for
  Report the average matching score over all images

Algorithm 1 shows our method for scanpath evaluation. Shortly, first for an image some clusters are derived from its human fixation map and then scanpath of each subject is coded into a string using these clusters. Then for a saliency map, an inhibition of return (IOR) mechanism is used to generate a sequence. Using Needleman-Wunsch [53] string matching algorithm, a model's scanpath is compared against a subject's scanpath and then the average score over all subjects is calculated (Supplement).

Results of scanpath prediction over Toronto and MIT datasets are shown in Fig. 7. Over the Toronto dataset, Tavakoli, GBVS, and Judd performed significantly better then the other models. Over the MIT dataset, GBVS, Judd, and AWS ranked on top. To investigate the relationship between fixation location and sequence prediction, we plotted these two scores versus each other (Fig. 8). There is a moderate linear correlation between these two scores. While there is not a single winning model using both scores, some models score well using both (e.g., AWS, HouNIPS, AIM, and SDDR). A random white noise map achieves the lowest sequence score (about 0.1 on Toronto and 0.05 on MIT) while a Gaussian model scores very well as it has some overlap with all subject sequences. Fig. 9.b shows sample images and their corresponding sequence measures.

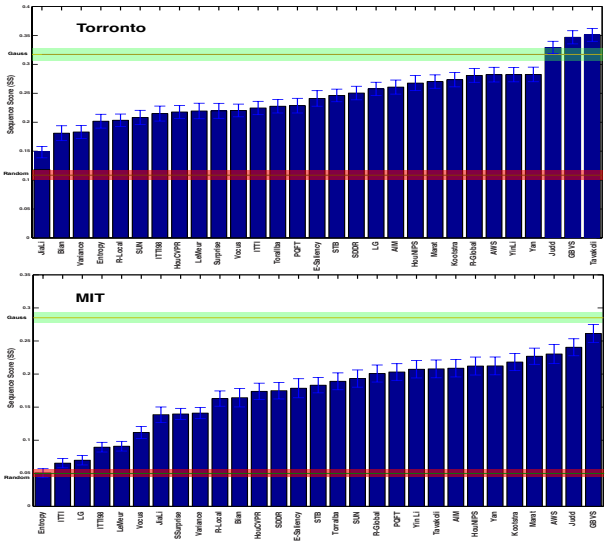

Figure 7. Sequence prediction results over Toronto and MIT datasets.

5. Model comparison over applications

Since models are already good at predicting fixations, some new measures are necessary to draw distinctions among them. An application of saliency modeling is to decode task (illustrated by the classic study of Yarbus [63]), stimulus category, or different populations of human subjects (e.g., control vs. ADHD patients vs. normal subjects, elderly vs. young). Here, as an example we intend to decode the category of the stimulus from features augmented from statistics of fixations, saccades, and saliency at fixations. Note that here we intentionally do not use image features to keep our method general for different applications.

The task is to decode the stimulus category of a subset of images from the NUSEF dataset containing 5 categories over a total of 409 images (Event: 36, Face: 52, Nude: 20, Portrait: 123, Others: 178). For each image we extract a 872 dimensional vector. Features from statistics of **fixations** include: *fixation histogram (256D) and fixation duration histogram (60 bins)*, from **saccades**: *saccade length (50 bins), saccade orientation (36 bins), saccade duration (60 bins), saccade velocity (50 bins), and saccade slope (30 bins)*, from **saliency**: *saliency map (300D), saliency histogram at fixated locations (10 bins), and Top 10 salient locations*. To compute the histograms for a given image, we initially compute corresponding features (e.g., saccade velocity, etc.) for each observer and quantize the values into several bins. Later, the histogram of saccade statistic for an image is computed from all the observers and is L1 normalized. Fixation histogram is made by dividing the image into a grid pattern (16×16) and counting the number of fixations in each grid. Saliency-based features include histogram of saliency values at fixation points (10 bins), vectorized saliency map of size 20×15 (a vector of size 1×300) and coordinates of ten most salient points obtained by applying IOR to the saliency map (a vector of size 1×20).

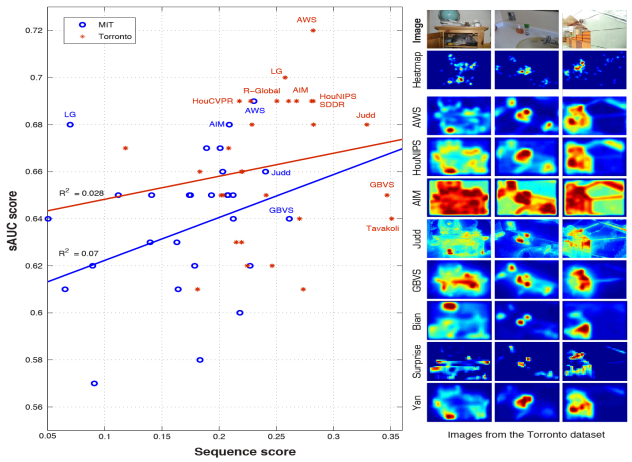

Figure 8. Fixation location prediction vs. sequence prediction accuracies.

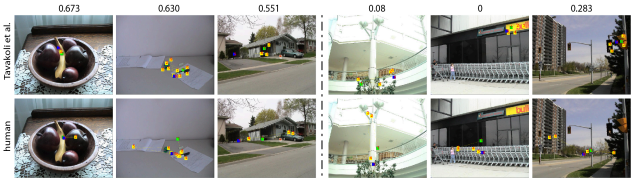

Figure 9. Three best (left columns) and worst (right columns) sequence score example images. Scores are overlaid on top of images. The top (bottom) row shows corresponds to the Tavakoli model (human). Fixation orders are marked by numbers, green and blue rectangles are used to visualize the first and last fixations, respectively.

Classification was performed using a multi-class SVM classifier using binary C-SVC support vector classification with RBF kernel. Initially, we picked 20 images from each scene category and optimized libsvm parameters (but not weights) using a 5-fold cross validation scheme (over 100 runs). Later, we categorized images by a 5-fold cross validation. This process was repeated 10 times. To cope with the imbalanced data over categories, we performed oversampling to reach the same number of samples in each category (i.e., number of samples in the largest class). For measuring Naive Bayes chance, we shuffled class labels and trained the SVM classifier again.

Fig. 10.a shows decoding results using saliency maps of four models and the MEP map. Using all features (and ITTI98 saliency maps), the best decoding accuracy is 0.512 followed by AIM (+ all features = 0.435). Note that the chance level here is 0.2 (Naive Bayes chance: SVM trained with shuffled labels is 0.36). MEP combined with other features scores 0.432 (performance alone is: 0.38) indicating different average fixation locations over stimulus categories. Fig. 10 shows performance of individual features in classification. Fixation histogram, saliency (here MEP), and

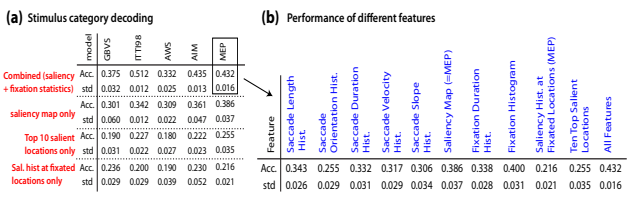

Figure 10. a) Stimulus category decoding for five categories of the NUSEF dataset, b) Breakdown of performance over different features. Saliency map here is the MEP.

saccade length histogram were more useful than the others. Confusion matrices of models show better accuracy over *other*, *portrait*, and *face* stimuli (See supplement).

6. Conclusions and future directions

Learned lessons: Our comparisons[2] show that in general AWS, LG, HouNIPS, Judd, Rarity-G (smoothed version), AIM, and Torralba models performed higher than other models. Note that some recent high performing models (e.g., ΔQDCT [59], CovSal [60]) and benchmarking efforts were not considered here (e.g., [29, 58]). Analysis of scores shows that CC and NSS suffer from center-bias and conclusions should not be based just on them. The shuffled AUC score tackles the center-bias issue with Gaussian model nearly at the chance level (sAUC $\simeq$ 0.5). This comes with the drawback of giving less credit to central fixations. Smoothing saliency maps affects performance and should be taken into account for fair model comparison. We found that some stimulus categories are harder for models (e.g., nature, nude, and portrait) which warrant more attention in future works. Finally, we showed that it is feasible to decode the stimulus category from a feature vector combined from saliency, saccade, and fixation statistics. This could be because of two reasons: similar saliency patterns across scenes of a category and/or systematic/semantic biases of fixations in each category. In this regard, it will also be interesting to test the feasibility of predicting whether a scene is natural or man-made from saliency and fixations.

We believe it is important to constantly measure the gap between the IO model and models to find out in which directions models lag behind human performance. Our analyses over IO model indicate: (1) Observers show poor agreement for complex cluttered scenes such as landscapes, buildings, and street views and (2) Observers show good agreement on scenes with people, animals, faces, and text.

Future directions: Our results show a small gap (but statistically significant) between the best models and human performance (e.g., AWS = 0.72, Human IO = 0.73 on Toronto dataset) indicating a significant progress in this field over the last few years. However, there is a danger that we are fitting ourselves to existing datasets working in a closed world. Thus we believe it is important to gather larger datasets, especially over new stimulus categories. Investigating cases where the current best models fail can help discover additional cues that guide human saccades and should be added to models (e.g., motion, vanishing point, actor and action, focus of expansion, signs, social cues including gaze and hand direction, image clutter, object importance, and text). For example, it has been shown that humans look at the center objects [61, 62]. Some saliency models implicitly emphasize the central parts of objects (e.g, [17]). Explicit central object-bias may lead to even higher fixation predic-

[2]Our online challenge: https://sites.google.com/site/saliencyevaluation/.

tion accuracies. Another direction is finding better ways to combine local and global saliences (i.e., regions with the same global saliency could have different local saliences).

We showed that, from statistics of fixations, saccades, and saliency at fixations, it is possible to decode the stimulus category. Future extensions would be designing standard and more challenging decoding scenarios and considering other features such as saccade sequence (scanpath) information for better discriminating tasks, populations of patients, or stimulus category. Our results are particularly important, if we notice that a recent study [57] reported a failure in decoding subject's task from fixations. Another promising research direction is designing better saliency evaluation scores which: (1) are able to better distinguish fixated vs. non-fixated locations, and (2) are able to discount confounding parameters such as center-bias.

This work was supported by NSF (CCF-1317433 and CMMI-1235539) and ARO (W911NF-11-1-0046 and W911NF-12-1-0433).

References

[1] A. Borji and L. Itti. State-of-the-art in Visual Attention Modeling. *PAMI*, 2013.

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

[3] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. PAMI*, 20(11):1254-1259, 1998.

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

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

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

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

[8] G. Kootstra, A. Nederveen, and B. de Boer. Paying attention to symmetry. *British Machine Vision Conference*, 2008.

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

[10] N.D.B. Bruce and J.K. Tsotsos. Saliency based on information maximization. *Advances in Neural Information Processing Systems (NIPS)*, 2005.

[11] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. *Computer Vision and Pattern Recognition (CVPR)*, 2007.

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

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

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

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

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

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

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

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

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

[21] O. Le Meur, P. Le Callet, and D. Barba. Predicting visual fixations on video based on low-level visual features. *Vision Research*, 47(19), 2007.

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

[23] S. Marat, T. Ho-Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué. Modeling saliency to predict gaze direction for short videos. *IJCV*, 2009.

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

[25] 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.

[26] J. Li, Y. Tian, T. Huang, and W. Gao. Probabilistic multi-task learning for visual saliency estimation in video. *IJCV*, 90(2):150-165, 2010.

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

[28] 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): 2007.

[29] O. Le Meur and T. Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavioral Research*, 2012.

[30] A. Borji, D. N. Sihite, and L. Itti. Quantitative Analysis of Human-Model Agreement in Visual Saliency Modeling: A Comparative Study. *IEEE Transactions on Image Processing*, 2012.

[31] D. Noton and L. Stark. Scanpaths in Saccadic Eye Movements While Viewing and Recognizing Patterns. *Vision Research*, 11, 929-942, 1971

[32] C.M. Privitera and L. Stark. Algorithms for defining visual regions-of-interest: comparison with eye fixations. *PAMI*, 2000.

[33] L. Itti, L. and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40, 2000.

[34] H.R. Tavakoli, E. Rahtu, and J. Heikkilä. Fast and efficient saliency detection using sparse sampling and kernel density estimation. *SCIA*, 2011.

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

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

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

[38] A. Borji and L. Itti. Exploiting Local and Global Patch Rarities for Saliency Detection. *CVPR*, 2012.

[39] A. Borji, Boosting Bottom-up and Top-down Visual Features for Saliency Estimation, *CVPR*, 2012.

[40] J. B. Huang and N. Ahuja, Saliency Detection via Divergence Analysis: A Unified Perspective, *ICPR*, 2012.

[41] T. Judd, F. Durand, and A. Torralba. A Benchmark of Computational Models of Saliency to Predict Human Fixations. *MIT technical report*, 2012.

[42] T. Ho-Phuoc, L. Alacoque, A. Dupret, A. Guerin-Dugue, and A. Verdant. A unified method for comparison of algorithms of saliency extraction, *SPIE*, 2012.

[43] X. Hou, J. Harel, and C. Koch. Image Signature: Highlighting Sparse Salient Regions. *PAMI*, 2012.

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

[45] K. Humphrey, G. Underwood, and T. Lambert. Salience of the lambs: A test of the saliency map hypothesis with pictures of emotive objects, *J. of vision*, 2012.

[46] P. Lang, , M. Bradley, and B. Cuthbert. (iaps): Affective ratings of pictures and instruction manual. *Technical report*, University of Florida, 2008.

[47] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga. Saliency Estimation Using a Non-Parametric Low-Level Vision Model. *CVPR*, 2011.

[48] M. Spain and P. Perona. Measuring and predicting object importance. *International Journal of Computer Vision*, 2010.

[49] A. Khosla, J. Xiao, A. Torralba, and A. Oliva. Memorability of Image Regions. *NIPS*, 2012.

[50] W. Einhäuser, M. Spain, and P. Perona. Objects Predict Fixations Better Than Early Saliency. *Journal of Vision*, 2008.

[51] R.J. Peters, A. Iyer, L. Itti, C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 2005.

[52] M.S. Gide and L.J. Karam. Comparative evaluation of visual saliency models for quality assessment task, *6th Int. Workshop on Video Processing*, 2012.

[53] S.B. Needleman, C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Molecular Biology*, 1970.

[54] A. Borji, D. N. Sihite, and L. Itti. Probabilistic Learning of Task-Specific Visual Attention, *CVPR*, 2012.

[55] S. Winkler and R. Subramanian, Overview of Eye tracking Datasets, *Quality of Multimedia Experience (QoMEX*, 2013.

[56] U. Engelke, H. Liu, J. Wang, P. Le Callet, I. Heynderickx, H.J Zepernick, and A. Maeder, A Comparative Study of Fixation Density Maps, *IEEE T. IP*, 2013.

[57] M.R. Greene, T. Liu, J.M. Wolfe, Reconsidering Yarbus: A failure to predict observers' task from eye movement patterns, *Vision Research*, 2012.

[58] J.B Huang and N. Ahuja, Saliency Detection via Divergence Analysis: A Unified Perspective, *International Conference on Pattern Recognition*, 2012.

[59] B. Schauerte and R. Stiefelhagen, Quaternion-based spectral saliency detection for eye fixation prediction, *ECCV*, 2012.

[60] E. Erdem and A. Erdem, Visual saliency estimation by nonlinearly integrating features using region covariances, *Journal of Vision*, 13:4, 1-20, 2013.

[61] A. Nuthmann and J. M. Henderson, Object-based attentional selection in scene viewing. *Journal of Vision*, 10(8):20, 1-19, 2010.

[62] A. Borji, D. N. Sihite, and L. Itti. Objects do not predict fixations better than early saliency: A re-analysis of Einhäuser et al.'s data, *Journal of Vision*, 2013.

[63] A., Yarbus, Eye movements and vision.. New York: Plenum, 1967.