Salient Object Detection: A Benchmark

Ali Borji, Dicky N. Sihite, and Laurent Itti

Department of Computer Science, University of Southern California
http://ilab.usc.edu/~borji

Abstract. Several salient object detection approaches have been published which have been assessed using different evaluation scores and datasets resulting in discrepancy in model comparison. This calls for a methodological framework to compare existing models and evaluate their pros and cons. We analyze benchmark datasets and scoring techniques and, for the first time, provide a quantitative comparison of 35 state-of-the-art saliency detection models. We find that some models perform consistently better than the others. Saliency models that intend to predict eye fixations perform lower on segmentation datasets compared to salient object detection algorithms. Further, we propose combined models which show that integration of the few best models outperforms all models over other datasets. By analyzing the consistency among the best models and among humans for each scene, we identify the scenes where models or humans fail to detect the most salient object. We highlight the current issues and propose future research directions.

1 Introduction

Visual saliency is the ability of a vision system (human or machine) to select a certain subset of visual information for further processing. This mechanism serves as a filter to select only the *interesting* information related to current behaviors or tasks to be processed while ignoring irrelevant information.

Recently, salient object detection has attracted a lot of interest in computer vision as it provides fast solutions to several complex processes. Firstly, it detects the most salient and attention-grabbing object in a scene, and then it segments the whole extent of that object. The output usually is a map where the intensity of each pixel represents the probability of that pixel belonging to the salient object. This problem in its essence is a segmentation problem but slightly differs from the traditional general image segmentation. While salient object detection models segment only the salient foreground object from the background, general segmentation algorithms partition an image into regions of coherent properties. Salient object detection methods also differ from other saliency models that aim to predict scene locations where a human observer may fixate. Since saliency models, whether they address segmentation or fixation prediction, both generate saliency maps; they are interchangeably applicable.

The value of saliency detection methods lies in their applications in many fields including: object detection and recognition [30][9], image compression [59], video

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part II, LNCS 7573, pp. 414–429, 2012.

summarization [60], and photo collage [20][17]. A comparison of some image retargeting techniques (some based on salient object detection [13][35]) is available at: http://people.csail.mit.edu/mrub/retargetme/.

Some work has been published on quantitative evaluation of general segmentation algorithms [44][48][1]. To the authors' best knowledge, such attempt for benchmarking salient object segmentation methods has not been reported. Unfortunately, these methods have often been evaluated on different datasets, which in some cases are small and not easily accessible. The lack of published benchmarks causes discrepancy in quantitative comparison of competing models. Not only does a benchmark allow researchers to compare their models with other algorithms, but it also helps identify the chief factors affecting performance. This could result in an even faster performance improvement.

2 Related Works

Here we provide a short summary of the main trends in saliency detection. The interested reader can refer to the extensive reviews for more details [67][61].

As a pioneer, Itti *et al.* [2] derived bottom-up visual saliency using center-surround differences across multi-scale image features. Ma and Zhang [51] proposed an alternative local contrast method using a fuzzy growth model. Harel *et al.* [3] used graph algorithms and a measure of dissimilarity to achieve efficient saliency computation with their Graph Based Visual Saliency (GBVS) model. Liu *et al.* [33] used conditional random field to learn regions of interest using three features: multi-scale contrast, center-surround histogram, and color spatial-distribution. More recently, Goferman *et al.* [15] simultaneously modeled local low-level clues, global considerations, visual organization rules, and high-level features to highlight salient objects along with their contexts. Zhai and Shah [12] defined pixel-level saliency by contrast to all other pixels. However, for efficiency they used only luminance information, thus ignoring distinctive clues in other channels. Achanta *et al.* [16] proposed a frequency-tuned method that directly defines pixel saliency using the color differences from the average image color. Visual saliency is equated to discrimination in [62] and extended to bottom-up mechanism in the pre-attentive biological vision. Spectral components in an image have been explored to detect visual saliency [5][34][56]. In Hou and Zhang [5], the gist of the scene is represented with the average Fourier envelope and the differential spectral components are used to extract salient regions. This is replaced by the phase spectrum of the Fourier transform in [34] because it is more effective and computationally efficient. Some researchers including Bruce and Tsotsos [4] and Zhang *et al.* [10] attempted to define visual saliency based on information theory. Some others have further used graph-cut or grab-cub algorithms to refine borders of their saliency maps and count for salient object contours [23][13]. While some methods define visual saliency in a local way (e.g., Itti *et al.* [2], SEO [8], GBVS [3], AWS [7], and DAKlein [25]),

[1] http://www.wisdom.weizmann.ac.il/$\sim$vision/Seg_Evaluation_DB

some others are based on global rarity of image regions over the entire scene (e.g., AIM [4], SUN [10], HouNIPS [6], HouCVPR [5], and RC [13]).

Object-based theories of attention propose that humans attend to objects and high-level concepts. Inspired by these cognitive findings, some models (e.g., Judd *et al.* [1]) have used object detectors such as faces, humans, animals, and text to detect salient locations. Some models address saliency detection in the spatio-temporal domain by employing motion, flicker, optical flow (e.g., [65]), or interest points learned from the image regions at fixated locations (e.g., [63]). Since the research in this area is rather new and the few existing models are in their early phases, thus we leave their quantitative evaluation for the future.

Recently a new trend called *active visual segmentation* has emerged with the intention to segment a region that contains a fixation point (Mishra *et al.* [49]). Their framework combines monocular cues (color/intensity/texture) with stereo and/or motion, in a cue-independent manner. Similarly, Siagian and Koch [50] also proposed a new approach for active segmentation by combining boundary and region information. We consider adding them to promote this new trend.

3 The Salient Object Detection Benchmark

Saliency Detection Models. In this work, as the initial seed, we focus on those models that are easily accessible, attained good accuracies, or have been highly referred. Software for some models was already available online. For others, we contacted their creators for the code; the authors then either sent us the source code to compile or the executables. Some authors, however, preferred to run their models on our stimuli and hence send us back the saliency maps. In order to achieve a thorough model comparison, we intend to open an online challenge where modelers could contribute by submitting their results.

We compare three categories of models: 1) those aiming to detect and segment the most salient object in a scene (emphasized more here), 2) active segmentation approaches, and 3) models that address fixation prediction. Table 1 shows the list of models from the first two categories, and table 2 shows category 3.

Datasets. We choose 5 benchmark datasets based on the following criteria: 1) being widely-used, 2) having size and stimulus variety, and 3) containing different biases such as number of annotators, number of salient objects, and center-bias. Due to specialty of various datasets, it is likely that model rankings may differ across datasets. Hence, to come up with a fair comparison, it is recommended to run models over several datasets and draw objective conclusions. A model is considered to be good, if it performs well over almost all datasets.

Fig. 1 provides explanation of the 5 datasets used here as well as sample images of the five smallest and largest objects from each. Fig. 2 shows *Mean Annotation Position*: $MAP = (1/UV)\sum_{u=1}^{U}\sum_{v=1}^{V} s_{uv}$ averaged over U images and annotated bounding boxes of V subjects (s_{uv}). There is a strong center-bias in the single-object datasets, most probably due to the tendency of photographers to frame interesting objects at the image center [11]. Similarly, there are two peaks (at the left and the right) in the images with two salient objects

Table 1. Compared salient object detection models (checked) sorted chronologically. Abbreviations: {M: Matlab, C: C/C++, S: Sent saliency maps}. w and h: image width/height. DB shows the datasets that we have results over them. JiaLiSal is applied to 100 and 1000 images of ASD and MSRA, respectively. $max\ X$: Preserve the aspect ratio while resizing the bigger dimension to X.

#	**Acronym** (Model)	**Ref.**	**Pub/Year**	**Code**	**Resolution**	**DB**	Avl.
1	**IO**: Inter-observer model	-	-	M	$w \times h$	All	✓
2	**MAP**: Mean Annotation Position	-	-	M	500×500	All	✓
3	**MZ**: Ma and Zhang	[51]	ACM-M/2003	S	$w \times h$	ASD	✓
4	**LC**: Zhai and Shah	[18]	ACM-M/2006	C	$w \times h$	All	✓
5	**salLiu**: Liu *et al.*	[33]	CVPR/2007	M	max 200	All	✓
6	**AC**: Achanta *et al.*	[14]	ICVS/2008	M	$w \times h$	All	✓
7	**MSSS**: Achanta and Süsstrunk	[55]	ICIP/2009	M	$w \times h$	All	✓
8	**FTS**: Achanta *et al.*	[16]	CVPR/2009	M	$w \times h$	All	✓
9	**EDS**: Rosin	[19]	PR/2009	C	$w \times h$	All	✓
10	Gopalakrishnan *et al.*	[34]	CVPR/2009	-	-	-	-
11	Marchesotti *et al.*	[35]	ICCV/2009	-	-	-	-
12	**Valenti**: Valenti *et al.*	[40]	ICCV/2009	S	$w \times h$	ASD/MSRA	✓
13	**Goferman**: Goferman *et al.*	[15]	CVPR/2010	M	max 250	All	✓
14	**PMehrani**: Mehrani and Veksler	[23]	BMVC/2010	S	$w \times h$	ASD/SED1	✓
15	Rahtu *et al.*	[29]	ECCV/2010	-	-	-	-
16	Khuwuthyakorn *et al.*	[28]	ECCV/2010	-	-	-	-
17	Zhang *et al.*	[21]	IEEE TOM/2010	-	-	-	-
18	**JiaLiSal**: Jia Li *et al.*	[36]	IJCV/2010	S	$[w\ h]/16$	ASD/MSRA	✓
19	**LiuICIP**: Liu *et al.*	[53]	ICIP/2010	S	$w \times h$	ASD	✓
20	**MichalGazit**: Gazit *et al.*	[37]	ECCV-W/2010	M	$w \times h$	All	✓
21	**DAKlein**: Klein and Frintrop	[25]	ICCV/2011	S	$w \times h$	All	✓
22	**MengW**: M. Wang *et al.*	[18]	CVPR/2011	S	$w \times h$	ASD	✓
23	Feng *et al.*	[22]	ICCV/2011	-	-	-	-
24	Deng and Luo	[39]	OE/2011	-	-	-	-
25	Lu *et al.*	[24]	ICCV/2011	-	-	-	-
26	L. Wang *et al.*	[26]	ICCV/2011	-	-	-	-
27	**SVO**: Chang *et al.*	[27]	ICCV/2011	M	$w \times h$	All	✓
28	**CBsal**: Jiang *et al.*	[31]	BMVC/2011	M	$w \times h$	All	✓
29	**RC**: M.M. Cheng *et al.*	[13]	CVPR/2011	C	$w \times h$	All	✓
30	**HC**: M.M. Cheng *et al.*	[13]	CVPR/2011	C	$w \times h$	All	✓
31	**Materias**: Li *et al.*	[36]	BMVC/2011	M	$w \times h$	All	✓
32	**LiuIETIP**: Liu *et al.*	[42]	IEEE TIP/2011	S	$w \times h$	ASD	✓
33	**Mishra**: Mishra *et al.*	[49]	PAMI/2011	C	$w \times h$	All	✓
34	**SRS1**: Siagian and Koch	[50]	Submitted.	C	$w \times h$	All	✓

Table 2. Compared saliency models originally developed for eye fixation prediction

#	**Acronym** (Model)	**Ref.**	**Pub/Year**	**Code**	**Resolution**	**DB**	Avl.
1	**ITTI**: Itti *et al.*	[2]	PAMI/1998	C	$w/16 \times h/16$	All	✓
2	**ITTI98**: Itti *et al.* (maxNorm)	[2]	PAMI/1998	C	$w/16 \times h/16$	All	✓
3	**AIM**: Bruce and Tsotsos	[4]	NIPS/2005	M	$w/2 \times h/2$	All	✓
4	**GBVS**: Harel *et al.*	[3]	NIPS/2006	M	$w \times h$	All	✓
5	**HouCVPR**: Hou and Zhang	[5]	CVPR/2007	M	64×64	All	✓
6	**HouNIPS**: Hou and Zhang	[6]	NIPS/2008	M	$w \times h$	All	✓
7	**SUN**: Zhang *et al.*	[10]	JOV/2008	M	$w/2 \times h/2$	All	✓
8	**PQFT**: Guo and Zhang	[56]	TIP/2009	M	400×400	All	✓
9	**SEO**: Seo and Milanfar	[8]	JOV/2009	M	$w \times h$	All	✓
10	**AWS**: Diaz *et al.*	[7]	ACIVS/2009	M	$w/2 \times h/2$	All	✓
11	**Judd**: Judd *et al.*	[1]	ICCV/2009	M	$w \times h$	All	✓

(SED2). Fig. 2.a shows entropy of images. ASD and MSRA contain more cluttered scenes. Histogram of normalized object sizes (object size/image size) is plotted in Fig. 2.b. It shows there are few images with large objects in these datasets. Objects usually range from small to medium size, about 30% of the whole image (with MSRA containing larger objects on average). Fig. 2.c shows subject agreement for an image which is defined as:

$$r = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{|s_i \cap s_j|}{|s_i \cup s_j|} \quad (1)$$

where s_i and s_j are annotations of the $i-th$ and $j-th$ subjects (of n subjects). Above score has the well-defined lower bound of 0 when there is no overlap in segmentations of users and upper-bound of 1 when they have perfect overlap. As Fig. 2 shows, subjects have higher agreement over MSRA, SED1, and SED2 datasets (about 70% of r values are above 0.9) compared to the SOD dataset.

Proposed Combined Models. Since different models are based on different hypotheses and algorithms, it is likely that combining evidences from them may enhance the saliency detection accuracy. Here, we investigate such an idea. Let $p(x_f)$ represent the probability of an image pixel x being part of the salient foreground object (i.e., 2D normalized saliency map [0 1]). Let $p(x_f|M_i)$ be such evidence from the $i-th$ model. Assuming independence among models (i.e., $p(x_f|M_i, M_j) \propto p(x_f|M_i)p(x_f|M_j)$), then a Naive Bayesian evidence accumulation would be:

$$p(x_f|M_1, M_2, \cdots, M_K) \propto \frac{1}{Z} \prod_{k=1}^{K} p(x_f|M_k) \quad (2)$$

where K is the number of models and Z is chosen in a way that the final map is a probability density function (pdf). Since, a very small value by only a single model suppresses all evidences from the other models in the multiplication case (Eq. 2), we also consider another combination scheme using linear summation:

$$p(x_f|M_1, M_2, \cdots, M_K) \propto \frac{1}{Z} \sum_{k=1}^{K} \mathcal{G}\big(p(x_f|M_k)\big) \quad (3)$$

where $\mathcal{G}(x)$ is one of three combination functions: 1) *identity* (*i.e.*, $\mathcal{G}(x) = x$), 2) $exp(x)$, and 3) $-1/log(x)$. We chose exp and $-1/log$ functions to avoid negative values and weigh the highly salient regions more, assuming that models are more accurate at the peaks of their saliency maps.

Evaluation Scores. Similar to [16], we calculate the precision-recall (PR) curve by varying a threshold on the intensity values [0:0.05:1] and generating a binary saliency map. Since MSRA dataset has bounding boxes, we first fit a rectangle to the thresholded saliency map, fill it, and then calculate scores using bounding boxes. We also report the F-Measure defined as: $F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 Precision + Recall}$.

Here, as in [16] and [13], we use $\beta^2 = 0.3$ to weigh precision more than recall. We also calculate the ROC and AUC (Area Under ROC Curve) results from *true positive rates* and *false positive rates* obtained during the calculation of PR.

4 Experiments and Results

Baseline Models. We implemented two simple yet powerful baseline models: 1) MAP explained in Sec. 3 and Fig. 2, and 2) Human Inter-observer (IO) 'model' which is the aggregated map of annotations by other subjects (excluding the one under test) for each image. The IO model provides an upper bound for other models since humans usually agree in annotating the most salient object.

Fig. 1. Sample images from the datasets. Top row shows the five smallest objects and bottom row shows the five largest objects from each dataset. 1) **ASD** [16]: This dataset contains 1,000 images from the MSRA dataset. Authors have manually segmented the salient object (contour) within the user-drawn rectangle to obtain binary masks. Link: `http://ivrgwww.epfl.ch/$\sim$achanta/`. 2) **MSRA** [33]: This dataset (part B of the original dataset) includes 5,000 images containing labeled rectangles from nine users drawing a rectangular shape around what they consider the most salient object. There is a large variation among images including natural scenes, animals, indoor, outdoor, resolution, etc. Link: `http://research.microsoft.com/en-us/um/people/jiansun/SalientObject/salient_object.htm` . 3) **SED** [47]: This dataset contains two parts. The first one, *single object database*(**SED1**), has 100 images containing only one salient object similar to the ASD. But in the second one, *two objects database*(**SED2**), there are two salient objects in each image (100 images). Our purpose in employing this dataset is to evaluate the accuracy of models when there is more than one object in a scene is to evaluate accuracy of models over more complex stimuli. It is still not clear how the models developed over single-object datasets will scale up in more general cases. Each of one-object and two-object datasets contain 100 images. Link: `http://www.wisdom.weizmann.ac.il/$\sim$vision/Seg_Evaluation_DB`. 4) **SOD** [46]: This dataset is a collection of salient object boundaries based on the Berkeley Segmentation Dataset (BSD). Seven subjects are asked to choose the salient object(s) in 300 images. This dataset contains many images with several objects making it challenging for models. Link: `http://elderlab.yorku.ca/$\sim$vida/SOD/index.html`. Another interesting dataset called *Imgsal* contains both human fixations and tagging of objects with different sizes. Link: `http://www.cim.mcgill.ca/$\sim$lijian/database.htm`.

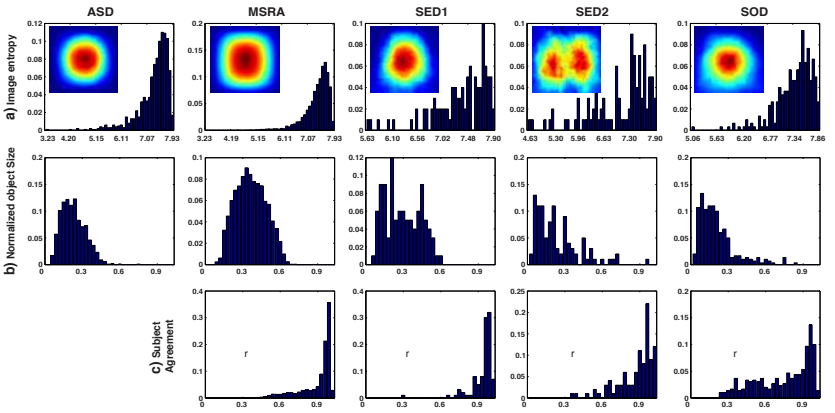

Fig. 2. a) histogram of image entropy (inset: location prior MAP), b) distribution of normalized object sizes, c) distribution of annotators agreement ($0 \leqslant r \leqslant 1$). Note that the ASD dataset has only one annotator.

Fig. 3 shows precision-recall, F-measure, ROC curves, and AUC values for all models over five datasets. Model rankings using F-measure usually do not match with the other scores (although very similar). Therefore, we focus on drawing conclusions based on ROC, AUC, and PR scores as they are more consistent.

Accuracy of Baseline Models. There is still a large gap between existing models and the IO over all datasets which indicates room for improvement. The IO model performs lower on the SOD. This also could be verified from Fig. 2 where subjects show less agreement since there is no unique salient object in many images. MAP model stands somewhere in the middle, better than some models indicating that a simple center-biased model could capture a lot of salient regions. MAP performs the lowest on the SED2 because two objects do not always happen at exact locations of the two left and right peaks of the MAP model. Random predictor, a map with the value for each pixel taken uniformly random from the range [0 1], scores 0.5 on AUC and provides a theoretical lower-bound for models. All models perform above chance level in all cases.

Model Rankings. Table 3 shows the best four models from salient object detection and fixations prediction categories over 5 datasets. Based on average rankings using AUC, SVO, Goferman, CBsal, and RC rank as the top four in order. For fixation prediction, on average the four best models are: GBVS, AIM, HouNIPS, and AWS. Among salient object detection models, SRS1, Michal-Gazit, Mishra, and EDS usually rank at the bottom over datasets. LC and AC models also perform poorly using AUC and PR. SalLiu is a good model in terms of AUC score and F-measure but not PR. This might be due to the binary nature of its saliency maps. Goferman is usually good using AUC but not as good using PR. Low performance of MichalGazit is probably due to its very sparse maps. MAP model does well over SOD dataset indicating existence of center-bias. Our results show that SOD is the most difficult dataset for models (they perform

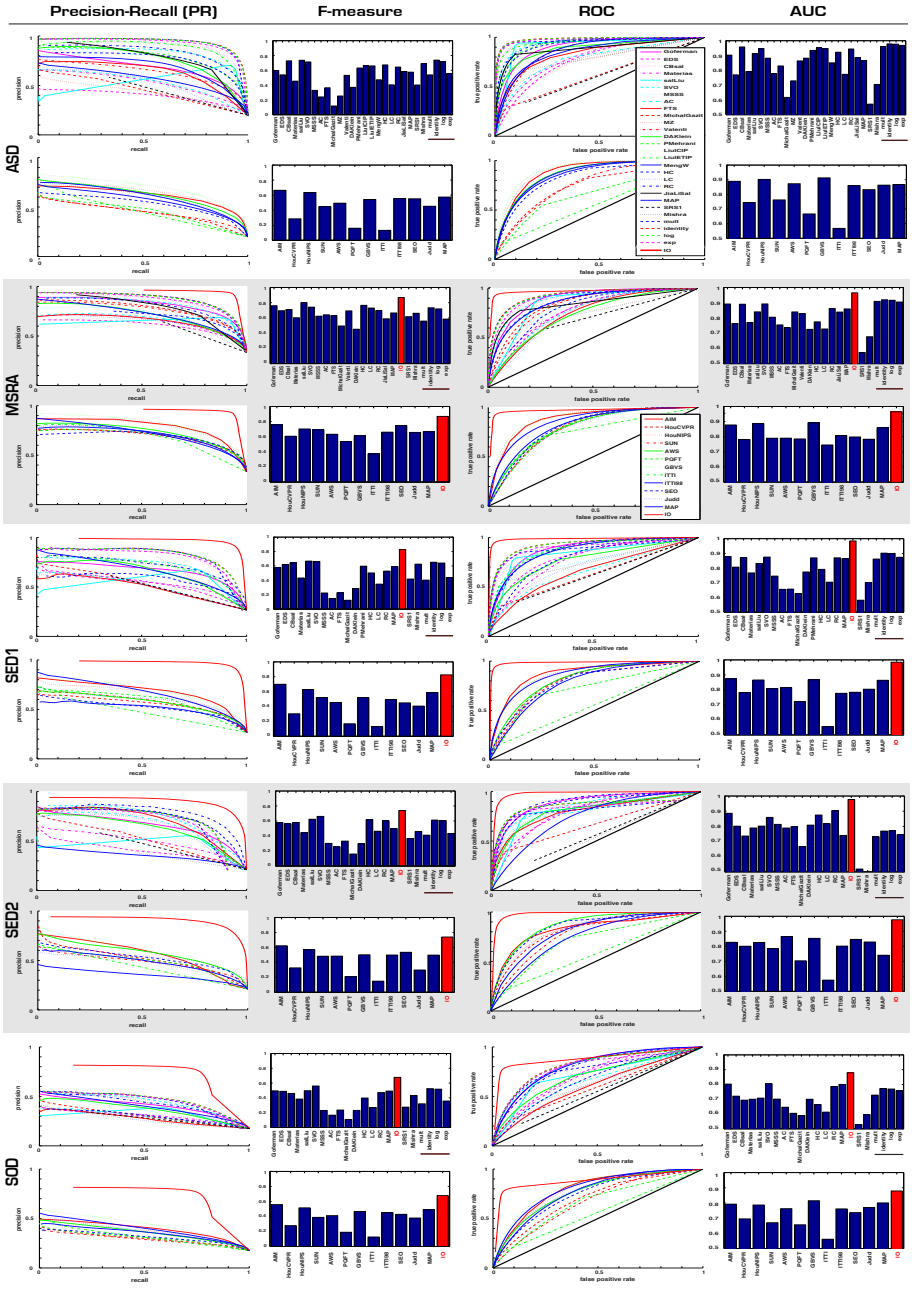

Fig. 3. PR, F-measure, ROC, and AUC scores of saliency models: Top row for each dataset shows saliency segmentation and bottom row shows fixation prediction models. *mult* indicates multiplication of models, and *identity*, *log*, *exp* indicate linear summation of identity, logarithmic, and exponential functions, respectively (see text).

poorly and score on a narrow band) while ASD and MSRA (the biggest dataset with rectangular annotations) datasets are the easiest ones.

MAP model ranks in the middle among fixation prediction models. On SOD, MAP model works very well, right below the best model, repeatedly indicating high center bias in this dataset (Fig. 2). Using F-measure AIM consistently ranks above other models. ITTI and PQFT rank at the bottom over all datasets. The ITTI98 model is based on the same principles of the ITTI model but uses *maxNorm* normalization: For each feature map, find the global max M and the average m of all other local maxima. Then, weigh the map by $(M-m)^2$. Please refer to [67] for a full investigation of fixation prediction models.

Table 3. Model rankings over categories and datasets using AUC. Gof. = Goferman.

#	Salient object detection models					Fixation prediction models				
	ASD	**MSRA**	**SED1**	**SED2**	**SOD**	**ASD**	**MSRA**	**SED1**	**SED2**	**SOD**
1	CBsal	CBsal	Gof.	RC	SVO	GBVS	GBVS	AIM	AWS	GBVS
2	LiuICIP	SVO	SVO	Gof.	Gof.	HouNIPS	HouNIPS	GBVS	GBVS	MAP
3	SVO	Gof.	CBsal	HC	MAP	AIM	AIM	MAP	SEO	AIM
4	LiuIETIP	RC	PMehrani	SVO	RC	AWS	MAP	HouNIPS	AIM	HouNIPS

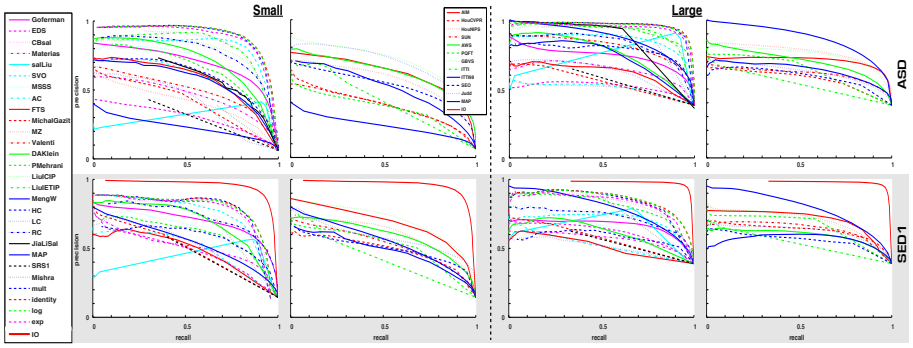

Fig. 4. Accuracy of models over small and large objects from ASD and SED1 datasets

Models built originally for fixation prediction, on average, perform lower than models specifically built to detect and segment the most salient object in a scene. Best fixation prediction models perform better than poor saliency detection models. Why does performance accuracy of the two categories of approaches differ over segmentation datasets? The reason lies on the amount of true positives vs. false positives. Segmentation approaches try to generate white salient regions to capture more of the true positives. On the other hand, fixation prediction models are very selective and generate few false positives (there are not many fixations on the image). In a separate study, we noticed that fixation prediction models perform better than the saliency detection models over eye fixation datasets

(for fixation prediction) [67]. Active segmentation algorithms score consistently below the other two categories. The main reason is the dependency of these models on the initial seed which sometimes may not happen on the most salient object due to the spatial outliers in the image.

Accuracy of Combined Models. Our combined models (using CBsal, SVO, and RC) score the highest in many cases supporting our claim in evidence integration. Overall, our combined models rank the best in the following order: *identity*, *log*, *exp*, and *mult*. Over **SED2** and **SOD** datasets, combined models perform lower compared to single object datasets but still outperform many models. According precision-recall curves, our models stand on top (except SED2). This is because of CBsal model perform poorly on these datasets causing the performance to drop for combined models. Note that our selection of which models to combine was purely based on the ASD dataset and not by over-fitting results to all datasets. It is possible that combining best models over each dataset will outperform all models over that dataset. We found that the combination of the best two models (CBsal and SVO) still works as good as (slightly below) combining the three best models (supplement).

The Role of Object Size. It is more challenging to obtain high precision-recall for small objects than large objects [16][25]: an algorithm that selects the whole image obtains 80% precision with 100% recall if an object occupies 80% of the image. We compare accuracy of models over 100 images of ASD (50 for SED1) with the smallest and 100 (50 for SED1) images with the largest objects (Fig. 1 shows samples). On average, the object occupied 6.2% (15% for SED1), 38.2% (39% for SED1) of the image area, respectively for small and large objects. The resulting PR curves for ASD and SED1 datasets are shown in Fig. 4 (see supplement for other DBs). Models (specially MAP) score higher on large objects. IO scores higher on large objects thus showing higher subject agreement. Combined models still perform higher than other models in both cases (difference is more pronounced over small objects). Good models (e.g., CBsal, SVO, and LiuIETIP) still perform well with the exception of SVO that shows a noticeable performance drop over large objects. Fixation prediction model rankings differ over both cases. While GBVS is the best over small objects, HouNIPS (over ASD) and AIM (over SED1) are the best over large objects. MAP model outperforms all fixation prediction models over large objects.

The Role of Annotation Consistency. To check whether annotation consistency affects accuracy, we selected (according to Eq. 1) 100 most and 100 least consistent images (50 for SED datasets) of all datasets and calculated the scores shown in Fig. 6. As expected, IO and MAP models perform very high over most consistent images (on average for all other models). Combined models are at the top in both cases (except SED2 in the least consistent case).

Easy/Difficult Stimuli for Models. Here, we study the easiest and most difficult scenes for 11 models that did well in all cases. For each model, we sorted the stimuli based on their AUC score and chose the top (easiest) and bottom (hardest) five images (supplement). We noticed that models have many easy and difficult scenes in common (Fig. 7). The top five stimuli usually have one

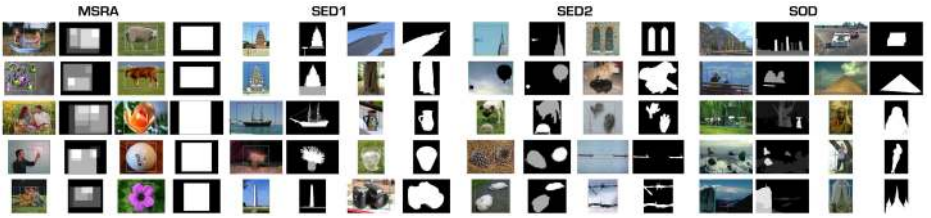

Fig. 5. Five images with least (and most) annotation consistency from datasets

vivid salient object at the center with a distinctive color from the background. The bottom five stimuli often contain objects in a textured background, objects composed of several different parts, or objects that attract top-down attention (e.g., text, faces, human, social interactions, gaze direction, or animals).

Analysis of Map Smoothing and Center-Bias. Here, we investigate the role of map smoothing (blurring) as it has been shown to affect scoring for fixation prediction in the past [10]. We convolve saliency maps of models with variable-sized Gaussian kernels and calculate the scores. We also add (separately) central Gaussian kernels to the saliency maps. AUC scores are shown in Fig. 8. With smoothing, scores slightly change but qualitative trends and model rankings stay the same, hence not affecting our conclusions. The reason why smoothing changes fixation prediction but not salient object detection accuracy is because: 1) there is uncertainty in fixations such that they often do not land on the exact intended locations, and 2) in salient object detection, scores are calculated using image regions while in fixation prediction, they are calculated by sampling maps from eye positions. Shown in Fig. 2, all datasets have center-bias similar to the eye movement datasets ([67][11]). From Fig. 8 (right side), we conclude that adding center-bias raises the accuracy of low-performing models while it decreases the accuracy of good models. However, this change in accuracy is not significant and does not alter model rankings.

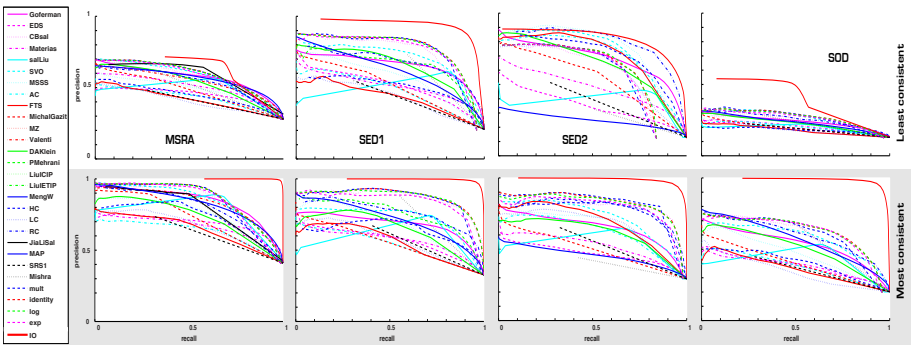

Fig. 6. Accuracy of saliency detection models over least and most consistent images

Fig. 7. Left) Easiest stimuli for 11 best models (Fig. 3), Right) Most difficult stimuli

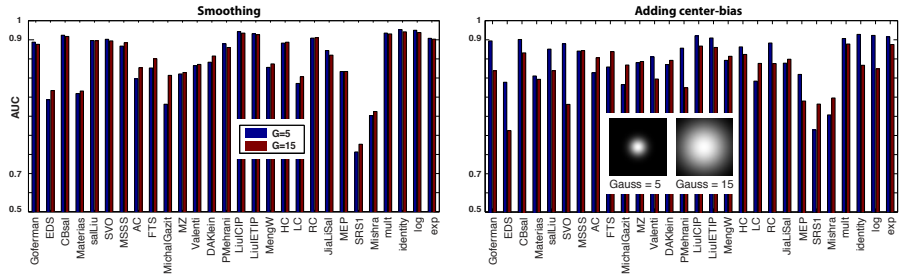

Fig. 8. Left: scores of models for maps smoothed with 2 Gaussian kernel sizes over ASD dataset. Right: model accuracy by adding Gaussian center-bias to center of a map.

Analysis of Model Similarity. Following Judd [64], here we measure the similarity among salient object detection models. We build a similarity matrix Q with the element ij measured as follows. For each image of a dataset, calculate the correlation coefficients (CC) of saliency maps between models i and j. Then the q_{ij} will be the average of all CCs over all images. Over the ASD dataset, (HC, RC), (LiuICIP, LiuIETIP), (MSSS, FTS), (SVO, Goferman, Valenti, AWS), (HouNIPS, AIM, GBVS) are the most similar ones to each other. We calculate dissimilarity as 1 minus similarity (thus ranging between [-1 0]), and use multi-dimensional scaling (MDS) analysis to represent models in a 2D space (Fig. 9 for ASD). AC, MZ, ITTI, and SRS1 models are most different from the others (see supplement).

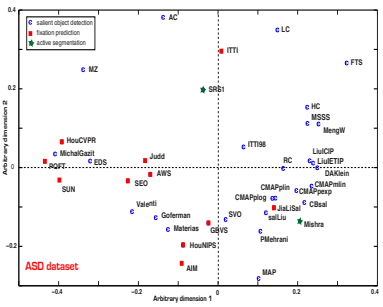

Fig. 9. Measuring model similarities

Analysis of Human-Model Agreement. Finally, over the MSRA dataset, we plot in Fig. 10 human vs. model agreement (5 best models including: CBsal, SVO, LiuICIP, LiuIETIP, and Goferman) for each image. Aligning with Fig. 2, most of the subjects show annotation consistency while models are less consistent. We inspect cases of agreement/disagreement between and within humans and models: 1) Images that both humans and models agree, usually have one clear object with different features from the background rendering them bottom-up salient. 2) Images that humans and models disagree (within the same group),

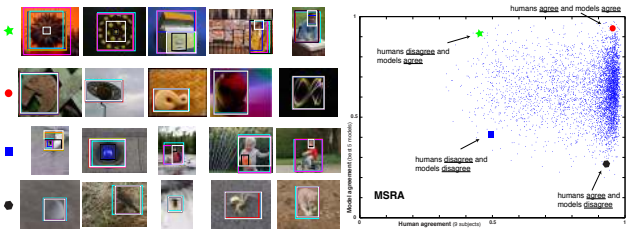

Fig. 10. Analysis of human model agreement over MSRA dataset

usually are not easy to define the most salient object. Salient objects have several parts in a crowded background. The most interesting case is when humans (within) disagree with models agree (and vice versa). 3) Images for which models disagree usually have textured backgrounds with salient objects sharing similar features with the background. However, this does not disrupt the top-down mechanisms by which humans decide the most salient object. 4) Images that humans disagree but models agree, usually contain salient objects with multiple parts and different features from the background. This makes detecting the salient object with several parts easy for models but difficult for humans. Overall, there are not many cases for which humans disagree.

5 Conclusions

Based on the extensive experiments over 5 datasets, we conclude that SVO [27], Goferman [15], CBsal [31], RC [13], and Liu *et al.* [53][52] (LiuICIP and LiuIETIP) work better than the others overall. Salient object detection models (Table 1) perform better than fixation prediction models (Table 2). Map smoothing is not a big challenge to scoring as opposed to the fixation prediction. SOD has 300 images and is the hardest dataset for models which encourages further effort in the future. Many models share the easiest and the most difficult stimuli. Although model rankings remain the same over small and large objects, model accuracy is higher over large objects. Further, models work better on the most consistent images. Scenes containing objects in textured and cluttered backgrounds are challenging for many models. There are cases where the level of agreement on salient object is low among humans but high for the models. Analysis of model similarity shows that some models use different saliency detection mechanisms. Focusing on these models may inspire further development.

We showed that a simple integration scheme generalizes well across datasets. It suggests using more sophisticated combination techniques (e.g., by selecting and combining different models) which may enhance accuracies. While most of the models try to correctly segment the object regions, only recently researchers (e.g., [46]) have started to account for image boundary. Contemporary saliency detection datasets suffer from the drawback that they contain images with few (oftentimes one) close-up objects. Future work should investigate collecting datasets with more variety including cluttered scenes with multiple objects. Majority of existing models are only applicable to static images. Further research is needed to scale up current models or build new models in

the spatio-temporal domain. Models have different implementation languages (C++, Matlab) which makes analysis of computational complexity challenging. While previous progress is promising, further work is needed to bridge the existing gap among current models and human performance. To ease this process and initiate a collaborative effort, we share our results, data, and software in our website.

References

1. Judd, T., Ehinger, K., Durand, F.: Learning to predict where humans look. In: ICCV (2009)
2. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. PAMI 20(11), 1254–1259 (1998)
3. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: NIPS (2006)
4. Bruce, N.D.B., Tsotsos, J.K.: Saliency based on information maximization. In: NIPS (2005)
5. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: CVPR (2007)
6. Hou, X., Zhang, L.: Dynamic attention: Searching for coding length increments. In: NIPS (2008)
7. Garcia-Diaz, A., Fdez-Vidal, X.R., Pardo, X.M., Dosil, R.: Decorrelation and Distinctiveness Provide with Human-Like Saliency. In: Blanc-Talon, J., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2009. LNCS, vol. 5807, pp. 343–354. Springer, Heidelberg (2009)
8. Seo, H.J., Milanfar, P.: Static and space-time visual saliency detection by self-resemblance. Journal of Vision 9, 1–27 (2009)
9. Walther, D., Koch, C.: Modeling attention to salient proto-objects. Neural Net. (2006)
10. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: SUN: A Bayesian framework for saliency using natural statistics. JOV (2008)
11. Tatler, B.W.: 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)
12. Zhai, Y., Shah, M.: Visual attention detection in video sequences using spatiotemporal cues. ACM Multimedia (2006)
13. Cheng, M.M., Zhang, G.X., Mitra, N.J., Huang, X., Hu, S.M.: Global contrast based salient region detection. In: CVPR (2011)
14. Achanta, R., Estrada, F.J., Wils, P., Süsstrunk, S.: Salient Region Detection and Segmentation. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) ICVS 2008. LNCS, vol. 5008, pp. 66–75. Springer, Heidelberg (2008)
15. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. In: CVPR (2010)
16. Achanta, R., Hemami, S., Estrada, F., Süsstrunk, S.: Frequency-tuned salient region detection. In: CVPR (2009)
17. Wang, J., Sun, J., Quan, L., Tang, X., Shum, H.Y.: Picture collage. In: CVPR (2006)
18. Wang, M., Konrad, J., Ishwar, P., Jing, Y., Rowley, H.: Image saliency: from intrinsic to extrinsic context. In: CVPR (2011)
19. Rosin, P.L.: A simple method for detecting salient regions. Pattern Rec. (2009)

20. Goferman, S., Tal, A., Zelnik-Manor, L.: Puzzle-like collage. In: EuroGraphics (2010)
21. Zhang, W., Wu, Q.M.J., Wang, G., Yin, H.: An adaptive computational model for salient object detection. IEEE Trans. on Multimedia 12(4) (2010)
22. Feng, J., Wei, Y., Tao, L., Zhang, C., Sun, J.: Salient object detection by composition. In: ICCV (2011)
23. Mehrani, P., Veksler, O.: Saliency segmentation based on learning and graph cut. In: BMVC (2010)
24. Lu, Y., Zhang, W., Lu, H., Xue, X.: Salient object detection using concavity context. In: ICCV (2011)
25. Klein, D.A., Frintrop, S.: Center-surround divergence of feature statistics for salient object detection. In: ICCV (2011)
26. Wang, L., Xue, J., Zheng, N., Hua, G.: Automatic Salient object extraction with contextual cue. In: ICCV (2011)
27. Chang, K.Y., Liu, T.L., Chen, H.T., Lai, S.H.: Fusing generic objectness and visual saliency for salient object detection. In: ICCV (2011)
28. Khuwuthyakorn, P., Robles-Kelly, A., Zhou, J.: Object of Interest Detection by Saliency Learning. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 636–649. Springer, Heidelberg (2010)
29. Rahtu, E., Kannala, J., Salo, M., Heikkilä, J.: Segmenting Salient Objects from Images and Videos. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 366–379. Springer, Heidelberg (2010)
30. Kanan, C., Cottrell, G.: Robust classification of objects, faces, and flowers using natural image. In: CVPR (2010)
31. Jiang, H., Wang, J., Yuan, Z., Liu, T., Zheng, N., Li, S.: Automatic salient object segmentation based on context and shape prior. In: BMVC (2011)
32. Li, J., Tian, Y., Huang, T., Gao, W.: A dataset and evaluation methodology for visual saliency in video. In: Int. Conf. on Multimedia and Expo., pp. 442–445 (2009)
33. Liu, T., Sun, J., Zheng, N., Tang, X., Shum, H.: Learning to detect a salient object. In: CVPR (2007)
34. Gopalakrishnan, V., Hu, Y., Rajan, D.: Random walks on graphs to model saliency in images. In: CVPR (2009)
35. Marchesotti, L., Cifarelli, C., Csurka, G.: A framework for visual saliency detection with applications to image thumbnailing. In: ICCV (2009)
36. Li, J., Levine, M.D., An, X., He, H.: Saliency detection based on frequency and spatial domain analysis. In: BMVC (2011)
37. Holtzman-Gazit, M., Zelnik-Manor, L., Yavneh, I.: Salient edges: A multi scale approach. In: ECCV, Workshop on Vision for Cognitive Tasks (2010)
38. Luo, Y., Yuan, J., Xue, P., Tian, Q.: Saliency Density Maximization for Object Detection and Localization. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part III. LNCS, vol. 6494, pp. 396–408. Springer, Heidelberg (2011)
39. Deng, Q., Luo, Y.: Edge-based method for detecting salient objects. Opt. Eng. 50 (2011)
40. Valenti, R., Sebe, N., Gevers, T.: Image saliency by isocentric curvedness and color. In: ICCV (2009)
41. Zhang, Q., Liu, H., Shen, J., Gu, G., Xiao, H.: An improved computational approach for salient region detection. Journal of Computers (2010)
42. Li, H., Ngan, K.N.: A co-saliency model of image pairs. IEEE Trans. Image Process (2011)
43. Ge, F., Wang, S.: New benchmark for image segmentation evaluation. Journal of Electronic Imaging 16(3) (2007)

44. Estrada, F.J., Jepson, A.D.: Benchmarking image segmentation algorithms. IJCV (2009)
45. Ancuti, C.O., Ancuti, C., Bekaert, P.: CVPR (2011)
46. Movahedi, V., Elder, J.H.: Design and perceptual validation of performance measures for salient object segmentation. In: POCV (2010)
47. Alpert, S., Galun, M., Basri, R., Brandt, A.: Image segmentation by probabilistic bottom-up aggregation and cue integration. In: CVPR (2007)
48. Ge, F., Wang, S., Liu, T.: Image-segmentation evaluation from the perspective of salient object extraction. In: CVPR (2006)
49. Mishra, A.K., Aloimonos, Y., Fah, C.L., Kassim, A.: Active visual segmentation. IEEE Trans. PAMI (2011)
50. Siagian, C., Koch, C.: Salient segmentation using contours and region growing (submitted)
51. Ma, Y.F., Zhang, H.J.: Contrast-based image attention analysis by using fuzzy growing. ACM Multimedia, 374–381 (2003)
52. Liu, Z., Xue, Y., Yan, H., Zhang, Z.: Efficient saliency detection based on Gaussian models. IET Image Processing 5(2), 122–131 (2011)
53. Liu, Z., Xue, Y., Shen, L., Zhang, Z.: Nonparametric saliency detection using kernel density estimation. In: ICIP, pp. 253–256 (2010)
54. Li, J., Tian, Y., Huang, T., Gao, W.: Probabilistic multi-task learning for visual saliency estimation in video. IJCV 90(2), 150–165 (2010)
55. Achanta, R., Susstrunk, S.: Saliency detection for content-aware image resizing. In: ICIP (2009)
56. Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and Its applications in image and video compression. IEEE Trans. on Image Processing (2010)
57. Huang, T.H., Cheng, K.Y., Chuang, Y.Y.: A collaborative benchmark for region of interest detection algorithms. In: CVPR (2009)
58. Masciocchi, C.M., Mihalas, S., Parkhurst, D., Niebur, E.: Everyone knows what is interesting: salient locations which should be fixated. Journal of Vision (2009)
59. Itti, L.: Automatic Foveation for Video Compression using a neurobiological model of visual attention. IEEE Trans. Image Process (2004)
60. Ma, Y., Hua, X., Lu, L., Zhang, H.: A generic framework of user aattention model and its application in video summarization. IEEE Trans. Multimedia (2005)
61. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Trans. Pattern Anal. Mach. Intell (2012)
62. Gao, D., Vasconcelos, N.: Discriminant saliency for visual recognition from cluttered scenes. In: NIPS (2004)
63. Kienzle, W., Franz, M.O., Schölkopf, B., Wichmann, F.A.: Center-surround patterns emerge as optimal predictors for human saccade targets. J. Vision (2009)
64. Judd, T.: Understanding and predicting where people look. Phd Thesis, MIT (2011)
65. Itti, L., Dhavale, N., Pighin, F.: SPIE (2003)
66. Koehler, K., Guo, F., Zhang, S., Eckstein, M.: Vision Science Society (2011)
67. Borji, A., Sihite, D.N., Itti, L.: Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. IEEE Trans. Image Processing (2012)