Salient Object Detection: A Benchmark

Ali Borji, Dicky N. Sihite, and Laurent Itti
University of Southern California, Computer Science, Neuroscience

http://ilab.usc.edu/publications/doc/Borji_etal12eccv.pdf
Supported by the National Science Foundation, the General Motors Corporation, and the Army Research Office

(1) Introduction

* 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.

* 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.

(2) Benchmark

Saliency detection models:

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 shows sample images from used datasets.

Evaluation Scores:

Similar to [1], 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.

(3) Models & Data

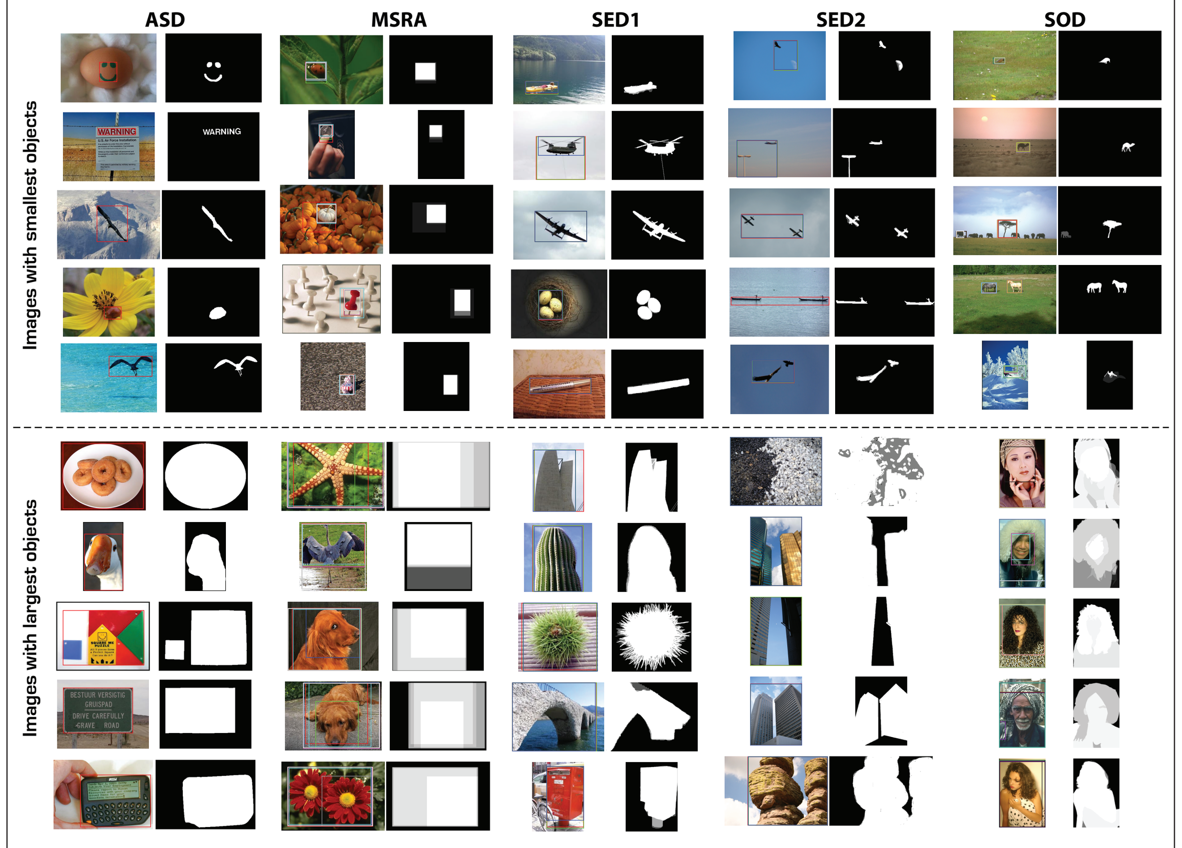
Table 1. Compared salient object detection models (checked) sorted chronologically. Abbreviations: {M: Matlab, C: C/C++, S: Sent saliency maps}. w and h: image width/height. DB shows the datasets that we have results over them. JiaLiSal is applied to 100 and 1000 images of ASD and MSRA, respectively.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

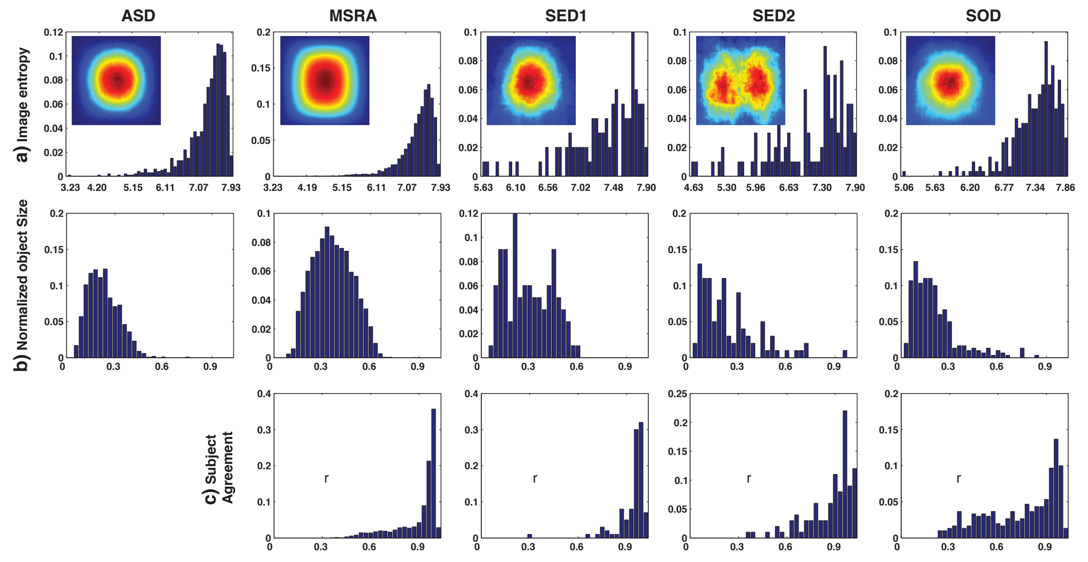
Fig.1 Sample images from the datasets. Top row shows the five smallest objects and bottom row shows the five largest objects from each dataset. 1) ASD [1]: This dataset contains 1,000 images. 2) MSRA [2]: This dataset (part B of the original dataset) includes 5,000 images. 3) SED [3]: This dataset contains two parts (SED1 and SED2)



(4) Experiments and Results

Baseline Models.
We implemented two simple yet powerful baseline models:
1) MAP (mean annotation position; see 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. 2. a) histogram of image entropy (inset: location prior MAP), b) distribution of normalized object sizes, c) distribution of annotators agreement (0 <= r <= 1). Note that the ASD dataset has only one annotator.



Based on the average rankings using AUC meric SVO, Goferman, CBsal, and RC stand at 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, MichalGazit, 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.

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. 3. Accuracy of models over small and large objects from ASD and SED1 data

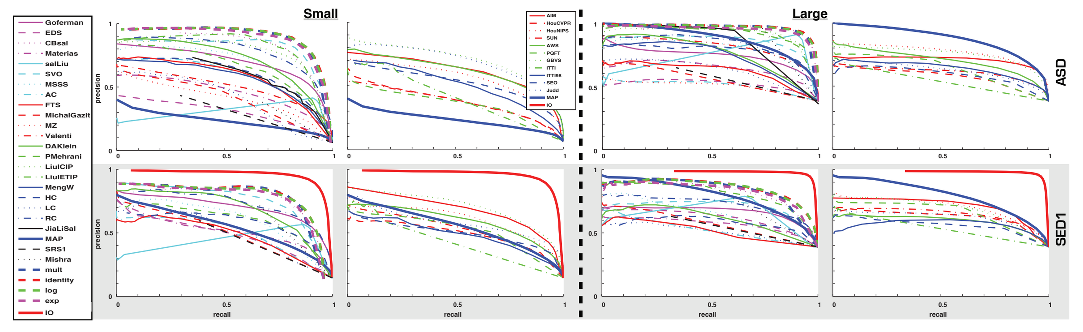


Fig. 4. Accuracy of saliency detection models over least and most consistent images

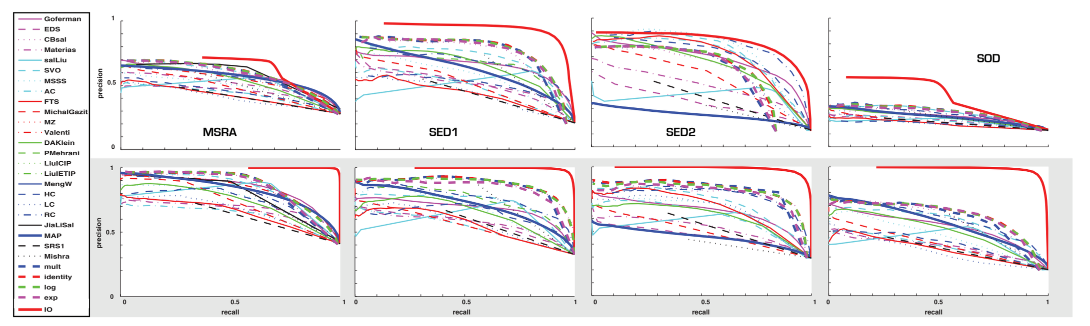
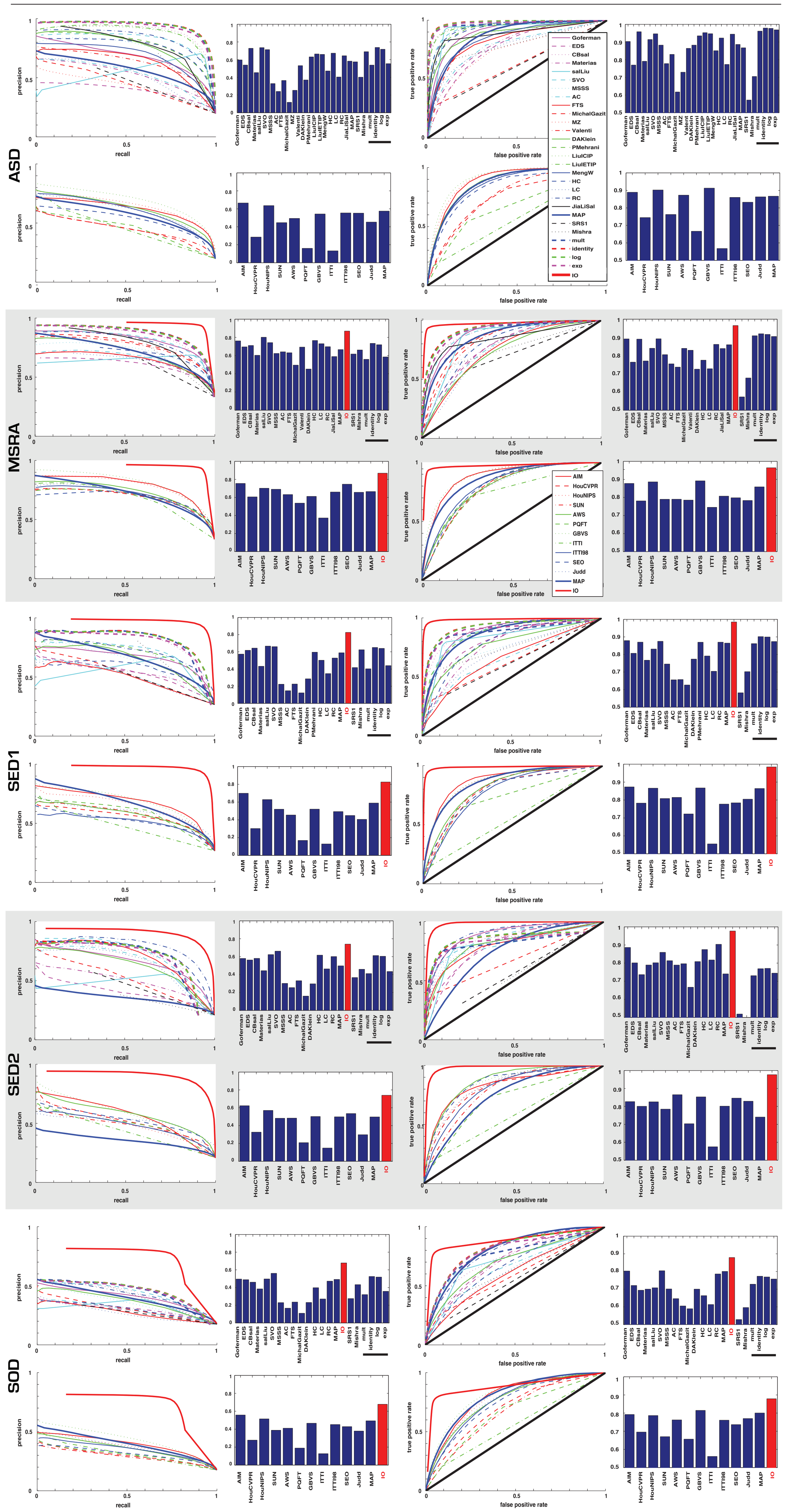


Fig. 5. PR, F-measure, ROC, and AUC scores of saliency models: Top row for each dataset shows saliency segmentation and bottom row shows fixation prediction models. mult indicates multiplication of models, and identity, log, exp indicate linear summation of identity, logarithmic, and exponential functions.



(5) Results cntd.

Fig. 6. Five images with least (and most) annotation consistency from datasets

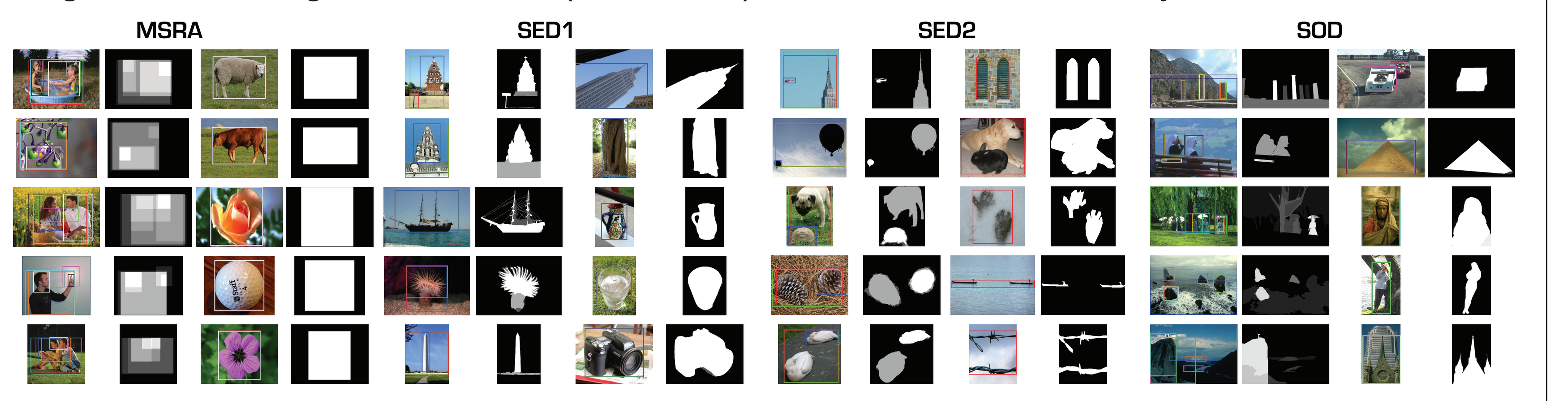
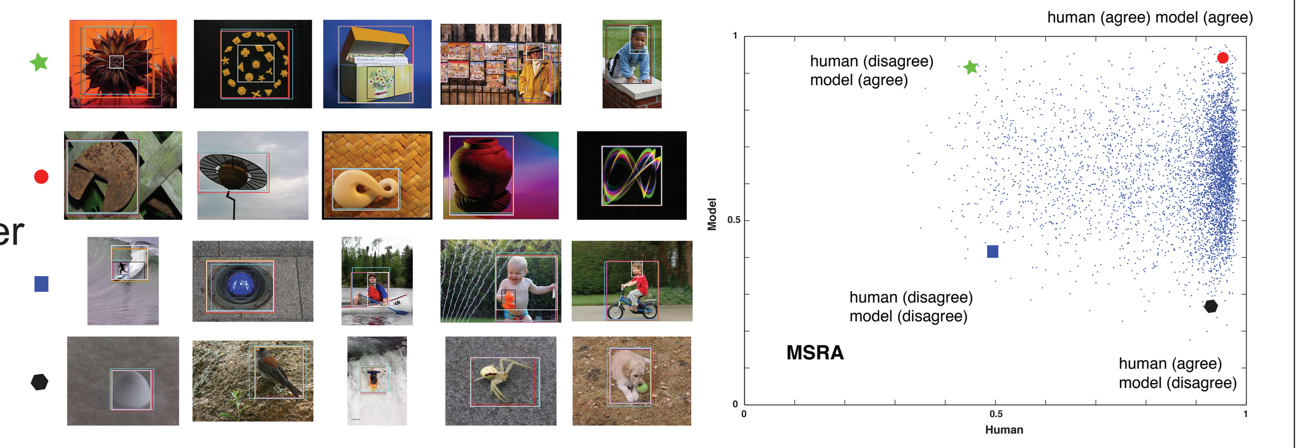


Fig. 7. Left) Easiest stimuli for 11 best models (Fig. 5), Right) Most difficult stimuli

Fig. 8. Analysis of human model agreement over MSRA dataset



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 bottomup salient. 2) Images that humans and models disagree (within the same group), 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.

(6) Conclusions

Based on the extensive experiments over 5 datasets, we conclude that SVO, Goferman, CBsal, RC, and Liu et al. (LiuICIP and Liu-IETIP) 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.

[1] Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: CVPR (2009)
[2] Liu, T., Sun, J., Zheng, N., Tang, X., Shum, H.: Learning to detect a salient object. In: CVPR (2007)
[3] Alpert, S., Galun, M., Basri, R., Brandt, A.: Image segmentation by probabilistic bottom-up aggregation and cue integration. In: CVPR (2007)
[4] Movahedi, V., Elder, J.H.: Design and perceptual validation of performance measures for salient object segmentation. In: POCV (2010)