Analysis of scores, datasets and models in visual saliency prediction

Ali Borji, Hamed R.-Tavakoli, Dicky N. Sihite, Laurent Itti
University of Southern California / University of Oulu
{borji, sihtie, itti}@usc.edu hrezazad@ee.oulu.fi

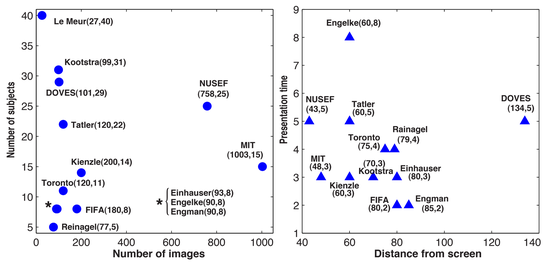
Introduction

A large number of methods has been proposed for predicting where people look in scenes. But due to lack of an exhaustive coherent benchmarking system, to address several issues such as evaluation measures, center-bias, map characterstics, and data set bias, a lot of inconsistencies still exist.

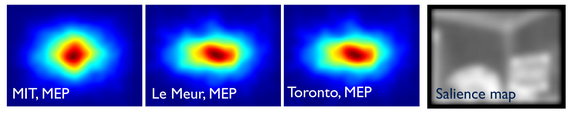
Contributions

1. Discussing current challenges and directions in saliency modeling such as evaluation matrices, dataset bias, model parameters, etc.
2. Comparing 32 models and their pros and cons in a unified quantitative framework over 4 widely-used data sets for fixation prediction (for both regular and affective data) as well as scanpath prediction.
3. Stimuli/task decoding using saliency and fixation statistics.

Datasets

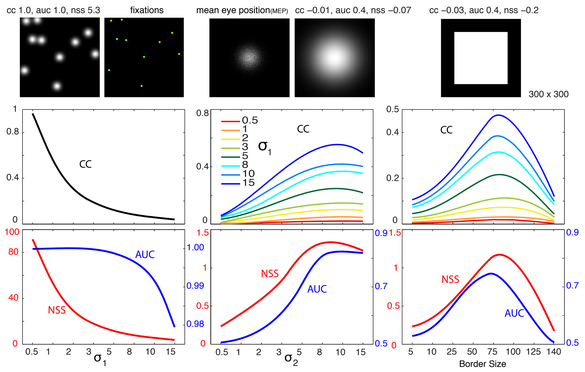


MIT, NUSEF, Kootstra and Toronto

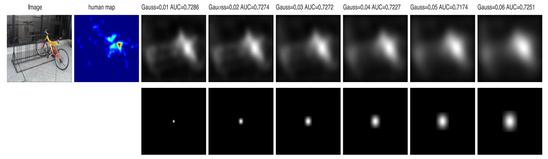


Center-Bias and border-effect phenomena

Tricking the metric



Smoothing



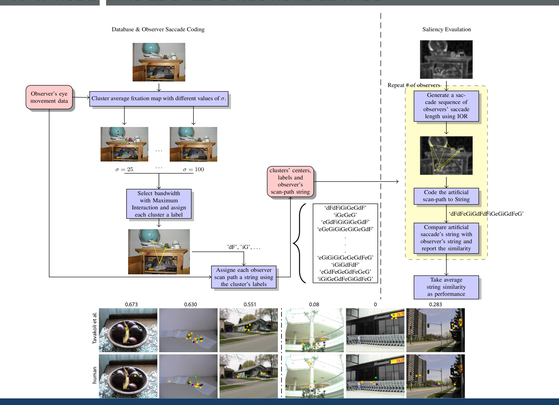
Other issues

re-parametrization & well-defined bounds

	upper bound	lower bound	nonlinear shifts
AUC	x	x	—
KL	—	x	x
CC	x	x	—
NSS	x	—	—

Note: There are 4 different AUC metrics. sAUC is robust toward center-bias and border-effect.

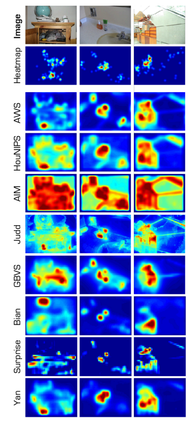
Scanpath Evaluation



The Benchmark

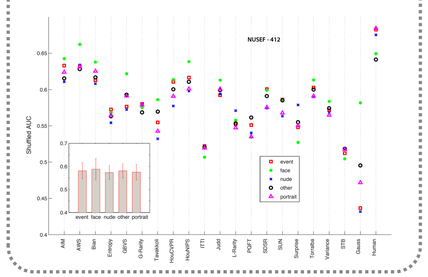
Fixation Prediction

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	-

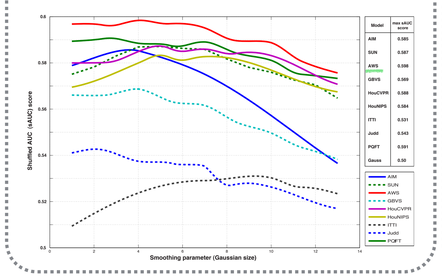


Images from the Torronto dataset

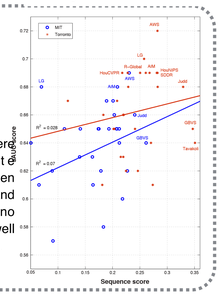
Class categories



Affective Data

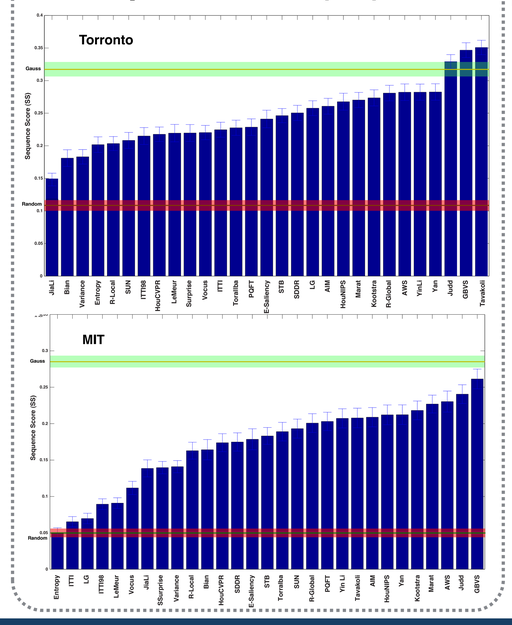


sAUC vs. SS



Note: Although there is a moderate relationship between scan path score and sAuc, there is no method to score well using both scores.

Scanpath scores (SS)



Conclusions & Discussions

1. Our comparisons show that in general AWS, LG, HouNIPS, Judd, Rarity-G (smoothed version), AIM, and Torralba models performed higher than other models.
2. 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 ≈ 0.5).
3. Some stimulus categories are harder for models (e.g., nature, nude, and portrait) which warrant more attention in future works.
4. It is feasible to decode the stimulus category from a feature vector combined from saliency, saccade, and fixation statistics.
5. Our results show a small gap (but statistically significant) between the best models and human performance.

ICCV'2013, Sydney, Australia.

https://sites.google.com/site/saliencyevaluation/