**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, sihite, 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 characteristics, 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

---

**Tricking the metric**

**The Benchmark**

**Fixation Prediction**

---

**Smoothing**

**Class categories**

**Affective Data**

---

**Other issues**

re-parametrization & well-defined bounds

<table>
<thead>
<tr>
<th>AUC</th>
<th>upper bound</th>
<th>lower bound</th>
<th>nonlinear shifts</th>
</tr>
</thead>
<tbody>
<tr>
<td>KL</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>CC</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>NSS</td>
<td>X</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

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

**Scanpath Evaluation**

**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/