Computer vision vs. human vision: What can be learned?

Ali Borji and Laurent Itti
University of Southern California, Los Angeles, CA 90089

The computer vision community has made rapid advances in several areas recently. In some restricted cases (e.g., where variability is low), computers even outperform humans for tasks such as frontal-view face recognition, fingerprint recognition, change detection, etc. A current trend is harvesting increasingly larger and unbiased datasets (e.g., ImageNet, SUN, Flicker, LabelME), constructing features/algorithms from these data, and designing suitable scores to gauge progress. The past successes have created the hope that maybe one day we will be able solve the hard problem of vision without having humans in the picture. Several previous studies, under the names of *humans in the loop*, *human debugging*, *finding weak links* (and often using AMT), have used humans to estimate the relative strengths and weaknesses of different algorithm components. Here we take a more systematic approach, comparing 14 computer vision models on 7 datasets using 5 different tests. The first two tests regard scene categorization using color photographs and line drawings. The third test addresses invariance properties of models on animal vs. non-animal recognition. The fourth test is about local vs. global information in the context of recognizing jumbled scenes. The final test involves object recognition over two large datasets, Caltech 256 and SUN. We learn that:

1. Models outperform humans in rapid categorization tasks, indicating that discriminative information is in place but humans do not have enough time to extract them. Models outperform humans on jumbled scenes and score relatively high in absence of (less) global information, which hints that they miss opportunistic local discriminative features.

2. We find that some models and edge detection methods are more efficient on line drawings and edge maps. Our analysis helps objectively assess the power of edge detection algorithms to extract meaningful structural features for classification, which hints toward two new directions. First, it provides another objective metric (in addition to conventional F-measure) for evaluating edge detection methods (i.e., an edge detection method serving better classification accuracy is favored). Second, it will help study which structural components of scenes are more important. For example, the fact that long contours are more informative can be used to build better feature detectors.

3. While models are far from human performance over object and scene recognition on natural scenes, even classic models show high performance and correlation with humans on sketches. The simplicity of sketches is a great opportunity to transcend models and discover mechanisms of biological object recognition. Another direction in this regard is to augment color, line, and spatial information for building better gist models (e.g., similar to geo_map).

4. Consistent with the literature, we find that some models (e.g., HOG, SSIM, geo/texton, and GIST) perform well. We find that they also resemble humans better. GIST, a model of scene recognition works better than many models over both Caltech-256 and Sketch datasets. HMAX has the 2nd best correlation on sketches and achieves a high accuracy.

5. Invariance analysis shows that only sparseSIFT and geo_color are invariant to in-plane rotation with the former having higher accuracy. On test 4, LBP has the highest $d^{'}$ and is the most similar model to humans over original images but it fails on rotated images.

Please refer to our paper for more details:
A. Borji and L. Itti, "Computer vision vs. human vision: What can be learned?", *CVPR*, 2014.

Model	siagianItti07	HMAX	denseSIFT	dSIFT_pyr	geo_color	geo_map8x8	geo_texton	GIST	gistPadding	HOG	HOG_pyr	LBP	LBP_pyr	LBPHF	LBPHF_pyr	line_hist	sparseSIFT	SSIM	SSIM_pyr	texton	texton_pyr	tiny_image
SUN	7.43	7	21.5	-	9.14	6.02	**23.5**	16.3	13.7	**27.2**	-	18.0	-	12.8	-	5.7	11.5	**22.5**	-	17.6	-	5.54
Caltech-256	16.5	12	29.4	28.4	4.9	5.3	20.3	27.4	25.1	**33.3**	**32.7**	20.7	20.5	17.6	17.8	6.54	20.4	**30.2**	25.0	29.9	27.8	13
Sketch	-	**55**	7.6	43.4	1.68	30.6	23.4	53.7	53.6	21.2	52.3	12.8	48.9	9.6	43.3	15.1	24.9	27.5	**56.2**	23.1	**56.9**	27.2
Animal/Non-Anim.	-	75.8	84.4	83.6	73.7	72.5	78.8	81.5	81	84	84.2	83.1	**85.7**	83.1	**85.8**	74.5	80.7	**84.9**	84.7	78.3	78.6	65
Similarity rank	13.6	13.6	9.3	12.6	10.2	13.8	**8.4**	11.2	12.4	**5.6**	**9.2**	10.0	10.2	11.7	10.0	13.0	11.8	**9.2**	10.8	9.6	**9.2**	18.9

Table 1. Classification accuracy (first 4th rows) and human-model similarity.