Human vs. Computer in Scene and Object Recognition

Ali Borji and Laurent Itti {borji,itti}@usc.edu
University of Southern California, Computer Science, Psychology, and Neuroscience

http://ilab.usc.edu/publications/doc/Borji_Itti14cvpr.pdf

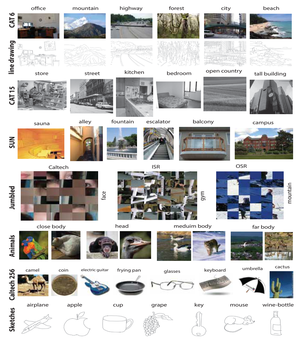
Supported by the National Science Foundation, the General Motors Corporation, the Army Research Office, and Office of Naval Research

Introduction & Motivation

- Here, to help focus research efforts onto the hardest unsolved problems, and bridge computer and human vision, we define a battery of 5 tests that measure the gap between human and machine performances in several dimensions.

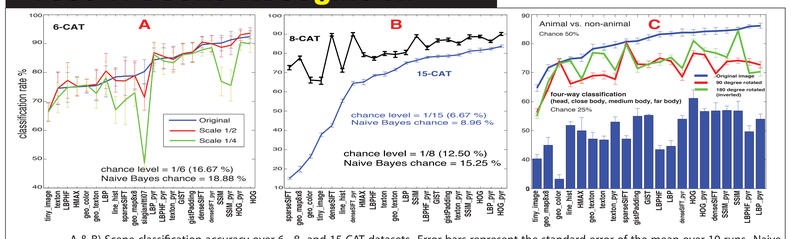
- Cases where machines are superior motivate us to design new experiments to understand mechanisms of human vision, and to reason about its failure. Cases where humans are better inspire computational researchers to learn from humans.

- In some applications (e.g., human-machine interaction or personal robots), perfect accuracy is not necessarily the goal; rather, having the same type of behavior (e.g., failing in cases where humans fail too) is favorable.



Sample images from scene and object datasets used in this study.

Test 1: Scene Recognition

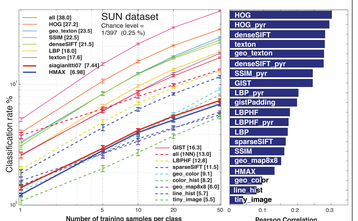


A & B) Scene classification accuracy over 6-, 8- and 15-CAT datasets. Error bars represent the standard error of the mean over 10 runs. Naive Bayes chance is simply set by the size of the largest class. All models work well above chance level. C) Top: animal vs. non-animal (distractor images) classification. Bottom: classification of target images. 4-way classification is only over target scenes (and not distractors).

- **A & B:** We find that HOG, SSIM, texton, denseSIFT, LBP, and LBHPF outperform other models (accuracy above 70%). We note that spatial feature integration (i.e., x_pyr for the model x) enhances accuracies.

- **C:** Animal vs. Non-Animal: All models perform above 70%, except tiny image. Human accuracy here is about 80%. Interestingly, some models exceed human performance here.

- SUN dataset: Models that performed well on small datasets (although they degrade heavily) still rank on top. GIST model works well here (16.3%) but below top contenders: HOG, texton, SSIM, denseSIFT, and LBP (or their variants). Models ranking at the bottom, in order, are tiny image, line hist, geo color, HMAX, and geo map8x8.



Performances and correlations on SUN dataset. We randomly chose n = {1, 5, 10, 20, 50} images per class for training and 50 for test.

Learned Lessons

1) Models outperform humans in rapid categorization tasks, indicating that discriminative information is in place but humans do not have enough time to extract it. Models outperform humans on jumbled images and score relatively high in absence of (less) global information.

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 new directions.

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.

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.

5) Invariance analysis shows that only sparseSIFT and geo_color are invariant to in-plane rotation with the former having higher accuracy (our 3rd test). GIST, a model of scene recognition works better than many models over both Caltech-256 and Sketch datasets.

Test 2: Recognition of Line Drawings and Edge Maps

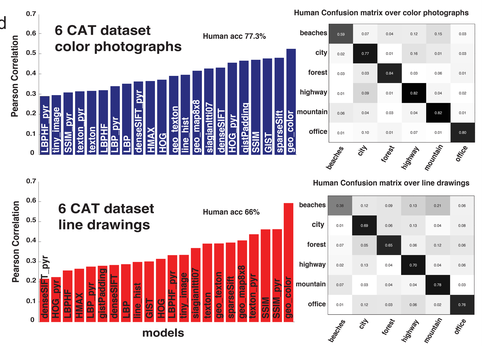
Line Drawings

- Scenes were presented to subjects for 17-87 ms in a 6-alternative force choice task (human acc= 77.3%).

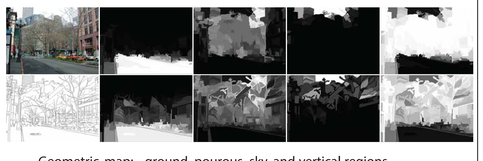
- On color images, geo_color, sparseSIFT, GIST, and SSIM showed the highest correlation (all with classification accuracy ≥ 75%), while tiny images, texton, LBHF, and LBP showed the least. Over the SUN dataset, HOG, denseSIFT, and texton showed high correlation with human CM.

- It seems that those models that take advantage of regional histogram of features (e.g., denseSIFT, GIST, geo_x; x=map or color) or heavily rely on edge histograms (texton and HOG) show higher correlation with humans on color images (although low in magnitude).

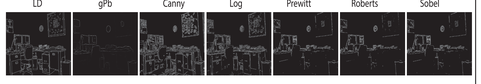
- Over line drawings: As in images, geo_color, SSIM, and sparseSIFT correlate well with humans.To our surprise, geo_color worked well.



Human-model agreement on the 6-CAT dataset. See our paper and its supplement for confusion matrices of models.

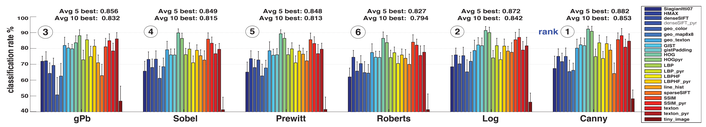


Geometric map: ground, pourous, sky, and vertical regions.



Edge maps for a sample image .

Edge Maps



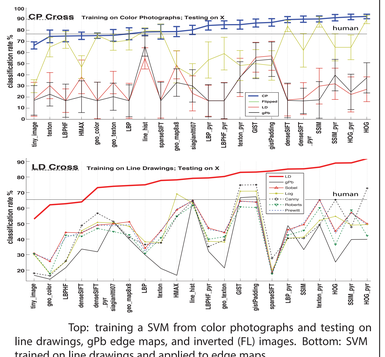
Scene classification results using edge detected images over 6-CAT dataset. Canny edge detector leads to best accuracies followed by the log and gPb methods.

- A majority of models perform > 70% on line drawings which is higher than human performance (similar pattern on images with human=77.3% and models > 80%).

- SVM trained on images and tested on line drawings: Some models (e.g., line hists, GIST, geo map, sparseSIFT) better generalize to line drawings.

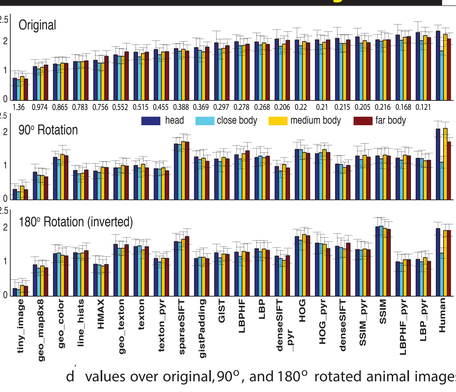
- SVM trained on line drawings and tested on edge maps: Surprisingly, averaged over all models, Sobel and Canny perform better than gPb. GIST, line hists, and HMAX were the most successful models using all edge detection methods. sparseSIFT, LBP, geo_color, and geo_texton were the most affected ones.

- Models using Canny technique achieved the best scene classification accuracy.



Top: training a SVM from color photographs and testing on line drawings, gPb edge maps, and inverted (FL) images. Bottom: SVM trained on line drawings and applied to edge maps.

Test 3: Invariance Analysis



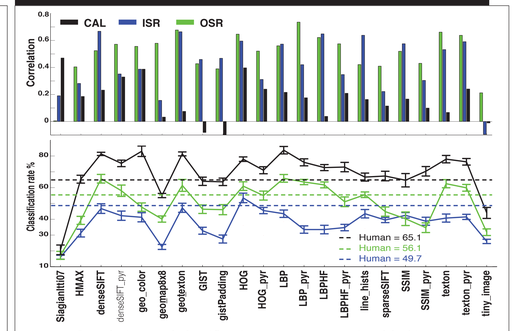
d' values over original, 90°, and 180° rotated animal images.

- A majority of models are invariant to scaling while few are drastically affected with a large amount of scaling (e.g., siagianItti07, SSIM, line hists, and sparseSIFT).

- Interestingly, LBP here shows a similar pattern as humans across four stimulus categories (i. e., max for head, min for close body).

- Some models show higher similarity to human disruption over the four categories of the animal dataset: sparseSIFT, SSIM, and HOG.

Test 4: Local vs. Global Information



Correlation and classification accuracy over jumbled images.

- As expected, models based on histograms are less influenced (e.g ., geo color, line hist, HOG, texton, and LBP).

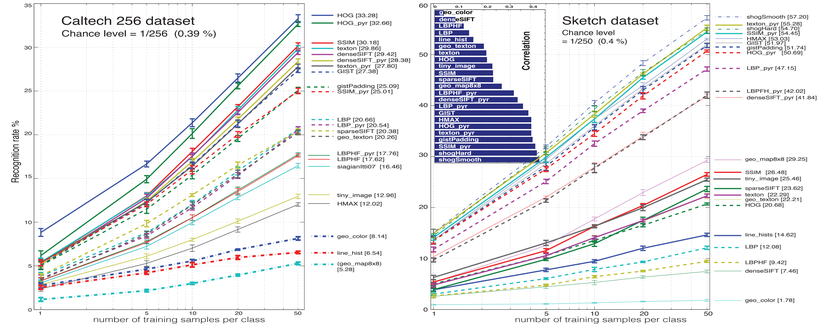
- Models correlate higher with humans over scenes (OSR and ISR) than objects, and better on outdoor scenes than indoors.

- Some models, which use global feature statistics, show high correlation only on scenes but very low on objects (e.g., GIST, texton, geo map, and LBP), since they do not capture object shape or structure.

Test 5: Object Recognition

- On Caltech-256, HOG achieves the highest accuracy about 33.28% followed by SSIM, texton, and dense SIFT.

- GIST which is specifically designed for scene categorization achieves 27.4% accuracy, better than some models specialized for object recognition (e.g., HMAX).



Left: Object recognition performance on Caltech-256 dataset. Right: Recognition rate and correlations on Sketch dataset.

- On sketch images, the shogSmooth model, specially designed for recognizing sketch images, outperforms others (acc=57.2%). Texton histogram and SSIM ranked second and fourth, respectively. HMAX did very well (in contrast to Caltech-256), perhaps due to its success in capturing edges, corners, etc.

- Overall, models did much better on sketches than on natural objects (results are almost 2 times higher than the Caltech-256). Here, similar to the Caltech-256, features relying on geometry (e.g., geo_map) did not perform well.

Summary

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

Classification results corresponding to 50 training and (50 over SUN and remaining images over Caltech-256 and Sketch) testing images per class
Animal vs. non-Animal corresponds to classification of 600 target vs. 600 distractor images . Top three models on each dataset are highlighted in red.