Human vs. computer in scene and object recognition

Ali Borji* and Laurent Itti*†
Departments of Computer Science* and Psychology†, Neuroscience Graduate Program†
University of Southern California, Los Angeles, CA 90089
{borji,itti}@usc.edu http://ilab.usc.edu/borji/

Abstract

Several decades of research in computer and primate vision have resulted in many models (some specialized for one problem, others more general) and invaluable experimental data. 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 (generalization across scene categories, generalization from images to edge maps and line drawings, invariance to rotation and scaling, local/global information with jumbled images, and object recognition performance). We measure model accuracy and the correlation between model and human error patterns. Experimenting over 7 datasets, where human data is available, and gauging 14 well-established models, we find that none fully resembles humans in all aspects, and we learn from each test which models and features are more promising in approaching humans in the tested dimension. Across all tests, we find that models based on local edge histograms consistently resemble humans more, while several scene statistics or "gist" models do perform well with both scenes and objects. While computer vision has long been inspired by human vision, we believe systematic efforts, such as this, will help better identify shortcomings of models and find new paths forward.

1. Introduction

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 Amazon Mechanical Turk) [27, 26, 36, 28, 34], have used humans to estimate the relative strengths and weaknesses of different algorithm components (e.g., [18, 36]). Here we take a more systematic approach, comparing 14 computer vision models on 7 datasets using 5 different tests. We focus on two difficult central vision problems, namely object and scene recognition. While current computer vision models have difficulty on these problems, humans solve them almost effortlessly.

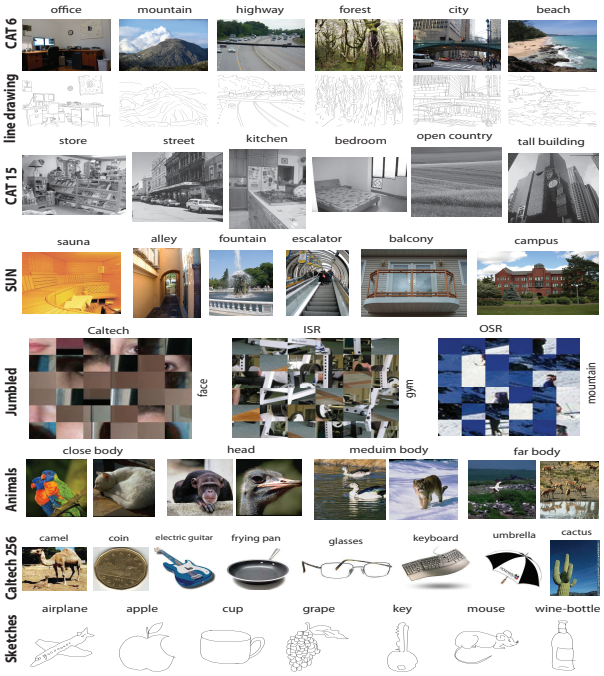

Figure 1. Sample images from scene and object datasets used in this study.

What can we learn from comparing human and machines? First, such a comparison helps diagnose cases where humans and machines disagree. Cases where machines are superior motivate experimentalists 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. Thus we believe that accuracy alone is not enough to judge

which models perform better (e.g., the main goal of existing challenges such as PASCAL or Caltech), and error patterns should be also part of the equation, as some machine errors may or may not be considered potential show-stoppers (helps achieving a graceful degradation). The intuition behind using error patterns is that two scenes or objects (from different categories) might be confused with each other if they share representations that humans perceive as similar or equivalent [25, 10]. Second, in some applications (e.g., humanizing machine behavior in 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. Here, we focus on visual perception (and even more specifically on the more tractable problem of scene and object recognition), rather than on broader visual cognition.

We organize 5 tests: The first two 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. Comparing with previous studies (e.g., [18]) of biological plausibility of recognition models, here we investigate several models on large benchmark datasets with direct respect to behavioral data.

2. Elements of Our Comparison

Statistics of 7 datasets (5 for scenes and 2 for objects) used here are described in the next section (Fig. 1), followed by our model evaluation part. Please see supplementary for details on human data gathering protocols on datasets.

2.1. Human studies

6-CAT: [25] includes 3,866 color photographs: beaches (553), city (648), forest (730), highway (563), mountain (720), and offices (652) with resolution of 800×600 pixels. Line drawings of about 80 images per category (475 total) were taken from a small part of the Lotus Hill Dataset, onto which contours were traced by trained artists.

15-CAT: [8] contains 4,578 gray-level images of size about 300×200 including coast (360), forest (328), highway (260), mountain (374), street (292), tall buildings (356), bedroom (216), kitchen (210), inside city (308), open country (410 images), suburb (241), industrial (311), living room (289), industry (308), and store (315). We call the first 8 categories as the 8-CAT dataset [10].

SUN: [15] is an extensive scene database containing 899 categories and 130,519 images. We experiment with 397 well-sampled categories to evaluate state-of-the-art algorithms and their accuracy versus humans. The number of images varies across categories, but there are at least 100 images per category, and 108,754 images in total. Xiao et al. [15] measured human scene classification accuracy using Amazons Mechanical Turk (MT). For each SUN category they measured human accuracy on 20 distinct test scenes, for a total of $397 \times 20 = 7,940$ experiments. In our work, we focus on results collected from the 13 "good workers" who performed at least 100 HITs and had accuracy above 95% on the relatively easy first level of the hierarchy. These workers accounted for just over 50% of all HITs [15].

Animals: [21] contains 600 target images (256×256 pixels; 150 *close body*, 150 *far body*, 150 *head*, and 150 *medium body*) and 600 distractors. An advantage of this dataset is offering human accuracy on 90° and 180° rotated images, allowing invariance analysis of models. Fourteen subjects were presented a stimulus (gray-level image) for 20 *ms*, followed by a blank screen for 30 *ms* (i.e., SOA of 50 ms), and followed by a mask for 80 *ms*. Subjects ended the trial with an answer of 'yes' or 'no' by pressing one of two keys.

Jumbled images: [23] Humans are able to recognize jumbled images as those in Fig. 1, 4^{th} row [3]. Indeed, Parikh [23] showed a a majority-vote accumulation over human classification of the individual blocks is a good predictor of human responses of the entire jumbled images. This dataset contains human performances on 3 image sets: 1) OSR, 384 outdoor scenes from the 8 categories of 8-CAT, 2) ISR, 300 indoor scenes [5] from bathroom, bedroom, dining room, gym, kitchen, living room, theater and staircase categories, and 3) CAL: Caltech objects (50 images from each of 6 categories aeroplane, car-rear, face, ketch, motorbike, and watch). Data have been collected in 3 cases, 1) *intact original images*, 2) *jumbled by variable block size* (used here), and 3) *individual blocks*. Here, we utilize the jumbled images over 6×6 blocks (Fig. 1).

Caltech-256: [17] This dataset, one of the most challenging datasets for object recognition, corrected some of the deficiencies of Caltech-101 dataset by introducing a diverse set of lighting conditions, poses, backgrounds, sizes, and camera systematics. It contains 30,607 images, with each category (out of 256) having a minimum of 80 images.

Sketch images: [16] This dataset contains non-expert sketches of everyday objects such as teapot or car. There are 20,000 unique sketches evenly distributed over 250 categories (i.e., 80 images per category). In a perceptual study [16], humans were able to correctly identify the object category of a sketch 73.1% of the time (chance is 0.4%). Given a random sketch, participants were asked to select the best fitting category from the set of 250 object categories. To avoid the frustration of scrolling through a list of 250 categories for each query, categories were organized in an intuitive 3-level hierarchy, containing 6 1st-level and 27 2nd-level categories such as animals, buildings, and musical instruments. There were a total of 5,000 HITs, each requiring workers to identify 4 sketches from random categories.

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
Reference	[20]	[22]	[8]	[8]	[15]	[6]	[15]	[10]	[15, 10]	[4]	[4]	[9]	[9]	[2]	[2]	[7]	[13]	[12]	[12]	[11]	[11]	[14]
Feat. dimension	714	4096	300	6300	3920	256	2560	512	512	300	6300	59	1239	38	798	230	2000	300	6300	512	10572	3072
Run time	0.59	4.27	4.48	-	0.85	2.80	5.92	0.46	1.12	**0.29**	-	**0.32**	-	0.34	-	-	0.66	3.52	-	5.21	-	**0.004**

Table 1. Employed models. Run times are in sec/image, averaged over 1K images randomly chosen from the 15-CAT, on a PC with 6-core AMD 2435 with 32 GB RAM. tiny_images is the fastest model followed by HOG and LBP. Slowest ones are geo_texton, texton, and HMAX (CPU version). dSIFT stands for denseSIFT.

2.2. Computational models of visual recognition

We run 14 models that have been shown to perform well on previous benchmarks. Models are listed in Table 1. Some are specifically designed for scene classification (e.g., GIST [10]: the output energy of a bank of 24 Gabor-like filters tuned to 8 orientations at 4 different scales on a 4×4 grid; similar core computation for gistPadding) while others are proposed for object recognition (e.g. HOG [4], HMAX [22], LBP [9]). HOG and SIFT models work well for both scene and object recognition [4, 8, 15].

In HOG, histograms of oriented gradients on each node of a grid are computed, then a descriptor is built for each one. The resultant descriptors are stacked on a 2×2 grid on the image. The descriptors are quantized into 300 visual words by k-means. sparseSIFT features [13] are extracted from MSER and Hessian-affine interest points (IP) and are clustered into a 2K word dictionary (1K per IP). denseSIFT features [8] are densely extracted using a flat window at two scales (4 and 8 pixel radii) on a regular grid at steps of 5 pixels and are clustered into a 300 dictionary. LBP [9] and LBPHF [2] are powerful texture features based on occurrence histogram of local binary patterns.

SSIM [12] measures local self-similarities of a scene layout. SSIM descriptors are computed on a regular grid and then quantized in 3 radial bins and 10 angular bins, obtaining 30D vectors. The descriptors are then quantized into 300 visual words by k-means. tiny_image features [14] serves as a baseline here and is simply the downsampled and linearized color image (i.e., $32 \times 32 \times 3$). line_hist features [7] are histograms based on the statistics of detected lines from an edge detector- one with bins corresponding to line angles and one with bins corresponding to line lengths. Texton histogram [11] is a 512D histogram by assigning each pixel's set of filter responses to the nearest texton dictionary entry (from a universal texton dictionary [11])[1].

Geometric Probability Map (geo_map8$\times$8) computes geometric class probabilities for image regions, e.g., ground, vertical, porous, and sky [6], and similarly for Geometry Specific Histograms (e.g., geo_texton). geo_color is composed of color histograms (joint histograms of color in CIE Lab color space (4, 14, and 14 bins, respectively [15]) over geometric maps.

HMAX is based on the feed-forward computations in the hierarchy of the visual ventral stream. We chose HMAX due to its previous success at surpassing the state of the art with generalization over object classes (faces, cars, handwritten digits, and pedestrians) [22]. For each image, a feature vector is computed by concatenating responses of a fixed subset of 4,096 S2 model units. Lastly, siagianItti07 is constructed from the maps of a saliency model [24, 29, 30]. We consider 4 scales for each orientation pyramid, 6 scales for each color pyramid, and 6 scales for intensity. For each map, averages in each patch of grid sizes $n \times n$ (here $n \in \{1, 2, 4\}$) are calculated (thus 21 values). The output is thus a $(4 \times 4 + 6 \times 2 + 6 \times 1) \times 21 = 714$D vector.

We borrowed some model codes from the SUN page (http://people.csail.mit.edu/jxiao/SUN/) with the same kernels they used (see [15] for more details).

3. Experiments and Results

We train 1-vs-all SVM classifiers following a cross validation procedure by dividing each dataset into 10 folds. For each of our tests, a model confusion matrix (CM) is derived over each fold and then the average CM over all folds is computed. The trace of this matrix indicates the accuracy. The Pearson correlation coefficient of the model CM and the human CM measures the human-model similarity/agreement. Note that we discard diagonal entries to limit the analysis to error cases. In each test, we first report accuracies and then analyze correlations.

3.1. Test 1: Generalization on scene categorization

Our first test addresses visual discrimination and representation power of humans and algorithms. Classification accuracies over 3 classic datasets (6-, 8-, and 15-CAT) are shown in Fig. 2. Although model rankings vary across datasets possibly due to different categories and thus feature statistics, some patterns can be observed. For example, in alignment with [15], we find that HOG, SSIM, texton, denseSIFT, LBP, and LBHPF outperform other models (accuracy above 70%). Increasing the number of classes from 6 and 8 to 15, drastically hinders performances of some models: tiny_image, geo_map8x8, geo_color, and sparseSIFT. This is perhaps due to confusion of additional similar classes in 15-CAT. Overall, we note that spatial feature integration (i.e., x_pyr for the model x) enhances accuracies. All models perform above chance level.

We now proceed to the animal categorization task using data of [21]. We follow the random-split procedure as in [21] by first splitting the set of 1,200 (animal and non-

[1] For some models, after visual word representation/quantization, 3-level spatial histograms are computed on grids of 1×1, 2×2, and 4×4 and are then augmented. We denote them with x_pyr (e.g., HOG_pyr) [15].

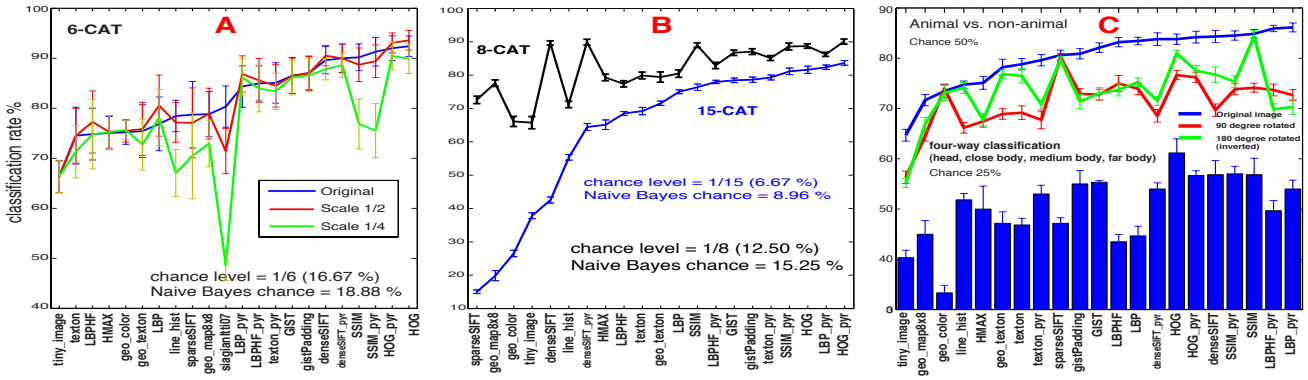

Figure 2. A & B) Scene classification accuracy over 6-, 8- and 15-CAT datasets. Error bars represent the standard error of the mean over 10 runs. Naive Bayes chance is simply set by the size of the largest class. All models work well above chance level. C) Top: animal vs. non-animal (distractor images) classification. Bottom: classification of target images. 4-way classification is only over target scenes (and not distractors). See Fig. 1 for sample images.

animal) images into 2 sets of 600 images each (one for training and the other for test). We then trained a SVM and applied it to the test set. The whole procedure was repeated 20 times. Results are shown in Fig. 2.C. For animal vs. non-animal (collapsed across all 4 categories; top curves in Fig. 2.C), models that worked well on scene categories (Figs. 2.A & 2.B) ranked on top here, for models such as: LBP, LBPHF, SSIM, HOG, and denseSIFT (with performance ranging in a narrow band). SIFT with sparse sampling does not perform as well as denseSIFT in alignment with [15, 8]. All models perform above 70%, except tiny_image (chance=50%). Human accuracy here is about 80%. Interestingly, some models exceed human performance in this task. On the 4-way categorization, similar to Figs. 2.A & 2.B, HOG is the best model followed by SSIM and denseSIFT. Although some models can tell whether an animal exists in the scene or not well (e.g., LBP, LBPHF, texton, and geo_color) they fail to separate categories of scenes with only animals (i.e., 4-way classification).

Moving to the large-scale SUN dataset (Fig. 3), 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. The low performance of HMAX can be because we used fixed learned patches (not from SUN dataset). Further, HMAX is essentially an object recognition model. Ranking of sparseSIFT swings. It ranks high on 6-CAT and animal vs. non-animal classification, but performs low on 15-CAT, 4-way animal categorization, and SUN datasets.

Fig. 4 shows human-model correlation over the 6-CAT dataset. Human CMs were borrowed from [25] where subjects were presented scenes for 17-87 *ms* in a 6-alternative force choice task (left panel; human acc= 77.3%). On this task, geo_color, sparseSIFT, GIST, and SSIM showed the highest correlation (all with classification accuracy $\geq 75\%$), while tiny_images, texton, LBHF, and LBP showed the least

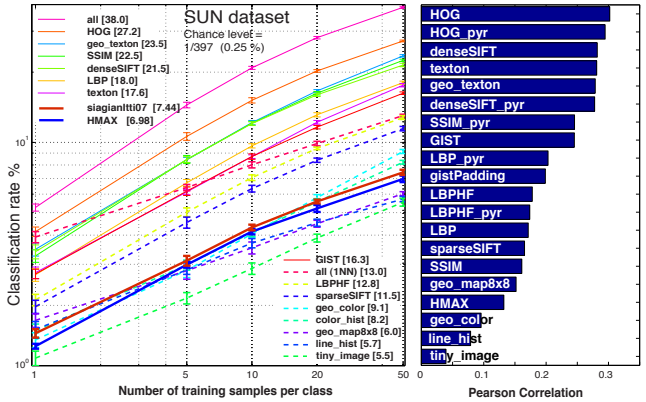

Figure 3. Performances and correlations on SUN dataset. We randomly choose $n \in \{1, 5, 10, 20, 50\}$ images per class for training and 50 for test.

correlation. Over the SUN dataset (CM taken from [15] website), HOG, denseSIFT, and texton showed high correlation with human CM. GIST performed well on all 3 datasets. As we expected, correlations are high on 6-CAT and lower on the SUN dataset. Overall, 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 (although still low in magnitude). Note that we do not expect to see the same ranking in correlations across 6-CAT and SUN since humans tasks were different on these datasets: rapid categorization on 6-CAT vs. and regular classification on SUN. While performing well, some models don't resemble humans much (e.g., LBP), suggesting that their internal representations may not match that of humans very well.

3.2. Test 2: Test of early vision: Recognition of line drawings and edge maps

Our second test regards the low-level representation capabilities of humans and machines focusing on line drawings. Humans are able to classify a scene from its line drawings even for a very short presentation [25, 19] (accuracy of 66%). Performance of models over 6-CAT line drawings are shown in Fig. 7 bottom panel. Majority of models per-

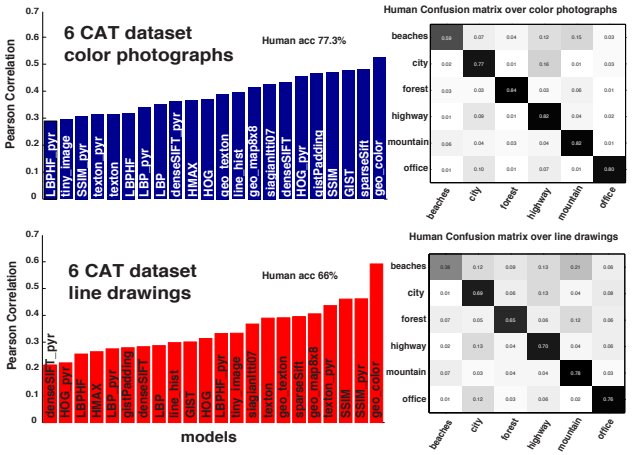

Figure 4. Human-model agreement on the 6-CAT dataset. See supplement for confusion matrices of models.

form above 70% which is higher than human performance (similar pattern over color photographs with human=77.3% and models $> 80\%$). While all models show a performance drop on line drawings, their ranking is nearly the same as in color images except for few models. Some models improve (line_hist and sparseSIFT) while some others (denseSIFT, line_hist, and LBP) degrade.

Fig. 4 (bottom) shows correlations over line drawings. As in images, geo_color, SSIM, and sparseSIFT did well here. To our surprise, geo_color worked well on both line drawings and color images. To investigate this, in Fig. 5, we illustrate geometric maps for an image and its line drawing and observe that both look similar. We also see similar patterns over the feature histograms over two images (supplement). High accuracy of geo_color over line drawings (62%) suggests that when trained, geo_color can mimic human CM over line drawings (supplement). We think discrepancy in model ranking here might be partly due to differences of human confusions over images and drawings (e.g., different confusions between mountain vs. forest).

To study how much structural information models retain, we trained a SVM on color images and tested it on line drawings (separate train and test sets). Fig. 7 shows that performance of a majority of models drops significantly. Some (e.g., line_hists, GIST, geo_map, sparseSIFT) better generalize to line drawings (but not necessarily in the same way as humans, due to differences in error patterns). gPb doesn't work well here as it actively suppresses texture edges which are quite discriminative for scene classification.

Further, to evaluate how well edge maps can simulate human line drawings, we applied a SVM trained from line drawings to edge maps using 6 prevalent edge detection methods (Fig. 7; bottom panel). Surprisingly, averaged over all models, Sobel and Canny perform better than gPb [33]. While gPb aims to retain important and possibly identity-preserving edges (Fig. 6), as it turns out, this does not help classification much using models. It seems that models were better able to extract discriminative features from edge

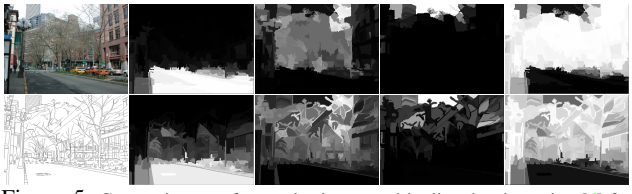

Figure 5. Geometric maps for a color image and its line drawing using [6] for ground, pourous, sky, and vertical regions. See supplement for feature histograms.

Figure 6. Edge maps for the top left image in Fig. 1.

maps. 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. sparseSIFT is dramatically reduced here possibly because some interest points are not preserved in edge maps. Finally, Fig. 8 shows the direct classification accuracy using each algorithm's edge maps. Canny technique achieved the best accuracy (with the highest average accuracy on 5 and 10 top models). gPb did not do the best here, supporting our earlier argument. HOG, SSIM, and texton are the best using all edge detectors which is in alignment with model performances over line drawings (Fig. 7).

3.3. Test 3: Invariance analysis

While from the design of models, we know about their invariance properties, it is worth testing them empirically. We perform two types of invariance analysis: scaling and in-plane rotation (see [18] for in-depth invariance analysis). Fig. 2.A shows results of applying a SVM trained on original scenes to 0.5 and 0.25 scaled images over 6-CAT dataset. 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).

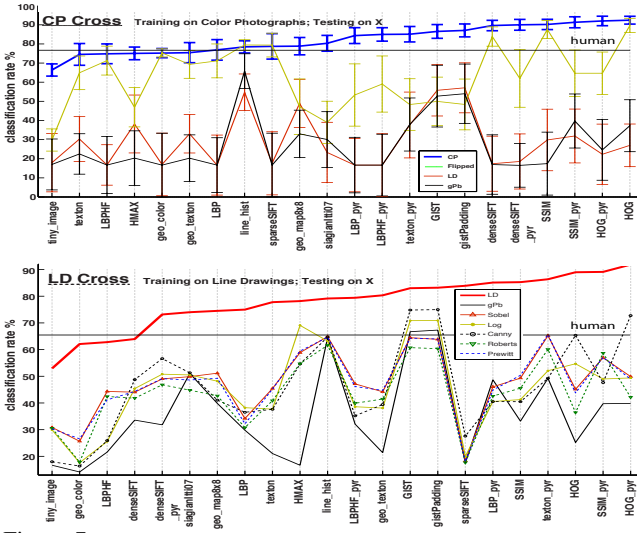

Figure 7. Top: training a SVM from color photographs and testing on line drawings, gPb edge maps, and inverted (FL) images. Bottom: SVM trained on line drawings and applied to edge maps.

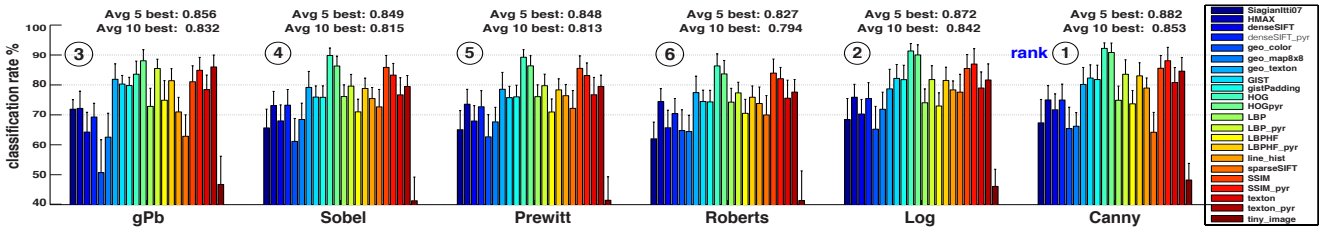

Figure 8. Scene classification results using edge detected images over 6-CAT dataset. Canny leads to best accuracies followed by the log and gPb methods.

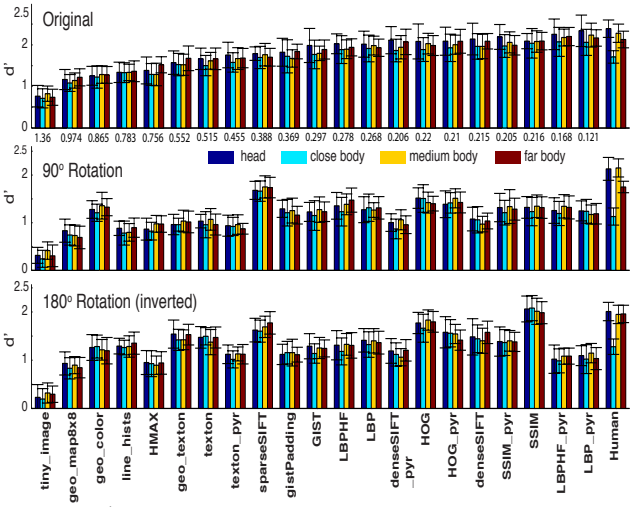

Figure 9. d' values over original, 90°, and 180° rotated animal images. Numbers under the x-axis at the top panel show the mean distance between model and human over 4 categories over original images. Models were trained on original images and tested on rotated ones.

For rotation invariance, we tested a binary SVM trained on 0° animal scenes to 90° and 180° rotated images using different train and test sets (Fig. 2.C). Some models are invariant to both types of rotations (e.g., sparseSIFT, geo_color) while some are only invariant to 180° rotation (e.g., line_hist, HOG, and texton). Some other models are influenced by both rotations (e.g., LBP, GIST, SSIM, HMAX, denseSIFT) and In Fig. 7, a trained SVM on original images (6-CAT) was applied to inverted images. Some models are invariant (e.g., HOG, SSIM, SIFT) while others are influenced (e.g., HMAX, GIST, tiny_image).

Fig. 9 shows a break-down of human-model agreement over categories of animal dataset. Following Serre *et al.* [21] we measure agreement by d' which is a monotonic function that combines both the hit and false-alarm rates of each observer or a model. ($d' = Z(H) - Z(F)$, where Z is the inverse of the cumulative normal distribution). Some models have d' values close to humans (LBP, LBPHF, and SSIM on original images). 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. In alignment with Fig. 2.C, sparseSIFT was not affected by rotation while SSIM was only affected by 90° rotation. Shown in Fig. 9 humans are less affected by rotation compared to models, suggesting that more research is necessary to achieve better rotation invariance in models.

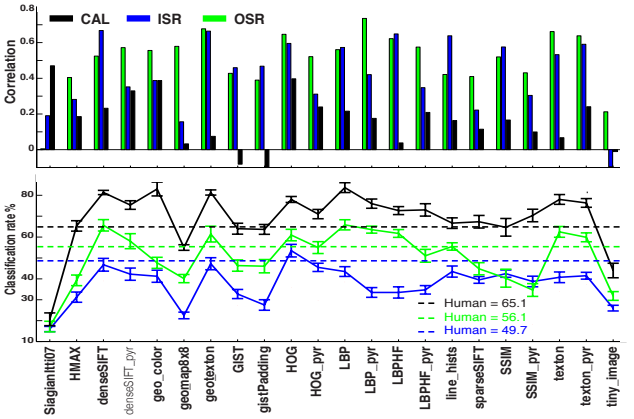

Figure 10. Correlation and classification accuracy over jumbled images.

3.4. Test 4: Local or global information: Recognition of jumbled images

Human performance on jumbled images depends on the level of image blocking [23] (here 65%). Model accuracies (trained and tested on jumbled images) are shown in Fig. 10. Models did well on CAL jumbled objects, with majority surpassing human accuracy. Model performances are higher in outdoor scenes (OSR) than indoor scenes (ISR). While models perform about the same as humans on OSR they score lower than humans on ISR. In addition to strong artificial vertical and horizontal gradients, because of the blocking, jumbling may hinder more those models that need some sort of alignment (e.g., siagianItti07, GIST, denseSIFT, geo_map, SSIM, HMAX, and tiny_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. This indicates that current models mainly utilize global scene statistics (also utilized by humans [10]), rather than category-specific local features. Averaged over OSR and ISR scenes and consistent with results in Fig. 4, HOG, geo_color, denseSIFT, and texton features show the highest correlation with humans. 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

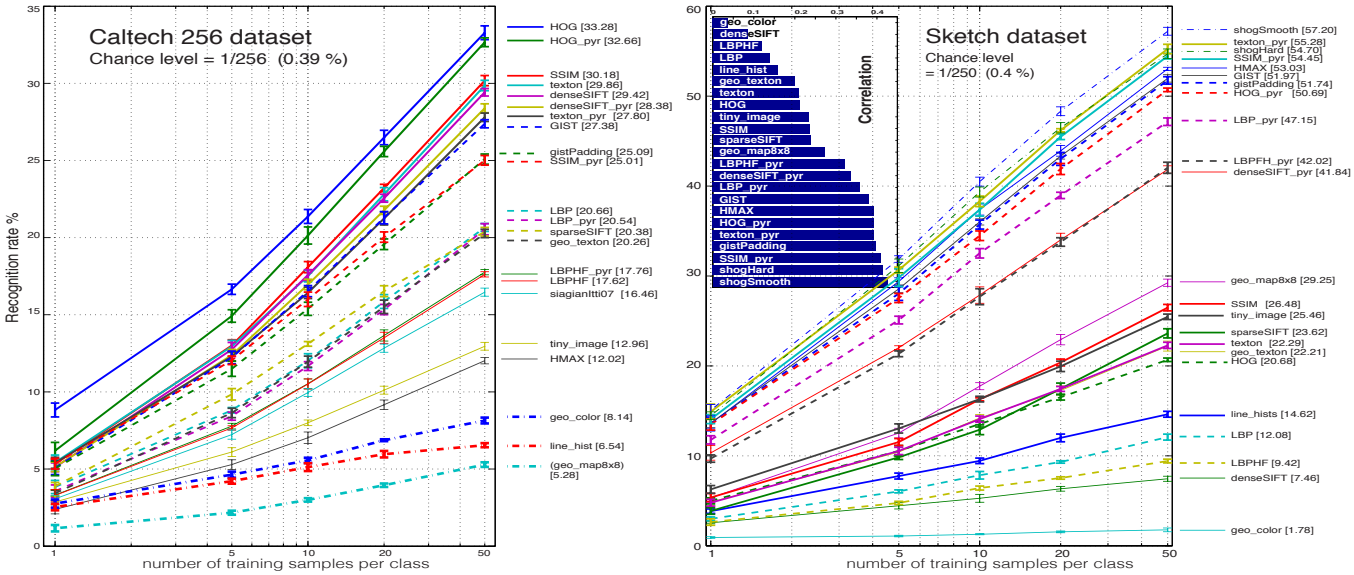

Figure 11. Left: Object recognition performance on Caltech-256 dataset. Right: Recognition rate and correlations on Sketch dataset.

3.5. Test 5: Object recognition performance

Fig. 11 reports multi-class recognition results on Caltech-256 and Sketch datasets. We randomly choose $n \in \{1, 5, 10, 20, 50\}$ images per class for training, and the remaining images per class for testing.

On Caltech-256, HOG achieves the highest accuracy about 33.28% followed by SSIM, texton, and denseSIFT. GIST which is specifically designed for scene categorization achieves 27.4% accuracy, better than some models specialized for object recognition (e.g., HMAX). Note that higher accuracies for HMAX have been reported by customizing this model (e.g., [17, 21]). denseSIFT outperforms the sparseSIFT. Interestingly, the very simple tiny_image model works well above chance (∼13%). Geometric feature maps did not work well since, as opposed to scenes, objects do not contain regions that can be aligned spatially (e.g., sky or ground). These results (here max accuracy about 33%) are in alignment with previous reports using classic feature detectors [31, 32]. In [31], authors were able to obtain accuracy about 50% using sophisticated feature combination techniques. Nevertheless, still best models fall short of human object recognition ability.

On sketch images, the shogSmooth model, specially designed for recognizing sketch images [16], 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. which are more explicit on sketch images than natural scenes (In alignment with its performance shown in Fig. 7.bottom). Contrary to our expectation, GIST, which includes color and texture processing, did well and close to HMAX. LBP did not work comparatively well due to lack of texture on sketches. 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. All models are above chance. Correlation analysis (using the 50 training/testing case) shows higher similarities for higher performing models with shog, HMAX, GIST, texton, and HOG on top. Humans perform about 17% better than the best model (human acc=73%), suggesting that they may be relying on additional features.

Similarity rank: To summarize, we show the average rank of models over 5 tests in the last row of Table 2. The lower the similarity rank, the better. HOG, geo_texton, SSIM, and texton performed the best, yet their overall score is low (i.e., no model does well on all tests). Table 2 also lists a summary of classification accuracies.

4. Discussions and Conclusions

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 it [19]. Models outperform humans on jumbled images and score relatively high in absence of (less) global information. Explicit addition of opportunistic local discriminative features, that humans often use, may enhance accuracy of models. **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 [25] can be used to build better feature detec-

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 2. Classification results corresponding to 50 training and (50 over SUN and remaining images over Caltech-256 and Sketch) testing images per class (Figs. 3 and 11). Animal vs. non-Animal corresponds to classification of 600 target vs. 600 distractor images (Fig. 2.C). Top three models on each dataset are highlighted in bold.

tors. **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 (our 3rd test). 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.

We argued that both accuracy and confusion matrices are important in evaluating models. On the one hand, high performing models may not show good correlation with humans which warrants further inspection. One could propose other alternatives to highly correlated CMs, e.g., looking at which exemplars are difficult to classify (instead of looking at misses at the category level). On the other hand, highly correlated CMs could occur even when the absolute performance (e.g., classification accuracy) is quite different. Contrasting humans and machines, although helpful, has its own challenges for two reasons. First, there exist many models (some we did not consider here e.g., new deep learning methods [35]), with several parameters (e.g., normalization, pooling sizes, kernels), sometimes yielding to quite different scores. Second, similarly human studies have been designed for specific purposes and hypotheses (with different settings) and it is not trivial to directly use them for model evaluation. This calls for extensive collaboration among experimental and computational vision researchers.

This work was supported by NSF (CCF-1317433 and CMMI-1235539) and ARO (W911NF-11-1-0046 and W911NF-12-1-0433).

References

[1] A. Turing. Computing Machinery and Intelligence. Mind LIX (236). 1950.

[2] T. Ahonen, J. Matas, C. He, and M. Pietikaïnen. Rotation invariant image description with local binary pattern histogram fourier features. In SCIA, 2009.

[3] J. Vogel, A. Schwaninger, C. Wallraven and H. H. Blthoff. Categorization of Natural Scenes: Local vs. Global Information. *APGV*, 2006.

[4] N. Dalal and B. Triggs. Histogram of oriented gradient object detection. *Computer Vision and Pattern Recognition*, 2005.

[5] A. Quattoni and A. Torralba. Recognizing Indoor Scenes. *CVPR*, 2009.

[6] D. Hoiem, A. Efros, and M. Hebert. Geometric context from a single image. *ICCV*, 2005.

[7] J. Kosecka and W. Zhang. Video compass, *ECCV*, 2002.

[8] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, *CVPR*, 2006.

[9] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, *PAMI*, 2002.

[10] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope, *Intl. J. Computer Vision*, 2001.

[11] L. Renninger and J. Malik. When is scene recognition just texture recognition? *Vision Research*, 2004.

[12] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. *CVPR*, 2007.

[13] J. Sivic and A. Zisserman. Video data mining using configurations of viewpoint invariant regions. *CVPR*, 2004.

[14] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: a large database for non-parametric object and scene recognition. *IEEE PAMI*, 2008.

[15] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. SUN Database: Large-scale Scene Recognition from Abbey to Zoo. *Computer Vision and Pattern Recognition*, 2010.

[16] M. Eitz, J. Hays, and M. Alexa. How Do Humans Sketch Objects? *ACM Transactions on Graphics, Proc. SIGGRAPH*, 2012.

[17] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. *Tech. rep., California Institute of Technology*, 2007.

[18] N. Pinto, D.D. Cox, and J.J. DiCarlo. Why is Real-World Visual Object Recognition Hard? *Plos computational biology*, 2008.

[19] M.C. Potter, E.I. Levy. Recognition memory for a rapid sequence of pictures. *J Exp Psychol*, 1969.

[20] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *PAMI*, 2007.

[21] T. Serre, A. Oliva, and T. Poggio. A Feedforward Architecture Accounts for Rapid Categorization, *PNAS*, 2007.

[22] T. Serre, L. Wolf, and T. Poggio. Object Recognition with Features Inspired by Visual Cortex, *CVPR*, 2005.

[23] D. Parikh. Recognizing Jumbled Images: The Role of Local and Global Information in Image Classification, *ICCV*, 2011.

[24] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 1998.

[25] D.B. Walther, B. Chai, E. Caddigan, D.M. Beck, L. Fei-Fei. Simple line drawings suffice for functional MRI decoding of natural scene categories. *PNAS*, 2011.

[26] D. Parikh, C. L. Zitnick and T. Chen. Exploring Tiny Images: The Roles of Appearance and Contextual Information for Machine and Human Object Recognition. *IEEE PAMI*, 2012.

[27] D. Parikh and C. L. Zitnick. Finding the Weakest Link in Person Detectors. *CVPR*, 2011.

[28] R. Mottaghi, S. Fidler, J. Yao, R. Urtasun, and D. Parikh. Analyzing Semantic Segmentation Using Hybrid Human-Machine CRFs. *CVPR*, 2013.

[29] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Trans. PAMI*, 2013.

[30] A. Borji, D.N Sihite and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. *IEEE Tran. IP.*, 2012.

[31] P. Gehler and S. Nowozin. On Feature Combination for Multiclass Object Classification. *ICCV*, 2009.

[32] C. Kanan and G. Cottrell. Robust Classification of Objects, Faces, and Flowers Using Natural Image Statistics. *CVPR*, 2010.

[33] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE T-PAMI*, 2011.

[34] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. *ECCV*, 2012.

[35] Y Bengio, A Courville, and P Vincent, Representation learning: A review and new perspectives, *IEEE T-PAMI*, 2013.

[36] F. Fleuret, T. Li, C. Dubout, E.K. Wampler, S. Yantis, and D. Geman. Comparing machines and humans on a visual categorization test. *PNAS*, 2011.