Human vs. computer in scene and object recognition
Supplementary Material

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/

1. Introduction

This supplementary material accompanies the paper "Computer vision vs. human vision: What can be learned?" submitted to CVPR 2014. It contains the followings:

- Human data gathering protocols on used datasets.
- Edge maps using 6 edge detectors, shown in Fig. 1.
- Sample jumbled scenes, shown in Fig. 2.
- Sample distractor scenes from the Animals dataset, shown in Fig. 3.
- geo_color model performance and confusion Matrix over the 6-CAT dataset, shown in Fig. 4.
- Model confusion matrices over color images from the 6-CAT dataset, shown in Fig. 5.
- Model confusion matrices over line drawings from the 6-CAT dataset, shown in Fig. 6.
- Human confusion matrix over Sketch images, shown in Fig. 7.
- Human confusion matrix over SUN dataset, shown in Fig. 8.
- Feature distributions for the image shown in Fig.5 of the main text and it's line drawing, shown in Figs. 9, 11, 13, 12, 15, 16, 10, 18, 17, 19, 14.
- Histogram of black and white pixels across 6 categories of the 6-CAT dataset, shown in Fig. 20.
- Average similarity rank of models, shown in Fig. 21.

2. Human Data Gathering Protocols

6-CAT: [6][1] Human Behavior Experiment [6]: The procedure was a Six-alternative force choice task in which subjects viewed a blank gray screen (fixating at the central cross) for 500 ms. This was followed by stimulus presentation (SOA) which was variable and was adjusted for each subject for him to reach 65% accuracy (stair-case strategy: i.e., increasing SOA until subject reaches 65% accuracy). Subjects then were shown a scrambled-phase image for 500 ms and a blank gray mask for 2000 ms. Line drawings and color photographs were randomly mixed. SOA ranged from 16.67 ms to 86.67 ms for which average accuracy over subjects were 77% over color photographs and 67% over line-drawings.

Fig. 1 shows sample images from this dataset (line drawings) and their corresponding edge detected maps from probabilistic edge detection map [7] (shown here as gPb), Canny, Log, Prewitt, Roberts, and Sobel. Please see the first figure in the paper for original images.

Animals: [8][2] In this task, Animal- vs. non-animal-categorization task, four (balanced) classes of stimuli were used and were manually arranged into four groups (150 images each) based on the distance of the animal from the camera: head (close-up), close-body (animal body occupying the whole image), medium-body (animal in scene context), and far-body (small animal or groups of animals). A set of matching distractors (Fig. 3) (300 each from natural and artificial scenes) was selected so as to prevent human observers and the computational models from relying on low-level cues.

Experimental Procedure: A stimulus (gray-level image) was flashed 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.

[1] http://vision.stanford.edu/projects/sceneclassification/index.html
[2] http://cbcl.mit.edu/software-datasets/serre/SerreOlivaPoggioPNAS07/index.htm

Jumbled images: [5][3] Data was collected through Mechanical Turk (MT). Each trial contained an image and subjects were supposed to report the category of the scene (The image was displayed to the subjects along with a list of radio buttons indicating the categories to select from). There were no constraints on subjects' response times. The images were resized such that the largest dimension was 256 pixels. Jumbled images were created by dividing an image into $n \times n$ non-overlapping blocks ($n \times n$ in our paper), and then randomly shuffling them. Fig. 2 shows sample images from the four image sets in this dataset.

Sketch images: [4][4] This dataset contains non-expert sketches of everyday objects such as 'teapot' or 'car'. There are 20,000 unique sketches evenly distributed over 250 object categories (i.e., 80 images per category). In a perceptual study [4], humans were able to correctly identify the object category of a sketch 73.1% of the time (chance is 0.4%). Experimental method: Given a random sketch, participants were asked to select the best fitting category from the set of 250 object categories. They had unlimited time, although they were naturally incentivized to work quickly for greater pay. 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 top-level and 27 second-level categories such as 'animals', 'buildings' and 'musical instruments'. There were a total of 5,000 HITs to MT, each requiring workers to identify 4 sketches from random categories. Fig. 7 shows human confusion matrices over sketch images.

SUN: [3][5] Since measuring human classification accuracy with nearly 400 categories is difficult and to help participants know which labels are available, Xiao et al. [3] grouped the 397 scene categories in a 3-level tree. Participants navigated through an over complete three-level hierarchy to arrive at a specific scene type (e.g. 'bedroom') by making relatively easy choices (e.g. 'indoor' versus 'outdoor natural' versus 'outdoor man-made' at the first level).

Xiao et al. [3], measured human scene classification accuracy using Amazon's Mechanical Turk (AMT). For each SUN category they measured human accuracy on 20 distinct test scenes, for a total of $397 \times 20 = 7,940$ experiments (HITs: Human Intelligence Tasks). Fig. 8 shows human accuracy and confusion matrices calculated from good workers.

In our work, we focus on results collected from the 'good workers' who performed at least 100 HITs and have accuracy greater than 95% on the relatively easy first level of the hierarchy the leaf-level accuracy rises to 68.5%. These 13 'good workers' accounted for just over 50% of all HITs.

[3]http://ttic.uchicago.edu/∼dparikh/publications.htm.
[4]http://cybertron.cg.tu-berlin.de/eitz/projects/classifysketch/
[5]http://groups.csail.mit.edu/vision/SUN/

3. geo_color Model Performance

Our results showed that geo_color model[6] achieves the best correlation with human behavior over line drawings on the 6-CAT dataset. Here, we analyze the performance of this model in more detail. Over scrutiny shows similar feature distributions across color images and line drawings; more using geo_color, denseSIFT, LBP, LBPHF, SSIM, Texton, geo_texton, geo_map, ans less using sparseSIFT, HOG. Fig. 4 shows the confusion matrix of this model. Note that also the high classification accuracy over line drawings is not surprising as distributions of black and white pixels differs dramatically across classes. Further, high correlation of geo_color with humans suggests that that humans may rely on regional features more than global scene histograms.

The reason why geo_color works well on line drawings is that categories of 6-CAT dataset are well-separated using only binary pixel statistics. Fig. 20 shows the histogram of black regions (black pixels belonging to line drawings) over images from different categories. As it shows, this simple feature can distinguish classes to a good extent. Note that geo_color only extracts statistics from the binary image and there is no other texture or colored pixels in line-drawing images.

4. Average Similarity Rank of Models

Fig. 21 shows similarity ranks of models on each of the following tests (plus the average similarity rank over tasks):

- 6-CAT; CP (Color Photographs)
- 6-CAT; LD (Line Drawings)
- SUN
- Animal vs. non-Animal
- Invariance; Animal-180°
- Invariance; Animal-90°
- Jumbled Caltech
- Jumbled ISR
- Jumbled OSR
- Sketch

Results are sorted according to the average similarity rank. Note that the lower the similarity score the better.

[6]Color histograms are joint histograms of color in CIE Lab color space (4, 14, and 14 bins, respectively and are calculated over geometric regions; ground, sky, porous, vertical) [3].

References

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

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

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

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

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

[6] D.B. Walther, B. Chai, E. Caddigan, D.M. Beck, and L. Fei-Fei. Simple line drawings suffice for functional MRI decoding of natural scene categories. *Proceedings of the National Academy of Sciences (PNAS)*, 2011.

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

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

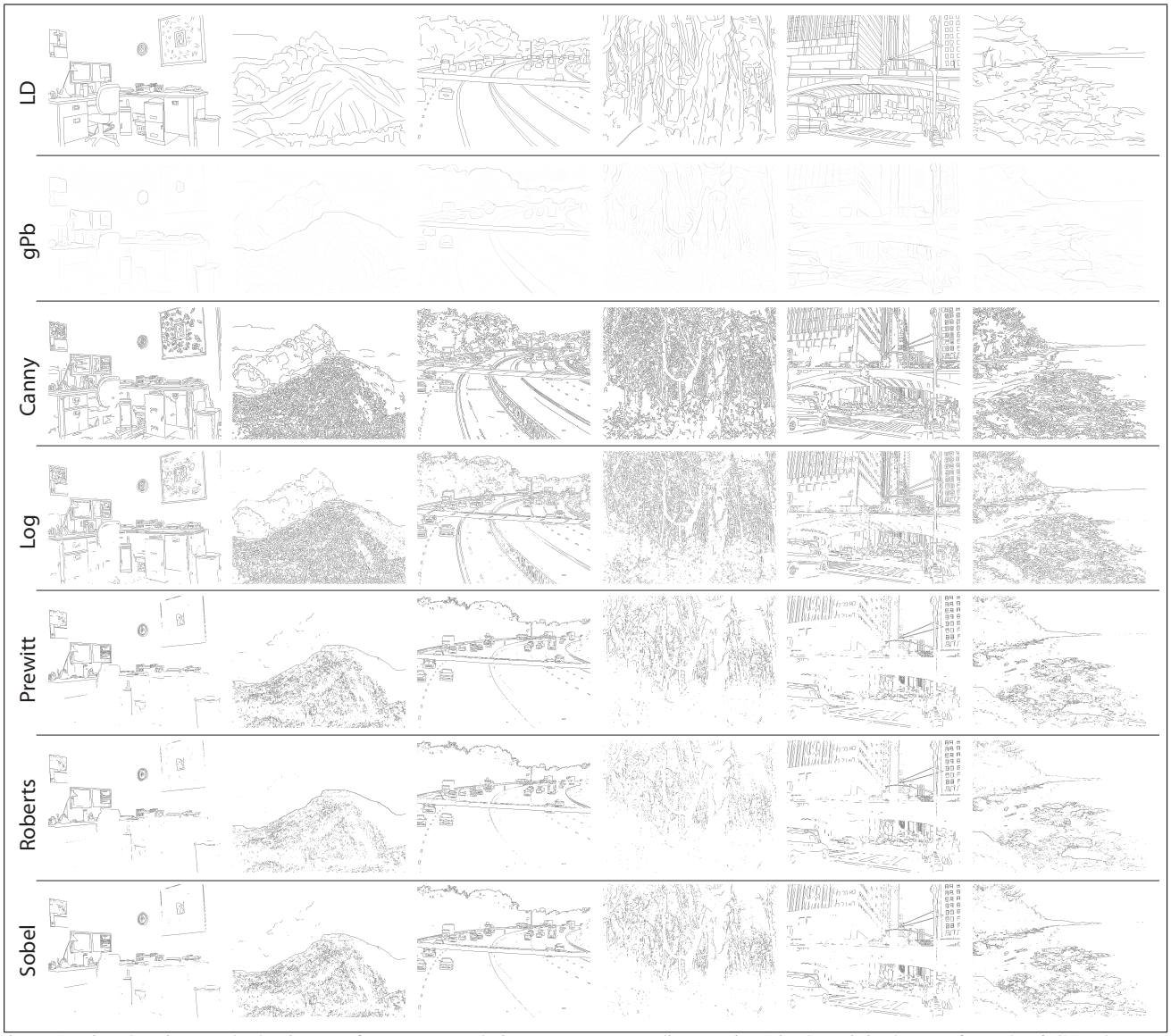

Figure 1. Line drawings and edge images for some sample images; corresponding to Fig. 1 in the original paper from the 6-CAT dataset.

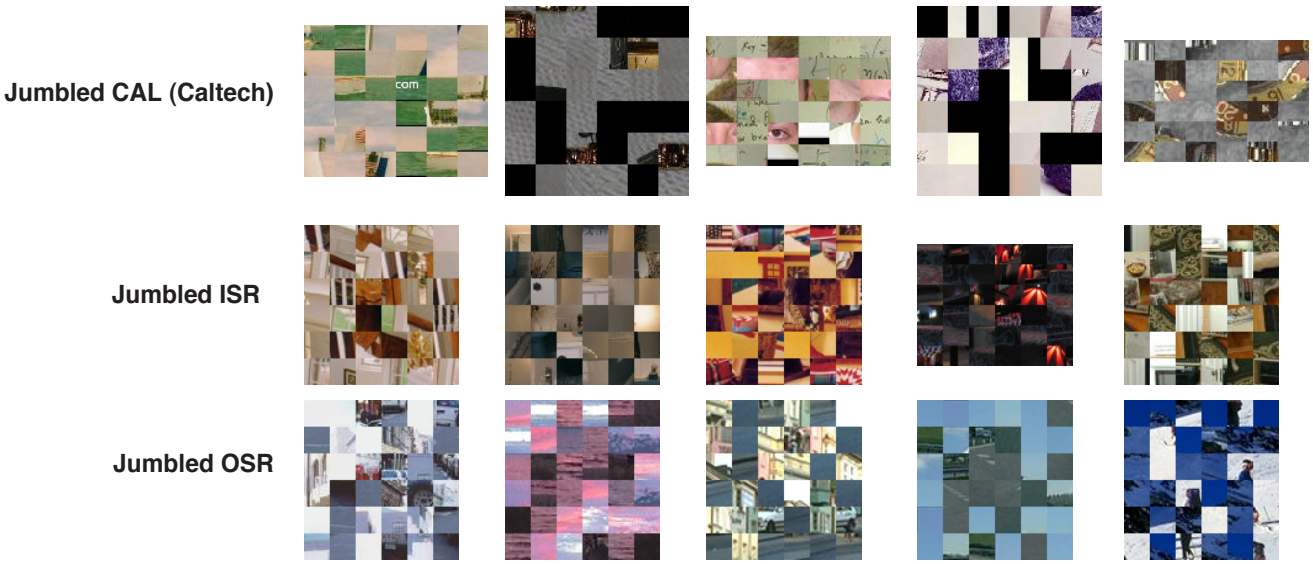

Figure 2. Sample images from the jumbled images dataset [5].

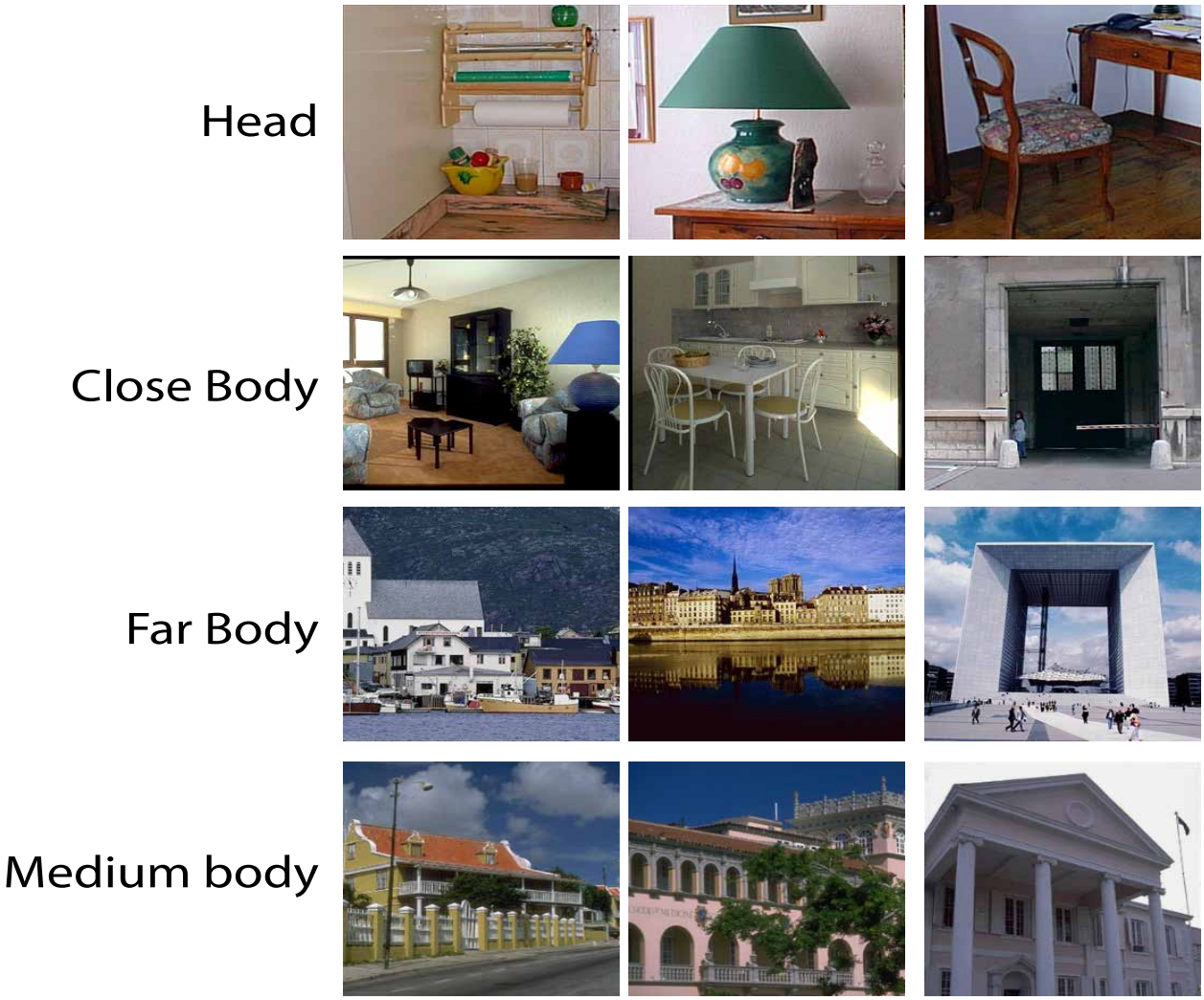

Figure 3. Sample distractor scenes corresponding to animal scenes in Fig. 1 of the main text (Test 3).

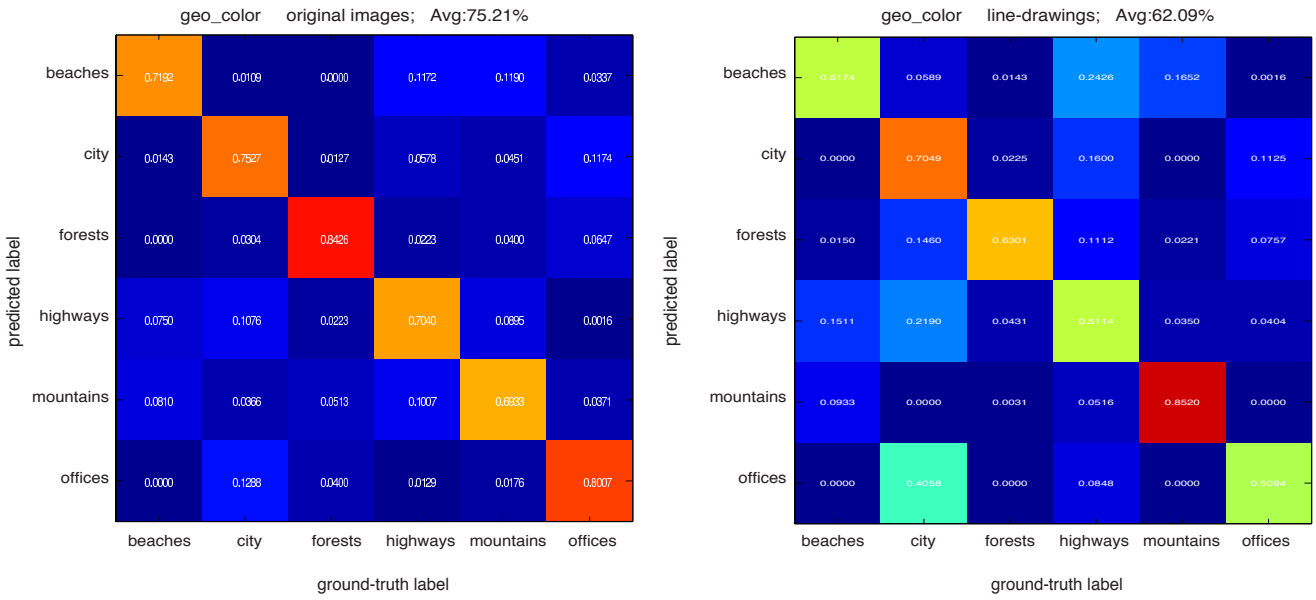

Figure 4. Confusion matrix of the geo_color model over color images (left) and line drawings (right) over the 6-CAT dataset.

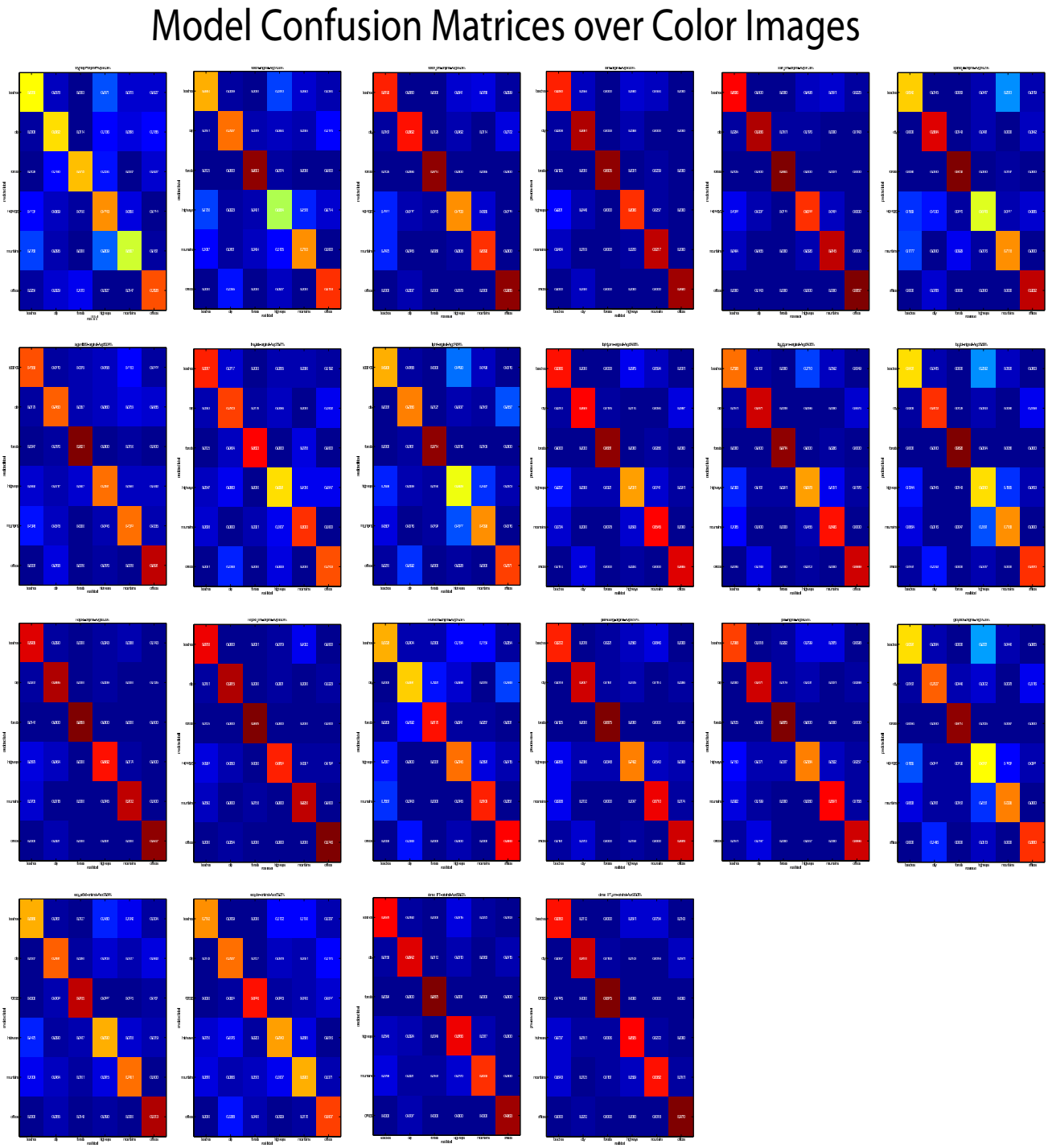

Figure 5. Confusion matrices of models over original scenes of the 6-CAT dataset.

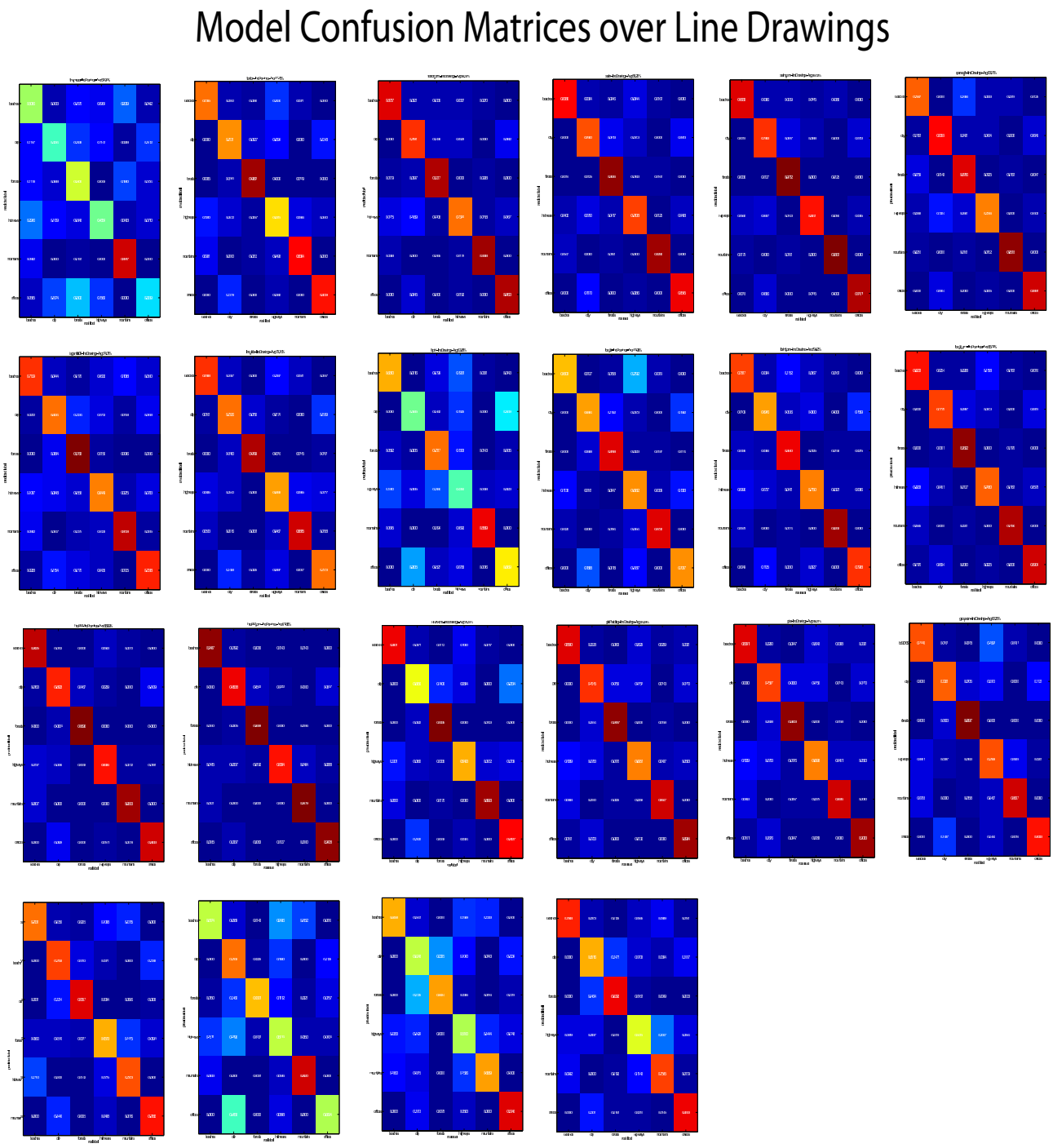

Figure 6. Confusion matrices of models over line drawings of the 6-CAT dataset.

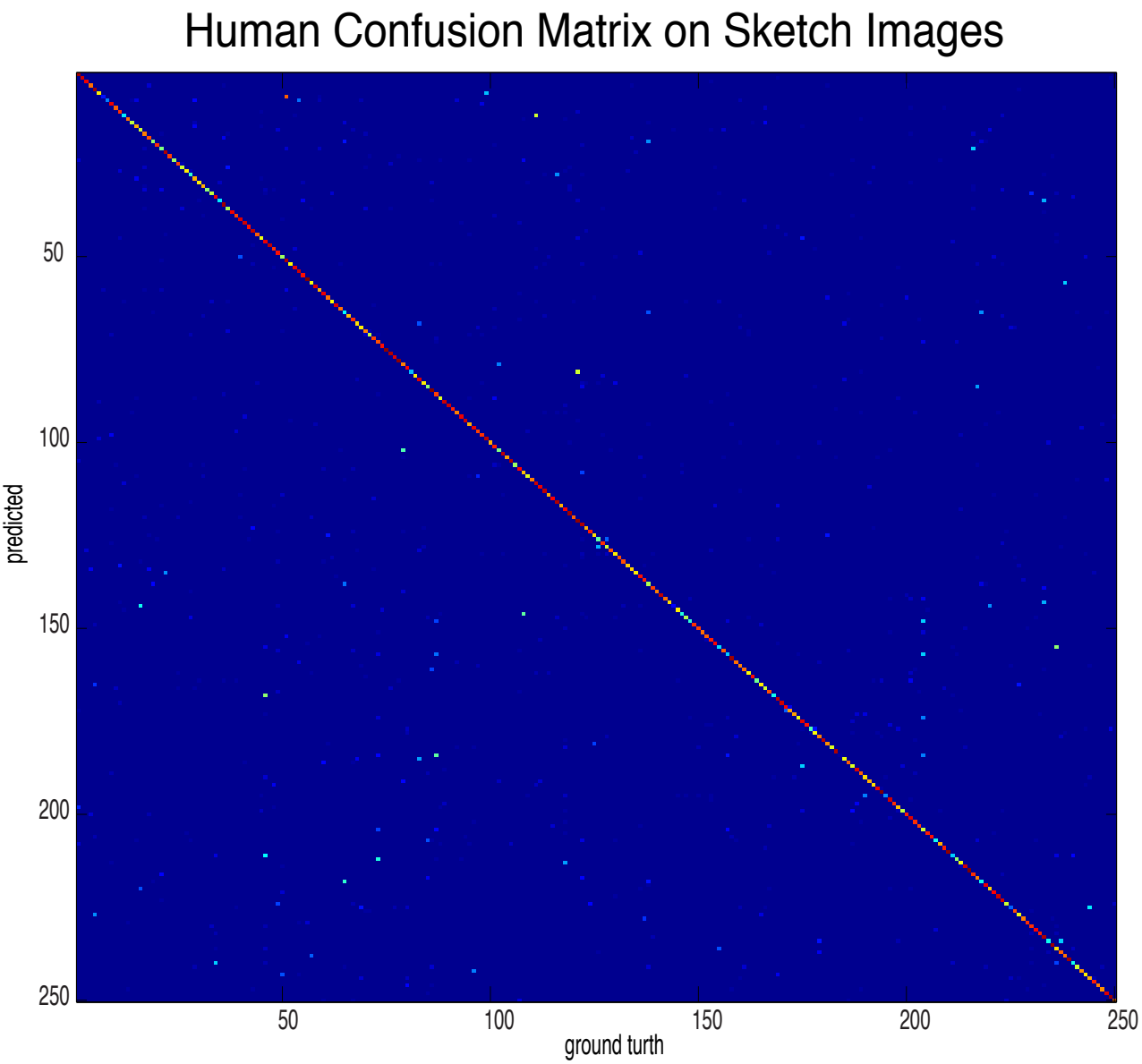

Figure 7. Human confusion matrix over sketch dataset [4].

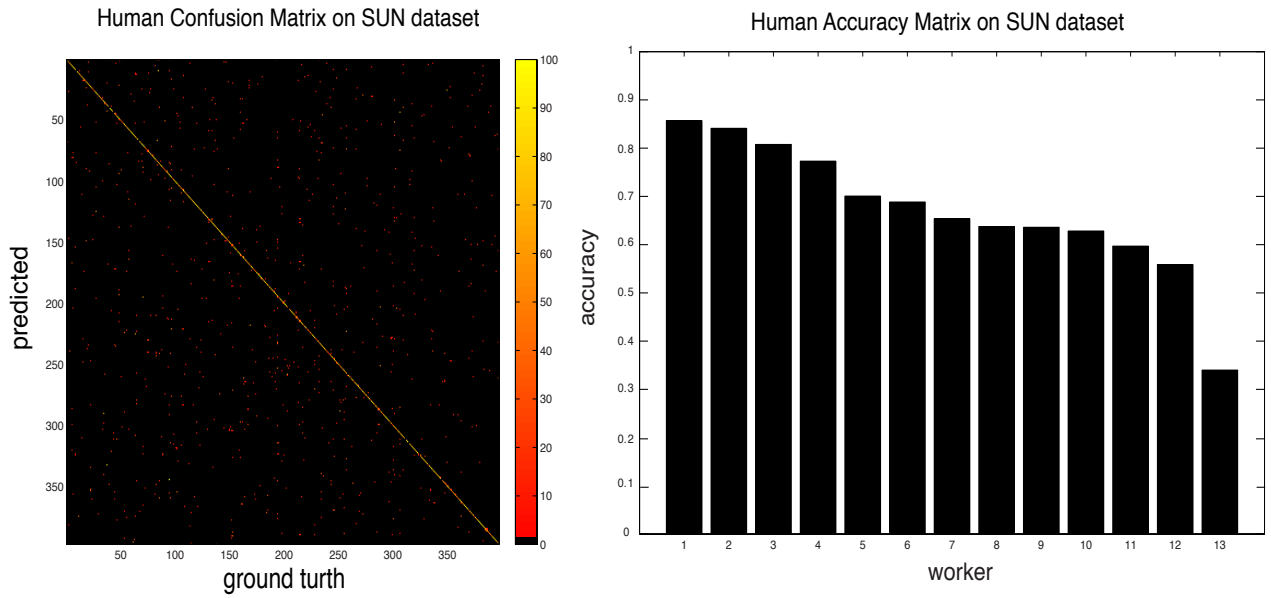

Figure 8. Left: Human confusion matrix over SUN dataset (only for 13 good workers). Right: Human scene classification accuracy over SUN dataset for individual workers.

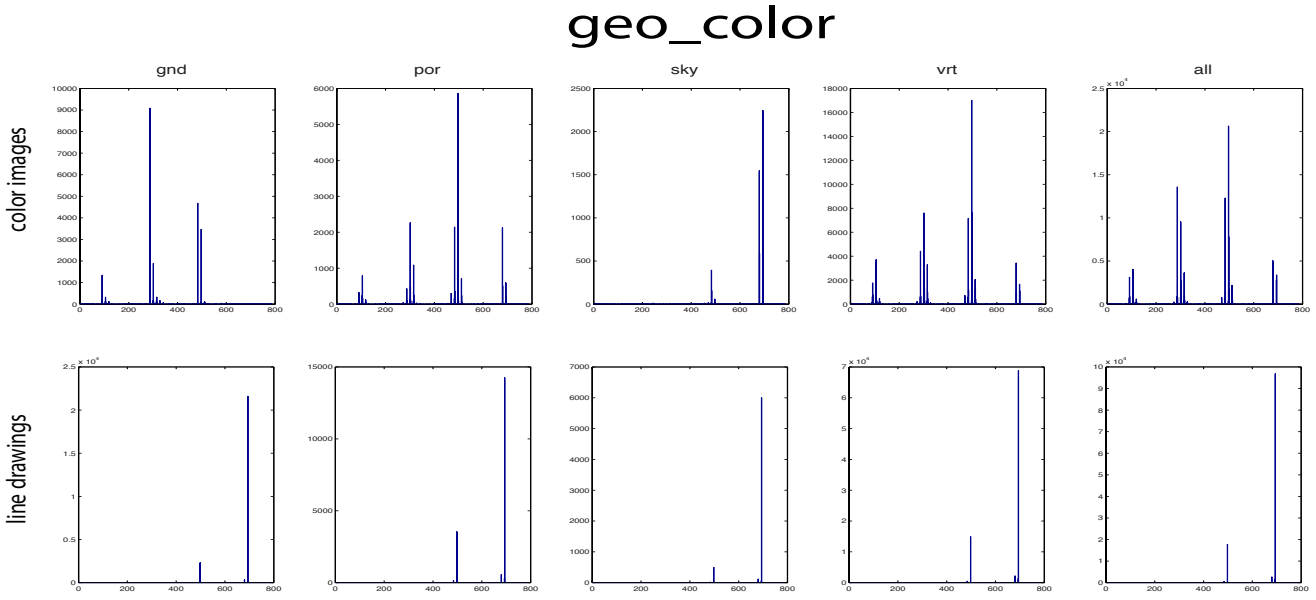

Figure 9. Histogram of geo_colro features for the images shown in Figure 5 of the main text (top: original scene; bottom: it's line drawing).

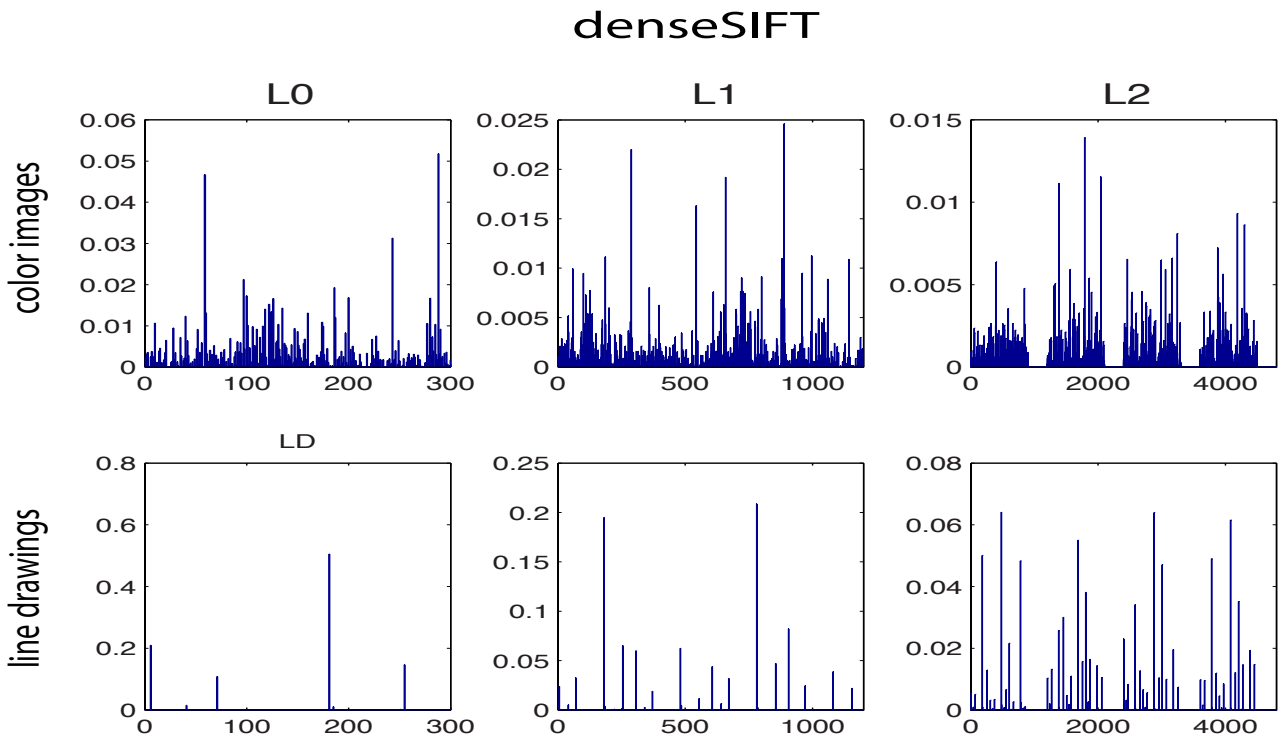

Figure 10. Histogram of denseSIFT features for the images shown in Figure 5 of the main text.

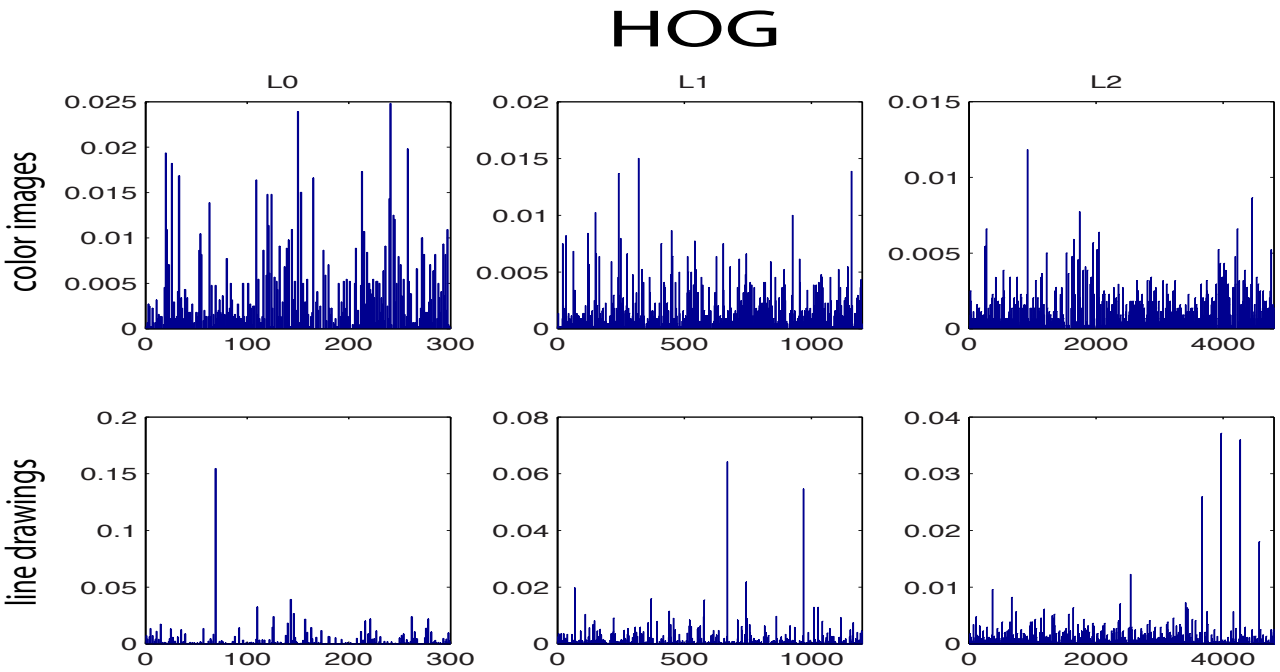

Figure 11. Histogram of HOG features for the images shown in Figure 5 of the main text.

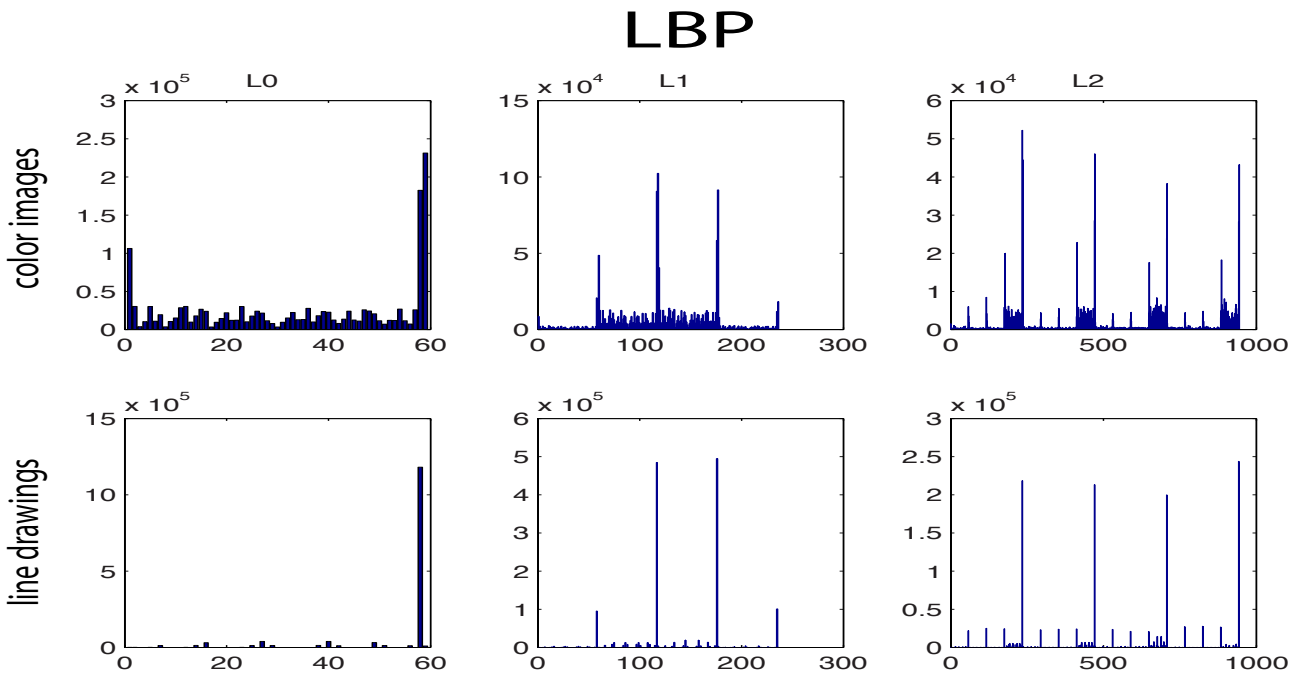

Figure 12. Histogram of LBP features for the images shown in Figure 5 of the main text.

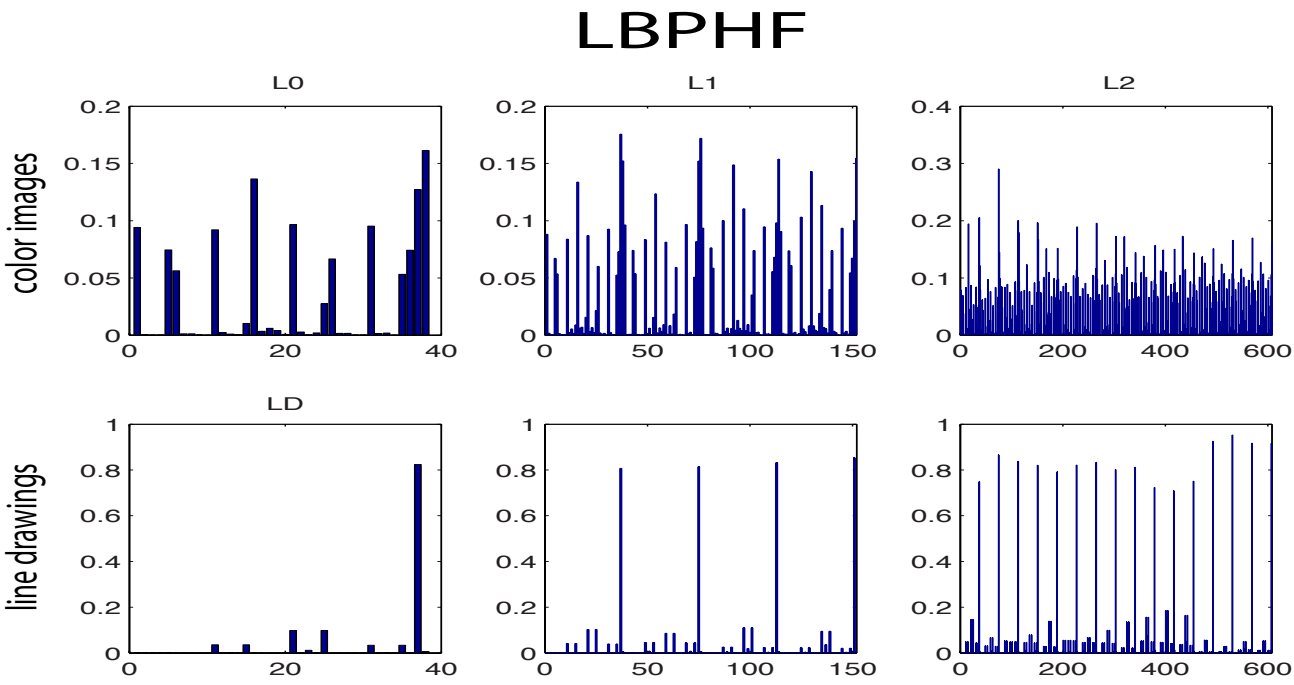

Figure 13. Histogram of LBPHF features for the images shown in Figure 5 of the main text.

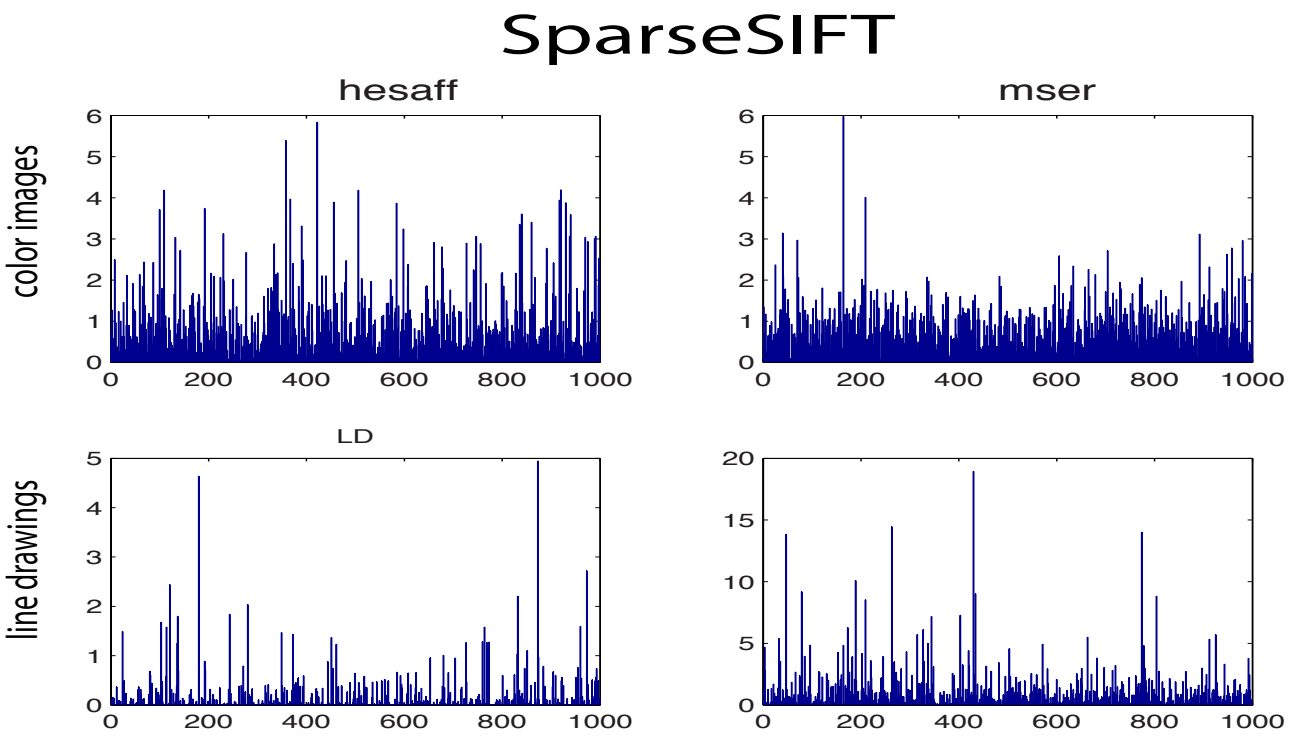

Figure 14. Histogram of sparseSIFT features for the images shown in Figure 5 of the main text.

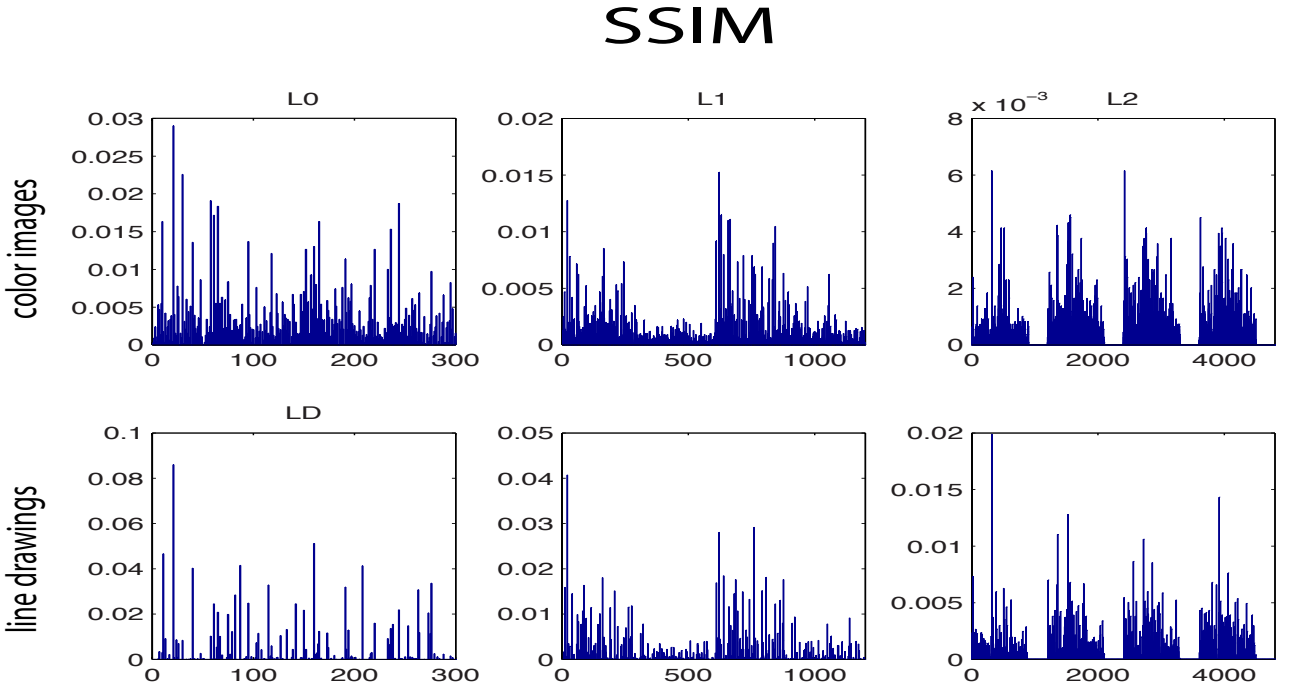

Figure 15. Histogram of SSIM features for the images shown in Figure 5 of the main text.

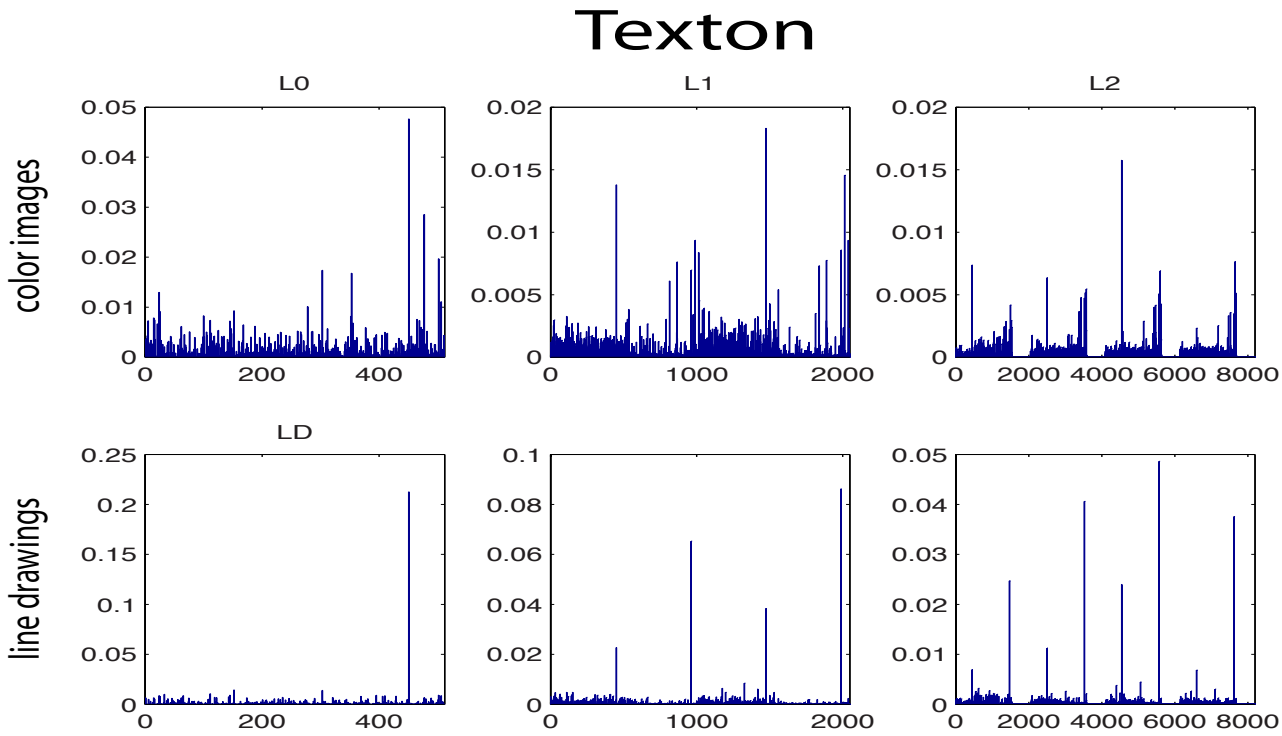

Figure 16. Histogram of Texton features for the images shown in Figure 5 of the main text.

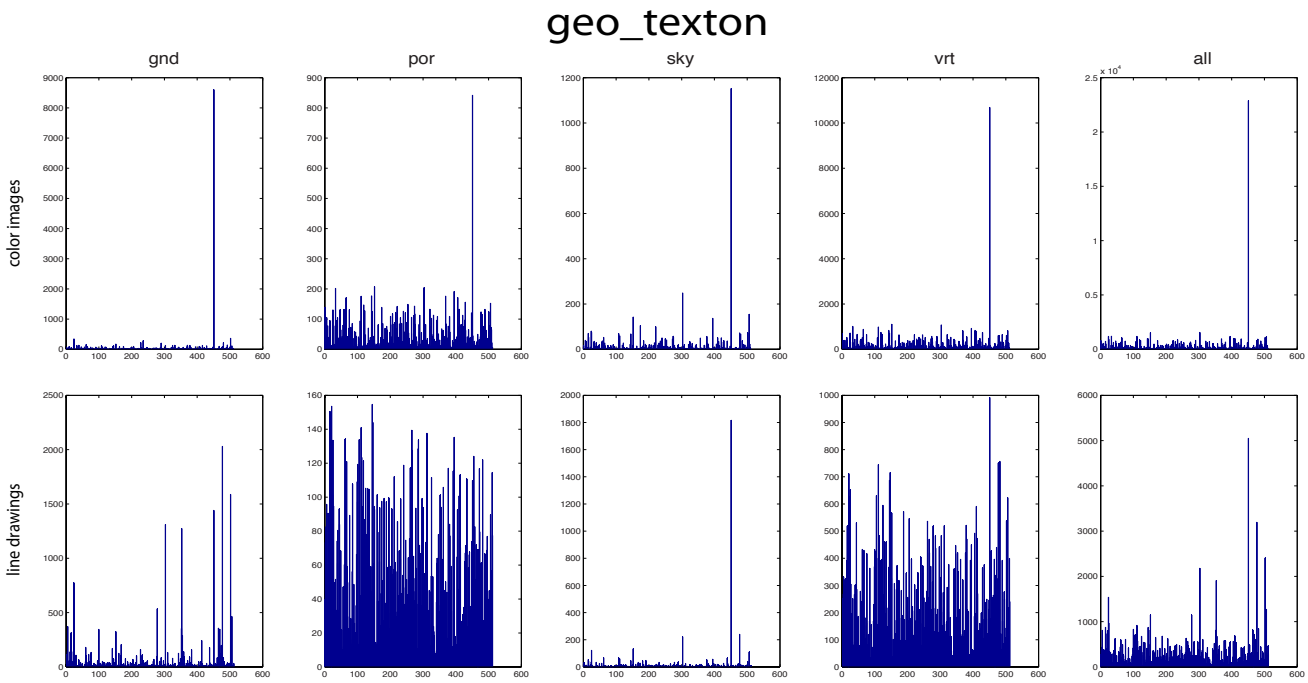

Figure 17. Histogram of geo_texton features for the images shown in Figure 5 of the main text.

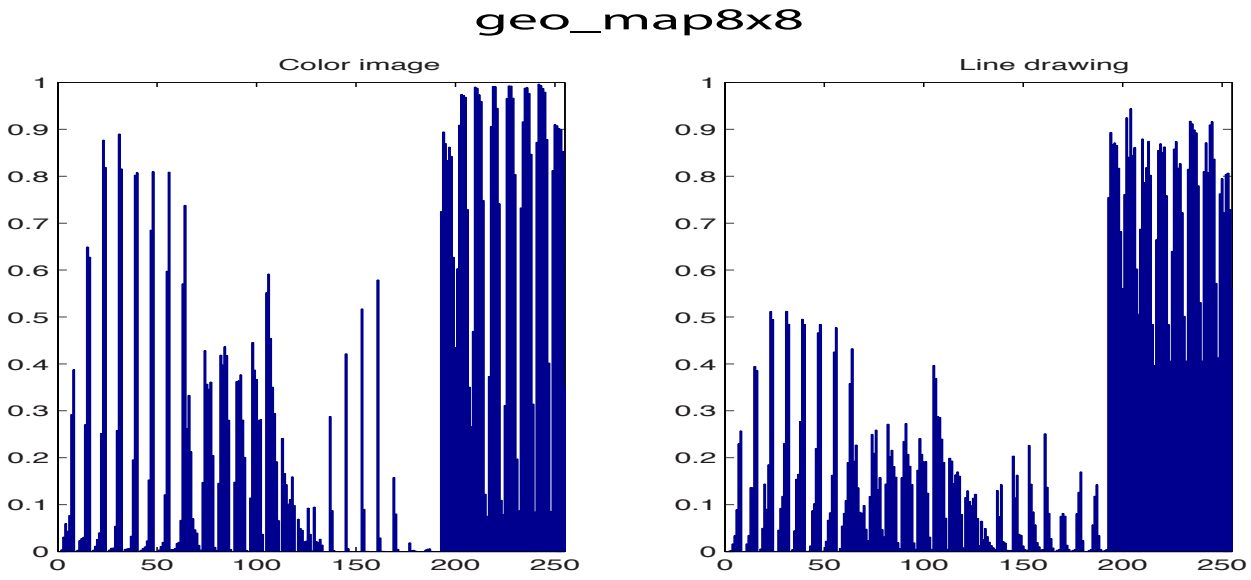

Figure 18. Histogram of geo_map features for the images shown in Figure 5 of the main text.

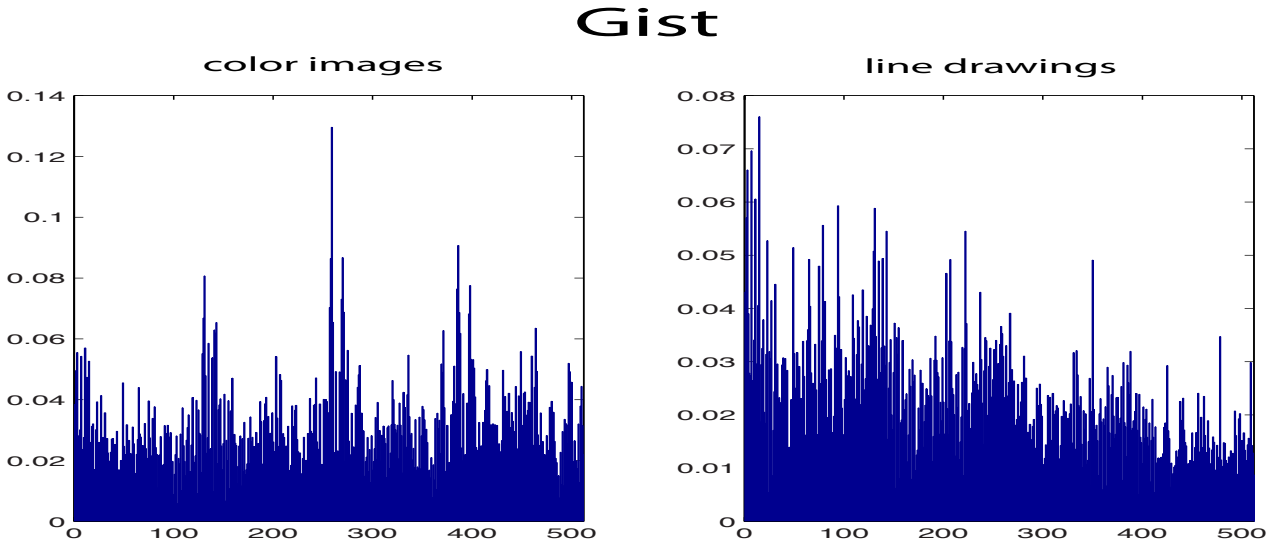

Figure 19. Histogram of GIST features for the images shown in Figure 5 of the main text.

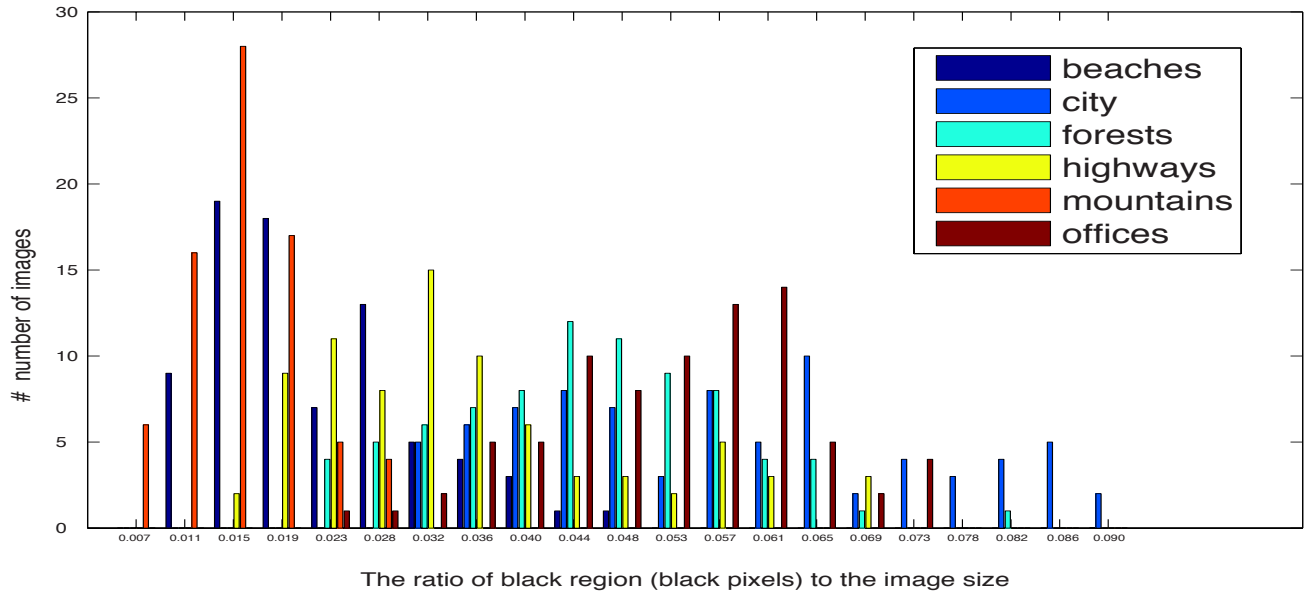

Figure 20. Histogram of ratio of black image regions (in binary line-drawings) over images of the 6-CAT dataset.

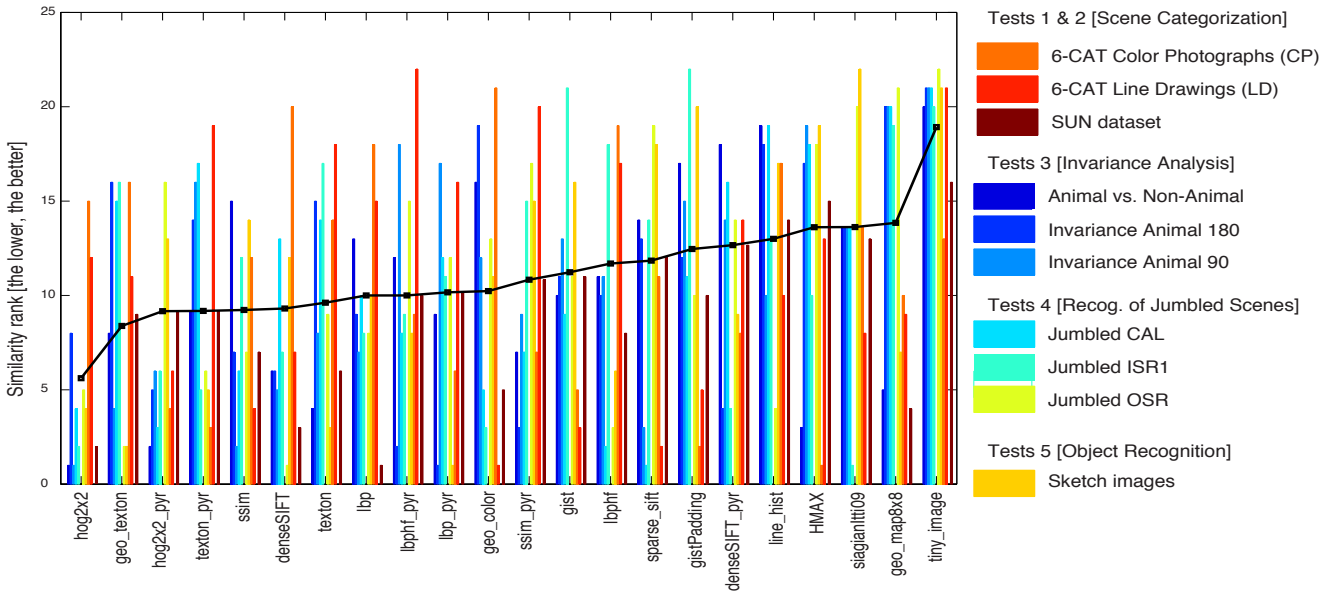

Figure 21. Similarity ranks of models over five tests as well as the average similarity rank over models. The lower the similarity rank, the better.