Learning about Canonical Views from Internet Image Collections Supplemental Material

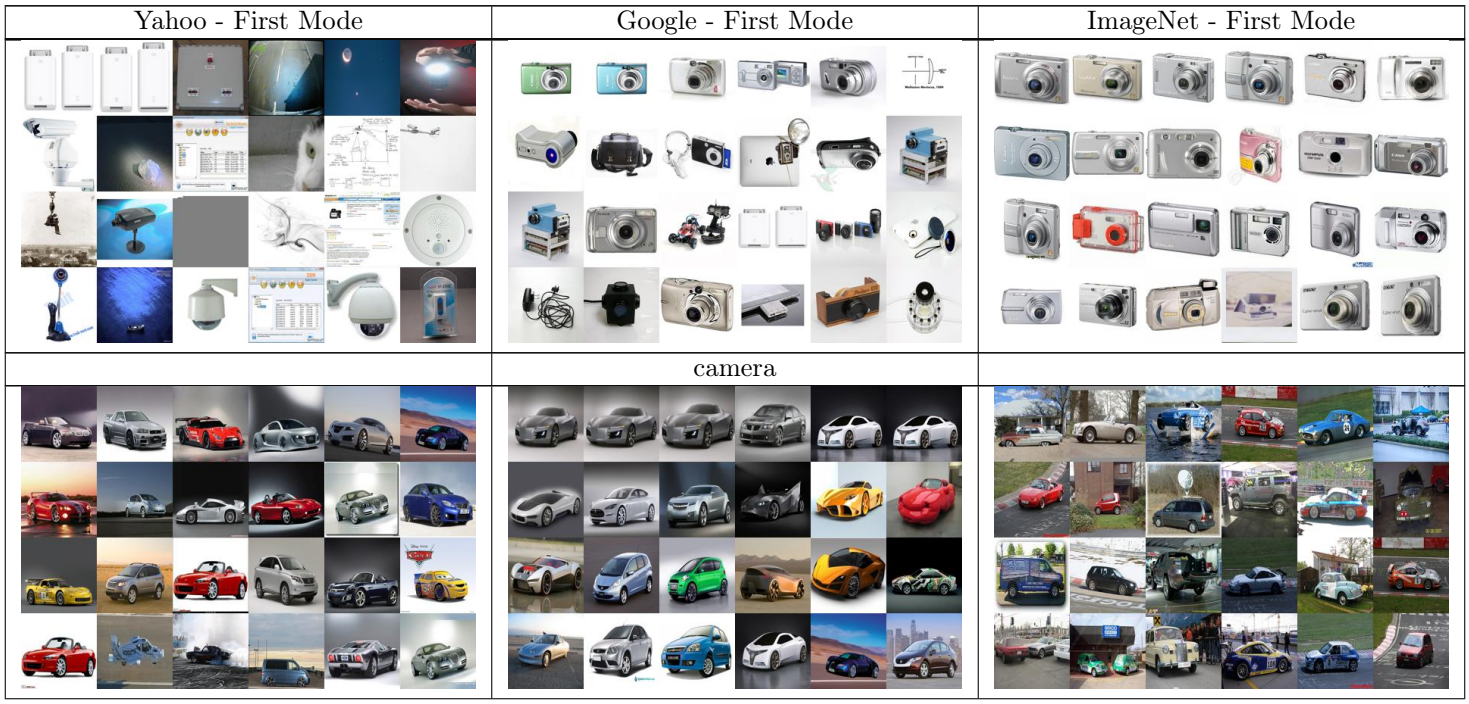
Elad Mezuman
Interdisciplinary Center for Neural Computation
Edmond & Lily Safra Center for Brain Sciences
Hebrew University of Jerusalem

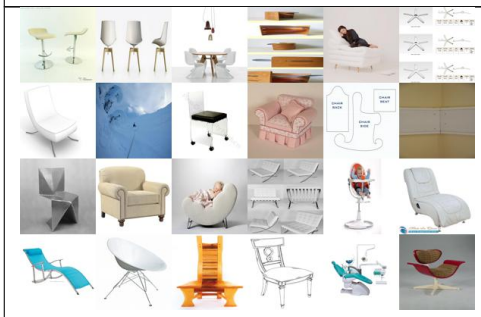
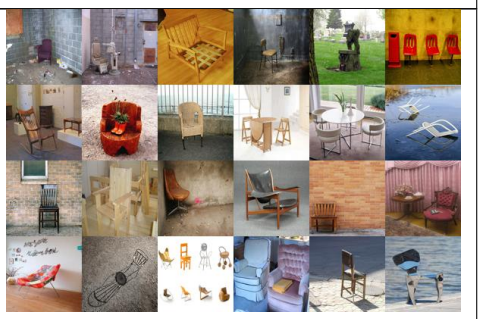
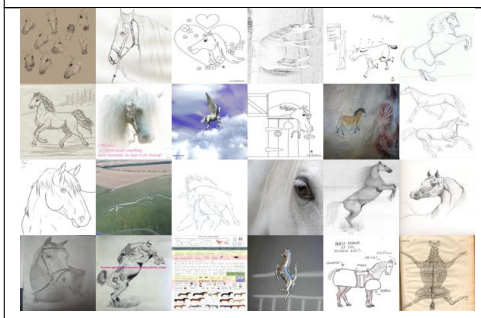
Yair Weiss
School of Computer Science and Engineering
Edmond & Lily Safra Center for Brain Sciences
Hebrew University of Jerusalem

In the supplemental material we present the results of our experiments. We present the output of the automatic phase before it was human validation phase, i.e. we present the modes in the GIST space; for some of the categories it might not correspond to a mode in the view space, thus it would have been rejected in the human validation step.

1. Section 1 contains comparison of our results on the categories for which Palmer *et al.* [1] found the canonical view, using human subjects ("Palmer's categories") for three different Internet image search engines - Yahoo, Google and ImageNet [2].
2. Section 2 contains results for "Palmer's categories" where we used cropped images from ImageNet [2]; the images were cropped using the bounding boxes ImageNet supplies for several categories. The first two modes found by our method are presented.
3. Section 3 contains results for categories inspired by the seminal work of Rosch *et al.* [3] ("Rosch's categories"), where they studied human recognition for categories in different level of abstraction where the source of images was ImageNet. Two first modes are presented.
4. Section 4 contains selected results for categories of mammals from ImageNet; the images were cropped first according to the bounding boxes ImageNet supplies.
5. Section 5 contains results for a subset of "Palmer's categories" for which we have downloaded images from two Google "local" sites (google.de and google.es) using the translated name of the object to the local language (German and Spanish).
6. Section 6 contains the control experiment for white background

1 "Palmer's categories", Yahoo vs. Google vs. ImageNet

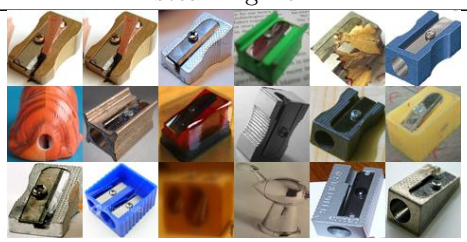
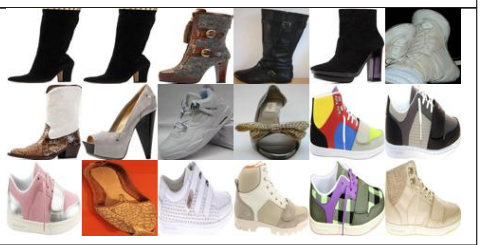


Yahoo - First Mode	Google - First Mode	ImageNet - First Mode
	car	
		
	chair	
		
	clock	
		
	horse	
		
	house	
		
	steaming iron	

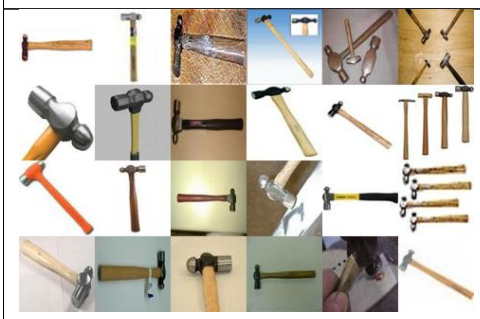
Yahoo - First Mode	Google - First Mode	ImageNet - First Mode
		
	manual pencil sharpener	
		
	piano	
		
	shoe	
		
	teapot	
		
	rotary telephone	

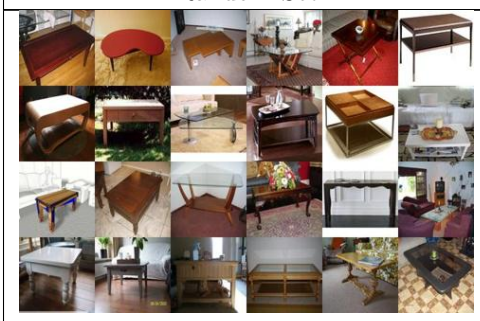
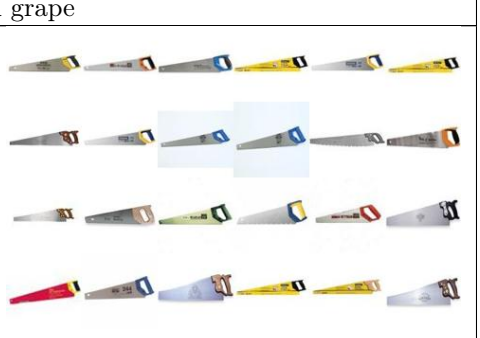
2 "Palmer's categories", cropped images from ImageNet, two first modes

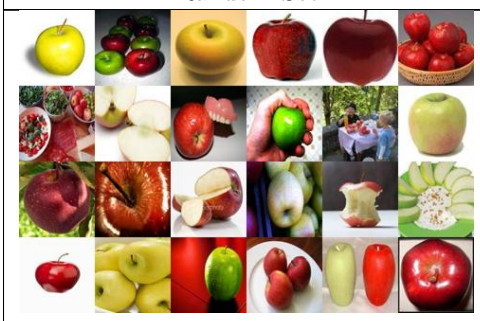
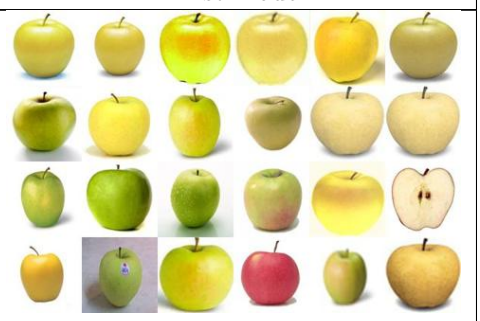
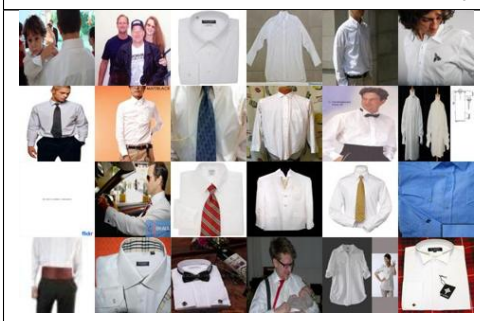
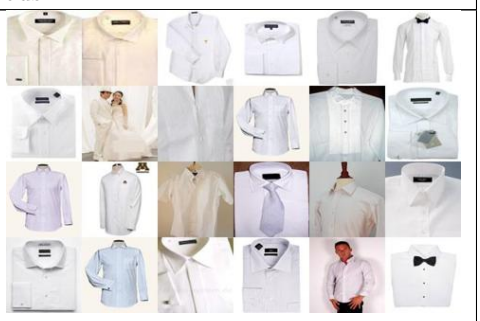
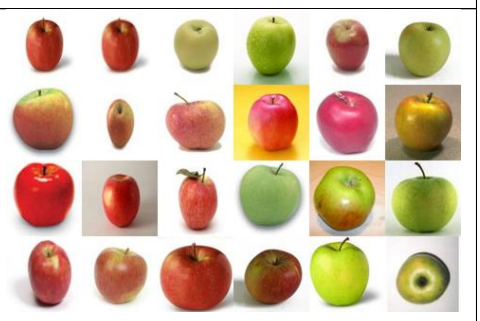
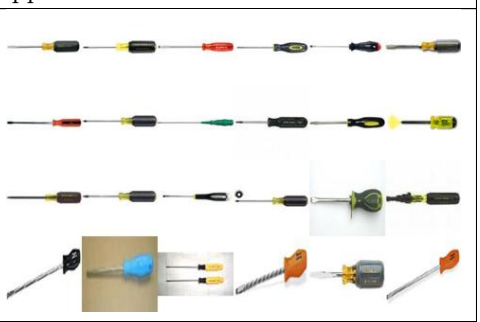
Random Set	First Mode	Second Mode
		
	camera	
		
	car	
		
	chair	
		
	clock	
		
	horse	
		
	house	

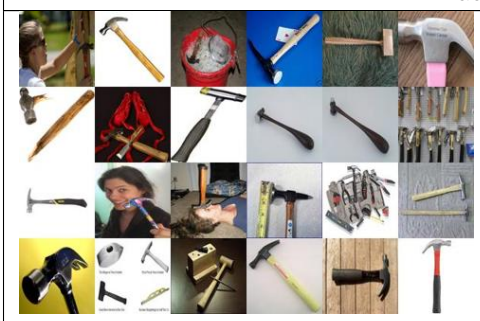
Random Set	First Mode	Second Mode
		
	steaming iron	
		
	manual pencil sharpener	
		
	piano	
		
	shoe	
		
	teapot	
		
	rotary telephone	

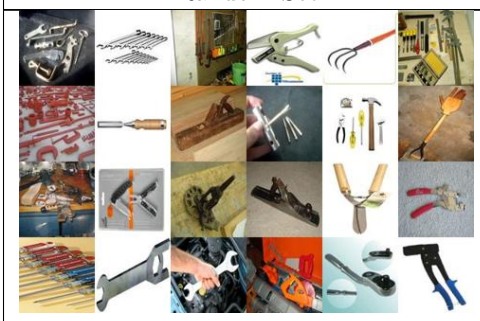
3 "Rosch's categories", source: ImageNet (full images)

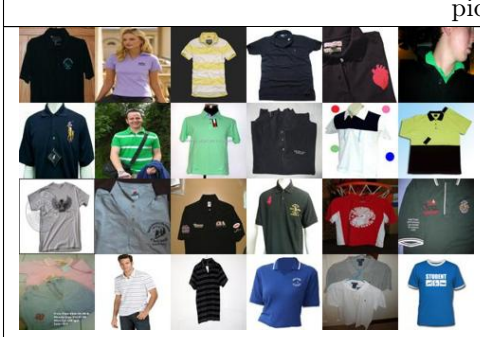
Random Set	First Mode
	
apple	
	
athletic sock	
	
ball-peen hammer	
	
bongo	
	
carpenter's hammer	

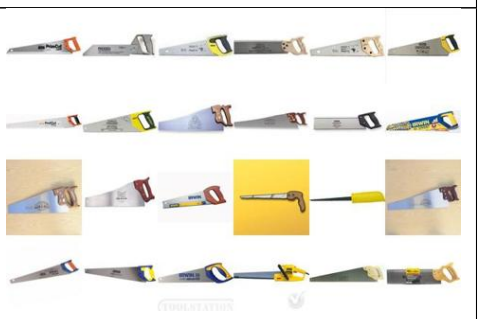
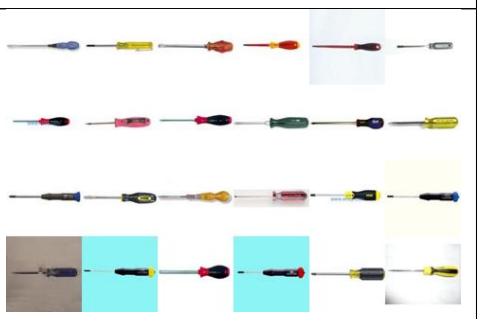
Random Set	First Mode
	
coffee table	
	
concert grand	
	
Concord grape	
	
crosscut saw	
	
damson	

Random Set	First Mode
	
Delicious	
	
dress shirt	
	
drum	
	
eating apple	
	
flat tip screwdriver	

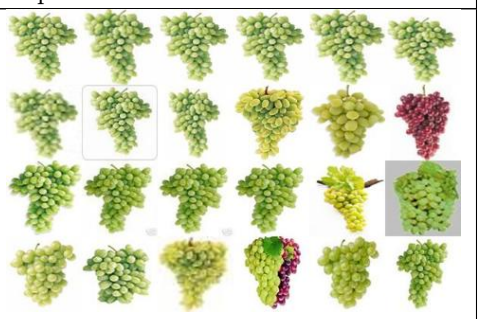
Random Set	First Mode
	
floor lamp	
	
fruit	
	
grand piano	
	
hacksaw	
	
hammer	

Random Set	First Mode
	
hand tool	
	
knee-high	
	
ladder-back	
	
lamp	
	
maple	

Random Set	First Mode
	
McIntosh	
	
motor vehicle	
	
Phillips screwdriver	
	
pickup	
	
polo shirt	

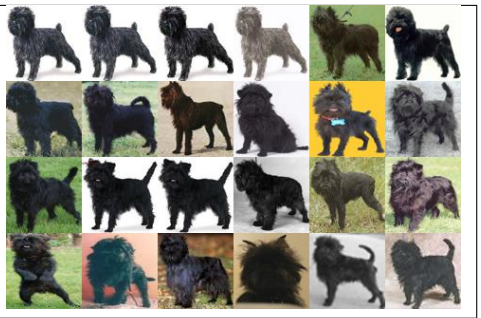
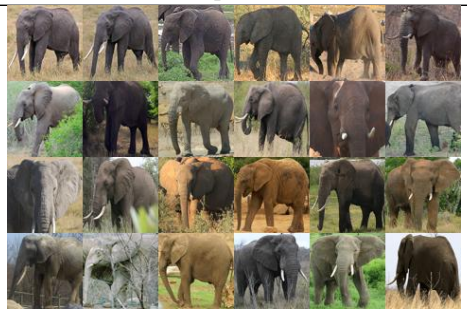
Random Set	First Mode
	
Red Delicious	
	
saw	
	
screwdriver	
	
sedan	
	
shirt	

Random Set	First Mode
	
silver maple	
	
slacks	
	
snare drum	
	
sock	
	
sugar maple	

Random Set	First Mode
	
swivel chair	
	
table lamp	
	
Thompson Seedless	
	
tool	
	
truck	

Random Set	First Mode
	
upright	
	
Victoria plum	

4 Mammals, selected categories, source: ImageNet (cropped images)

Random Set	First Mode	Second Mode
		
	affenpinscher	
		
	African elephant	
		
	Australian terrier	
		
	bison	
		
	black-and-tan coonhound	

bullock

cavy

cayuse

cheetah

chow

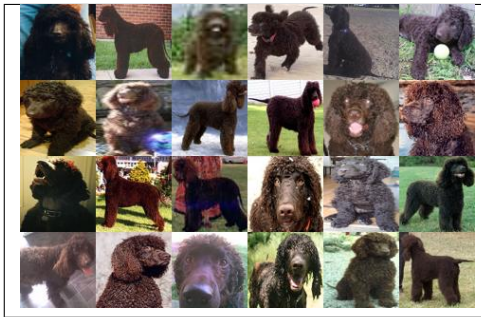
eland

flat-coated retriever

German shepherd

gnu

hamster



Irish water spaniel

kob

monotreme

Persian cat

pinche

pronghorn

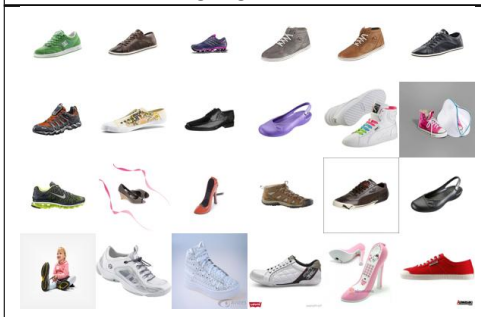
rodent

steeplechaser

tiger

uakari

5 “Palmer’s categories”, German vs. English vs. Spanish

google.de	google.com	google.es
		
schuh	shoe	zapatos
		
teekanne	teapot	tetera

(a) sedan, source: ImageNet

(b) shoe, source: ImageNet

Figure 1: Smooth background effect. Among images with smooth background there is a variation of views - indicating that the view in the image is more important than the smooth background for the method we used.

6 Control - Will a method that simply chooses images with uniform background can also find canonical views?

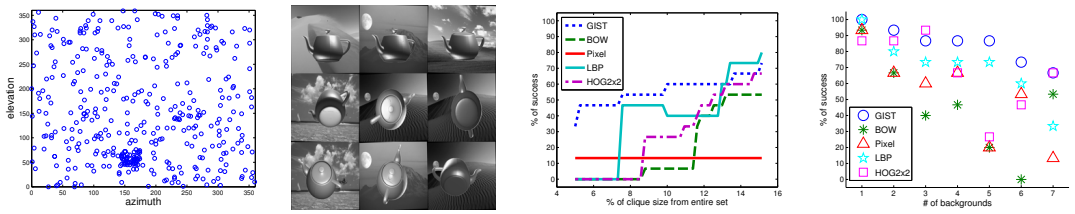
As can be seen in the collages we present for Palmer and Rosch categories, the most frequent view chosen by our method often has a white, or uniform background. Will a method that simply chooses images with uniform background can also find canonical views? Fig. 1 shows this is not the case. Among images with smooth backgrounds there is still a large variation in views.

7 Experiments on Synthetic Dataset

We created a synthetic dataset with images of teapots from different viewpoints (see Fig. 2 a and b) using matlab. In addition to images with random view we added a cluster of teapots all with similar azimuth and elevation (up to 30° variance in each, example images can be seen in the first row of Fig. 2b), this view is considered the preferred view and the ground truth for evaluation. The teapots were placed around the center of one out of 7 backgrounds.

Comparison Between Different Image Descriptors In the first set of experiments we compared different image descriptors; we wanted to check the distance of which can serve as a proxy for view distance. The descriptions we used are: GIST [4], Bag Of Visual Words [5], simple pixel image descriptor, Local Binary Patterns (LBP) descriptor [6] and HOG2x2 [7]. In each experiment, we approximated the most frequent view using Parzen estimator on one of the five descriptors' spaces. In addition to the most frequent image in the specific descriptor's space we took the 14 images closest to it (L2 distance in the descriptor space). The percent of success is the percent of images out of these 15 images which are from the ground truth view. In the first experiment we varied the ratio of images from the preferred view to the total number of images in the dataset. The results are presented in Fig. 2c. The performance achieved using GIST descriptors is much better than the performance achieved using the other descriptors. It also can be seen that even when the percent of the ground truth images out of the entire set is relatively small (6%) 50% of the images selected are from the ground truth canonical view, and when the set size increases the performances improves.

In the second experiment we wanted to test the effect of the difficulty of the dataset and the effect of the background on the results. We created 7 datasets, consisting of 350 images from a random view and 50 more from the preferred view. Each dataset had a different number of backgrounds from 1-7. The results are shown in Fig. 2d. Again the performances achieved using the GIST descriptor are good and better than when using other descriptors. We can see that indeed the performance decreases when the number of backgrounds increases but stays above 60%.



(a) (b) (c) (d)

Figure 2: Finding an image descriptor which captures the viewpoint of the image. We created a dataset containing images of teapots from random views plus a clique of teapot images from around the same view - which is considered the ground truth preferred view. (a) Distribution of 400 teapot images by viewpoint, the most likely view is around azimuth 160° and elevation 60°. (b) Example of the images in the synthetic dataset, images from the preferred view are in the first row. (c-d) Comparison of success for 5 image descriptors, GIST (blue), Bag of Words (green), simple pixel descriptor (red), LBP (cyan) and HOG2x2 (magenta). The matching percentage of the most likely image and its 14 closest neighbors to the ground truth is presented for (c) different size of the set of ground truth images (background number fixed to 7) and (d) different number of backgrounds in the dataset (set size fixed to 12.5%).

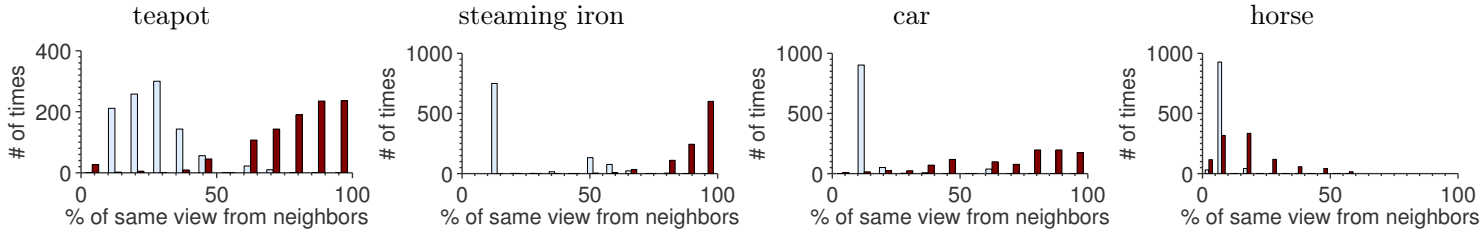


Figure 3: Discriminating between image sets with, or without, a preferred view. Histograms of the percentage of images among the nearest neighbors of the most frequent image, which share the same view with it. The color of the bars indicates whether :(red) preferred view existed in the imageset or (sky) was not.

As a final validation, we checked in which percentage of the experiments the most frequent view chosen using the different descriptors was indeed from the ground truth view. We did it for both the first and second experiments under all conditions (58 in total). When using the GIST the most frequent view chosen was in 100% of the cases from the ground truth view. For the other descriptors we got: LBP 83%, HOG2x2 72%, BOW 50% and pixel descriptor 9%.

We concluded that the GIST similarity is the best proxy for view similarity out of the ones we tested. Our next experiments where done using GIST descriptor.

Can we find the most frequent view directly from images? The second batch of experiments was done on the synthetic dataset and on four real world categories from Google("car", "horse" and “steaming iron”), for which we manually found the view of the object in each image (see section 2 in the paper). In each of the datasets we conducted 1,000 experiments where we sampled 200 images according to the original distribution of views in each dataset (for the teapot the size of the preferred view group was 15%). On the teapot, steaming iron and car datasets the most frequent image found using our method was from the ground truth view in 94%, 99% and 95% (respectively) of the experiments. On the horse category the method usually did not find the preferred view (12%).

Can we know if a preferred view exist? Next, we validated the second phase of our algorithm. In this phase a human is required to look on the set of images of the modes found using the Parzen density estimator on GIST space and decide if a frequent view was found; he does it by deciding if most of the images are from the same view. We compared the percent of images that are from the same view in the first mode in two kinds of experiments: (1) experiments in which the images where sampled from the original distribution and (2) experiments where the images were sampled uniformly among the views - in this experiment no preferred view existed (for the car and horse categories we used 100 images which were classified as 'noise or unique view'). We used 9 nearest neighbors (k). Distributions of the percent of images in the first mode that are from the same view in the 2 sets of experiments, over the 1,000 trials, are presented in Fig. 4. These results indicate that if we run our method and 50% or more of the most frequent image and its nearest neighbors are from the same view, we can say with confidence: There is a preferred view in this image set and we have found it. If we find a smaller percentage we cannot be sure whether it is because no preferred view exist in the image set or the signal is not strong enough (like in the horse category). Thus, when our method is used on several categories it underestimates the number of categories with preferred view.

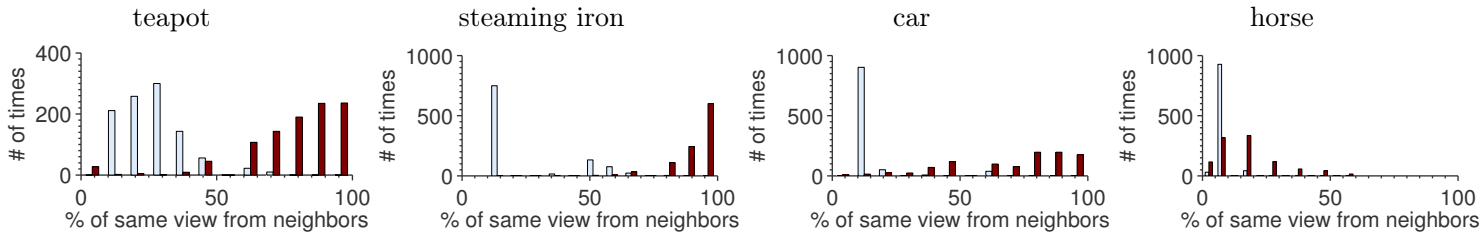


Figure 4: Discriminating between image sets with, or without, a preferred view. Histograms of the percentage of images among the nearest neighbors of the most frequent image, which share the same view with it. The color of the bars indicates whether :(red) preferred view existed in the imageset or (sky) was not.

References

[1] S. Palmer, E. Rosch, and P. Chase. Canonical perspective and the perception of objects. *Attention and performance IX*, pages 135–151, 1981. 1

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 1

[3] E. Rosch, C.B. Mervis, W.D. Gray, D.M. Johnson, and P. Boyes-Braem. Basic objects in natural categories* 1. *Cognitive psychology*, 8(3):382–439, 1976. 1

[4] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. 22

[5] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR 2006*. 22

[6] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on PAMI*, 2002. 22

[7] J. Xiao, J. Hays, K.A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR 2010*. 22