Novel Dataset for Fine-Grained Image Categorization: Stanford Dogs

Aditya Khosla Nityananda Jayadevaprakash Bangpeng Yao Fei-Fei Li
Computer Science Department, Stanford University, Stanford, CA
{aditya86, bangpeng, feifeili}@cs.stanford.edu nityananda@gmail.com

1. Introduction

We introduce a 120 class Stanford Dogs dataset, a challenging and large-scale dataset aimed at fine-grained image categorization. Stanford Dogs includes over 22,000 annotated images of dogs belonging to 120 species. Each image is annotated with a bounding box and object class label. Fig. 1 shows examples of images from Stanford Dogs. This dataset is extremely challenging due to a variety of reasons. First, being a fine-grained categorization problem, there is little inter-class variation. For example the basset hound and bloodhound share very similar facial characteristics but differ significantly in their color, while the Japanese spaniel and papillion share very similar color but greatly differ in their facial characteristics. Second, there is very large intra-class variation. The images show that dogs within a class could have different ages (e.g. *beagle*), poses (e.g. *blenheim spaniel*), occlusion/self-occlusion and even color (e.g. *Shih-tzu*). Furthermore, compared to other animal datasets that tend to exist in natural scenes, a large proportion of the images contain humans and are taken in man-made environments leading to greater background variation. The aforementioned reasons make this an extremely challenging dataset.

1.1. Comparison to Other Datasets

There have been a number of other datasets used for fine-grained visual categorization [6] including Caltech-UCSD 200 Birds (CUB-200) dataset [4], PASCAL Action Classification [2] and People-Playing Musical Instruments (PPMI) [5]. Tbl. 1 shows some properties of existing datasets in comparison with our proposed dataset. Unlike previous datasets, ours consists of a large number of classes (120) with a large number of images per class (150-200).

This allows for rigorous testing of algorithms under various experimental settings. It would allow us to identify the dependence of algorithms on the amount of data available per class. This can also allow us to test the limitations of the fine-grained visual categorization problem. Can the performance be improved significantly with more data? Can existing object recognition algorithms be used without modification if provided with sufficient data? Is the performance of proposed algorithms limited by the size of data or design of algorithm? These are some of the questions we hope to be able to address more adequately using this dataset by applying the training and testing techniques described in Sec 3.

Dataset	No. of classes	No. of images	Images per class	Visibility varies?	Bounding boxes?
CUB-200 [4]	**200**	6033	30	**Yes**	**Yes**
PPMI [5]	24	4800	**200**	No	**Yes**
PASCAL [2]	9	1221	**135**	**Yes**	**Yes**
Stanford Dogs	**120**	**20580**	**180**	**Yes**	**Yes**

Table 1. Comparison of our data set and the other existing fine-grained categorization datasets on still images. "Visibility" variation refers to the variation of visible body parts of the humans/animals in the dataset, e.g. in some images the full human body is visible, while in some other images only the head and shoulder are visible. Bold font indicates relatively larger scale datasets or larger image variations.

2. Image Collection And Annotation

The images and bounding boxes were downloaded from ImageNet [1]. The classes were selected to be leaf nodes, under the *'Canis familiaris'* node, that contain a single species of dogs. Nodes containing images from multiple species (e.g. *puppy*) were removed. Only images of $200 * 200$ pixels or larger were kept. Each image was examined to confirm whether or not it matched images from Wikipedia and shared similar features to the other images in the same category. Degenerate or unusual images (distorted colors, very blurry or noisy, largely occluded, extreme close-ups) were removed manually. All duplicated images, within and between categories, were removed. The bounding boxes on ImageNet [1] are annotated and verified through Amazon Mechanical Turk.

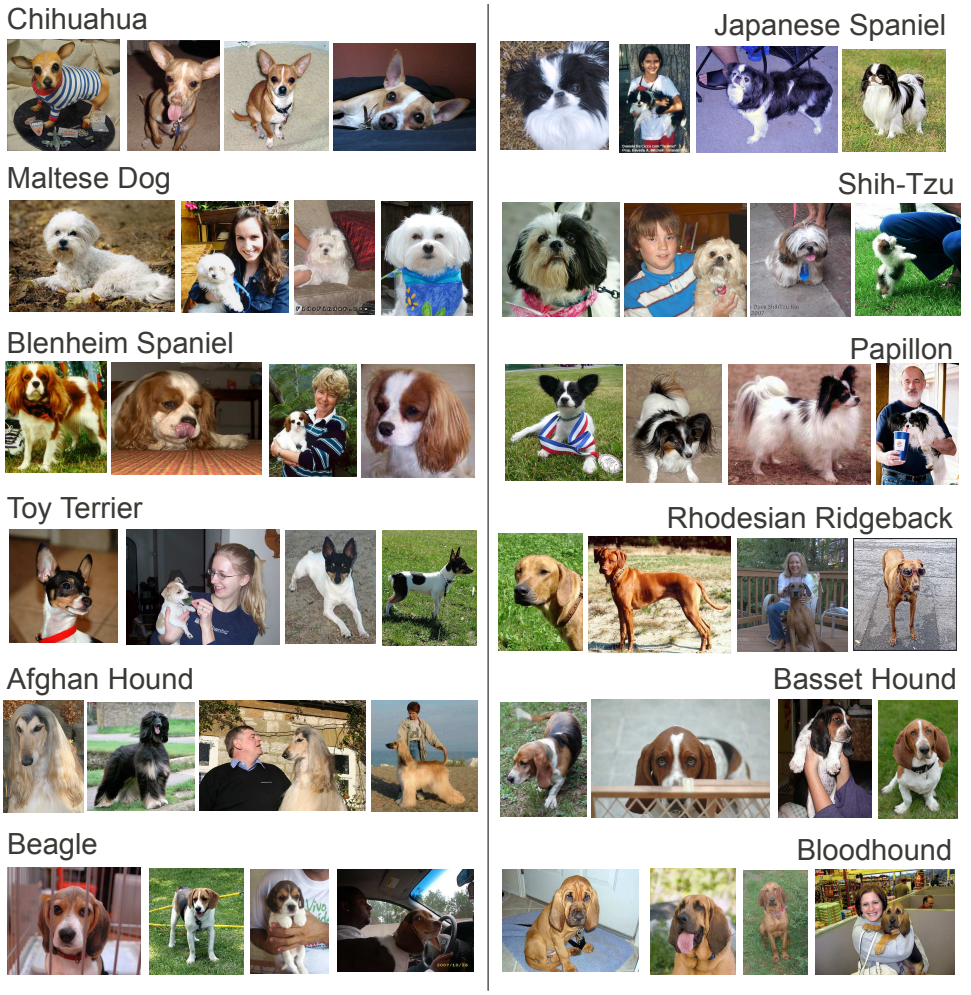

Figure 1. Four random example images from 12 of the 120 dog categories. We observe significant pose/visibility variation and background clutter in all the classes. In addition, there is large intra-class variation in appearance of the dogs. For example, the same species don't share a consistent color of fur or are of different sizes because of age (i.e. puppy vs fully-grown). Furthermore, their appearance is often modified by humans by placing articles of clothing or cutting/growing the fur. Images from all categories are available for viewing and download on the dataset website.

3. Training and Testing

We split the database into training/testing data and specify our evaluation methodology. This will allow for the testing of the amount of data required for each algorithm, and identify their region of peak performance. For each class, 100 images are used for training, and the remaining are used for testing (at least 50). The training and testing splits are fixed and available on the dataset website.

In addition, we follow a similar training/testing methodology as Caltech-101 [3]. We vary the number of training images (N_{train}) used while keeping the test set fixed. We propose the use of $N_{train} = \{15, 30, 60, 100\}$ i.e. we randomly sample a set of N_{train} images per class from the complete training set. When $N_{train} < 100$, the experiment is repeated 10 times to produce an average result. The identity of the training examples used in each case is available on the website to ensure that all results are directly comparable. All information related to the dataset, together with baseline results are available at the dataset website:

http://vision.stanford.edu/aditya86/StanfordDogs/

References

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

[2] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results.

[3] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In *Workshop on Generative-Model Based Vision*, 2004.

[4] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical report, California Institute of Technology, 2010.

[5] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, San Francisco, CA, June 2010.

[6] B. Yao*, A. Khosla*, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *CVPR*, Colorado Springs, CO, June 2011. (*-indicates equal contribution).