Cats and Dogs

Omkar M Parkhi[1,2] Andrea Vedaldi[1] Andrew Zisserman[1] C. V. Jawahar[2]

[1]Department of Engineering Science,
University of Oxford,
United Kingdom
{omkar,vedaldi,az}@robots.ox.ac.uk

[2]Center for Visual Information Technology,
International Institute of Information Technology,
Hyderabad, India
jawahar@iiit.ac.in

Abstract

We investigate the fine grained object categorization problem of determining the breed of animal from an image. To this end we introduce a new annotated dataset of pets, the Oxford-IIIT-Pet dataset*, covering 37 different breeds of cats and dogs. The visual problem is very challenging as these animals, particularly cats, are very deformable and there can be quite subtle differences between the breeds.*

We make a number of contributions: first, we introduce a model to classify a pet breed automatically from an image. The model combines shape, captured by a deformable part model detecting the pet face, and appearance, captured by a bag-of-words model that describes the pet fur. Fitting the model involves automatically segmenting the animal in the image. Second, we compare two classification approaches: a hierarchical one, in which a pet is first assigned to the cat or dog family and then to a breed, and a flat one, in which the breed is obtained directly. We also investigate a number of animal and image orientated spatial layouts.

These models are very good: they beat all previously published results on the challenging ASIRRA test (cat vs dog discrimination). When applied to the task of discriminating the 37 different breeds of pets, the models obtain an average accuracy of about 59%, a very encouraging result considering the difficulty of the problem.

1. Introduction

Research on object category recognition has largely focused on the discrimination of well distinguished object categories (*e.g*, airplane vs cat). Most popular international benchmarks (*e.g*, Caltech-101 [22], Caltech-256 [26], PASCAL VOC [20]) contain a few dozen object classes that, for the most part, are visually dissimilar. Even in the much larger ImageNet database [18], categories are defined based on a high-level ontology and, as such, any visual similarity between them is more accidental than systematic. This work concentrates instead on the problem of *discriminating different breeds of cats and dogs*, a challenging example of fine grained object categorization in line with that of previous work on flower [15, 32, 33, 39] and animal and bird species [14, 27, 28, 43] categorization. The difficulty is in the fact that breeds may differ only by a few subtle phenotypic details that, due to the highly deformable nature of the bodies of such animals, can be difficult to measure automatically. Indeed, authors have often focused on cats and dogs as examples of highly deformable objects for which recognition and detection is particularly challenging [24, 29, 34, 45].

Beyond the technical interest of fine grained categorization, extracting information from images of pets has a practical side too. People devote a lot of attention to their domestic animals, as suggested by the large number of social networks dedicated to the sharing of images of cats and dogs: Pet Finder [11], Catster [4], Dogster [5], My Cat Space [9], My Dog Space [10], The International Cat Association [8] and several others [1, 2, 3, 12]. In fact, the bulk of the data used in this paper has been extracted from annotated images that users of these social sites post daily (Sect. 2). It is not unusual for owners to believe (and post) the incorrect breed for their pet, so having a method of automated classification could provide a gentle way of alerting them to such errors.

The first contribution of this paper is the introduction of a large annotated collection of images of 37 different breeds of cats and dogs (Sect. 2). It includes 12 cat breeds and 25 dog breeds. This data constitutes the benchmark for pet breed classification, and, due to its focus on fine grained categorization, is complementary to the standard object recognition benchmarks. The data, which is publicly available, comes with rich annotations: in addition to a breed label, each pet has a pixel level segmentation and a rectangle localising its head. A simple evaluation protocol, inspired by the PASCAL VOC challenge, is also proposed to enable the comparison of future methods on a common grounds (Sect. 2). This dataset is also complementary to the subset of ImageNet used in [27] for dogs, as it contains additional annotations, though for fewer breeds.

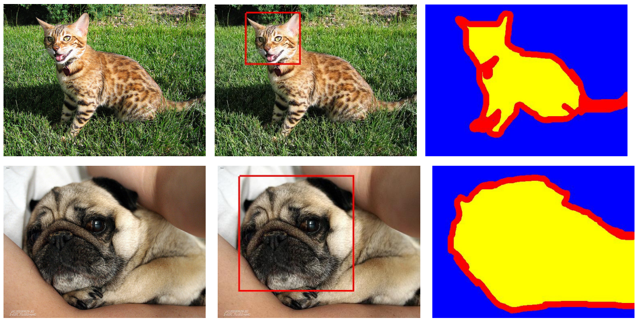

Figure 1. **Annotations in the Oxford-IIIT Pet data.** From left to right: pet image, head bounding box, and trimap segmentation (*blue*: background region; *red*: ambiguous region; *yellow*: foreground region).

The second contribution of the paper is a model for pet breed discrimination (Sect. 3). The model captures both shape (by a deformable part model [23, 42] of the pet face) and texture (by a bag-of-visual-words model [16, 30, 38, 44] of the pet fur). Unfortunately, current deformable part models are not sufficiently advanced to represent satisfactorily the highly deformable bodies of cats and dogs; nevertheless, they can be used to reliably extract stable and distinctive *components* of the body, such as the pet face. The method used in [34] followed from this observation: a cat's face was detected as the first stage in detecting the entire animal. Here we go further in using the detected head shape as a part of the feature descriptor. Two natural ways of combining the shape and appearance features are then considered and compared: a flat approach, in which both features are used to regress the pet's family and the breed simultaneously, and a hierarchical one, in which the family is determined first based on the shape features alone, and then appearance is used to predict the breed conditioned on the family. Inferring the model in an image involves segmenting the animal from the background. To this end, we improved on our previous method on of segmentation in [34] basing it on the extraction of superpixels.

The model is validated experimentally on the task of discriminating the 37 pet breeds (Sect. 4), obtaining very encouraging results, especially considering the toughness of the problem. Furthermore, we also use the model to break the ASIRRA test that uses the ability of discriminating between cats and dogs to tell humans from machines.

2. Datasets and evaluation measures

2.1. The Oxford-IIIT Pet dataset

The *Oxford-IIIT Pet dataset* is a collection of 7,349 images of cats and dogs of 37 different breeds, of which 25 are dogs and 12 are cats. Images are divided into training, validation, and test sets, in a similar manner to the PASCAL VOC data. The dataset contains about 200 images for each breed (which have been split randomly into 50 for training, 50 for validation, and 100 for testing). A detailed list of breeds is given in Tab. 1, and example images are given in Fig. 2. The dataset is available at [35].

Dataset collection. The pet images were downloaded from Catster [4] and Dogster [5], two social web sites dedicated to the collection and discussion of images of pets, from Flickr [6] groups, and from Google images [7]. People uploading images to Catster and Dogster provide the breed information as well, and the Flickr groups are specific to each breed, which simplifies tagging. For each of the 37 breeds, about 2,000 – 2,500 images were downloaded from these data sources to form a pool of candidates for inclusion in the dataset. From this candidate list, images were dropped if any of the following conditions applied, as judged by the annotators: (i) the image was gray scale, (ii) another image portraying the same animal existed (which happens frequently in Flickr), (iii) the illumination was poor, (iv) the pet was not centered in the image, or (v) the pet was wearing clothes. The most common problem in all the data sources, however, was found to be errors in the breed labels. Thus labels were reviewed by the human annotators and fixed whenever possible. When fixing was not possible, for instance because the pet was a cross breed, the image was dropped. Overall, up to 200 images for each of the 37 breeds were obtained.

Annotations. Each image is annotated with a breed label, a pixel level segmentation marking the body, and a tight bounding box about the head. The segmentation is a trimap with regions corresponding to: foreground (the pet body), background, and ambiguous (the pet body boundary and any accessory such as collars). Fig. 1 shows examples of these annotations.

Evaluation protocol. Three tasks are defined: pet family classification (Cat vs Dog, a two class problem), breed classification given the family (a 12 class problem for cats and a 25 class problem for dogs), and breed and family classification (a 37 class problem). In all cases, the performance is measured as the average per-class classification accuracy. This is the proportion of correctly classified images for each of the classes and can be computed as the average of the diagonal of the (row normalized) confusion matrix. This means that, for example, a random classifier has average accuracy of $1/2 = 50\%$ for the family classification task, and of $1/37 \approx 3\%$ for the breed and family classification task. Algorithms are trained on the training and validation subsets and tested on the test subset. The split between training and validation is provided only for convenience, but can be disregarded.