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 example 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.

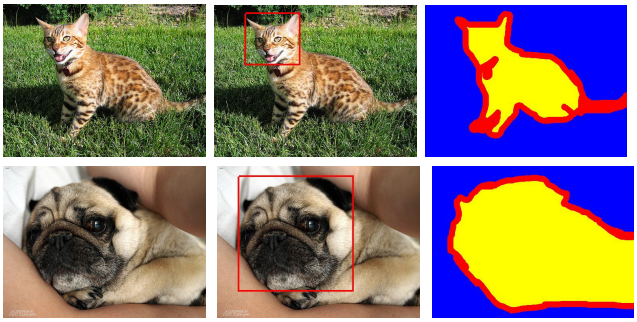


Figure 1. **Annotations in the Oxford-IIIT Pet data.** From left to right: pet image, head bounding box, and trimap segmentation (*blue*: background region; *red*: ambiguous region; *yellow*: foreground region).

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.

Breed	Training	Validation	Test	Total	Breed	Training	Validation	Test	Total
Abyssinian	50	50	98	198	English Setter	50	50	100	200
Bengal	50	50	100	200	German Shorthaired	50	50	100	200
Birman	50	50	100	200	Great Pyrenees	50	50	100	200
Bombay	49	47	88	184	Havanese	50	50	100	200
British Shorthair	50	50	100	200	Japanese Chin	50	50	100	200
Egyptian Mau	47	46	97	190	Keeshond	50	50	99	199
Maine Coon	50	50	100	200	Leonberger	50	50	100	200
Persian	50	50	100	200	Miniature Pinscher	50	50	100	200
Ragdoll	50	50	100	200	Newfoundland	50	46	100	196
Russian Blue	50	50	100	200	Pomeranian	50	50	100	200
Siamese	50	49	100	199	Pug	50	50	100	200
Sphynx	50	50	100	200	Saint Bernard	50	50	100	200
American Bulldog	50	50	100	200	Samoyed	50	50	100	200
American Pit Bull Terrier	50	50	100	200	Scottish Terrier	50	50	99	199
Basset Hound	50	50	100	200	Shiba Inu	50	50	100	200
Beagle	50	50	100	200	Staffordshire Bull Terrier	50	50	89	189
Boxer	50	50	99	199	Wheaten Terrier	50	50	100	200
Chihuahua	50	50	100	200	Yorkshire Terrier	50	50	100	200
English Cocker Spaniel	50	46	100	196	**Total**	1846	1834	3669	7349

Table 1. **Oxford-IIIT Pet data composition**. The 12 cat breeds followed by the 25 dog breeds.

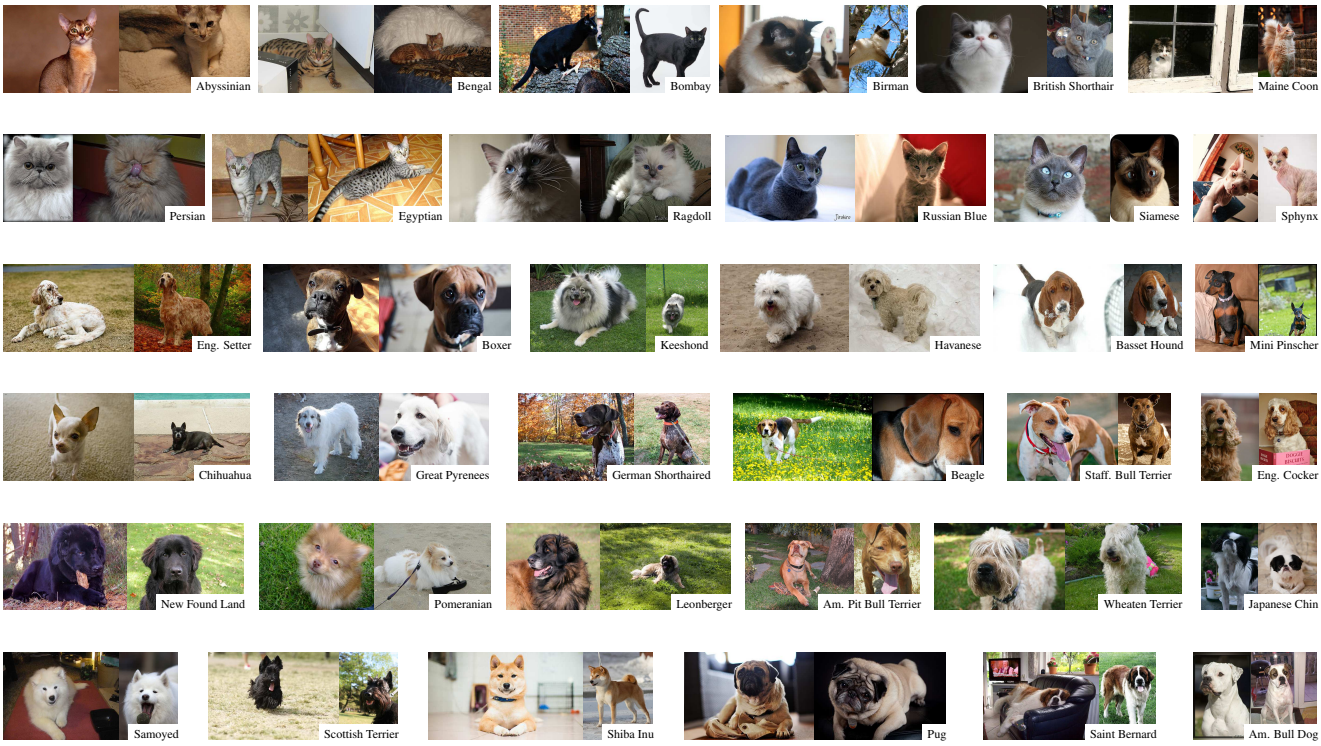


Figure 2. **Example images from the Oxford-IIIT Pet data.** Two images per breed are shown side by side to illustrate the data variability.

2.2. The ASIRRA dataset

Microsoft Research (MSR) proposed the problem of discriminating cats from dogs as a test to tell humans from machines, and created the *ASIRRA* test ([19], Fig. 3) on this basis. The assumption is that, out of a batch of twelve images of pets, any machine would predict incorrectly the family of at least one of them, while humans would make no mis-

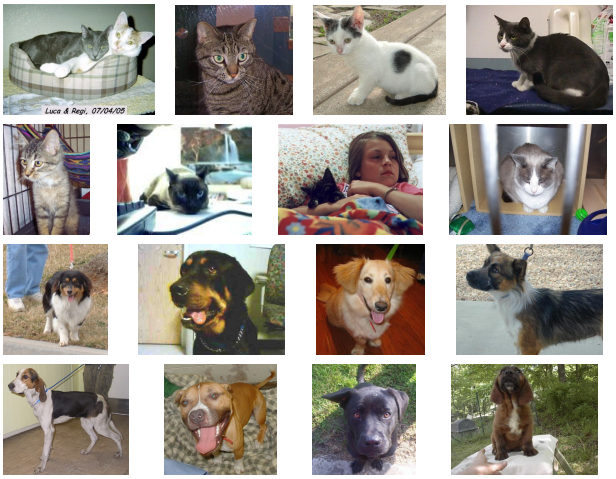


Figure 3. **Example images from the MSR ASIRRA dataset.**

takes. The ASIRRA test is currently used to protect a number of web sites from the unwanted access by Internet bots. However, the reliability of this test depends on the classification accuracy α of the classifier implemented by the bot. For instance, if the classifier has accuracy $\alpha = 95\%$, then the bot fools the ASIRRA test roughly half of the times ($\alpha^{12} \approx 54\%$).

The complete MSR ASIRRA system is based on a database of several millions images of pets, equally divided between cats and dogs. Our classifiers are tested on the $24{,}990$ images that have been made available to the public for research and evaluation purposes.

3. A model for breed discrimination

The breed of a pet affects its size, shape, fur type and color. Since it is not possible to measure the pet size from an image without an absolute reference, our model focuses on capturing the pet shape (Sect. 3.1) and the appearance of its fur (Sect. 3.2). The model also involves automatically segmenting the pet from the image background (Sect. 3.3).

3.1. Shape model

To represent shape, we use the deformable part model of [23]. In this model, an object is given by a root part connected with springs to eight smaller parts at a finer scale. The appearance of each part is represented by a HOG filter [17], capturing the local distribution of the image edges; inference (detection) uses dynamic programming to find the best trade-off between matching well each part to the image and not deforming the springs too much.

While powerful, this model is insufficient to represent the flexibility and variability of a pet body. This can be seen by examining the performance of this detector on the cats and dogs in the recent PASCAL VOC 2011 challenge data [20]. The deformable parts detector [23] obtains an Average Precision (AP) of only 31.7% and 22.1% on cats and dogs respectively [20]; by comparison, an easier category such as bicycle has AP of 54% [20]. However, in the PASCAL VOC challenge the task is to detect the *whole body* of the animal. As in the method of [34], we use the deformable part model to detect certain stable and distinctive *components* of the body. In particular, the head annotations included in the Oxford-IIIT Pet data are used to learn a deformable part model of the cat faces, and one of the dog faces ([24, 29, 45] also focus on modelling the faces of pets). Sect. 4.1 shows that these shape models are in fact very good.

3.2. Appearance model

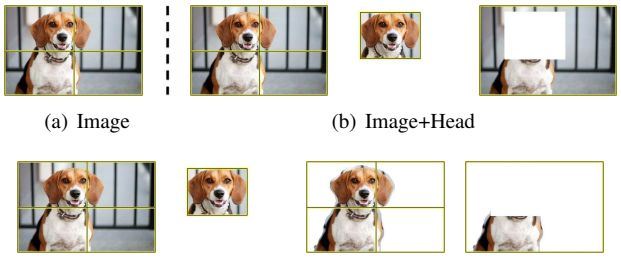
To represent texture, we use a bag-of-words [16] model. Visual words [38] are computed densely on the image by extracting SIFT descriptors [31] with a stride of 6 pixels and at four scales, defined by setting the width of the SIFT spatial bins to 4, 6, 8, and 10 pixels respectively. The SIFT features have constant orientation (*i.e*, they are not adapted to the local image appearance). The SIFT descriptors are then quantized based on a vocabulary of $4{,}000$ visual words. The vocabulary is learned by using k-means from features randomly sampled from the training data. In order to obtain a descriptor for the image, the quantized SIFT features are pooled into a spatial histogram [30], which has dimension equal to $4{,}000$ times the number of spatial bins. Histograms are then l^1 normalized and used in a support vector machine (SVM) based on the exponential-χ^2 kernel [44] for classification.

Different variants of the spatial histograms can be obtained by placing the spatial bins in correspondence of particular geometric features of the pet. These layouts are described next and in Fig. 4:

Image layout. This layout consists of five spatial bins organized as a 1×1 and a 2×2 grids (Fig. 4a) covering the entire image area, as in [30]. This results in a $20{,}000$ dimensional feature vector.

Image+head layout. This layout adds to the *image layout* just described a spatial bin in correspondence of the head bounding box (as detected by the deformable part model of the pet face) as well as one for the complement of this box. These two regions do *not* contain further spatial subdivisions (Fig. 4b). Concatenating the histograms for all the spatial bins in this layout results in a $28{,}000$ dimensional feature vector.

Image+head+body layout. This layout combines the spatial tiles in the *image layout* with an additional spatial bin



(a) Image (b) Image+Head

(c) Image+Head+Body

Figure 4. **Spatial histogram layouts.** The three different spatial layouts used for computing the image descriptors. The image descriptor in each case is formed by concatenating the histograms computed on the individual spatial components of the layout. The spatial bins are denoted by yellow-black lines.

in correspondence of the pet head (as for the *image+head layout*) a swell as other spatial bins computed on the foreground object region and its complement, as described next and in Fig. 4c. The foreground region is obtained either from the automatic segmentation of the pet body or from the ground-truth segmentation to obtain a best-case baseline. The foreground region is subdivided into five spatial bins, similarly to the *image layout*. An additional bin obtained from the foreground region with the head region removed and no further spatial subdivisions is also used. Concatenating the histograms for all the spatial bins in this layout results in a $48{,}000$ dimensional feature vector.

3.3. Automatic segmentation

The foreground (pet) and background regions needed for computing the appearance descriptors are obtained automatically using the grab-cut segmentation technique [36]. Intialization of grab-cut segmentations was done using cues from the overgementation of an image (*i.e*, superpixels) similar to the method of [15]. In this method, a SVM classifier is used to assign superpixels a confidence score. This confidence score is then used to assign superpixels to a foreground or background region to initilaze the grab-cut iteration. We used Berkeley's ultrametric color map (UCM) [13] for obtaining the superpixels. Each superpixel was described by a feature vector comprising the color histogram and Sift-BoW histogram computed on it. Superpixels were assigned a score using a linear-SVM [21] which was trained on the features computed on the training data. After this initialization, grab-cut was used as in [34]. The improved initialization achieves segmentation accuracy of 65% this improving over our previous method [34] by 4% and is about 20% better than simply choosing all pixels as foreground (*i.e*, assuming the pet foreground entirely occupies the image). (Tab. 2). Example segmentations produced by our method on the Oxford-IIIT Pet data are shown in Fig. 5.

Method	Mean Segentation Acccuracy
All foreground	45%
Parkhi *et al.* [34]	61%
This paper	65%

Table 2. **Performance of segmentation schemes.** Segmentation accuracy computed as intersection over union of segmentation with ground truth.

Dataset	Mean Classification Accuracy
Oxford-IIIT Pet Dataset	38.45%
UCSD-Caltech Birds	6.91%
Oxford-Flowers102	53.71%

Table 3. **Fine grained classification baseline.** Mean classification accuracies obtained on three different datasets using the VLFeat-BoW classification code.

4. Experiments

The models are evaluated first on the task of discriminating the family of the pet (Sect. 4.1), then on the one of discriminating their breed given the family (Sect. 4.2), and finally discriminating both the family and the breed (Sect. 4.3). For the third task, both hierarchical classification (*i.e*, determining first the family and then the breed) and flat classification (*i.e*, determining the family and the breed simultaneously) are evaluated. Training uses the Oxford-IIIT Pet train and validation data and testing uses the Oxford-IIIT Pet test data. All these results are summarized in Tab. 4 and further results for pet family discrimination on the ASIRRA data are reported in Sect. 4.1. Failure cases are reported in Fig. 7.

Baseline. In order to compare the difficulty of the Oxford-IIIT Pet dataset to other Fine Grained Visual Categorization datasets, and also to provide a baseline for our breed classification task, we have run the publicly available VLFeat [40] BoW classification code over three datasets: Oxford Flowers 102 [33], UCSD-Caltech Birds [14], and Oxford-IIIT Pet dataset (note that this code is a faster successor to the VGG-MKL package [41] used on the UCSD-Caltech Birds dataset in [14]). The code employs a spatial pyramid [30], but does not use segmentation or salient parts. The results are given in Table 3.

4.1. Pet family discrimination

This section evaluates the different models on the task of discriminating the family of a pet (cat Vs dog classification).

Shape only. The maximum response of the cat face detector (Sect. 3.1) on an image is used as an image-level score for the cat class. The same is done to obtain a score for

.	Shape	Appearance		Classification Accuracy (%)				
		layout type	using ground truth	family	breed (S. 4.2)		both (S. 4.3)	
				(S. 4.1)	cat	dog	hier.	flat
1	✓	–	–	94.21	NA	NA	NA	NA
2	–	Image	–	82.56	52.01	40.59	NA	39.64
3	–	Image+Head	–	85.06	60.37	52.10	NA	51.23
4	–	Image+Head+Body	–	87.78	64.27	54.31	NA	54.05
5	–	Image+Head+Body	✓	88.68	66.12	57.29	NA	56.60
6	✓	Image	–	94.88	50.27	42.94	42.29	43.30
7	✓	Image+Head	–	95.07	59.11	54.56	52.78	54.03
8	✓	Image+Head+Body	–	94.89	63.48	55.68	55.26	56.68
9	✓	Image+Head+Body	✓	95.37	66.07	59.18	57.77	59.21

Table 4. **Comparison between different models.** The table compares different models on the three tasks of discriminating the family, the breed given the family, and the breed and family of the pets in the Oxford-IIIT Pet dataset (Sect. 2). Different combinations of the shape features (deformable part model of the pet faces) and of the various appearance features are tested (Sect. 3.2, Fig. 4).

the dog class. Then a linear SVM is learned to discriminate between cats and dogs based on these two scores. The classification accuracy of this model on the Oxford-IIIT Pet test data is 94.21%.

Appearance only. Spatial histograms of visual words are used in a non-linear SVM to discriminate between cats and dogs, as detailed in Sect. 3.2. The accuracy depends on the type of spatial histograms considered, which in turn depends on the layout of the spatial bins. On the Oxford-IIIT Pet test data, the *image layout* obtains an accuracy of 82.56%; adding head information using *image+head layout* yields an accuracy of 85.06%. Using *image+head+body layout* improves accuracy by a further 2.7% to 87.78%. An improvement of 1% was observed when the ground-truth segmentations were used in place of the segmentations estimated by grab-cut (Sect. 3.2). This progression indicates that the more accurate the localization of the pet body, the better is the classification accuracy.

Shape and appearance. The appearance and shape information are combined by summing the exp-χ^2 kernel for the appearance part (Sect. 3.2) with a linear kernel on the cat scores and a linear kernel on the dog scores. The combination boosts the performance by an additional 7% over that of using appearance alone, yielding approximately 95.37% accuracy (Table 4, rows 5 and 9), with all the variants of the appearance model performing similarly.

The ASIRRA data. The ASIRRA data does not specify a training set, so we used models trained on the Oxford-IIIT Pet data and the ASIRRA data was used only for testing. The accuracy of the shape model on the ASIRRA data is 92.9%, which corresponds to a 42% probability of breaking the test in a single try. For comparison, the best accuracy reported in the literature on the ASIRRA data is 82.7% [25], which corresponds to just a 9.2% chance of breaking the test. Due to lack of sufficient training data to train appearance models for ASIRRA data, we did not evaluate these models on ASIRRA dataset.

Method	Mean Class. Accuracy
Golle *et al.* [25]	82.7%
This paper (Shape only)	92.9%

Table 5. **Performance on ASIRRA Data.** Table shows performance achieved on task of pet family classification posed by the ASIRRA challenge. Best results obtained by Golle [25] were obtained using 10000 images from the data. 8000 for training and 2000 for testing. Our test results are shown on 24990 images in the ASIRRA dataset.

4.2. Breed discrimination

This section evaluates the models on the task of discriminating the different breeds of cats and dogs given their family. This is done by learning a multi-class SVM by using the 1-vs-rest decomposition [37] (this means learning 12 binary classifiers for cats and 25 for dogs). The relative performance of the different models is similar to that observed for pet family classification in Sect. 4.1. The best breed classification accuracies for cats and dogs are 63.48% and 55.68% respectively, which improve to 66.07% and 59.18% when the ground truth segmentations are used.

4.3. Family and breed discrimination

This section investigates classifying both the family and the breed. Two approaches are explored: *hierarchical classification*, in which the family is decided first as in Sect. 4.1, and then the breed is decided as in Sect. 4.2, and *flat classification*, in which a 37-class SVM is learned directly, using the same method discussed in Sect. 4.2. The relative per-

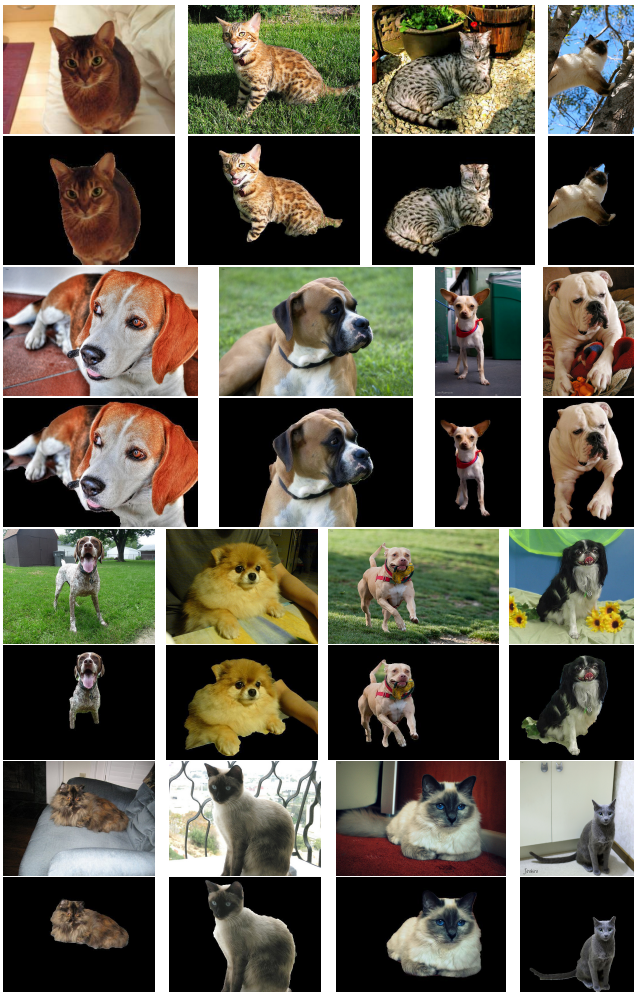


Figure 5. **Example segmentation results on Oxford-IIIT Pet dataset.** The segmentation of the pet from the background was obtained automatically as described in Sect. 3.3.

formance of the different models is similar to that observed in Sect. 4.1 and 4.2. Flat classification is better than hierarchical, but the latter requires less work at test time, due to the fact that fewer SVM classifiers need to be evaluated. For example, using the appearance model with the *image, head, image-head layouts* for 37 class classification yeilds an accuracy of 51.23%, adding the shape information hierarchically improves this accuracy to 52.78%, and using shape and appearance together in a flat classification approach achieves an accuracy 54.03%. The confusion matrix for the best result for breed classification, corresponding to the last entry of the eight row of Table 4 is shown in Fig. 4.

5. Summary

This paper has introduced the PET dataset for the fine-grained categorisation problem of identifying the family

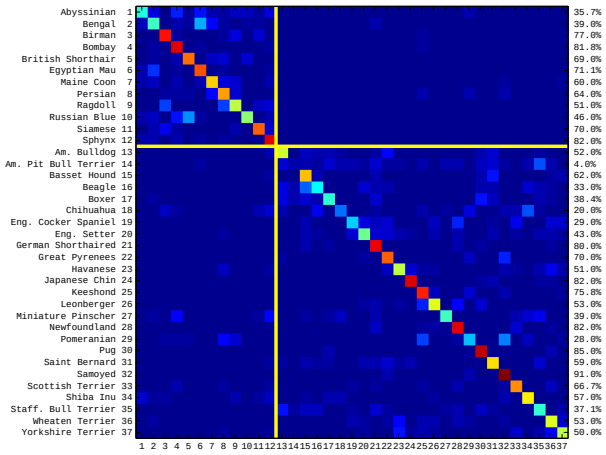


Figure 6. **Confusion matrix for breed discrimination.** The vertical axis reports the ground truth labels, and the horizontal axis to the predicted ones (the upper-left block are the cats). The matrix is normalized by row and the values along the diagonal are reported on the right. The matrix corresponds to the breed classifier using shape features, appearance features with the *image, head, body, body-head layouts* with automatic segmentations, and a 37-class SVM. This is the best result for breed classification, and corresponds to the last entry of row number 8 in Tab. 4.

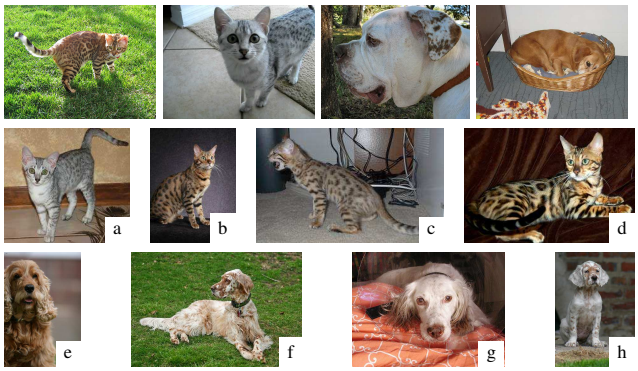


Figure 7. **Failure cases** for the model using appearance only (*image layout*) in Sect. 4.2. *First row:* Cat images that were incorrectly classified as dogs and viceversa. *Second row:* Bengal cats (b–d) classified as Egyptian Mau (a). *Third row:* English Setter (f–h) classified as English Cocker Spaniel (e).

and breed of pets (cats and dogs). Three different tasks and corresponding baseline algorithms have been proposed and investigated obtaining very encouraging classification results on the Oxford-IIIT Pet test data. Furthermore, the baseline models were shown to achieve state-of-the-art performance on the ASIRRA challenge data, breaking the test with 42% probability, a remarkable achievement considering that this dataset was *designed* to be challenging for machines.

Acknowledgements. We are grateful for financial support from EU Project AXES ICT-269980 and ERC grant VisRec no. 228180.

References

[1] American kennel club. http://www.akc.org/.
[2] The cat fanciers association inc. http://www.cfa.org/Client/home.aspx.
[3] Cats in sinks. http://catsinsinks.com/.
[4] Catster. http://www.catster.com/.
[5] Dogster. http://www.dogster.com/.
[6] Flickr! http://www.flickr.com/.
[7] Google images. http://images.google.com/.
[8] The international cat association. http://www.tica.org/.
[9] My cat space. http://www.mycatspace.com/.
[10] My dog space. http://www.mydogspace.com/.
[11] Petfinder. http://www.petfinder.com/index.html.
[12] World canine organisation. http://www.fci.be/.
[13] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. From contours to regions: An empirical evaluation. In *Proc. CVPR*, 2009.
[14] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *Proc. ECCV*, 2010.
[15] Y. Chai, V. Lempitsky, and A. Zisserman. Bicos: A bi-level co-segmentation method for image classification. In *Proc. ICCV*, 2011.
[16] G. Csurka, C. R. Dance, L. Dan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Proc. ECCV Workshop on Stat. Learn. in Comp. Vision*, 2004.
[17] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.
[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. CVPR*, 2009.
[19] J. Elson, J. Douceur, J. Howell, and J. J. Saul. Asirra: A CAPTCHA that exploits interest-aligned manual image categorization. In *Conf. on Computer and Communications Security (CCS)*, 2007.
[20] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html.
[21] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9, 2008.
[22] L. Fei-Fei, R. Fergus, and P. Perona. A Bayesian approach to unsupervised one-shot learning of object categories. In *Proc. ICCV*, 2003.
[23] P. F. Felzenszwalb, R. B. Grishick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2009.
[24] F. Fleuret and D. Geman. Stationary features and cat detection. *Journal of Machine Learning Research*, 9, 2008.
[25] P. Golle. Machine learning attacks against the asirra captcha. In *15th ACM Conference on Computer and Communications Security (CCS)*, 2008.
[26] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical report, California Institute of Technology, 2007.
[27] A. Khosla, N. Jayadevaprakash, B. Yao, and F. F. Li. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, CVPR*, 2011.
[28] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proc. CVPR*, 2009.
[29] I. Laptev. Improvements of object detection using boosted histograms. In *Proc. BMVC*, 2006.
[30] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bag of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. CVPR*, 2006.
[31] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, 1999.
[32] M.-E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *Proc. CVPR*, 2006.
[33] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proc. ICVGIP*, 2008.
[34] O. Parkhi, A. Vedaldi, and A. Zisserman. The truth about cats and dogs. In *Proc. ICCV*, 2011.
[35] O. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. The Oxford-IIIT PET Dataset. http://www.robots.ox.ac.uk/~vgg/data/pets/index.html, 2012.
[36] C. Rother, V. Kolmogorov, and A. Blake. "grabcut" — interactive foreground extraction using iterated graph cuts. In *ACM Trans. on Graphics*, 2004.
[37] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.
[38] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, 2003.
[39] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *Proc. ICCV*, 2007.
[40] A. Vedaldi and B. Fulkerson. VLFeat library. http://www.vlfeat.org/, 2008.
[41] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *Proc. ICCV*, 2009.
[42] A. Vedaldi and A. Zisserman. Structured output regression for detection with partial occulsion. In *Proc. NIPS*, 2009.
[43] P. Welinder, S. Branson, T. Mita, C. Wah, and F. Schroff. Caltech-ucsd birds 200. Technical report, Caltech-UCSD, 2010.
[44] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 2007.
[45] W. Zhang, J. Sun, and X. Tang. Cat head detection - how to effectively exploit shape and texture features. In *Proc. ECCV*, 2008.