Dog Breed Classification Using Part Localization

Jiongxin Liu[1], Angjoo Kanazawa[2], David Jacobs[2], and Peter Belhumeur[1]

[1] Columbia University
[2] University of Maryland

Abstract. We propose a novel approach to fine-grained image classification in which instances from different classes share common parts but have wide variation in shape and appearance. We use dog breed identification as a test case to show that extracting corresponding parts improves classification performance. This domain is especially challenging since the appearance of corresponding parts can vary dramatically, *e.g.*, the faces of bulldogs and beagles are very different. To find accurate correspondences, we build exemplar-based geometric and appearance models of dog breeds and their face parts. Part correspondence allows us to extract and compare descriptors in like image locations. Our approach also features a hierarchy of parts (*e.g.*, face and eyes) and breed-specific part localization. We achieve 67% recognition rate on a large real-world dataset including 133 dog breeds and 8,351 images, and experimental results show that accurate part localization significantly increases classification performance compared to state-of-the-art approaches.

1 Introduction

Image classification methods follow a common pipeline in which a set of features are extracted from an image and fed to a classifier. These features are often extracted at generic locations or keypoints within the image, sampling both object and background, with the hope that these locations will reveal something about the class. However, for fine-grained classification, background regions may contain more noise than useful contextual information about identity. Moreover, while generic sampling may be useful in capturing larger scale information that differentiates very different classes, it can miss the details that are needed to distinguish between classes that are similar in appearance.

We argue and demonstrate in this paper that fine-grained classification can be improved if the features used for classification are localized at object parts. While such localization across wide categories of objects may not yet be possible, we will show that within a particular category of objects – domestic dogs – such localization is both possible and helpful in significantly improving the recognition accuracy over state-of-the-art methods.

The domestic dog (*Canis lupus familiaris*) displays "greater levels of morphological and behavioral diversity than have been recorded for any land mammal" [1]. The dog's diversity in its visual appearance would seem to present significant challenges to part localization. However, we show that using appearance-based

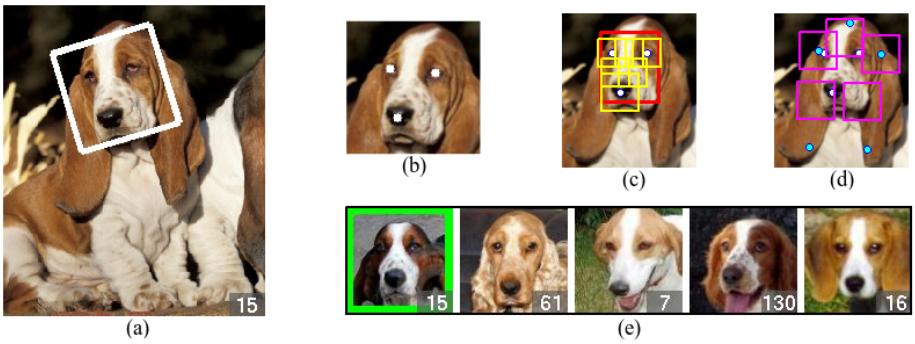


Fig. 1. (a) Given an image, our method automatically detects the dog's face, (b) localizes eyes and nose of the face, (c) aligns the face and extracts greyscale SIFT features (yellow windows) and a color histogram (red window), (d) infers remaining part locations from exemplars (cyan dots) to extract additional SIFT features (magenta windows), and (e) predicts the breed (green box) along with the next best guesses from left to right. The numbers correspond to breed names listed in Table 1.

sliding window detectors and a probabilistic consensus of geometric models, we can accurately detect dog faces and localize their face parts. In addition, many subsets of dog breeds are quite similar in appearance (*e.g.*, beagle and basset hound) making identification very challenging. We also show that the use of image descriptors aligned with common parts allows us to overcome this challenge in real-world images.

Determination of dog breeds provides an excellent domain for fine grained visual categorization experiments. After humans, dogs are possibly the most photographed species. With the advent of image search engines, millions of dog images can be easily obtained. These images are nearly infinite in their variety, showing dogs of all shapes, sizes, and colors, under differing illumination, in innumerable poses, and in just about any location. Success in this domain will certainly lead to further success in the broader domain of automatic species identification [2–5], contributing to the ever growing area of biodiversity studies. In addition, there are some practical applications to the automatic processing of dog images. Dog face detectors can be used for autofocus. For personal photos, detection and identification of dog breeds can facilitate autotagging and image search.

Dog breed identification is representative of a set of fine-grained classification problems in which corresponding object parts exist across classes, but the geometry and appearance of these parts vary significantly. To cope with this problem, first we use a sliding window detector to locate dog faces. Then we accurately localize the eyes and nose by combining appearance-based detection with a large exemplar-based set of geometric models, building on the consensus of models approach of [6]. Using this small set of face parts (eyes and nose), we are able to align an image with models for each dog breed, and hypothesize breed-specific

locations of additional parts, such as ears, whose position and appearance vary greatly between breeds. We then extract image features at these locations for use in classification.

To train and evaluate this system, we have created a labeled dataset of 8,351 real-world images of 133 American Kennel Club (AKC) recognized dog breeds. The images are downloaded from Google, Image-net and Flickr. In each image, parts on the dog faces are localized using Amazon's Mechanical Turk (MTurk). These parts include both eyes, the nose, the tips of both ears, and three points marking the top of the dog's head between the ears. Due to different pose, lighting, expression, and intra-breed variation in the dataset, instances from the same breed can look quite different. Yet, in experiments with this dataset, we show that we can automatically determine a dog breed with 67% accuracy on the first guess – many of our errors are due to the close visual similarity of some breeds. Overall, we find that we can place the correct breed among the top 10 guesses with more than 93% accuracy. We also show that by using descriptors aligned to parts, we significantly outperform other state-of-the-art fine-grained classification methods on this dataset.

Our paper makes the following novel contributions:

- We show how a probabilistic consensus of exemplars approach, which has previously been applied only to part detection, can be extended to perform object classification.
- We show how class-specific object parts can be inherited from the exemplars and used to improve classification accuracy.
- We create a new and extensive 8,351 image dataset including not only class labels for 133 dog breeds but also 66,808 part labels (Eight per image).
- We design a complete working vision system released as a free iPhone app.

2 Related Work

There has been a great deal of recent interest in methods for visual classification [7–13]. Our work specifically addresses fine-grained categorization, in which discrimination between a large number of similar classes is difficult. In addressing this problem, [14] mines discriminative features with randomized sampling, and [3] uses a multiple kernel framework to combine kernels in a way that is most discriminative for each class. [2] also uses the framework as part of an interactive system for bird species identification. [5] identifies plant species using images of leaves, and along with [3], relies on segmentation to localize the object of interest before extracting descriptors.

Especially relevant to our approach is [4], which uses the poselet framework [15] to localize the head and body of birds, enabling part-based feature extraction. Our work is in some ways complementary to this, in that [4] focuses on developing methods of using large, articulated parts while our approach finds parts describable at point locations. We also make use of a hierarchical approach in which first the face, and then parts of the face are found, and make use of class-specific part localization to look for parts such as the ears, which are

extremely variable in position and appearance. One reason for this difference in emphasis is that we have found that for species such as dogs, much of the information about identity is contained in the appearance of the face. (While overall size is also important, it is difficult to extract from unconstrained images.) Also highly relevant is the contemporaneous work of [16], which determines the breed of cats and dogs. They introduce a dataset containing 37 breeds (25 breeds of dogs) and combine a deformable part model and texture-based representations of fur to identify breed.

There is also a vast literature on face detection, though to our knowledge this work has mostly focused on human faces. Haar and Haar-like wavelets have been widely used in a cascaded Adaboost classifier [17]. As we will show, these detectors are not as effective for dog faces, presumably due to their much greater variation in geometry and appearance. We instead use a more powerful, sliding window support vector machine (SVM) [6, 18]. [19] has also recently addressed the problem of detection of cats and dogs. However, they focus on the head and its relation to the full body, while we focus here on only the face, as slightly more than half of dog images do not contain the full body. A good deal of work has also addressed the localization of fiducial points and parts of human faces [20–22, 6]. [6] proposes to combine the output of local detectors with a non-parametric set of models to localize an extensive list of face parts.

Once parts are localized, we build image descriptors using SIFT [23], centering descriptors at matching locations. Many methods have used bag-of-words approaches with SIFT descriptors, in which localization is not important or at least not the focus (*e.g.* [7]), while other approaches have related image descriptors to capture spatially localized information that can then be grouped together [18, 9] to capture geometric relations.

Parts and attributes have been widely used in face recognition [24, 25]. These methods also try to increase the discriminative ability of features by extracting information at local regions. Fiducial points are directly used in some face recognition work. [26] builds an automatic face recognition system where Gabor filters are applied to extract descriptors around the detected fiducials. [27] studies the strong correlation between eye localization error and the face recognition rate.

3 Dog Breed Dataset

We have created a dataset[1] of natural images of dogs, downloaded from sources such a Flickr, Image-Net, and Google. The dataset contains 133 breeds of dogs with 8,351 images. The images were not filtered, except to exclude images in which the dog's face did not have both eyes visible. Sample faces from all the breeds are shown in Fig. 2 and the list of breed names is shown in Table 1. Not only is there great variation across breeds – making detection a challenge, but there is also great variation within breeds – making identification a challenge. See the blowup of sample images of Breed 97: the Lakeland terrier. Note the variations in color, ear position, fur length, pose, lighting and even expression.

[1] The dataset is available at http://faceserv.cs.columbia.edu/DogData/

Fig. 2. Representative faces from all the breeds. The breeds are numbered according to the alphabetical order of names.

Fig. 3. Sample dog images from our dataset, with parts labeled by MTurk workers.

Each of the images was submitted to MTurk to have the breed verified by multiple workers. Afterward, each image was submitted again to MTurk to have parts of the dog's face labeled. Eight points were labeled in each image by three separate workers. If there was gross disagreement amongst the workers in the locations of these points, we resubmitted the image again for relabeling. The points that were labeled were the eyes, the nose, the tips of both ears, the top of the head, and the inner bases of the ears. In Fig. 3, we show the average location, over three workers, for these eight points.

4 Dog Face Detection

We have created a dog face detector capable of detecting faces of all the dog breeds in our dataset. Although we do not view our dog face detector as a

1:Affenpinscher (80)	28:Bluetick coonhound (44)	55:Curly-coated retriever (63)	82:Havanese (76)	109:Norwegian elkhound (56)
2:Afghan hound (73)	29:Border collie (93)	56:Dachshund (82)	83:Ibizan hound (58)	110:Norwegian lundehund (41)
3:Airedale terrier (65)	30:Border terrier (65)	57:Dalmatian (89)	84:Icelandic sheepdog (62)	111:Norwich terrier (55)
4:Akita (79)	31:Borzoi (70)	58:Dandie dinmont terrier (63)	85:Irish red and white setter (46)	112:Nova scotia duck tolling retriever (67)
5:Alaskan malamute (96)	32:Boston terrier (81)	59:Doberman pinscher (59)	86:Irish setter (66)	113:Old english sheepdog (49)
6:American eskimo dog (80)	33:Bouvier des flandres (56)	60:Dogue de bordeaux (75)	87:Irish terrier (82)	114:Otterhound (44)
7:American foxhound (63)	34:Boxer (80)	61:English cocker spaniel (76)	88:Irish water spaniel (64)	115:Papillon (79)
8:American staffordshire terrier (82)	35:Boykin spaniel (66)	62:English setter (66)	89:Irish wolfhound (66)	116:Parson russell terrier (38)
9:American water spaniel (42)	36:Briard (81)	63:English springer spaniel (66)	90:Italian greyhound (73)	117:Pekingese (60)
10:Anatolian shepherd dog (62)	37:Brittany (62)	64:English toy spaniel (49)	91:Japanese chin (71)	118:Pembroke welsh corgi (66)
11:Australian cattle dog (83)	38:Brussels griffon (71)	65:Entlebucher mountain dog (53)	92:Keeshond (55)	119:Petit basset griffon vendeen (39)
12:Australian shepherd (83)	39:Bull terrier (87)	66:Field spaniel (41)	93:Kerry blue terrier (44)	120:Pharaoh hound (49)
13:Australian terrier (58)	40:Bulldog (66)	67:Finnish spitz (42)	94:Komondor (55)	121:Plott (35)
14:Basenji (86)	41:Bullmastiff (86)	68:Flat-coated retriever (79)	95:Kuvasz (61)	122:Pointer (40)
15:Basset hound (92)	42:Cairn terrier (79)	69:French bulldog (64)	96:Labrador retriever (54)	123:Pomeranian (55)
16:Beagle (74)	43:Canaan dog (62)	70:German pinscher (59)	97:Lakeland terrier (62)	124:Poodle (62)
17:Bearded collie (77)	44:Cane corso (80)	71:German shepherd dog (78)	98:Leonberger (57)	125:Portuguese water dog (42)
18:Beauceron (63)	45:Cardigan welsh corgi (66)	72:German shorthaired pointer (60)	99:Lhasa apso (53)	126:Saint bernard (37)
19:Bedlington terrier (60)	46:Cavalier king charles spaniel (84)	73:German wirehaired pointer (52)	100:Lowchen (42)	127:Silky terrier (51)
20:Belgian malinois (78)	47:Chesapeake bay retriever (67)	74:Giant schnauzer (51)	101:Maltese (60)	128:Smooth fox terrier (38)
21:Belgian sheepdog (80)	48:Chihuahua (68)	75:Glen of imaal terrier (55)	102:Manchester terrier (36)	129:Tibetan mastiff (60)
22:Belgian tervuren (59)	49:Chinese crested (63)	76:Golden retriever (80)	103:Mastiff (72)	130:Welsh springer spaniel (55)
23:Bernese mountain dog (81)	50:Chinese shar-pei (62)	77:Gordon setter (54)	104:Miniature schnauzer (53)	131:Wirehaired pointing griffon (37)
24:Bichon frise (77)	51:Chow chow (78)	78:Great dane (50)	105:Neapolitan mastiff (39)	132:Xoloitzcuintli (33)
25:Black and tan coonhound (46)	52:Clumber spaniel (61)	79:Great pyrenees (74)	106:Newfoundland (62)	133:Yorkshire terrier (38)
26:Black russian terrier (51)	53:Cocker spaniel (59)	80:Greater swiss mountain dog (57)	107:Norfolk terrier (58)	
27:Bloodhound (80)	54:Collie (71)	81:Greyhound (70)	108:Norwegian buhund (33)	

Table 1. List of breed names. Each breed name is numbered, with the number of images shown to the right.

technical contribution of this paper, it is a necessary component of our complete vision system and is described briefly here.

The detector is a SVM regressor with greyscale SIFT [23] descriptors as features. Eight SIFT descriptors are extracted at fixed positions relative to the center point. The positions and scales are chosen to roughly align with the geometry of a dog's face (eyes and nose). Once extracted, these descriptors are then concatenated into a single 1024-dimensional feature vector for our SVM regressor. We use 4,700 positive examples for training.

For each negative training sample, we randomly rescale the image and randomly choose a location that is at least the inter-ocular distance away from the above described center point on the dog's face. We then extract the same eight SIFT descriptors at this non-face location. As negative examples are plentiful, we use 13,000 negative examples. With both positive and negative samples in hand, we train our SVM regressor using an RBF kernel.

As the SVM regressor is trained at a fixed rotation and scale, at detection time we must search not only over location, but also over rotation and scale. We threshold and merge the repeated detections with non-maximum suppression for each rotation separately, and choose the detection window with the highest score if multiple windows collide at a certain location.

4.1 Dog Face Detection: Experiments

Following [17], we also implemented a cascaded AdaBoost detector with Haar-like features to compare with our SVM-based detector. While this detector has seen much success in detecting human faces, it is sometimes plagued by unwanted false positives. Perhaps due to the extreme variability in geometry and appearance of dog faces, this weakness in the cascaded Adaboost detectors is exacerbated in the dog face domain. Even training on considerably more data and using 20 cascades, we could not create a detector with a desirable ratio of

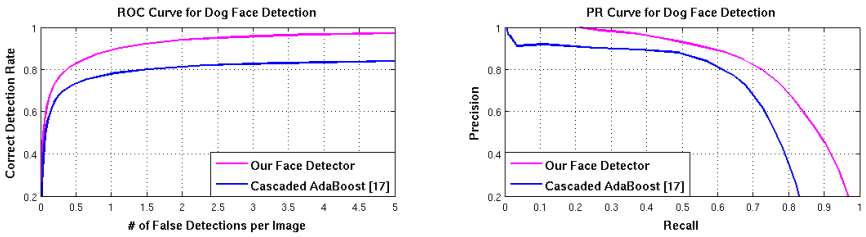


Fig. 4. The ROC and Precision/Recall (PR) curves for dog face detection.

true positives to false positives. We compare the performance of our detector with an Adaboost detector on 3,500 held out images in Figure 4.

5 Localization of Face Parts

To localize parts of the dog face, we build on the consensus of models approach of [6]. The method accurately localizes the eyes and nose; we handle more challenging dog parts during the breed identification process (Section 6). We combine low-level detectors with labeled images that model part locations. We first train a sliding window SVM detector for each dog part. If we let I denote a query image, let $\mathbf{p}^I$ denote the locations of the parts in the image, and let C denote the detector responses for the parts in I, then our goal is to compute

$$\hat{\mathbf{p}}^I = \arg\max_{\mathbf{p}^I} P(\mathbf{p}^I|C). \tag{1}$$

In [6], probable locations for these parts are dictated by exemplars that have been manually labeled. These exemplars help create conditional independence between different parts. Let $\mathbf{p}_k$ be the locations of the parts in the k^{th} exemplar image, Eq. 1 is then rewritten as

$$\hat{\mathbf{p}}^I = \arg\max_{\mathbf{p}^I} \sum_{k=1}^{m} \int_{t\in T} \prod_{i=1}^{n} P(\Delta\mathbf{p}_{k,t}^{(i)}) P(\mathbf{p}^{(i)I}|C^{(i)}) dt. \tag{2}$$

Here the summation is over all m exemplars, *i.e.*, in our case over all labeled examples of parts of dogs' faces. The integral is over similarity transformations t of the exemplars. $\mathbf{p}_{k,t}^{(i)}$ denotes the part i's location in the k^{th} model, transformed by t, and $\Delta\mathbf{p}_{k,t}^{(i)}$ denotes the difference in location of the part i in the query image from that of the transformed exemplar. This amounts to introducing a generative model of part locations in which a randomly chosen example is transformed and placed in the image with noise added. After assuming independence of parts in the deviation from the model, we then marginalize the model out.

This optimization is then solved by a RANSAC-like procedure, in which a large number of exemplars are randomly selected and fit to the modes of the

Fig. 5. An example of part detection. Left: (a) Original image overlaid with heat maps from part detectors. Red is used for the left eye, green for the right eye, and yellow for the nose; better scores are shown as brighter. (b) Detected parts using the maximum value of the detection scores. (c) Detected parts using the method described in this paper. Right: Mean localization error divided by the inter-ocular distance.

detector output. For each hypothesis in which a model is transformed into the image, the model part locations are combined with the detector output; the best fitting matches then pool information, creating a consensus about the part locations. We form the consensus for part i as

$$\hat{\mathbf{p}}^{(i)I} = \arg\max_{\mathbf{p}^{(i)I}} \sum_{k,t \in \mathcal{M}} P(\Delta\mathbf{p}_{k,t}^{(i)}) P(\mathbf{p}^{(i)I} | C^{(i)}) \tag{3}$$

where $\Delta\mathbf{p}_{k,t}^{(i)}$ is modeled as a Gaussian distribution, and we sum over the best fits $\mathcal{M}$ between the exemplars and detectors that are produced by RANSAC.

5.1 Localization of Face Parts: Experiments

We have evaluated the accuracy of part detection on our dataset. We compare our full model to a simpler system that locates a part at the mode of the SVM-based sliding window detector for each part. We also compare our results to the agreement of human labelers by determining the distance between the location indicated by one human labeler and the average of the other two. We show qualitative results and make a quantitative comparison in Fig. 5. Note that our complete model improves over the results of just using a low-level detector, and that our localization error is better than the agreement among human labelers.

6 Breed Identification

Our classification algorithm focuses entirely on the face of the dog. This is partly because the face is largely a rigid object, simplifying the problem of comparing images of different dogs. However, we are also guided by the intuition that dog breeds are largely identifiable from their face. A dog's body shape is not only difficult to identify and often not present in images, but also offers little additional information except in a more extreme cases (*e.g.*, dachshunds).

If we denote the breed of a dog by B, our goal is to compute

$$\hat{B} = \arg\max_B P(B|I). \tag{4}$$

Let the part locations in the query image I be given by $\mathbf{p}^I$. Then

$$\hat{B} = \arg\max_B \int P(B|I, \mathbf{p}^I) P(\mathbf{p}^I|I) d\mathbf{p}^I. \tag{5}$$

Here we integrate over all possible locations of the parts $\mathbf{p}^I$ in the image I.

If these locations can be accurately localized, then $P(\mathbf{p}^I|I)$ is approximately a delta function about the true locations of the parts. Then if we write

$$\hat{\mathbf{p}}^I = \arg\max_{\mathbf{p}^I} P(\mathbf{p}^I|I), \tag{6}$$

we have

$$\hat{B} = \arg\max_B P(B|I, \hat{\mathbf{p}}^I) P(\hat{\mathbf{p}}^I|I). \tag{7}$$

Note that $P(\hat{\mathbf{p}}^I|I)$ is independent of B, so that

$$\hat{B} = \arg\max_B P(B|I, \hat{\mathbf{p}}^I). \tag{8}$$

This means that we can break our problem into two parts. First, we must compute $\arg\max_{\mathbf{p}^I} P(\mathbf{p}^I|I)$ as explained in the previous section. Next we must compute $\arg\max_B P(B|I, \hat{\mathbf{p}}^I)$. Note that

$$P(B|I, \hat{\mathbf{p}}^I) = \frac{P(I|B, \hat{\mathbf{p}}^I) P(B|\hat{\mathbf{p}}^I)}{P(I|\hat{\mathbf{p}}^I)} \tag{9}$$

where the denominator $P(I|\hat{\mathbf{p}}^I)$ is a constant that does not affect which breed will maximize the probability. So

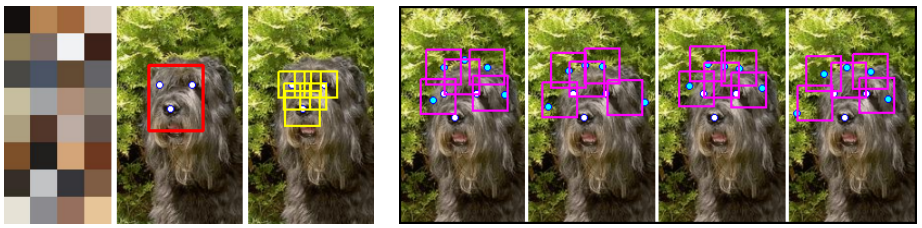
$$\hat{B} = \arg\max_B P(I|B, \hat{\mathbf{p}}^I) P(B|\hat{\mathbf{p}}^I). \tag{10}$$

However, our knowledge of what constitutes a breed is completely given by our set of labeled exemplar images. We divide the information in these images into two parts. First, we let $\mathbf{p}^B$ denote the known locations of the parts of all exemplars for breed B. Then we let D^B denote descriptors characterizing the appearance of the exemplars for breed B. These descriptors are extracted at corresponding part locations given by $\mathbf{p}^B$. So we can rewrite Eq. 10 as

$$\hat{B} = \arg\max_B P(I|D^B, \mathbf{p}^B, \hat{\mathbf{p}}^I) P(D^B, \mathbf{p}^B|\hat{\mathbf{p}}^I) \tag{11}$$

In approximating this, we assume that the breed appearance descriptors D^B are independent of their positions, and we have a uniform distribution over breeds. This allows us to rewrite Eq. 11 as

$$\hat{B} = \arg\max_B P(I|D^B, \mathbf{p}^B, \hat{\mathbf{p}}^I) P(\mathbf{p}^B|\hat{\mathbf{p}}^I) \tag{12}$$



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

Fig. 6. (a) The cluster centers used to create the color histogram. (b) The window used to extract the color histogram based on detected locations of the eyes and nose (white dots). (c) The SIFT descriptor windows (yellow) dictated by the eyes and nose. (d) Four different sets of inferred locations (cyan dots) and the SIFT descriptor windows (magenta) dictated by these.

This suggests that we compute $P(I|D^B, \mathbf{p}^B, \hat{\mathbf{p}}^I)$ by measuring how well the appearance of the query image at and around the part locations given by $\mathbf{p}^I$ agrees with the appearance of our exemplars in their corresponding locations $\mathbf{p}^B$.

To do this, we train one vs. all SVMs for each breed B, and we use two types of features: greyscale SIFT descriptors and a color histogram. We want to center the collection of SIFT features at places dictated by the part locations. However, at this point we are only able to locate the eyes and nose with high accuracy. Since the other parts are more breed-specific, we infer the locations of the remaining parts from exemplars of breed B when generating the negative training samples. During testing, for each breed we choose r exemplars whose eyes and nose locations are closest to the query's after alignment with a similarity transform. These exemplars are the ones that are most likely in the same pose as the query image. Consequently, when we use these similarity transformations to infer the location of additional face parts, these are likely to align with those of the query image of the same breed. For example, Fig. 6 (d) shows the detected locations (white dots) for the eyes and nose, and four different sets of inferred locations for the remaining parts (cyan dots).

To extract features, we center 3 SIFT descriptors at the eyes and nose. We center another 3 descriptors at the 3 midpoints along the lines connecting the eyes and nose. The windows for the 6 SIFT descriptors are shown in yellow in Fig. 6 (c). We place an additional 5 descriptors at the bases of the ears and the midpoints of the lines connecting the eyes with the other inferred parts. The windows for the additional 5 SIFT descriptors are shown in magenta in Fig. 6 (d). The color histogram is computed over a rectangular region centered on the dog's face, shown in red in Fig. 6 (b). The histogram is created using 32 color centers computed from a k-means clustering across all exemplar images of all dogs, shown in Fig. 6 (a). The 32 color features along with the 11 SIFT features are concatenated to produce a 1440-dimensional feature vector.

Fig. 7. Classification examples. Testing samples are in the first column, the closest 10 breeds based on our method are shown to the right.

Given a query image, the selected exemplars produces r feature vectors for each breed. We evaluate each of these using our one vs. all SVM and allow the best scoring feature vector to represent the breed. In practice, we choose $r = 10$.

The second probability in Eq. 12 can be computed directly from the distribution of part locations in our exemplars. Since we are aligning the eyes with a similarity transform, only the relative location of the nose could carry information about the breed. But we have not found it helpful to include this.

Finally, we also use our part detection to assist in improving face detection. We consider the five face windows that the detector scores highest. In each of these windows, we compute the part locations. Then, we multiply the face detector score by the geometric mean of the detection scores for the parts, selecting the window with the highest score.

6.1 Breed Identification: Experimental Results

There are 133 breeds in our dataset of 8,351 images. We randomly split the images of each breed in a fixed ratio to get 4,776 training images and 3,575 test images. We double the size of the training data by reflection.

Fig. 7 gives qualitative results for some query images. For each query in the first column, we overlay the image with the detected part locations for the eyes and nose. To better show the performance, we rank the breeds based on their

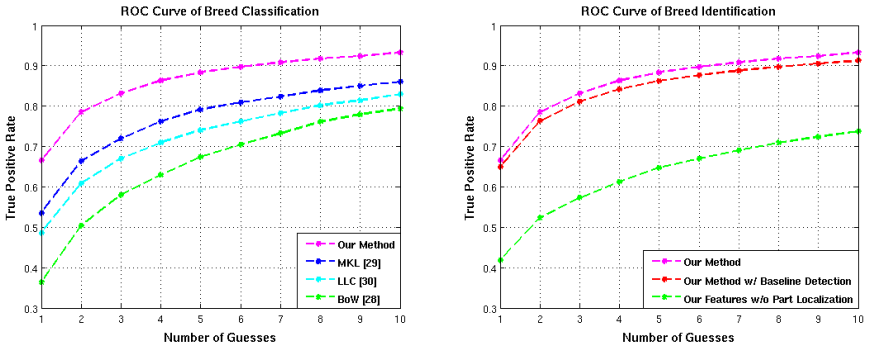
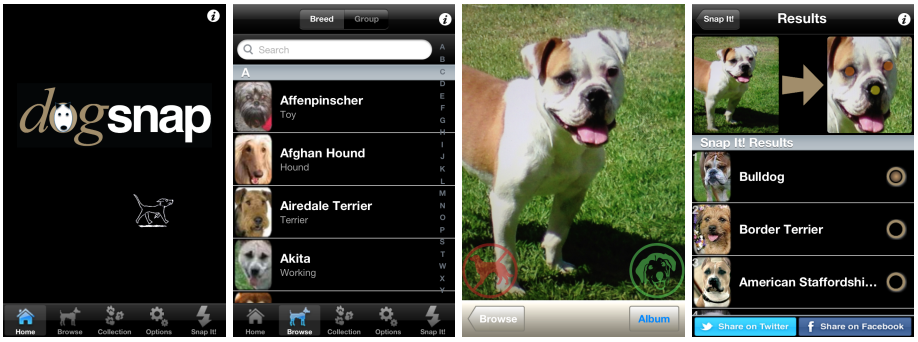


Fig. 8. Performance curves for breed identification, showing how often the correct breed appears among the top 1–10 guesses. On the left we show our method compared with three other methods. On the right we show three variations of our method. The first uses our feature set sampled on a grid within our detected face window. The second uses part localization, but applied to only the highest scoring window from the face detector. The third – and best – uses both the part scores and face detection scores to select the best face window.

probability score. Our system works with high accuracy, failing mostly when the face detection fails (these failures are excluded in these examples) or when the parts detection fails on samples in which fur completely occludes the eyes.

We compare with three other methods: a Bag of Words (BoW) model with spatial tiling [28], a multiple kernel learning (MKL) approach [29] used in bird recognition [2], and locally constrained linear coding (LLC) [30] also applied to a bird dataset [14]. We apply each of these methods inside a cropped window found by selecting the location and size from the highest face detector score within the image; if we use the uncropped images, the performance of all methods is poor – below 20% on the first guess. This gives each method the benefit of face detection and allows us to evaluate the additional gain produced by our system using part detection. In Fig. 8-left we show performance curves for all methods. Our method significantly outperforms existing approaches getting the breed identification corrrect 67% of the time on the first guess vs. 54% for MKL, 49% for LLC, and 36% for BoW. In Fig. 8-right we show three variants of our approach. As a baseline, we use our feature set, extracted on a grid, rather than at part locations. We can see that the use of parts results in a substantial improvement in performance. We can also see that the use of parts to improve face detection itself results in a further improvement in performance, eliminating approximately 20% of the errors for the top ten guess.

To facilitate experimentation by ourselves and others with this algorithm, we have created and released a free iPhone app for dog breed identification (see Fig. 9). The app allows a user to photograph a dog and upload its picture to a server for face detection, part detection, and breed identification.



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

Fig. 9. Screenshots of our iPhone app. (a) Home screen. (b) Browse screen with the dog breeds. (c) Dog camera. (e) Detected dog face and parts, with results.

7 Conclusion

One might expect that fine-grained classification problems would be extremely difficult, that telling a beagle from a basset hound would be much harder than telling a car from a computer mouse. Our main contribution is to show that much of this difficulty can be mitigated by the fact that it is possible to establish accurate correspondences between instances from a large family of related classes. We combine features that can be effectively located using generic feature models with breed specific models of part locations. An additional contribution is the creation of a large, publicly available dataset for dog breed identification, coupled with a practical system that achieves high accuracy in real-world images.

Acknowledgements

This research is supported by the National Science Foundation under Grant No. 1116631.

References

1. Spady, T.C., Ostrander, E.A.: Canine behavioral genetics: Pointing out the phenotypes and herding up the genes. AJHG **82** (2008) 10–18
2. Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P., Belongie, S.: Visual recognition with humans in the loop. Proc. ECCV (2010)
3. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. Proc. 6th Indian Conf. on Computer Vision, Graphics and Image Processing (2008) 722–729
4. Farrell, R., Oza, O., Zhang, N., Morariu, V., Darrell, T., Davis, L.: Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. Proc. ICCV (2011)

5. Belhumeur, P., Chen, D., Feiner, S., Jacobs, D., Kress, W., Ling, H., Lopez, I., Ramamoorthi, R., Sheorey, S., White, S., Zhang, L.: Searching the worlds herbaria: A system for visual identification of plant species. Proc. ECCV (2008) 116–129
6. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. Proc. CVPR (2011)
7. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. Work. on Stat. Learning in Comp. Vis., ECCV (2004) 1–22
8. Jurie, F., Triggs, B.: Creating efficient codebooks for visual recognition. Proc. ICCV (2005)
9. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. Proc. CVPR (2006) 2169–2178
10. Gehler, P., Nowozin, S.: On feature combination for multiclass object classication. Proc. CVPR (2009)
11. Zhou, X., Yu, K., Zhang, T., Huang, T.S.: Image classification using super-vector coding of local image descriptors. Proc. ECCV (2010)
12. Deselaers, T., Ferrari, V.: Visual and semantic similarity in imagenet. Proc. CVPR (2011)
13. Sadeghi, M.A., Farhadi, A.: Recognition using visual phrases. Proc. CVPR (2011)
14. Yao, B., Khosla, A., Fei-Fei, L.: Combining randomization and discrimination for fine-grained image categorization. Proc. CVPR (2011)
15. Bourdev, L., Maji, S., Brox, T., Malik, J.: Detecting people using mutually consistent poselet activations. Proc. ECCV (2010)
16. Parkhi, O., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. Proc. CVPR (2012)
17. Viola, P., Jones, M.: Robust real-time object detection. IJCV **57** (2001) 137–154
18. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. Proc. CVPR **1** (2005) 886–893
19. Parkhi, O., Vedaldi, A., Jawahar, C.V., Zisserman, A.: The truth about cats and dogs. In: Proc. ICCV. (2011)
20. Cristinacce, D., Cootes, T.: Feature detection and tracking with constrained local models. Proc. BMVC (2006) 929–938
21. Milborrow, S., Nicolls, F.: Locating facial features with an extended active shape model. Proc. ECCV (2008) 504–513
22. Saragih, J.M., Lucey, S., Cohn, J.F.: Face alignment through subspace constrained mean-shifts. Proc. ICCV (2009)
23. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **20** (2004)
24. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. Proc. ICCV (2009)
25. Yin, Q., Tang, X., Sun, J.: An associate-predict model for face recognition. Proc. CVPR (2011) 497–504
26. Arca, S., Campadelli, P., Lanzarotti, R.: A face recognition system based on automatically determined facial fiducial points. Pattern Recognition **39** (2006) 432–443
27. Campadelli, P., Lanzarotti, R., Lipori, G.: Precise eye localization through a general-to-specific model definition. Proc. BMVC (2006)
28. Vidaldi, A., Zisserman, A.: Image classification practical. `http://www.robots.ox.ac.uk/~vgg/share/practical-image-classification.htm` (2011)
29. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. Proc. ICCV (2009) 606–613
30. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. Proc. CVPR (2009) 3360–3367