Muhammad Umar Salman
20100120

18th April 2020

Cats and Dogs

AUTHORS

1. Omkar M Parkhi
2. Andrea Vedaldi
3. Andres Zisserman
4. C. V. Jawahar

SUMMARY

This study investigates the fine-grained object categorization in determining the pet family and breed from an image. It has collected the data from the annotated Oxford-IIIT-Pet dataset which covers 37 different breeds of cats (12) and dogs (25). The study uses 200 images for each breed, where 50 are used in training set, 50 in validation set and 100 in the test set. The evaluation protocol that the study has gone with is what they call an 'average per class classification accuracy' which is the diagonal of the row normalized confusion matrix.

To distinguish between the deformable and subtle differences between the breeds, the study uses 2 features.

1. **Shape** which is captured by the deformable part model detecting the pet face.
2. **Appearance** which is captured by the bag of words model that describes the fur of the pet.

This study compares two approaches of classification using these features. The first being the **flat approach** which directly classifies the pet's family and breed simultaneously (37 class problem) and the second being the **hierarchal approach** where the pet family is determined first (2 class problem) then the breed given the family which is either a (12 or 25 class problem).

The study later talks a lot about the details on the two models. Firstly, the shape model which uses a deformable part model in which an object is given by a root part connected with eight small parts where each part is represented by a HOG filter which captures the local distribution of the image edges. However, this model isn't flexible towards the whole pet's body and focuses on modelling the faces of the pet. Yet in the PASCAL VOC challenge (where you had to detect the whole body) it performed pretty well.

Secondly, the appearance model which uses a bag of words model, where visual words are computed by extracting SIFT descriptors, these SIFT descriptors are then quantized using 4000 visual words learned by k-means. These quantized SIFT features are then pooled into a spatial histogram which is L1 normalized and used in an SVM (exponential kernel) for classification. Different variants of spatial histograms can be obtained in correspondence to features of the pet to form 3 different layouts mentioned below.

Three different spatial histogram layouts were used to compute the image descriptors.

1. Image
2. Image + Head
3. Image + Head + Body.

The foreground region (the pet in the image) and background region were needed for computing the appearance descriptors using automatic segmentation where SVM classifier assigns super pixels a confidence score to assign it to either the foreground or the background, this method achieved an accuracy of 65%, 4% more than the last best.

Classification models were then constructed for different combinations of layout type with or without shape and with or with the ground truth segmentation. The table looked as the following:

Shape (Yes/No)	Layout Type (Combinations of the 3 layouts)	Using Ground Truth (Yes/No)

When the model was only trained on the Shape model it gave 94.21% classification accuracy of distinguishing between pet family on the Oxford-IIIT Pet test and also did extremely well in the ASIRRA dataset challenge with a new best of 92.9% mean class accuracy which corresponds to a 42% probability of breaking the test in a single try without even training on the ASSIRA dataset.

When the model was trained only on the Appearance model it gave the its best result with the layout combination: Image + Head + Body with an accuracy of 87.78, an improvement by 2.7% on the layout: Image + Head. However, this accuracy increased by 1% with the inclusion of ground truth segmentation.

The overall best accuracy of all models was including shape and where layout type was Image + Head + Body and using ground truth segmentation instead of automatic segmentation which was a stunning 95.37% accuracy on the Pet family classification. All in all, we can also see that the flat approach did slightly better in all accuracies for all model compared to the hierarchal approach however noting that the latter required less work at test time.