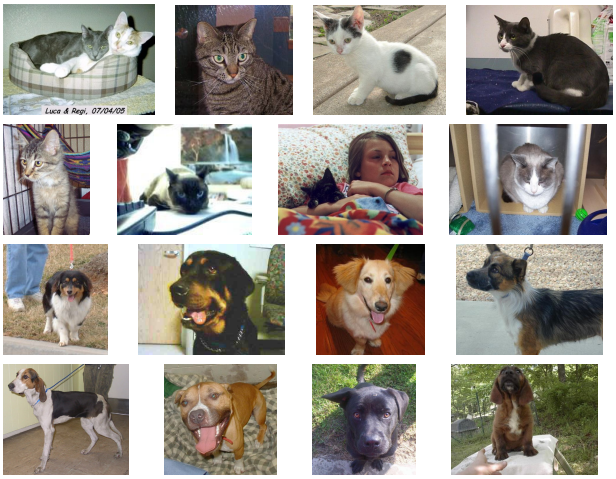

Figure 3. **Example images from the MSR ASIRRA dataset.**

of at least one of them, while humans would make no mistakes. 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 4 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 on 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