for $i = 1, \ldots, N$ and $k = 1, \ldots, K$. The rest of the model follows the generative model of the IBP infinite factor analysis.

Our model for learning image attributes will be based on this infinite sparse factor analysis model. We describe experimental setup, inference algorithm, results, and evaluation in the next sections.

4 Experimental Setup

4.1 Datasets

Currently, our factor analysis models do not explicitely capture spatial information among image pixels. Therefore, it is reasonble for us to select datasets whose images are roughly aligned. Given this information about the datasets, we can compute feature descriptions directly from images without the need to detect interest points. We perform inference algorithms on two datasets. One is the Caltech101 dataset [7], which consists of pictures of objects belonging to 101 categories. There are about 40 to 800 images per category, but most categories have about 50 images. In our experiments, we are focusing on the following seven categories: face ($N = 435$), chair ($N = 61$), lamp ($N = 61$), emu ($N = 53$), flamingo ($N = 67$), windsor chair ($N = 53$), and background ($N = 468$).

The other dataset comes from Sudderth et al. [22]. It consists of images from 16 categories, which fall into three groups: seven animal faces, five animal profiles, and four wheeled vehicles. Figure 3 [22] shows sample images from each category. The dataset also includes images from two additional categories: cannon and human face, as well as the background category. For convenience, we will refer to this dataset as the sixteen-category dataset.

Figure 4: The sixteen-category dataset [22], which consists of a total of $1,885$ images with at least 50 images per category. Images in the dataset come from searches, the Caltech 101 and Weizmann Institute [3] datasets.

4.2 Feature Description

One of the most important steps in any vision task is selecting image representations. Using pixel intensities to represent an image has disadvantages because image intensities are sensitive to small shifts, rotations, and changes in illumination. Instead, our model exploits Histogram of Oriented Gradients (HOG) descriptors [4], which have been widely used in computer vision for the purpose of object detection. They have also been successfully applied to attribute recognition tasks [6, 5]. We expect HOG descriptors to be good at describing parts and textures of an object.

HOG descriptors describe an image with the distribution of intensity gradients or edge directions. To achieve the implementation of these descriptors, we divide the image into regions and then compute histogram of gradient directions for each region. We use the specific implementation from Felzenszwalb et al. [8] with bin size set to 8. In the end, each image in Caltech101 can be represented by a 693-dimensional HOG feature vector, while each image in the sixteen-category dataset can be represented by a $1,089$-dimensional vector.