Figure 4: Inferred distributions over words for LabelMe data, as a function of inferred image category. The letters correspond to paths in Figure 6.

We evaluate the proposed model on the image-classification task, similar to as considered in Li et al. (2010). A set of 800 randomly selected images are held out as testing images from the 8 classes, each class with 100 testing images. Each image is represented by the estimated distribution over all the nodes in the entire hierarchy. Only nodes that are associated to the image have nonzero values in the distribution. We calculate the χ²-distances between the node distribution of the testing images and those of the training images. The KNN algorithm ($K$ is set to be 50) is then applied to obtain the class label.

Figure 5(a) shows the confusion matrix of classification, with an average classification accuracy of 78.3%, compared with 76% in Li et al. (2011). In all of the above examples the dictionary learning was applied directly to the observed pixel values within a given patch, with no 

Figure 5(b) we show the confusion matrix of the model based on SIFT features, with an average accuracy of 76.9%, slightly better than the results reported in Wang et al. (2009) (but here there is no need to tune the number of VQ codes). This also demonstrates that performing dictionary learning directly on the patches, rather than via a state-of-the-art feature extraction method, yields highly competitive results.

in performance manifested with 150 or 400 codes, for example. To further test the proposed model, we considered the same classification experiment on MSRC data, which is characterized by 10 classes. Five images per class were randomly chosen as testing data, and the remaining images are treated as training data to learn a hierarchical structure. An average accuracy of 64% is obtained with the proposed model, compared with 60% using that in Li et al. (2010), where the codebook size is set to be 200. These experiments indicate that the proposed model typically does better than that in Li et al. (2010) for the classification task, even when we optimize the latter with respect to the number of codes.

The experiments above have been performed in 64-bit Matlab on a machine with 2.27 GHz CPU and 4 Gbyte RAM. One MCMC sample of the proposed model takes approximately 4, 2, 8 and 10 minutes respectively for the MNIST, Face, MSRC and LabelMe experiments (in which we simultaneously analyzed respectively 1000, 698, 320, and 800 total images). Note that while these model learning times are relatively expensive, model testing (after the tree and dictionary are learned) is very fast, this employed for the aforementioned classification task. To scale the model up to larger numbers of training images, we may perform variational Bayesian inference rather than sampling, and employ online-learning methods (Hoffman et al., 2010; Carvalho et al., 2010).

6 Conclusions

The nested Dirichlet process has been integrated with dictionary learning to constitute a new hierarchical topic model for imagery. The dictionary learning may be employed on the original image pixels, or on features from any image feature extractor. If words are available, they may be utilized as well, with word-dependent usage probabilities inferred for each path through the tree. The model infers both the tree depth and width. Encouraging qualitative and quantitative results have been demonstrated for analysis of many of the traditional datasets in this field, with comparisons provided to other related published methods.

476