A Bayesian Hierarchical Model for Learning Natural Scene Categories

Li Fei-Fei, Pietro Perona
Arpit Shrivastava, Pranav Maneriker
Guide: Prof. Vinay Namboodiri

Department of Computer Science and Engineering, Indian Institute of Technology Kanpur

Problem

Humans are extremely proficient at perceiving natural scenes and understanding their contents. This paper is discusses a techniques for learning natural scene categories.

Abstract

The authors propose a novel approach to learn and recognize natural scene categories. The method does not require experts to annotate the training set. They represent the image of a scene by a collection of local regions, denoted as codewords obtained by **unsupervised learning**. Each region is represented as part of a "theme".

Introduction

- Classify a scene without first extracting objects.
- The key idea is to use intermediate representation (themes)before classifying scenes.
- In previous work, such themes were learnt from hand-annotations of experts, while method in this paper learns the theme distributions as well as the codewords distribution over the themes without supervision.
- The authors introduce the generative Bayesian hierarchical model for scene categories.

Approach

- An image is modelled as a collection of local patches. Each patch is represented by a codeword from a large vocabulary of codewords.
- The model is an adaptation to vision of ideas proposed by Blei et al. in the context of document analysis (Latent Dirichlet Allocation).

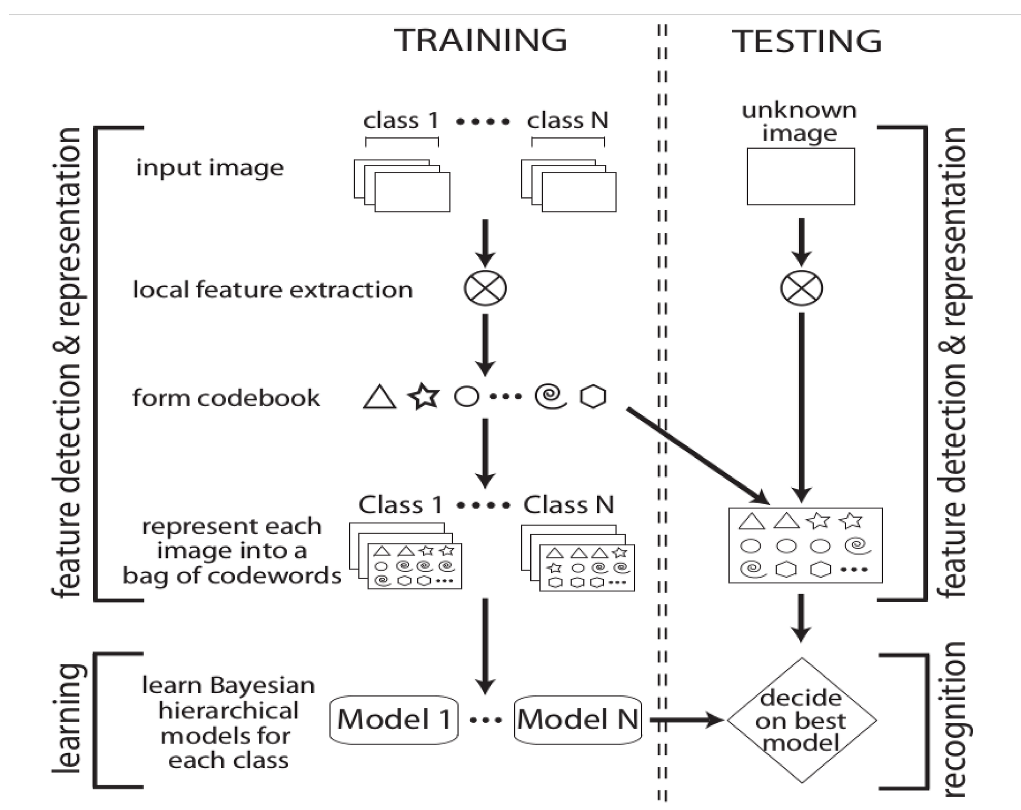

Figure 1: Flow chart of the algorithm.

Model

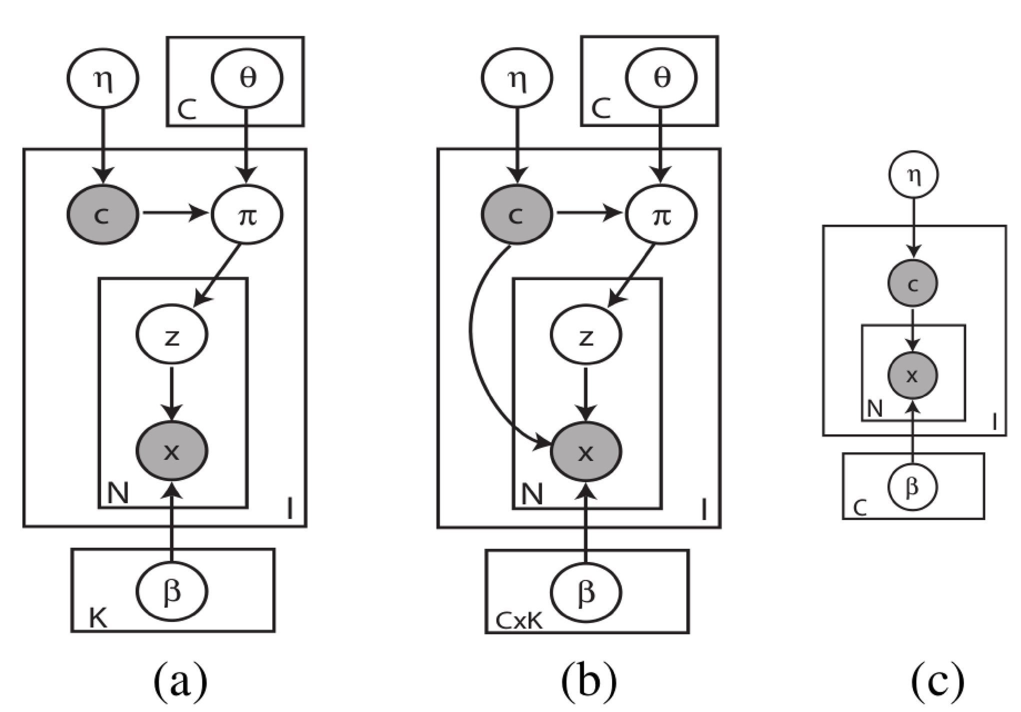

Figure 2: (a) Theme Model 1 for scene categorization that shares both the intermediate level themes as well as feature level codewords. (b) Theme Model 2 for scene categorization that shares only the feature level codewords; (c) Traditional texton model

$$
\begin{aligned}
p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\pi}, c | \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\beta}) &= p(c|\boldsymbol{\eta}) p(\boldsymbol{\pi}|c, \boldsymbol{\theta}) \cdot \\
& \prod_{n=1}^{N} p(z_n|\boldsymbol{\pi}) p(x_n|z_n, \boldsymbol{\beta}) \\
p(c|\boldsymbol{\eta}) &= \mathrm{Mult}(c|\boldsymbol{\eta}) \\
p(\boldsymbol{\pi}|c, \boldsymbol{\theta}) &= \prod_{j=1}^{C} Dir(\boldsymbol{\pi}|\boldsymbol{\theta}_{j\cdot})^{\delta(c,j)} \\
p(z_n|\boldsymbol{\pi}) &= \mathrm{Mult}(z_n|\boldsymbol{\pi}) \\
p(x_n|z_n, \boldsymbol{\beta}) &= \prod_{k=1}^{K} p(x_n|\boldsymbol{\beta}_{k\cdot})^{\delta(z_n^k, 1)}
\end{aligned}
$$

Learning

The authors maximize the log likelihood term $logp(x|\theta, \beta, c)$ by estimating the optimal θ and β. The learning is done using variational inference. The algorithm is used is the EM algorithm iterated until the model parameter values converge.

Classification

An unknown image is first represented by a collection of patches, or codewords. Given x, we would like to compute the probability of each scene class.

$$p(c|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta}) \propto p(\boldsymbol{x}|c, \boldsymbol{\theta}, \boldsymbol{\beta}) p(c|\boldsymbol{\eta}) \propto p(\boldsymbol{x}|c, \boldsymbol{\theta}, \boldsymbol{\beta})$$

$$p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\beta}, c) = \int p(\boldsymbol{\pi}|\boldsymbol{\theta}, c) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\boldsymbol{\pi}) p(x_n|z_n, \beta) \right) d\boldsymbol{\pi}$$

this equation is not tractable and a wide range of approximate inference algorithms can be considered, including Laplace approximation, variational approximation and MCMC method for solving it.

Dataset & Experimental Setup

- Dataset contains 13 categories of natural scenes.
- Average size of each image is approximately 250 ÃŮ 300 pixels
- scenes were split randomly into two separate sets of images, N (100) for training and the rest for testing
- The performance metric is the average value of the diagonal entries of the confusion table.

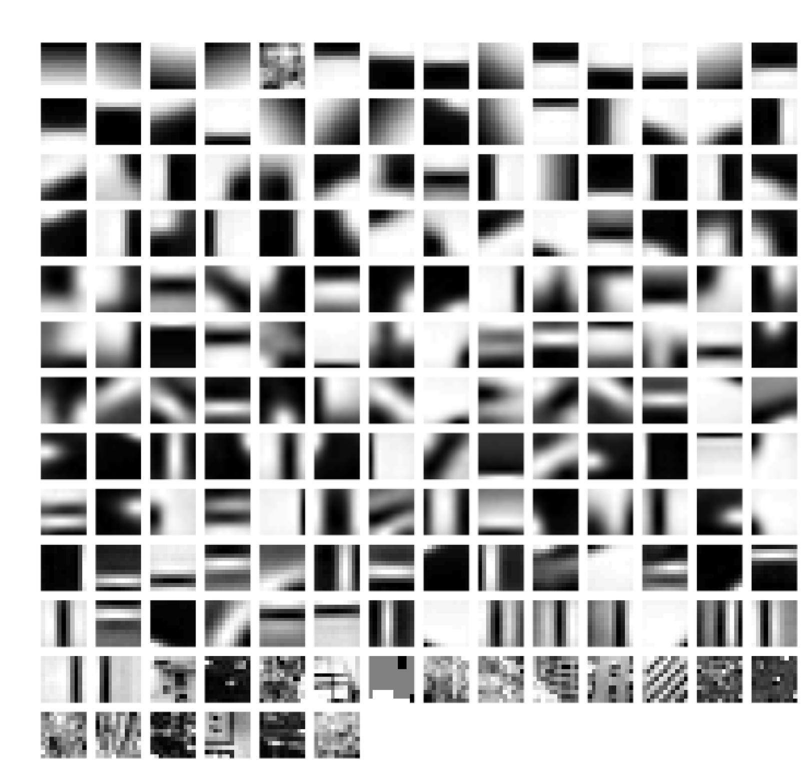

Figure 3: Codebook obtained from 650 training examples from all 13 categories (50 images from each category). Image patches are detected by a sliding grid and random sampling of scales

Results

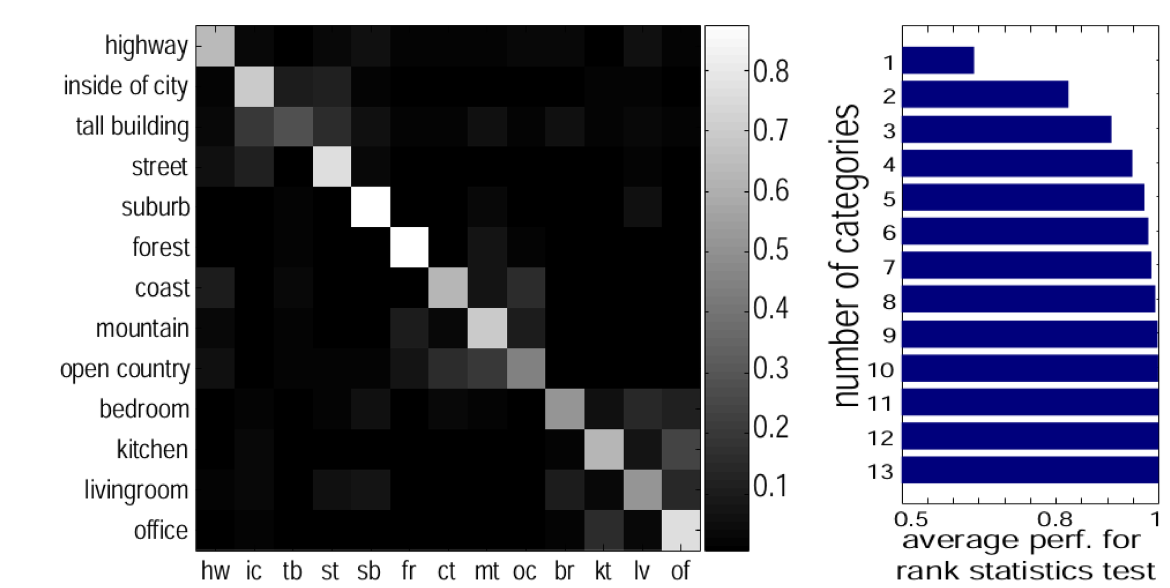

Figure 4: Left: Confusion table Right: Rank statistics

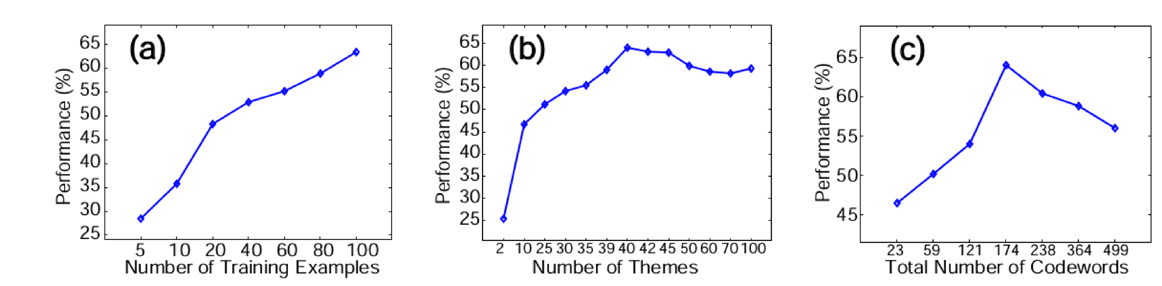

Figure 5: Performance with different parameters

Feature detection and representation

Descriptor	Grid	Random	Saliency	DoG
11 x 11 Pixel	64.0	47.5	45.5	N/A
128-dim Sift	65.2	60.7	53.1	52.5

References

[1] Li Fei-Fei and Pietro Perona.
A bayesian hierarchical model for learning natural scene categories.
In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524–531. IEEE, 2005.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan.
Latent dirichlet allocation.
the Journal of machine Learning research, 3:993–1022, 2003.