Discovering Objects and their Location in Images

Josef Sivic1, Bryan C. Russell2, Alexei A. Efros3, Andrew Zisserman1 and William T. Freeman2
1Oxford University                     2MIT 3Carnegie Mellon University

Introduction

Goal: Discover visual object categories and their segmentation given a collection of unlabeled images

Approach: 1) Represent an image as a collection of visual words
2) Apply topic discovery models from statistical text analysis

Image representation

Represent an image as a histogram of “visual words”

Examples of visual words

- Detect affine covariant regions
- Represent each region by a SIFT descriptor
- Build visual vocabulary by k-means clustering (K~1,000)
- Assign each region to the nearest cluster centre

Examples of visual words

- Detect affine covariant regions
- Represent each region by a SIFT descriptor
- Build visual vocabulary by k-means clustering (K~1,000)
- Assign each region to the nearest cluster centre

Overview

Find visual words

Form histograms

Documents

Discover topics

The topic discovery models

Probabilistic Latent Semantic Analysis (pLSA) [Hofmann’99]

\[ P(z|d) = \sum_z P(z)P(w|z)P(d|w) \]

pLSA Model fitting:

Find topic vectors \( P(z|d) \) common to all documents and mixture coefficients \( P(z|d) \) specific to each document. Fit model by maximizing likelihood of data using EM.

Latent Dirichlet Allocation (LDA) [Blei et al.’03]

Treat multinomial weights over topics as random variables. Fit model using Gibbs sampling [Griffiths and Steyvers’04].

Experiment I: Caltech Dataset

Four object categories: faces, motorbikes, airplanes and cars rear (total of 3,190 images) and 900 background images

Image Classification

Assign each image to a topic with the highest \( P(z|d) \)

Learn \( K = (5, 6, 7) \) topics

Background is better modeled by multiple topics

Performance on novel images is comparable with semisupervised method of [Fergus et al.’03]

Experiment II: MIT dataset

- 2873 images, learn 10 topics

Results

Results shown only for pLSA. LDA had very similar performance.

Experiment III: Application to image retrieval

Learn topic vectors on Caltech database

Represent new query image in terms of learned topic vectors

Improve localization using doubles

Form a new vocabulary from pairs of locally co-occurring regions

Doublet formation

Doublet example I

Doublet example II

All detected visual words

Single segmentation

Doublet segmentation