Discovering Objects and their Location in Images

Josef Sivic\(^1\), Bryan C. Russell\(^2\), Alyosha Efros\(^3\), Andrew Zisserman\(^1\) and William T. Freeman\(^2\)

\(^1\)Oxford University \hspace{1cm} \(^2\)CSAIL MIT \hspace{1cm} \(^3\)Carnegie Mellon University

**Introduction**

**Goal:** Discover visual object categories given an collection of unlabelled 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”

- Detact 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 center

Examples of visual words

- Five samples from a motorbike visual word
- Five samples from an airplane visual word

Overview

- Find visual words
- Form histograms
- Discover topics

**The topic discovery model**

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

Model fitting:

Find topic vectors \( P(w|z) \) common to all documents and mixture coefficients \( P(z|d) \) specific to each document.

\[
P(w_i|d_j) = \sum_z P(z_i|d_j)P(w_i|z_k)
\]

Fit pLSA parameters by maximizing likelihood of all images using EM

\[
L = \prod_{i=1}^{N} P(w_i|d_j)\prod_{i=1}^{M} P(d_j|w_i)
\]

- \( L \) = number of visual words
- \( N \) = number of images

**Results**

**Image Classification**

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

- Experiment I.
  - 4 object categories (faces, motorbikes, airplanes, cars) and background images
  - Learn \( K = (5,6,7) \) topics
  - Background is better modelled by multiple topics

- Two most likely visual words for each learned topic shown by five samples in a row

- Experiment II: MIT dataset
  - 2873 images, learn 10 topics
  - 4 of the 10 learned topics shown by the 5 most probable images for each topic

**Segmentation**

For a given word \( w_i \) in document \( d_j \) examine posterior probability over topics

\[
P(z_i|w_i,d_j) = \frac{P(w_i|z_k)P(z_k|d_j)}{\sum_k P(w_i|z_k)P(z_k|d_j)}
\]

- Faces
- Motorbikes
- Airplanes
- Cars
- Background I
- Background II
- Background III

Visual words colour coded according to the topic with the highest probability

- Example motorbike segmentation
- Example airplane segmentation

Improving localization using doublets

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

- Doublet example I
- Doublet example II
- Singlet segmentation
- Doublet segmentation

- Example motorbike segmentation
- Example airplane segmentation

Examples of visual words

- Faces
- Motorbikes
- Airplanes
- Background I
- Background II
- Background III

Other logos?