Learning Object Categories from Google’s Image Search

Fergus, Perona, Fei-Fei, Zisserman

Introduction

Contributions
- TSI-pLSA - a translation and scale invariant pLSA model
- Unsupervised learning on training set collected from Google image search and therefore unlabeled

Related work
- Discovering objects and their locations in images
  - Sivic, Russell, Efros, Zisserman, Freeman
- pLSA for object category recognition and segmentation
- A visual category filter for Google images
- Fergus, Perona, Zisserman
- Reranking of Google images by learning a model

Main challenge
- Noisy training data: Less than 15% of the images returned by Google are related to the keyword
- Large variations in scale, position, and pose

Idea
- Build pLSA model with a number of topics
- Visual words of an image will fall under a common topic
- Visual words of positive examples will be similar
- Find this topic using a validation set of less noisy data

Overview

pLSA

Generative model
- Choose document (image) d with probability P(d)
- Choose topic z with probability P(z|d)
- Choose word w with probability P(w|z)
- Thus, P(d, z, w) = P(d)P(z|d)P(w|z)

Eg. Face

Learning
- Using EM
  - E-step: estimate P(z|d, w)
    - Associates z with the image and features
  - M-step: update P(z|d) and P(w|z)
    - Visual words from an image tend to fall under the same topic

Recognition
- Fix P(w|z) and estimate P(z|d) using EM

Drawbacks
- Spatial information is not used
- Multiple instances of a category cannot be captured
Generative model summary

- Choose document d with probability P(d)
- Choose topic z with probability P(z/d)
- Choose word w, location x with probability P(w,x/z)

Thus,

\[
P(w,x,z) = P(d)P(z|d)P(w|x,z)
\]

Drawback

- Uses absolute location of feature
- Not translation or scale invariant

Location of feature calculated with respect to object centroid

\[
x_{object} = \text{mean}, \quad \text{scale} = \text{variance}
\]

A grid of X locations within the bounding box and one background bin for feature location

Object centroid and scale captured in 4-vector latent variable c

Marginalizing over the entire range of c is not feasible

Small set of c values estimated during learning and recognition

Selecting the final classifier

Visual words of positive examples should belong to a common topic

Validation set will perform best under this common topic

Selecting the number of topics Z

Chosen empirically

ROC vs number of topics plotted for best topic under validation set and best topic under test set

Some Issues

- Selecting the final classifier
- Visual words of positive examples should belong to a common topic
- Validation set will perform best under this common topic
- Selecting the number of topics Z
- Chosen empirically
- ROC vs number of topics plotted for best topic under validation set and best topic under test set
Datasets

- **Training**
  - Google dataset
    - Images automatically downloaded from Google image search using the category name
  - Validation set – first five images from image search in 7 different languages
  - Other
    - Manually gathered frames from Caltech and Pascal datasets
- **Testing**
  - Manually gathered frames from Caltech and Pascal datasets

Parameters

- 700 regions per image using 4 different region detectors
  - Because the method requires large number of data for parameter estimation
- SIFT descriptor of 72 dimensions
  - Larger histogram bins more appropriate for object categorization
- K-means clustering with $k=350$ to obtain 350 visual words
- Number of grid positions $X_{fg} = 37$
- Number of topics $Z = 8$

Experiments and Results

- **Experiment 1** (standard datasets)
  - Caltech
    - 50 training images from Caltech, 2 topics
    - No clear winner
  - Pascal
    - Foreground and background images combined into one training set – unsupervised learning
    - 6 topics and best topic chosen using performance on foreground only images
    - TSI-pLSA performs better than the other methods
- **Experiment 2** (Google data)
  - Training images from Google image search
  - 8 topics and best topic chosen using Google validation set
  - TSI-pLSA performs better than the other methods except Google and background
  - ABS-pLSA and TSI-pLSA not rotation invariant
  - 6 topics and best chosen using performance on foreground only images
  - TSI-pLSA performs better than the other methods

- **Experiment 3**
  - Comparison with other supervised methods
  - TSI-pLSA is slightly worse than the other methods but it is unsupervised
- **Experiment 4**
  - Improving Google’s Image search
  - Best topic from 8 topics trained on raw Google data

Conclusions

- All three methods work on unlabeled Google dataset and automatically collected validation set and TSI-pLSA performs best
- TSI-pLSA identifies multiple instances of objects in images
- Can be used to rank images returned by Google
- TSI-pLSA performs badly when objects are rotated