Learning Object Categories from Google's Image Search

Fergus, Perona, Fei-Fei, Zisserman

Presented by
Sudheendra

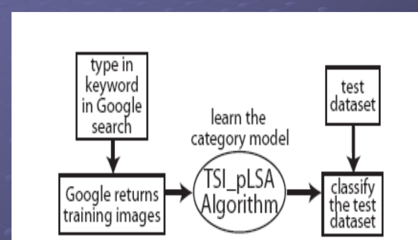
Introduction

- 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

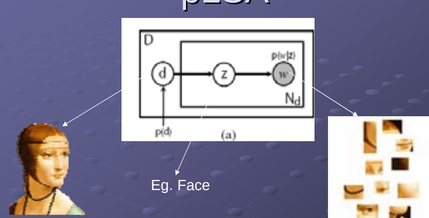
Introduction

- 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

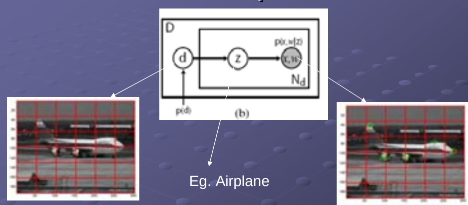


- Generative model
  - Choose document (image) d with probability P(d)
  - Choose topic z with probability P(z/d)
  - Choose word with probability P(w/z)
  - Thus, $P(d,w) = P(d)P(w|d)$, where $P(w|d) = \sum_{z \in Z} P(w|z)P(z|d)$.

pLSA

- Learning
  - Using EM
  - E-step: estimate P(z | d, w)
    - Associates z with the image and feature
  - 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

ABS-pLSA

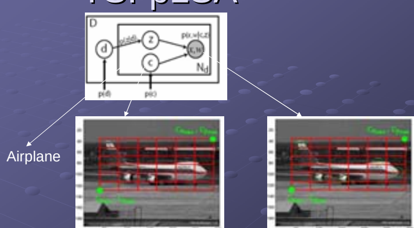


- Generative model summary
  - Choose document d with probability P(d)
  - Choose topic z with probability P(z/d)
  - Choose word w, location x with probability P(w,x/z)
  - Thus, $P(w,x,d) = \sum_{z=1}^{Z} P(w,x|z)P(z|d)P(d)$

ABS-pLSA

- Quantize the image into X bins
- Include spatial location with word to produce topic variable
- EM steps similar to pLSA
- Drawback
  - Uses absolute location of feature
  - Not translation or scale invariant

TSI-pLSA

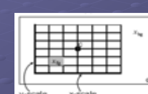


- Generative model summary
  - Choose document d with probability d
  - Choose topic z with probability P(z/d)
  - Choose word w, relative position x over all possible values of centroid c with probability P(w,x/z) calculated as

$$P(w,x|z) = \sum_c P(w,x,c|z) = \sum_c P(w,x|c,z)P(c)$$

TSI-pLSA

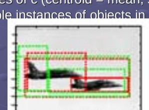
- Location of feature calculated with respect to object centroid



  - x-scale and y-scale along with the centroid specify the bounding box of the object
  - A grid of X_{fg} 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

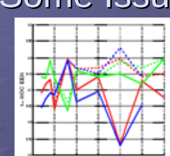
TSI-pLSA

- Learning
  - Estimating c values
    - Standard pLSA run on the training set
    - k = (1 … K) gaussians fitted to the locations of features weighted by p(w | z) to obtain k values of c (centroid = mean, scale = variance)
    - Captures multiple instances of objects in image



  - EM: as before with P(w, x | z) estimated by marginalizing over above found c values
- Recognition
  - Estimating c values
    - Similar to above, $(\hat{P}(w|z) = \sum_{x_{fg}} P(w,x|z))$ used to weight gaussians
  - EM: Lock P(w | z) and iterate to find P(z | d) summed over the found c values

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
    - 50-50 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 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 in all categories except Guitar
  - • ABS-pLSA and TSI-pLSA are not rotation invariant

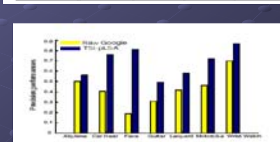
	pLSA	ABS	TSI	pLSA	ABS	TSI
Category	Prep.	Prep.	Prep.	Google	Google	Google
(A)irplane	17.7	13.2	4.7	24.7	17.2	15.5
(C)ars Rear	2.0	0.2	0.7	21.0	13.2	16.0
(F)ace	22.1	11.5	17.0	20.3	36.4	20.7
(G)uitar	9.3	10.0	14.4	17.6	62.0	31.8
(L)eopard	12.0	12.0	11.0	15.0	16.0	13.0
(M)otorbike	19.0	6.0	7.0	15.2	18.5	6.2
(W)rist watch	21.6	7.7	15.5	21.0	20.5	19.9
PASCAL Car	31.7	33.0	25.8	-	-	-
PASCAL Motorbike	33.7	30.2	25.7	-	-	-

Experiments and Results

- Experiment 3
  - Comparison with other supervised methods
  - TSI-pLSA is slightly worse than the other methods but it is unsupervised

Dataset	TSI-pLSA	[10]	[18]	[15]
Expt. B	Img. labels	Img. labels	Img. labels	Segmented
PASCAL Car	25.8 / 0.062	-	-	34.2 / 0.181
PASCAL Motorbike	25.7 / 0.249	-	-	31.7 / 0.341
Expt. C	None	Img. labels	Img. labels	Segmented
Airplane	15.5	7.0	11.1	-
Cars Rear	16.0	9.7	8.9	6.1
Face	20.7	3.6	6.5	-
Leopard	13.0	10.0	-	-
Motorbike	6.2	6.7	7.8	6.0

- 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