Discovering objects and their location in images

Josef Sivic et al.
ICCV 2005

Presented by Elden Yu, with quite some slides borrowed

Goal: Given a collection of unlabelled images,
discover visual object categories
and their segmentation

Which images contain the same object(s) ?

Where is the object in the image?

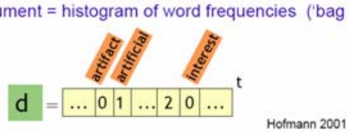
Visual object classes

Motivation

- Is it possible to learn object classes without supervision?
  - Results in speech recognition highlight the importance of huge amounts of training data
  - Annotation is expensive
  - Much more unsupervised data than labeled data
- The success in text mining
  - The bag of words representation
  - Latent semantic analysis

Analogy: Discovering topics in text collections

Document = histogram of word frequencies ('bag of words' model)



Hofmann 2001

Term-document matrix

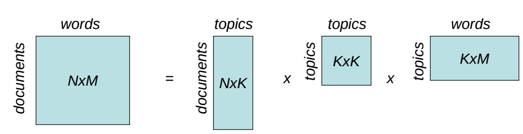
Hofmann: *Probabilistic latent semantic analysis*

Blei et al.: *Latent Dirichlet Allocation*

Griffiths and Steyvers: *Finding Scientific Topics*

Latent Semantic Analysis

- D = {d1,…,dN} N documents
- W = {w1,…,wM} M words
- Nij = #(di,wj) NxM co-occurrence term-document matrix



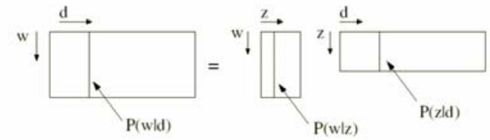
- A kind of Singular Value Decomposition to represent each document by its top K topics instead of by its M words

Probabilistic Latent Semantic Analysis (pLSA)

[Hofmann '99]

d … documents (images)
w … visual words
z … topics ('objects')

$$P(w_i|d_j) = \sum_{k=1}^{K} P(z_k|d_j)P(w_i|z_k)$$



odel fitting: find topic vectors P(w|z) common to all documents, and mixture coefficients P(z|d) specific to each documen

Learning the pLSA parameters

Observed counts of word *i* in document *j*

$$L = \prod_{i=1}^{M}\prod_{j=1}^{N} P(w_i|d_j)^{n(w_i,d_j)}$$

Unlike LSA, pLSA does not minimize any type of 'squared deviation.' The parameters are estimated in a probabilistically sound way.

Maximize likelihood of data using EM.
Minimize KL divergence between empirical distribution and model

EM for pLSA

- E-step: compute posterior probabilities for the latent variables

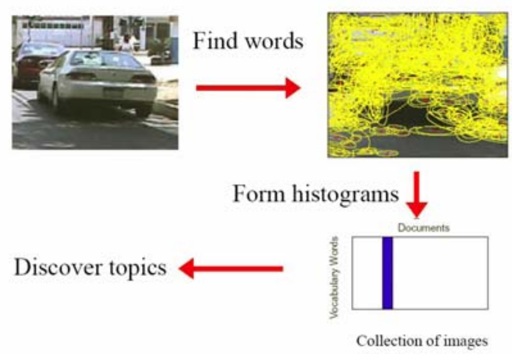
$$P(z_k|d_i,w_j) = \frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{l=1}^{K} P(w_j|z_l)P(z_l|d_i)}.$$

- M-step: maximize the expected complete data log-likelihood

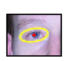
$$P(w_j|z_k) = \frac{\sum_{i=1}^{N} n(d_i,w_j)P(z_k|d_i,w_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N} n(d_i,w_m)P(z_k|d_i,w_m)}.$$

$$P(z_k|d_i) = \frac{\sum_{j=1}^{M} n(d_i,w_j)P(z_k|d_i,w_j)}{n(d_i)}.$$

Visual object discovery - overview

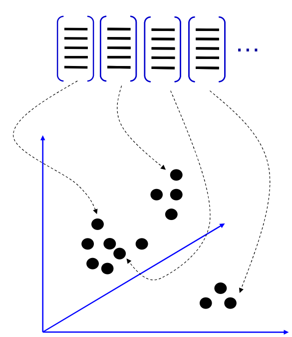


Visual Words

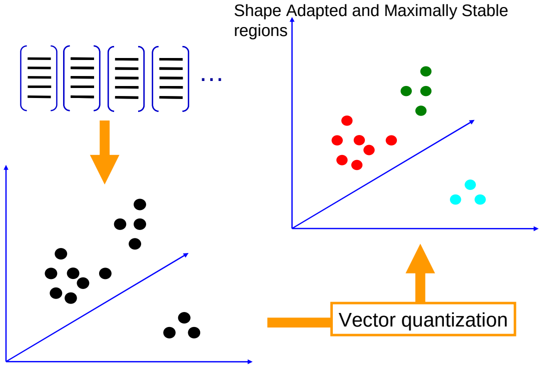



- Vector Quantized SIFT descriptors computed in regions
- Regions come from elliptical shape adaptation around interest point, and from the maximally stable regions of Matas et al.
- Both are elliptical regions at twice their detected scale

Building a Vocabulary



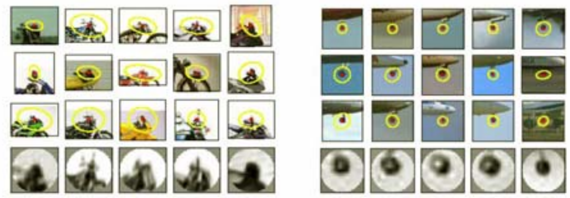
K-means clustering of 300K regions to get about 1K clusters for each of Shape Adapted and Maximally Stable regions



Visual words

Build visual vocabulary by k-means clustering (K~2,000) from Caltech4

Capture some intra-class variations



ikolaiczyk & Schmid'02 Schaffalitzky & Zisserman'02 Matas et al. '02 Lowe'99 Sivic & Zisserman'03]

Experiment on topic discovery

- For each image d_j
  - Compute $P(z_k|d_j)$ over k, and classify the image as containing object k according to the max of $P(z_k|d_j)$ over k
- With(1)/without(2) explicit background

Ex	Categories	K	pLSA %	pLSA #	KM baseline %	KM baseline #
(1)	4	4	98	70	72	908
(2)	4 + bg	5	78	931	56	1820
(2)*	4 + bg	6	76	1072	–	–
(2)*	4 + bg	7	83	768	–	–
(2)*	4 + bg-fxd	7	93	238	–	–

Table 1: Summary of the experiments. Column '%' shows the classification accuracy measured by the average of the diagonal of the confusion matrix. Column '#' shows the total number of misclassifications. See text for a more detailed description of the experimental results. In the case of (2)* the two/three background topics are allocated to one category. Evidently the baseline method performs poorly, showing the power of the pLSA clustering.

Experiment I.

Faces	435
Motorbikes	800
Airplanes	800
Cars (rear)	1155
Background	900
Total:	**4090**

Experiment on classifying new images

- For each image d_j
  - Compute $P(z_k|d_{test})$ over k, with the fold-in heuristic
  - Classify the image as containing object k according to the max of $P(z_k|d_j)$ over k
- Experiment (3) is a modification of (2)

True Class →	Faces	Motorb	Airplan	Cars rear
Topic 1 - Faces	99.54	0.25	1.75	0.75
Topic 2 - Motorb	0.00	96.50	0.25	0.00
Topic 3 - Airplan	0.00	1.50	97.50	0.00
Topic 4 - Cars rear	0.46	1.75	0.50	99.25

Table 3: Confusion table for unseen test images in experiment (3) – classification against images containing four object categories, but no background images. Note there is very little confusion between different categories. See text.

Experiment on classifying new images

- Experiment (4) is to compare with that of Fergus, with quite some hard heuristics

Percent ROC equal error rate

Category	pLSA	Constellation
Faces	**3.3**	3.6
Motorbikes	8.0	**6.7**
Airplanes	**1.6**	7.0
Cars	**7.0**	9.7

- Comparable performance to constellation model
- Level of supervision:
  - pLSA: one number (of topics)
  - Constellation: 400 labels for each category
- Also an indication of the level of difficulty of the Caltech 4 dataset

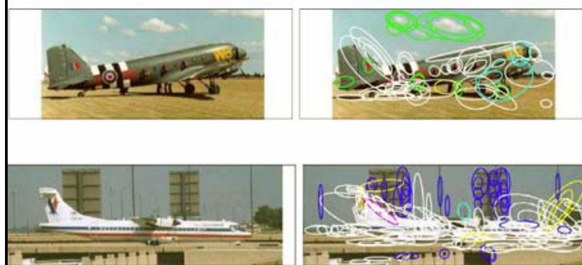
Image as a mixture of topics (objects)

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

Use posterior over topics
to classify individual visual words:

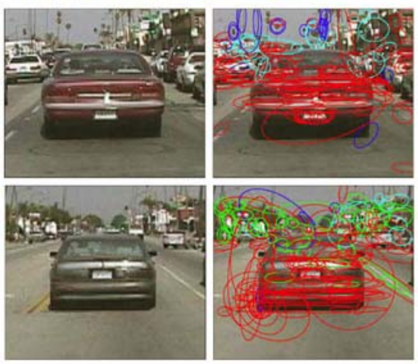
$$P(z_k|w_i,d_j) = \frac{P(w_i|z_k)P(z_k|d_j)}{\sum_{l=1}^{K} P(w_i|z_l)P(z_l|d_j)}$$

Example (sparse-) segmentations



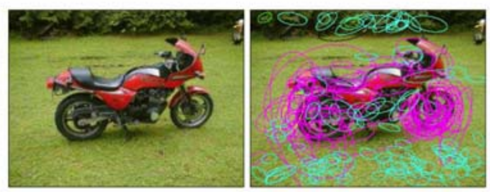
Faces Motorbikes Airplanes Cars Bg I, II, III

Example (sparse-) segmentations



Faces Motorbikes Airplanes Cars Bg I, II, III

So far: Image as a 'bag-of-visual-words'



Faces Motorbikes Airplanes Cars Bg I, II, III

Shortcomings:

- soft segmentation
- all spatial relations between visual words are lost

Use image segmentation to propose groupings of visual words