Learning about images from keyword-based Web search

CS 395T: Visual Recognition and Search
February 15, 2008
David Chen

Problems with traditional training data for object recognition

- Time-consuming and difficult to construct
  - Collect
  - Annotate
  - Align
  - Crop
- Bias in the types of images
- Does not reflect images encountered in the real world

Problems with traditional training data for object recognition

Caltech 101 “airplane”

Collecting images from the Web

- Pros
  - Large scale of freely available images
  - More representative of real-world images
- Cons
  - Lack of annotations
  - Data extremely noisy

Flickr Commons

mechanics, America, civil air patrol base, Maine, vintage, 1940s, historical photographs, slide film, 4x5, large format, LF, transparencies, transparency, CAP, Civil Air Patrol, Bar Harbor, Bar Harbor, ME, maintenance, rotary engine, propeller, fixed gear

General framework for object recognition

Gather raw data

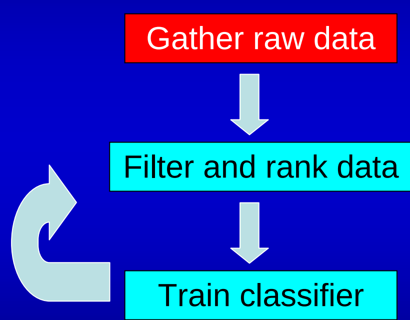
↓

Filter and rank data

↓

Train classifier

General framework for object recognition



Gathering raw data

- Image search engine
  - Extremely noisy
- Text search engine
  - Fairly robust result
  - Does not always return images
- Application-specific database
  - Bootstrapped to index the entire Web (Yeh, Tollmar, Darrell. CVPR 2004)

Image search engine

- Search with desired category name
- Search with additional words
  - Monkey zoo, monkey animal, monkey primate, monkey wild, monkey banana, etc
- Search in translated terms
  - Chinese, French, Spanish, Korean, etc

Image search engine

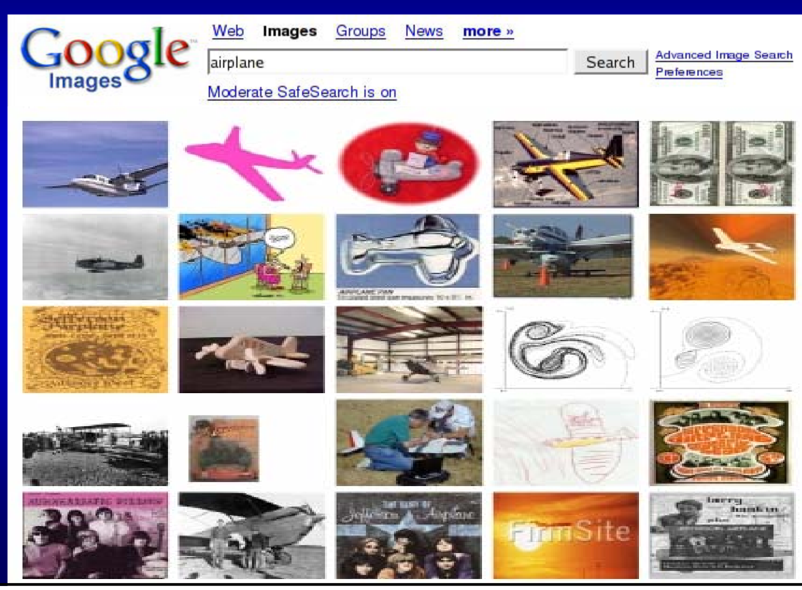


Image search engine

Search in translated terms

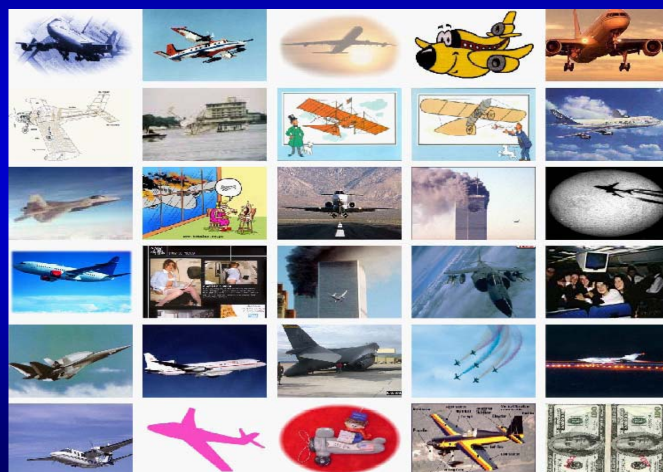
Flugzeug

Aeroplano

Avion

Avião

Airplane



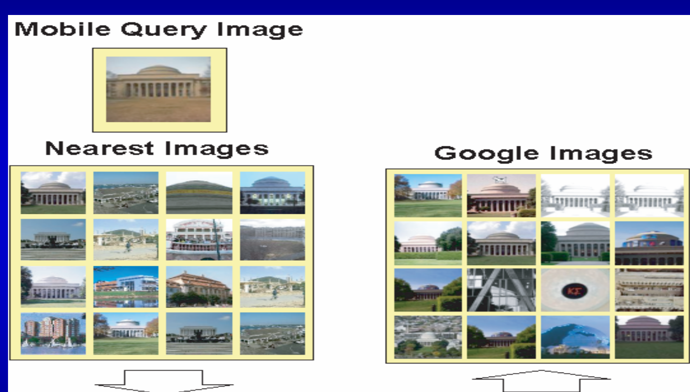
Text search engine

- Similar searching methods as image search engines
- Crawl returned pages for images
- Follow links on returned pages

Application-specific database

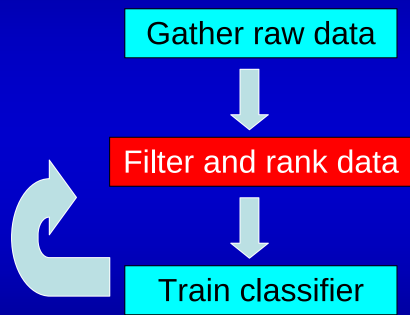
- A relatively small database of images
- Designed for quick image-based search
- Extract keywords from returned web pages
- Use extracted keywords to search text-based search engines

Application-specific database



MIT, story, engineering, kruckmeyer, boston, foundataion relations, MIT dome, da lucha, view realvideo, cancer research

General framework for object recognition



Removing Abstract Images

- Abstract images don’t look like realistic natural images
  - Drawings, non-realistic paintings, comics, casts or statues
- Difficult to do automatically

Removing Abstract Images

Train a SVM on hand-labeled dataset
(Schroff, Criminisi, Zisserman. ICCV 2007)

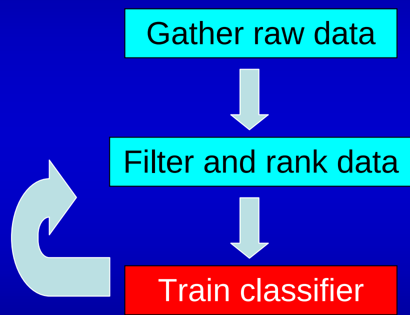
Drawings & Symbolic

Non Drawings & Symbolic

Ranking Images

- Use classifiers to rank the images
- Need data to train classifiers
- Train on a subset of higher precision data
- Build generic classifiers

General framework for object recognition



Features

Text

- Keyword used to search for the image
- HTML tag
- Context
- File name, directory

Image

- Kadir & Brady saliency operator
- Multi-scale Harris detector
- Difference of Guassians
- Edge based operator

Feature Representations

Text

- Binary Features
- TF-IDF
- Learning related words associated with the category
  - Using LDA (Berg, Forsyth. CVPR 2006)

Image

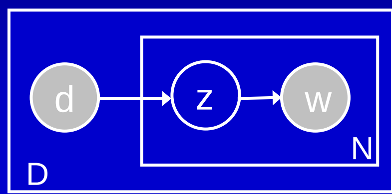
- SIFT
- Color histogram
- Energy spectrum
- Wavelet decompositions

Classifiers

- Bayesian network
- Hierarchical Bayesian text models
  - probabilistic Latent Semantic Analysis (pLSA)
  - Latent Dirichlet Analysis (LDA)
  - Hierarchical Dirichlet Processes (HDP)
- SVM
- Multiple instance learning (Vijayanarasimhan, Grauman. UTCS Tech report 2007)

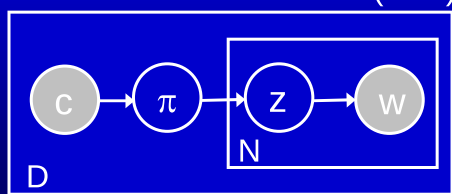
Hierarchical Bayesian text models

Probabilistic Latent Semantic Analysis (pLSA)



Hoffman, 1999

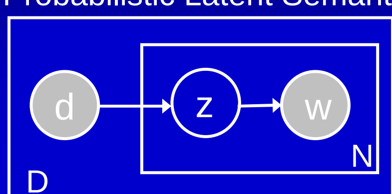
Latent Dirichlet Allocation (LDA)



Blei et al., 2001

Hierarchical Bayesian text models

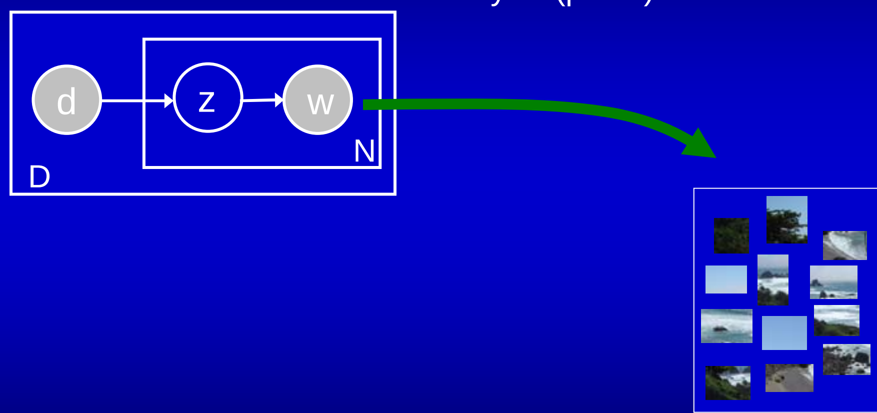
Probabilistic Latent Semantic Analysis (pLSA)



Sivic et al. ICCV 2005

Hierarchical Bayesian text models

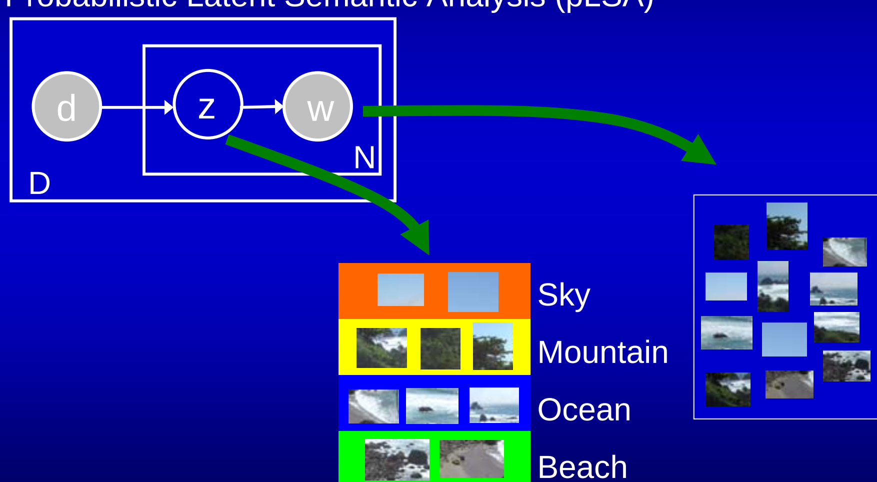
Probabilistic Latent Semantic Analysis (pLSA)



Sivic et al. ICCV 2005

Hierarchical Bayesian text models

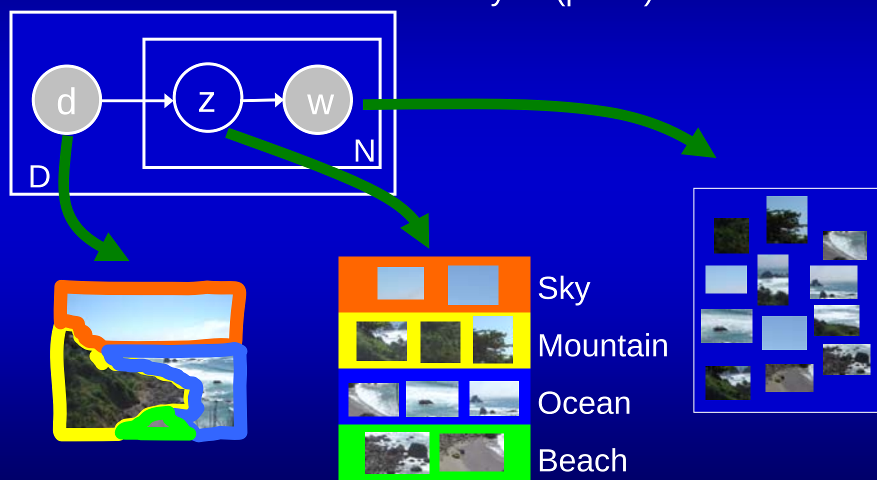
Probabilistic Latent Semantic Analysis (pLSA)



Sivic et al. ICCV 2005

Hierarchical Bayesian text models

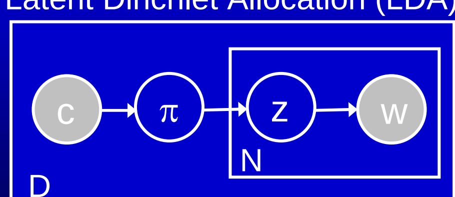
Probabilistic Latent Semantic Analysis (pLSA)



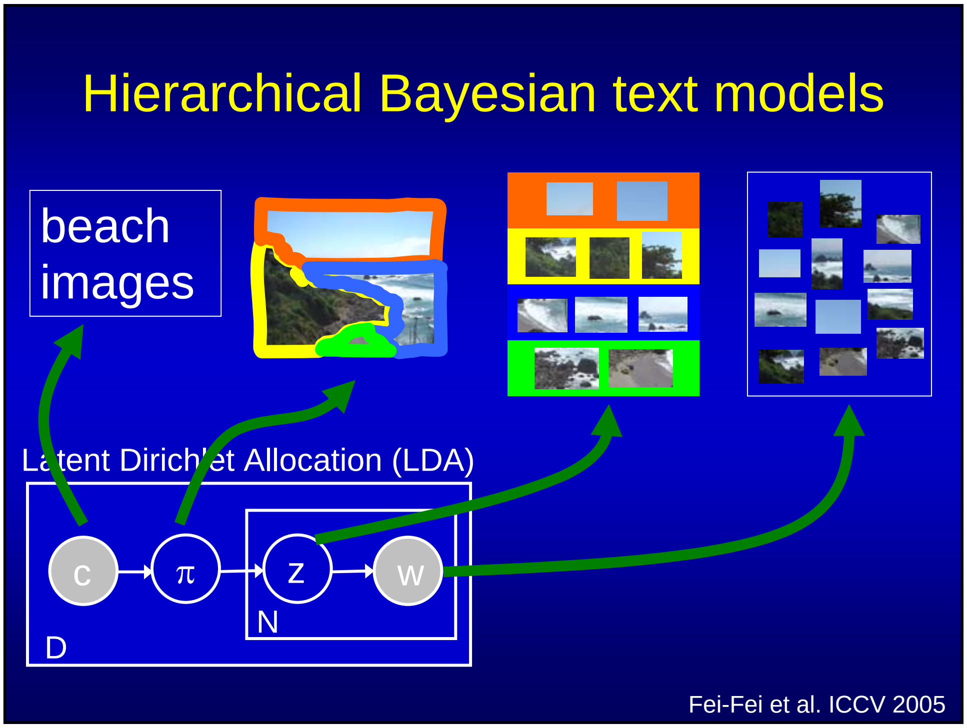
Sivic et al. ICCV 2005

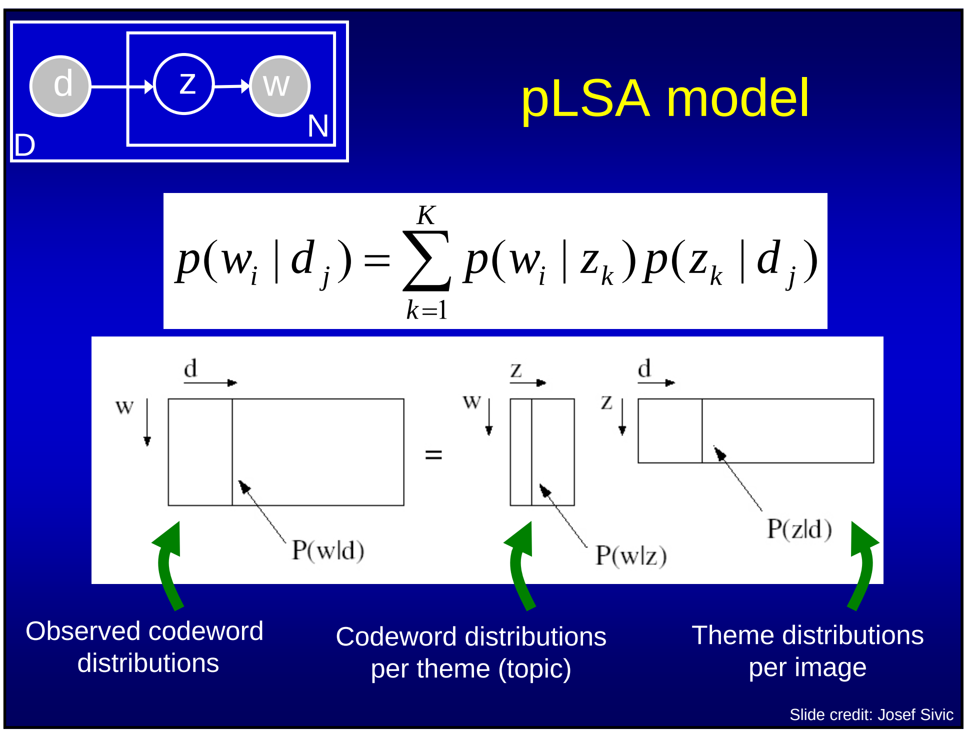
Hierarchical Bayesian text models

Latent Dirichlet Allocation (LDA)



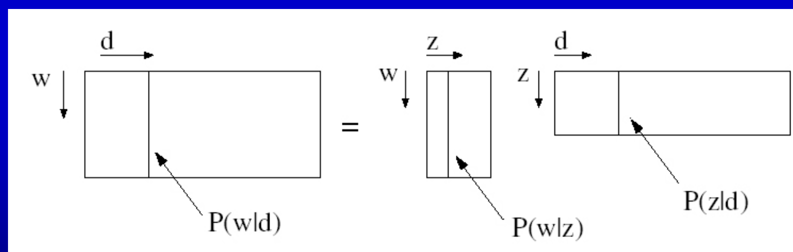
Fei-Fei et al. ICCV 2005


Hierarchical Bayesian text models
beach images
Latent Dirichlet Allocation (LDA)
c
π
z
w
N
D
Fei-Fei et al. ICCV 2005


d
z
w
N
D
pLSA model
$p(w_i \mid d_j) = \sum_{k=1}^{K} p(w_i \mid z_k) p(z_k \mid d_j)$
d
w
=
z
w
d
z
P(w|d)
P(w|z)
P(z|d)
Observed codeword distributions
Codeword distributions per theme (topic)
Theme distributions per image
Slide credit: Josef Sivic

Recognition using pLSA

$$z^* = \arg\max_z p(z \mid d)$$



Slide credit: Josef Sivic

Learning the pLSA parameters

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

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

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

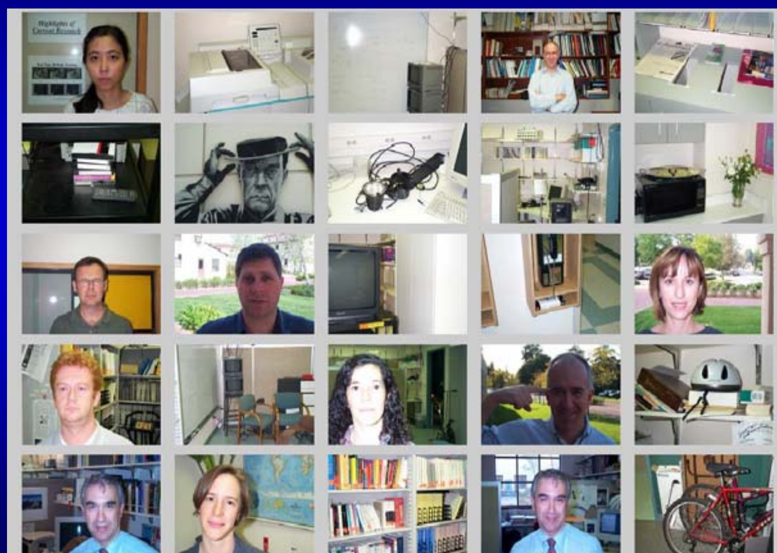
Maximize likelihood of data using EM

M … number of codewords

N … number of images

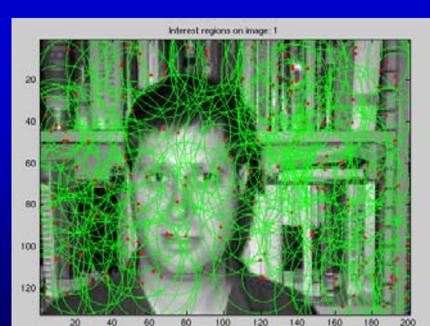
Slide credit: Josef Sivic

task: face detection – no labeling



Demo: feature detection

- Output of crude feature detector
  - Find edges
  - Draw points randomly from edge set
  - Draw from uniform distribution to get scale



Demo: learnt parameters

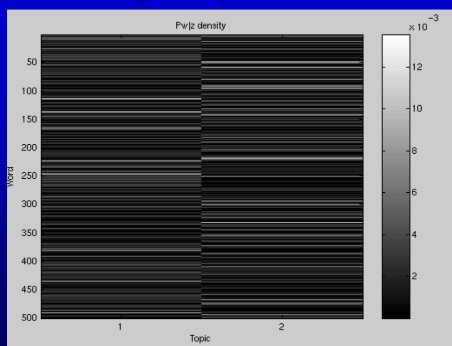
- Learning the model: do_plsa('config_file_1')
- Evaluate and visualize the model: do_plsa_evaluation('config_file_1')

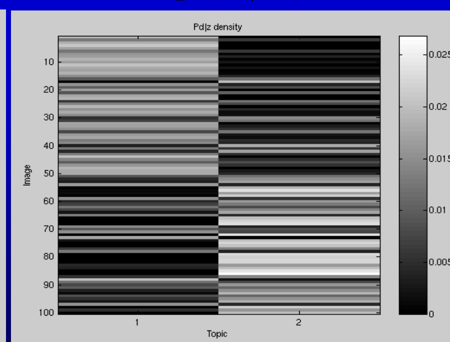
Codeword distributions per theme (topic)

$p(w \mid z)$

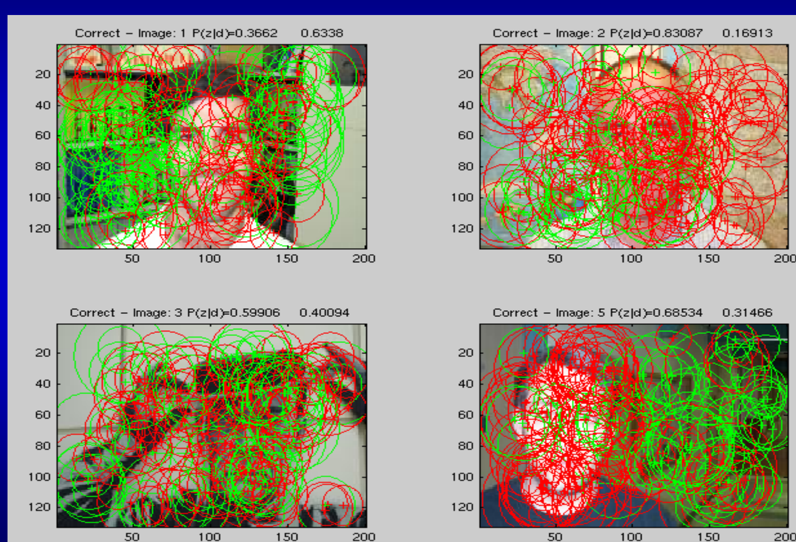
Theme distributions per image

$p(z \mid d)$





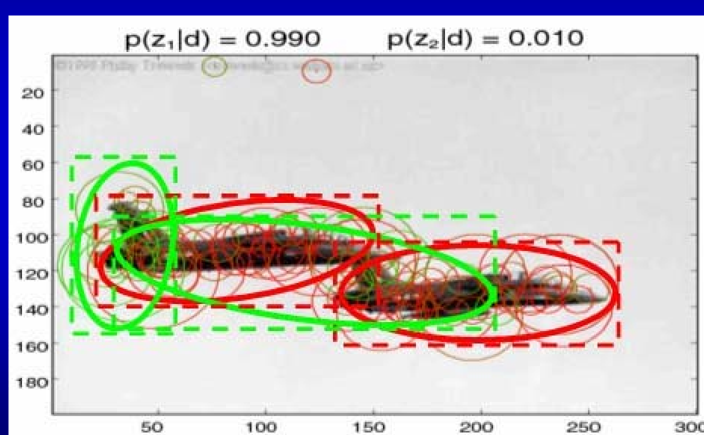
Demo: recognition examples



pLSA example

Fergus, Fei-Fei, Perona, Zisserman, ICCV 2005

pLSA example



Fergus, Fei-Fei, Perona, Zisserman, ICCV 2005

pLSA extensions

- Extended to incorporate position information (Fergus, Fei-Fei, Perona, Zisserman. ICCV 2005)
  - Absolute position pLSA
  - Translation and scale invariant pLSA
- Foreground and background distributions (van de Weijer, Schmid, Verbeek. ICCV 2007)

pLSA extensions

- User interaction to select relevant topics

(Berg, Forsyth. CVPR 2006)

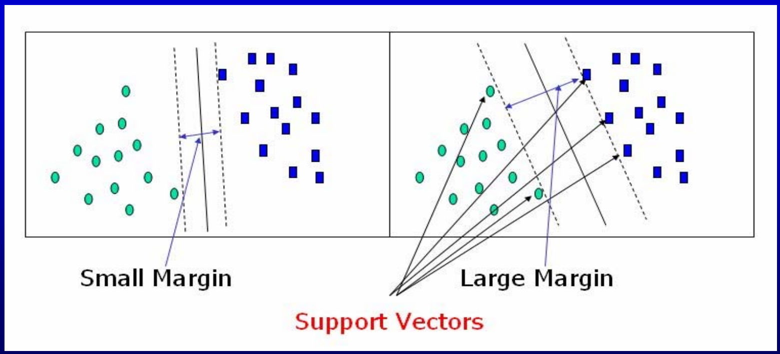
- Optional step to correct erroneous examples
  - Makes the results better when dataset is small
- Requires human in the loop

pLSA shortcomings

- Need to estimate number of topics
- Need to select which topic to use as classifier
- Does not always converge to the desired categories

Support Vector Machines

- Soft margin
- Robust to noise
- Attempt to maximize the margin

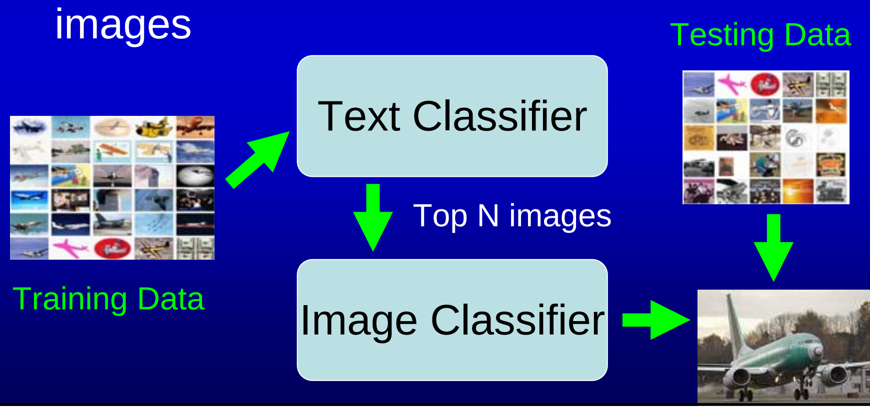


Multiple instance learning

- Robust to noisy training data
- Training data consists of bags of examples
- Positive bags contain at least one positive example
- Negative bags contain no positive examples

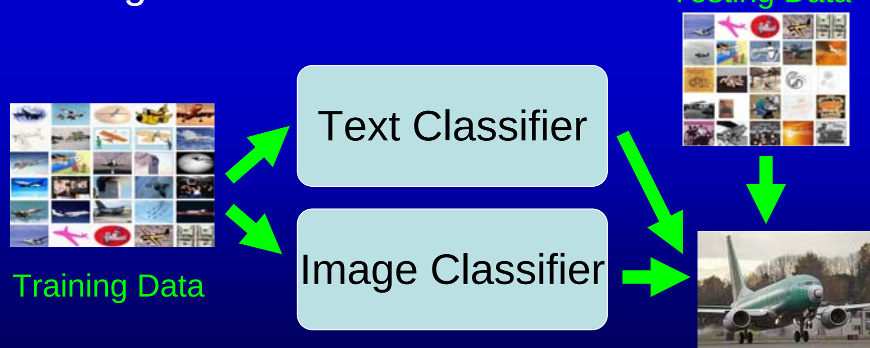
Combining text and image features

- Schroff, Criminisi, Zisserman. ICCV 2007
- Rank images using text features first
- Train image classifier on the top-ranked images

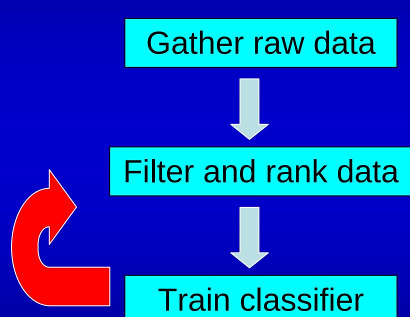


Combining text and image features

- Berg, Forsyth. CVPR 2006
- Voting-based approach
- Weigh score contributions from text and image classifications

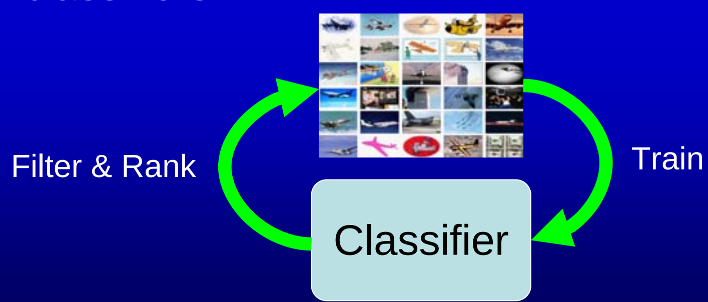


General framework for object recognition



Iterative training

- Use the trained classifier to filter the training data
- Better training data leads to better classifiers



Applications

- Building large datasets of images
- Ranking images from search results
- Building object recognition systems for many categories
- Learning color names
- Location recognition

Roadblocks

- Polysemy
  - Indiscriminative query terms
- Difficult images
  - Abstract images
  - Occlusions, clutter, variable lighting
  - Small portion of the image

Polysemy

Images related to the category

"Airplane"

Polysemy

Category names refer to several concepts

"Tiger"

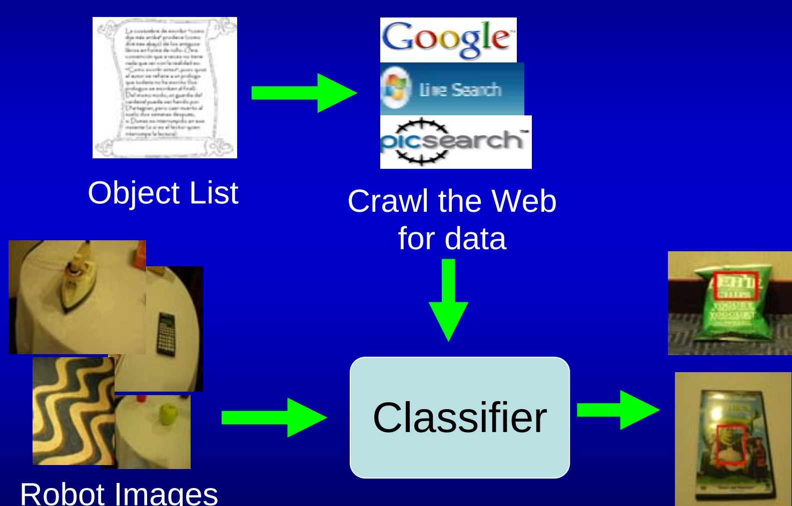
Conclusion

- Gather large amounts of images from Web
- Filter the results using both textual and visual information
- Build classifiers from filtered results
- Optionally reiterate the process
- Provides realistic training and testing data for object recognition
- Still faces many challenging problems

Semantic Robot Vision Challenge

- First contest was held at AAAI 2007
- Robot League
  - UBC LCI Robotics from University of British Columbia
  - Terrapins from University of Maryland
  - KSU Willie from Kansas State University
  - Sunflowers from University of Washington
- Software League
  - UIUC-Princeon
  - KSU Willie from Kansas State University

Semantic Robot Vision Challenge



Semantic Robot Vision Challenge

1. scientific calculator
2. Ritter Sport Marzipan
3. book “Harry Potter and the Deathly Hallows”
4. DVD “Shrek”
5. DVD “Gladiator”
6. CD “Hey Eugene” by Pink Martini
7. fork
8. electric iron
9. banana
10. green apple
11. red bell pepper
12. Lindt Madagascar
13. rolling suitcase
14. red plastic cup
15. Twix candy bar
16. Tide detergent
17. Pepsi bottle
18. yogurt Kettle Chips
19. upright vacuum cleaner

Semantic Robot Vision Challenge

- Relatively small number of images used
  - Specific objects: 3 – 15
  - General objects: 20 - 40
- Commercial images desired
  - Use blacklist to exclude amateur photos
  - Build detector of homogenous, monochromatic background
- Graphic filter
- Rank images base on intra-class similarity and inter-class dissimilarity

Semantic Robot Vision Challenge

Robot League

	Points	Images Returned	Non-zero overlap
University of British Columbia	13	15	7
University of Maryland	6	2	2
Kansas State University	0	3	0

Software League

Princeton-UIUC	5	10	7
Kansas State University	0	2	2

CD "Hey Eugene" by Pink Martini

DVD “Gladiator”

Pepsi bottle

red bell pepper

red plastic cup

Discussion topics

- How to deal with polysemy
- Different ways of combining textual and visual information
- How to deal with difficult images
  - Prune them
  - Better object recognition algorithms
- Better algorithms for building classifiers from noisy data