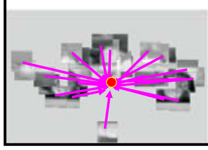
Part-based models & recognition with local features

Kristen Grauman
UT-Austin



Thursday, Nov 13

Upcoming schedule

- Tuesday 11/18
- Thursday 11/20
- Tuesday 11/25
- (Thursday 11/27: Thanksgiving)
- Tuesday 12/2
- Thursday 12/4: Last class: review, wrap-up

Shape

Motion & Tracking

- Saturday 12/13: Final exam

Pset 3 results

Andy Luong

Andy Luong

Andy Luong

Anush Moorthy

Anush Moorthy

Birgi Tamersoy

Birgi Tamersoy

Wei-Cheng Su

Birgi Tamersoy

Chia-Sheng Tsai

Chia-Sheng Tsai

Kristen Nishiguchi

Kristen Nishiguchi

Kristen Nishiguchi

Kristen Nishiguchi

Bricks region

Kristen Nishiguchi

Window region

Jeff Donahue

Jeff Donahue

Matthew deWet

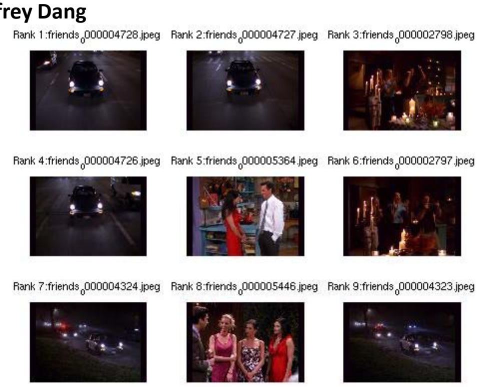
Cabinet region

Matthew deWet

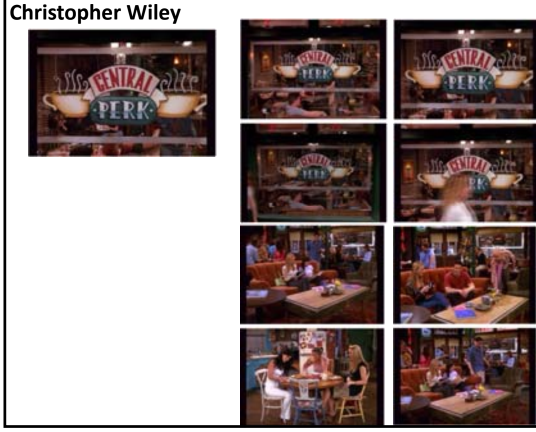
Region Query 2 – The Couch

Matthew deWet

Jeffrey Dang



Christopher Wiley



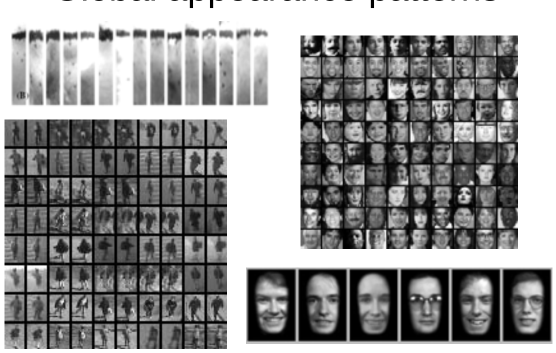
Last time

- Recognizing a window of appearance via classification
  - Nearest neighbors
  - SVMs
    - Applications to gender classification, pedestrian detection

Today

- Limitations of global appearance & sliding windows
- Categorization with local features:
  - Bag-of-words classification
  - Part-based models

Global appearance patterns



Global appearance, windowed detectors: The good things

- **Some classes well-captured by 2d appearance pattern**
- **Simple detection protocol to implement**
- **Good feature choices critical**
- **Past successes for certain classes**

Visual Object Recognition Tutorial

K. Grauman, B. Leibe

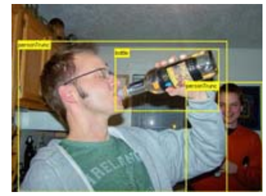
30

Limitations

- High computational complexity
  - For example: 250,000 locations x 30 orientations x 4 scales = 30,000,000 evaluations!
  - With so many windows, false positive rate better be low
  - If training binary detectors independently, means cost increases linearly with number of classes

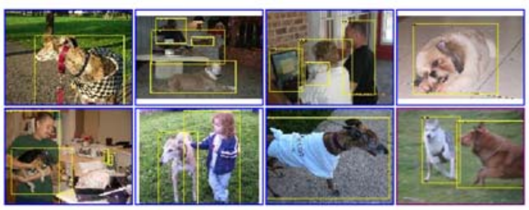
Limitations (continued)

- Not all objects are “box” shaped

Limitations (continued)

- Non-rigid, deformable objects not captured well with representations assuming a fixed 2d structure; or must assume fixed viewpoint
- Objects with less-regular textures not captured well with holistic appearance-based descriptions



Limitations (continued)

- If considering windows in isolation, context is lost


Sliding window


Detector’s view

Figure credit: Derek Hoiem

Limitations (continued)

- In practice, often entails large, cropped training set (expensive)
- Requiring good match to a global appearance description can lead to sensitivity to partial occlusions


Image credit: Adam, Rivlin, & Shimshoni

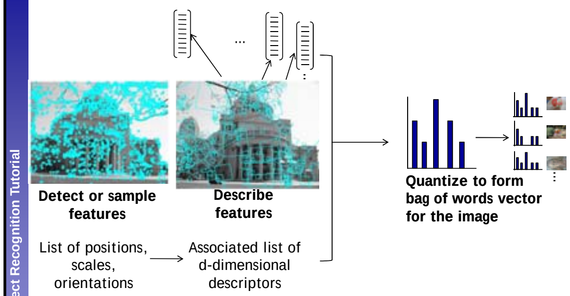
Models based on local features will alleviate some of these limitations…

Today

- Limitations of global appearance & sliding windows
- Categorization with local features:
  - Bag-of-words classification
  - Part-based models

Visual Object Recognition Tutorial

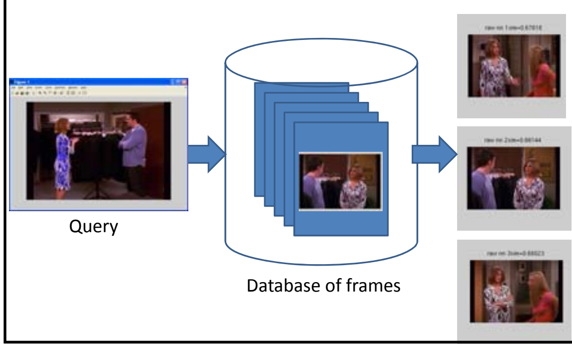
Recall: Local feature extraction



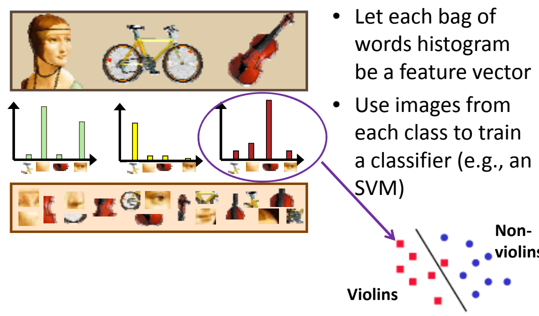
K. Grauman, B. Leibe

Indexing with bags-of-words

Measure similarity to all database items, rank.



Categorization with bags-of-words



- Let each bag of words histogram be a feature vector
- Use images from each class to train a classifier (e.g., an SVM)

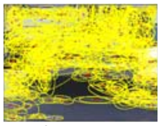
Sampling strategies

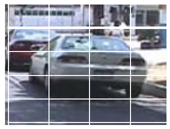
- Reliable local feature matches well-suited for recognition of instances (specific objects, scenes). Even a few (sparse) strong matches can be a good indicator for moderately-sized databases.

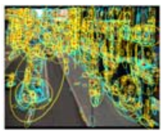
Sampling strategies

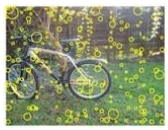
- For category-level recognition, we can't necessarily rely on having such exact feature matches; sparse selection of features may leave more ambiguity.

Sampling strategies


Sparse, at interest points


Dense, uniformly

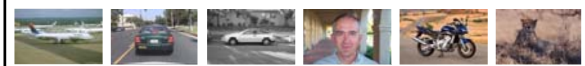

Multiple interest operators


Randomly

- *Some rules of thumb:*
- To find specific, textured objects, sparse sampling from interest points often more reliable.
- Multiple complementary interest operators offer more image coverage.
- For object categorization, dense sampling often offers better coverage.

Image credits: F-F. Li, E. Nowak, J. Sivic

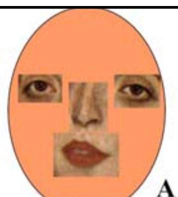
Categorization with bags-of-words

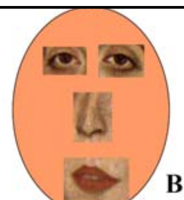


class	bag of features Zhang et al. (2005)	bag of features Willamowski et al. (2004)
airplanes	**98.8**	97.1
cars (rear)	98.3	**98.6**
cars (side)	**95.0**	87.3
faces	**100**	99.3
motorbikes	**98.5**	98.0
spotted cats	**97.0**	—

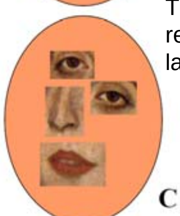
Have been shown to perform well in practice.

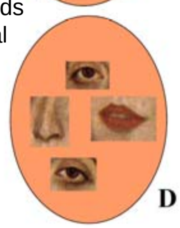
Source: Lana Lazebnik





The bag of words removes spatial layout.

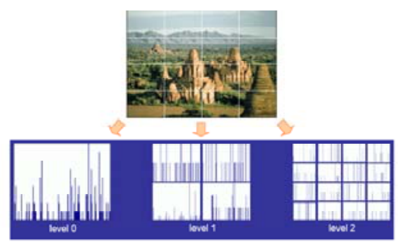




Slide by Bill Freeman, MIT

Introducing some loose spatial information

- A representation "in-between" orderless bags of words and global appearance: a spatial pyramid of bags-of-words.



Lazebnik, Schmid & Ponce, CVPR 2006

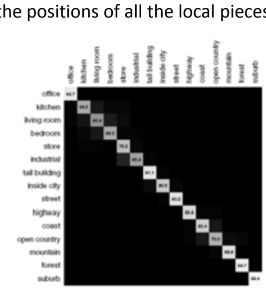
Introducing some loose spatial information

- Can capture **scene** categories well---texture-like patterns but with some variability in the positions of all the local pieces.

Lazebnik, Schmid & Ponce, CVPR 2006

Introducing some loose spatial information

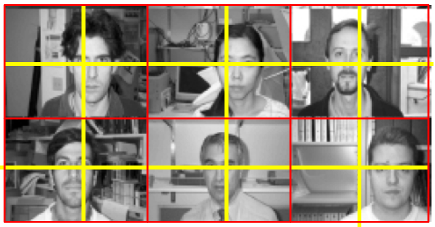
- Can capture **scene** categories well---texture-like patterns but with some variability in the positions of all the local pieces.



Confusion table

Lazebnik, Schmid & Ponce, CVPR 2006

Introducing some loose spatial information



- What will a grid binning of features over the whole image be sensitive to?

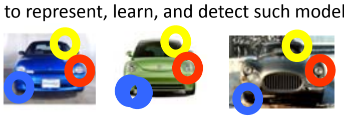
Part-based models

- Represent a category by common parts and their layout

Part-based models: questions

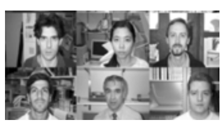
Some categories are well-defined by a collection of parts and their relative positions

- 1) How to represent, learn, and detect such models?



- 2) How can we learn these models in the presence of clutter?

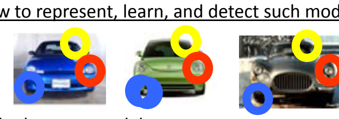
Vs.



Part-based models: questions

Some categories are well-defined by a collection of parts and their relative positions

- 1) How to represent, learn, and detect such models?



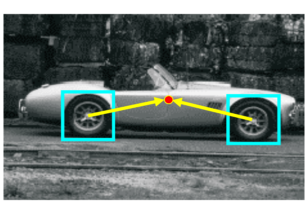
We'll look at two models:

- Generalized Hough with words ("Implicit Shape Model")
- Probabilistic generative model of parts & appearance ("Constellation model")

- 2) How can we learn these models in the presence of clutter?

Implicit shape models

- Visual vocabulary is used to index votes for object position [a visual word = "part"]



training image

visual codeword with displacement vectors

B. Leibe, A. Leonardis, and B. Schiele, Combined Object Categorization and Segmentation with an Implicit Shape Model, ECCV Workshop on Statistical Learning in Computer Vision 2004

Implicit shape models

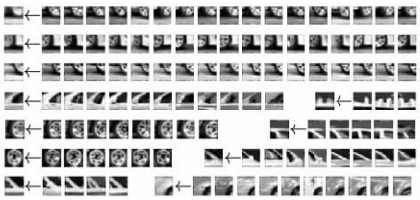
- Visual vocabulary is used to index votes for object position [a visual word = "part"]

test image

B. Leibe, A. Leonardis, and B. Schiele, Combined Object Categorization and Segmentation with an Implicit Shape Model, ECCV Workshop on Statistical Learning in Computer Vision 2004

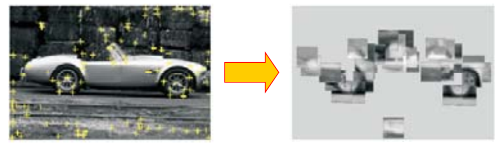
Implicit shape models: Training

1. Build vocabulary of patches around extracted interest points using clustering



Implicit shape models: Training

1. Build vocabulary of patches around extracted interest points using clustering
2. Map the patch around each interest point to closest word



Implicit shape models: Training

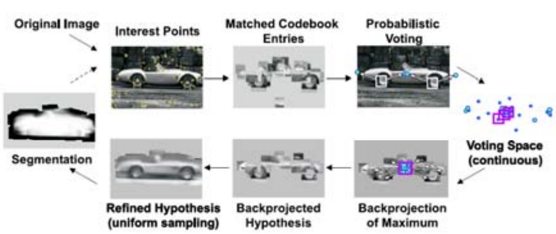
1. Build vocabulary of patches around extracted interest points using clustering
2. Map the patch around each interest point to closest word
3. For each word, store all positions it was found, relative to object center

Implicit shape models: Testing

1. Given new test image, extract patches, match to vocabulary words
2. Cast votes for possible positions of object center
3. Search for maxima in voting space
4. (Extract weighted segmentation mask based on stored masks for the codebook occurrences)

What is the dimension of the Hough space?

Implicit shape models: Testing



Visual Object Recognition Tutorial

Example: Results on Cows

Original image

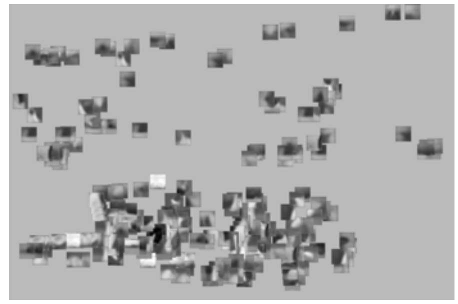
K. Grauman, B. Leibe

60

Example: Results on Cows

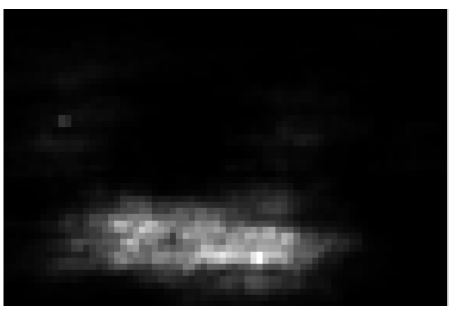
Interest points

Example: Results on Cows



Matched patches

Example: Results on Cows



Prob. Votes s

Example: Results on Cows

1st hypothesis

Example: Results on Cows

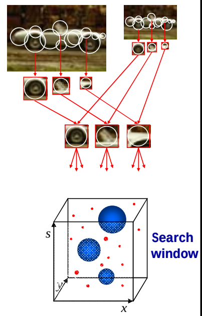
2nd hypothesis

Example: Results on Cows

3rd hypothesis

Scale Invariant Voting

- **Scale-invariant feature selection**
  - **Scale-invariant interest points**
  - **Rescale extracted patches**
  - **Match to constant-size codebook**
- **Generate scale votes**
  - **Scale as 3rd dimension in voting space**



Visual Object Recognition Tutorial

K. Grauman, B. Leibe

Detection Results

- **Qualitative Performance**
  - **Recognizes different kinds of objects**
  - **Robust to clutter, occlusion, noise, low contrast**

Visual Object Recognition Tutorial

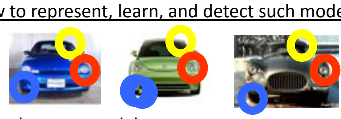
K. Grauman, B. Leibe

68

Part-based models: questions

Some categories are well-defined by a collection of parts and their relative positions

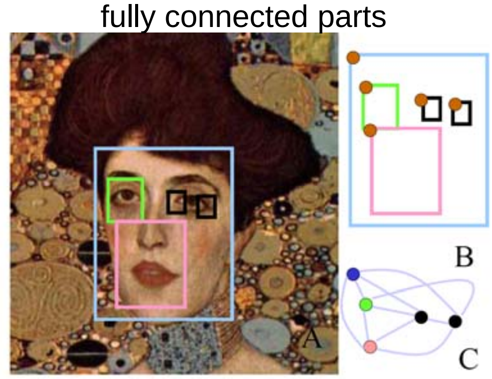
- 1) How to represent, learn, and detect such models?



We'll look at two models:
- Generalized Hough with words ("Implicit Shape Model")
- Probabilistic generative model of parts & appearance ("Constellation model")

- 2) How can we learn these models in the presence of clutter?

Part-based models: constellation of fully connected parts



Slide by Bill Freeman, MIT

Probabilistic constellation model

$$P(image \mid object) = P(appearance, shape \mid object)$$

Part descriptors

Part locations

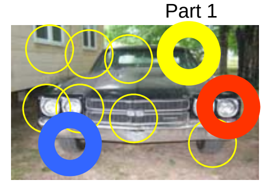
Candidate parts

Source: Lana Lazebnik

Probabilistic constellation model

$$P(image \mid object) = P(appearance, shape \mid object)$$

Part 1



Part 3

Part 2

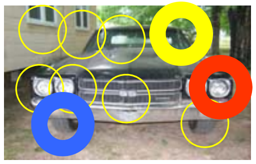
Source: Lana Lazebnik

Probabilistic constellation model

$$P(image \mid object) = P(appearance, shape \mid object)$$
$$= \max_h P(appearance \mid h, object)\, p(shape \mid h, object)\, p(h \mid object)$$

h: assignment of features to parts

Part 1



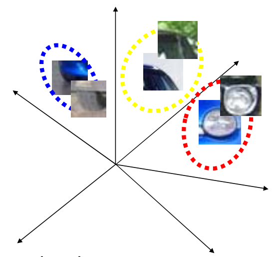
Part 3

Part 2

Source: Lana Lazebnik

Probabilistic constellation model

$$P(image \mid object) = P(appearance, shape \mid object)$$
$$= \max_h \boxed{P(appearance \mid h, object)}\, p(shape \mid h, object)\, p(h \mid object)$$



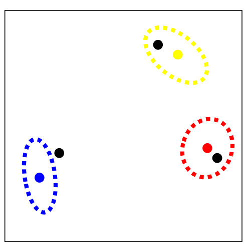
Distribution over patch descriptors

High-dimensional appearance space

Source: Lana Lazebnik

Probabilistic constellation model

$$P(image \mid object) = P(appearance, shape \mid object)$$
$$= \max_h P(appearance \mid h, object)\, \boxed{p(shape \mid h, object)}\, p(h \mid object)$$



Distribution over joint part positions

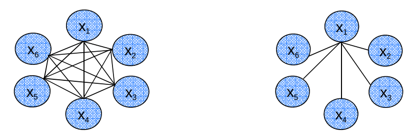
2D image space

Source: Lana Lazebnik

Visual Object Recognition Tutorial

Shape representation in part-based models

Fully connected constellation model	"Star" shape model
e.g. Constellation Model	e.g. ISM
Parts fully connected	Parts mutually independent
Recognition complexity: $O(N^P)$	Recognition complexity: O(NP)
Method: Exhaustive search	Method: Gen. Hough Transform



N image features, P parts in the model

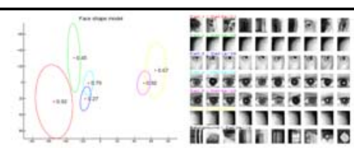
K. Grauman, B. Leibe

Slide credit: Rob Fergus

Example results from constellation model: data from four categories

Slide from Li Fei-Fei http://www.vision.caltech.edu/feifeili/Resume.htm

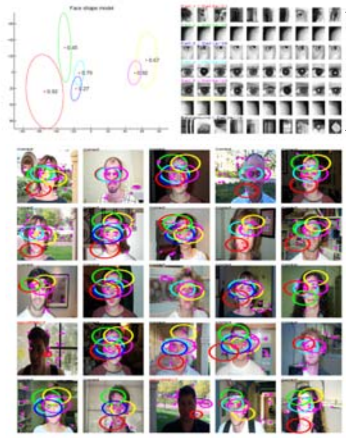
Face model



Appearance: 10 patches closest to mean for each part

Fergus et al. CVPR 2003

Face model



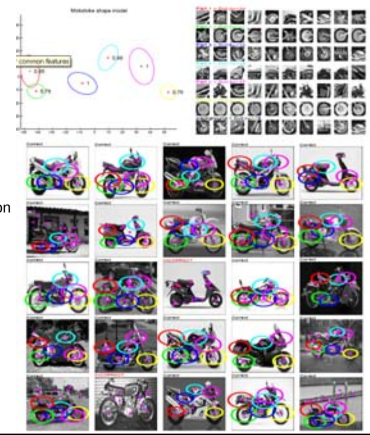
Appearance: 10 patches closest to mean for each part

Recognition results

Test images: size of circles indicates score of hypothesis

Fergus et al. CVPR 2003

Motorbike model



Appearance: 10 patches closest to mean for each part

Recognition results

Fergus et al. CVPR 2003

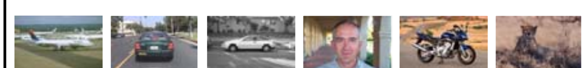
Spotted cat model

Appearance: 10 patches closest to mean for each part

Recognition results

Fergus et al. CVPR 2003

Comparison



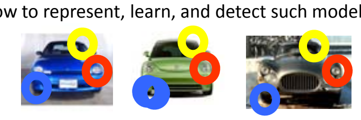
class	bag of features Zhang et al. (2005)	bag of features Willamowski et al. (2004)	Part-based model Fergus et al. (2003)
airplanes	**98.8**	97.1	90.2
cars (rear)	98.3	**98.6**	90.3
cars (side)	**95.0**	87.3	88.5
faces	**100**	99.3	96.4
motorbikes	**98.5**	98.0	92.5
spotted cats	**97.0**	—	90.0

Source: Lana Lazebnik

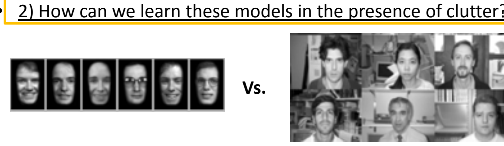
Part-based models: questions

Some categories are well-defined by a collection of parts and their relative positions

- 1) How to represent, learn, and detect such models?



- 2) How can we learn these models in the presence of clutter?



Weber, Welling, Perona., 2000.

Fig. 1. Which objects appear consistently in the left images, but not on the right side? Can a machine learn to recognize instances of the two object classes (*faces* and *cars*) without any further information provided?

Learning part-based models with "weak" supervision

Main idea:

- Use interest operator to detect small highly textured regions (on both fg and bg)
  - If training objects have similar appearance, these regions will often be similar in different training examples
- Cluster patches: large clusters used to select candidate fg parts
- Choose most informative parts while simultaneously estimating model parameters
  - Iteratively try different combinations of a small number of parts and check model performance on validation set to evaluate quality

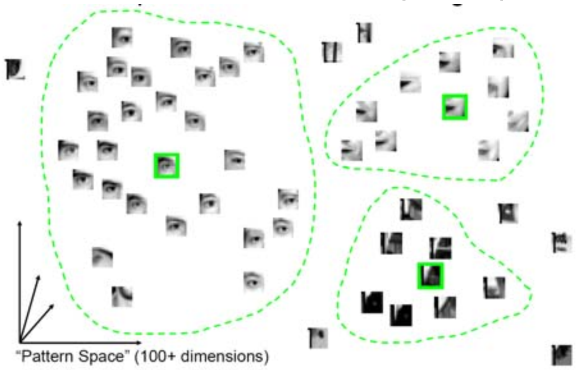
Weber, Welling, Perona, ECCV 2000.

Detect features

- Use a scale invariant detector (like DoG in SIFT detection)

From: Rob Fergus http://www.robots.ox.ac.uk/%7Efergus/

Cluster features in training examples



Slide by Bill Freeman, MIT

Candidate parts

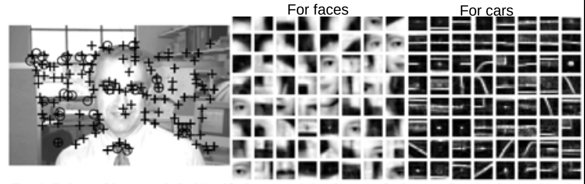


Fig. 3. Points of interest (left) identified on a training image of a human face in cluttered background using Förstner's method. Crosses denote corner-type patterns while circles mark circle-type patterns. A sample of the patterns obtained using k-means clustering of small image patches is shown for faces (center) and cars (right). The car images were high-pass filtered before the part selection process. The total number of patterns selected were 81 for faces and 80 for cars.

At this point, parts appear in both background and foreground of training images.

Weber, Welling, Perona. Unsupervised Learning of Models for Recognition, 2000.

Learning part-based models with "weak" supervision

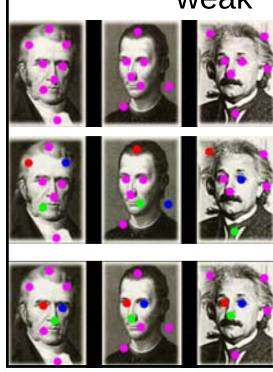
Which of the candidate parts define the class, and in what configuration?

Let's assume:

- We know number of parts that define the model (and can keep it small).
- Object of interest is only consistent thing somewhere in each training image.

Images from Rob Fergus

Learning part-based models with "weak" supervision



Which of the candidate parts define the class, and in what configuration?

Initialize model parameters randomly.

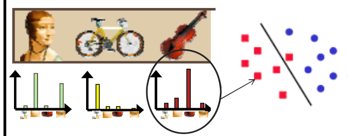
Iterate:

1. Find best assignment in the training images given the current parameters
2. Recompute parameters based on current features

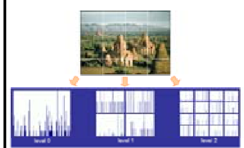
Today

- Limitations of global appearance & sliding windows
- Categorization with local features:
  - Bag-of-words classification
  - Part-based models

Recap (1 of 3)



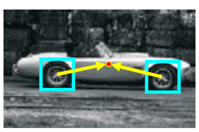
Bag of words a simple way to use local features for recognition (via classification)



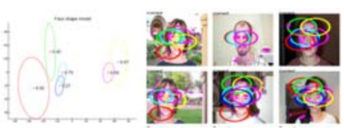
Rather than throw away all spatial information, can introduce a pyramid grid of bags of words within the image.

Recap (2 of 3)

Part-based models summarize a category's local appearance and relative structure:



- Generalized Hough with visual words as parts



- Probabilistic constellation model

Recap (3 of 3)

Learning from cluttered image examples: if we can collect examples with uncorrelated clutter in backgrounds, possible to automatically extract object parts of interest to learn category model.

Next time: shape