Computer Vision - Lecture 16

Part-based Models for Object Categorization

08.01.2015

Bastian Leibe
RWTH Aachen
http://www.vision.rwth-aachen.de

leibe@vision.rwth-aachen.de

Course Outline

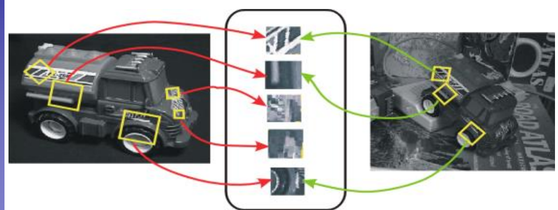
- Image Processing Basics
- Segmentation & Grouping
- Object Recognition
- Object Categorization I
  - Sliding Window based Object Detection
- Local Features & Matching
  - Local Features - Detection and Description
  - Recognition with Local Features
  - Indexing & Visual Vocabularies
- Object Categorization II
  - Bag-of-Words Approaches & Part-based Approaches
- 3D Reconstruction
- Optical Flow

Topics of This Lecture

- Recap: Specific Object Recognition with Local Features
  - Matching & Indexing
  - Geometric Verification
- Part-Based Models for Object Categorization
  - Structure representations
  - Different connectivity structures
- Bag-of-Words Model
  - Use for image classification
- Implicit Shape Model
  - Generalized Hough Transform for object category detection
- Deformable Part-based Model
  - Discriminative part-based detection

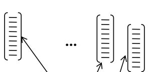
Recap: Recognition with Local Features

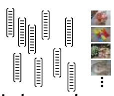
- Image content is transformed into local features that are invariant to translation, rotation, and scale
- Goal: Verify if they belong to a consistent configuration

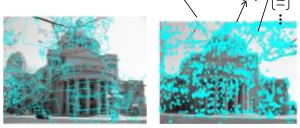


Local Features, e.g. SIFT

Recap: Indexing features





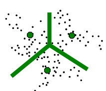


Index each one into pool of descriptors from previously seen images

Detect or sample features

Describe features

or



List of positions, scales, orientations → Associated list of d-dimensional descriptors

Match to quantized descriptors (visual words)

⇒ *Shortlist of possibly matching images + feature correspondences*

Extension: *tf-idf* Weighting

- Term frequency - inverse document frequency
  - Describe frame by frequency of each word within it, downweight words that appear often in the database
  - (Standard weighting for text retrieval)

Number of occurrences of word i in document d

Number of words in document d

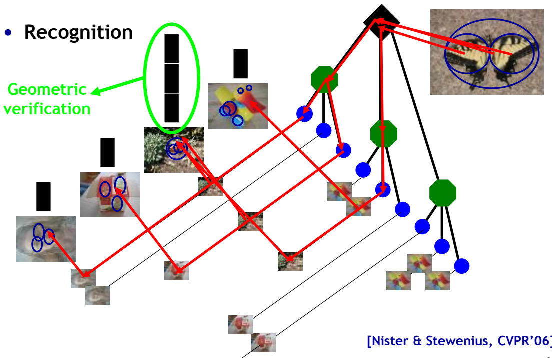
$$t_i = \frac{n_{id}}{n_d} \log \frac{N}{n_i}$$

Total number of documents in database

Number of occurrences of word i in whole database

Recap: Fast Indexing with Vocabulary Trees

- Recognition



[Nister & Stewenius, CVPR'06]

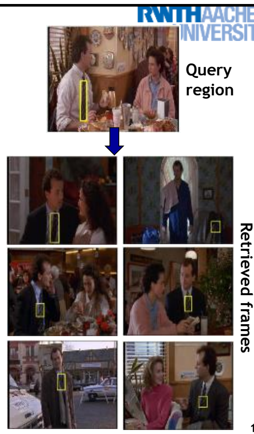
Application for Content Based Img Retrieval

- What if query of interest is a portion of a frame?

[Sivic & Zisserman, ICCV'03]

Video Google System

1. Collect all words within query region
2. Inverted file index to find relevant frames
3. Compare word counts
4. Spatial verification



Sivic & Zisserman, ICCV 2003

- Demo online at : http://www.robots.ox.ac.uk/~vgg/research/vgoogle/index.html

Collecting Words Within a Query Region

- Example: Friends

Query region:
pull out only the SIFT descriptors whose positions are within the polygon

Example Results

Query

More Results

Query

Retrieved shots

Recap: Geometric Verification by Alignment

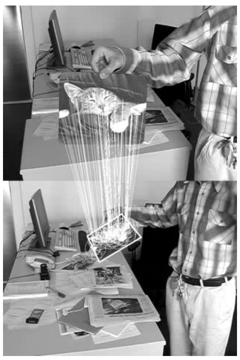
- Assumption
  - Known object, rigid transformation compared to model image
  - ⇒ *If we can find evidence for such a transformation, we have recognized the object.*
- You learned methods for
  - Fitting an ***affine transformation*** from ≥ 3 correspondences
  - Fitting a ***homography*** from ≥ 4 correspondences

Affine: solve a system $At = b$ Homography: solve a system $Ah = 0$

- Correspondences may be noisy and may contain outliers
  - ⇒ Need to use robust methods that can filter out outliers
  - ⇒ Use RANSAC or the Generalized Hough Transform

Applications: Aachen Tourist Guide

Applications: Fast Image Registration



Applications: Mobile Augmented Reality

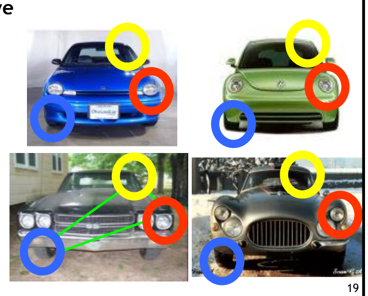
D. Wagner, G. Reitmayr, A. Mulloni, T. Drummond, D. Schmalstieg, Pose Tracking from Natural Features on Mobile Phones. In *ISMAR 2008*.

Topics of This Lecture

- Recap: Specific Object Recognition with Local Features
- Part-Based Models for Object Categorization
  - Structure representations
  - Different connectivity structures
- Bag-of-Words Model
  - Use for image classification
- Implicit Shape Model
  - Generalized Hough Transform for object category detection
- Deformable Part-based Model
  - Discriminative part-based detection

Recognition of Object Categories

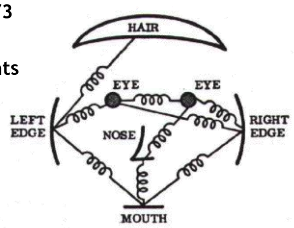
- We no longer have exact correspondences…
- On a local level, we can still detect similar parts.
- Represent objects by their parts
  - ⇒ Bag-of-features
- How can we improve on this?
  - Encode structure



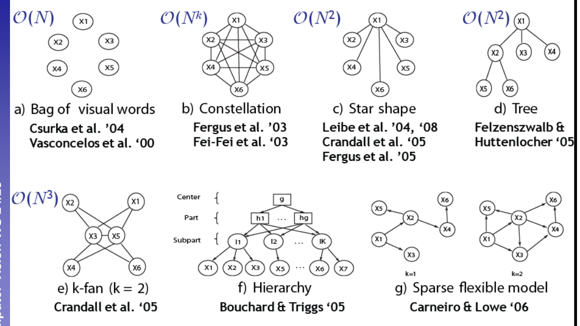
Slide credit: Rob Fergus

Part-Based Models

- Fischler & Elschlager 1973
- Model has two components
  - parts
    (2D image fragments)
  - structure
    (configuration of parts)



Different Connectivity Structures



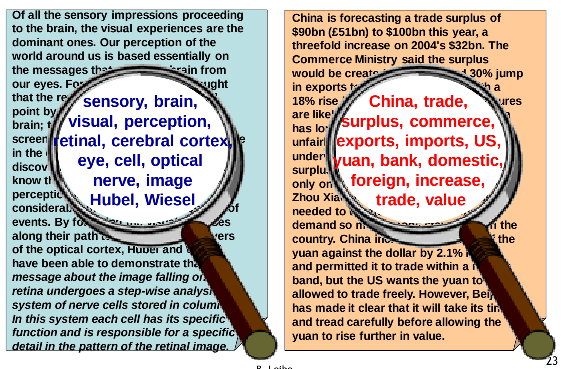
Slide adapted from Rob Fergus

Image from [Carneiro & Lowe, ECCV'06]

Topics of This Lecture

- Recap: Specific Object Recognition with Local Features
- Part-Based Models for Object Categorization
  - Structure representations
  - Different connectivity structures
- **Bag-of-Words Model**
  - Use for image classification
- Implicit Shape Model
  - Generalized Hough Transform for object category detection
- Deformable Part-based Model
  - Discriminative part-based detection

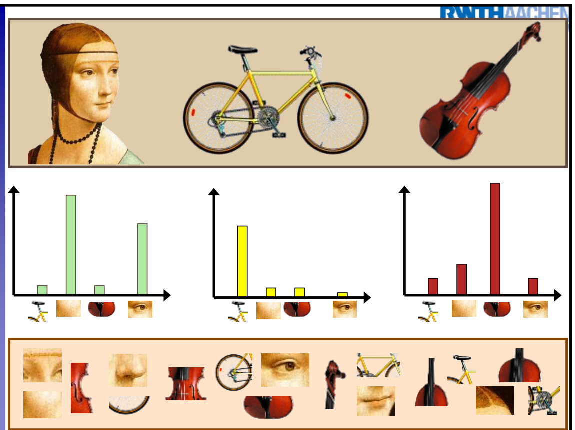
Analogy to Documents



Slide credit: Li Fei-Fei

Object → **Bag of 'words'**

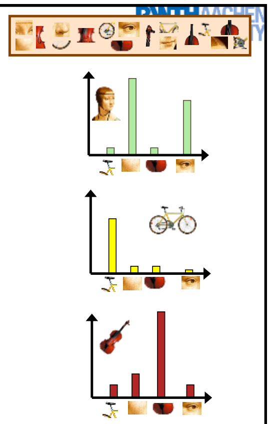
Source: ICCV 2005 short course, Li Fei-Fei



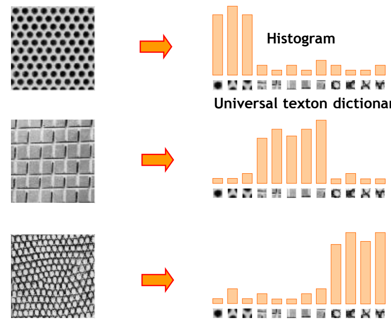
Source: ICCV 2005 short course, Li Fei-Fei

Bags of Visual Words

- Summarize entire image based on its distribution (histogram) of word occurrences.
- Analogous to bag of words representation commonly used for documents.
- Main difference to text retrieval: visual words are not given a priori, but obtained through clustering (e.g., using k-means)



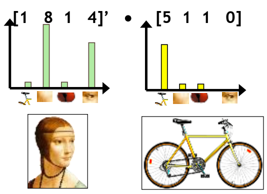
Similarly, Bags-of-Textons for Texture Repr.



Julesz, 1981; Cula & Dana, 2001; Leung & Malik 2001; Mori, Belongie & Malik, 2001; Schmid 2001; Varma & Zisserman, 2002, 2003; Lazebnik, Schmid & Ponce, 2003

Comparing Bags of Words

- We build up histograms of word activations, so any histogram comparison measure can be used here.
- E.g. we can rank frames by normalized scalar product between their (possibly weighted) occurrence counts
  - *Nearest neighbor* search for similar images.

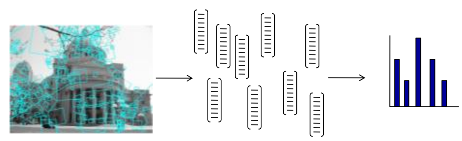


$$sim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}}$$

$\vec{d}_j$ $\vec{q}$

Learning/Recognition with BoW Histograms

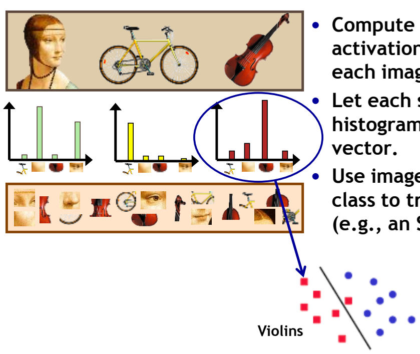
- Bag of words representation makes it possible to describe the unordered point set with a single vector (of fixed dimension across image examples)



- Provides easy way to use distribution of feature types with various learning algorithms requiring vector input.

Recap: Categorization with Bags-of-Words

- Compute the word activation histogram for each image.
- Let each such BoW histogram be a feature vector.
- Use images from each class to train a classifier (e.g., an SVM).



BoW for Object Categorization

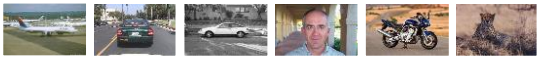
{face, flowers, building}

- Works pretty well for image-level classification

Csurka et al. (2004), Willamowski et al. (2005), Grauman & Darrell (2005), Sivic et al. (2003, 2005)

BoW for Object Categorization

Caltech6 dataset

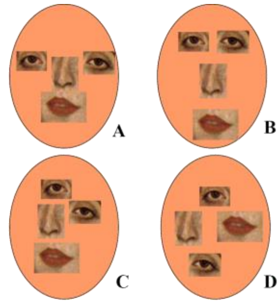


class	bag of features Zhang et al. (2005)	bag of features Willamowski et al. (2004)	Parts-and-shape model Fergus et al. (2003)
airplanes	**98.8**	97.1	90.2
cars (rear)	98.3	**98.6**	90.3
cars (side)	**95.0**	87.3	88.5
faces	**100**	99.3	96.4
motorbikes	**98.5**	98.0	92.5
spotted cats	**97.0**	—	90.0

- Good performance for pure classification (object present/absent)
  - Better than more elaborate part-based models with spatial constraints…
  - What could be possible reasons why?

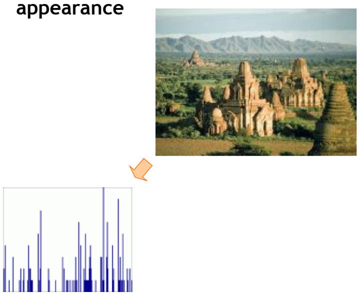
Limitations of BoW Representations

- The bag of words removes spatial layout.
- This is both a strength and a weakness.
- *Why a strength?*
- *Why a weakness?*



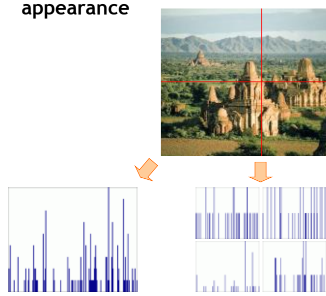
Spatial Pyramid Representation

- Representation in-between orderless BoW and global appearance



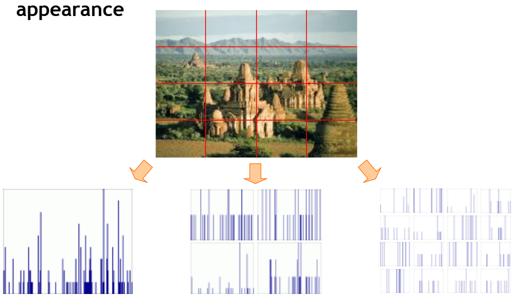
Spatial Pyramid Representation

- Representation in-between orderless BoW and global appearance



Spatial Pyramid Representation

- Representation in-between orderless BoW and global appearance



Summary: Bag-of-Words

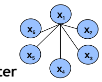
- Pros:
  - Flexible to geometry / deformations / viewpoint
  - Compact summary of image content
  - Provides vector representation for sets
  - Empirically good recognition results in practice
- Cons:
  - Basic model ignores geometry - must verify afterwards, or encode via features.
  - Background and foreground mixed when bag covers whole image
  - Interest points or sampling: no guarantee to capture object-level parts.
  - Optimal vocabulary formation remains unclear.

Topics of This Lecture

- Recap: Specific Object Recognition with Local Features
- Part-Based Models for Object Categorization
  - Structure representations
  - Different connectivity structures
- Bag-of-Words Model
  - Use for image classification
- Implicit Shape Model
  - Generalized Hough Transform for object category detection
- Deformable Part-based Model
  - Discriminative part-based detection

Implicit Shape Model (ISM)

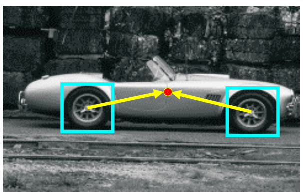
- Basic ideas
  - Learn an appearance codebook
  - Learn a star-topology structural model
    - Features are considered independent given obj. center



- Algorithm: probabilistic Gen. Hough Transform
  - Exact correspondences → Prob. match to object part
  - NN matching → Soft matching
  - Feature location on obj. → Part location distribution
  - Uniform votes → Probabilistic vote weighting
  - Quantized Hough array → Continuous Hough space

Implicit Shape Model: Basic Idea

- 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, Robust Object Detection with Interleaved Categorization and Segmentation, International Journal of Computer Vision, Vol. 77(1-3), 2008.

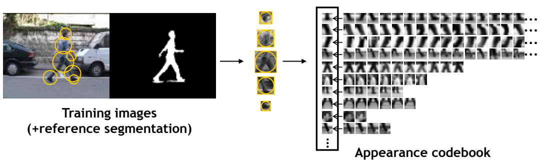
Implicit Shape Model: Basic Idea

- Objects are detected as consistent configurations of the observed parts (visual words).

Test image

B. Leibe, A. Leonardis, and B. Schiele, Robust Object Detection with Interleaved Categorization and Segmentation, International Journal of Computer Vision, Vol. 77(1-3), 2008.

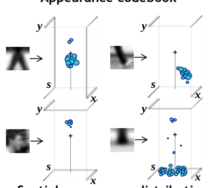
Implicit Shape Model - Representation



Training images (+reference segmentation)

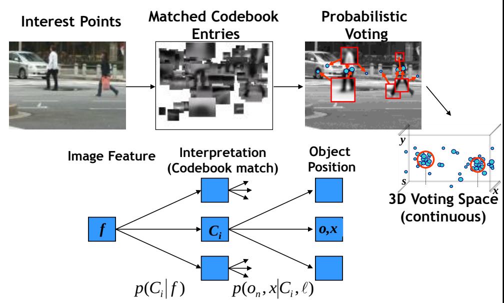
Appearance codebook

- Learn appearance codebook
  - Extract local features at interest points
  - Agglomerative clustering ⇒ codebook
- Learn spatial distributions
  - Match codebook to training images
  - Record matching positions on object



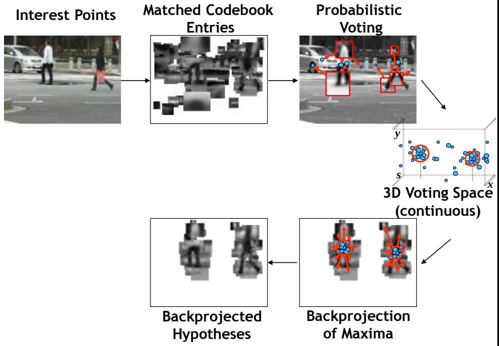
Spatial occurrence distributions

Implicit Shape Model - Recognition



Probabilistic vote weighting

Implicit Shape Model - Recognition



[Leibe, Leonardis, Schiele, SLCV'04; IJCV'08]

Example: Results on Cows

Original image

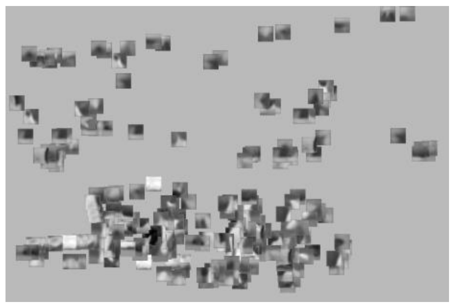
B. Leibe

Example: Results on Cows

Interest points

B. Leibe

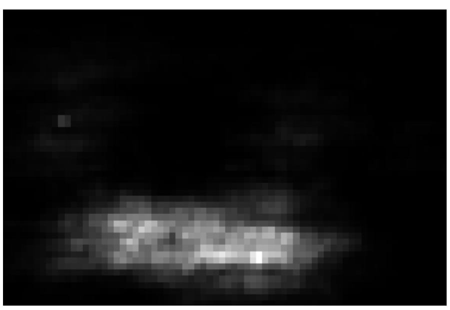
Example: Results on Cows



Matched patches

B. Leibe

Example: Results on Cows



Prob. Votes

B. Leibe

Example: Results on Cows

1st hypothesis

K. Grauman, B. Leibe

Example: Results on Cows

2[nd] hypothesis

Example: Results on Cows

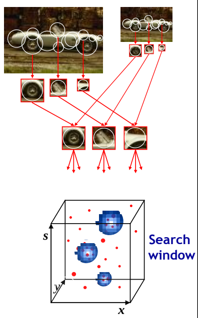
3[rd] hypothesis

Scale Invariant Voting

- Scale-invariant feature selection
  - Scale-invariant interest regions
  - Extract scale-invariant descriptors
  - Match to appearance codebook
- Generate scale votes
  - Scale as 3[rd] dimension in voting space

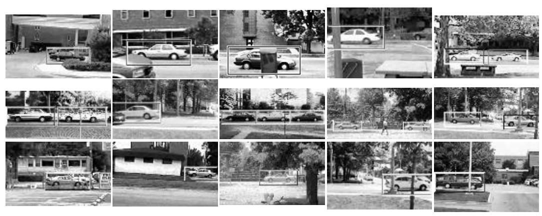
$$\begin{aligned} x_{vote} &= x_{img} - x_{occ}(s_{img}/s_{occ}) \\ y_{vote} &= y_{img} - y_{occ}(s_{img}/s_{occ}) \\ s_{vote} &= (s_{img}/s_{occ}). \end{aligned}$$

  - Search for maxima in 3D voting space

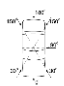


Detection Results

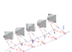
- Qualitative Performance
  - Recognizes different kinds of objects
  - Robust to clutter, occlusion, noise, low contrast



Detections Using Ground Plane Constraints



Battery of 5 ISM detectors for different car views

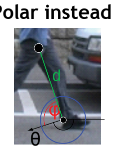


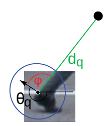
left camera 1175 frames

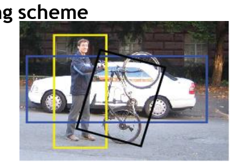
[Leibe, Cornelis, Cornelis, Van Gool, CVPR'07]

Extension: Rotation-Invariant Detection

- Polar instead of Cartesian voting scheme







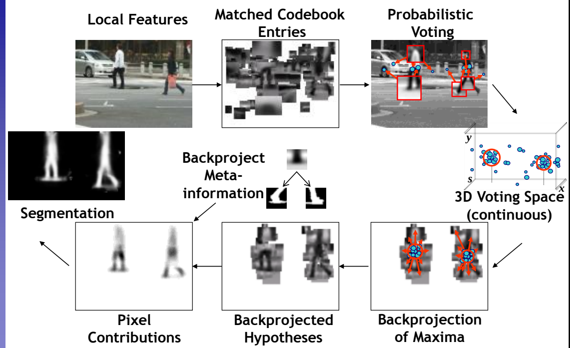
- Benefits:
  - Recognize objects under image-plane rotations
  - Possibility to share parts between articulations.
- Caveats:
  - Rotation invariance should only be used when it's really needed. (Also increases false positive detections)

[Mikolajczyk, Leibe, Schiele, CVPR'06]

Sometimes, Rotation Invariance Is Needed...

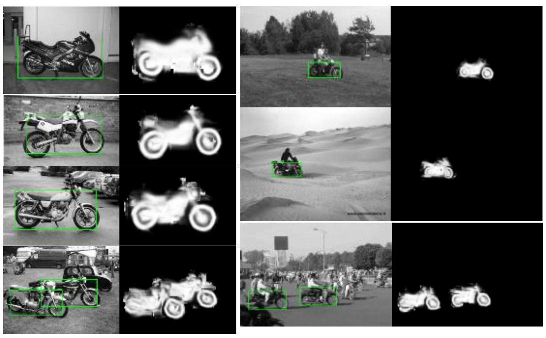
Figure from [Mikolajczyk et al., CVPR'06]

Implicit Shape Model - Segmentation



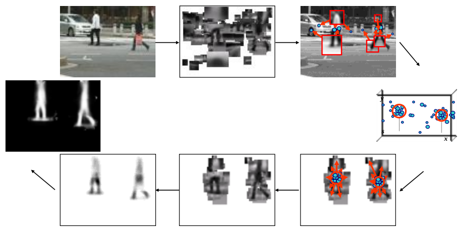
[Leibe, Leonardis, Schiele, DAGM'04; IJCV'08]

Example Results: Motorbikes



[Leibe, Leonardis, Schiele, SLCV'04; IJCV'08]

You Can Try It At Home...

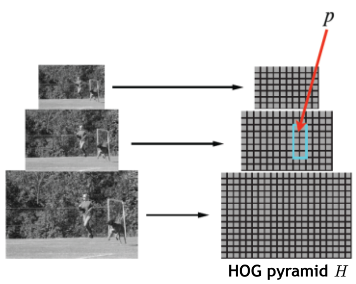


- Linux source code & binaries available
  - Including datasets & several pre-trained detectors
  - http://www.vision.rwth-aachen.de/software

Topics of This Lecture

- Recap: Specific Object Recognition with Local Features
- Part-Based Models for Object Categorization
  - Structure representations
  - Different connectivity structures
- Bag-of-Words Model
  - Use for image classification
- Implicit Shape Model
  - Generalized Hough Transform for object category detection
- **Deformable Part-based Model**
  - Discriminative part-based detection

Starting Point: HOG Sliding-Window Detector

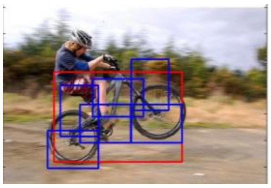


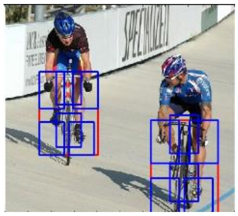
Score of F at position p is $F \cdot \phi(p, H)$

$\phi(p, H)$ = concatenation of HOG features from window specified by p.

- Array of weights for features in window of HOG pyramid
- Score is dot product of filter and vector

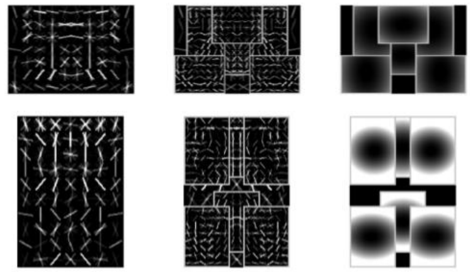
Deformable Part-based Models





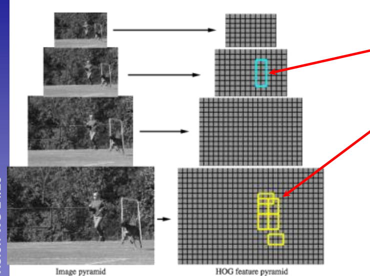
- Mixture of deformable part models (pictorial structures)
- Each component has global template + deformable parts
- Fully trained from bounding boxes alone

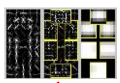
2-Component Bicycle Model



Root filters coarse resolution | Part filters finer resolution | Deformation models

Object Hypothesis





Score of filter: dot product of filter with HOG features underneath it

Score of object hypothesis is sum of filter scores minus deformation costs

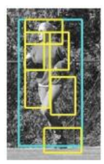
- Multiscale model captures features at two resolutions

Score of a Hypothesis

"data term" — "spatial prior"

$$\text{score}(p_0,\ldots,p_n) = \sum_{i=0}^{n} F_i \cdot \phi(H,p_i) - \sum_{i=1}^{n} d_i \cdot (dx_i^2, dy_i^2)$$

filters — displacements — deformation parameters



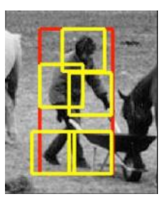
$$\text{score}(z) = \beta \cdot \Psi(H,z)$$

concatenation filters and deformation parameters

concatenation of HOG features and part displacement features

Recognition Model

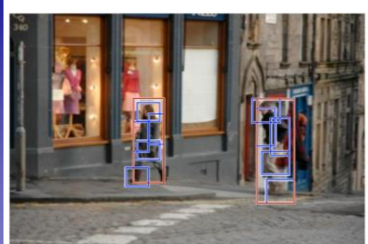
$$f_w(x) = w \cdot \Phi(x)$$

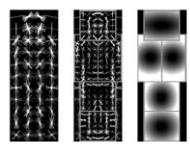


$$f_w(x) = \max_z w \cdot \Phi(x,z)$$

- z : vector of part offsets
- $\Phi(x,z)$: vector of HOG features (from root filter & appropriate part sub-windows) and part offsets

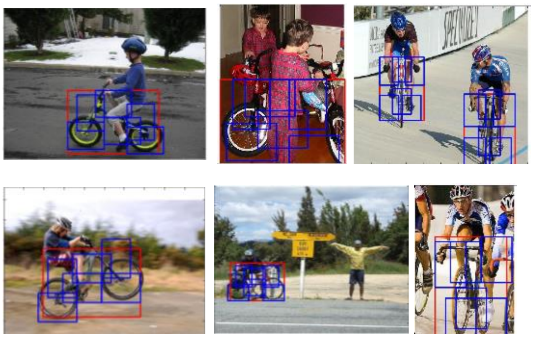
Results: Persons





- Results (after non-maximum suppression)
  - ~1s to search all scales

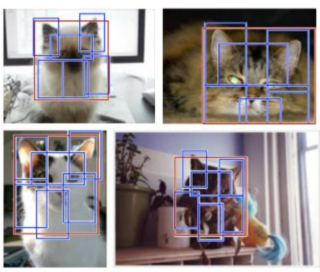
Results: Bicycles



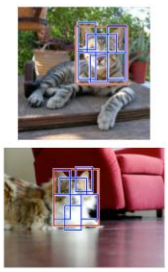
False Positives

- Bicycles

Results: Cats



High-scoring true positives



High-scoring false positives
(not enough overlap)

You Can Try It At Home…

- Deformable part-based models have been very successful at several recent evaluations.

⇒ Currently, state-of-the-art approach in object detection

- Source code and models trained on PASCAL 2006, 2007, and 2008 data are available here:
  http://www.cs.uchicago.edu/~pff/latent

References and Further Reading

- Details about the ISM approach can be found in
  - *B. Leibe, A. Leonardis, and B. Schiele,* Robust Object Detection with Interleaved Categorization and Segmentation, International Journal of Computer Vision, Vol. 77(1-3), 2008.
- Details about the DPMs can be found in
  - *P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan,* Object Detection with Discriminatively Trained Part Based Models, IEEE Trans. PAMI, Vol. 32(9), 2010.
- Try the ISM Linux binaries
  - http://www.vision.ee.ethz.ch/bleibe/code
- Try the Deformable Part-based Models
  - http://www.cs.uchicago.edu/~pff/latent