

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

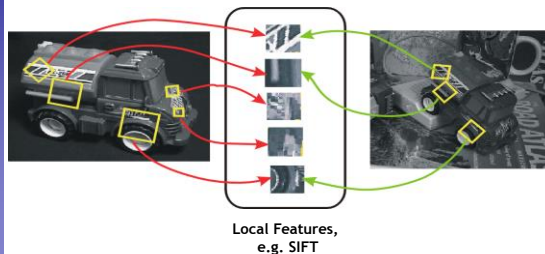
Topics of This Lecture

- 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

B. Leibe

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

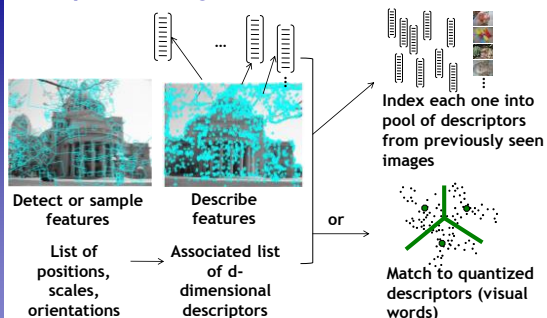


Local Features, e.g. SIFT

B. Leibe

Slide credit: David Lowe

Recap: Indexing features



⇒ Shortlist of possibly matching images + feature correspondences

B. Leibe

Slide credit: Kristen Grauman

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)

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

Number of occurrences of word i in document d → n_{id}

Number of words in document d → n_d

Total number of documents in database → N

Number of occurrences of word i in whole database → n_i

B. Leibe

Slide credit: Kristen Grauman

Computer Vision WS 14/15

Recap: Fast Indexing with Vocabulary Trees

RWTH AACHEN UNIVERSITY

- Recognition

Geometric verification

[Nister & Stewenius, CVPR'06]

Slide credit: David Nister

B. Leibe

8

Computer Vision WS 14/15

Application for Content Based Img Retrieval

RWTH AACHEN UNIVERSITY

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

Visually defined query

"Groundhog Day" [Rammis, 1993]

"Find this clock" →

"Find this place" →

→

Slide credit: Andrew Zisserman

B. Leibe

[Sivic & Zisserman, ICCV'03]

9

Computer Vision WS 14/15

Video Google System

RWTH AACHEN UNIVERSITY

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/~vgs/research/vgoogle/index.html>

Query region

Retrieved frames

Slide credit: Kristen Grauman

B. Leibe

10

Computer Vision WS 14/15

Collecting Words Within a Query Region

RWTH AACHEN UNIVERSITY

- Example: Friends

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

Slide credit: Kristen Grauman

B. Leibe

11

Computer Vision WS 14/15

Example Results

RWTH AACHEN UNIVERSITY

Query

raw nn 1 sim=0.56697 raw nn 2 sim=0.56163 raw nn 5 sim=0.54917

Slide credit: Kristen Grauman

B. Leibe

12

Computer Vision WS 14/15

More Results

RWTH AACHEN UNIVERSITY

Query

Retrieved shots

Slide credit: Kristen Grauman

B. Leibe

13

RWTH AACHEN UNIVERSITY

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 Homography: solve a system

$$At = b \qquad Ah = 0$$

- 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**

B. Leibe

Computer Vision WS 14/15

14

RWTH AACHEN UNIVERSITY

Applications: Aachen Tourist Guide



B. Leibe

Computer Vision WS 14/15

15

RWTH AACHEN UNIVERSITY

Applications: Fast Image Registration



B. Leibe

Computer Vision WS 14/15

16

RWTH AACHEN UNIVERSITY

Applications: Mobile Augmented Reality

Mobile Phone Augmented Reality

at
30 Frames per Second
using
Natural Feature Tracking

(all processing and rendering done in software)

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

B. Leibe

Computer Vision WS 14/15

17

RWTH AACHEN UNIVERSITY

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

B. Leibe

Computer Vision WS 14/15

18

RWTH AACHEN UNIVERSITY

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

Computer Vision WS 14/15

19

RWTH AACHEN UNIVERSITY

Part-Based Models

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

B. Leibe 20

RWTH AACHEN UNIVERSITY

Different Connectivity Structures

$O(N)$

a) Bag of visual words
Csurka et al. '04
Vasconcelos et al. '00

$O(N^2)$

b) Constellation
Fergus et al. '03
Fei-Fei et al. '03

$O(N^2)$

c) Star shape
Leibe et al. '04, '08
Crandall et al. '05
Fergus et al. '05

$O(N^2)$

d) Tree
Felzenszwalb & Huttenlocher '05

$O(N^3)$

e) k-fan (k = 2)
Crandall et al. '05

f) Hierarchy
Bouchard & Triggs '05

g) Sparse flexible model
Carneiro & Lowe '06

Slide adapted from Rob Fergus B. Leibe Image from [Carneiro & Lowe, ECCV'06]

RWTH AACHEN UNIVERSITY

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

B. Leibe 22

RWTH AACHEN UNIVERSITY

Analogy to Documents

Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages that reach our brain from our eyes. For this reason, we might think that the retina is the point by which the brain receives its information. In fact, the retina is only a screen on which the light falls. In the discovery of the structure of the retina, we know that the visual system does not know the perceptual events. By following the path of the optic nerves along their path to the optic chiasm, Hubel and Wiesel have been able to demonstrate that the message about the image falling on the retina undergoes a step-wise analysis. In this system each cell has its specific function and is responsible for a specific detail in the pattern of the retinal image.

China is forecasting a trade surplus of \$90bn (£51bn) to \$100bn this year, a threefold increase on 2004's \$32bn. The Commerce Ministry said the surplus would be created by a 30% jump in exports to \$180bn, a 18% rise in imports to \$90bn. China has long had a trade surplus, but a 30% jump in exports under Zhou Xiaochuan's leadership demanded so much attention from the country. China increased the yuan against the dollar by 2.1% and permitted it to trade within a narrow band, but the US wants the yuan to be allowed to trade freely. However, Beijing has made it clear that it will take its time and tread carefully before allowing the yuan to rise further in value.

Slide credit: Li Fei-Fei B. Leibe 23

RWTH AACHEN UNIVERSITY

Object

→

Bag of 'words'

Source: ICCV 2005 short course, Li Fei-Fei

RWTH AACHEN UNIVERSITY

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)

Computer Vision WS 14/15

Slide credit: Kristen Grauman

B. Leibe

Image credit: Li Fei-Fei

26

Similarly, Bags-of-Textons for Texture Repr.

Computer Vision WS 14/15

Slide credit: Svetlana Lazebnik

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

27

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.

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

Computer Vision WS 14/15

Slide credit: Kristen Grauman

B. Leibe

28

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.

Computer Vision WS 14/15

Slide credit: Kristen Grauman

B. Leibe

29

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).

Computer Vision WS 14/15

Slide adapted from Kristen Grauman

B. Leibe

30

BoW for Object Categorization

- Works pretty well for image-level classification

Computer Vision WS 14/15

Slide credit: Svetlana Lazebnik

B. Leibe


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

31

RWTH AACHEN UNIVERSITY

BoW for Object Categorization

Caltech6 dataset



class	bag of features		Parts-and-shape model
	Zhang et al. (2005)	Willamowski et al. (2004)	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

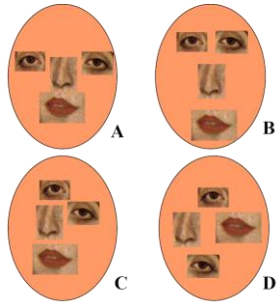
- Good performance for pure classification (object present/absent)
 - Better than more elaborate part-based models with spatial constraints...
 - What could be possible reasons why?

Computer Vision WS 14/15 32
Slide credit: Svetlana Lazebnik B. Leibe

RWTH AACHEN UNIVERSITY

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?

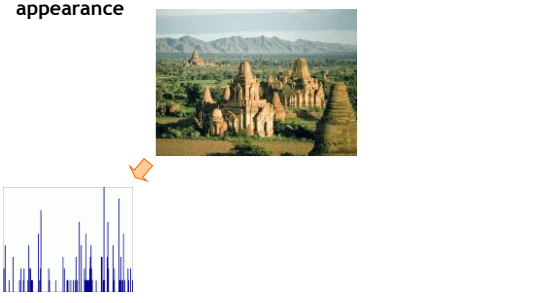


Computer Vision WS 14/15 33
Slide adapted from Bill Freeman B. Leibe

RWTH AACHEN UNIVERSITY

Spatial Pyramid Representation

- Representation in-between orderless BoW and global appearance

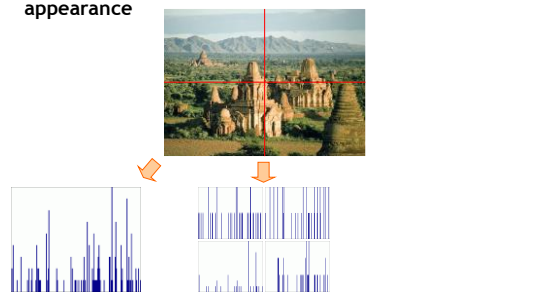


Computer Vision WS 14/15 35
Slide credit: Svetlana Lazebnik B. Leibe [Lazebnik, Schmid & Ponce, CVPR'06]

RWTH AACHEN UNIVERSITY

Spatial Pyramid Representation

- Representation in-between orderless BoW and global appearance

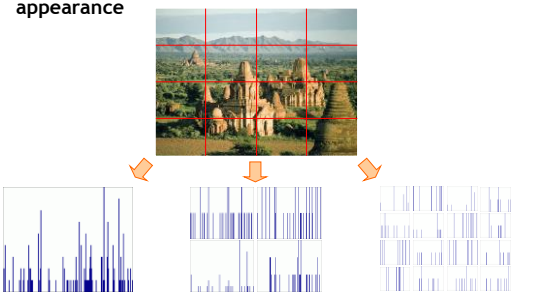


Computer Vision WS 14/15 36
Slide credit: Svetlana Lazebnik B. Leibe [Lazebnik, Schmid & Ponce, CVPR'06]

RWTH AACHEN UNIVERSITY

Spatial Pyramid Representation

- Representation in-between orderless BoW and global appearance



Computer Vision WS 14/15 37
Slide credit: Svetlana Lazebnik B. Leibe [Lazebnik, Schmid & Ponce, CVPR'06]

RWTH AACHEN UNIVERSITY

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.

Computer Vision WS 14/15 38
Slide credit: Kristen Grauman B. Leibe

RWTH AACHEN UNIVERSITY

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

39

RWTH AACHEN UNIVERSITY

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

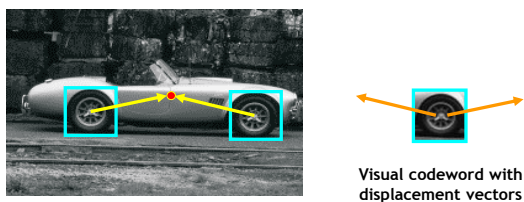


40

RWTH AACHEN UNIVERSITY

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

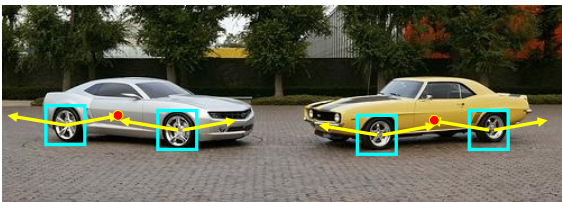
B. Leibe, A. Leonardis, and B. Schiele, [Robust Object Detection with Interleaved Categorization and Segmentation](#), International Journal of Computer Vision, Vol. 77(1-3), 2008.

41

RWTH AACHEN UNIVERSITY

Implicit Shape Model: Basic Idea

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



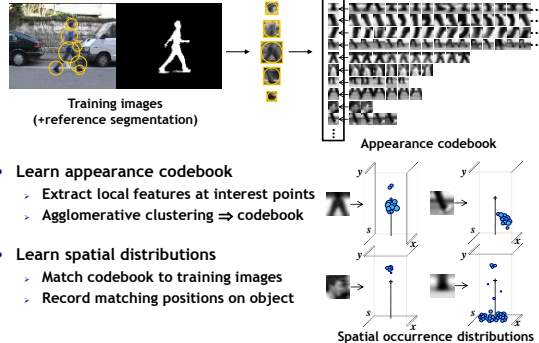
Test image

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.

42

RWTH AACHEN UNIVERSITY

Implicit Shape Model - Representation

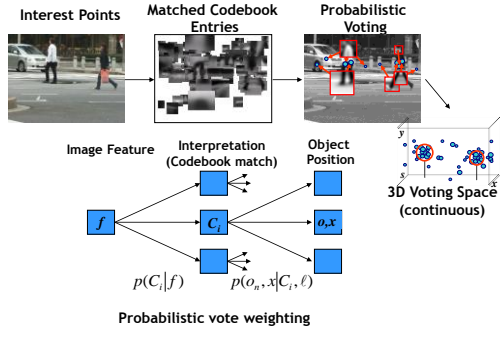


- Learn appearance codebook
 - Extract local features at interest points
 - Agglomerative clustering ⇒ codebook
- Learn spatial distributions
 - Match codebook to training images
 - Record matching positions on object

43

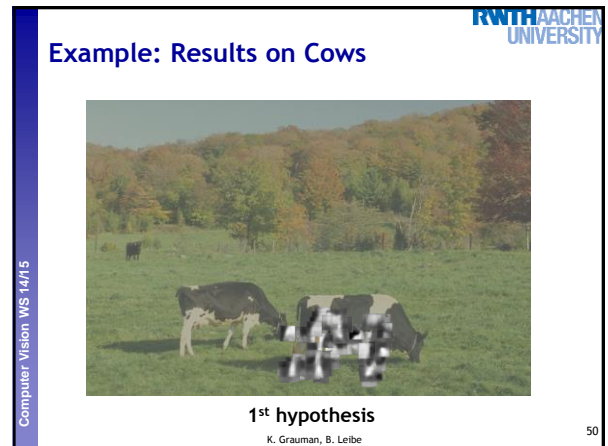
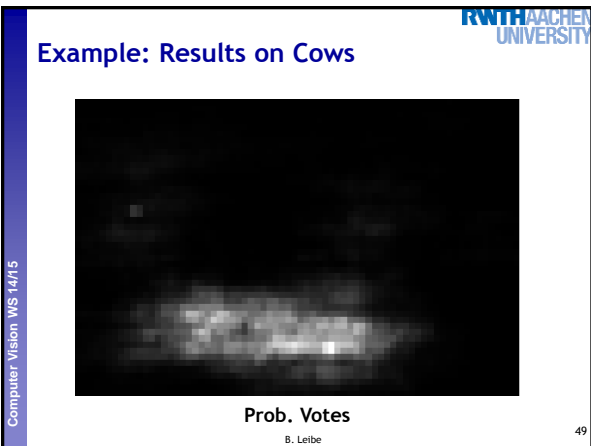
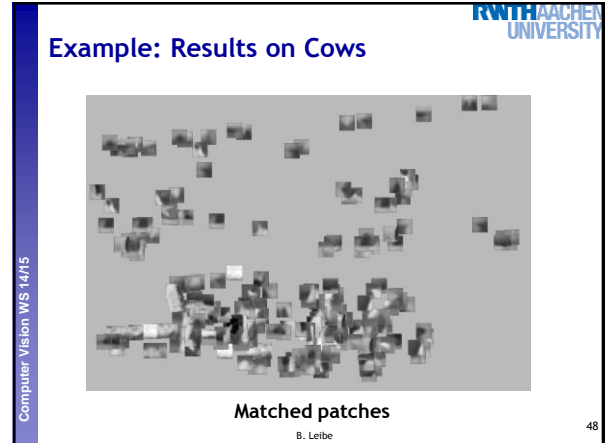
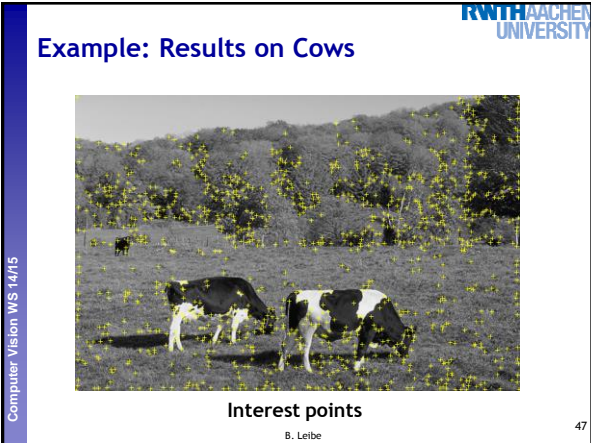
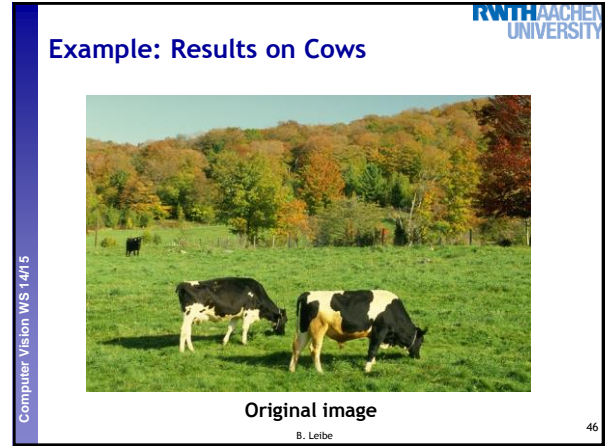
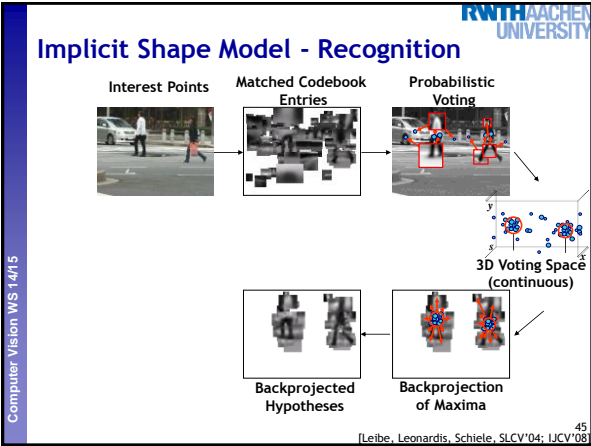
RWTH AACHEN UNIVERSITY

Implicit Shape Model - Recognition



3D Voting Space (continuous)

44



Computer Vision WS 14/15

RWTH AACHEN UNIVERSITY

Example: Results on Cows

2nd hypothesis

B. Leibe

51

Computer Vision WS 14/15

RWTH AACHEN UNIVERSITY

Example: Results on Cows

3rd hypothesis

B. Leibe

52

Computer Vision WS 14/15

RWTH AACHEN UNIVERSITY

Scale Invariant Voting

- Scale-invariant feature selection
 - Scale-invariant interest regions
 - Extract scale-invariant descriptors
 - Match to appearance codebook
- Generate scale votes
 - Scale as 3rd dimension in voting space
$$x_{vote} = x_{img} - x_{occ}(s_{img}/s_{occ})$$

$$y_{vote} = y_{img} - y_{occ}(s_{img}/s_{occ})$$

$$s_{vote} = (s_{img}/s_{occ})$$
 - Search for maxima in 3D voting space

B. Leibe

53

Computer Vision WS 14/15

RWTH AACHEN UNIVERSITY

Detection Results

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

Computer Vision WS 14/15

B. Leibe

56

Computer Vision WS 14/15

RWTH AACHEN UNIVERSITY

Detections Using Ground Plane Constraints

Battery of 5 ISM detectors for different car views

left camera
1175 frames

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

57

Computer Vision WS 14/15

RWTH AACHEN UNIVERSITY

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)

Computer Vision WS 14/15

B. Leibe [Mikolajczyk, Leibe, Schiele, CVPR'06]

58

RWTH AACHEN UNIVERSITY

Sometimes, Rotation Invariance Is Needed...

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

Computer Vision WS 14/15

RWTH AACHEN UNIVERSITY

Implicit Shape Model - Segmentation

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

Computer Vision WS 14/15

RWTH AACHEN UNIVERSITY

Example Results: Motorbikes

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

Computer Vision WS 14/15

RWTH AACHEN UNIVERSITY

You Can Try It At Home...

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

B. Leibe

Computer Vision WS 14/15

RWTH AACHEN UNIVERSITY

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

B. Leibe

Computer Vision WS 14/15

RWTH AACHEN UNIVERSITY

Starting Point: HOG Sliding-Window Detector

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

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

B. Leibe

Computer Vision WS 14/15

RWTH AACHEN UNIVERSITY

Deformable Part-based Models

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

65

RWTH AACHEN UNIVERSITY

2-Component Bicycle Model

Root filters coarse resolution Part filters finer resolution Deformation models

66

RWTH AACHEN UNIVERSITY

Object Hypothesis

Image pyramid HOG feature pyramid

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

67

RWTH AACHEN UNIVERSITY

Score of a Hypothesis

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

↑ 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

68

RWTH AACHEN UNIVERSITY

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

69

RWTH AACHEN UNIVERSITY

Results: Persons

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

70

RWTH AACHEN UNIVERSITY

Results: Bicycles

Slide adapted from Trevor Darrell B. Leibe

71

RWTH AACHEN UNIVERSITY

False Positives

- Bicycles

B. Leibe

72

RWTH AACHEN UNIVERSITY

Results: Cats

High-scoring true positives High-scoring false positives (not enough overlap)

Slide credit: Pedro Felzenszwalb

73

RWTH AACHEN UNIVERSITY

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>

B. Leibe

74

RWTH AACHEN UNIVERSITY

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>

75