

Computer Vision - Lecture 13

Indexing and Visual Vocabularies

12.12.2016

Bastian Leibe
 RWTH Aachen
<http://www.vision.rwth-aachen.de>
 leibe@vision.rwth-aachen.de

Computer Vision WS 16/17

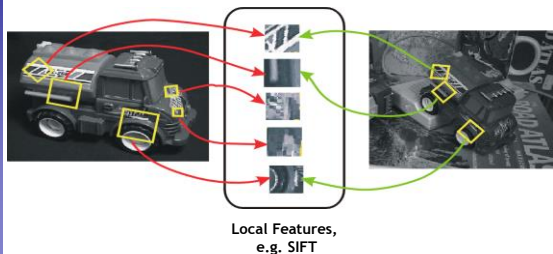
Course Outline

- 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

4

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



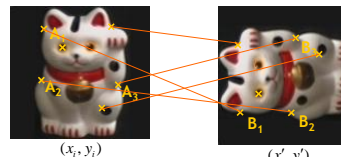
Slide credit: David Lowe

B. Leibe

6

Recap: Fitting an Affine Transformation

- Assuming we know the correspondences, how do we get the transformation?



$$\begin{bmatrix} x'_i \\ y'_i \end{bmatrix} = \begin{bmatrix} m_1 & m_2 \\ m_3 & m_4 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} + \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$$

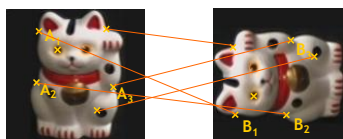
$$\begin{bmatrix} \dots & \dots & \dots & \dots & \dots & \dots \\ x_i & y_i & 0 & 0 & 1 & 0 \\ 0 & 0 & x_i & y_i & 0 & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \\ t_1 \\ t_2 \end{bmatrix} = \begin{bmatrix} \dots \\ x'_i \\ y'_i \\ \dots \end{bmatrix}$$

B. Leibe

7

Recap: Fitting a Homography

- Estimating the transformation



$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

$$\begin{bmatrix} x'' \\ y'' \\ z'' \end{bmatrix} = \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

$$x'' = \frac{1}{z'} x'$$

$$y'' = \frac{1}{z'} y'$$

$$x_A = \frac{h_{11} x_B + h_{12} y_B + h_{13}}{h_{31} x_B + h_{32} y_B + 1}$$

$$y_A = \frac{h_{21} x_B + h_{22} y_B + h_{23}}{h_{31} x_B + h_{32} y_B + 1}$$

Matrix notation $x' = Hx$

Slide credit: Krystian Mikolajczyk

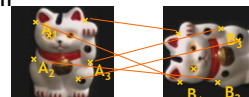
B. Leibe

8

Recap: Fitting a Homography

- Estimating the transformation

$$\begin{aligned} h_{11} x_B + h_{12} y_B + h_{13} - x_A h_{11} - x_A h_{12} y_B - x_A h_{13} &= 0 \\ h_{21} x_B + h_{22} y_B + h_{23} - y_A h_{11} - y_A h_{12} y_B - y_A h_{13} &= 0 \end{aligned}$$



$$\begin{bmatrix} x_B & y_B & 1 & 0 & 0 & 0 & -x_A x_B & -x_A y_B & -x_A \\ 0 & 0 & 0 & x_B & y_B & 1 & -y_A x_B & -y_A y_B & -y_A \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{bmatrix} \begin{bmatrix} h_{11} \\ h_{12} \\ h_{13} \\ h_{21} \\ h_{22} \\ h_{23} \\ h_{31} \\ h_{32} \\ h_{33} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \dots \end{bmatrix}$$

$$Ah = 0$$

Slide credit: Krystian Mikolajczyk

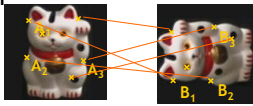
B. Leibe

9

Computer Vision WS 16/17 RWTH AACHEN UNIVERSITY

Recap: Fitting a Homography

- Estimating the transformation
- Solution:
 - Null-space vector of A
 - Corresponds to smallest eigenvector



$$Ah = 0$$

SVD

$$A = UDV^T = U \begin{bmatrix} d_{11} & \dots & d_{19} & \dots & v_{11} & \dots & v_{19} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{91} & \dots & d_{99} & \dots & v_{91} & \dots & v_{99} \end{bmatrix}^T$$

$h = \begin{bmatrix} v_{19} \\ \vdots \\ v_{99} \end{bmatrix}$ Minimizes least square error

Slide credit: Krystian Mikolajczyk B. Leibe 10

Computer Vision WS 16/17 RWTH AACHEN UNIVERSITY

Recap: Object Recognition by Alignment

- Assumption
 - Known object, rigid transformation compared to model image
 - \Rightarrow 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
 - \Rightarrow Need to use robust methods that can filter out outliers

Slide credit: B. Leibe 11

Computer Vision WS 16/17 RWTH AACHEN UNIVERSITY

Recap: Robust Estimation with RANSAC

RANSAC loop:

- Randomly select a *seed group* of points on which to base transformation estimate (e.g., a group of matches)
- Compute transformation from seed group
- Find *inliers* to this transformation
- If the number of inliers is sufficiently large, recompute least-squares estimate of transformation on all of the inliers

- Keep the transformation with the largest number of inliers

Slide credit: Kristen Grauman B. Leibe 12

Computer Vision WS 16/17 RWTH AACHEN UNIVERSITY

Problem with RANSAC


- In many practical situations, the percentage of outliers (incorrect putative matches) is often very high (90% or above).
- Alternative strategy: Generalized Hough Transform

Slide credit: Svetlana Lazebnik B. Leibe 15

Computer Vision WS 16/17 RWTH AACHEN UNIVERSITY

Strategy 2: Generalized Hough Transform

- Suppose our features are scale- and rotation-invariant
 - Then a single feature match provides an alignment hypothesis (translation, scale, orientation).

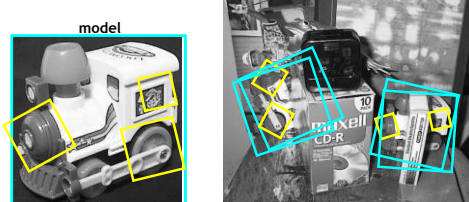


Slide credit: Svetlana Lazebnik B. Leibe 16

Computer Vision WS 16/17 RWTH AACHEN UNIVERSITY

Strategy 2: Generalized Hough Transform

- Suppose our features are scale- and rotation-invariant
 - Then a single feature match provides an alignment hypothesis (translation, scale, orientation).
 - Of course, a hypothesis from a single match is unreliable.
 - Solution: let each match vote for its hypothesis in a Hough space with very coarse bins.




Slide credit: Svetlana Lazebnik B. Leibe 17

RWTH AACHEN UNIVERSITY

Pose Clustering and Verification with SIFT

- To detect instances of objects from a model base:

1. Index descriptors
 - Distinctive features narrow down possible matches



18

Slide credit: Kristen Grauman B. Leibe Image source: David Lowe


Computer Vision WS 16/17

RWTH AACHEN UNIVERSITY

Pose Clustering and Verification with SIFT

- To detect instances of objects from a model base:

1. Index descriptors
 - Distinctive features narrow down possible matches
2. Generalized Hough transform to vote for poses
 - Keypoints have record of parameters relative to model coordinate system
3. Affine fit to check for agreement between model and image features
 - Fit and verify using features from Hough bins with 3+ votes




19

Slide credit: Kristen Grauman B. Leibe Image source: David Lowe

Computer Vision WS 16/17

RWTH AACHEN UNIVERSITY

Object Recognition Results



Background subtract for model boundaries Objects recognized Recognition in spite of occlusion

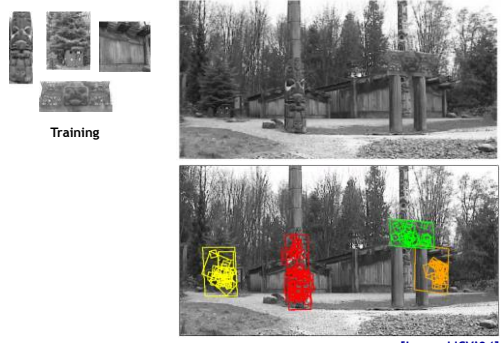
20

Slide credit: Kristen Grauman B. Leibe Image source: David Lowe

Computer Vision WS 16/17

RWTH AACHEN UNIVERSITY

Location Recognition



Training

[Lowe, IJCV'04]

21

Slide credit: David Lowe

Computer Vision WS 16/17

RWTH AACHEN UNIVERSITY

Topics of This Lecture

- Indexing with Local Features
 - Inverted file index
 - Visual Words
 - Visual Vocabulary construction
 - tf-idf weighting
- Bag-of-Words Model
 - Use for image classification

22

B. Leibe

Computer Vision WS 16/17

RWTH AACHEN UNIVERSITY

Application: Mobile Visual Search

Google Goggles in Action

Click the icons below to see the different ways Google Goggles can be used.




- Take photos of objects as queries for visual search

23

B. Leibe

Computer Vision WS 16/17

Computer Vision WS 16/17

Large-Scale Image Matching Problem

Database with thousands (millions) of images

- How can we perform this matching step efficiently?

B. Leibe 24

Computer Vision WS 16/17

Indexing Local Features

- Each patch / region has a descriptor, which is a point in some high-dimensional feature space (e.g., SIFT)

B. Leibe Figure credit: A. Zisserman

Computer Vision WS 16/17

Indexing Local Features

- When we see close points in feature space, we have similar descriptors, which indicates similar local content.

- This is of interest for many applications
 - E.g. Image matching,
 - E.g. Retrieving images of similar objects,
 - E.g. Object recognition, categorization, 3d Reconstruction,...

B. Leibe Figure credit: A. Zisserman

Computer Vision WS 16/17

Indexing Local Features

- With potentially thousands of features per image, and hundreds to millions of images to search, how to efficiently find those that are relevant to a new image?
- Low-dimensional descriptors (e.g. through PCA):
 - Can use standard efficient data structures for nearest neighbor search
- High-dimensional descriptors
 - Approximate nearest neighbor search methods more practical
- Inverted file indexing schemes

Slide credit: Kristen Grauman B. Leibe

Computer Vision WS 16/17

Indexing Local Features: Inverted File Index

- For text documents, an efficient way to find all pages on which a word occurs is to use an index...
- We want to find all images in which a feature occurs.
- To use this idea, we'll need to map our features to "visual words".

Computer Vision WS 16/17

Text Retrieval vs. Image Search

- What makes the problems similar, different?

Slide credit: Kristen Grauman

RWTH AACHEN
UNIVERSITY

Visual Words: Main Idea

- Extract some local features from a number of images ...

Computer Vision WS 16/17

B. Leibe

Slide credit: David Nister

RWTH AACHEN
UNIVERSITY

Visual Words: Main Idea

Computer Vision WS 16/17

B. Leibe

Slide credit: David Nister

RWTH AACHEN
UNIVERSITY

Visual Words: Main Idea

Computer Vision WS 16/17

B. Leibe

Slide credit: David Nister

RWTH AACHEN
UNIVERSITY

Visual Words: Main Idea

Computer Vision WS 16/17

B. Leibe

Slide credit: David Nister

RWTH AACHEN
UNIVERSITY

Each point is a local descriptor, e.g. SIFT vector.

Computer Vision WS 16/17

B. Leibe

Slide credit: David Nister

RWTH AACHEN
UNIVERSITY

Idea: quantize the feature space.

Computer Vision WS 16/17

B. Leibe

Slide credit: David Nister

RWTH AACHEN UNIVERSITY

Indexing with Visual Words

Map high-dimensional descriptors to tokens/words by quantizing the feature space

- Quantize via clustering, let cluster centers be the prototype "words"

Computer Vision WS 16/17

Slide credit: Kristen Grauman

B. Leibe

RWTH AACHEN UNIVERSITY

Indexing with Visual Words

Map high-dimensional descriptors to tokens/words by quantizing the feature space

- Determine which word to assign to each new image region by finding the closest cluster center.

Computer Vision WS 16/17

Slide credit: Kristen Grauman

B. Leibe

RWTH AACHEN UNIVERSITY

Visual Words

- Example: each group of patches belongs to the same visual word

Computer Vision WS 16/17

Figure from Sivic & Zisserman, ICCV 2003

Slide credit: Kristen Grauman

RWTH AACHEN UNIVERSITY

Visual Words

- Often used for describing scenes and objects for the sake of indexing or classification.

Computer Vision WS 16/17

Sivic & Zisserman 2003; Csurka, Bray, Dance, & Fan 2004; many others.

Slide credit: Kristen Grauman

B. Leibe

39

RWTH AACHEN UNIVERSITY

Inverted File for Images of Visual Words

Word number	List of image numbers
1	5, 10, ...
2	10, ...
...	...

When will this give us a significant gain in efficiency?

Computer Vision WS 16/17

Slide credit: Kristen Grauman

B. Leibe

Image credit: A. Zisserman

RWTH AACHEN UNIVERSITY

Example: Recognition with Vocabulary Tree

- Tree construction:

Computer Vision WS 16/17

Slide credit: David Nister

B. Leibe

[Nister & Stewenius, CVPR'06]

41

Computer Vision WS 16/17

Vocabulary Tree

- Training: Filling the tree

[Nister & Stewenius, CVPR'06]

Slide credit: David Nister B. Leibe 42

Computer Vision WS 16/17

Vocabulary Tree

- Training: Filling the tree

[Nister & Stewenius, CVPR'06]

Slide credit: David Nister B. Leibe 43

Computer Vision WS 16/17

Vocabulary Tree

- Training: Filling the tree

[Nister & Stewenius, CVPR'06]

Slide credit: David Nister B. Leibe 44

Computer Vision WS 16/17

Vocabulary Tree

- Training: Filling the tree

[Nister & Stewenius, CVPR'06]

Slide credit: David Nister B. Leibe 45

Computer Vision WS 16/17

Vocabulary Tree

- Training: Filling the tree

[Nister & Stewenius, CVPR'06]

Slide credit: David Nister B. Leibe 46

Computer Vision WS 16/17

Vocabulary Tree

- Recognition

RANSAC verification

[Nister & Stewenius, CVPR'06]

Slide credit: David Nister B. Leibe 47

RWTH AACHEN UNIVERSITY

Quiz Questions

- What is the computational advantage of the hierarchical representation vs. a flat vocabulary?
- What dangers does such a representation carry?

Computer Vision WS 16/17

48

RWTH AACHEN UNIVERSITY

Vocabulary Tree: Performance

- Evaluated on large databases
 - Indexing with up to 1M images
- Online recognition for database of 50,000 CD covers
 - Retrieval in ~1s (in 2006)
- Experimental finding that large vocabularies can be beneficial for recognition

[Nister & Stewenius, CVPR'06]

Computer Vision WS 16/17

49

RWTH AACHEN UNIVERSITY

Vocabulary Size

- Larger vocabularies can be advantageous...
- But what happens when the vocabulary gets too large?
 - Efficiency?
 - Robustness?

Computer Vision WS 16/17

50

RWTH AACHEN UNIVERSITY

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

Computer Vision WS 16/17

51

RWTH AACHEN UNIVERSITY

Summary: Indexing features

Computer Vision WS 16/17

52

RWTH AACHEN UNIVERSITY

Application for Content Based Img Retrieval

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

Visually defined query

"Find this clock" → [Image of a clock]

"Find this place" → [Image of a house]

"Groundhog Day" [Rammis, 1993]

Computer Vision WS 16/17


53

RWTH AACHEN UNIVERSITY


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



Query region



Retrieved frames

54

Computer Vision WS 16/17


B. Leibe

Slide credit: Kristen Grauman

RWTH AACHEN UNIVERSITY

Collecting Words Within a Query Region

- Example: Friends



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

55


Computer Vision WS 16/17

B. Leibe

Slide credit: Kristen Grauman


RWTH AACHEN UNIVERSITY

Example Results




Query


raw nn 1sim=0.56697



raw nn 2sim=0.56163



raw nn 5sim=0.54917



56


Computer Vision WS 16/17

B. Leibe


Slide credit: Kristen Grauman

RWTH AACHEN UNIVERSITY

More Results



Query



Retrieved shots

57

Computer Vision WS 16/17

B. Leibe

Slide credit: Kristen Grauman

RWTH AACHEN UNIVERSITY

Applications: Aachen Tourist Guide



59

Computer Vision WS 16/17

B. Leibe

RWTH AACHEN UNIVERSITY

Applications: Fast Image Registration



60

Computer Vision WS 16/17

B. Leibe

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)

Computer Vision WS 16/17

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

61

RWTH AACHEN UNIVERSITY

Topics of This Lecture

- Indexing with Local Features
 - Inverted file index
 - Visual Words
 - Visual Vocabulary construction
 - tf-idf weighting
- Bag-of-Words Model
 - Use for image classification

Computer Vision WS 16/17

62

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 come from our eyes. For that reason, it is not surprising that the most important point by which the brain receives information from the screen is the eye. In the discovery of the visual system, Hubel and Wiesel know that the perceptual system of the brain is not a simple filter of the external world. By following the path of the optical cortex, 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. Surpluses are likely to continue, as China has long had an unfair trade advantage under its surplus protection policy. Zhou Xiaochuan, who needed to be replaced in 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.

sensory, brain, visual, perception, retinal, cerebral cortex, eye, cell, optical nerve, image Hubel, Wiesel

China, trade, surplus, commerce, exports, imports, US, yuan, bank, domestic, foreign, increase, trade, value

Computer Vision WS 16/17

Slide credit: Li Fei-Fei

63

RWTH AACHEN UNIVERSITY

Object

→

Bag of 'words'

Computer Vision WS 16/17

Source: ICCV 2005 short course, Li Fei-Fei

RWTH AACHEN UNIVERSITY

Computer Vision WS 16/17

Source: ICCV 2005 short course, Li Fei-Fei

RWTH AACHEN UNIVERSITY

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.

Computer Vision WS 16/17

Slide credit: Kristen Grauman

Image credit: Li Fei-Fei

66

RWTH AACHEN UNIVERSITY

Similarly, Bags-of-Textons for Texture Repr.

Histogram
Universal texton dictionary

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

Slide credit: Svetlana Lazebnik

Computer Vision WS 16/17

RWTH AACHEN UNIVERSITY

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.

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

\vec{d}_j \vec{q}

B. Leibe Slide credit: Kristen Grauman

Computer Vision WS 16/17

RWTH AACHEN UNIVERSITY

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.

B. Leibe

Slide credit: Kristen Grauman

Computer Vision WS 16/17

RWTH AACHEN UNIVERSITY

Bags-of-Words for Classification

- 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).

Violins

B. Leibe

Slide adapted from Kristen Grauman

Computer Vision WS 16/17

RWTH AACHEN UNIVERSITY

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)

B. Leibe

Slide credit: Svetlana Lazebnik

Computer Vision WS 16/17

RWTH AACHEN UNIVERSITY

BoW for Object Categorization

Catech6 dataset

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

B. Leibe

Slide credit: Svetlana Lazebnik

Computer Vision WS 16/17

Computer Vision WS 16/17

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?

Slide adapted from Bill Freeman B. Leibe 73

Computer Vision WS 16/17

BoW Representation: Spatial Information

- A bag of words is an *orderless* representation: throwing out spatial relationships between features
- Middle ground:
 - Visual "phrases": frequently co-occurring words
 - Semi-local features: describe configuration, neighborhood
 - Let position be part of each feature
 - Count bags of words only within sub-grids of an image
 - After matching, verify spatial consistency (e.g., look at neighbors - are they the same too?)

Slide credit: Kristen Grauman B. Leibe 74

Computer Vision WS 16/17

Spatial Pyramid Representation

- Representation in-between orderless BoW and global appearance

Slide credit: Svetlana Lazebnik B. Leibe [Lazebnik, Schmid & Ponce, CVPR'06] 75

Computer Vision WS 16/17

Spatial Pyramid Representation

- Representation in-between orderless BoW and global appearance

Slide credit: Svetlana Lazebnik B. Leibe [Lazebnik, Schmid & Ponce, CVPR'06] 76

Computer Vision WS 16/17

Spatial Pyramid Representation

- Representation in-between orderless BoW and global appearance

Slide credit: Svetlana Lazebnik B. Leibe [Lazebnik, Schmid & Ponce, CVPR'06] 77

Computer Vision WS 16/17

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.

Slide credit: Kristen Grauman B. Leibe 78

References and Further Reading

- More details on RANSAC can be found in Chapter 4.7 of
 - R. Hartley, A. Zisserman
Multiple View Geometry in Computer Vision
2nd Ed., Cambridge Univ. Press, 2004
- Details about the Hough transform for object recognition can be found in
 - D. Lowe, [Distinctive image features from scale-invariant keypoints](#), *IJCV* 60(2), pp. 91-110, 2004
- Details about the Video Google system can be found in
 - J. Sivic, A. Zisserman,
[Video Google: A Text Retrieval Approach to Object Matching in Videos](#), ICCV'03, 2003.

