

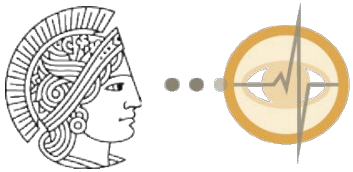
BoW: no spatial relationships

Part Based Object and People Detection

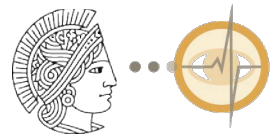
Cognitive Science Summerschool, Aug 27, 2oo9
Part 2: Bag of Words Models

Bernt Schiele

TU Darmstadt, Germany
http://www.mis.informatik.tu-darmstadt.de/
schiele@informatik.tu-darmstadt.de



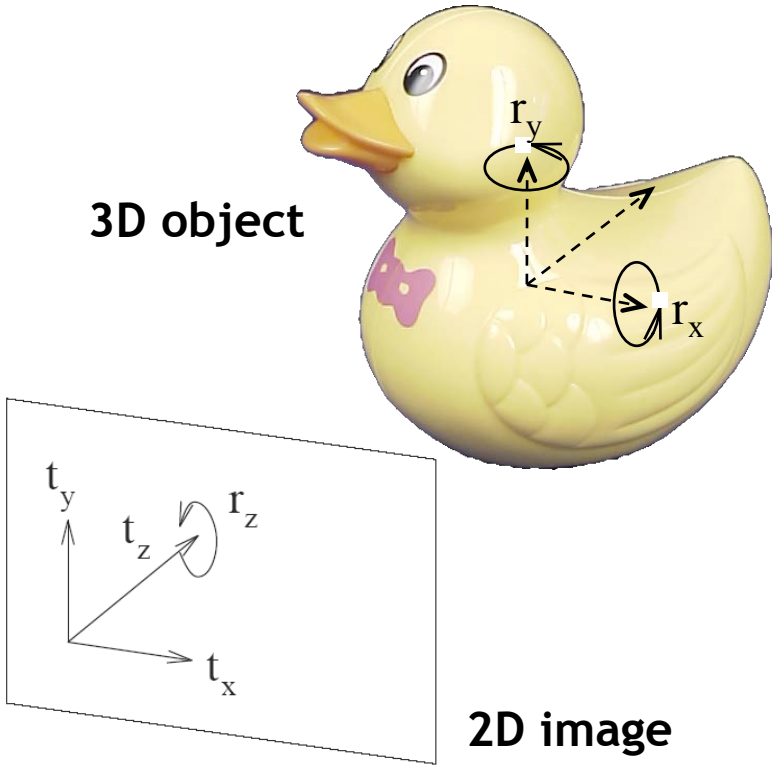
Overview Part 2:
Bag of Words (BoW) - Models



- Appearance-Based Recognition
  - a paradigm-shift in the 90's
  - PCA & first histogram-based models
- 'Today's' BoW-Models
  - local interest points
    - such as scale invariant interest points, ...
  - robust local features
    - such as SIFT, ...
  - discriminant classifiers
    - such as Support Vector Machines (SVM), ...

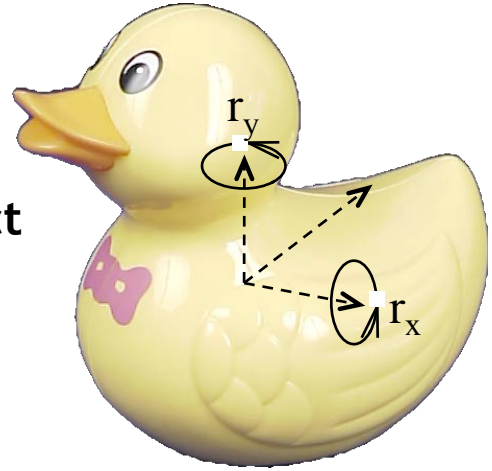
Appearance Based Recognition: Challenges

- Viewpoint changes
  - Translation
  - Scale changes
  - Image-plane rotation
  - Out-of-plane rotation
- Illumination
- Clutter
- Occlusion
- Noise



Appearance-Based Identification / Recognition

- Basic assumption
  - Objects can be represented by a set of images ("appearances").
  - For recognition, it is sufficient to just compare the 2D appearances.
  - No 3D model is needed.



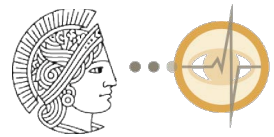




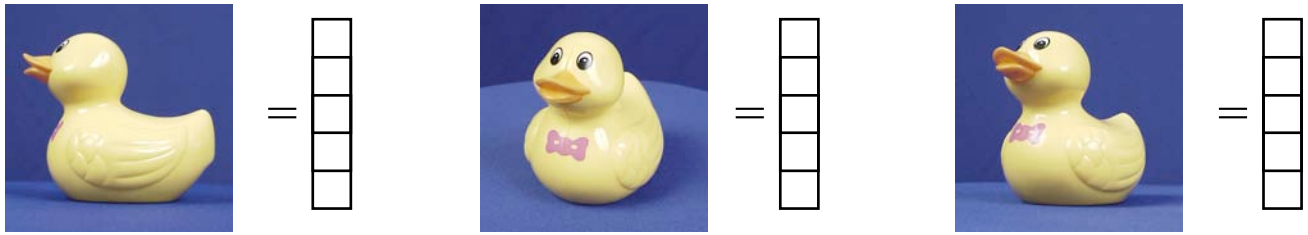


⇒ Fundamental paradigm shift in the 90's

Global Representation



- Idea
  - Represent each object (view) by a global descriptor.



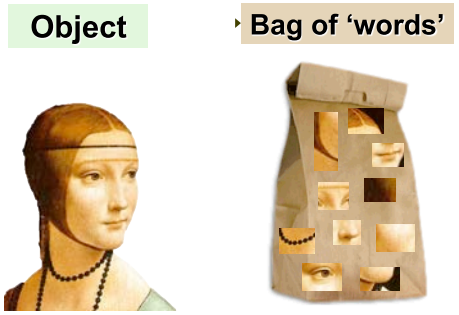
  - For recognizing objects, just match the (global) descriptors.
  - Some modes of variation are built into the descriptor, others have to be incorporated in the training data or the recognition process.
    - e.g. a descriptor can be made invariant to image-plane rotations, translation.
    - Other variations:
      - Viewpoint changes: Scale changes, Out-of-plane rotation
      - Illumination: Noise, Clutter, Occlusion

Appearance Based Models



- 1. Principal Component Analysis
  - Eigenfaces [Turk&Pentland'91]
  - PCA for Object Recognition [Murase&Nayar'95]

- 2. Statistics of Local Features
  - Color Histogram Approach [Swain&Ballard'91]
  - Multidimensional Receptive Field Histogram Approach [Schiele&Crowley'96-'00]
  - Bag of Words Approach [Csurka-et-al'04], [Tuytelaars&Schmid'07], ...



picture from Li Fei-Fei

Visual words distributions

Visual Codeword Dictionary:



BoW = Occurrence Histogram of Visual Codewords:

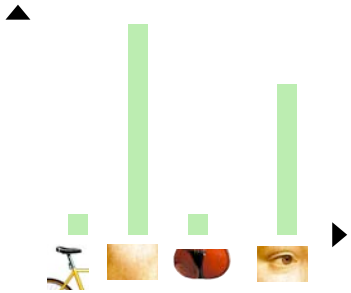


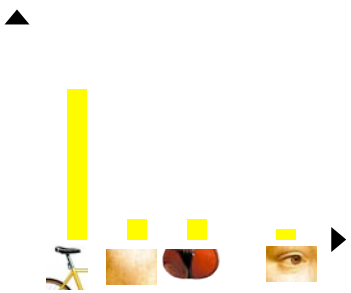


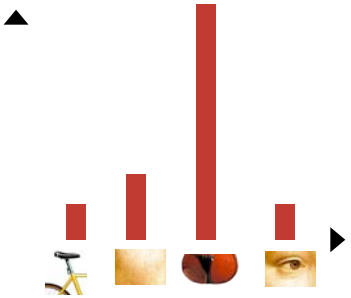


Bag of 'words'

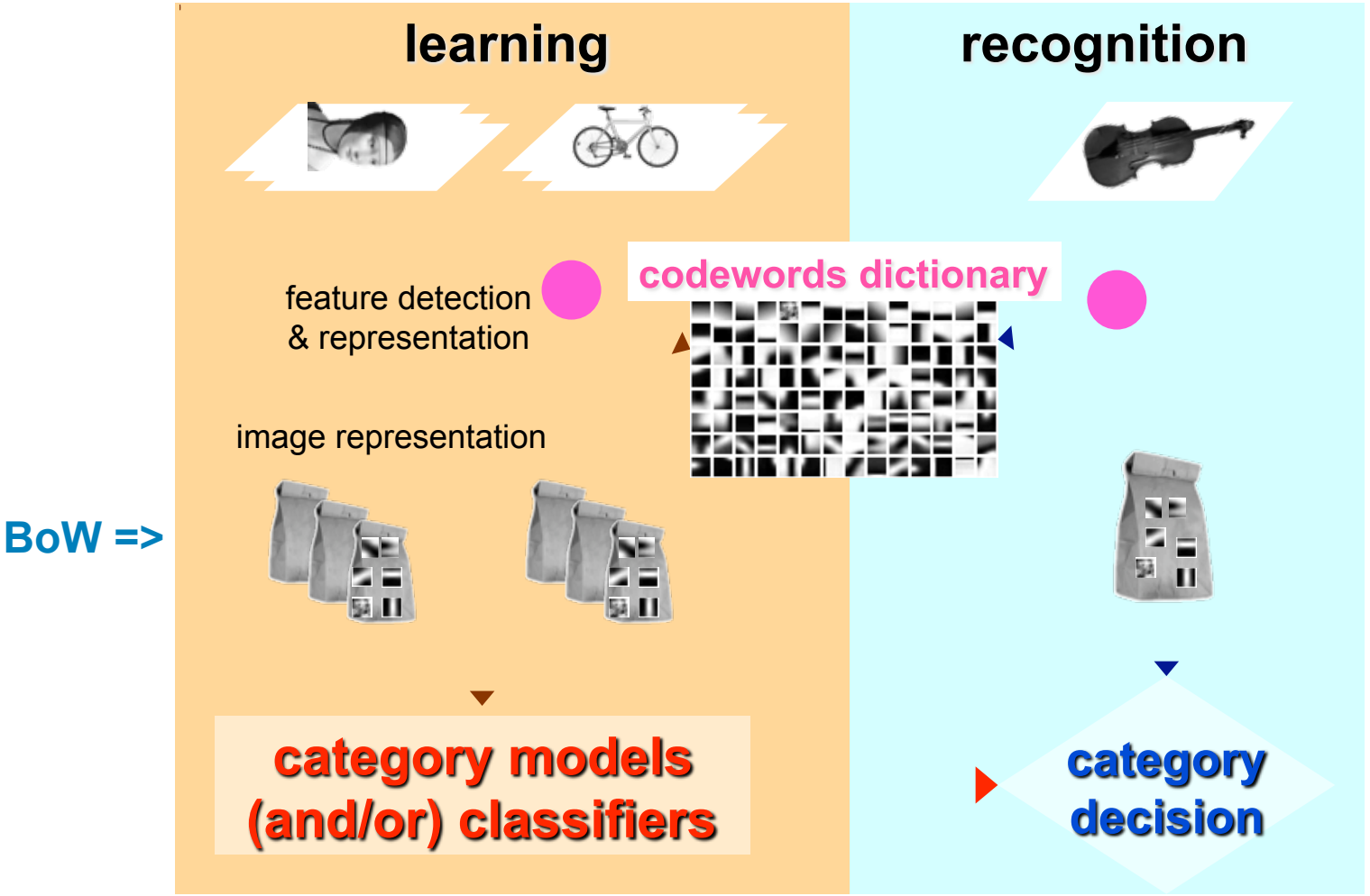








Bag-of-Words Model: Overview

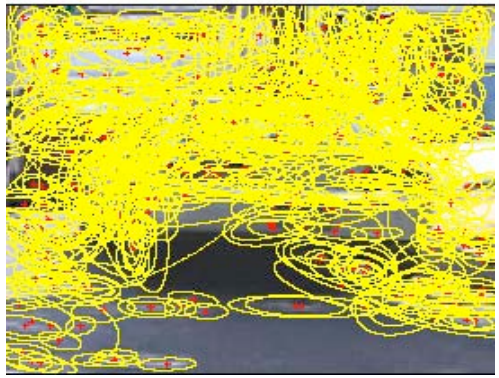


1. Feature detection and representation

- Regular grid:
  - Color Histogram Approach [Swain&Ballard'91]
  - Multidimensional Receptive Field Histograms [Schiele&Crowley'96-'00]



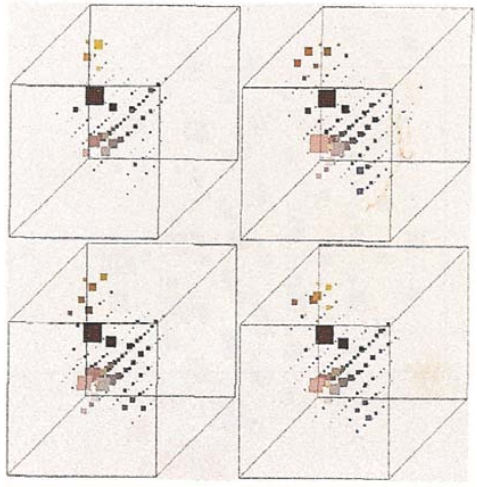
- Interest point detector:
  - use state-of-the-art interest point detector
    - e.g. scale- or affine-invariant
  - represent by using state-of-the-art features
    - e.g. SIFT, ShapeContext



Color Histograms: Use for Recognition

- Color:
  - Color stays constant under geometric transformations
  - Local feature
    - Color is defined for each pixel
    - Robust to partial occlusion





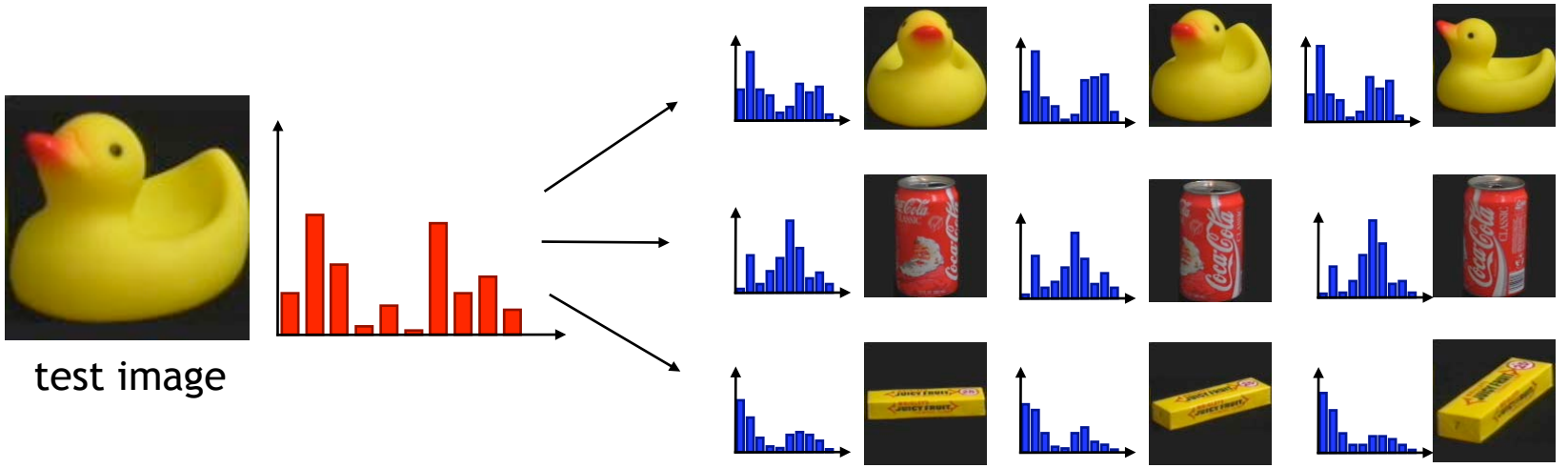
[Swain & Ballard, 1991]

Recognition using Histograms

- Simple algorithm

  "Nearest-Neighbor" strategy

  1. Build a set of histograms $H = \{M_1, M_2, M_3, \ldots\}$ for each known object
     - More exactly, for each view of each object
  2. Build a histogram T for the test image.
  3. Compare T to each $M_k \in H$
     - Using a suitable comparison measure
  4. Select the object with the best matching score
     - Or reject the test image if no object is similar enough.



test image

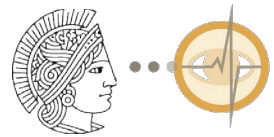
Color Histograms

- Recognition
  - Works surprisingly well
  - In the first paper (1991), 66 objects could be recognized almost without errors



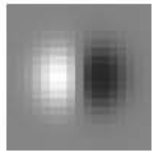
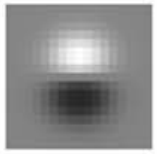
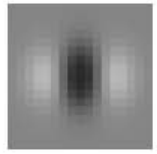
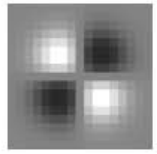
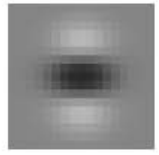
[Swain & Ballard, 1991]

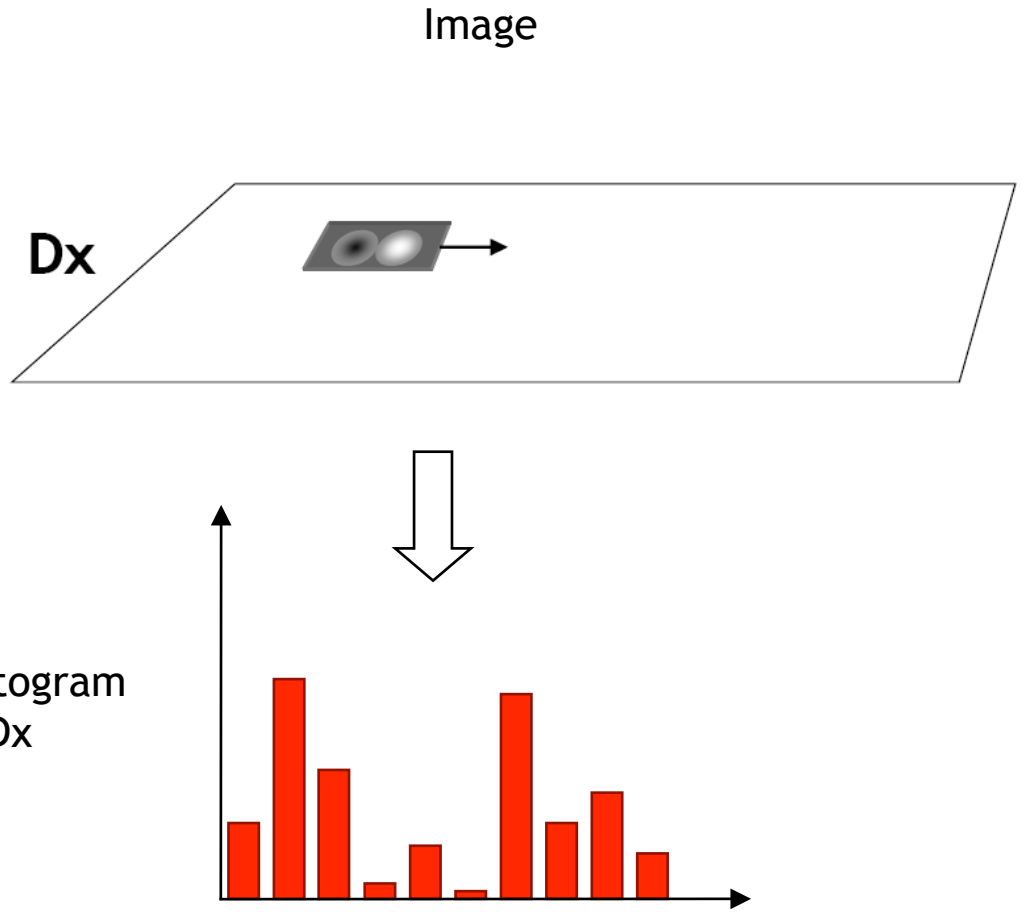
Discussion: Color Histograms



- Advantages
  - Invariant to object translations
  - Invariant to image rotations
  - Slowly changing for out-of-plane rotations
  - No perfect segmentation necessary
  - Histograms change gradually when part of the object is occluded
  - Possible to recognize deformable objects
    - e.g. pullover
- Problems
  - The pixel colors change with the illumination („color constancy problem“)
    - Intensity
    - Spectral composition (illumination color)
  - Not all objects can be identified by their color distribution.

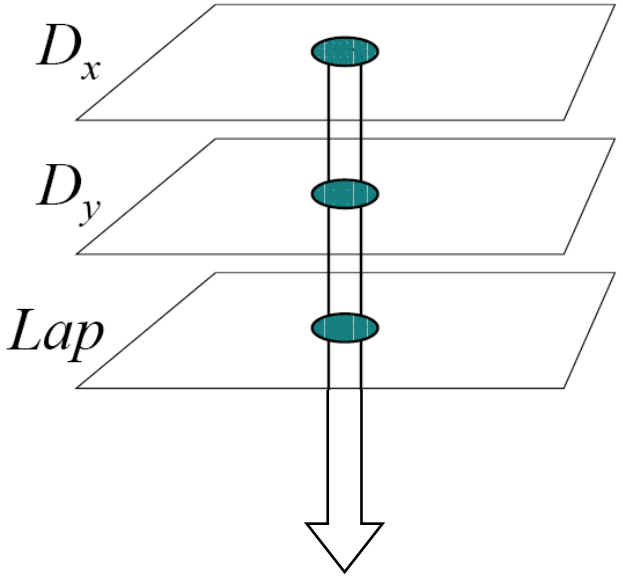
Generalization of the Idea

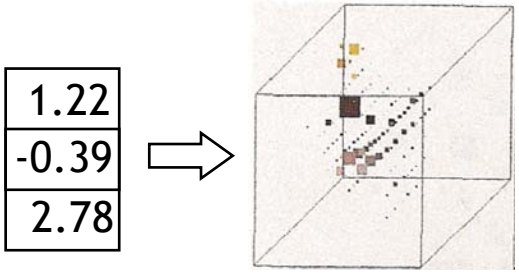
- Histograms of derivatives
  - Dx 
  - Dy 
  - Dxx 
  - Dxy 
  - Dyy 



Multidimensional Histograms

- Combination of several descriptors
  - Each descriptor is applied to the whole image.
  - Corresponding pixel values are combined into one feature vector.
  - Feature vectors are collected in a multidimensional histogram.



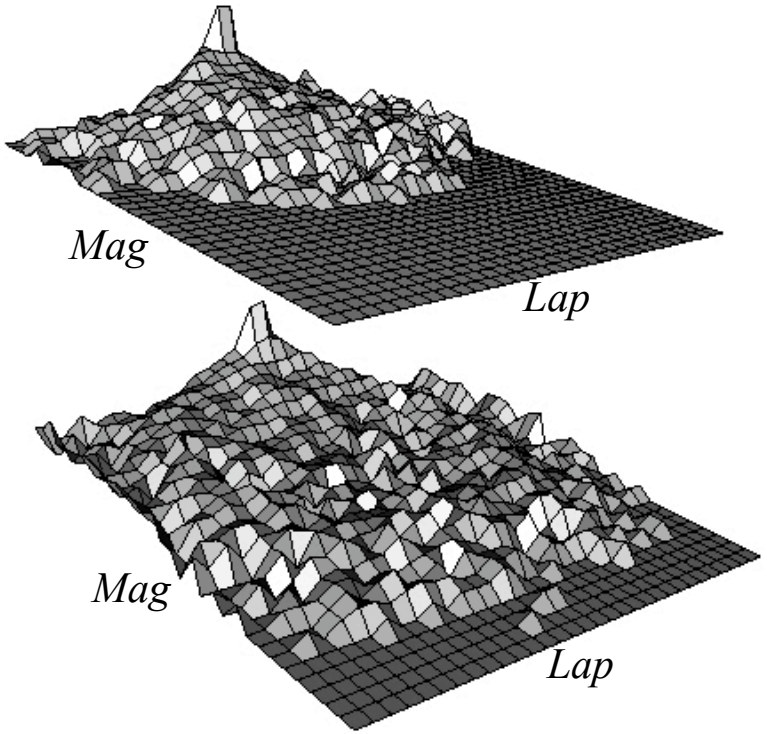


Multidimensional Histograms

- Examples

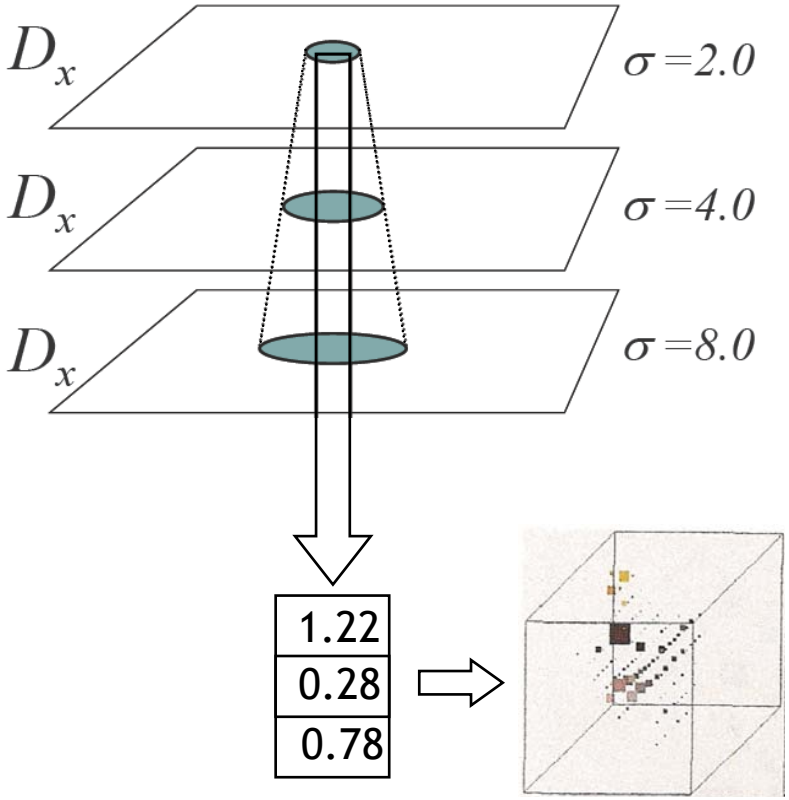
[Schiele & Crowley, 2000]



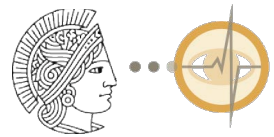


Multidimensional Histograms

- Combination of several scales
  - Descriptors are computed at different scales.
  - Each scale captures different information about the object.
  - Size of the support region grows with increasing σ.
  - Feature vectors capture both local details and larger-scale structures.



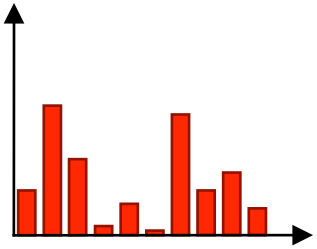
Probabilistic Recognition



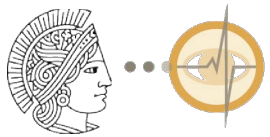
- Probability of object o_n given feature vector m_k

$$p(o_n|m_k) = \frac{p(m_k|o_n)p(o_n)}{p(m_k)} = \frac{p(m_k|o_n)p(o_n)}{\sum_i p(m_k|o_i)p(o_i)}$$

- with
  - $p(o_n)$ the a priori probability of object o_n,
  - $p(m_k)$ the a priori probability of feature vector m_k,
  - $p(m_k|o_n)$ the probability density function of object o_n.
    - directly given by (normalized) histogram !



Probabilistic Recognition (Naive Bayes)



- Joint probability for K independent feature vectors

$$p(o_n|\bigwedge_k m_k) = \frac{p(\bigwedge_k m_k|o_n)p(o_n)}{\sum_i p(\bigwedge_k m_k|o_i)p(o_i)}$$

$$= \frac{\prod_k p(m_k|o_n)p(o_n)}{\sum_i \prod_k p(m_k|o_i)p(o_i)}$$

- Assumption: all objects are equally probable
  - $\Rightarrow \quad p(o_i) = \frac{1}{N}$

$$\Rightarrow \quad p(o_n|\bigwedge_k m_k) = \frac{\prod_k p(m_k|o_n)}{\sum_i \prod_k p(m_k|o_i)}$$

Experimental Evaluation

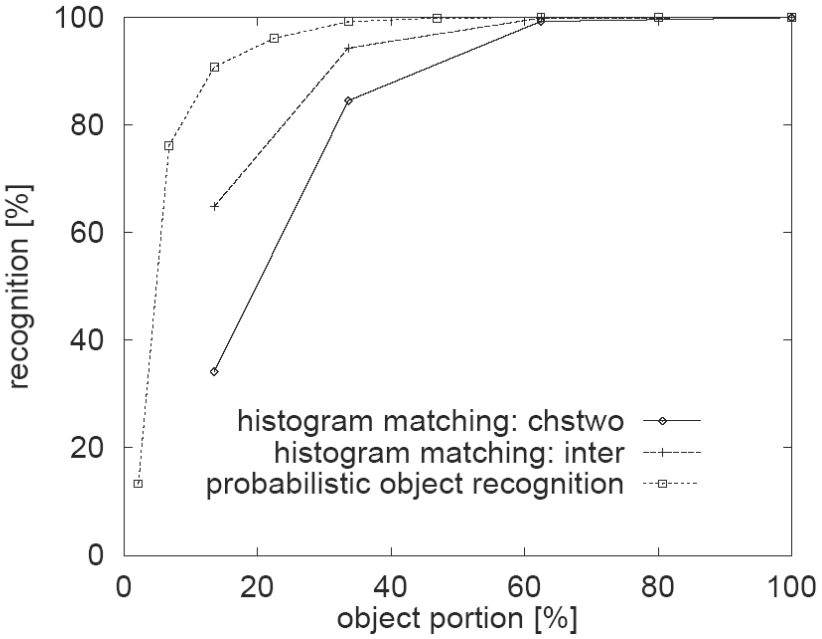
- Test database
  - 103 test objects
  - 1327 test images total
    - 607 images with scale changes and rotations for 83 objects
    - 720 images with different viewpoints for 20 objects
  - Use 6D descriptor
    - D_x-D_y with $\sigma_i=\{1,2,4\}$
    - explicitly trained for scale changes & rotations



[Schiele & Crowley, 2000]

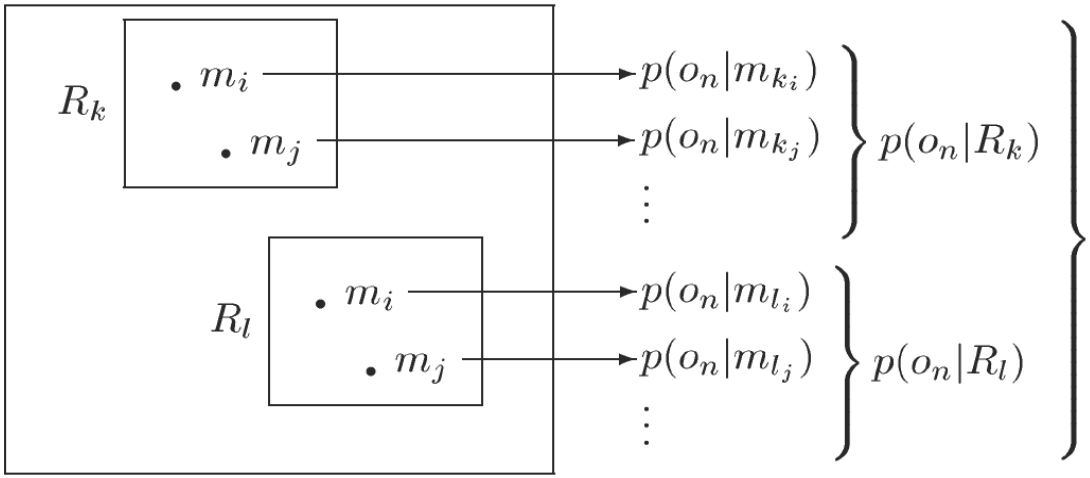
Experimental Evaluation

- Recognition under Partial Occlusion
  - Compare intersection (inter), $\chi 2$ (chstwo), and probabilistic recognition
- Results
  - Intersection more robust to occlusion than $\chi 2$
  - Probabilistic recognition most robust
    - 62% visibility ⇒ 100% recognition
    - 33% visibility ⇒ 99% recognition
    - 13% visibility ⇒ >90% recognition



[Schiele & Crowley, 2000]

Recognition of Multiple Objects



$$vote(o_n|Image) = \sum_k p(o_n|R_k)$$

- Local Appearance Hashing
  - Combination of the probabilistic recognition with a hash table
  - Only relatively small object region is needed for recognition.
    ⇒ Divide image into set of (overlapping) regions.
  - Each region votes for a single object.
  - Region votes are combined to vote for the presence of object on.

Recognition Results



Test image 1



First Match



Second Match



Third Match



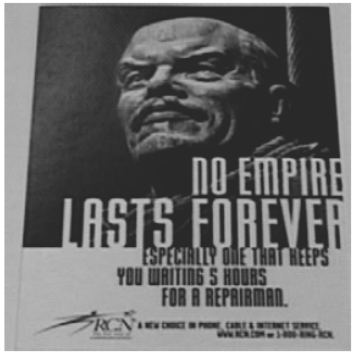
Test image 2



First Match



Second Match



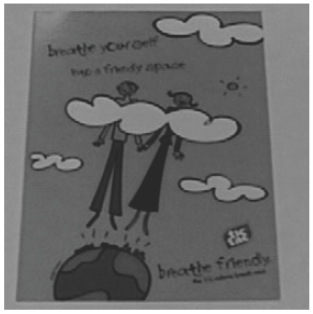
Third Match

[Schiele & Crowley, 2000]

Recognition Results



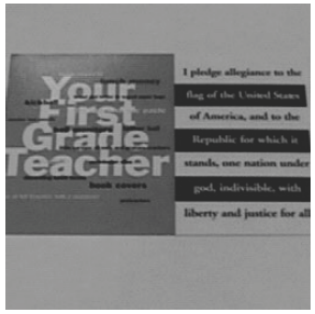
Test image 3



First Match



Second Match



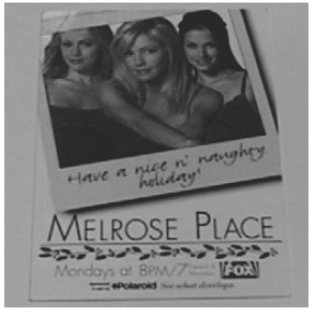
Third Match



Test image 4



First Match



Second Match



Third Match

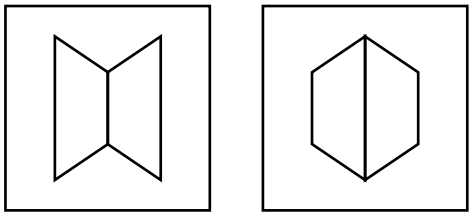


Fourth Match

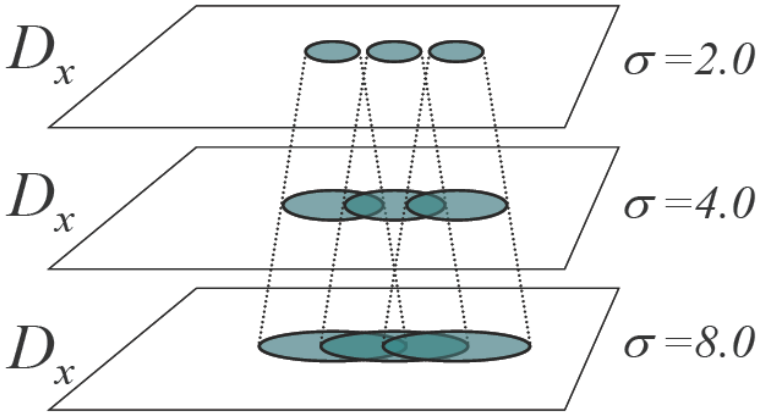
[Schiele & Crowley, 2000]

Why Does It Work So Well?

- Histogram Representation
  - Contains no structural description.
  - Many different objects should result in the same histograms.



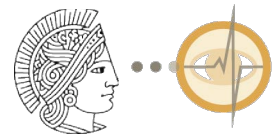
- But
  - Support regions of neighboring descriptors overlap.
  - Neighborhood relations are captured implicitly.



Museum's Guide…



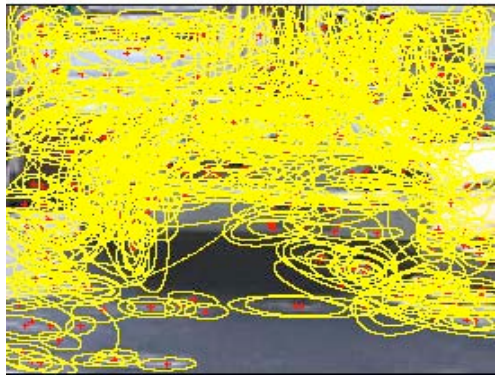
1. Feature detection and representation



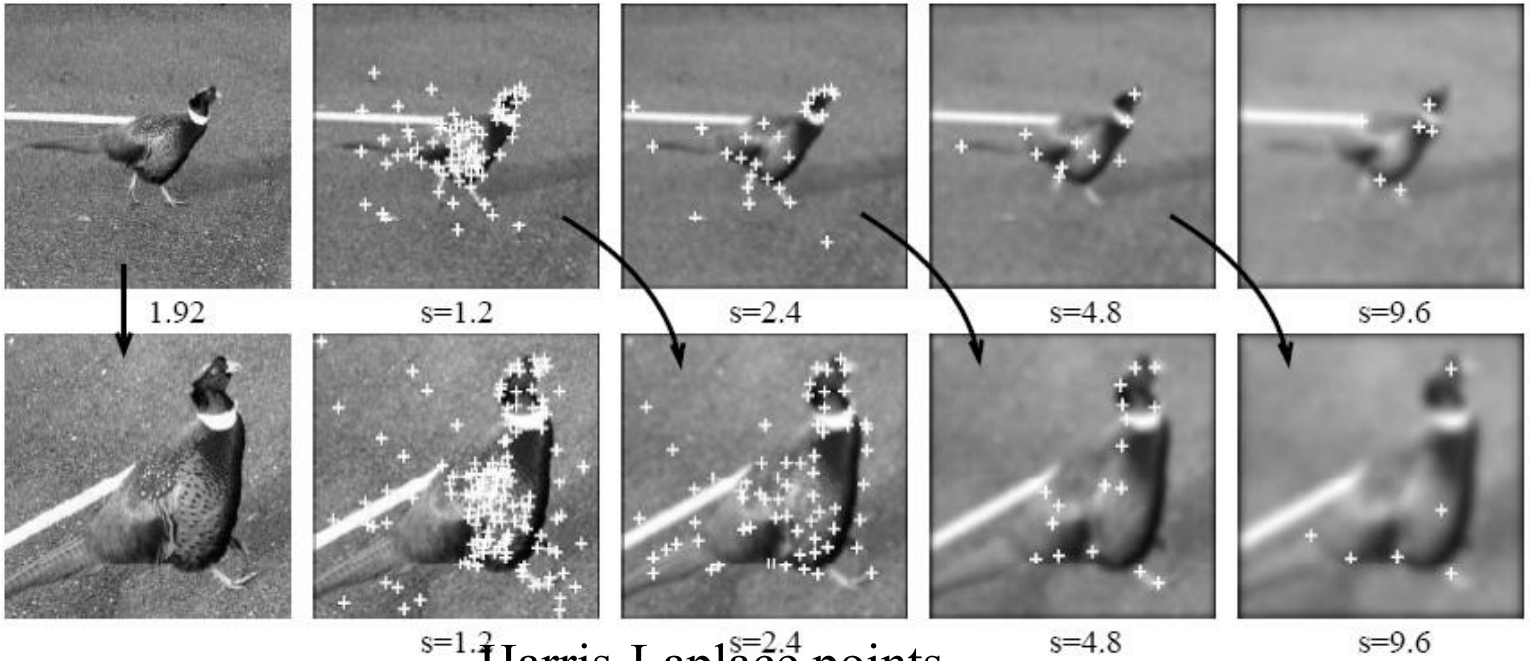
- Regular grid:
  - Color Histogram Approach [Swain&Ballard'91]
  - Multidimensional Receptive Field Histograms [Schiele&Crowley'96-'00]



- Interest point detector:
  - use state-of-the-art interest point detector
    - e.g. scale- or affine-invariant
  - represent by using state-of-the-art features
    - e.g. SIFT, ShapeContext

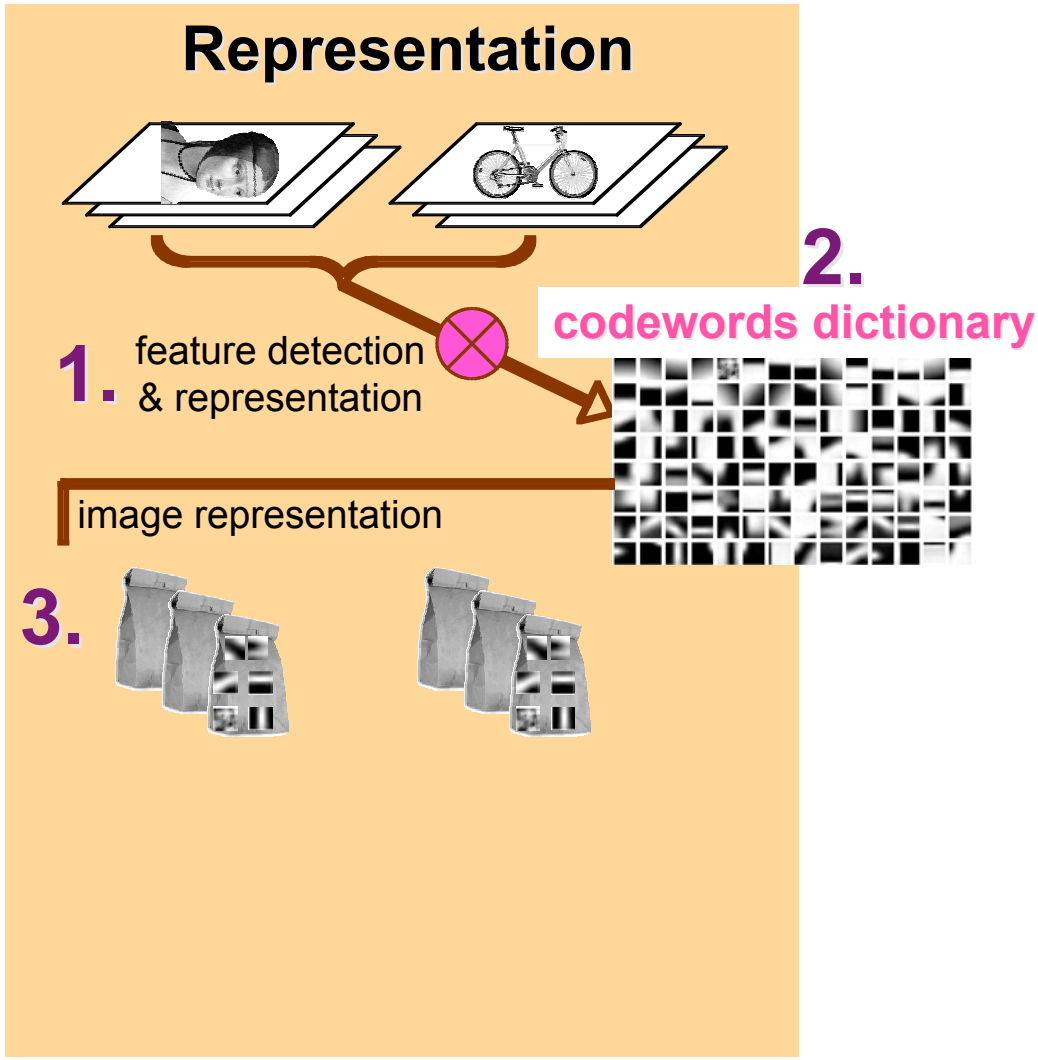


Scale invariant detectors
e.g. Harris-Laplace

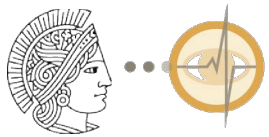


- Harris-Laplace Detector:
  - Detect Harris points over multiple scales
  - Select Harris points which maximize the Laplacian
    - i.e. Automatic scale selection

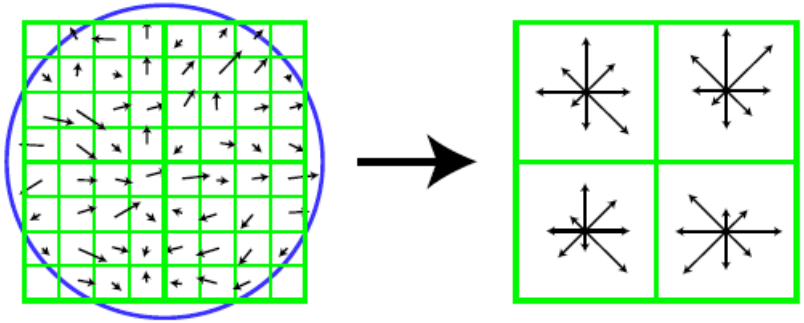
Harris-Laplace points


Representation
2.
codewords dictionary
1.
feature detection
& representation
image representation
3.

SIFT - Scale Invariant Feature Transform [Lowe]



- Interest Points:
  - Difference of Gaussians
- Feature Descriptor:
  - local histogram of 4x4 local orientation histograms (each over 16x16 pixels),
    - 8 orientations x 4 x 4 = 128 dimensions
  - example: 2x2 local orientation histogram (each of 4x4 pixels):

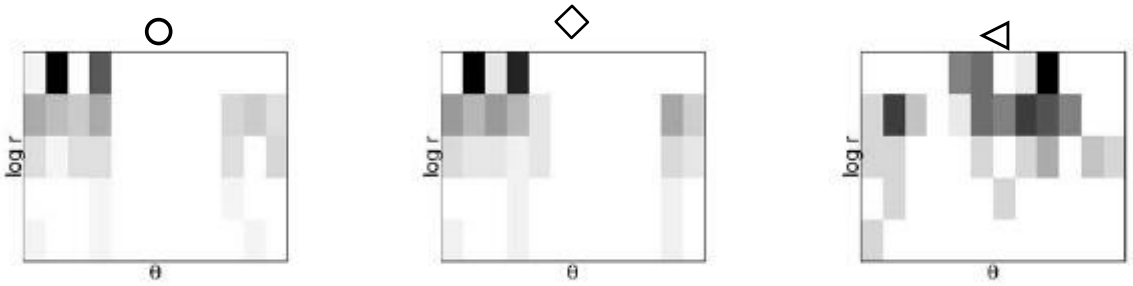


Local Descriptors

- Shape context
  - invariant – only when computed on normalized patches



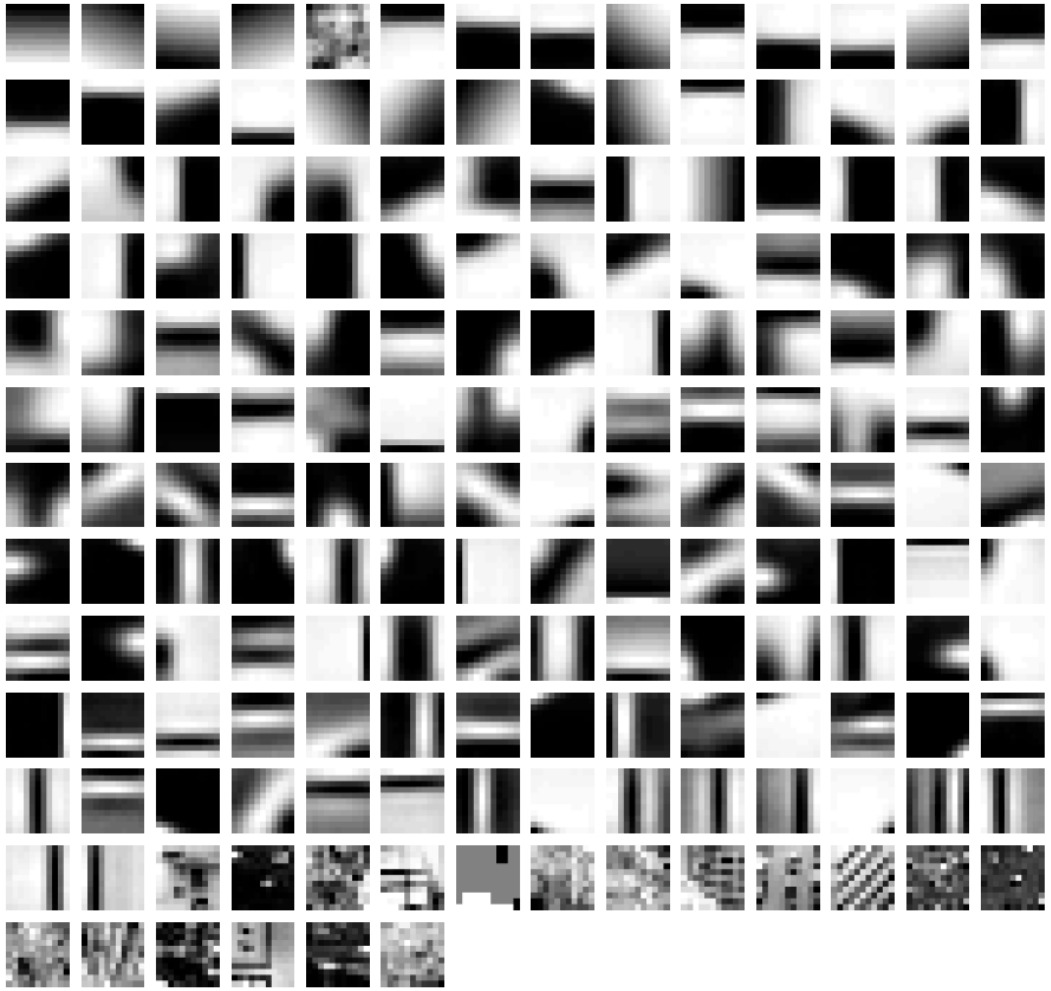
Log polar coordinate system



Histogram of edge points in log-polar coordinate system

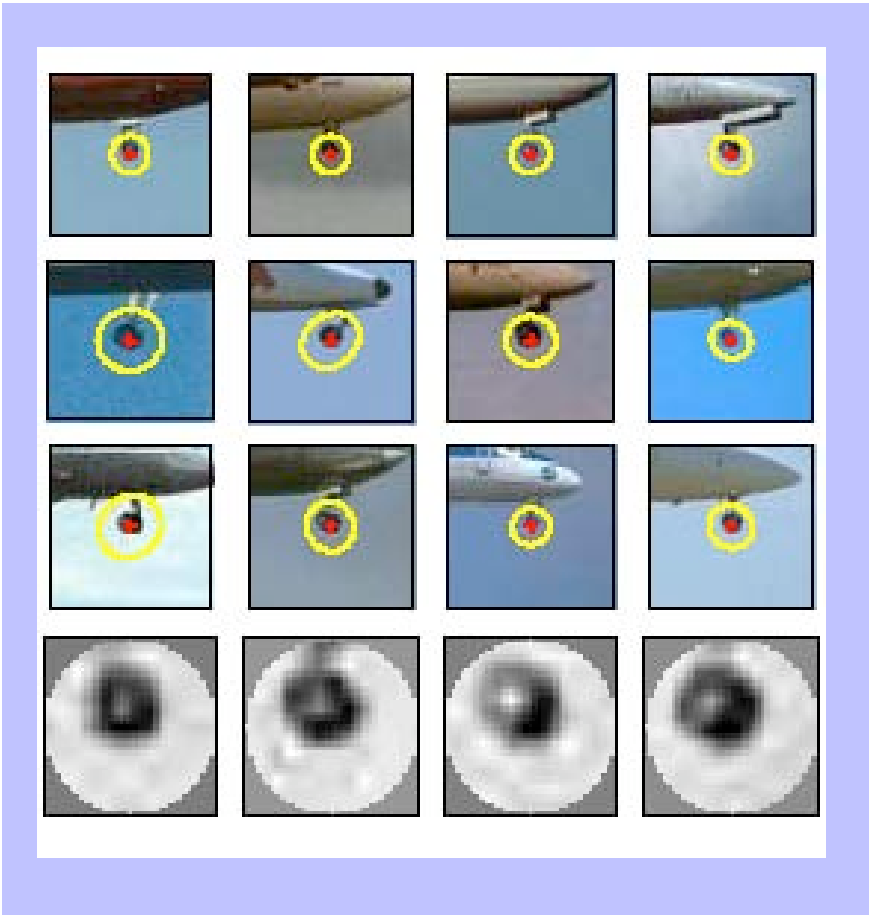


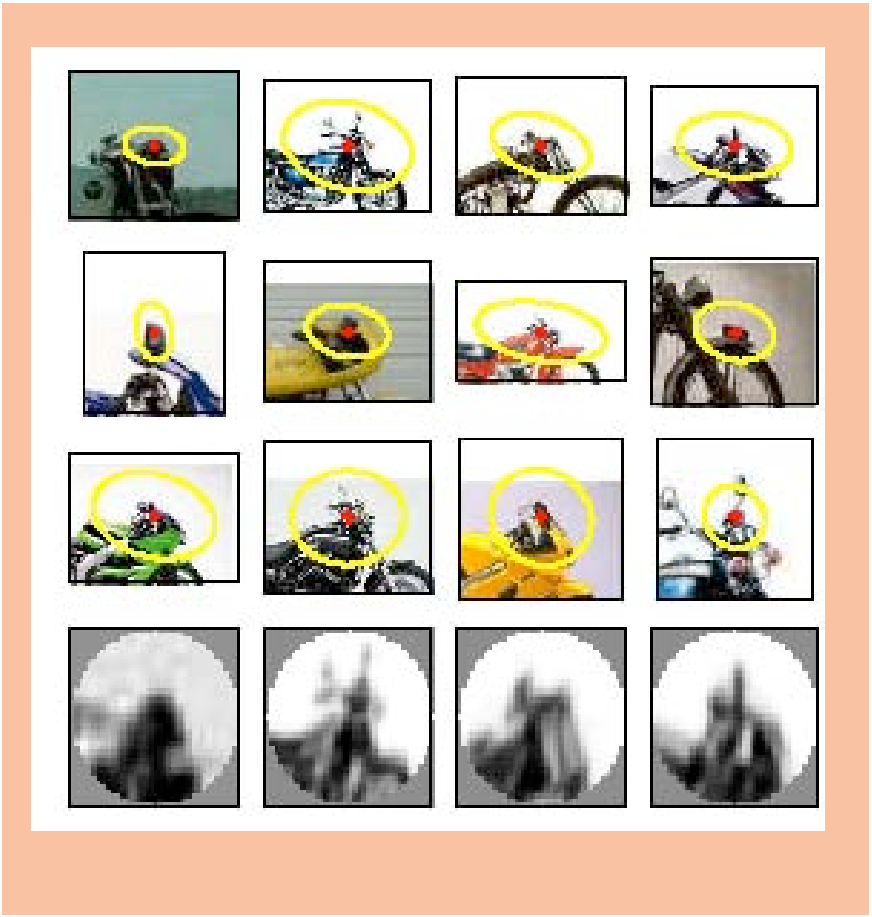
2. Codewords dictionary formation



Fei-Fei et al. 2005

Image patch examples of codewords





Sivic et al. 2005

3. Object / Image representation



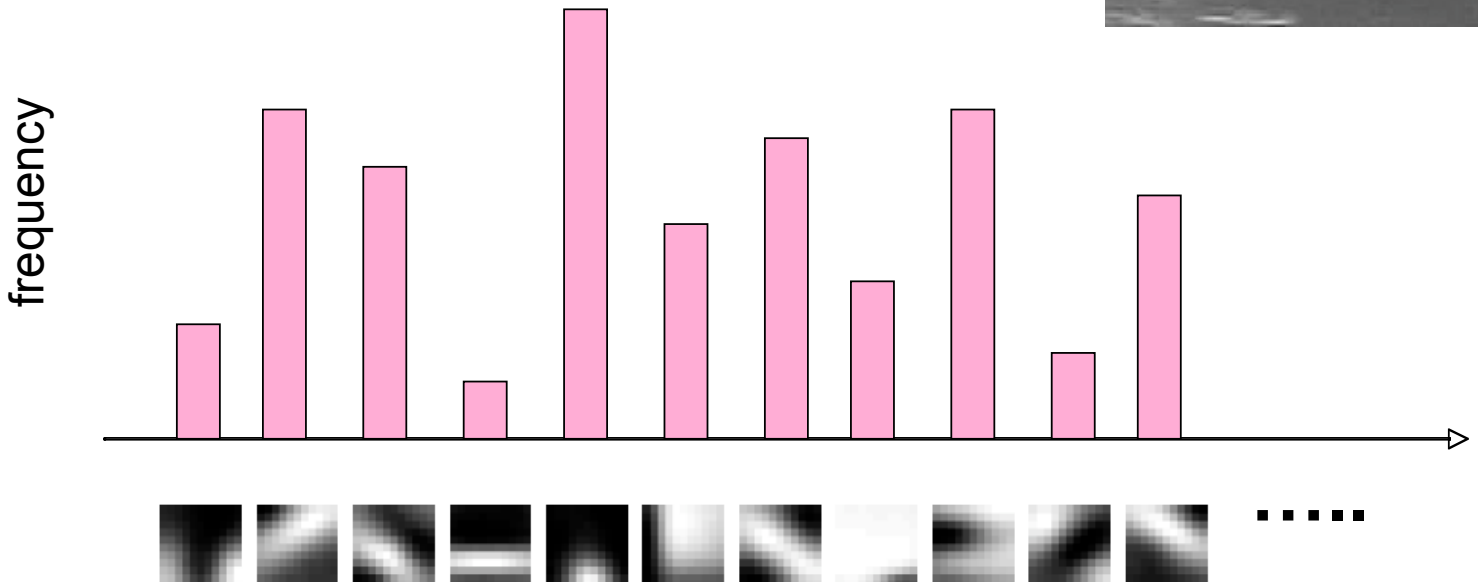


Image Classification with Naive Bayes

- Image dataset: 7 object categories, arbitrary views, partial occlusions



Csurka et al. 2004

Example of feature extraction



All features detected in the image



Features corresponding to two different visual words

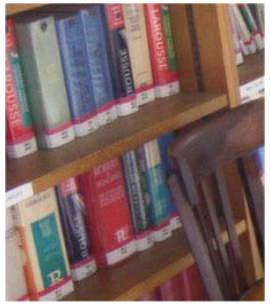
Csurka et al. 2004

Recognition results:

Table 1. Confusion matrix and the mean rank for the best vocabulary (*k=1000*).

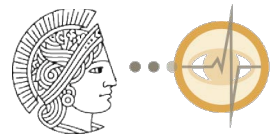
True classes →	*faces*	*buildings*	*trees*	*cars*	*phones*	*bikes*	*books*
faces	**76**	4	2	3	4	4	13
buildings	2	**44**	5	0	5	1	3
trees	3	2	**80**	0	0	5	0
cars	4	1	0	**75**	3	1	4
phones	9	15	1	16	**70**	14	11
bikes	2	15	12	0	8	**73**	0
books	4	19	0	6	7	2	**69**
Mean ranks	1.49	1.88	1.33	1.33	1.63	1.57	1.57

Examples of correctly classified images:



Summary & Discussion



- Bag-of-words representation:
  - Sparse representation of object category
  - Many machine learning methods are directly applicable.
  - Robust to occlusions
  - Allows sharing of representation between multiple classes
- Problems:
  - Localization of objects in images is problematic
  - Spatial distribution of visual words is not modeled, all these images have equal probability for bag-of-words methods:

