

“Bag of Words”



*A quiet meditation on the importance
of trying simple things first...*

16-721: Advanced Machine Perception

A. Efros, CMU, Spring 2009

Adopted from Fei-Fei Li, S.c. Zhu,
and L.Walker Renninger

Object



Bag of 'words'



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 eyes.

For a long time, the retinal image was considered as a movie screen. It is now known that the image is processed in a more complex way.

Following the discovery of the visual cortex, Hubel and Wiesel have demonstrated that the message about the image falling on the retina undergoes a fine analysis in a system of nerve cells stored in columns. In this system each cell has its specific function and is responsible for a specific detail in the pattern of the retinal image.

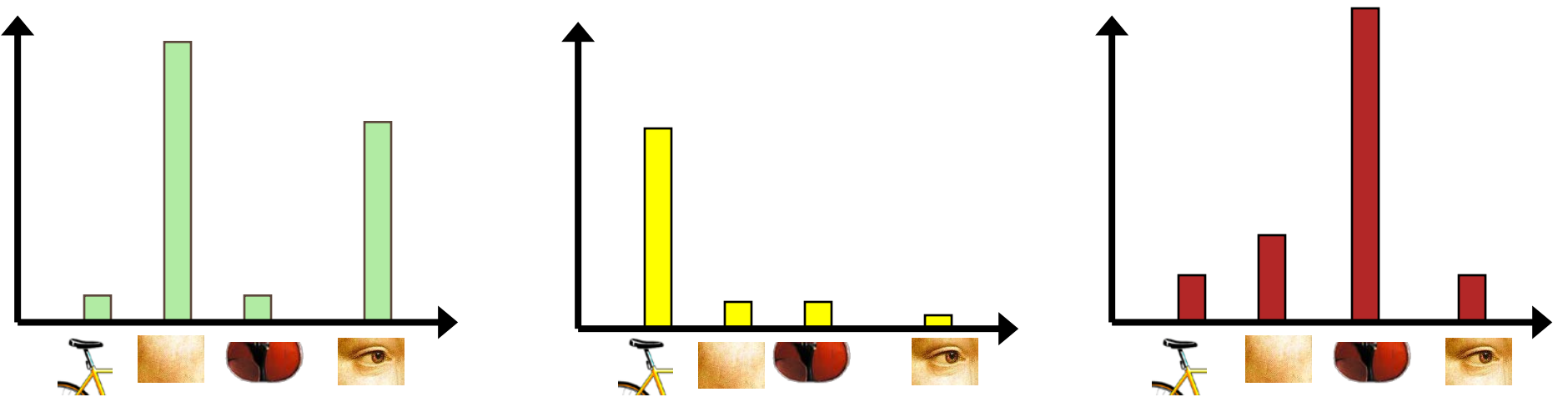
**sensory, brain,
visual, perception,
retinal, cerebral cortex,
eye, cell, optical
nerve, image
Hubel, Wiesel**

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 predicted 30% increase in exports to \$750bn, compared with \$560bn in 2004.

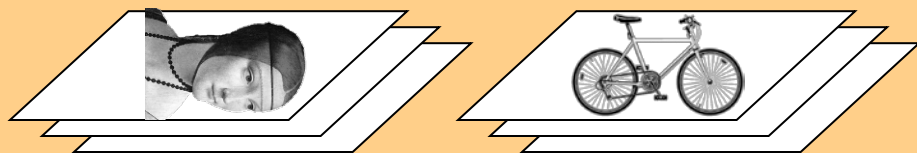
The increase in exports will be partly offset by a rise in imports to \$660bn. The government says the surplus will be a temporary phenomenon and will not annoy the US.

China's government has deliberately agreed to a 3% increase in the yuan is expected to rise to 6.8 yuan per dollar by the end of the year. The government also needs to increase the demand for the yuan in the country. China's government has also permitted it to trade within a narrow band but the US wants the yuan to be allowed to rise 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.

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



learning



feature detection
& representation

codewords dictionary

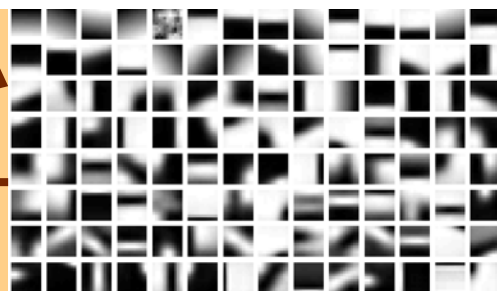
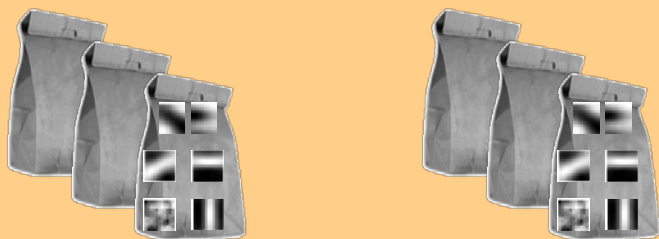


image representation



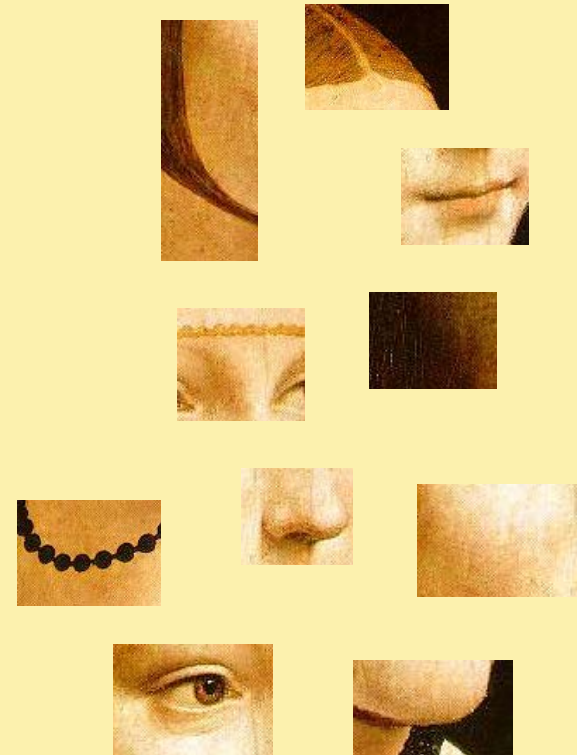
**category models
(and/or) classifiers**

recognition



**category
decision**

1.Feature detection and representation



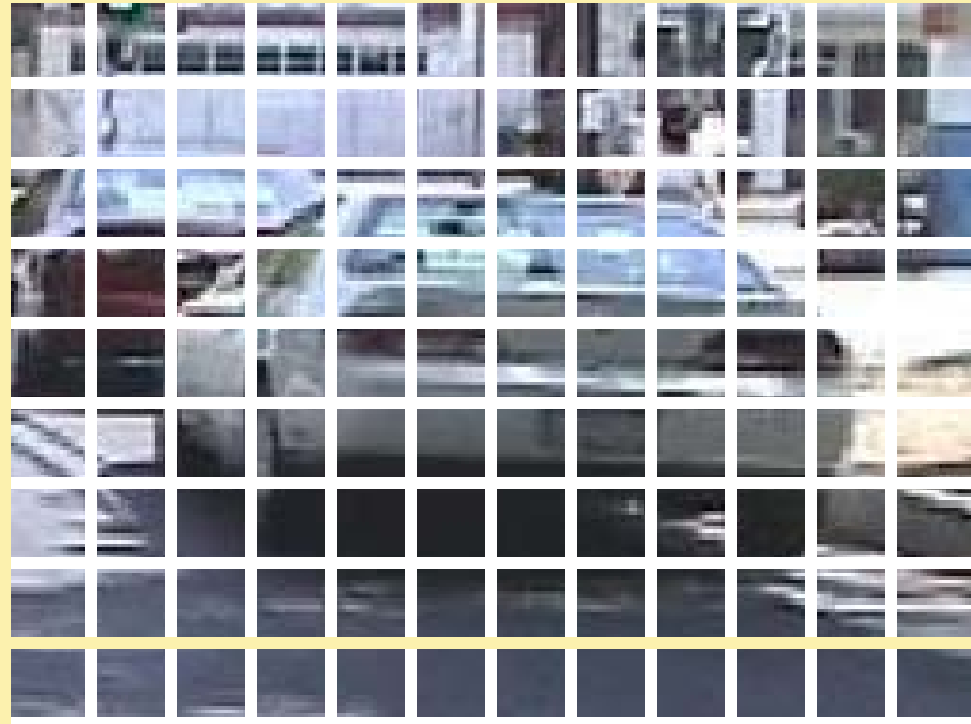
Feature detection

- Sliding Window
 - Leung et al, 1999
 - Viola et al, 1999
 - Renninger et al 2002



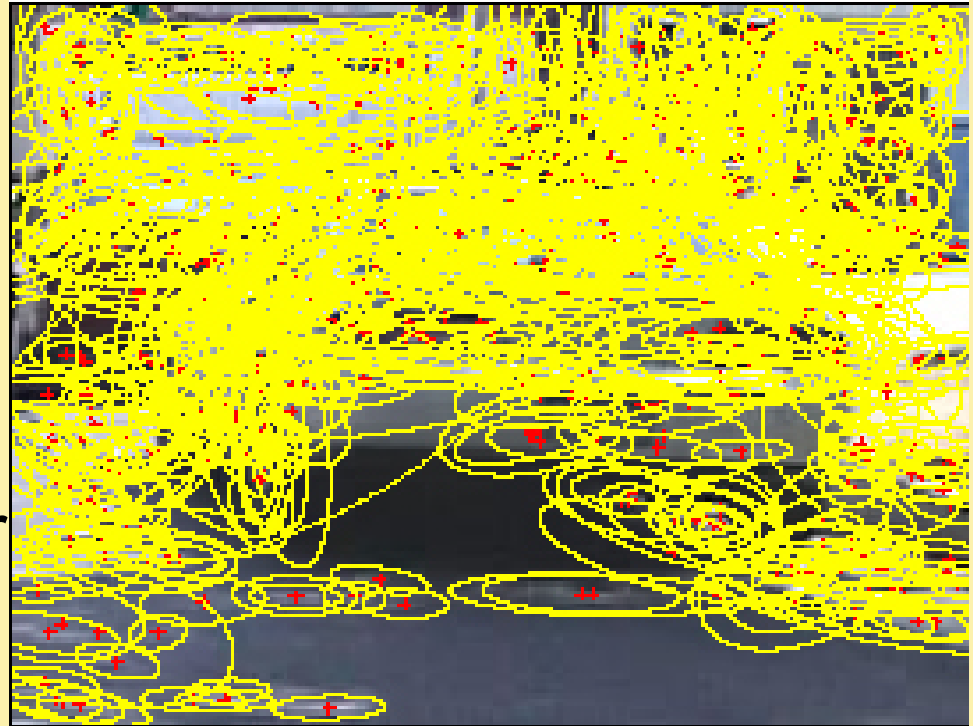
Feature detection

- Sliding Window
 - Leung et al, 1999
 - Viola et al, 1999
 - Renninger et al 2002
- Regular grid
 - Vogel et al. 2003
 - Fei-Fei et al. 2005



Feature detection

- Sliding Window
 - Leung et al, 1999
 - Viola et al, 1999
 - Renninger et al 2002
- Regular grid
 - Vogel et al. 2003
 - Fei-Fei et al. 2005
- Interest point detector
 - Csurka et al. 2004
 - Fei-Fei et al. 2005
 - Sivic et al. 2005



Feature detection

- Sliding Window
 - Leung et al, 1999
 - Viola et al, 1999
 - Renninger et al 2002
- Regular grid
 - Vogel et al. 2003
 - Fei-Fei et al. 2005
- Interest point detector
 - Csurka et al. 2004
 - Fei-Fei et al. 2005
 - Sivic et al. 2005
- Other methods
 - Random sampling (Ullman et al. 2002)
 - Segmentation based patches
 - Barnard et al. 2003, Russell et al 2006, etc.)

Feature Representation

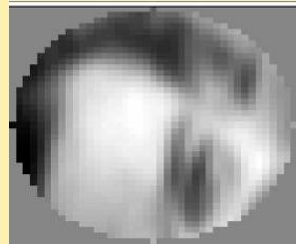
Visual words, aka textons, aka keypoints:
K-means clustered pieces of the image

- Various Representations:
 - Filter bank responses
 - Image Patches
 - SIFT descriptors

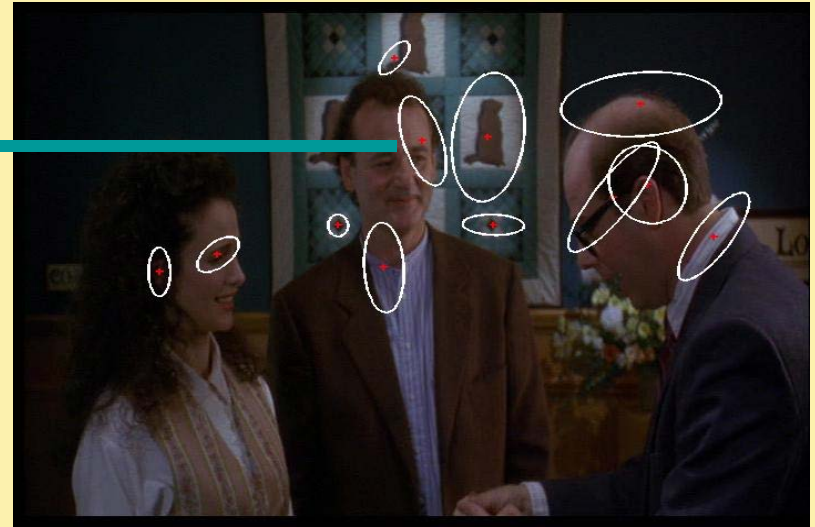
All encode more-or-less the same thing...

Interest Point Features


**Compute
SIFT
descriptor**
[Lowe'99]



**Normalize
patch**



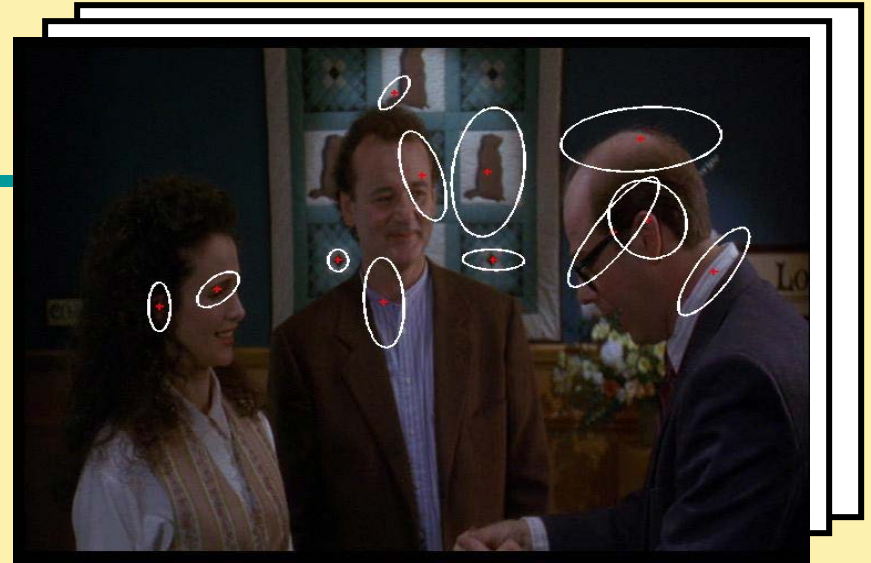
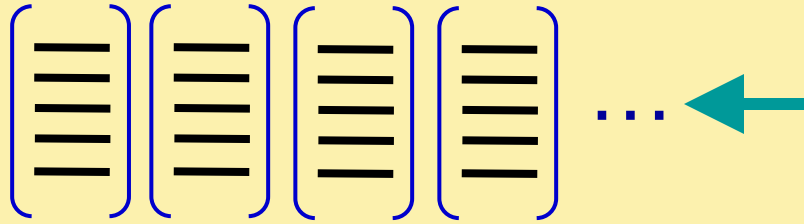
Detect patches

[Mikojczyk and Schmid '02]

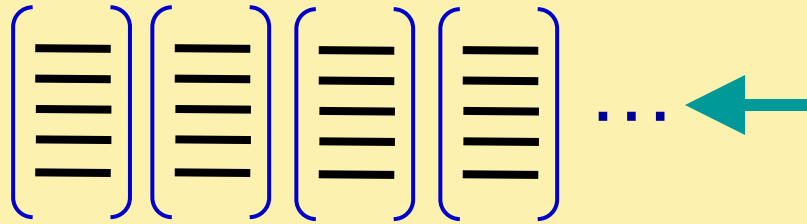
[Matas et al. '02]

[Sivic et al. '03]

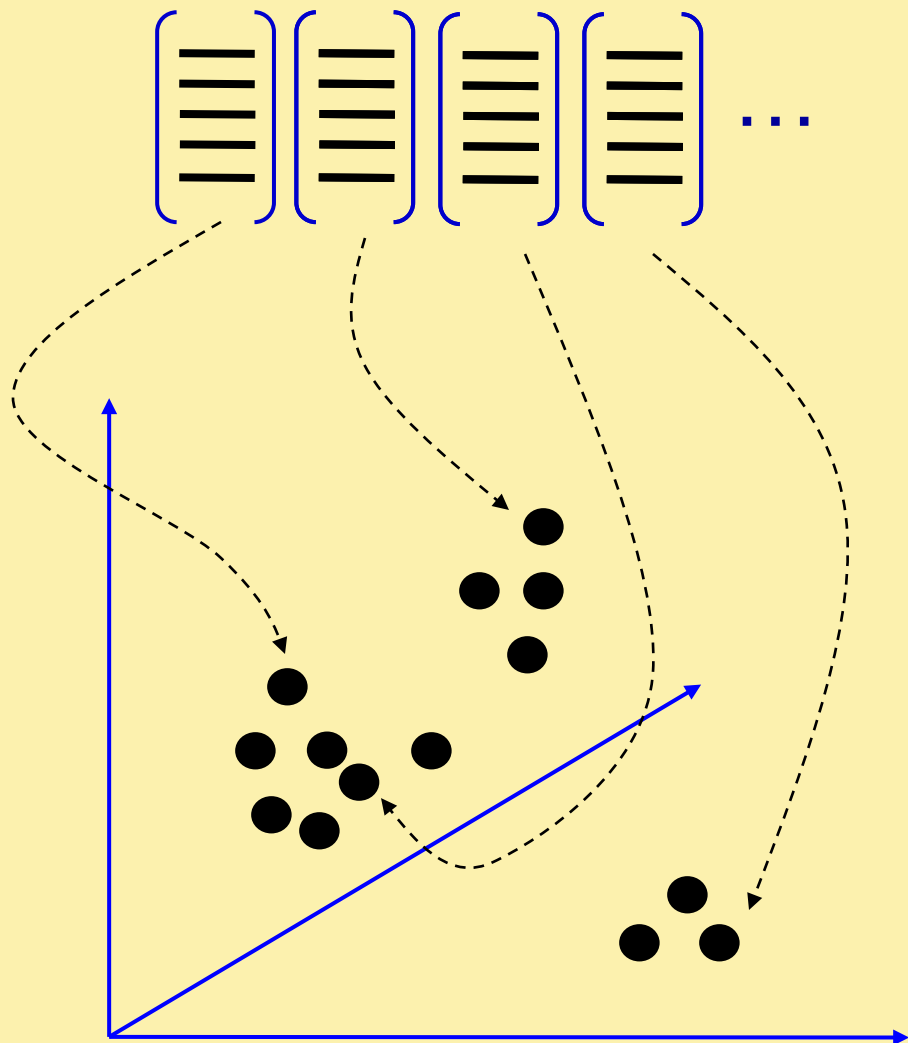
Interest Point Features



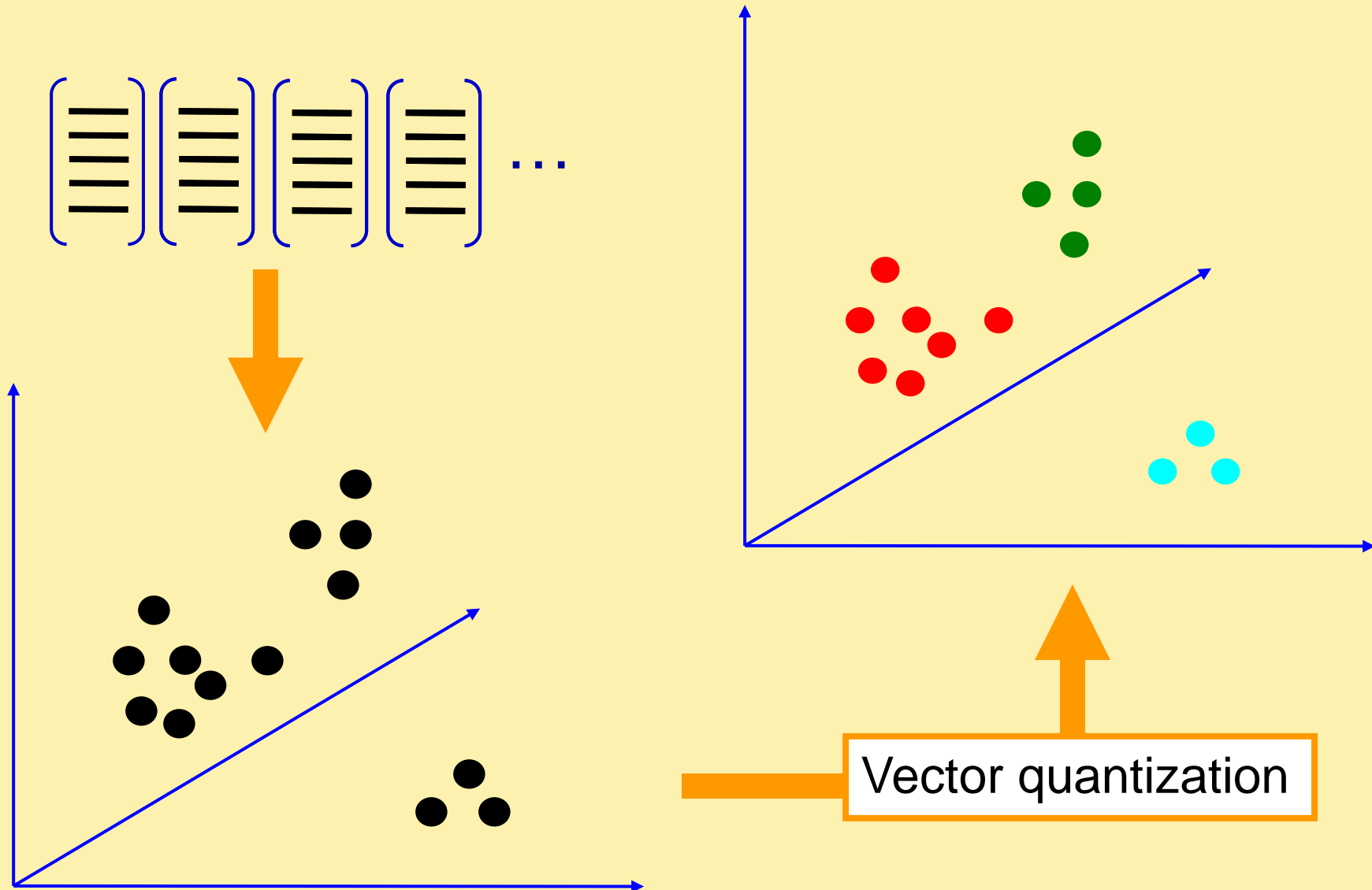
Patch Features



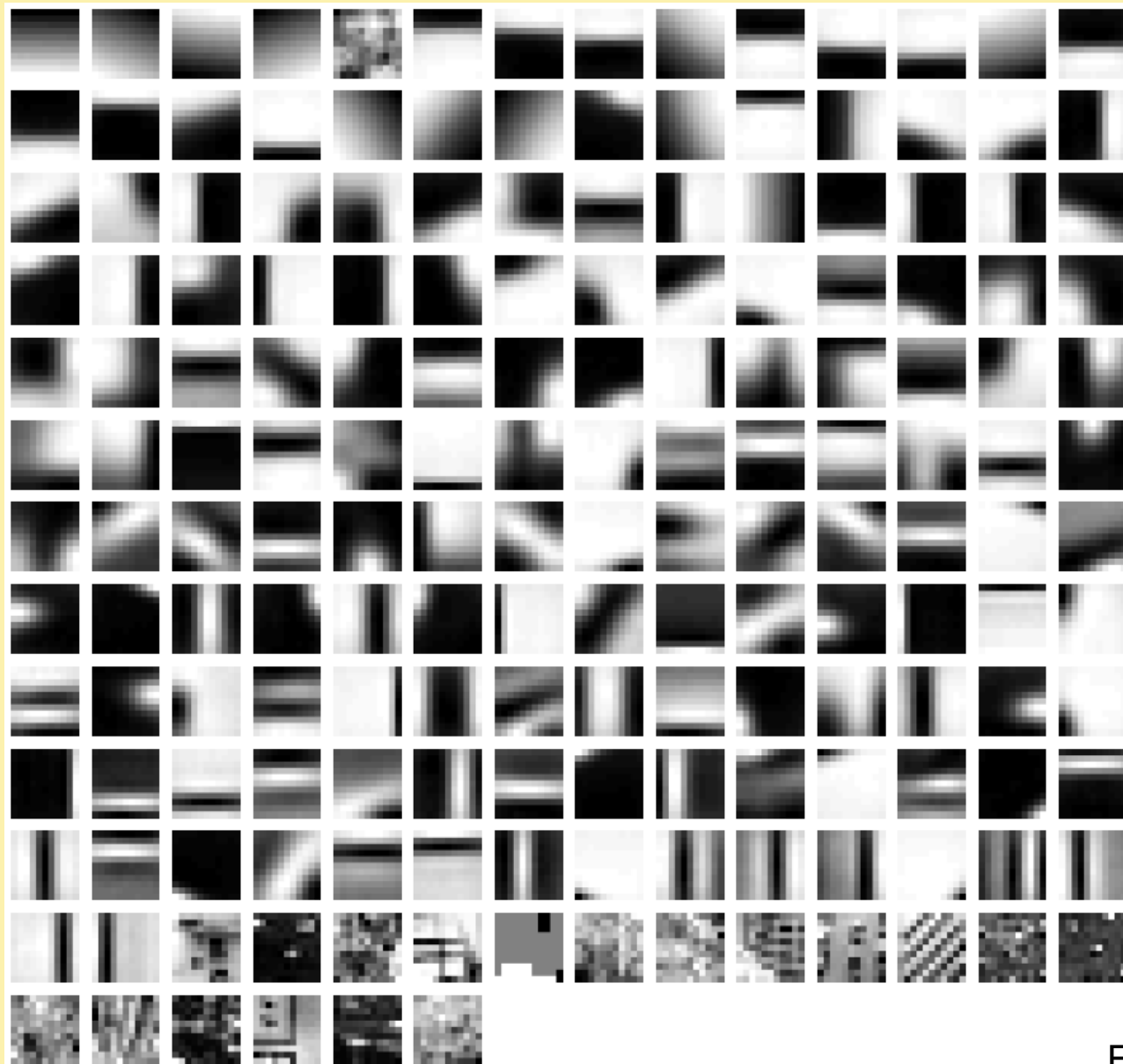
dictionary formation



Clustering (usually k-means)



Clustered Image Patches



Filterbank

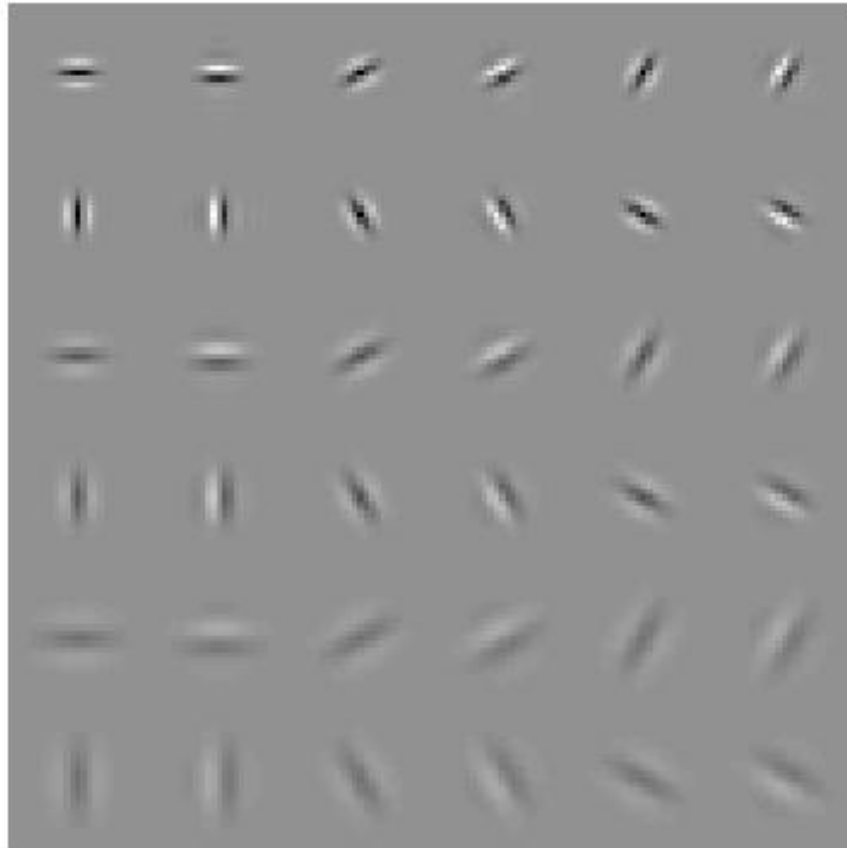
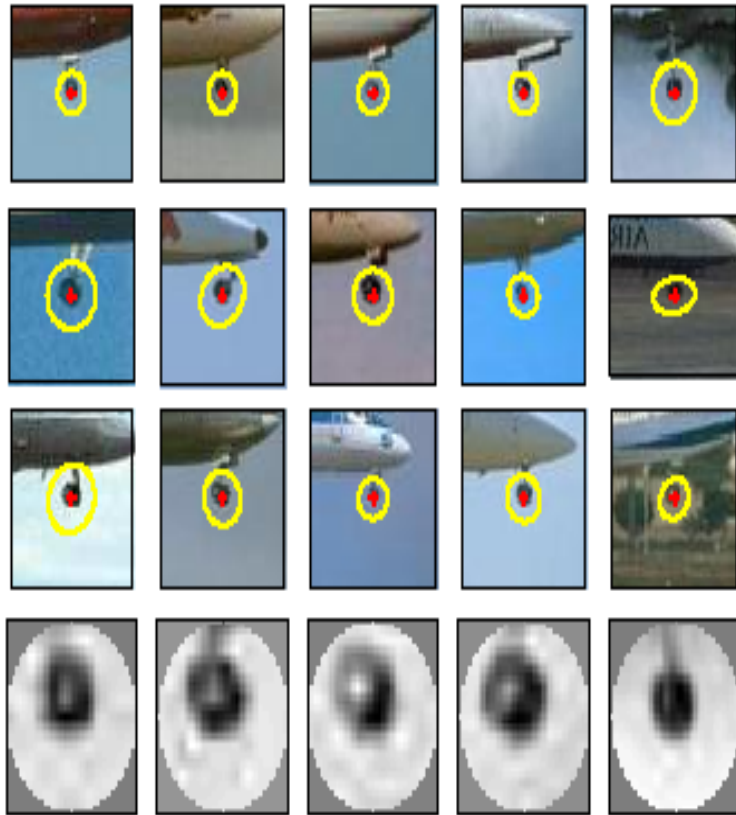
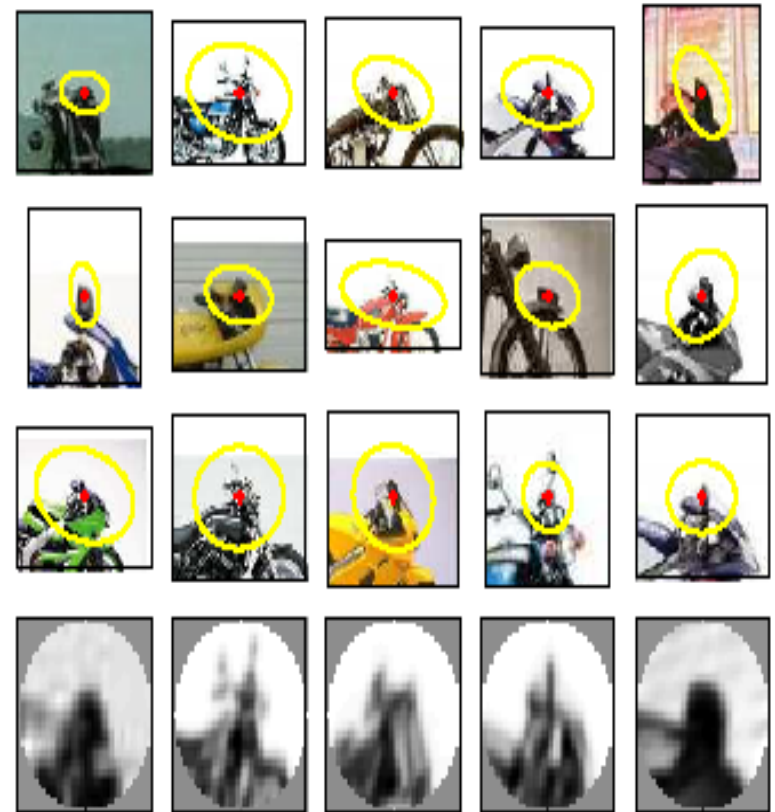


Image patch examples of codewords

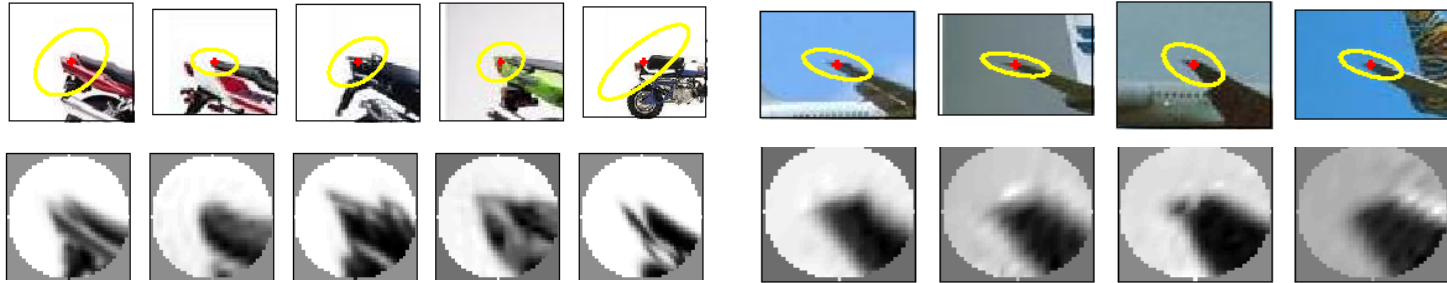
(a)



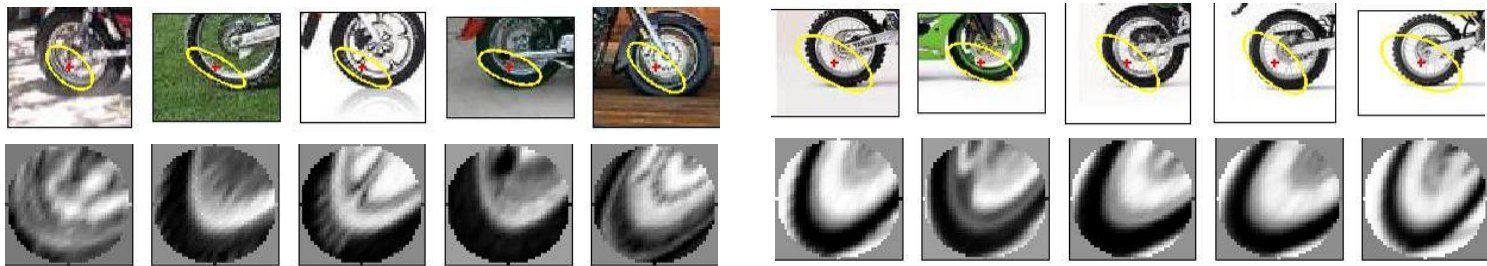
(b)



Visual synonyms and polysemy

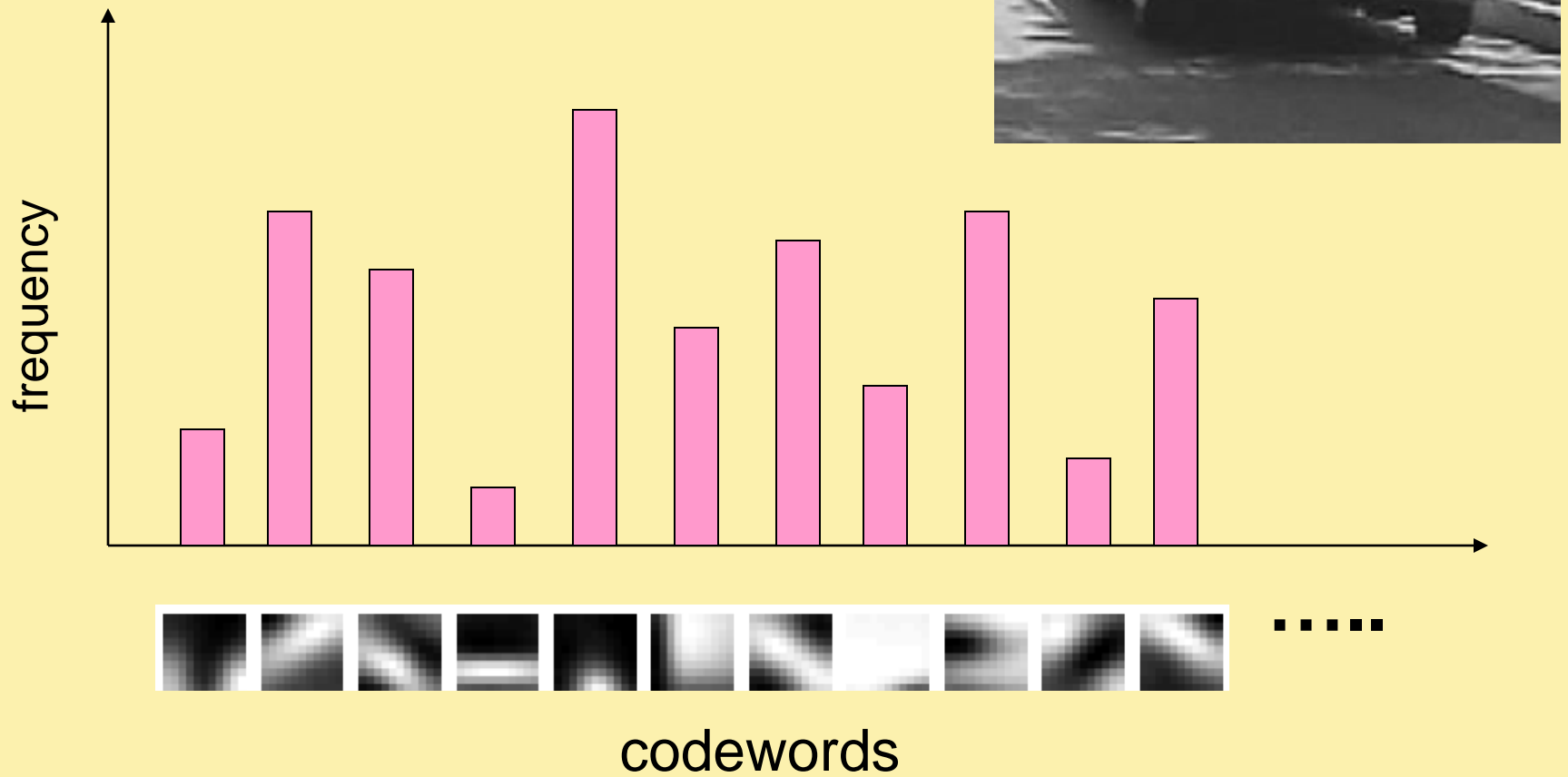


Visual Polysemy. Single visual word occurring on different (but locally similar) parts on different object categories.



Visual Synonyms. Two different visual words representing a similar part of an object (wheel of a motorbike).

Image representation



Scene Classification (Renninger & Malik)

beach



mountain



forest



city



street



farm



kitchen



livingroom



bedroom



bathroom



1. Bag of visual words model: recognizing object categories

Problem: Image Classification

Given:

- positive training images containing an object class, and



- negative training images that don't



Classify:

- a test image as to whether it contains the object class or not



?

Weakly-supervised learning

- Learn model from a set of training images containing object instances



- Know if image contains object or not
- But no segmentation of object or manual selection of features

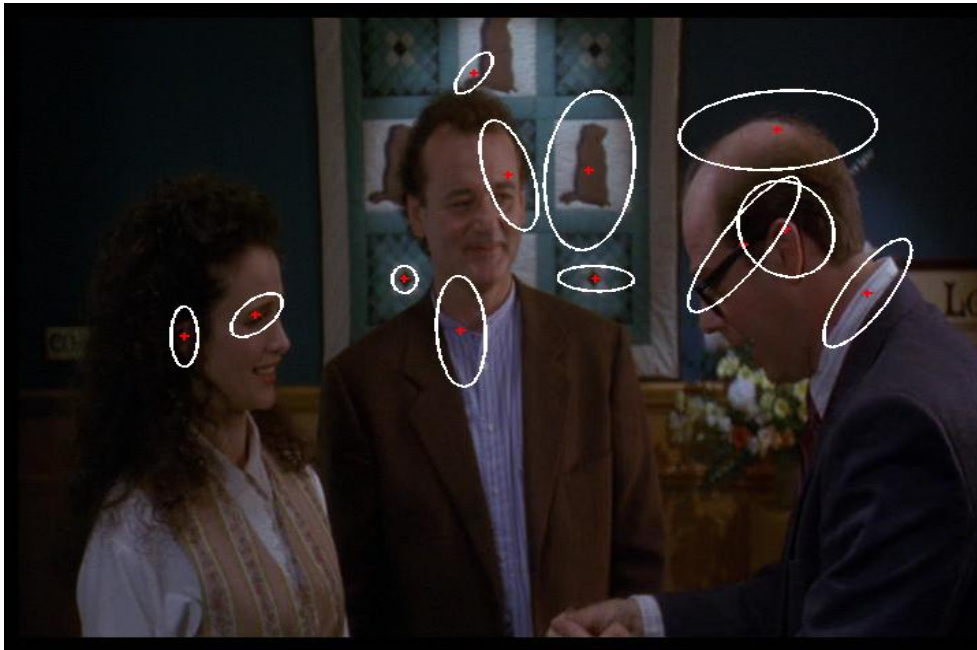
Three stages:

1. Represent each training image by a vector
 - Use a bag of visual words representation
2. Train a classifier to discriminate vectors corresponding to positive and negative training images
 - Use a Support Vector Machine (SVM) classifier
3. Apply the trained classifier to the test image

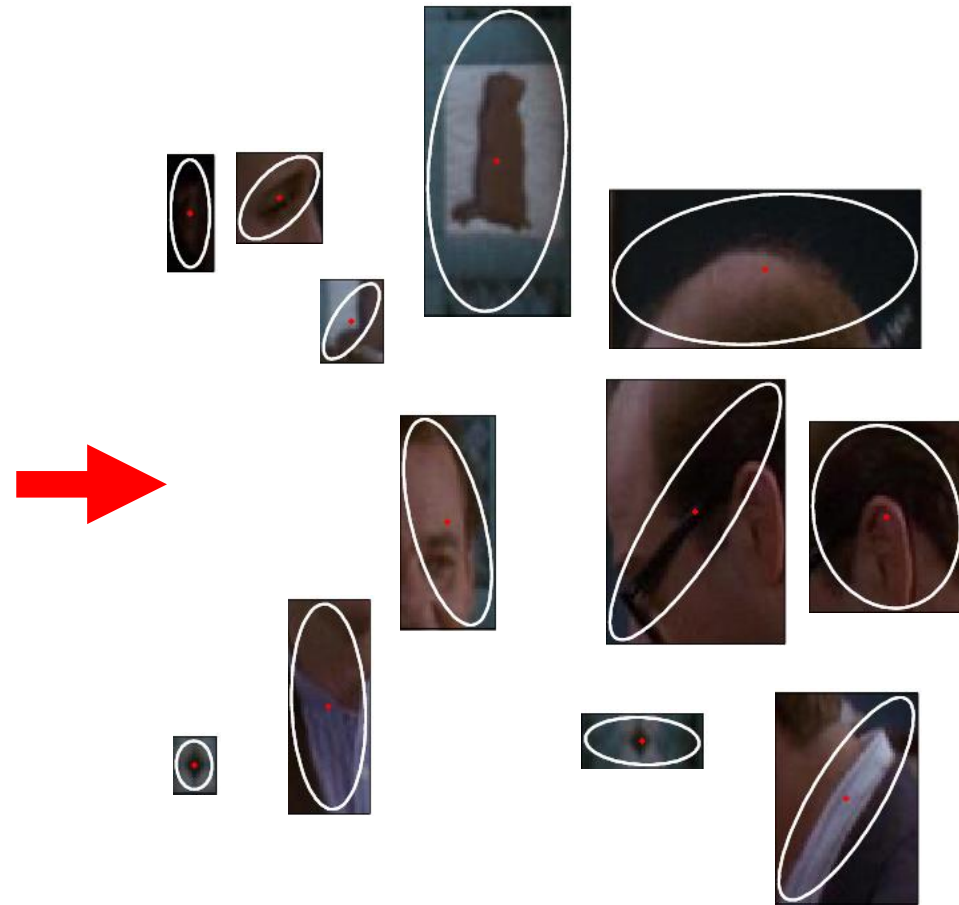
Representation: Bag of visual words

Visual words are 'iconic' image patches or fragments

- represent the frequency of word occurrence
- but not their position

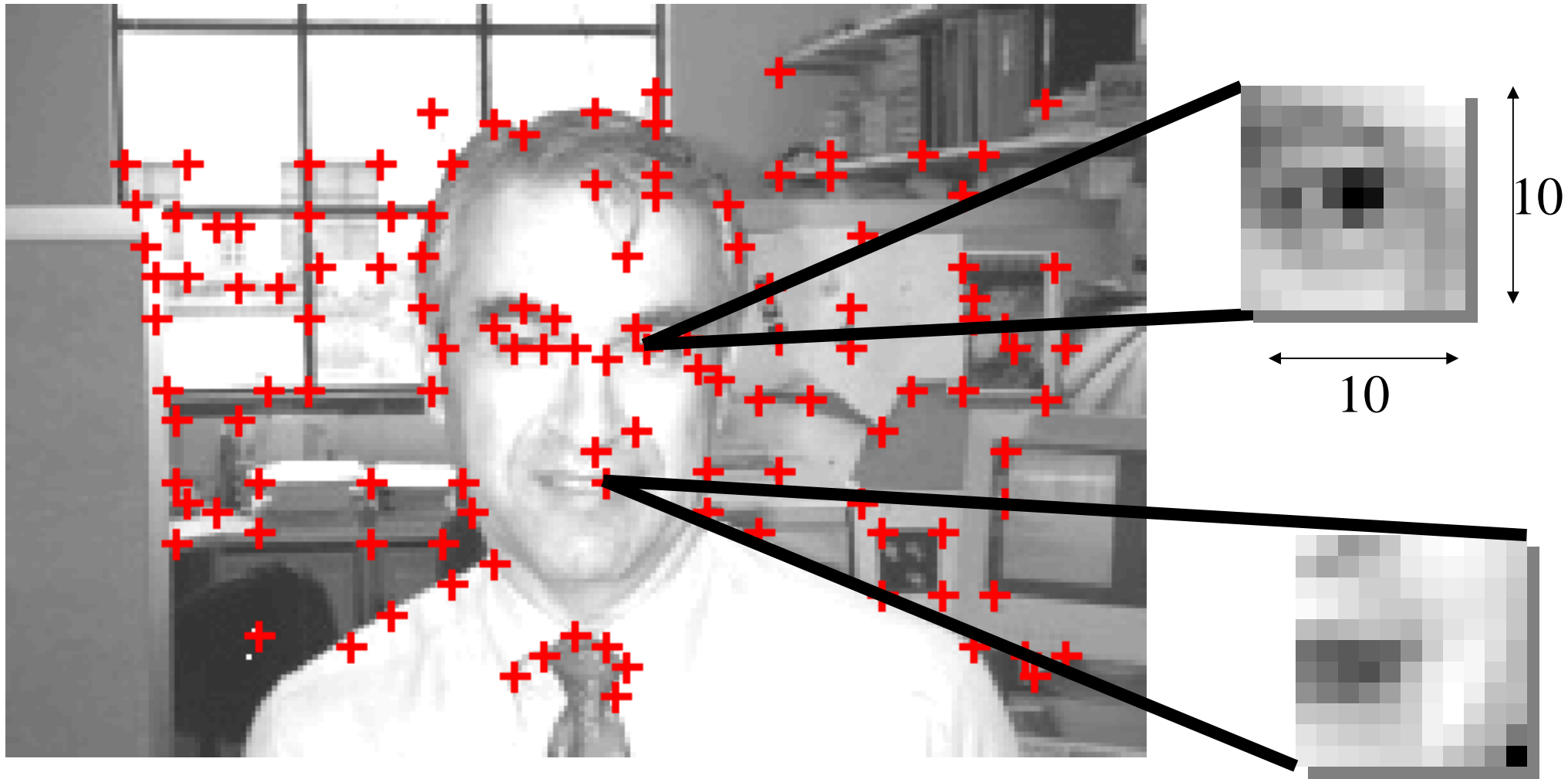


Image



Collection of visual words

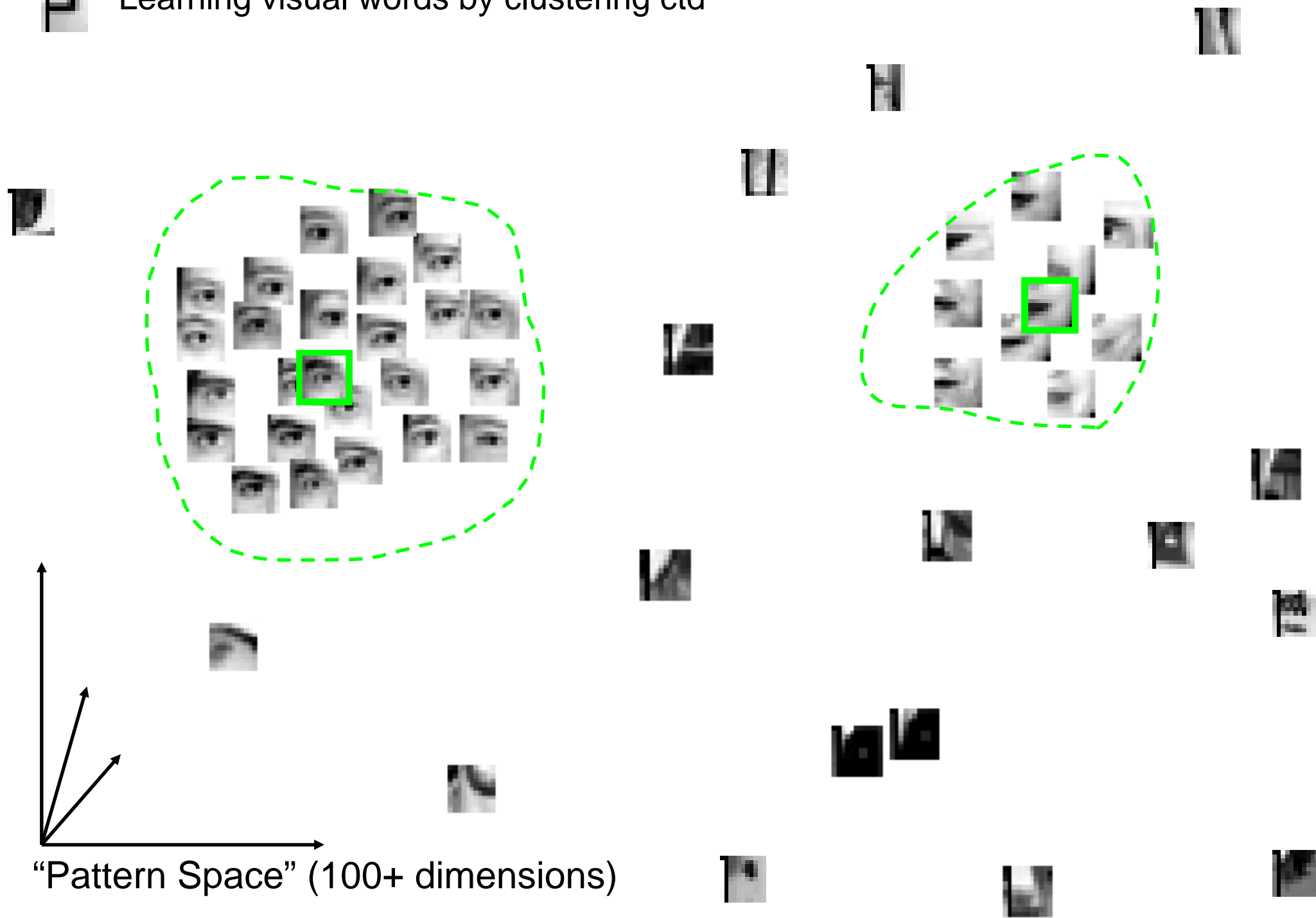
Example: Learn visual words by clustering



- Interest point features: textured neighborhoods are selected
- produces 100-1000 regions per image

Weber, Welling & Perona 2000

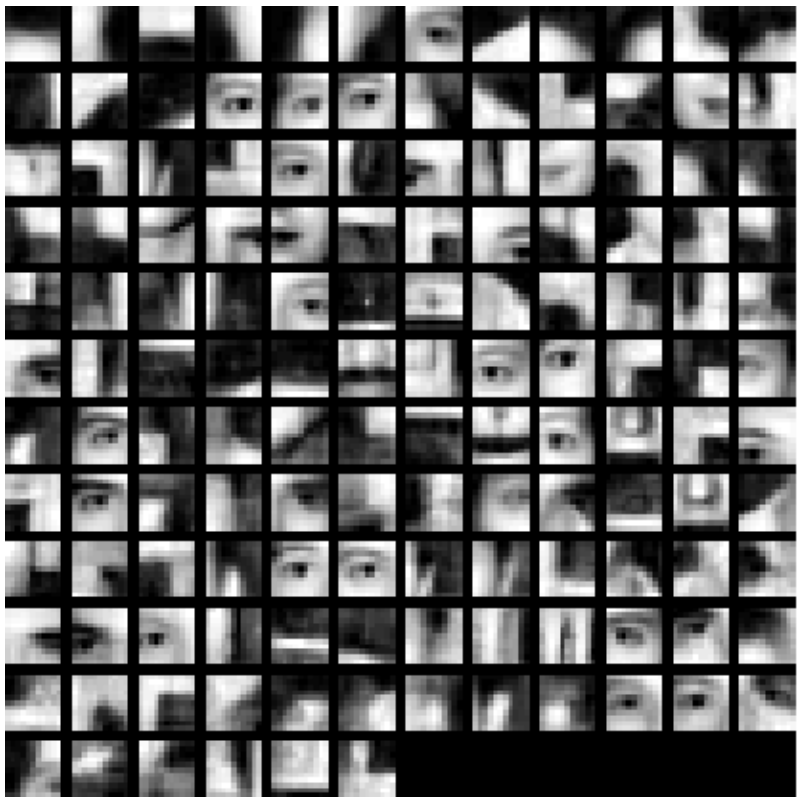
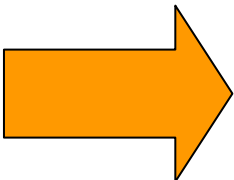
Learning visual words by clustering ctd



Example of visual words learnt by clustering faces



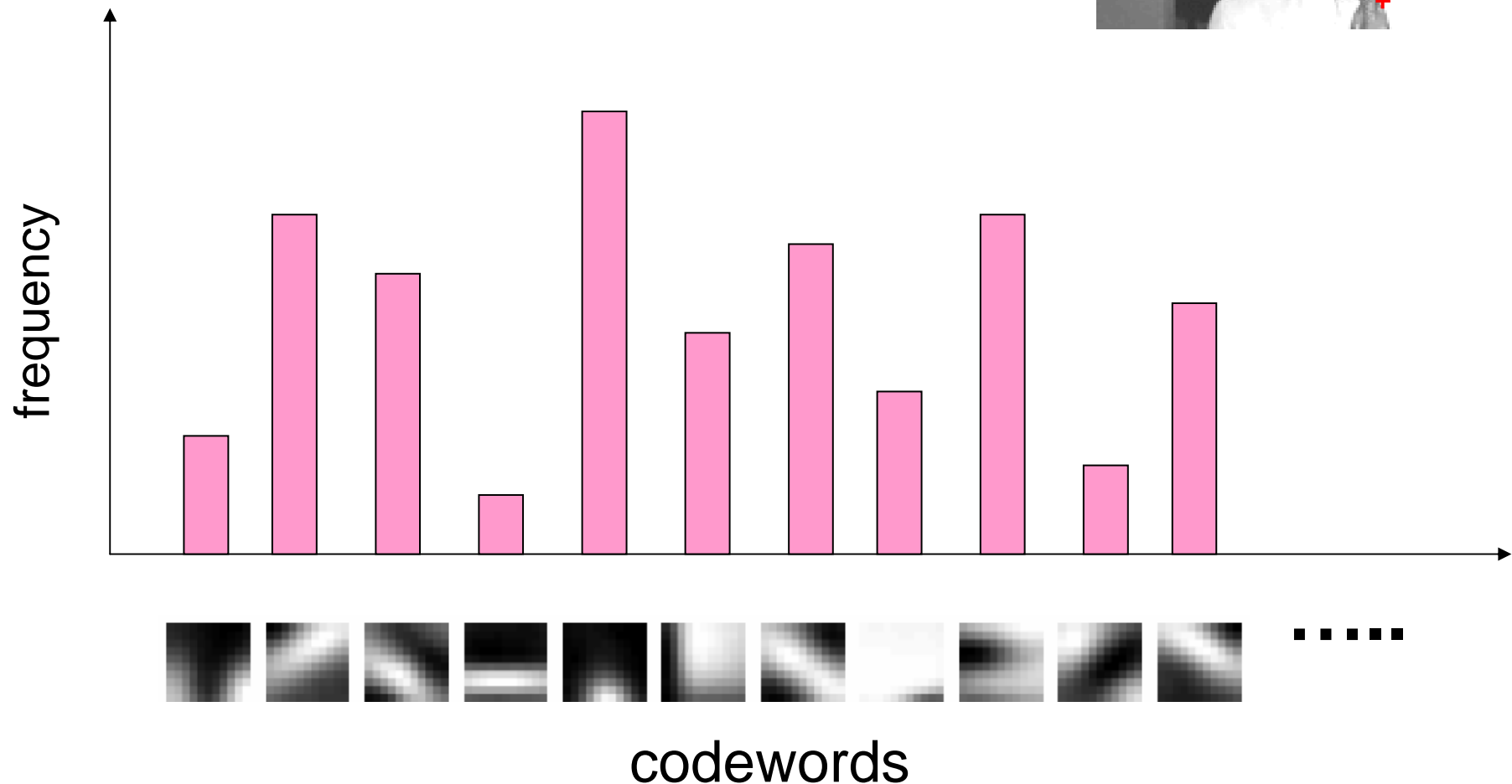
100-1000 images



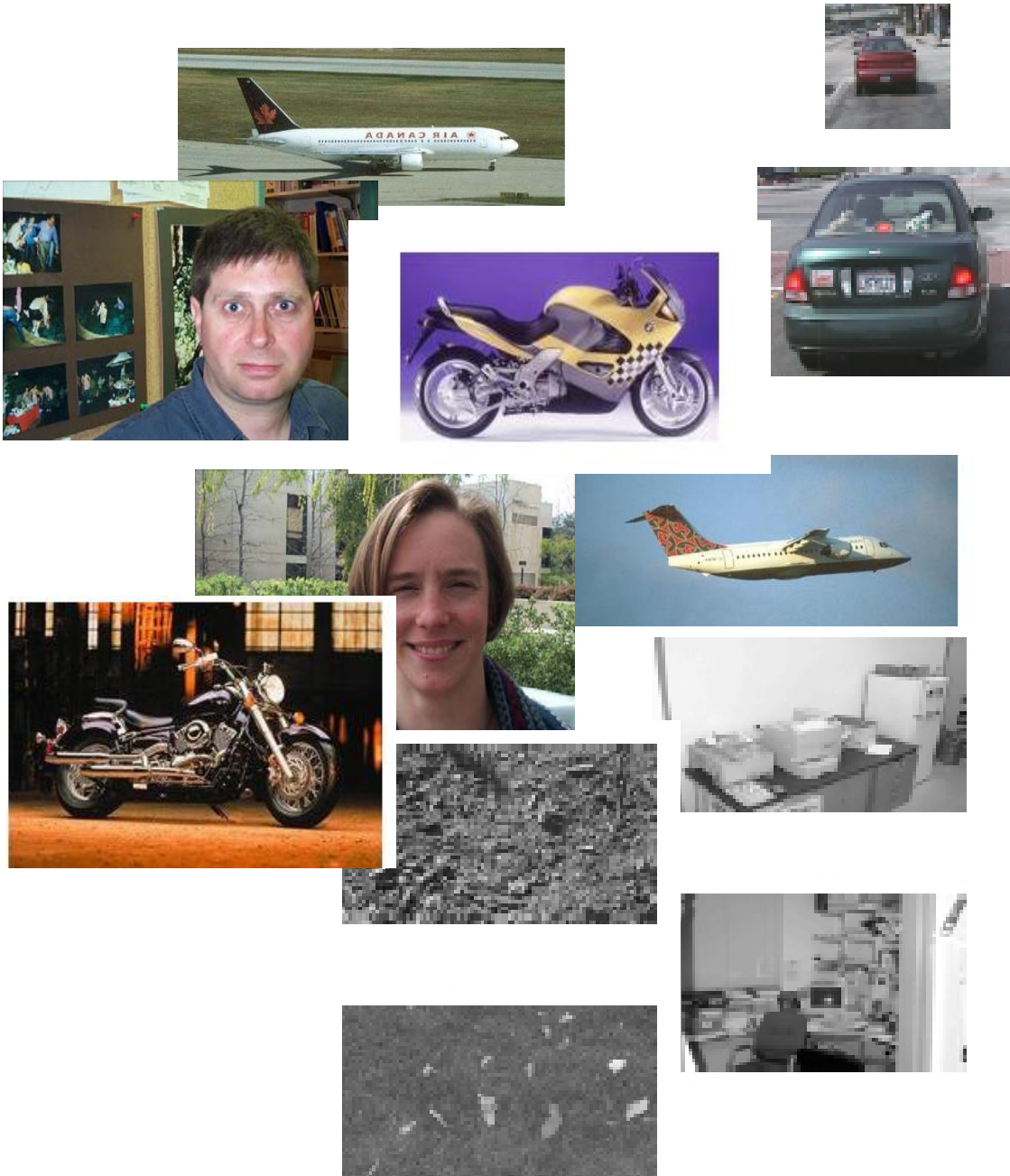
~100 visual words

Image representation – normalized histogram

- detect interest point features
- find closest visual word to region around detected points
- record number of occurrences, but not position



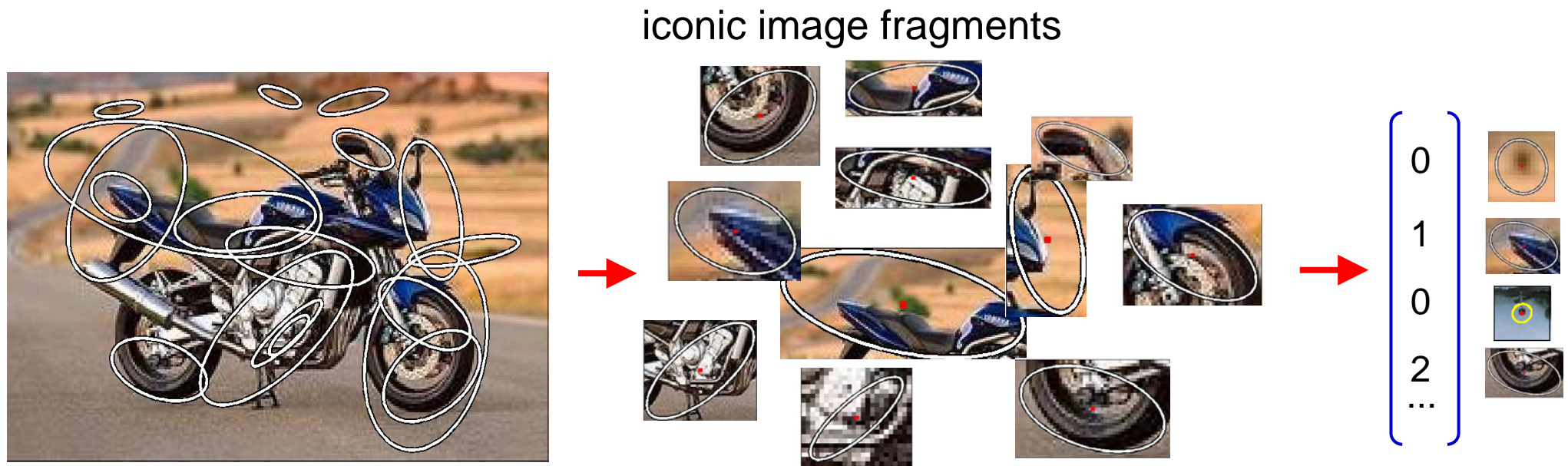
Example Image collection: four object classes + background



Faces	435
Motorbikes	800
Airplanes	800
Cars (rear)	1155
Background	900
Total:	4090

The “Caltech 5”

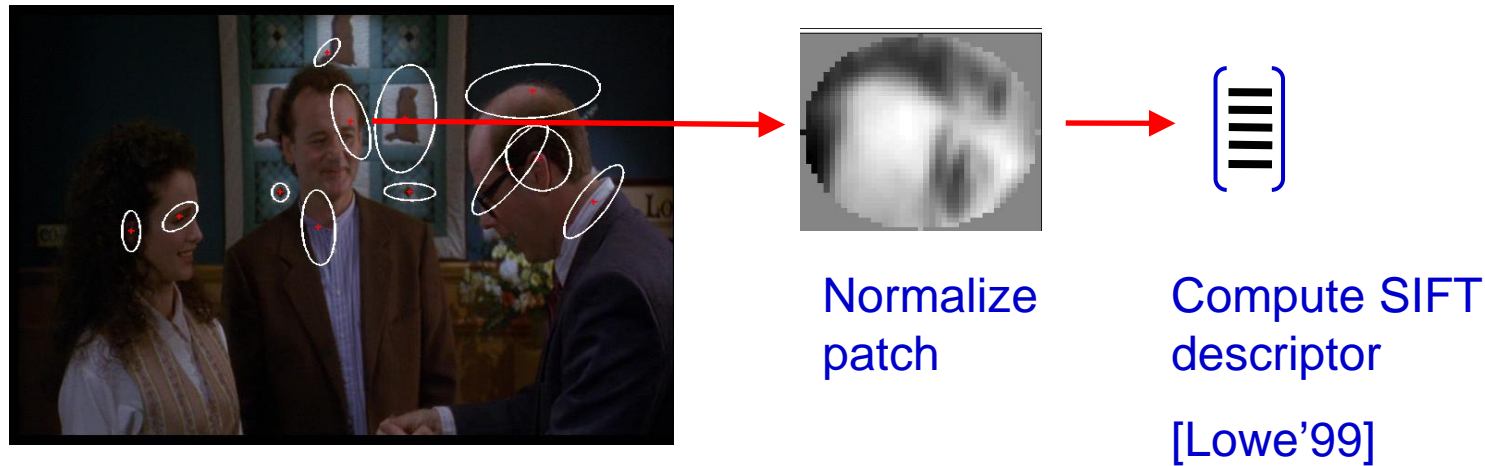
Represent an image as a histogram of visual words



Bag of words model

- Detect affine covariant regions
- Represent each region by a SIFT descriptor
- Build visual vocabulary by k-means clustering (K~1,000)
- Assign each region to the nearest cluster centre

Visual vocabulary for affine covariant patches

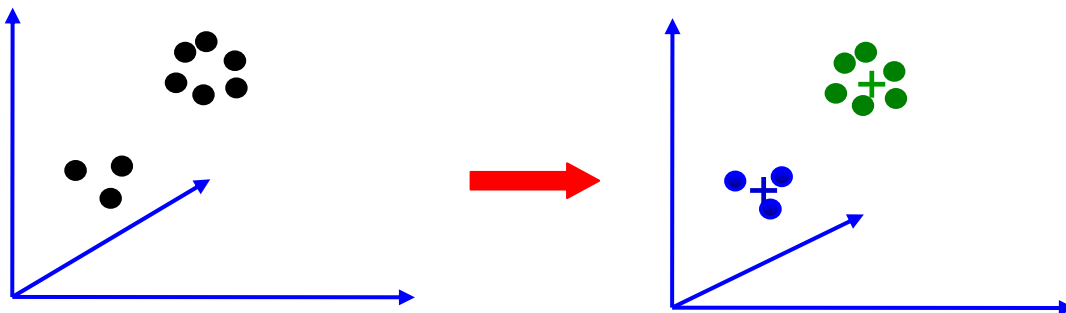


Detect patches

[Mikolajczyk and Schmid '02]

[Matas et al. '02]

Vector quantize descriptors from a set of training images using k-means



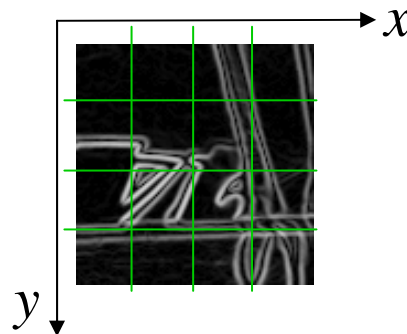
Descriptors – SIFT [Lowe'99]

distribution of the gradient over an image patch

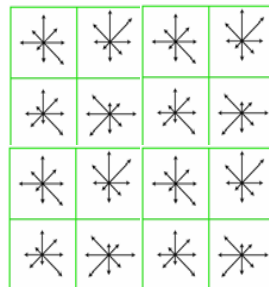
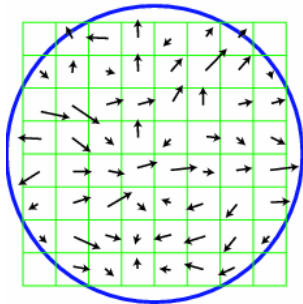
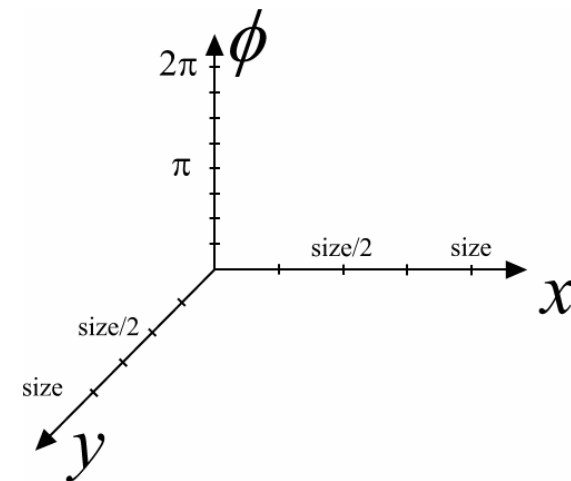
image patch



gradient



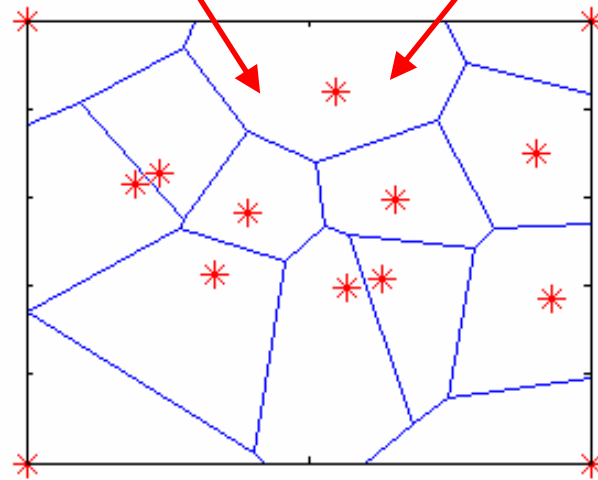
3D histogram



4x4 location grid and 8 orientations (128 dimensions)

very good performance in image matching [Mikolaczyk and Schmid'03]

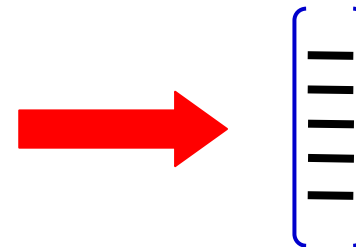
Vector quantize the descriptor space (SIFT)



The same visual word

Each image: assign all detections to their visual words

- gives bag of visual word representation
- normalized histogram of word frequencies
- also called 'bag of key points'



Visual words from affine covariant patches

Vector quantize SIFT descriptors to a vocabulary of iconic “visual words”.

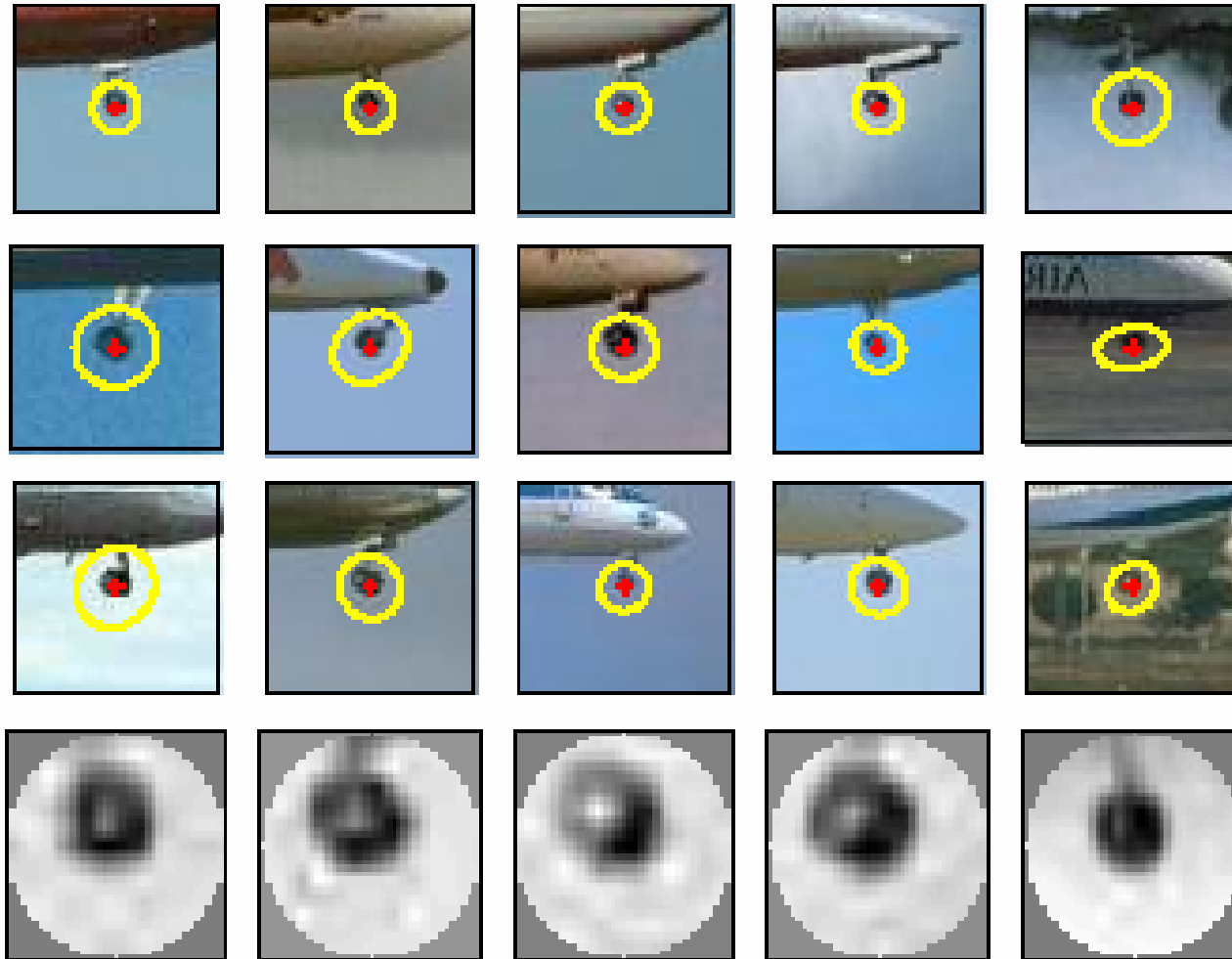
Design of descriptors makes these words invariant to:

- illumination
- affine transformations (viewpoint)

Size (granularity) of vocabulary is an important parameter

- fine grained – represent model instances
- coarse grained – represent object categories

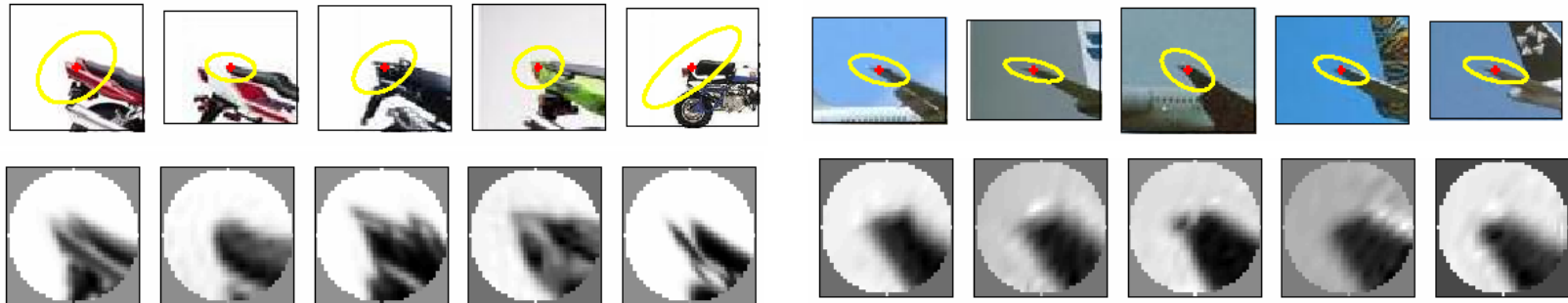
Examples of visual words



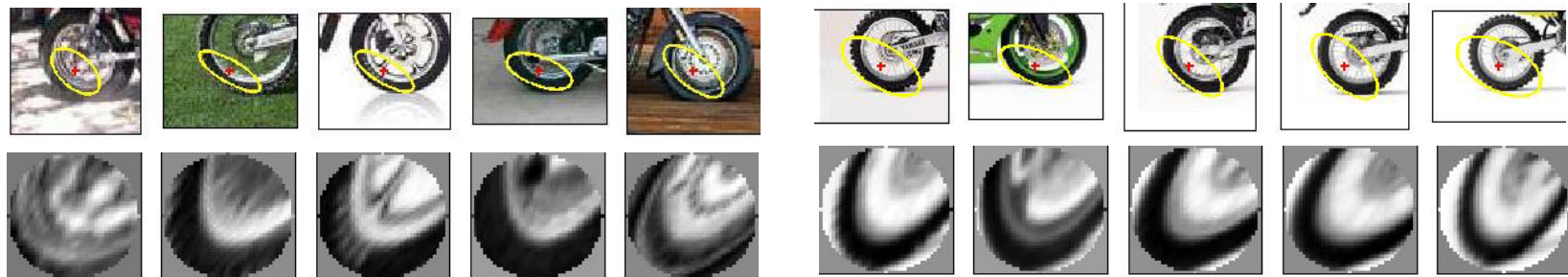
More visual words



Visual synonyms and polysemy



Visual Polysemy: Single visual word occurring on different (but locally similar) parts on different object categories.



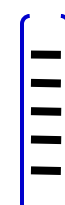
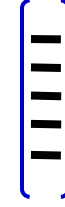
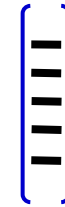
Visual Synonyms: Two different visual words representing a similar part of an object (wheel of a motorbike).

Training data: vectors are histograms, one from each training image

positive



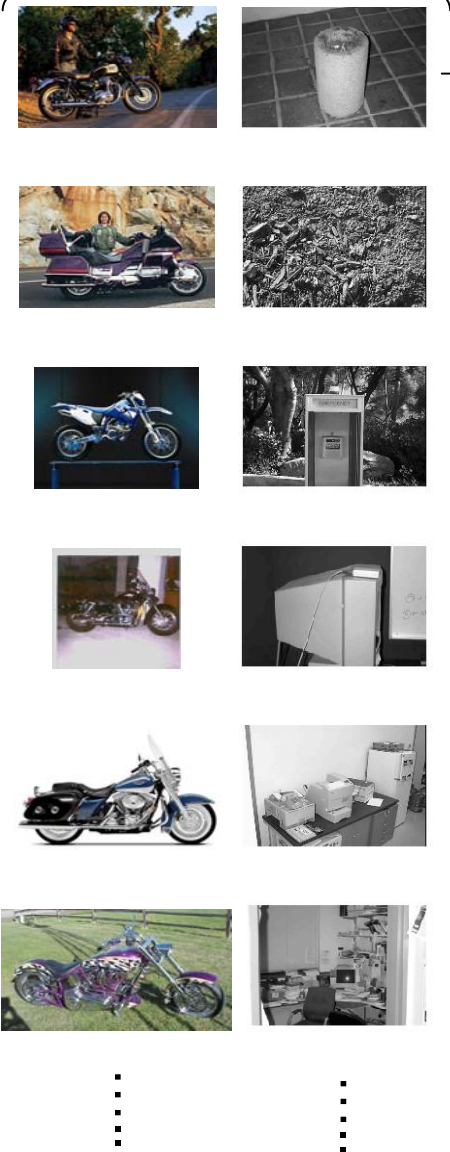
negative



Train classifier, e.g. SVM

Current Paradigm for learning an object category model

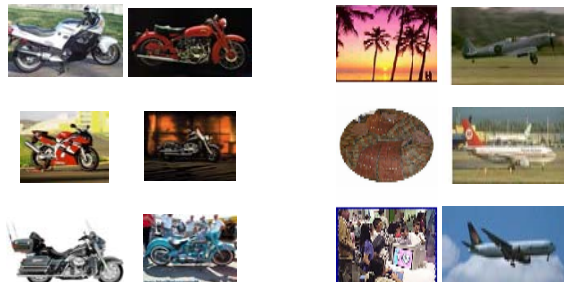
Manually gathered training images



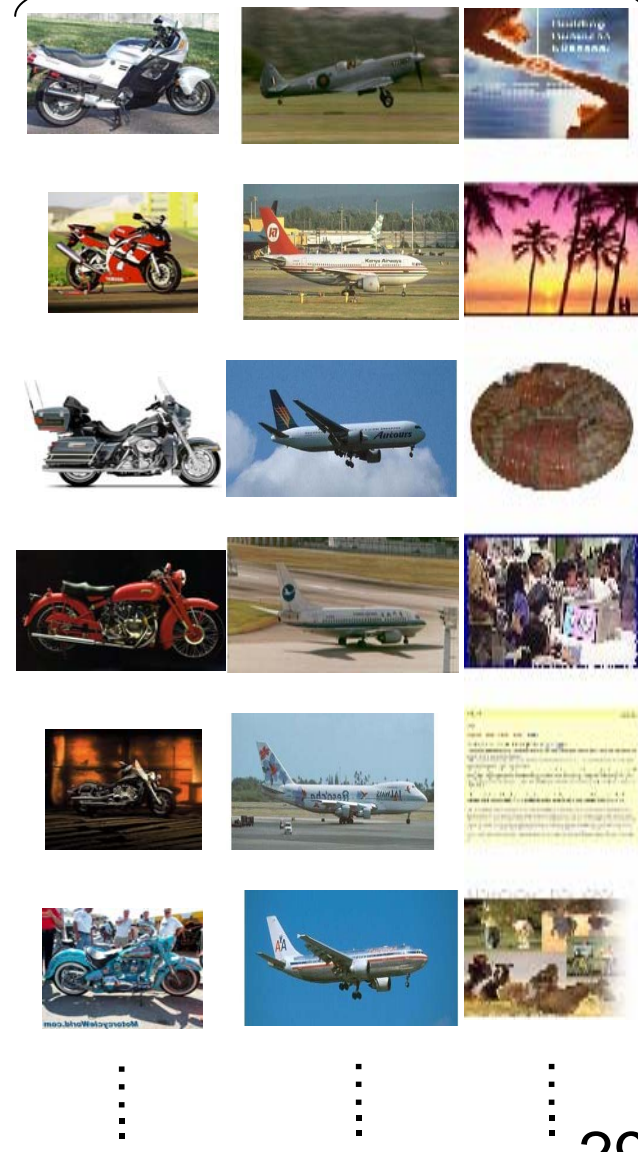
Visual words

Learn a visual category model

Evaluate classifier / detector



Test images



Example: weak supervision

Training

- 50% images
- No identification of object within image

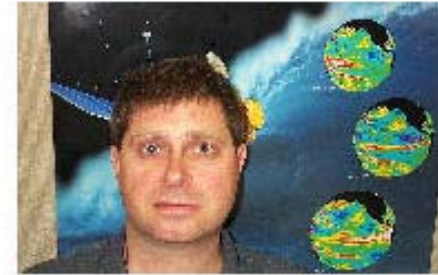
Motorbikes



Airplanes



Frontal Faces



Testing

- 50% images
- Simple object present/absent test

Cars (Rear)



Background



Learning

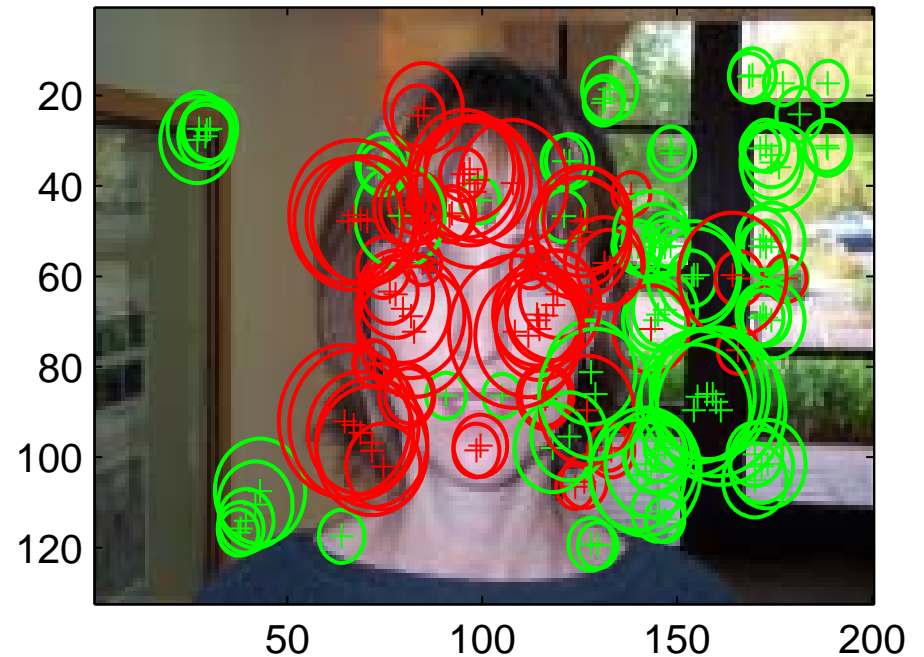
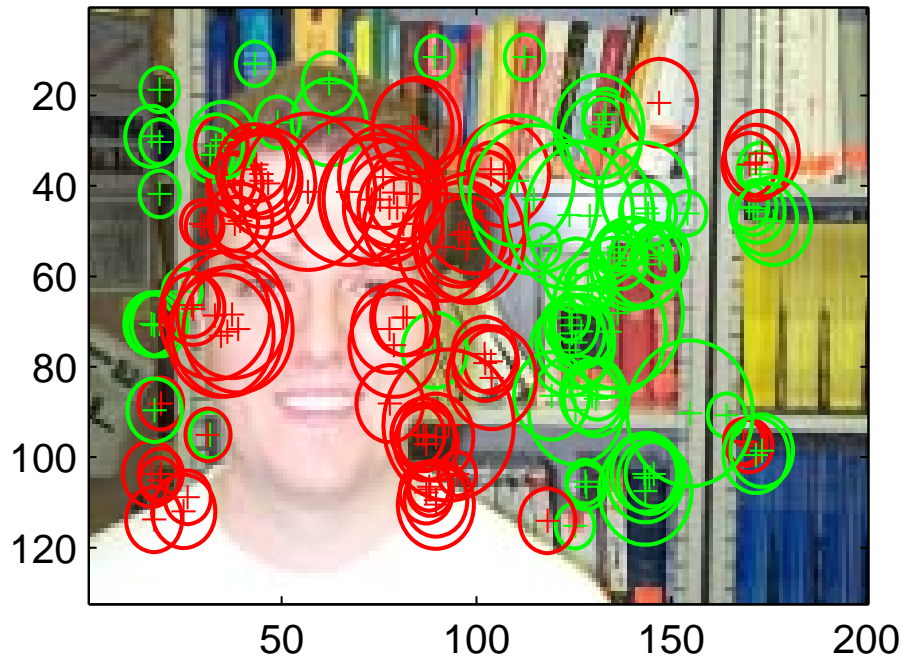
- SVM classifier
- Gaussian kernel using χ^2 as distance between histograms



Result

- Between 98.3 – 100% correct, depending on class

Localization according to visual word probability

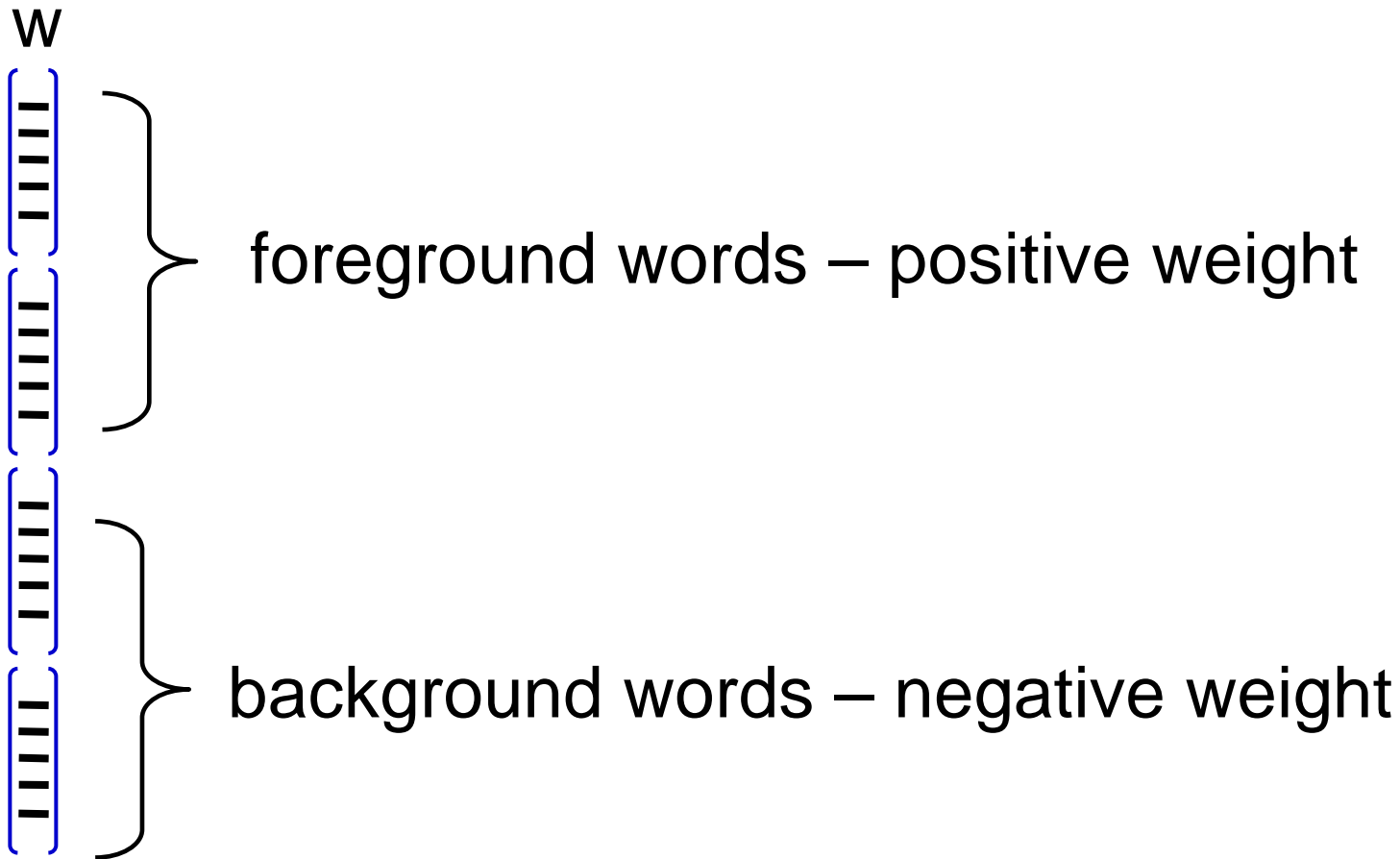
sparse segmentation



-  foreground word more probable
-  background word more probable

Why does SVM learning work?

- Learns foreground and background visual words



Bag of visual words summary

- Advantages:

- largely unaffected by position and orientation of object in image
- fixed length vector irrespective of number of detections
- Very successful in classifying images according to the objects they contain
- Still requires further testing for large changes in scale and viewpoint

- Disadvantages:

- No explicit use of configuration of visual word positions
- Poor at **localizing** objects within an image