CS231M · Mobile Computer Vision

Announcements

- P2 was released yesterday
- It is now due on May 8th (instead of May 6th)
- Please form your team by Wednesday this week;

CS231M · Mobile Computer Vision

Lecture 9

Recognition

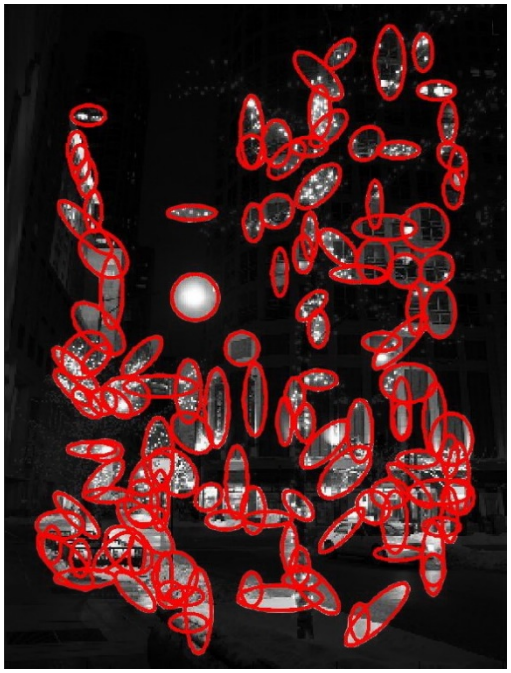
- Classification
- Detection
- Single instance detection and localization

Forsyth, Ponce “Computer vision: a modern approach”:
- Chapter 16, Sec. 16.1
- Chapters 6 (sec. 6.2)

Szeliski, “Computer Vision: algorithms and applications"
- Chapter 14, Sec. 14.1, Sec. 14.3, sec. 14.4

Several slides in this lecture are credit from L. Fei-Fei, R. Fergus, and A. Torralba. "Recognizing and Learning Object Categories, CVPR 2007 short course".

From low level to high level vision



Feature Detection

e.g. DoG

Feature Description

e.g. SIFT

- Estimation
- Matching
- Classification
- Detection

Classification or indexing

Is this an image of a bridge?

Image search engines

Google Image Search™

Picasa™

flickr™ GAMMA

webshots beta

bing

YouTube Broadcast Yourself™

Incogna

LTU technologies LTU

picsearch™

YAHOO!®

Detection

Does this image contain a bridge? [where?]

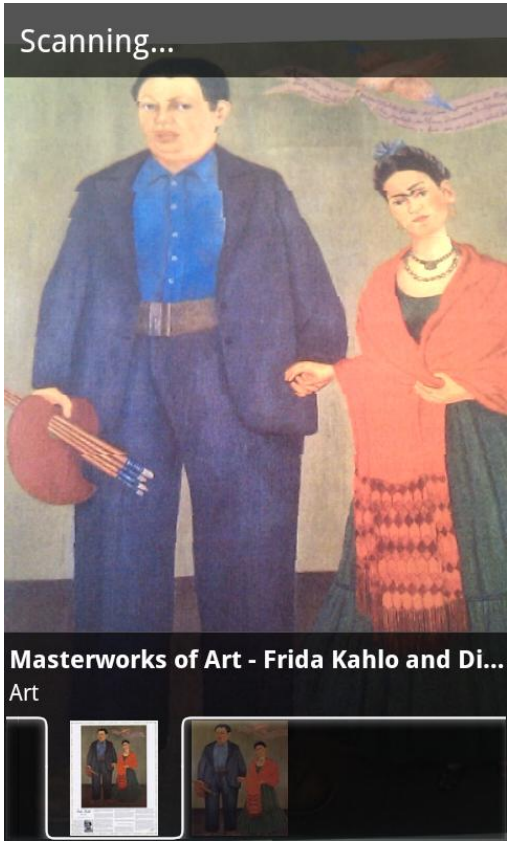
Face detection

Human body detection and gesture recognition

Single instance detection

Does this image contain the golden gate bridge? [where?]
Or which landmark does this image contain?

Visual search and landmarks recognition



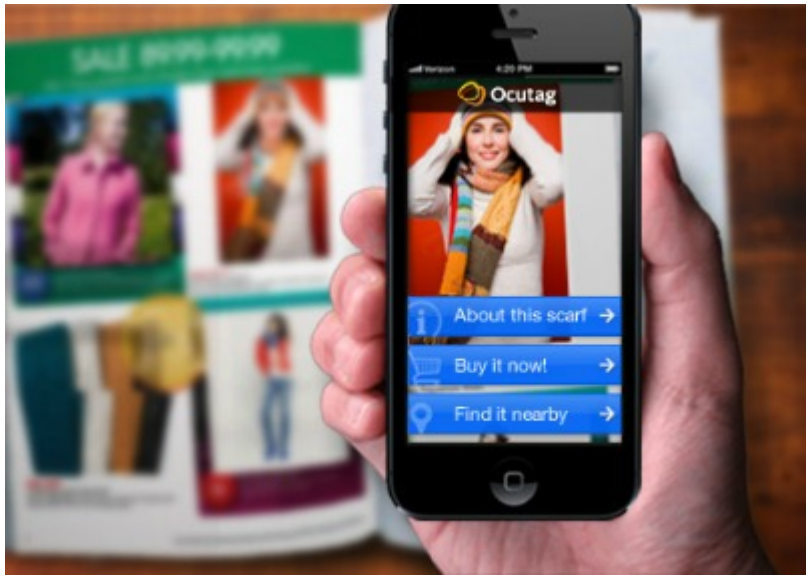
Google Goggles

Fingerprint identification

digitalPersona.

Face identifcation

Visual search and landmarks recognition



RICOH

A9

Challenges: illumination

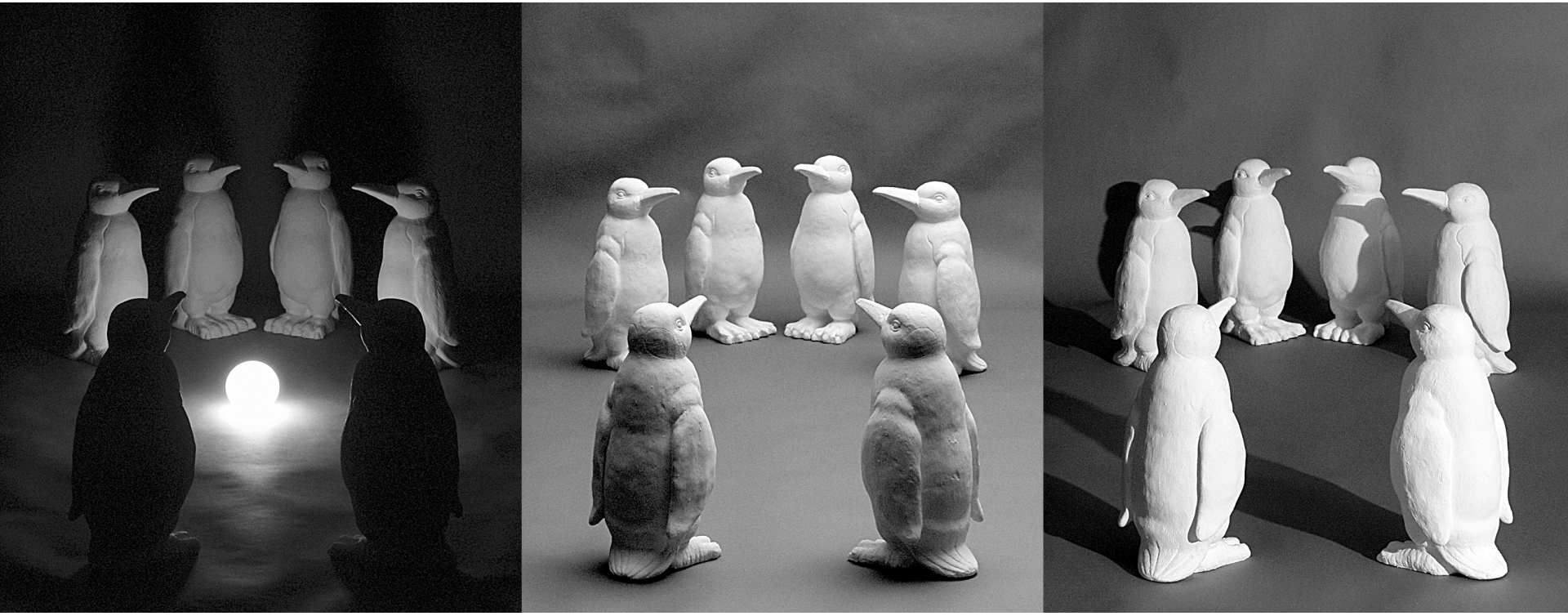


image credit: J. Koenderink

Challenges: scale

slide credit: Fei-Fei, Fergus & Torralba

Challenges: deformation

Challenges: occlusion

Magritte, 1957

slide credit: Fei-Fei, Fergus & Torralba

Challenges: background clutter

Kilmeny Niland. 1995

Challenges: viewpoint variation

Michelangelo 1475-1564

slide credit: Fei-Fei, Fergus & Torralba


~10,000 to 30,000

Challenges: intra-class variation

Recognition

- Classification
- Detection
- Single instance detection and localization

Classification or indexing

Is this an image of a bridge?

Object → Bag of 'words'

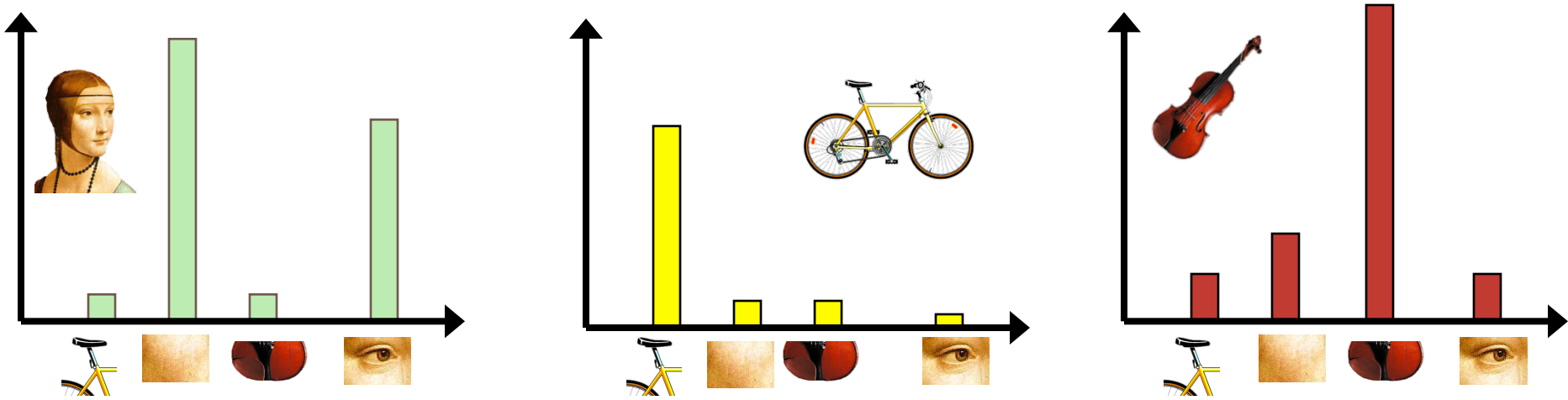
“Bag of Words” models

- Early “bag of words” models: mostly texture recognition
  - Cula & Dana, 2001; Leung & Malik 2001; Mori, Belongie & Malik, 2001; Schmid 2001; Varma & Zisserman, 2002, 2003; Lazebnik, Schmid & Ponce, 2003;
- Hierarchical Bayesian models for documents (pLSA, LDA, etc.)
  - Hoffman 1999; Blei, Ng & Jordan, 2004; Teh, Jordan, Beal & Blei, 2004
- Object categorization
  - Csurka, Bray, Dance & Fan, 2004; Sivic, Russell, Efros, Freeman & Zisserman, 2005; Sudderth, Torralba, Freeman & Willsky, 2005;
- Natural scene categorization
  - Vogel & Schiele, 2004; Fei-Fei & Perona, 2005; Bosch, Zisserman & Munoz, 2006

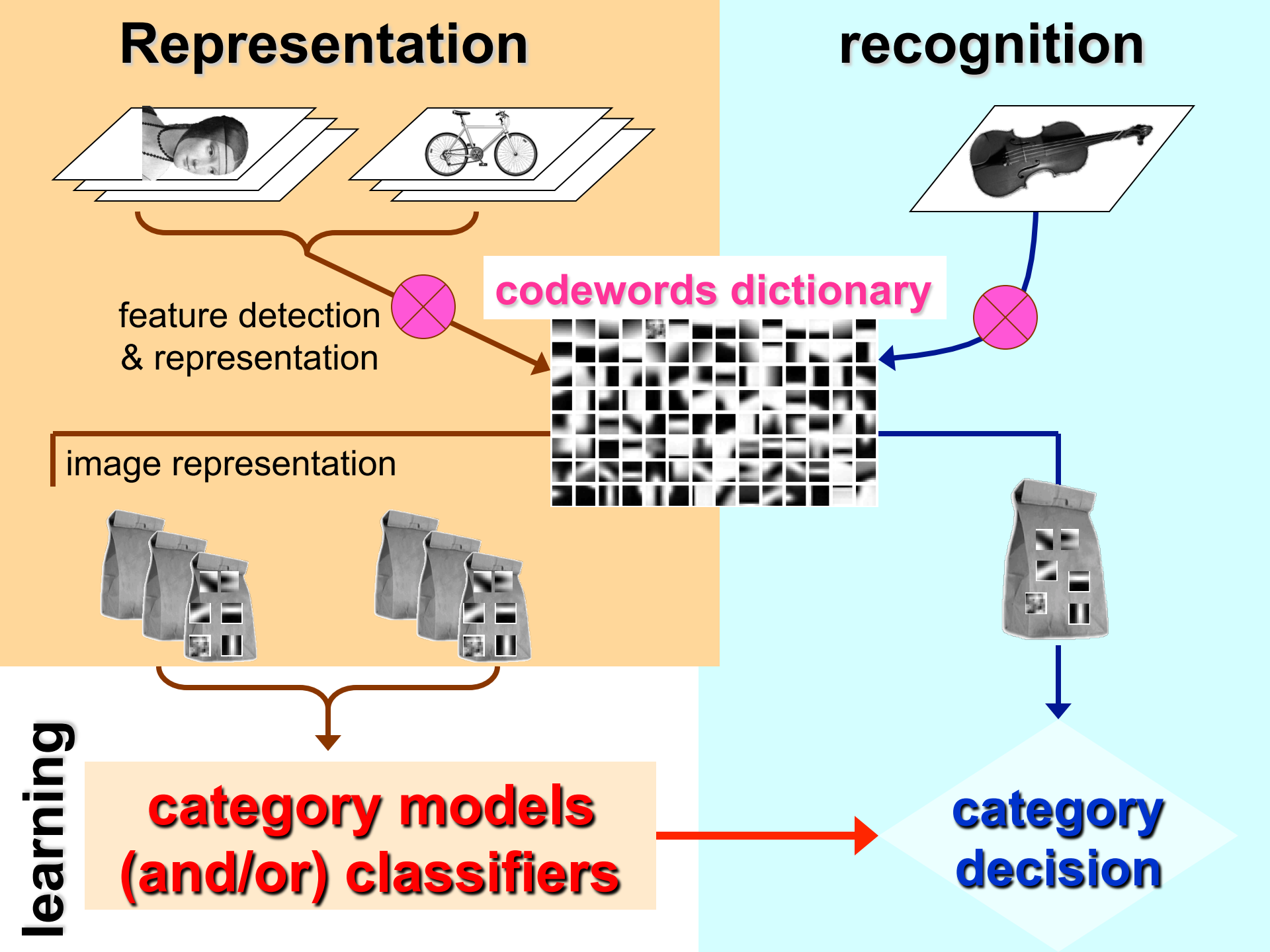
definition of “BoW”

– Independent features

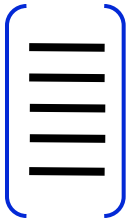
– histogram representation



codewords dictionary

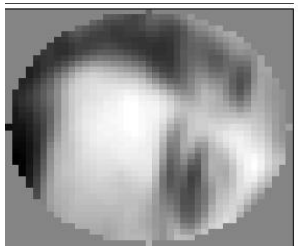

Representation
recognition
feature detection
& representation
codewords dictionary
image representation
learning
category models
(and/or) classifiers
category
decision

1.Feature detection and description

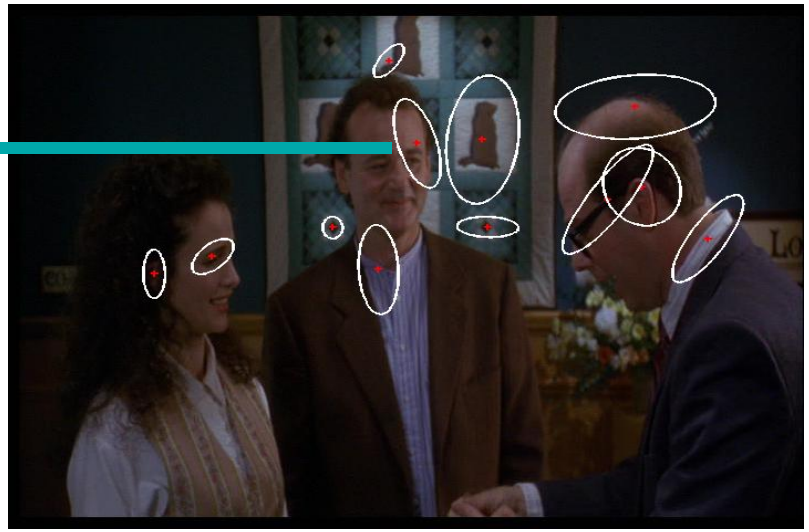


Compute SIFT descriptor

[Lowe'99]



Normalize patch



Detect patches

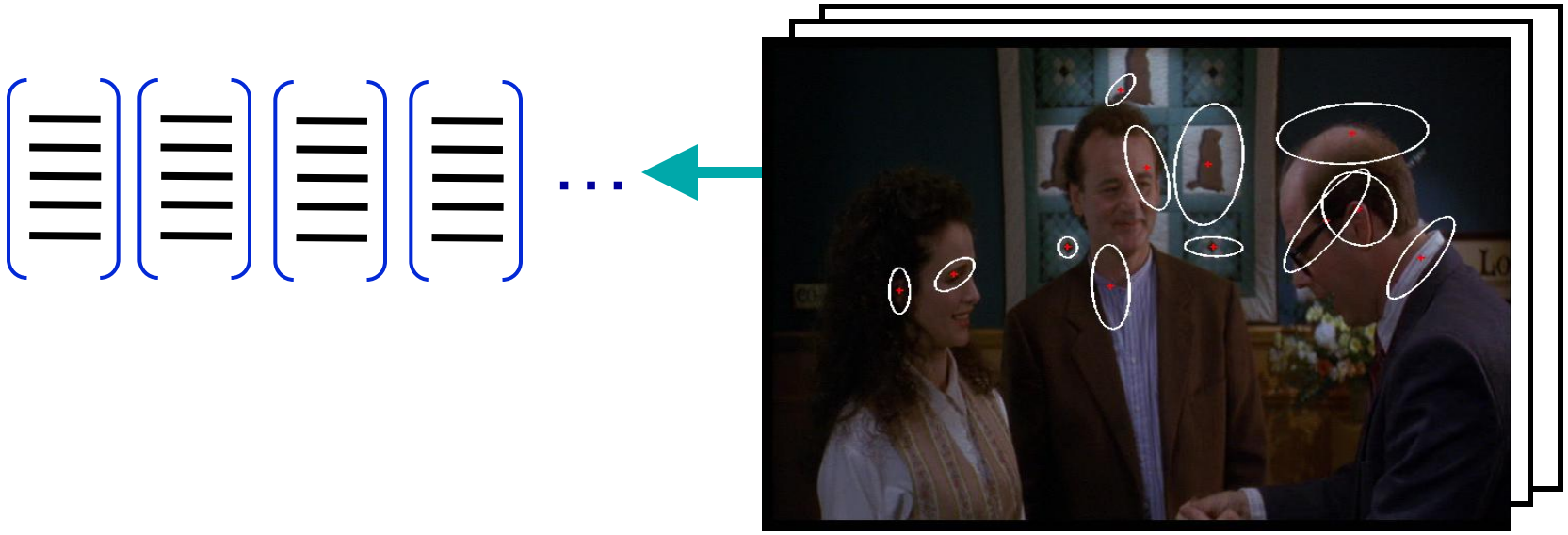
[Mikojaczyk and Schmid '02]

[Mata, Chum, Urban & Pajdla, '02]

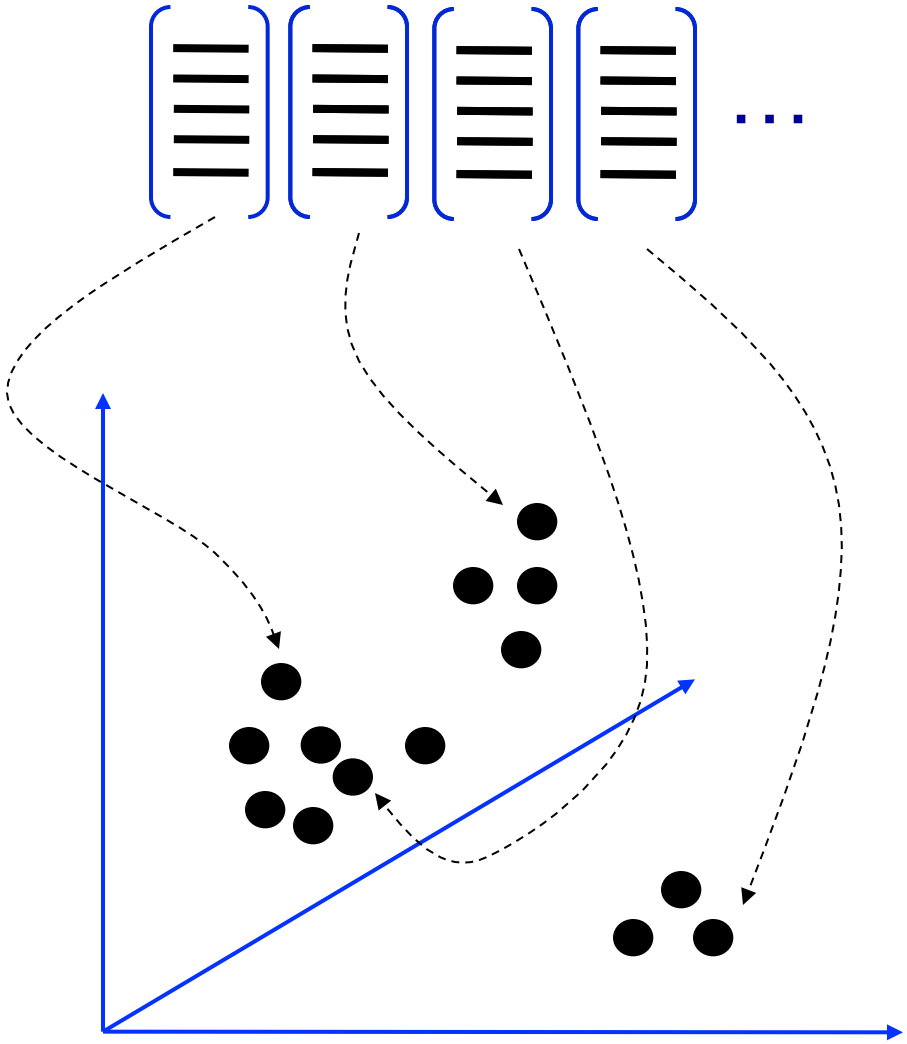
[Sivic & Zisserman, '03]

Slide credit: Josef Sivic

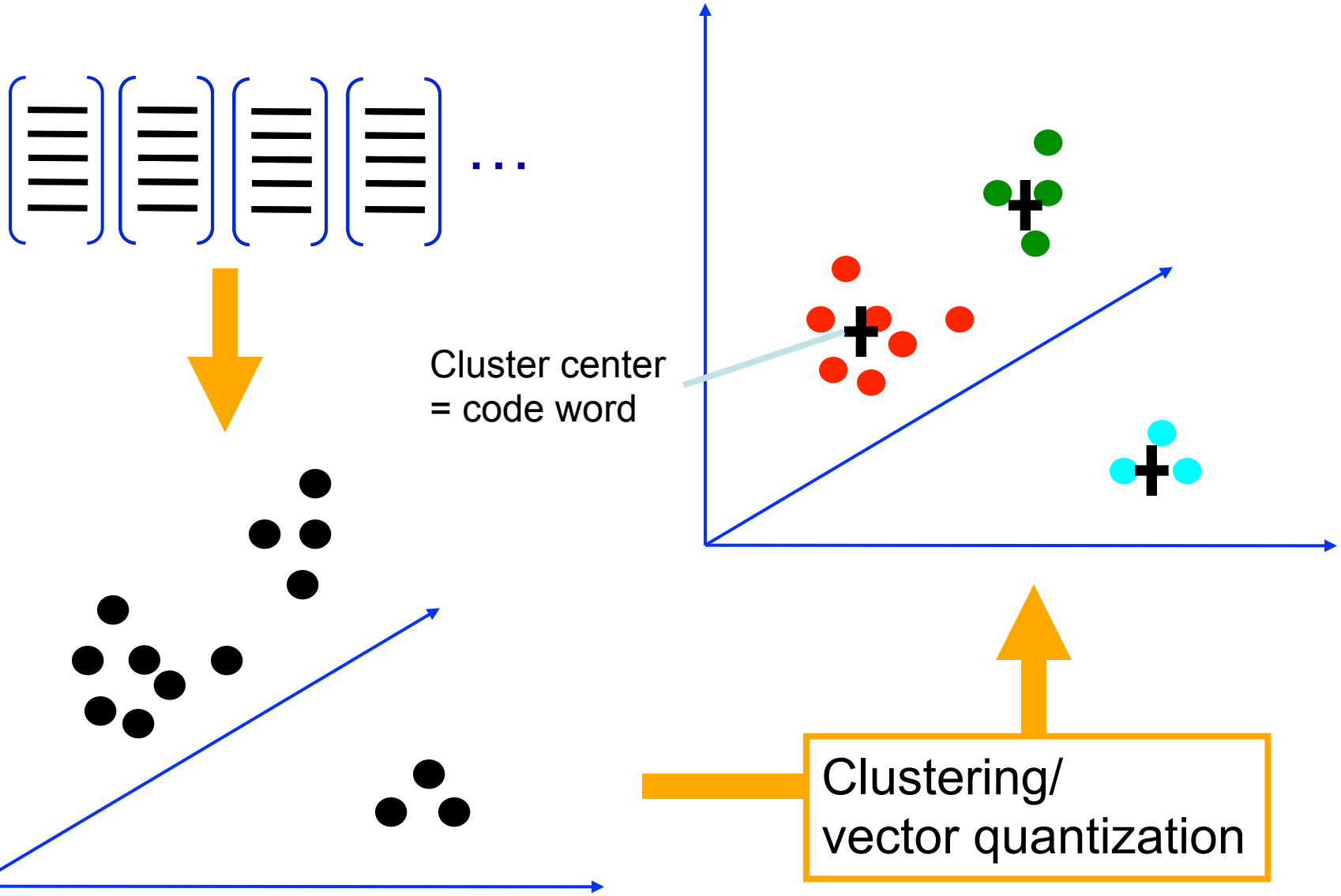
2. Codewords dictionary formation



2. Codewords dictionary formation

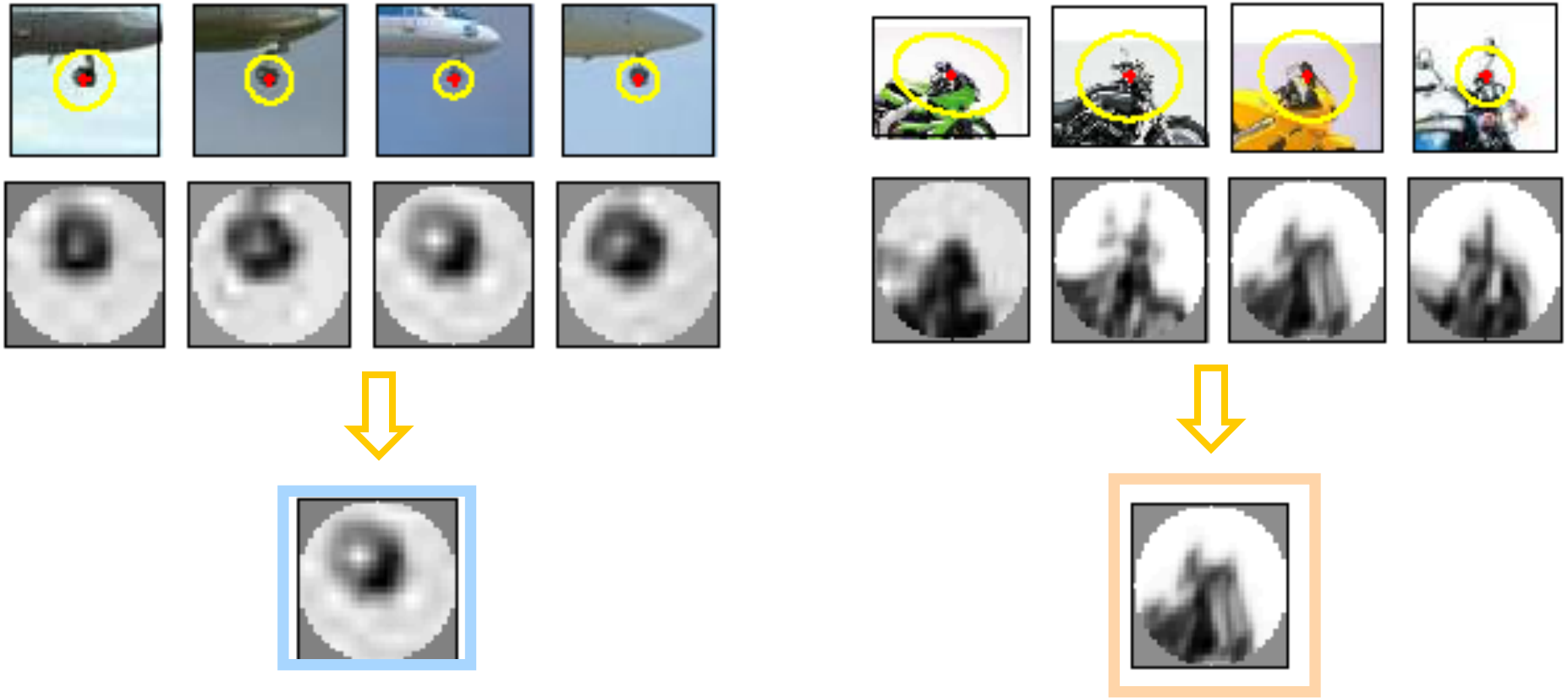


2. Codewords dictionary formation



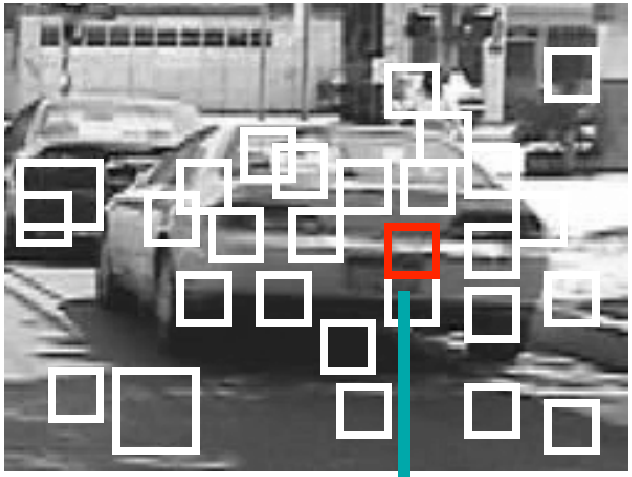
2. Codewords dictionary formation

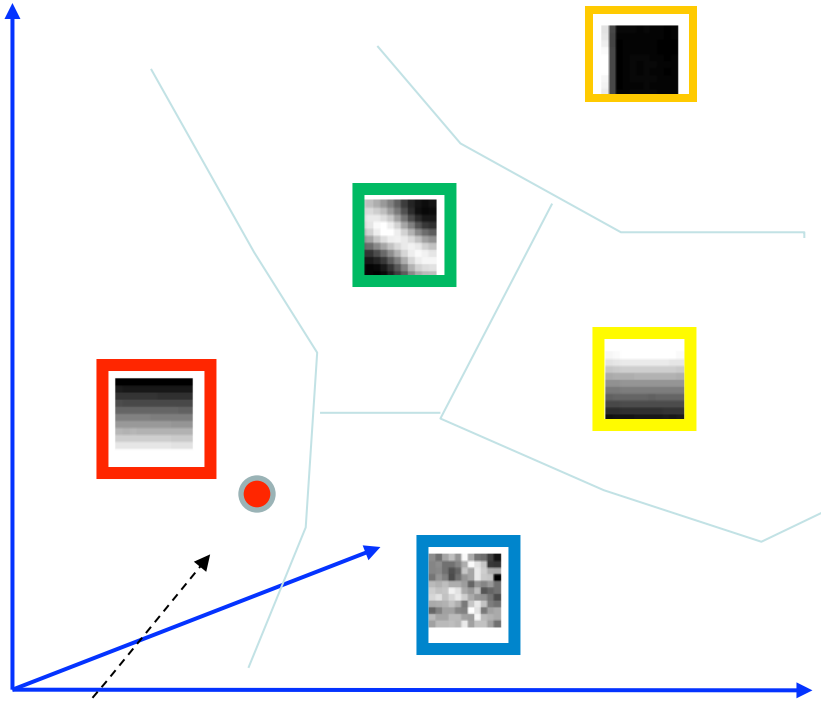
- **Image patch examples of codewords**

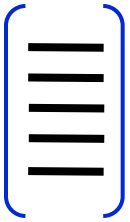


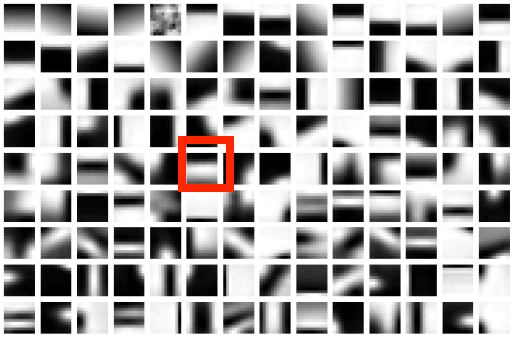
Sivic et al. 2005

3. Bag of word representation





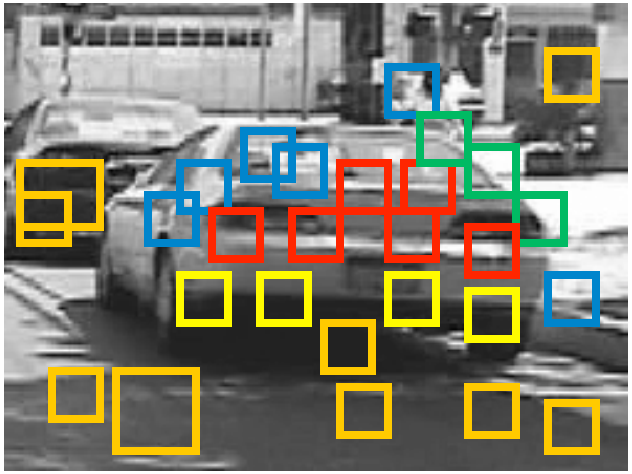


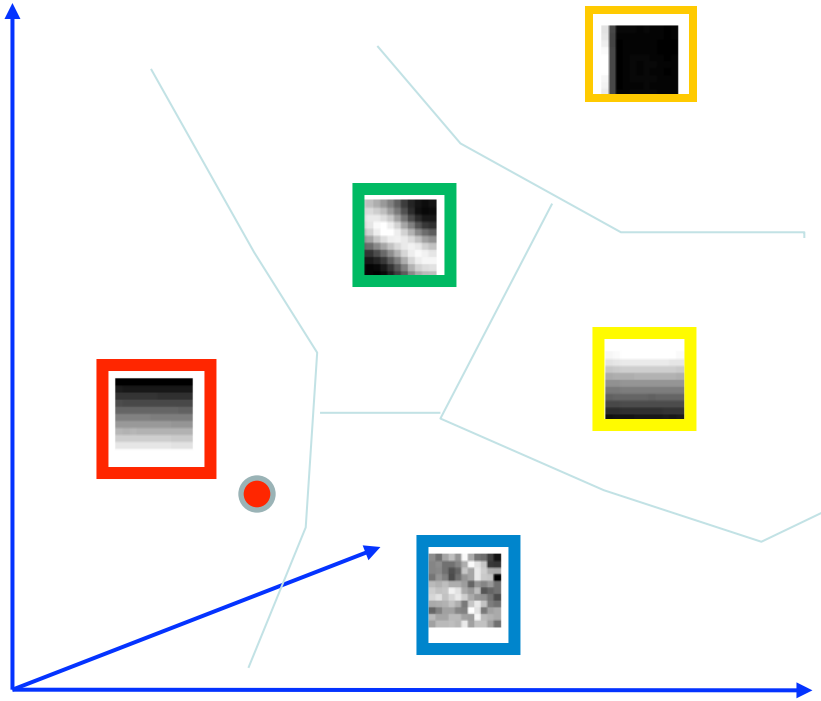


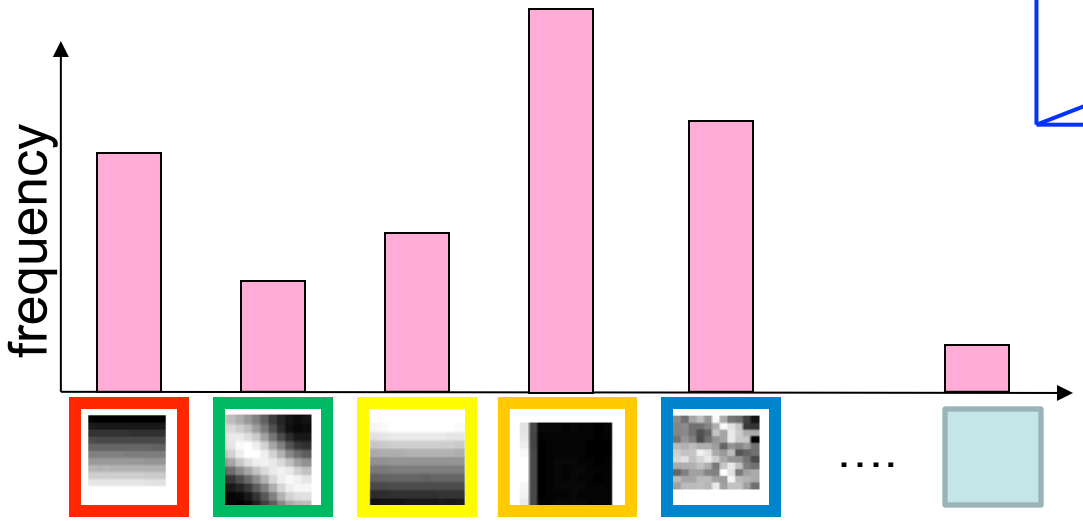
Codewords dictionary

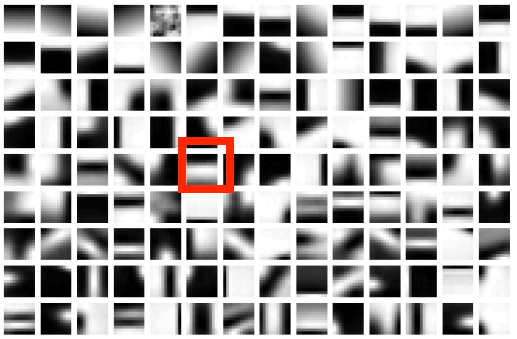
- Nearest neighbors assignment
- K-D tree search strategy

3. Bag of word representation



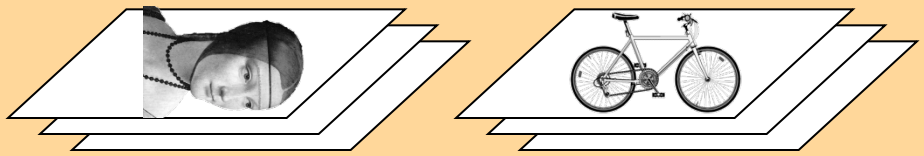






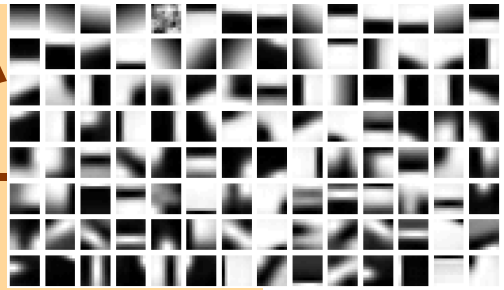
Codewords dictionary

Representation



2.

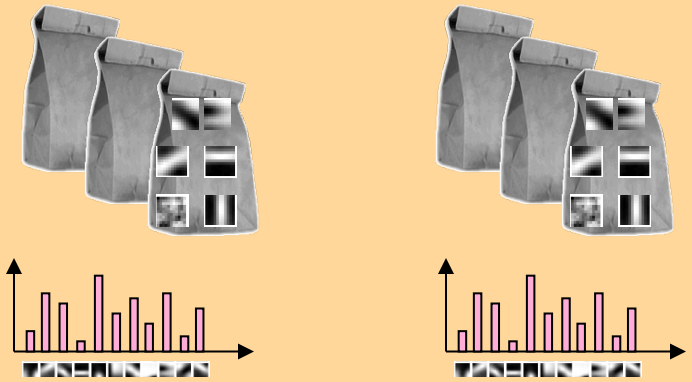
codewords dictionary



1. feature detection & representation

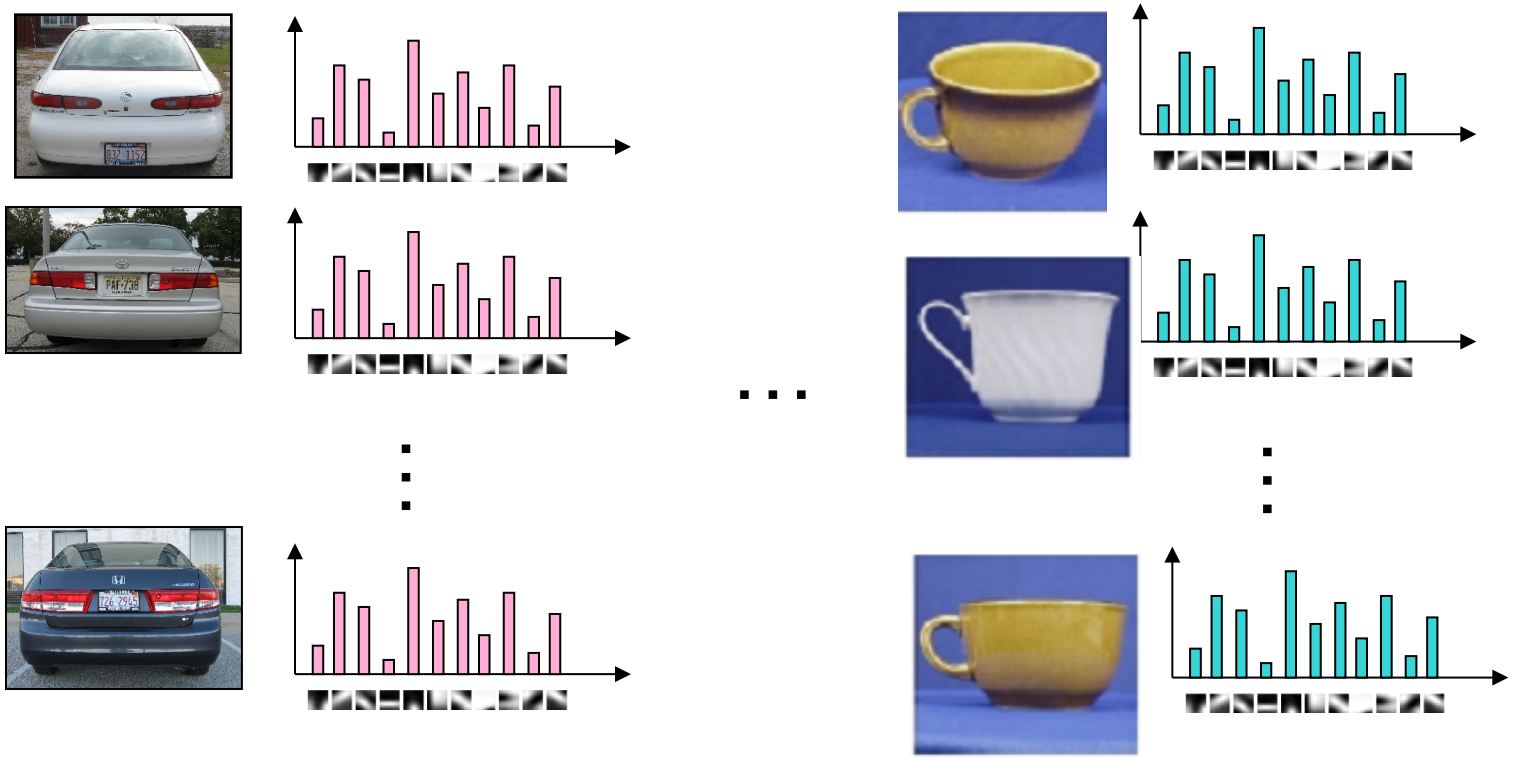
image representation

3.



category models

Category models

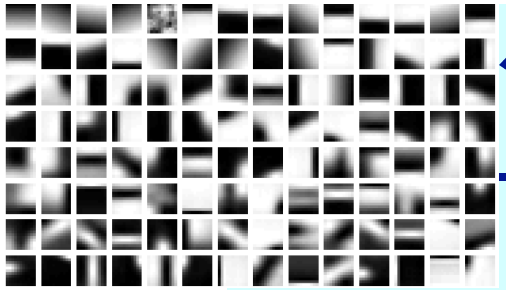


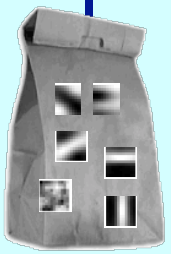
Class 1

Class N

Recognition

codewords dictionary





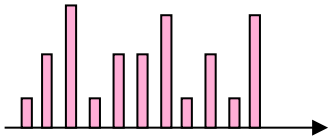
category models (and/or) classifiers

category decision

Major drawback of BOW models

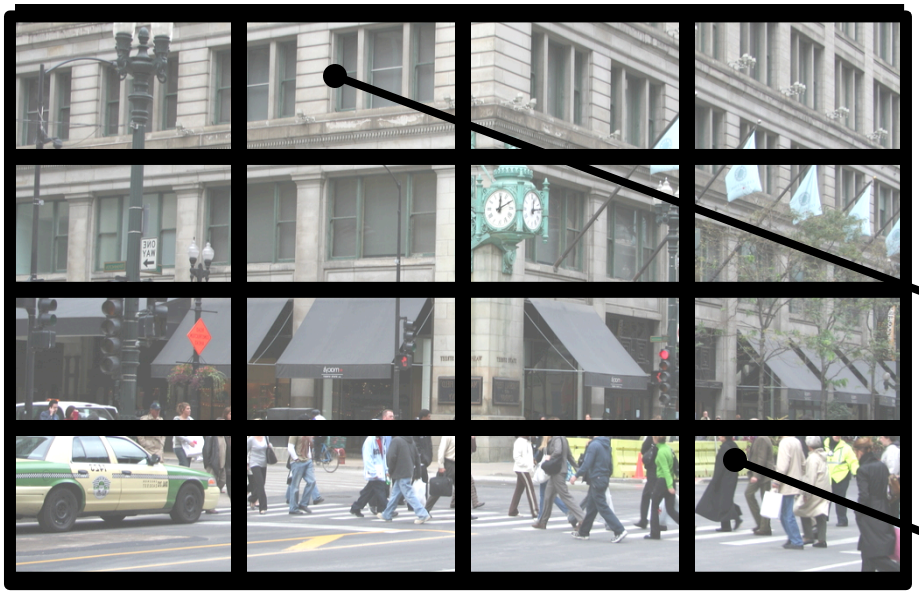
Don’t capture spatial information!

Spatial Pyramid Matching

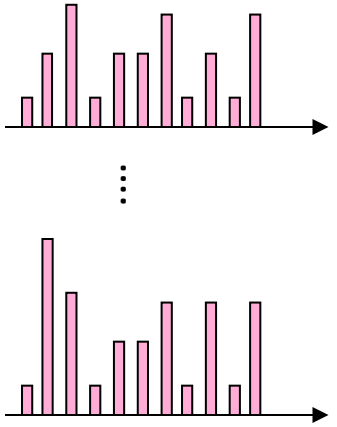


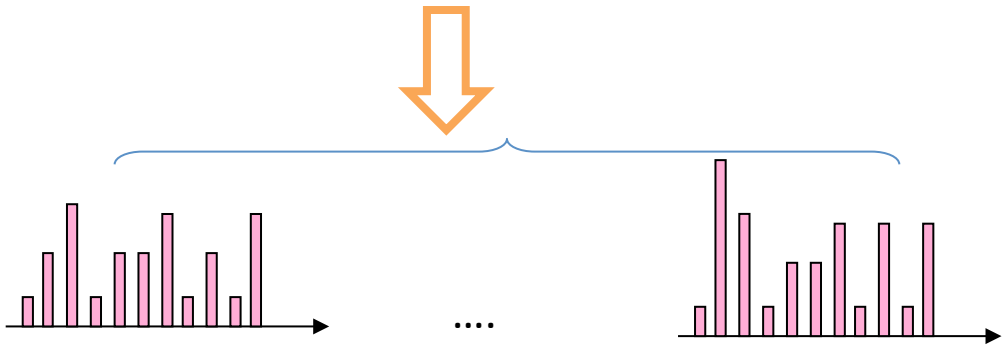
Class “street”

Spatial Pyramid Matching



- K. Grauman and T. Darrell 2005
- S. Lazebnik et al, 2006
- D. Nister et al. 2006,



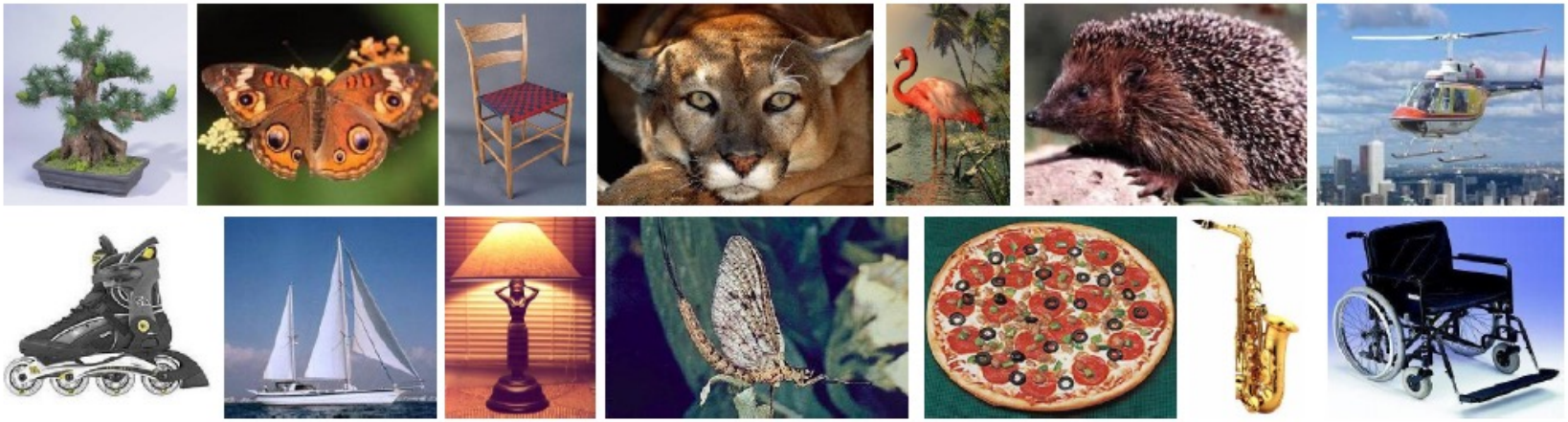


Class “street”

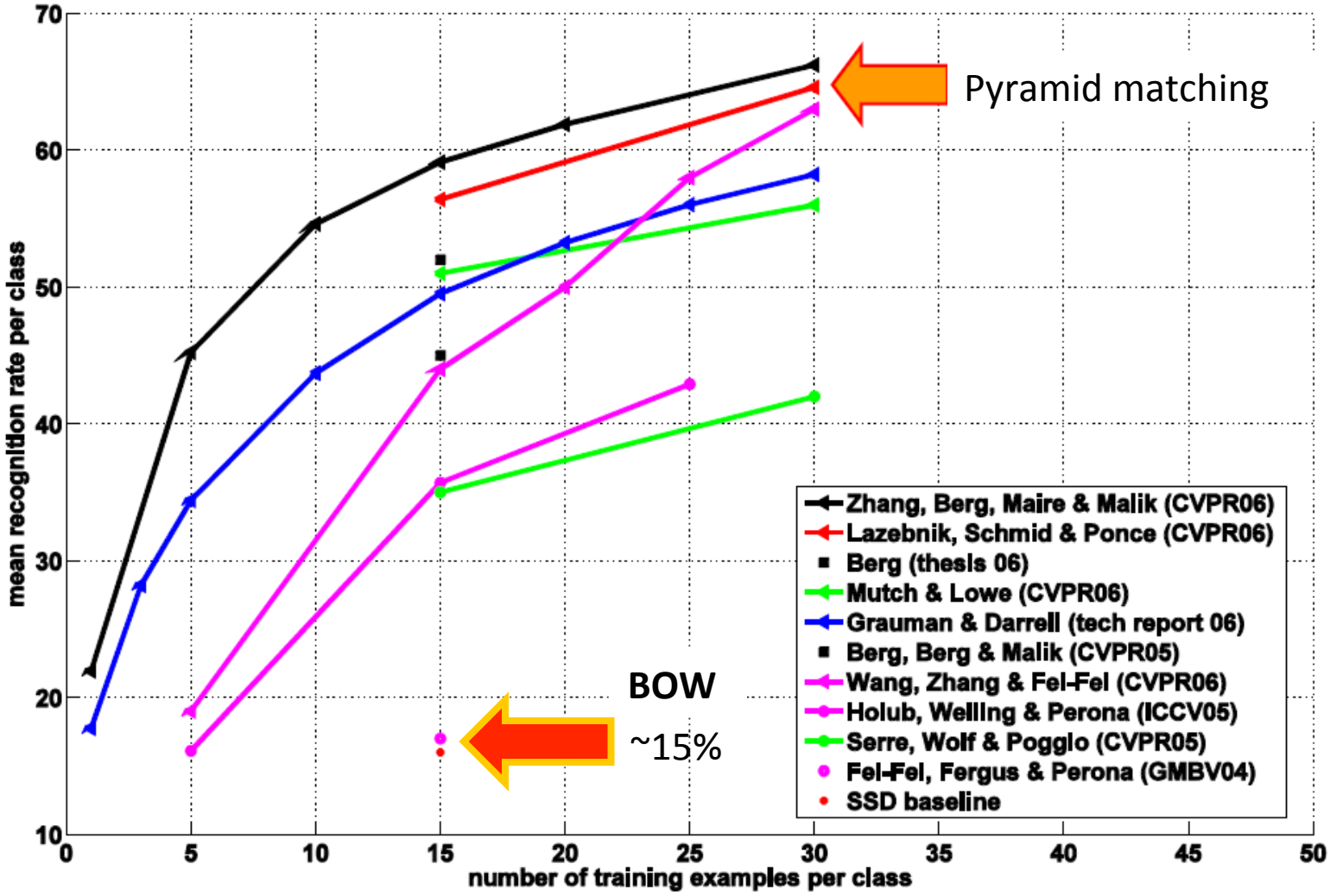
Caltech 101

Fei-Fei et al. (2004)

http://www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html



Caltech 101


Pyramid matching
BOW
~15%
mean recognition rate per class
number of training examples per class
Zhang, Berg, Maire & Malik (CVPR06)
Lazebnik, Schmid & Ponce (CVPR06)
Berg (thesis 06)
Mutch & Lowe (CVPR06)
Grauman & Darrell (tech report 06)
Berg, Berg & Malik (CVPR05)
Wang, Zhang & Fei-Fei (CVPR06)
Holub, Welling & Perona (ICCV05)
Serre, Wolf & Poggio (CVPR05)
Fei-Fei, Fergus & Perona (GMBV04)
SSD baseline

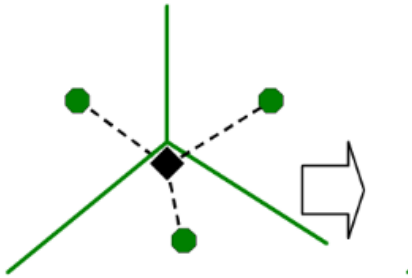
Major drawback of BOW models

- Don’t capture spatial information!
- As the number of images/classes to model increases, the dictionary size also increases
  - Computational cost of increasing the size of the vocabulary becomes very high!

Vocabulary tree

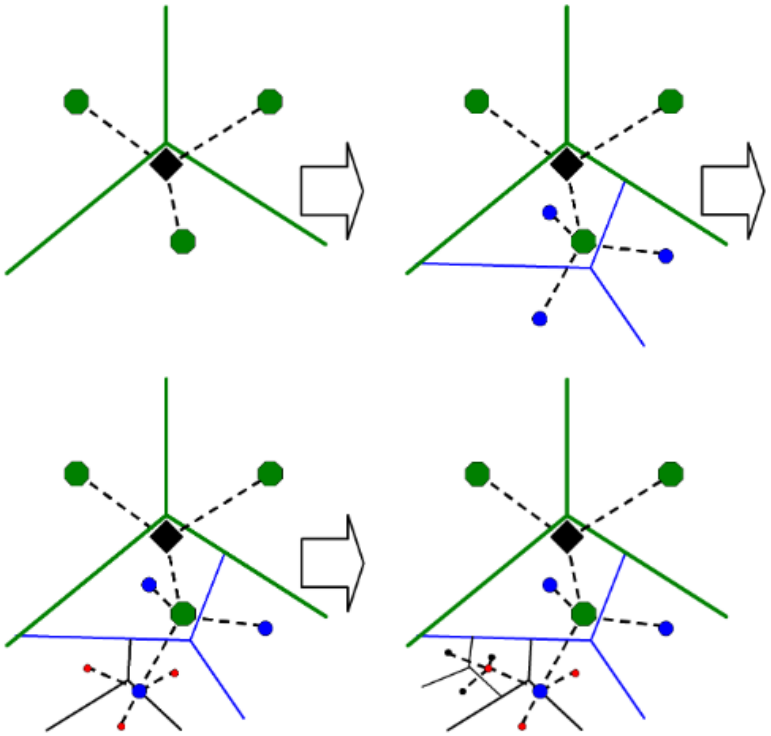
Scalable Recognition with a Vocabulary Tree. David Nistér and Henrik Stewénius. 2006

- Feature vectors are hierarchically clustered in into a k-way tree – also called vocabulary tree
    - Suppose I want to have a dictionary of 81 words



Vocabulary tree

Scalable Recognition with a Vocabulary Tree. David Nistér and Henrik Stewénius. 2006



- First, an initial k-means process is run on the training data, defining k cluster centers.
- The training data is then partitioned into k groups, where each group consists of the descriptor vectors closest to a particular cluster center
- The same process is then recursively applied to each group of descriptor vectors, recursively defining quantization cells by splitting each quantization cell into k new parts

Vocabulary tree

Scalable Recognition with a Vocabulary Tree. David Nistér and Henrik Stewénius. 2006

- Computational cost is logarithmic in the number of leaf nodes.
- Vocabularies of millions (e.g., 10^6) of codewords can be supported
  - Only 10 x 6 comparisons for quantizing each descriptor
  - Individual words can be made more discriminative

Vocabulary tree

With 40,000 images in the database, the retrieval is still real-time… (in 2006 !)

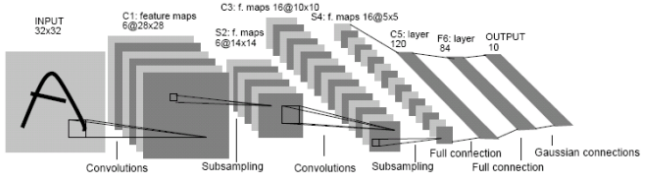
Classification methods

Nearest neighbor

10^6 examples

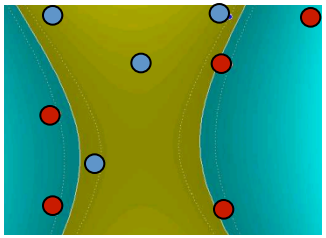
Shakhnarovich, Viola, Darrell 2003
Berg, Berg, Malik 2005...

Neural networks



LeCun, Bottou, Bengio, Haffner 1998
Rowley, Baluja, Kanade 1998
…

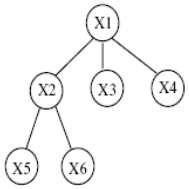
Support Vector Machines



Guyon, Vapnik, Heisele, Serre, Poggio…

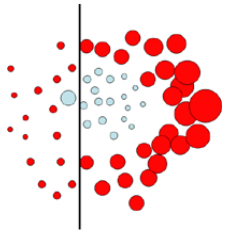
Latent SVM

Structural SVM



Felzenszwalb 00
Ramanan 03…

Boosting



Viola, Jones 2001,
Torralba et al. 2004,
Opelt et al. 2006,…

Courtesy of Vittorio Ferrari
Slide credit: Kristen Grauman

Slide adapted from Antonio Torralba

Recognition

- Classification
- Detection
- Single instance detection and localization

Detection

Does this image contain a face? [where?]

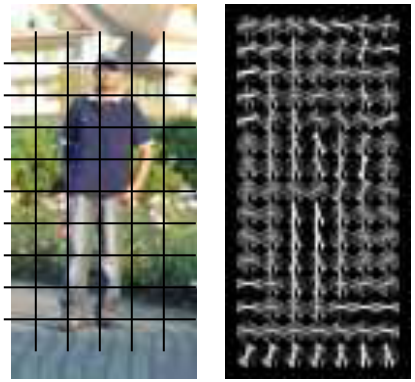
Detection

Does this image contain a face? [where?]

1. Slide a window in image [Viola, Jones 2001]
   - E.g., choose position, scale orientation
2. Compare it with a model/ template
   - Compute similarity to an example object or to a summary representation
3. Compute a score for each comparison
4. Retain position with max score

Model template:
HoG (Histogram of Oriented Gradients)

- Like SIFT, but…
  - Sampled on a dense, regular grid around the object
  - Gradients are contrast normalized in overlapping blocks

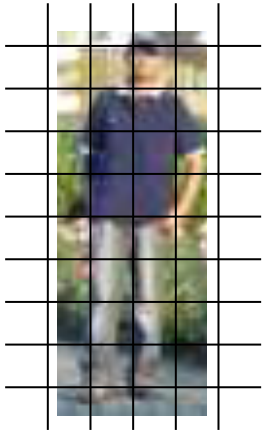


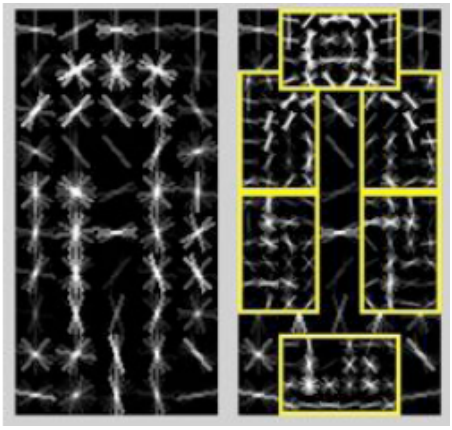
In OPEN CV: struct CV_EXPORTS HOGDescriptor

Navneet Dalal and Bill Triggs, Histograms of Oriented Gradients for Human Detection, CVPR05

Model template:
DPM (Deformable Part Model)

- Like HOG template, but…
  - Use a star-structured part-based model made of:
    - Root filter (similar to Dalal-Triggs)
    - Set of parts and an associated deformation model





More on this on Wed!
Guest lecture by Hyun Oh Song

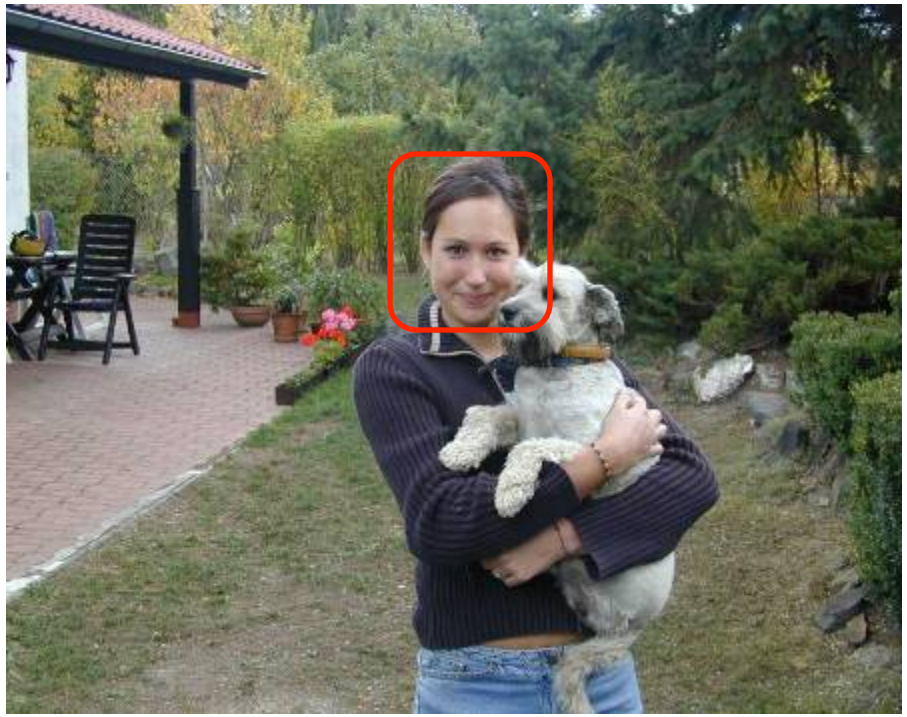
Felzenszwalb, et al., **Discriminatively Trained Deformable Part Models,**
http://people.cs.uchicago.edu/~pff/latent/

Detection

– Issue with Sliding Windows approaches:

- Computational complexity (x,y, S, θ, N of classes)

  - Beyond sliding windows (integral images) [Lampert et al 08, Alexe et al 10]



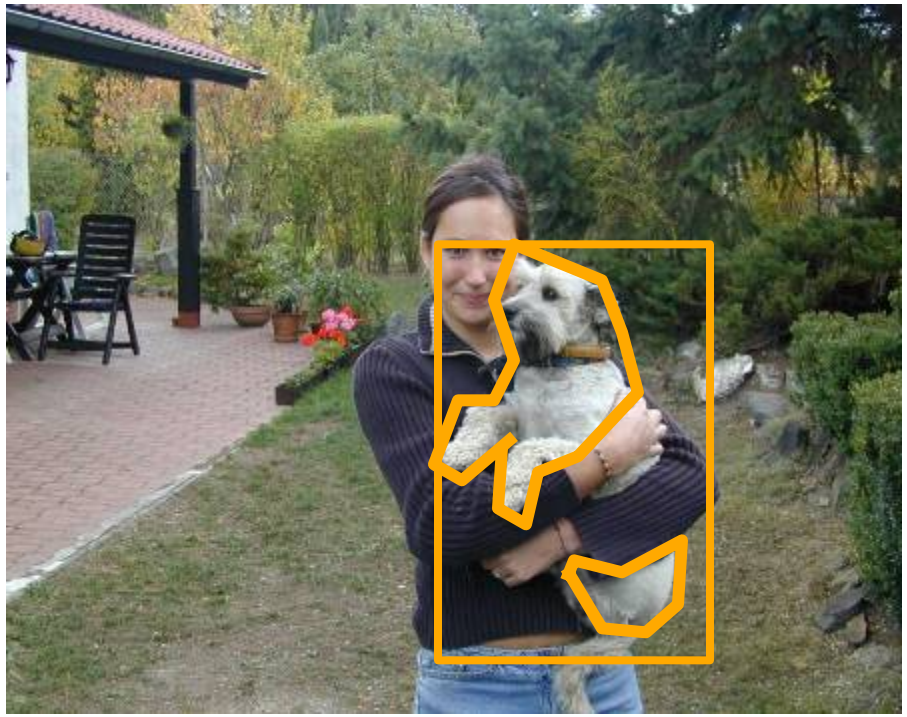
Detection

– Issue with Sliding Windows approaches:

- Computational complexity (x,y, S, θ, N of classes)

  - Beyond sliding windows (integral images) [Lampert et al 08, Alexe et al 10]

- Localization
  - Objects are not boxes



Detection

– Issue with Sliding Windows approaches:

- Computational complexity (x,y, S, θ, N of classes)

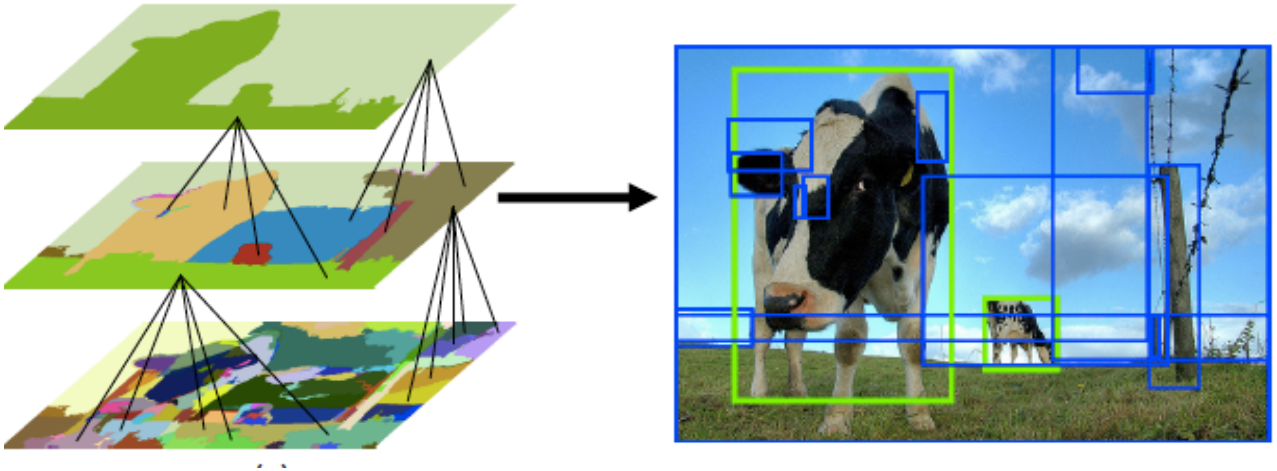
  - Beyond sliding windows (integral images) [Lampert et al 08, Alexe et al 10]

- Localization
  - Objects are not boxes
  - Prone to false positive

    Non max suppression:
    Canny ’86
    ….
    Desai et al , 2009

Detection

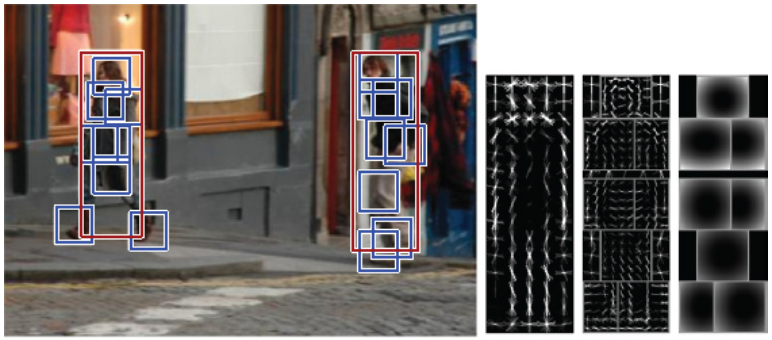
- Selective Search:



Selective Search: Sande et al 2011
segDPM: Fidler, Mottaghi, Yuille, Urtasun 2013

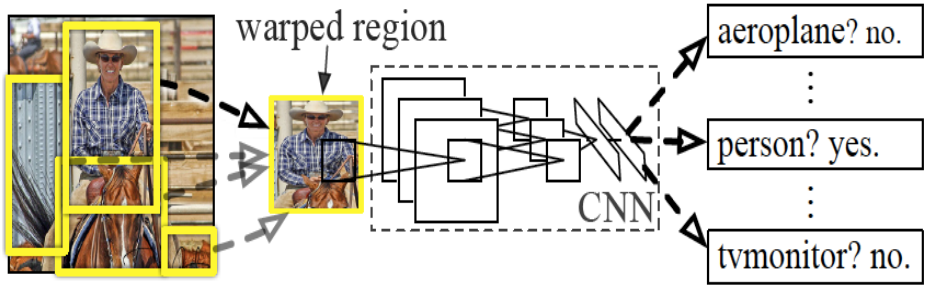
Object Detection

Deformable Part Models (DPM)



DPM: Felzenszwalb, Girshick, McAllester, Ramanan 2010
Sparselet: Song et al. 2012
Multi-Component model: Gu et al. 2012

Convolutional Neural Network (CNN)

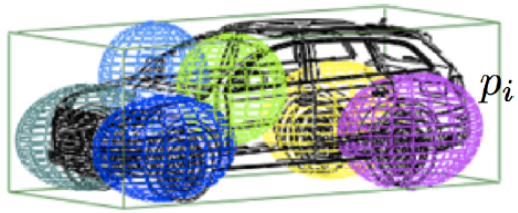


CNN: LeCun, Bottou, Bengio, Haffner 1998
Deep CNN: Krizhevsky, Sutskever, Hinton 2012
R-CNN: Girshick, J. Donahue, T. Darrell, J. Malik 2014

Boosting

Vila-Jones Detection: 2001
Regionlet: Wang et al 2013

3D Object Detection



ALM: Yu & Savarese, 2012
3D²PM: Pepik et al 2012
RGBD-CPMC: Lin et al 2013

Recognition

- Classification
- Detection
- Single instance detection and localization

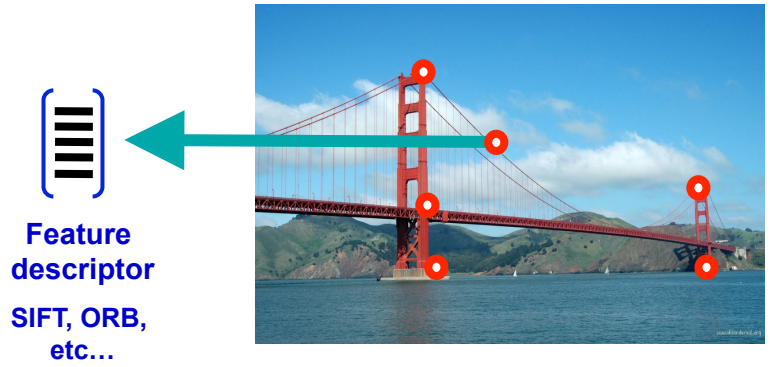
Single instance detection

- Does this image contain the golden gate bridge? [where?]
- Or which landmark does this image contain?

Recognizing single instances

-Representation

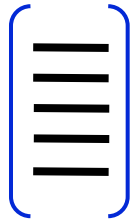
- Detectors and descriptors



-Model learning & Recognition

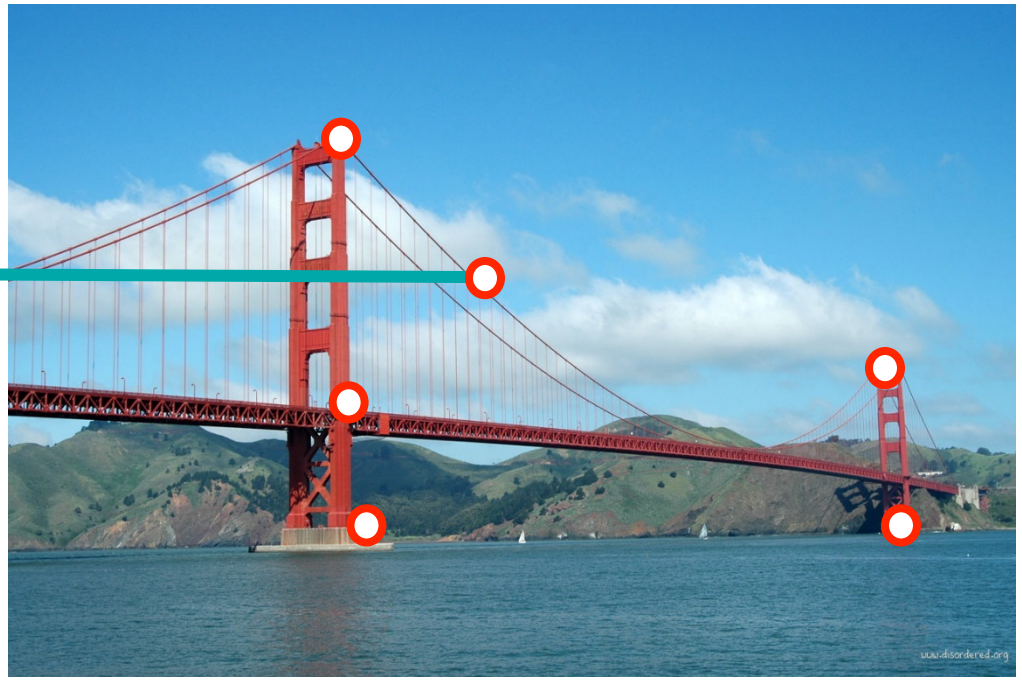
- Hypothesis generation
- Model verification

Representation



Feature descriptor

SIFT, ORB, etc…



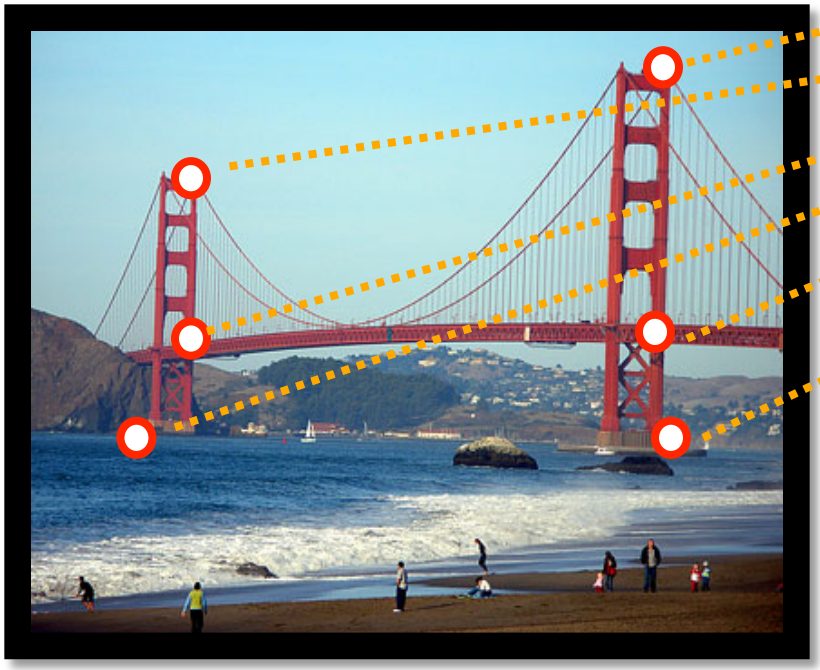
Recognition

Goal: given a query image I, match objects in the image against a collection of learnt object models

Recognition

Goal: given a query image I, match objects in the image against a collection of learnt object models

- Match features between query image I and object model
- Generate hypothesis with a few matches
- Verify hypothesis with all the remaining matches
- Select hypothesis with lowest fitting error



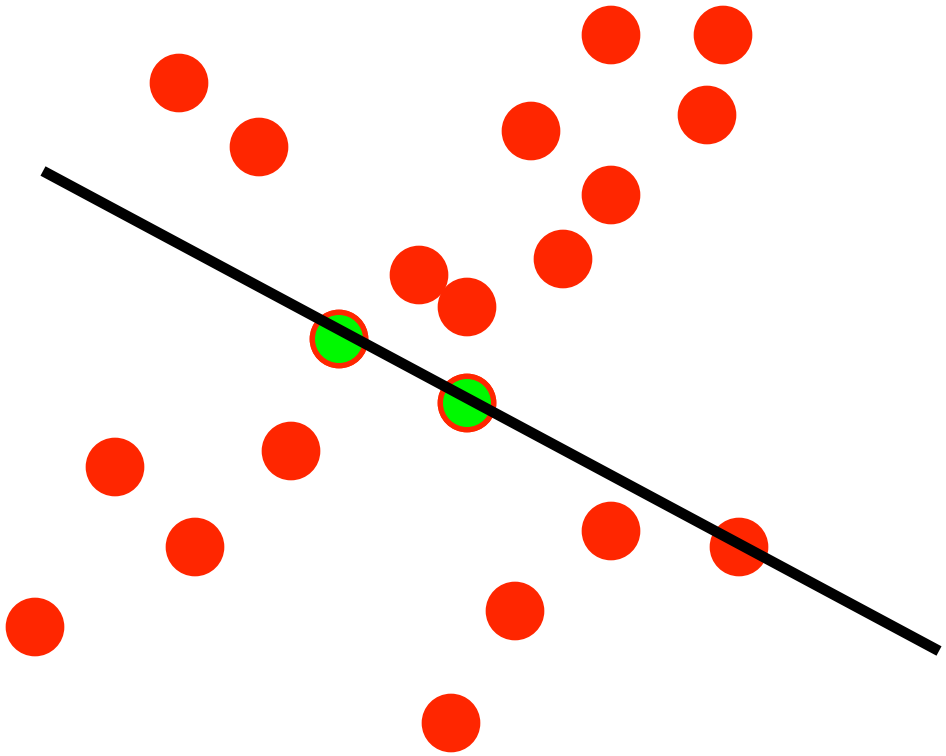
Recognition

- Which model to use?
- How generate hypotheses?
- How to verify these hypotheses

- Detecting planar objects
- Detecting arbitrary objects and estimate camera/object pose

RANSAC!

[Fisher & Bolles, 84]



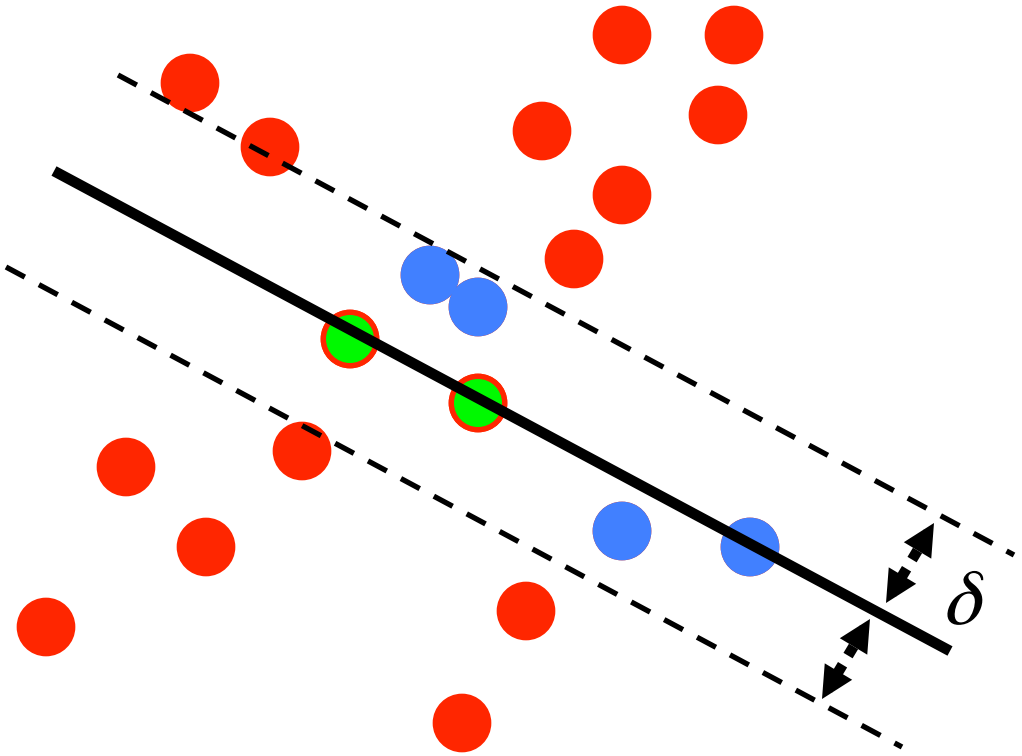
Sample set = set of points in 2D

Algorithm:

1. Select random sample of minimum required size to fit model [?] =[2]
2. Compute a putative model from sample set
3. Compute the set of inliers to this model from whole data set

Repeat 1-3 until model with the most inliers over all samples is found

RANSAC!



Sample set = set of points in 2D

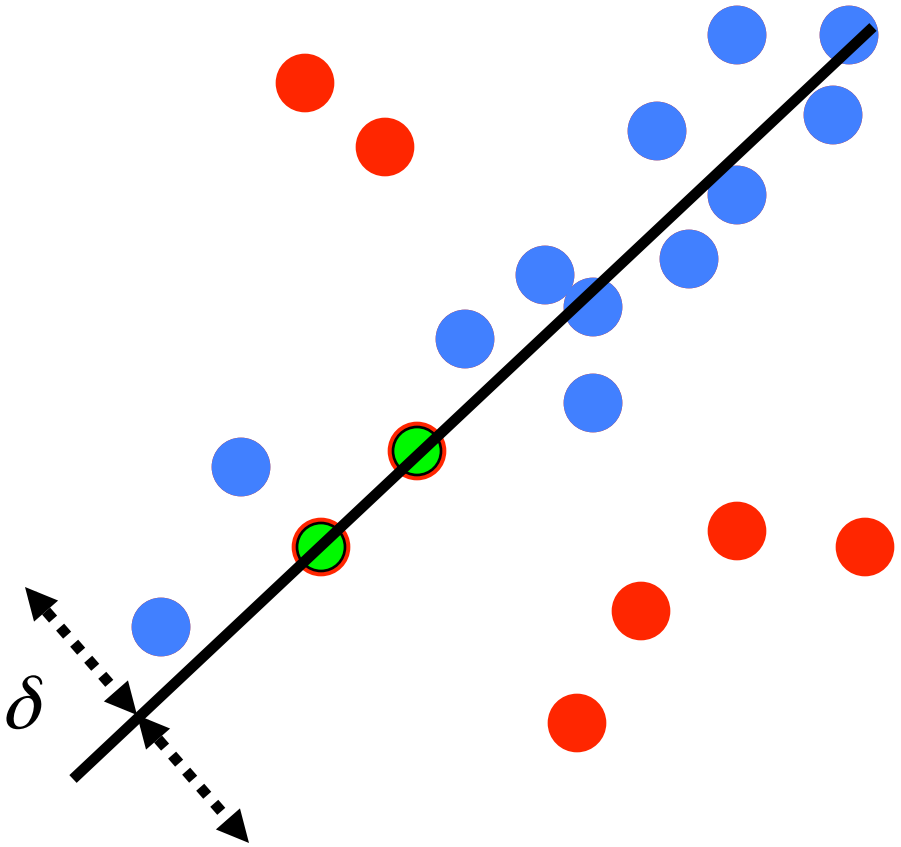
$|O| = 14$

Algorithm:

1. Select random sample of minimum required size to fit model [?] =[2]
2. Compute a putative model from sample set
3. Compute the set of inliers to this model from whole data set

Repeat 1-3 until model with the most inliers over all samples is found

RANSAC!



$|O| = 6$

Algorithm:

1. Select random sample of minimum required size to fit model [?]
2. Compute a putative model from sample set
3. Compute the set of inliers to this model from whole data set

Repeat 1-3 until model with the most inliers over all samples is found

Recognition

Goal: given a query image I, detect object instance and estimate its pose

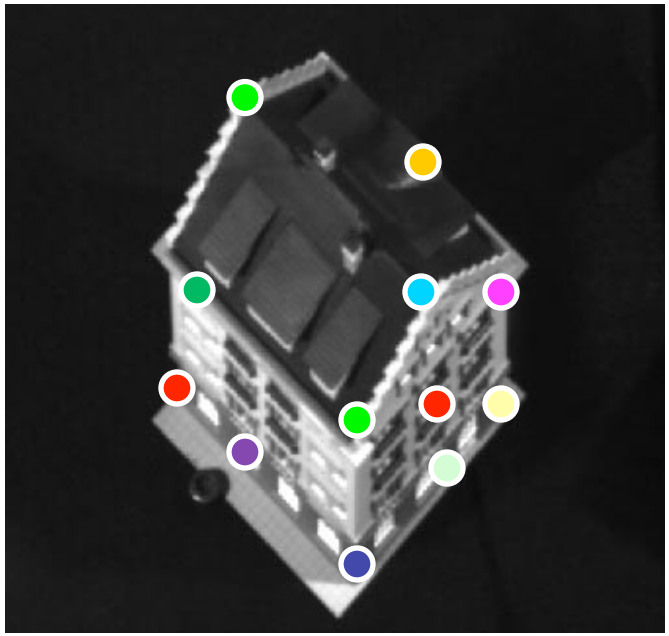
Equivalent to: from a collection of learnt object models, find object model that matches object in image

- Generate hypothesis
- Verify hypothesis
- Select hypothesis with lowest fitting error
- Generate recognition results

Recognition

Goal: given a query image I, find object model that matches with I

query

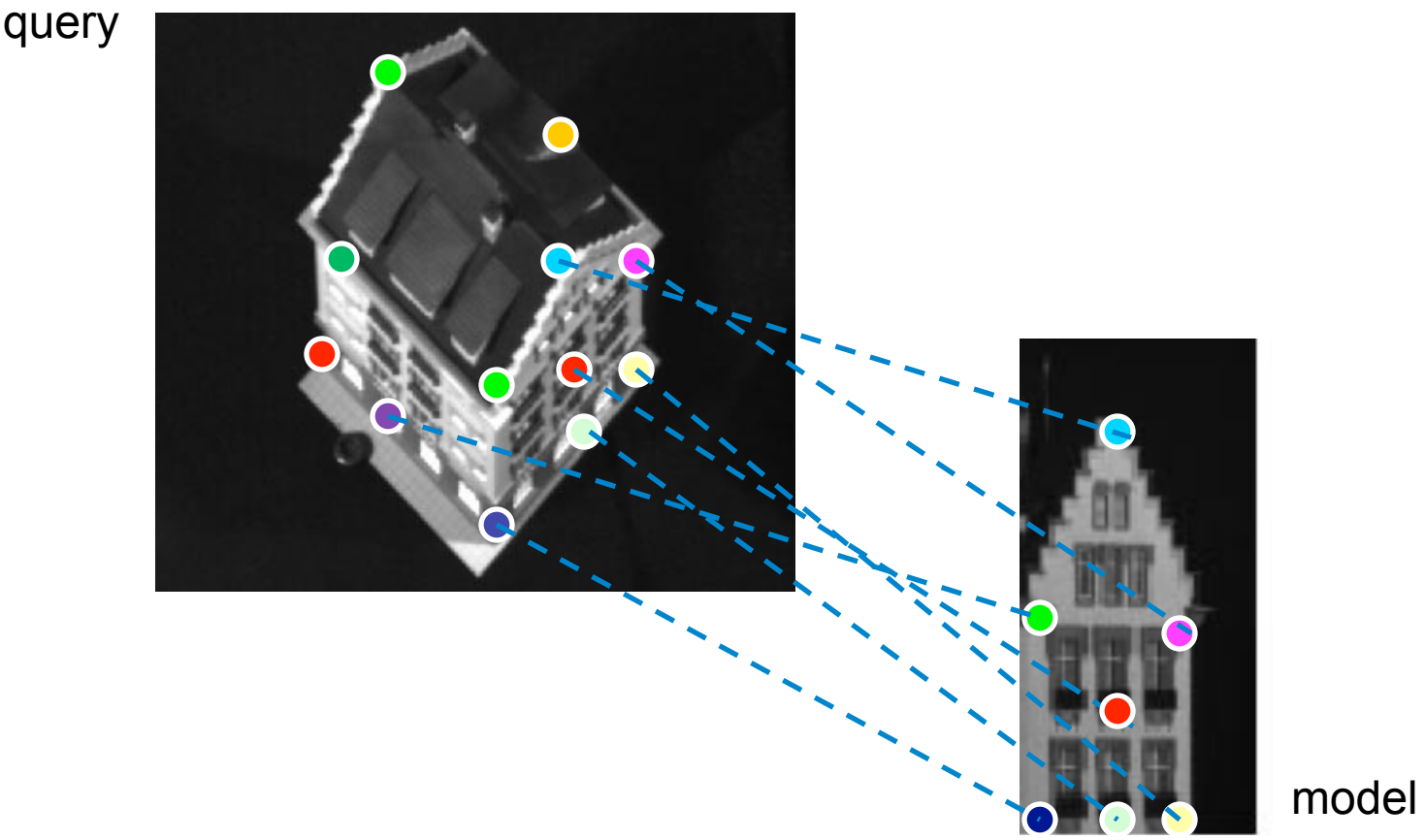


model

Recognition

Model: collection of points on planar surface

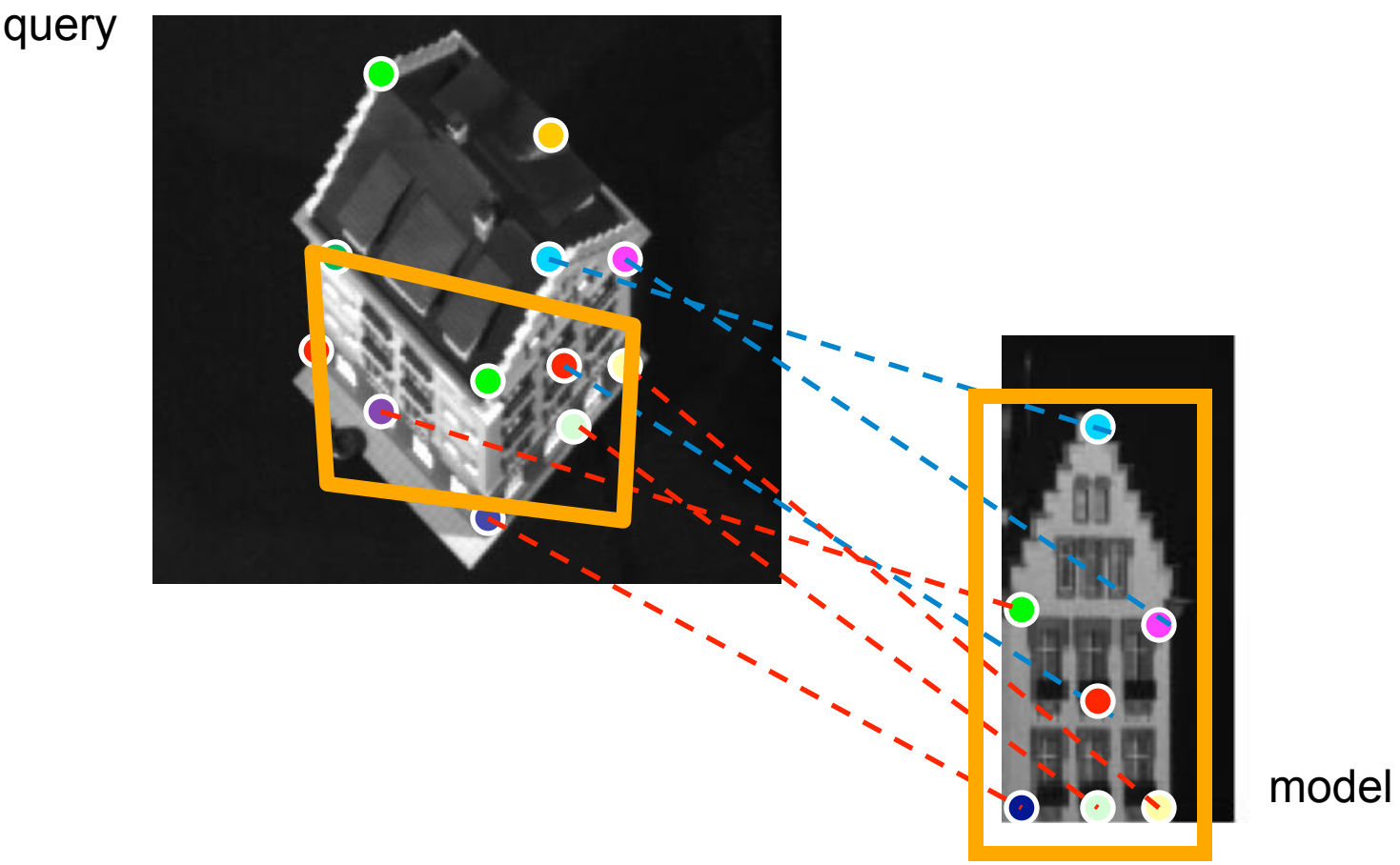
- Find matches between “model” points and “query” points
- Using N matches to fit homographic transformation
- If matches and selected model are correct, the fitting error is small



Recognition

Model: collection of points on planar surface

- Find matches between “model” points and “query” points
- Using N matches to fit homographic transformation
- If matches and selected model are correct, the fitting error is small

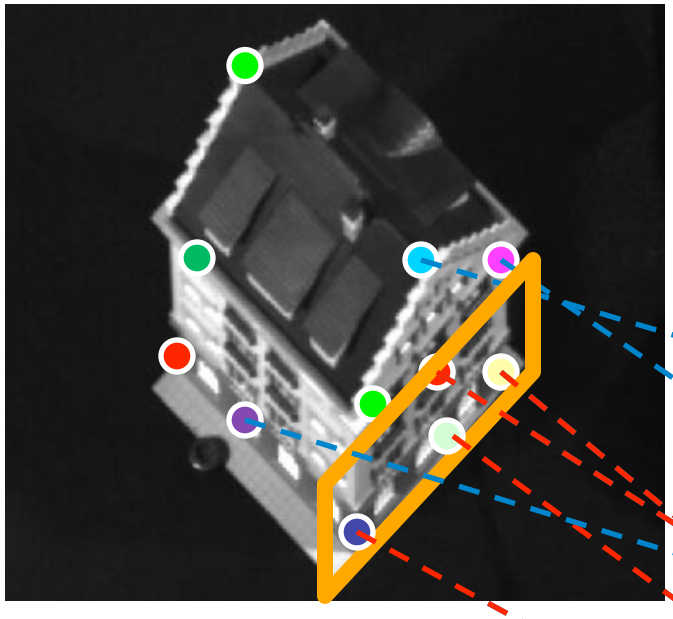


Recognition

Model: collection of points on planar surface

- Find matches between “model” points and “query” points
- Using N matches to fit homographic transformation
- If matches and selected model are correct, the fitting error is small

query



- Generate hypothesis
- Verify hypothesis
- Select hypothesis with lowest fitting error
- Generate recognition results

model

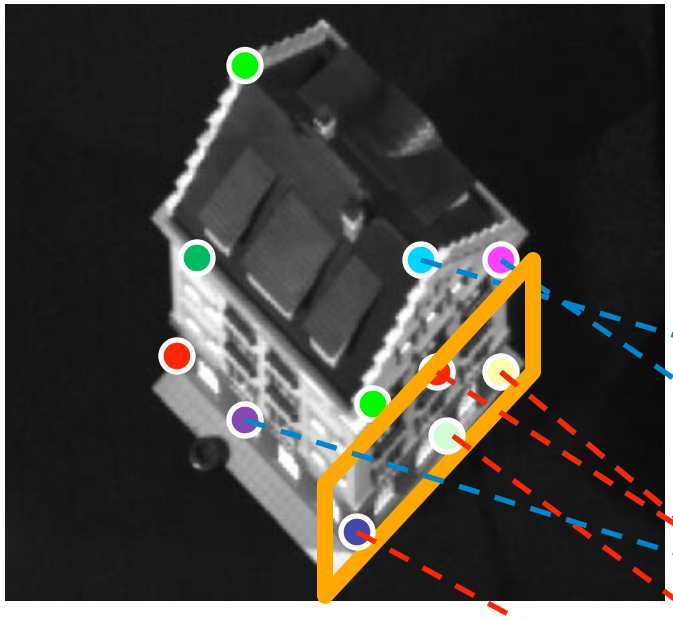
Verification: The hypothesis generates ***low*** fitting error

Recognition

Model: collection of points on planar surface

- Find matches between “model” points and “query” points
- Using N matches to fit homographic transformation
- If matches and selected model are correct, the fitting error is small

query

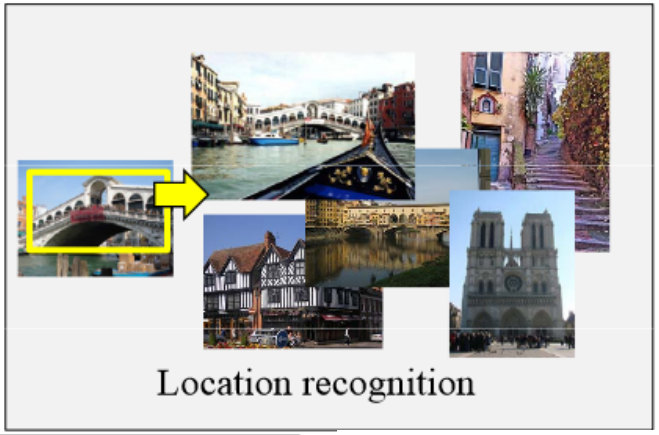


- Generate hypothesis
- Verify hypothesis
- Select hypothesis with lowest fitting error
- Generate recognition results

model

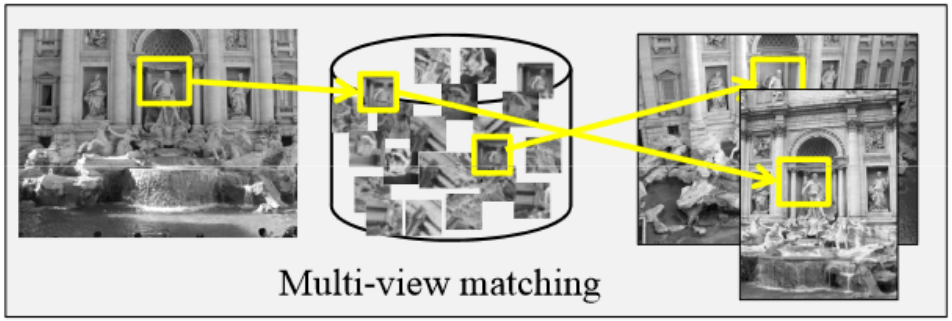
Verification: The hypothesis generates ***low*** fitting error

Large-scale visual search


Location recognition


Product search


Near duplicate detection


Multi-view matching

Visual link

Courtesy of Grauman and Fergus

Recent related work on large scale and efficient image search

- World-scale Mining of Objects and Events from Community Photo Collections. T. Quack, B. Leibe, and L. Van Gool. CIVR 2008.
- Total Recall II: Query Expansion Revisited. O. Chum, A. Mikulik, M. Perdoch, and J. Matas. CVPR 2011.
- Geometric Min-Hashing: Finding a (Thick) Needle in a Haystack, O. Chum, M. Perdoch, and J. Matas. CVPR 2009.
- Three Things Everyone Should Know to Improve Object Retrieval. R. Arandjelovic and A. Zisserman. CVPR 2012.
- Video Mining with Frequent Itemset Configurations. T. Quack, V. Ferrari, and L. Van Gool. CIVR 2006.
- Bundling Features for Large Scale Partial-Duplicate Web Image Search. Z. Wu, Q. Ke, M. Isard, and J. Sun. CVPR 2009.
- Total Recall: Automatic Query Expansion with a Generative Feature Model for Object Retrieval. O. Chum et al. CVPR 2007.
- Discovering Favorite Views of Popular Places with Iconoid Shift. T. Weyand and B. Leibe. ICCV 2011.
- Supervised Hashing with Kernels. W. Liu, J. Wang, R. Ji, Y. Jiang, S.-F. Chang. CVPR 2012
- Kernelized Locality Sensitive Hashing for Scalable Image Search, by B. Kulis and K. Grauman, ICCV 2009
- Image Webs: Computing and Exploiting Connectivity in Image Collections. K. Heath, N. Gelfand, M. Ovsjanikov, M. Aanjaneya, and L. Guibas. CVPR 2010.
- Improving Image-based Localization by Active Correspondence Search. T. Sattler, B. Leibe, L. Kobbelt. ECCV 2012.
- Learning Binary Projections for Large-Scale Image Search. K. Grauman and R. Fergus. Chapter to appear in Registration, Recognition, and
- Object Retrieval with Large Vocabularies and Fast Spatial Matching. J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, CVPR 2007. [pdf] [approx k-means code]
- City-Scale Location Recognition, G. Schindler, M. Brown, and R. Szeliski, CVPR 2007. [pdf]

Single instance object detection on a mobile device

•G. Takacs et al. "Outdoors augmented reality on mobile phone using loxel-based visual feature organization", MIR'08

•B. Girod, V. Chandrasekhar, D. M. Chen, N. M. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. S. Tsai and R. Vedantham, "Mobile Visual Search", IEEE Signal Processing Magazine, vol. 28, no. 4, pp. 61-76, July 2011.

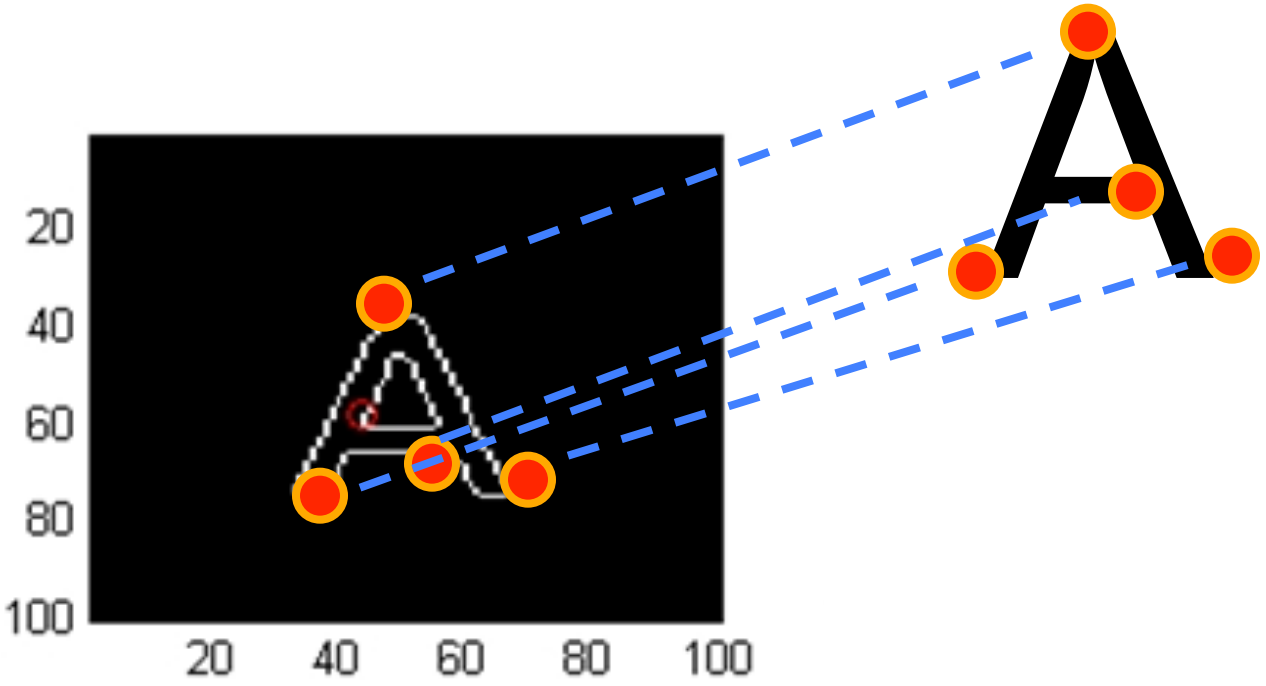
•J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object Retrieval with Large Vocabularies and Fast Spatial Matching," CVPR, 2007.

CS231M · Mobile Computer Vision

Next lecture:

- Object detection by DPM and sparselets

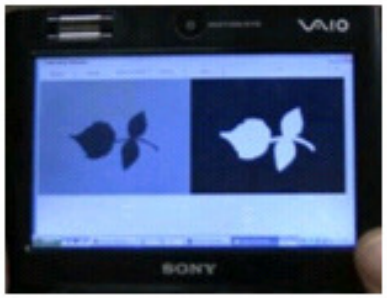
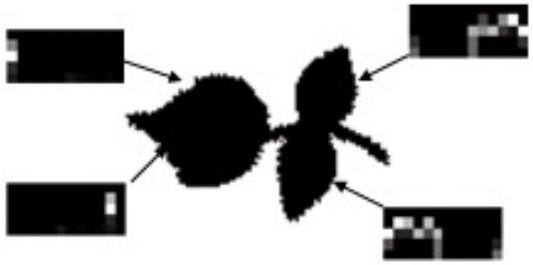
Shape matching



- Match shape against database
- Retrieve relevant information

- Shape context (Belongie et al 00)
- Shape Classification Using the Inner-Distance [Ling and Jacobs 07]

Shape matching

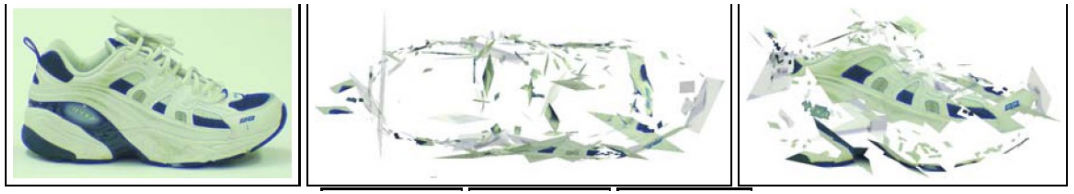





Searching the World's Herbaria: A System for the Visual Identification of Plant Species 2008.
S. Shirdhonkar, et al

Recognizing single instances

Goal: given a query image I, identify object model in the image I

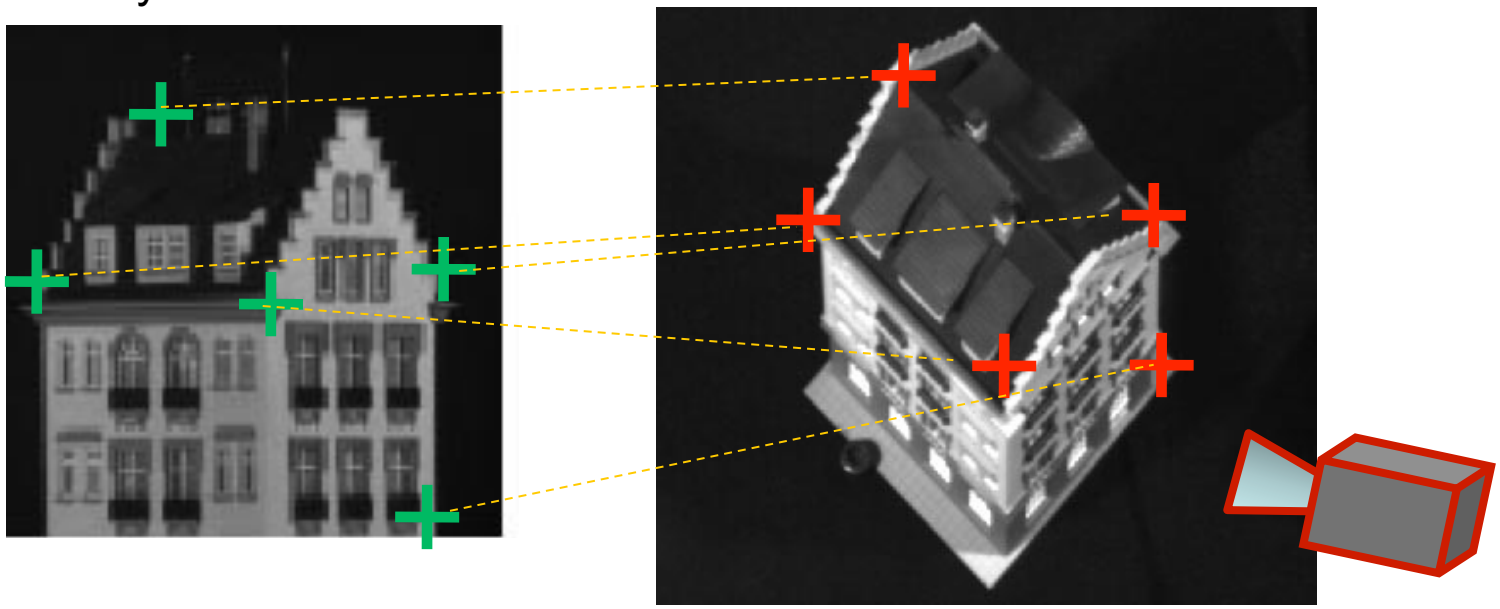
Model: collection of 3D points with descriptors



Rothganger et al. ‘03 ’06

Recognition

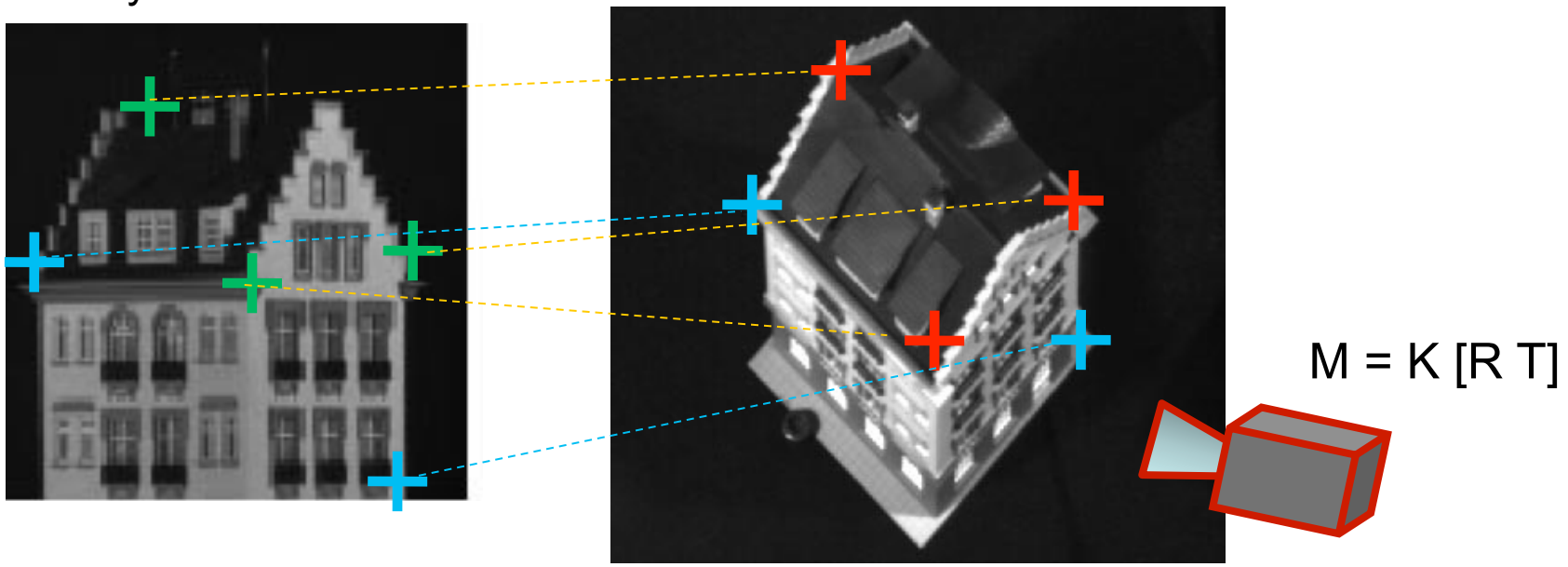
Class: toy house #3



1. Find matches between model and test image features

Recognition

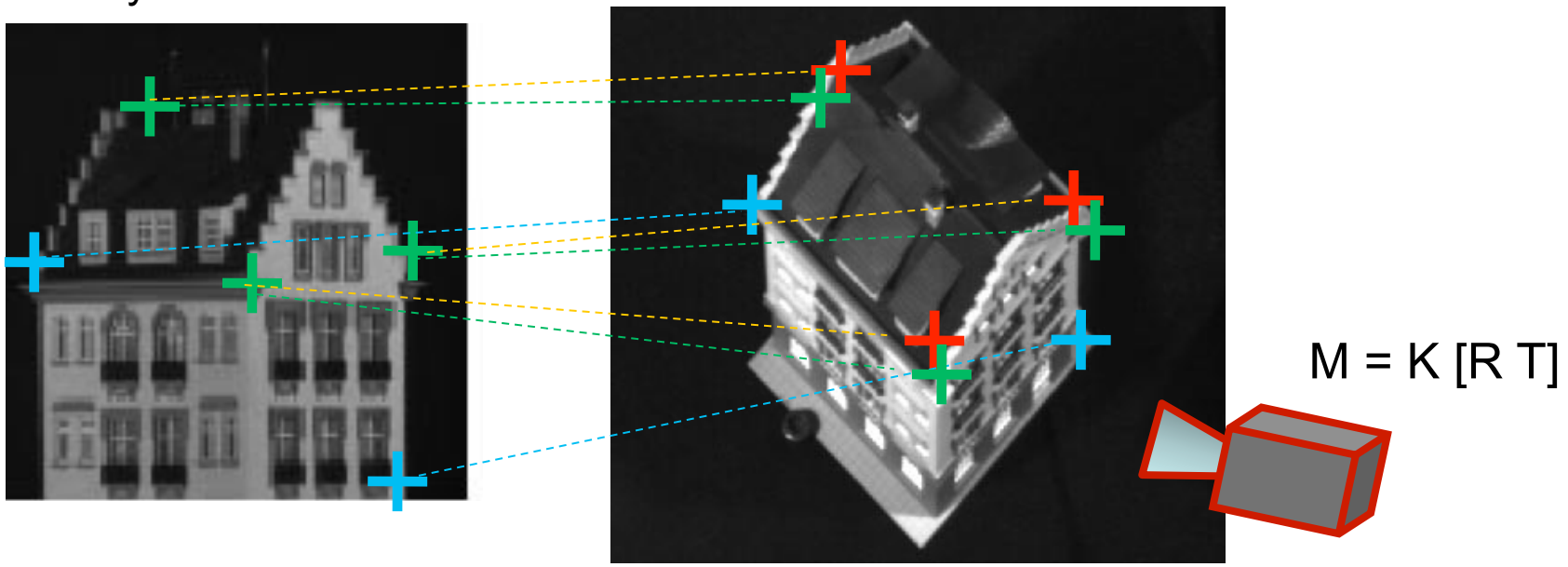
Class: toy house #3



1. Find matches between model and test image features
2. Generate hypothesis:
   - Compute transformation M from N matches (N=2; affine camera; key points with scale and rotation)
   - Generate hypothesis of object location and pose w.r.t. camera

Recognition

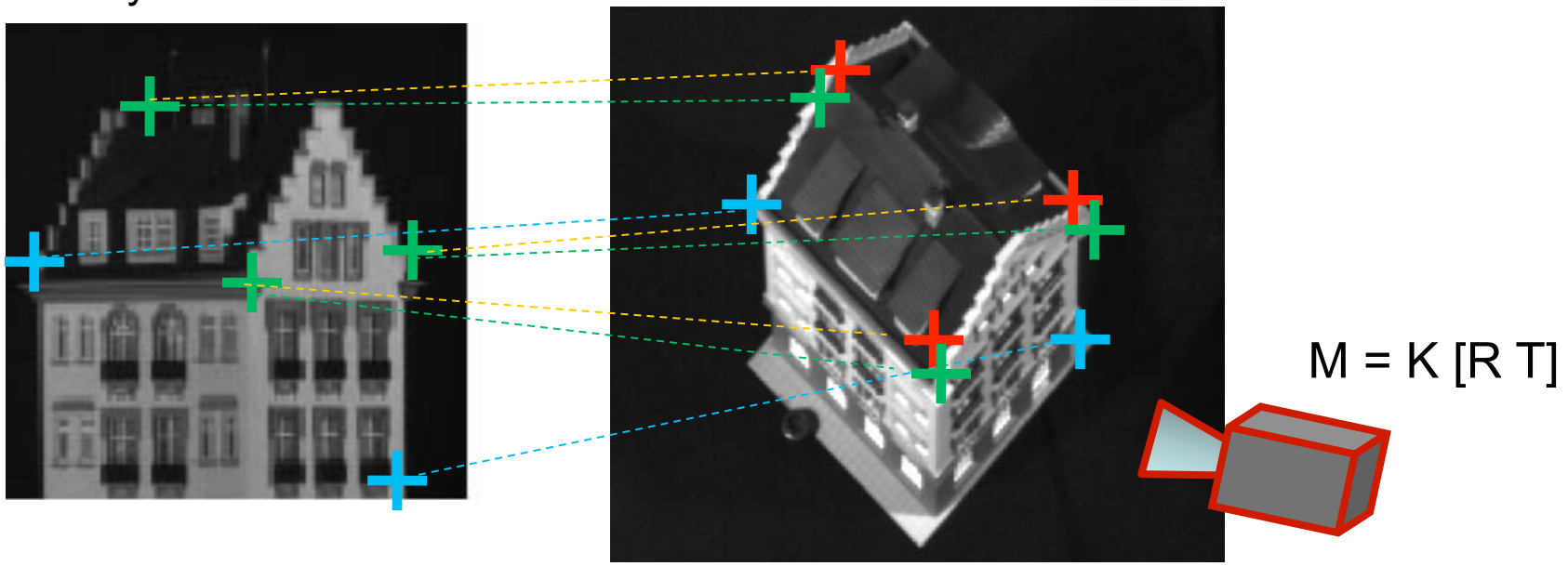
Class: toy house #3



1. Find matches between model and test image features
2. Generate hypothesis:
   - Compute transformation M from N matches
   - Generate hypothesis of object location and pose w.r.t. camera
3. Model verification
   - Use M to project other 3D model features into test image
   - Compute residual = D(projections, measurements)

Recognition

Class: toy house #3



4. Repeat steps 2 and 3 until residual doesn't decrease anymore
5. Repeat steps 1-4 for different object instances
6. M and C corresponding to min residual return the estimated object pose and object instance