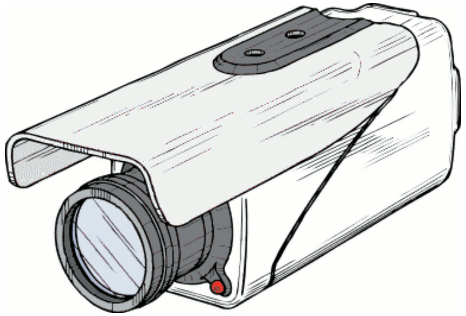
Univerza v Ljubljani

Machine Perception
Recognition and detection using local features

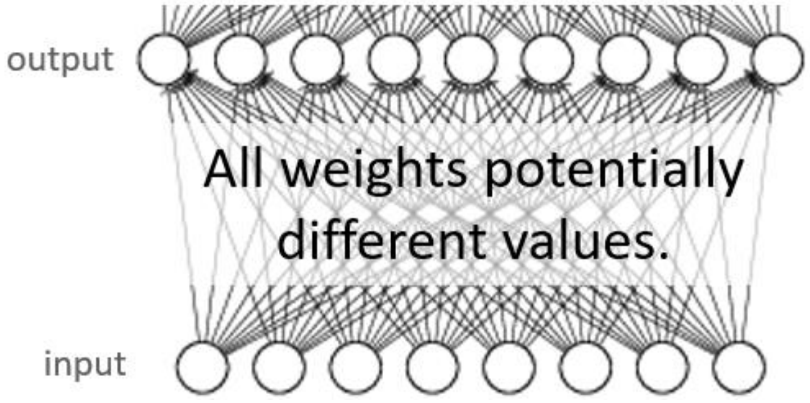


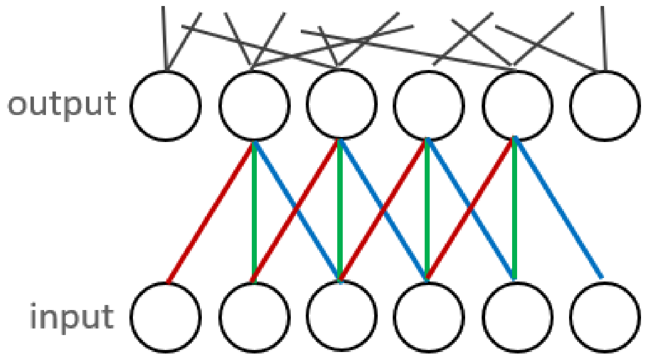
Matej Kristan

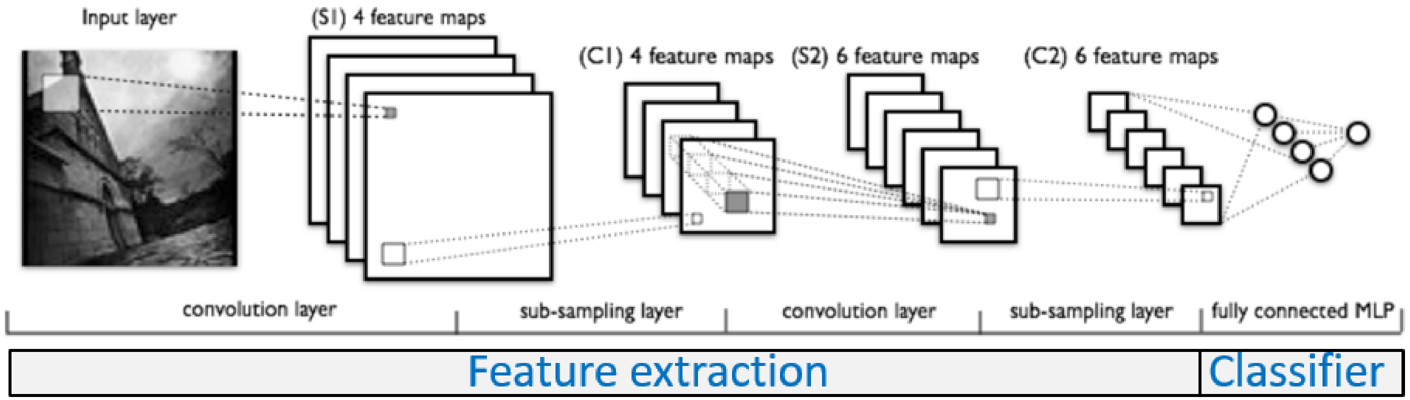
Laboratorij za Umetne Vizualne Spoznavne Sisteme,
Fakulteta za računalništvo in informatiko,
Univerza v Ljubljani

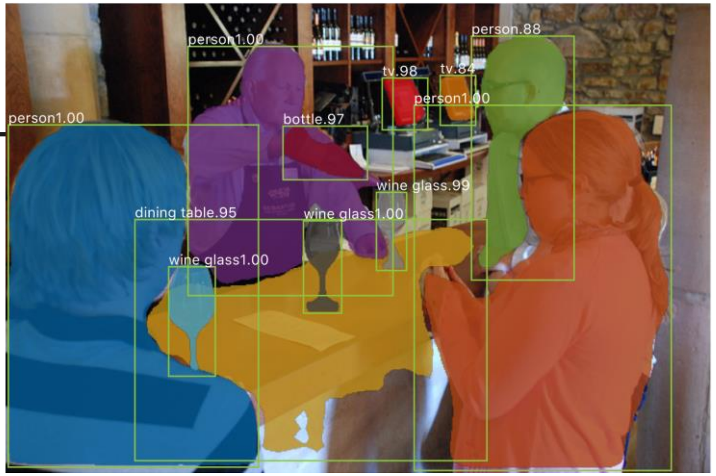
Previously at MP…

- End-to-end feature learning (CNNs) for recognition, detection, segmentation, …







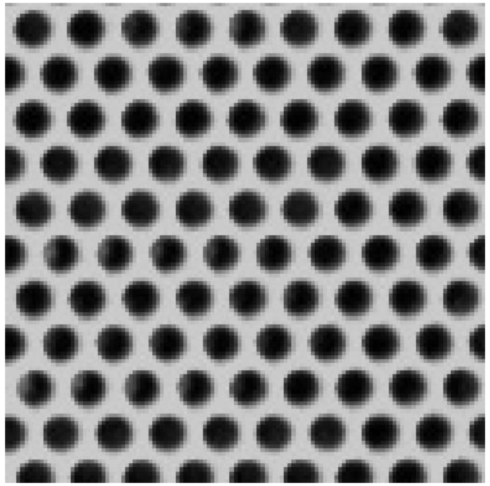


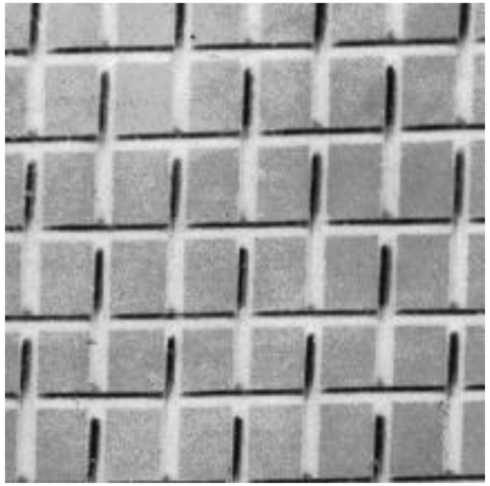
Machine perception

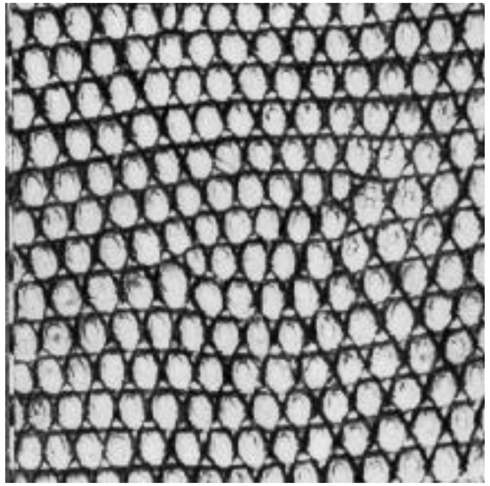
RECOGNITION USING LOCAL FEATURES: *BAG OF WORDS MODELS*

Intuition: texture recognition

- What is texture?
  - Could say: “spatially organized repeatable images”

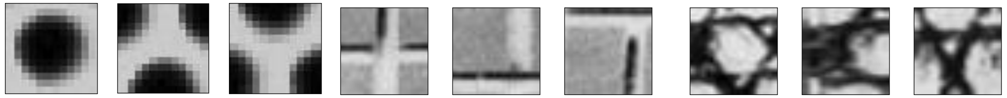




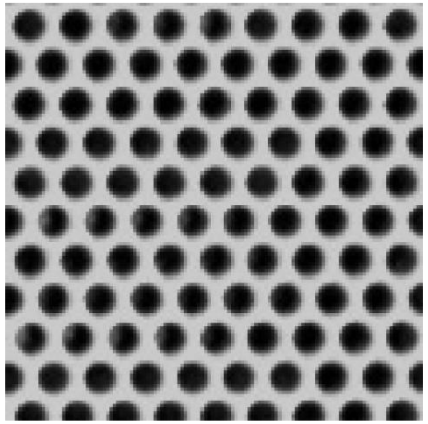


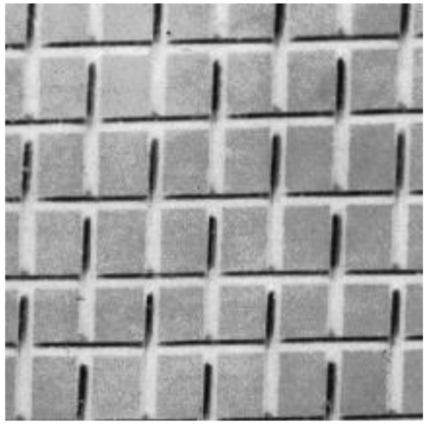
Intuition: texture recognition

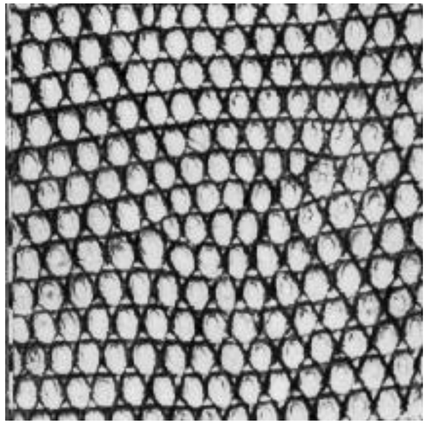
- Texture can be characterized by – textons (small „images“)



- For a random texture, the identity of the textons composing it is more important than their arrangement.

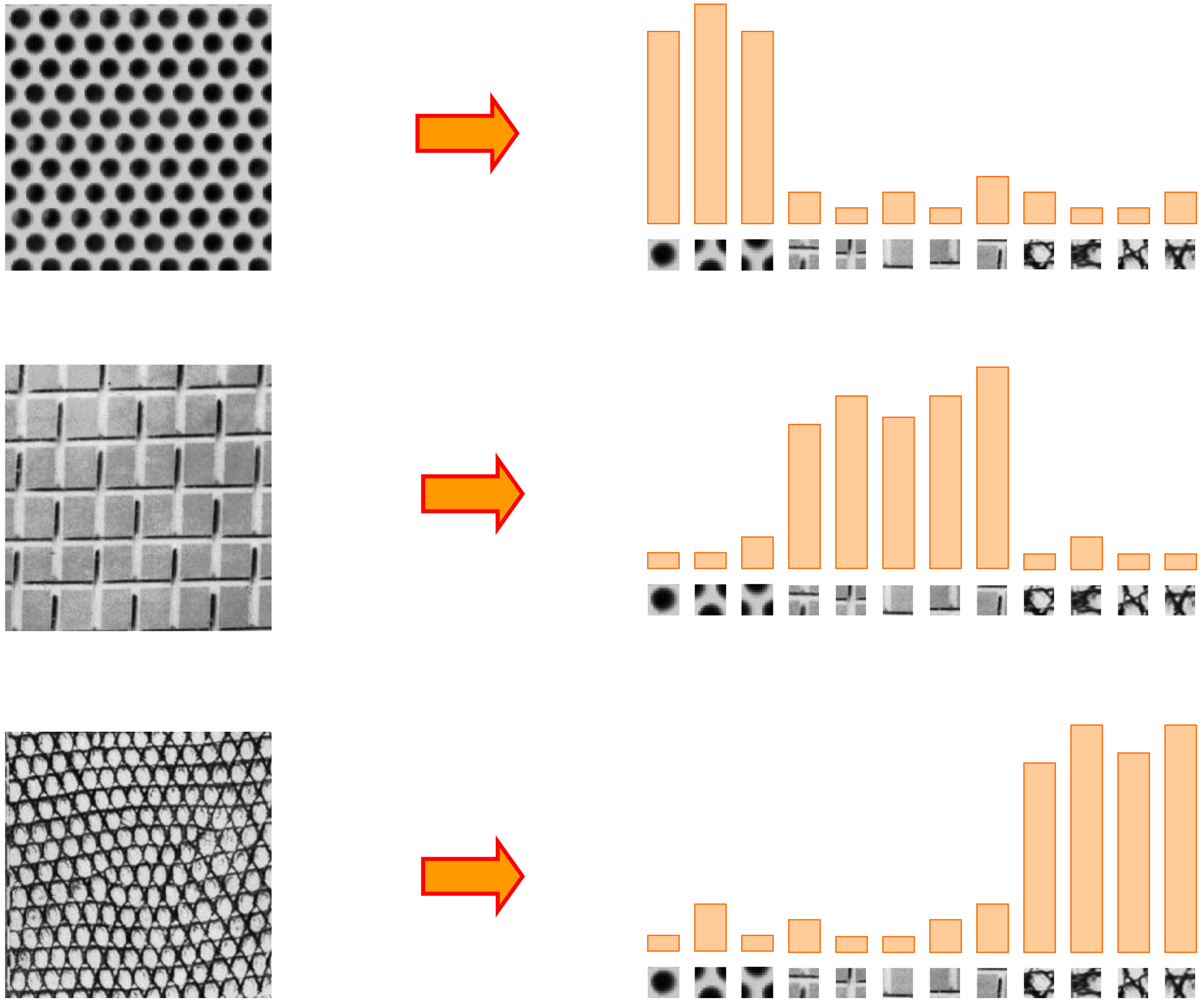






Julesz, 1981; Cula & Dana, 2001; Leung & Malik 2001; Mori, Belongie & Malik, 2001; Schmid 2001; Varma & Zisserman, 2002, 2003; Lazebnik, Schmid & Ponce, 2003

Intuition: texture recognition



Note the difference in the histograms!

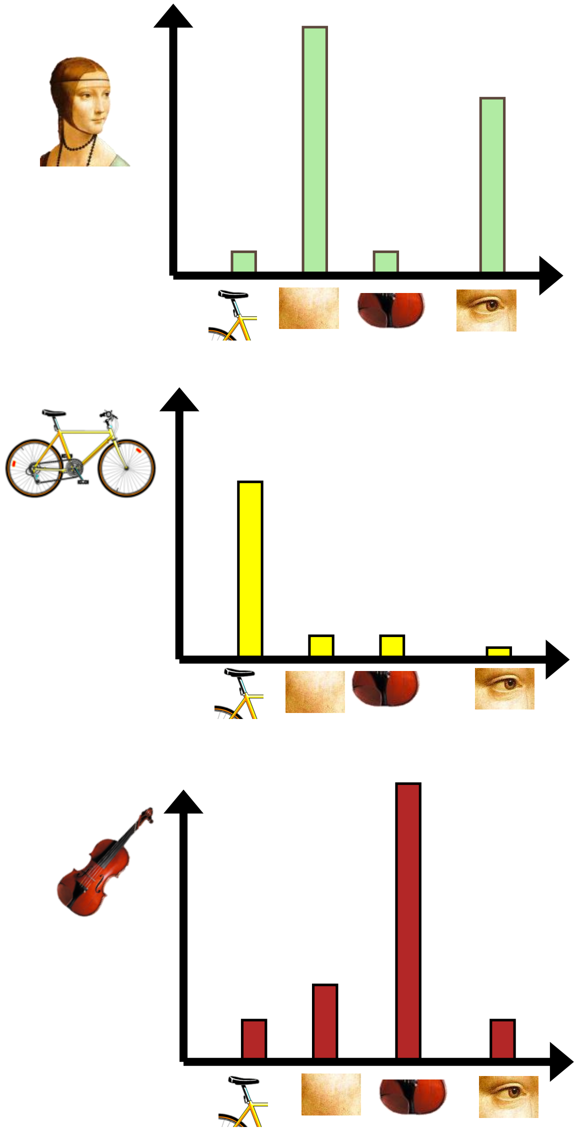
Julesz, 1981; Cula & Dana, 2001; Leung & Malik 2001; Mori, Belongie & Malik, 2001; Schmid 2001; Varma & Zisserman, 2002, 2003; Lazebnik, Schmid & Ponce, 2003

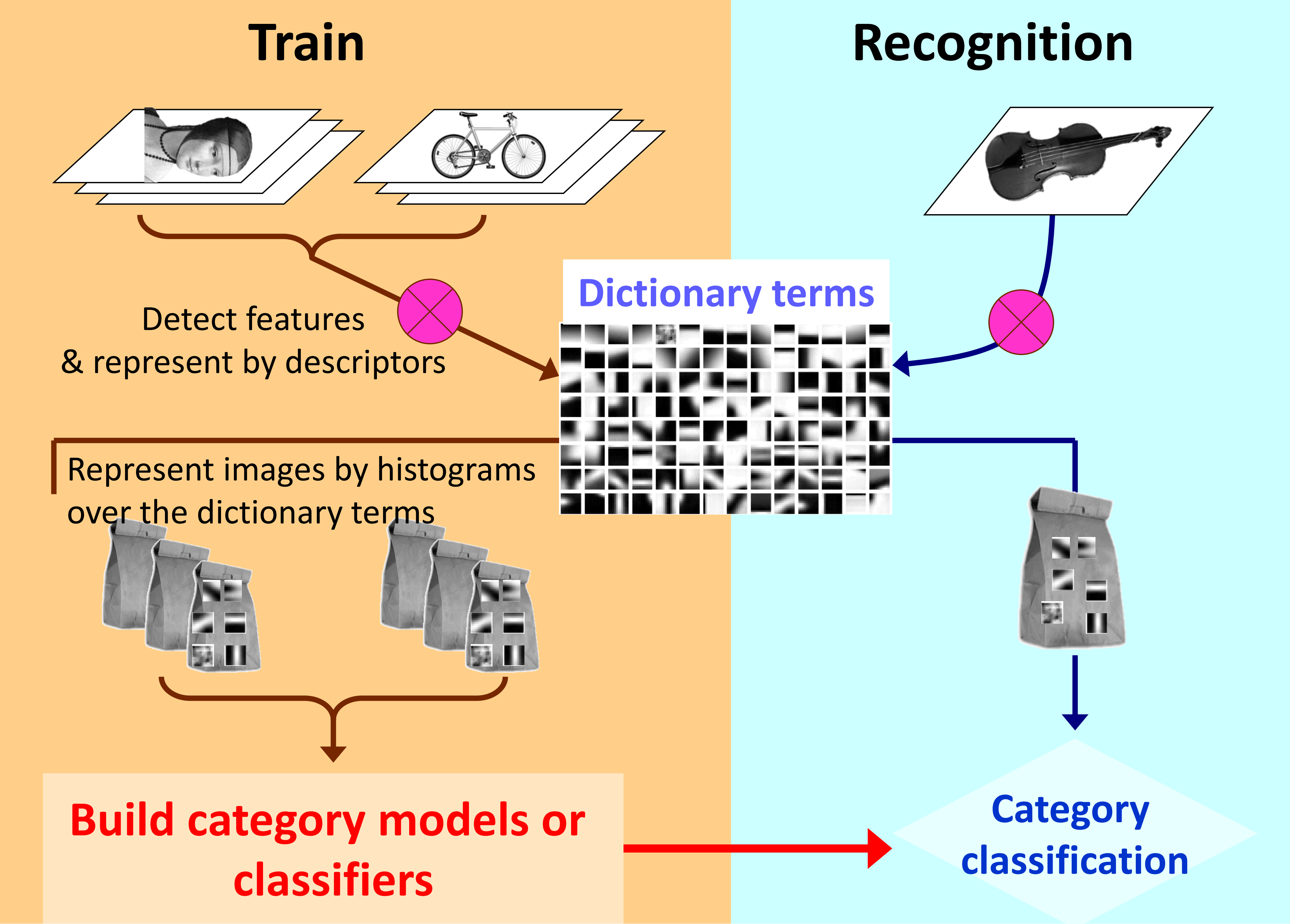
Bag of words models

Object → Bag of „words“

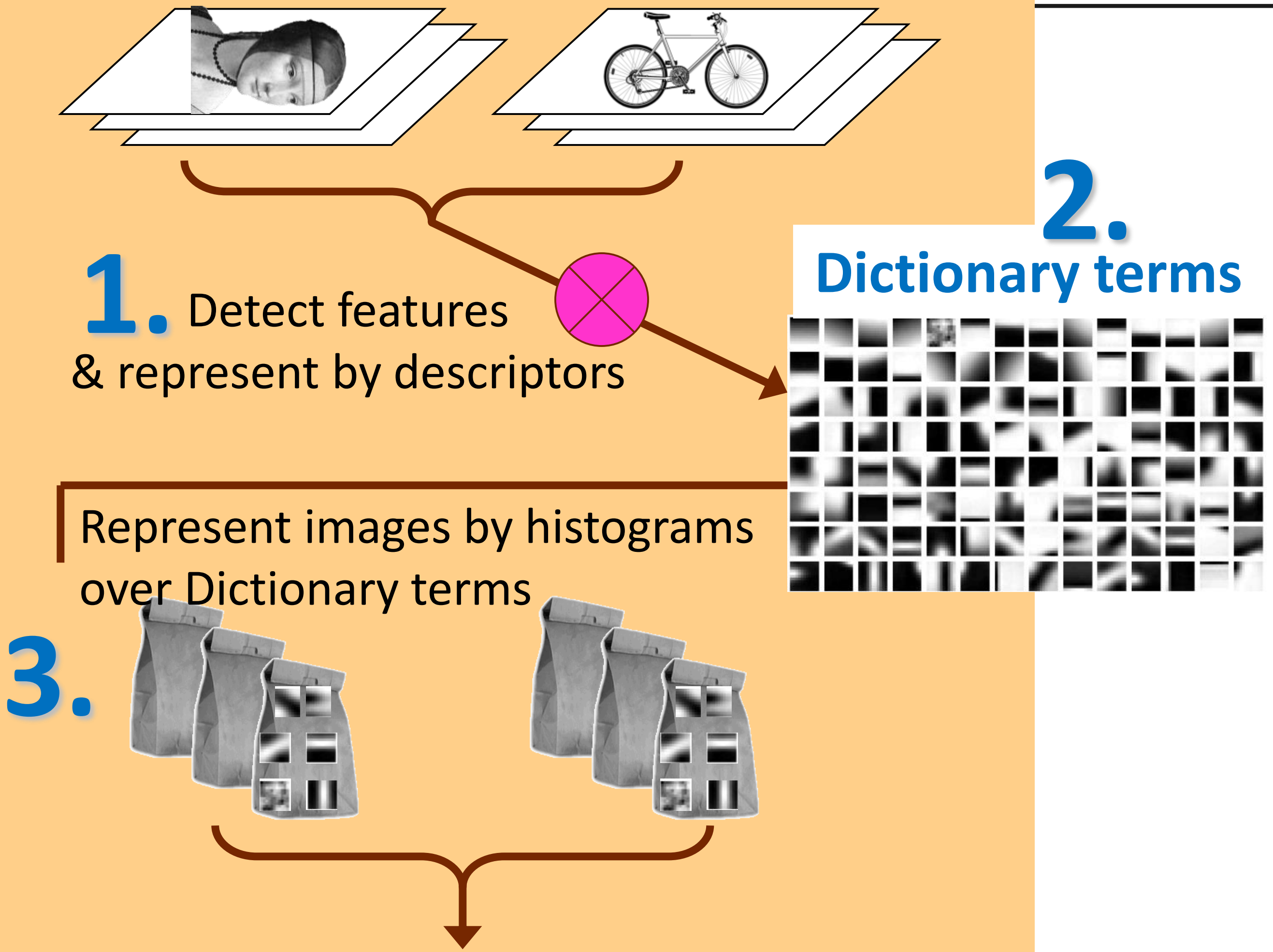
Bag of visual words

- Summarize an image by a distribution (histogram) over visual words.
- Analogous to text-based information retrieval systems – think of Google.
- Except: how to identify the "words"?



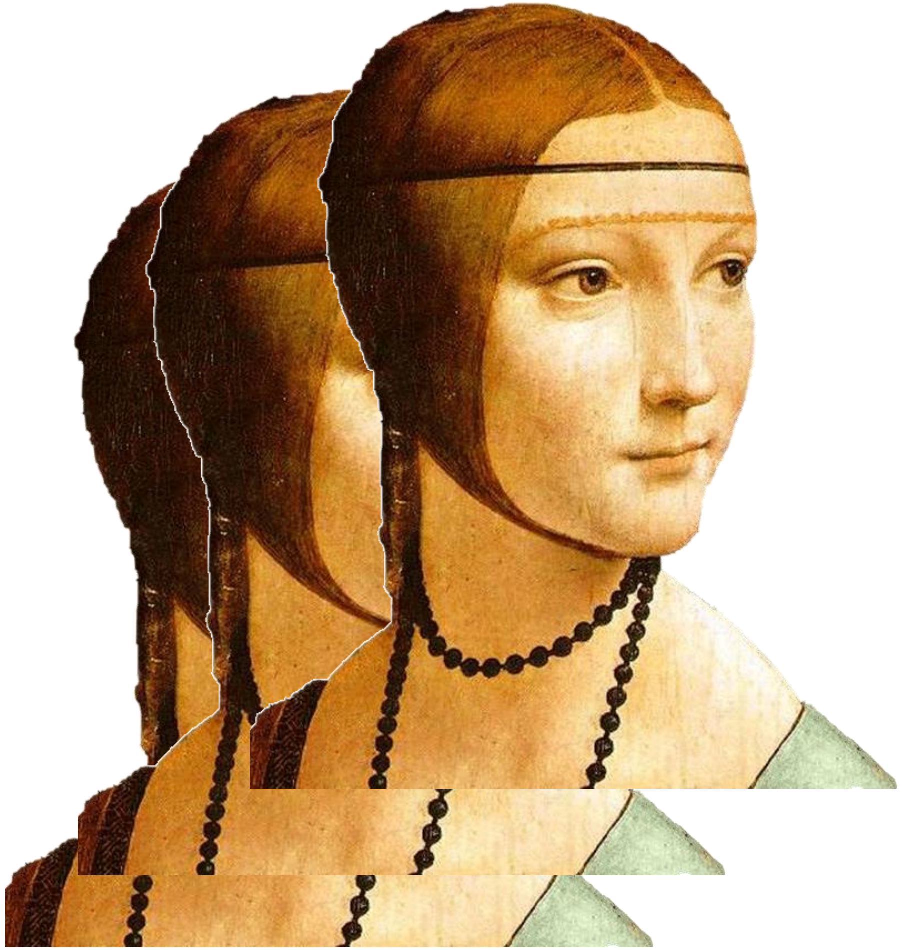

Train
Recognition
Dictionary terms
Detect features
& represent by descriptors
Represent images by histograms
over the dictionary terms
Build category models or
classifiers
Category
classification

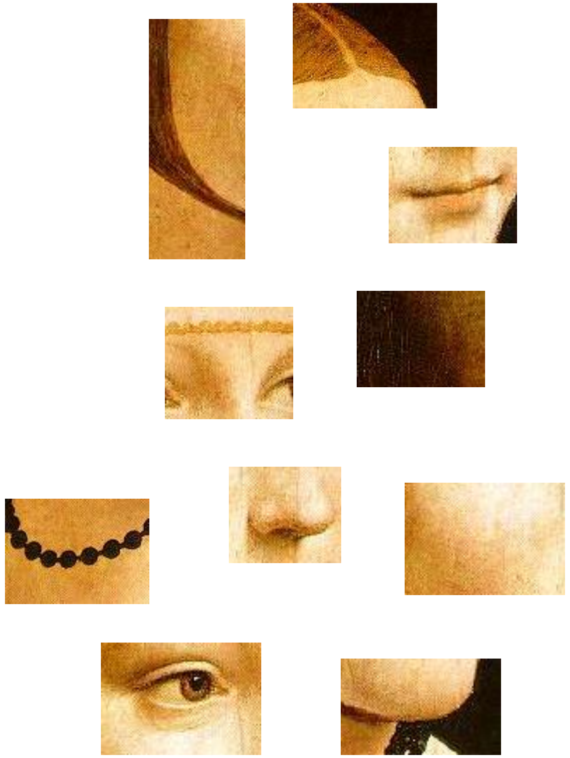
Train



Build category models or classifiers

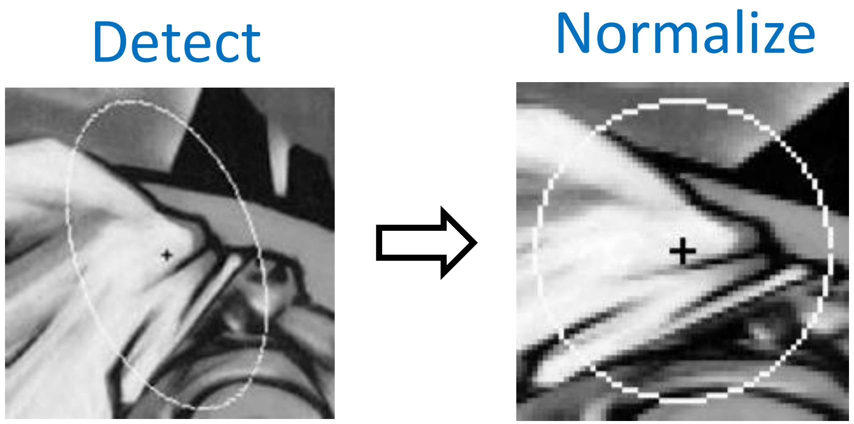
1. Feature detection & representation

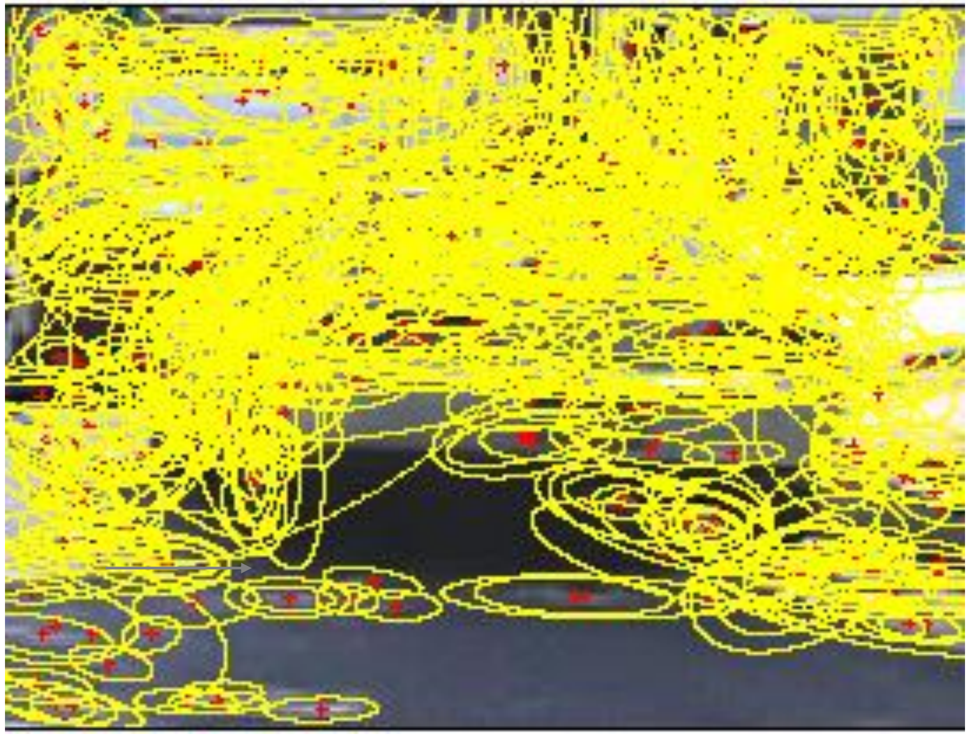




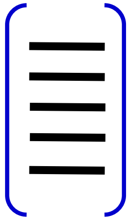
1.0 Feature detection & representation

- Use feature point detectors (we have studied quite a few)
  - E.g., SIFT
- Normalize each region to remove local geometric deformation



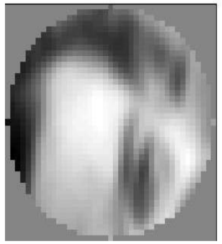


1.1 Feature detection & representation

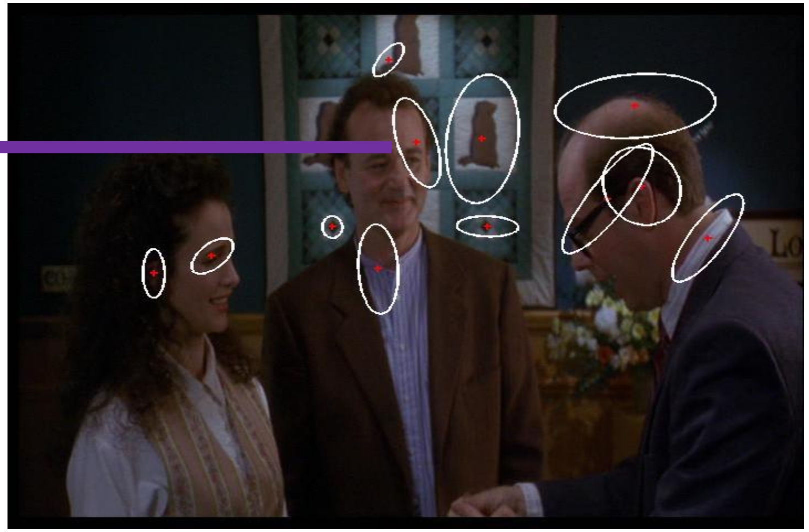


Calculate the SIFT descriptor

[Lowe'99]



Normalize the region



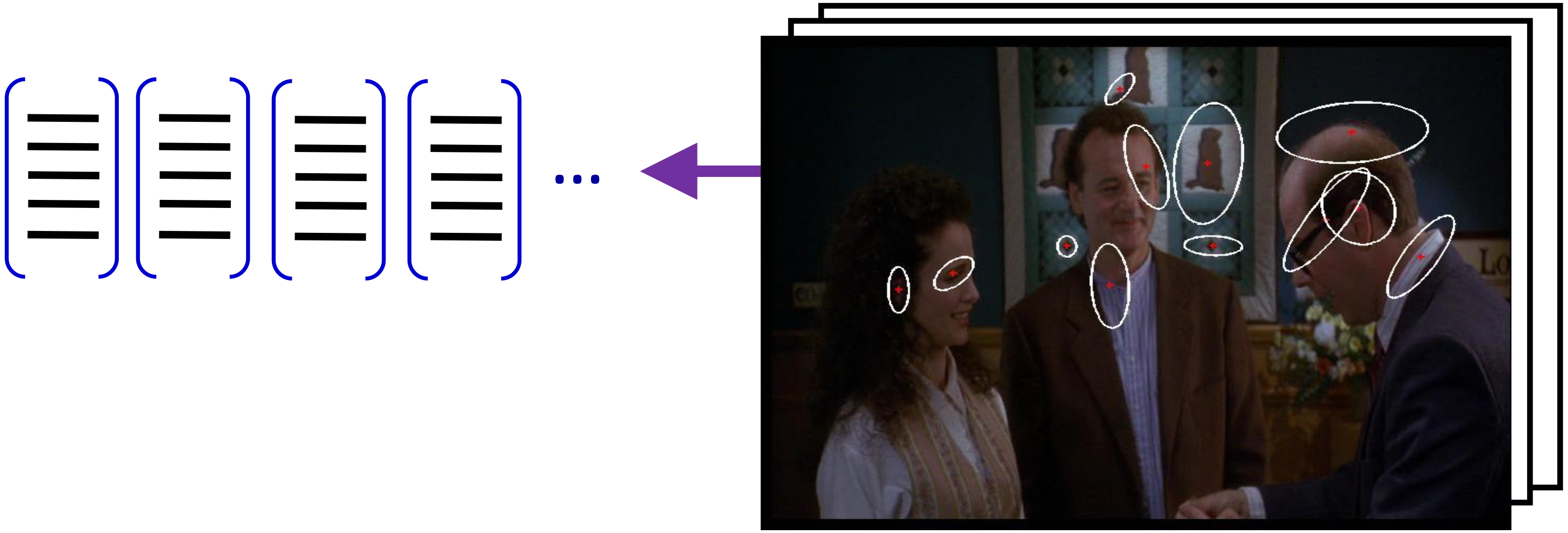
Detect regions

[Mikojaczyk and Schmid '02]

[Matas et al. '02]

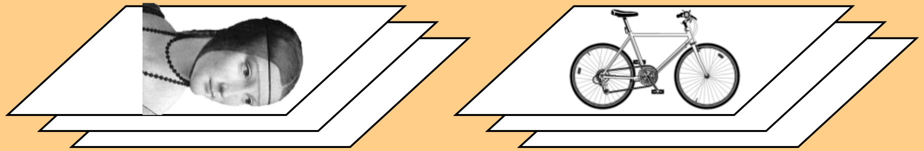
[Sivic et al. '03]

1.2 Feature detection & representation



Collect descriptors from all key-points from all training images.

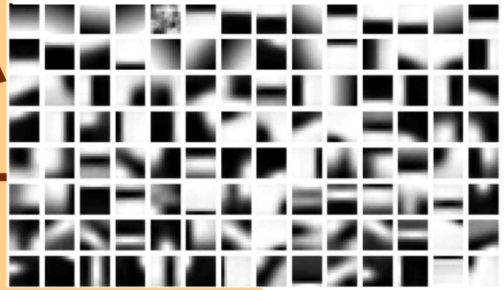
Train



Detect features
& represent by descriptors

2.

Dictionary terms

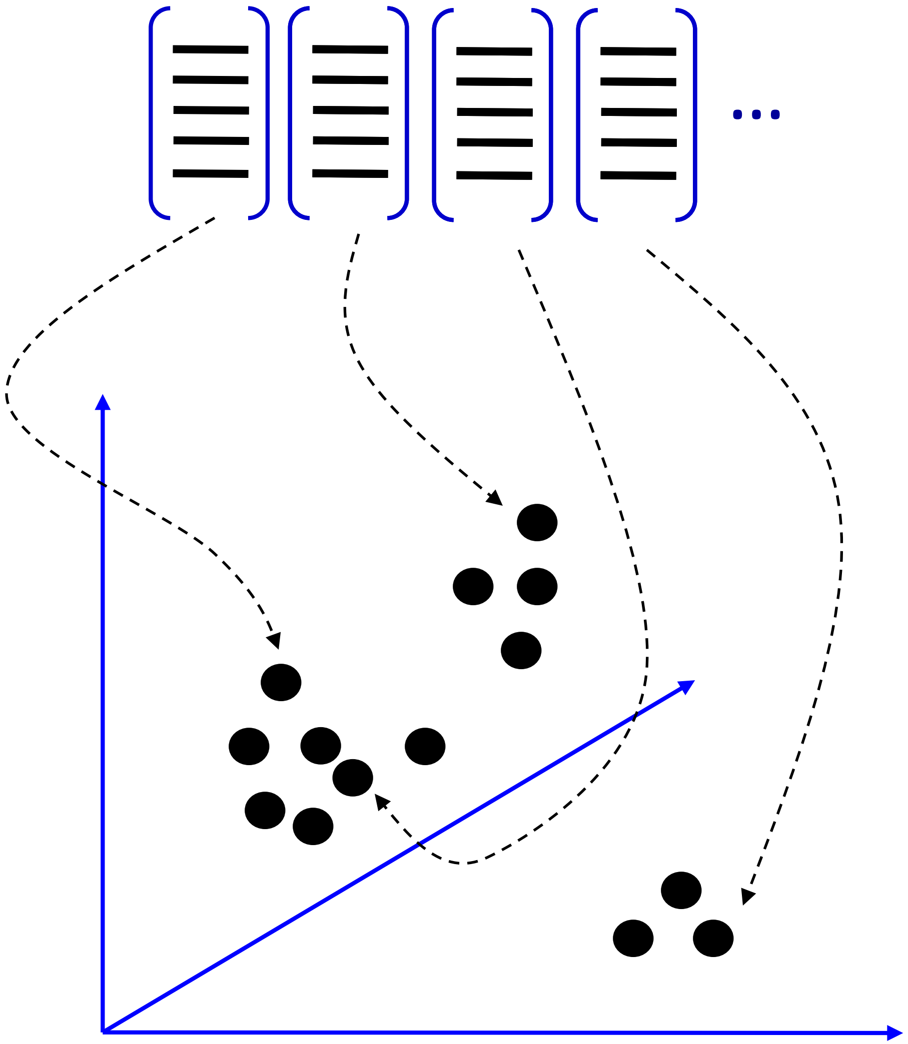


Create „word“ labels
from the extracted
SIFT descriptors...

Represent images by histograms
over Dictionary terms

Build category models or classifiers

2. Dictionary construction

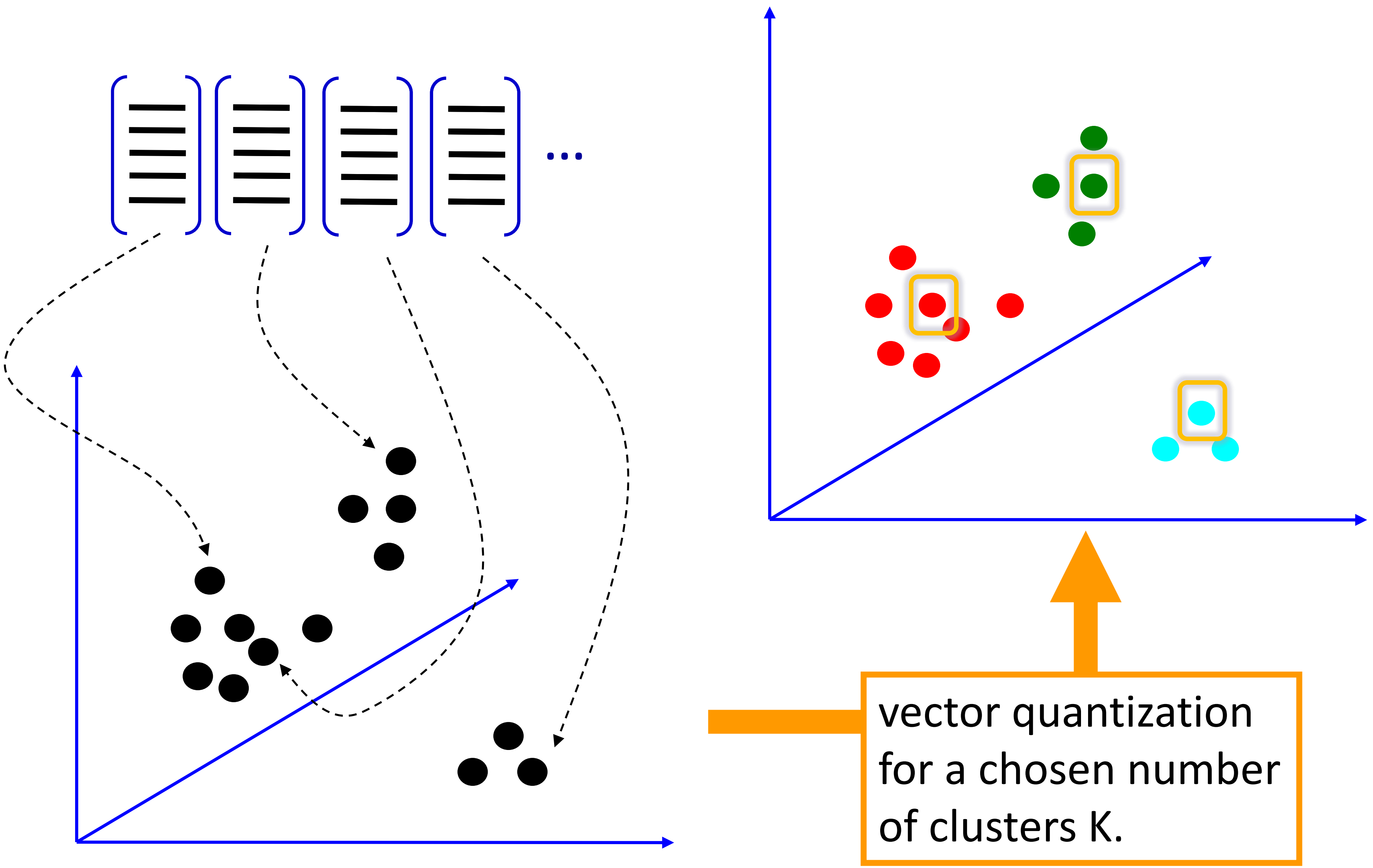


A SIFT descriptor is really a point in a high-dimensional space..

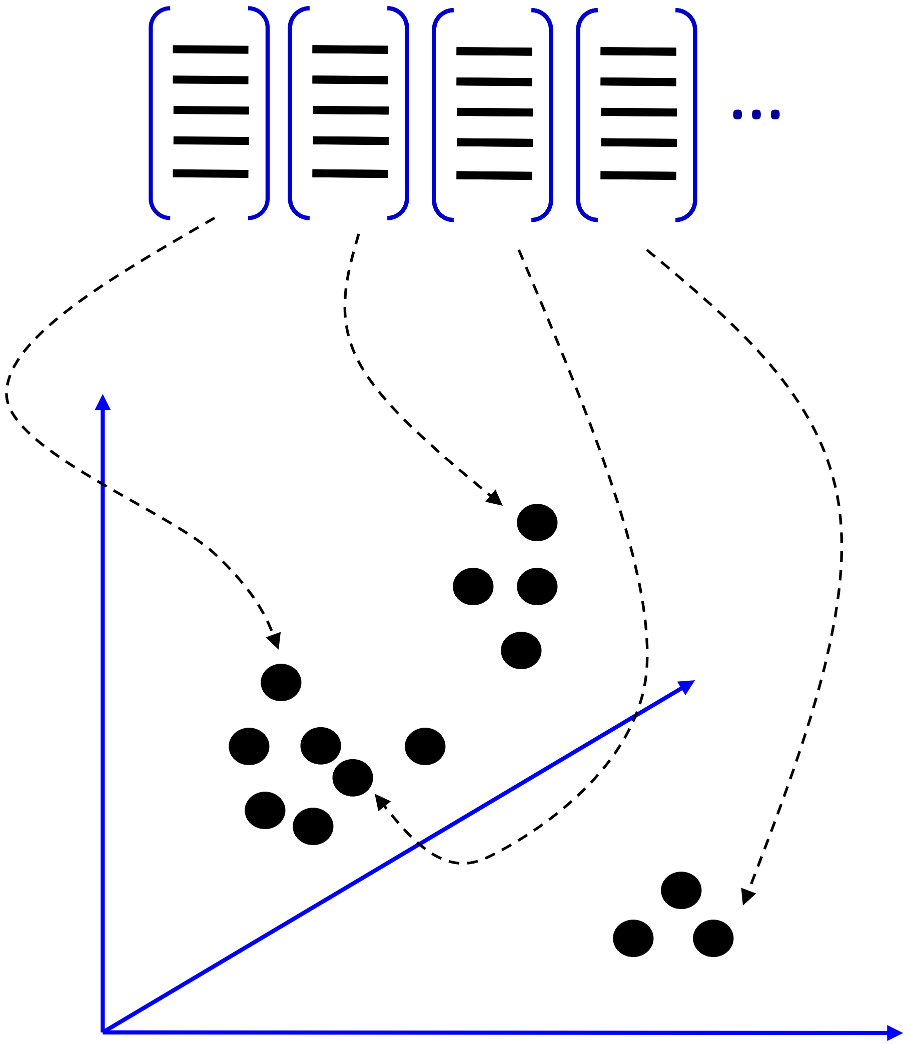
SIFTs corresponding to the same „visual word“ should be similar.

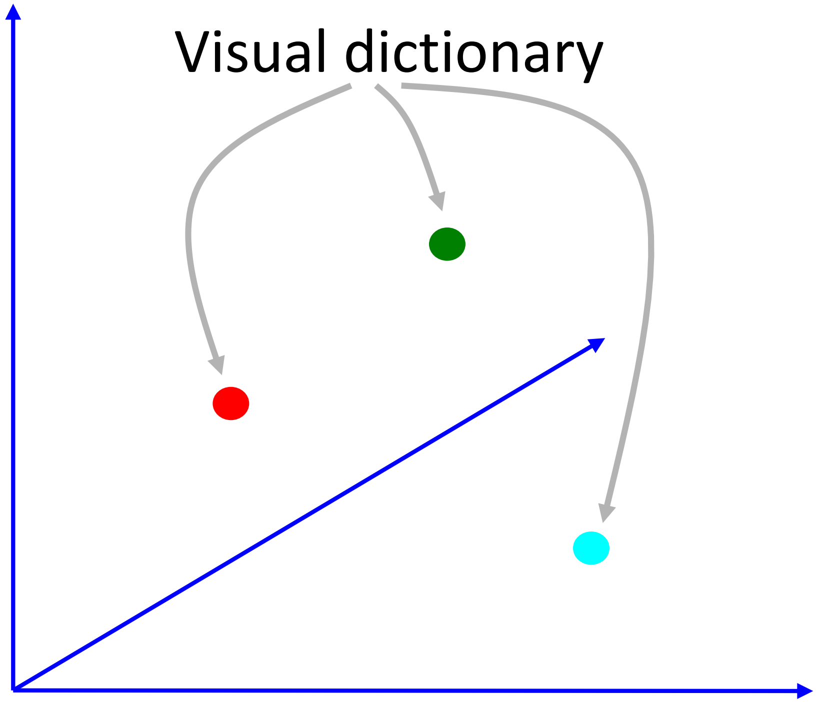
Similar SIFTs form clusters!

2. Dictionary construction



2. Dictionary construction

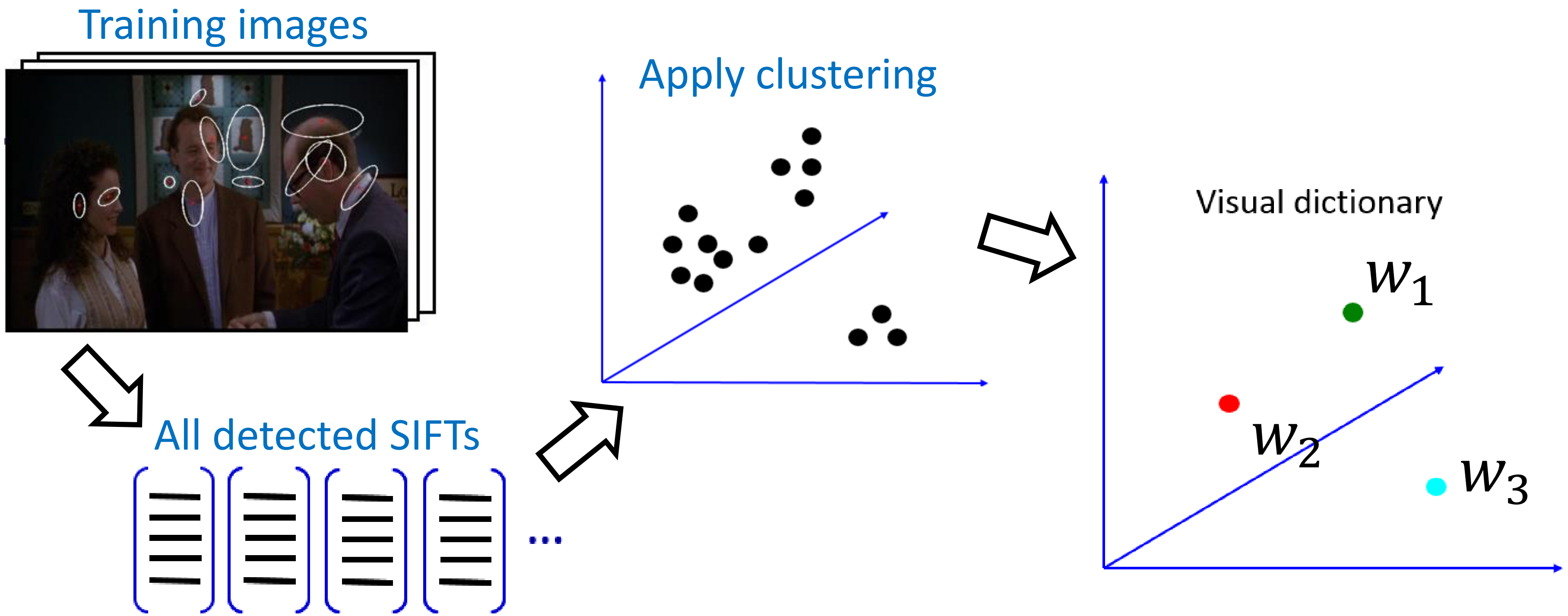




vector quantization for a chosen number of clusters K.

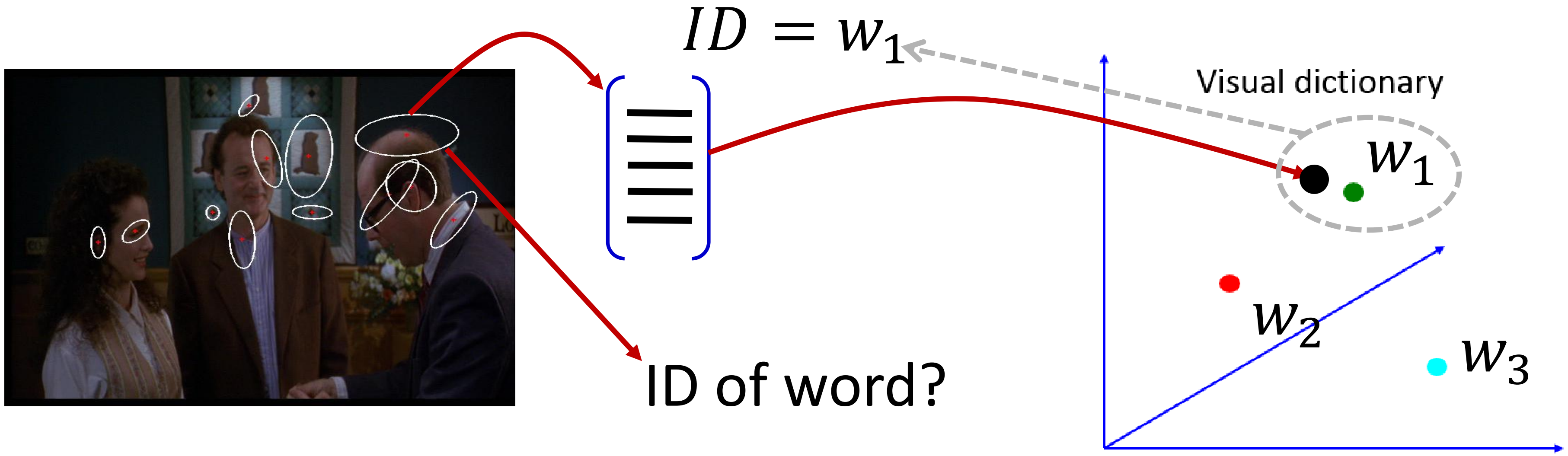
2.1 Clustering by vector quantization

- A standard approach to learning the visual codebook
  - K-means
  - Center of each cluster is the visual word (code vector)
  - Learn the code-book on separate training data (!!! This is learning stage!)

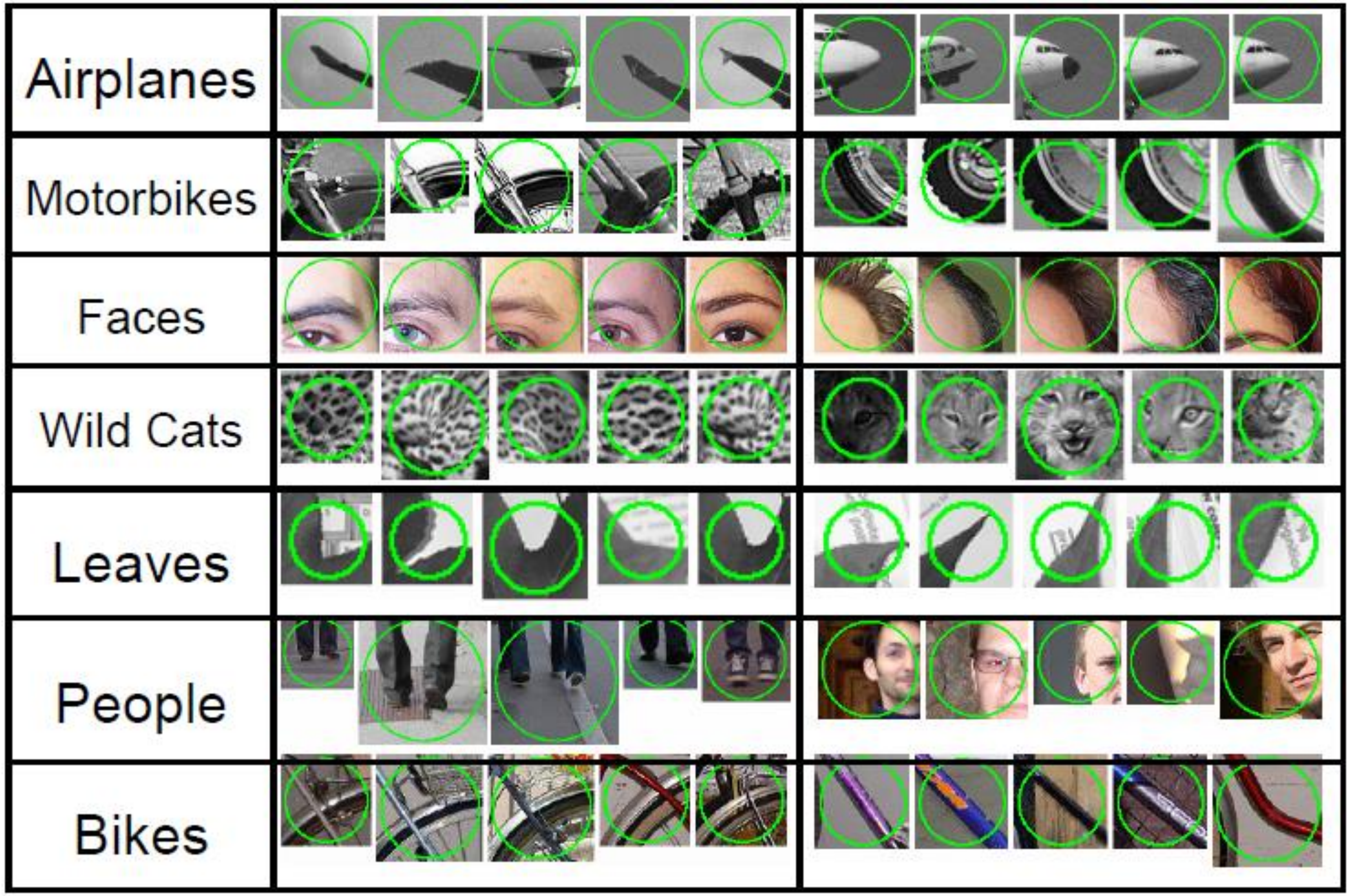


2.1 Clustering by vector quantization

- Apply codebook for feature quantization
  - Takes a feature vector (detected at key-point) and maps it to the index of the closest code vector.
  - Codebook = visual dictionary (vocabulary)
  - Code vector = visual word



2.2 Visual dictionary – example

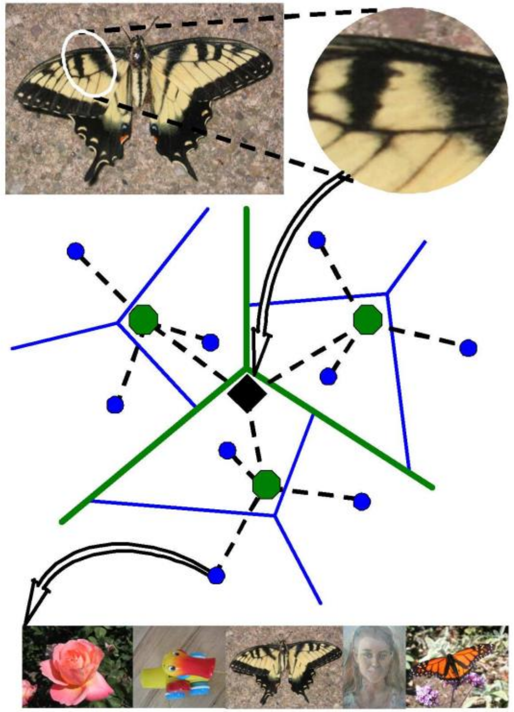


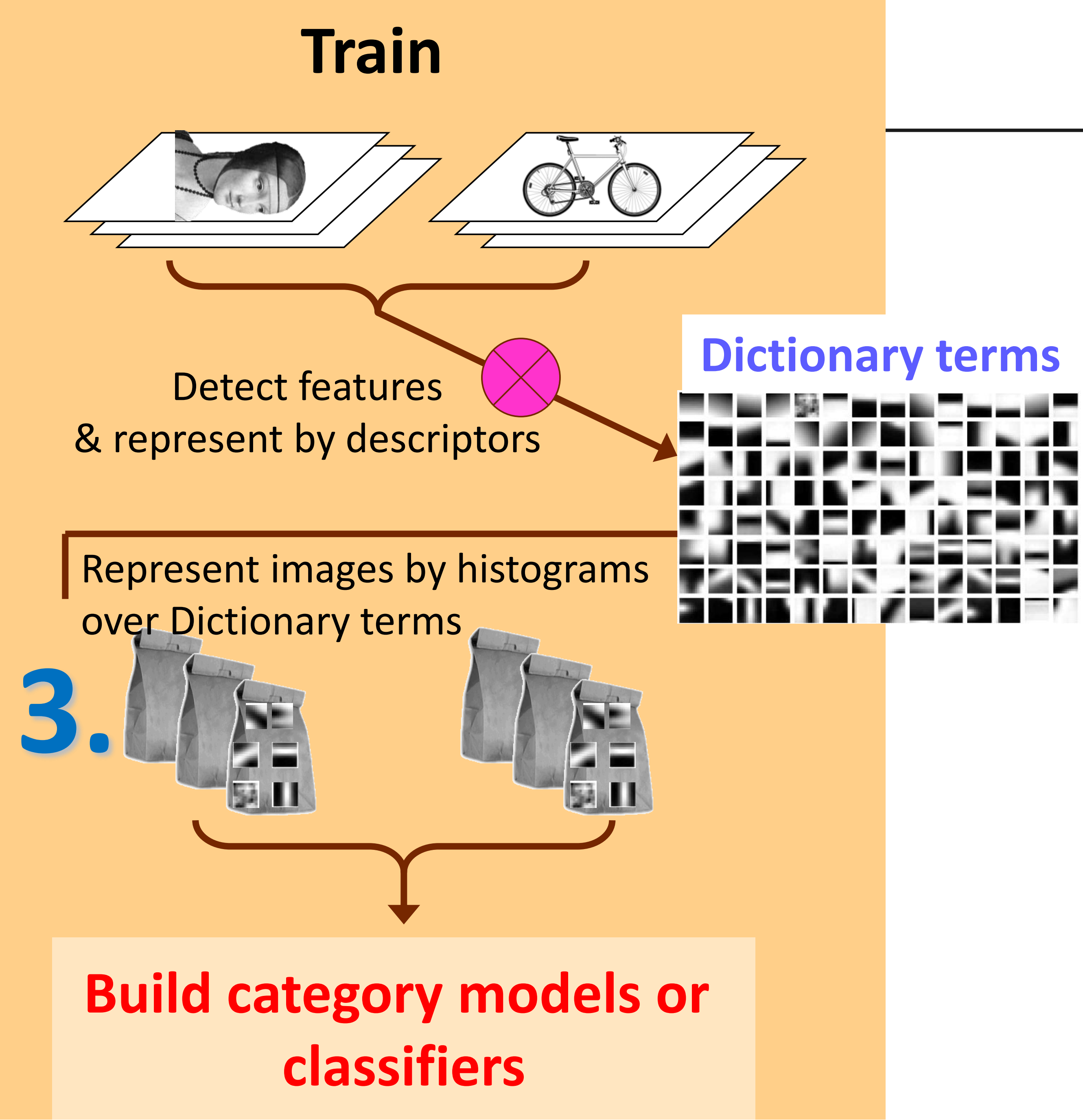
2.2 Visual dictionary – issues

- How to choose dictionary size?
  - Too small: visual words not expressive enough to describe all possible patches.
  - Too large: visual words too similar to discriminate well

- Computational efficiency in matching (need to compare many keypoints to many visual words in dictionary)
  - Vocabulary trees

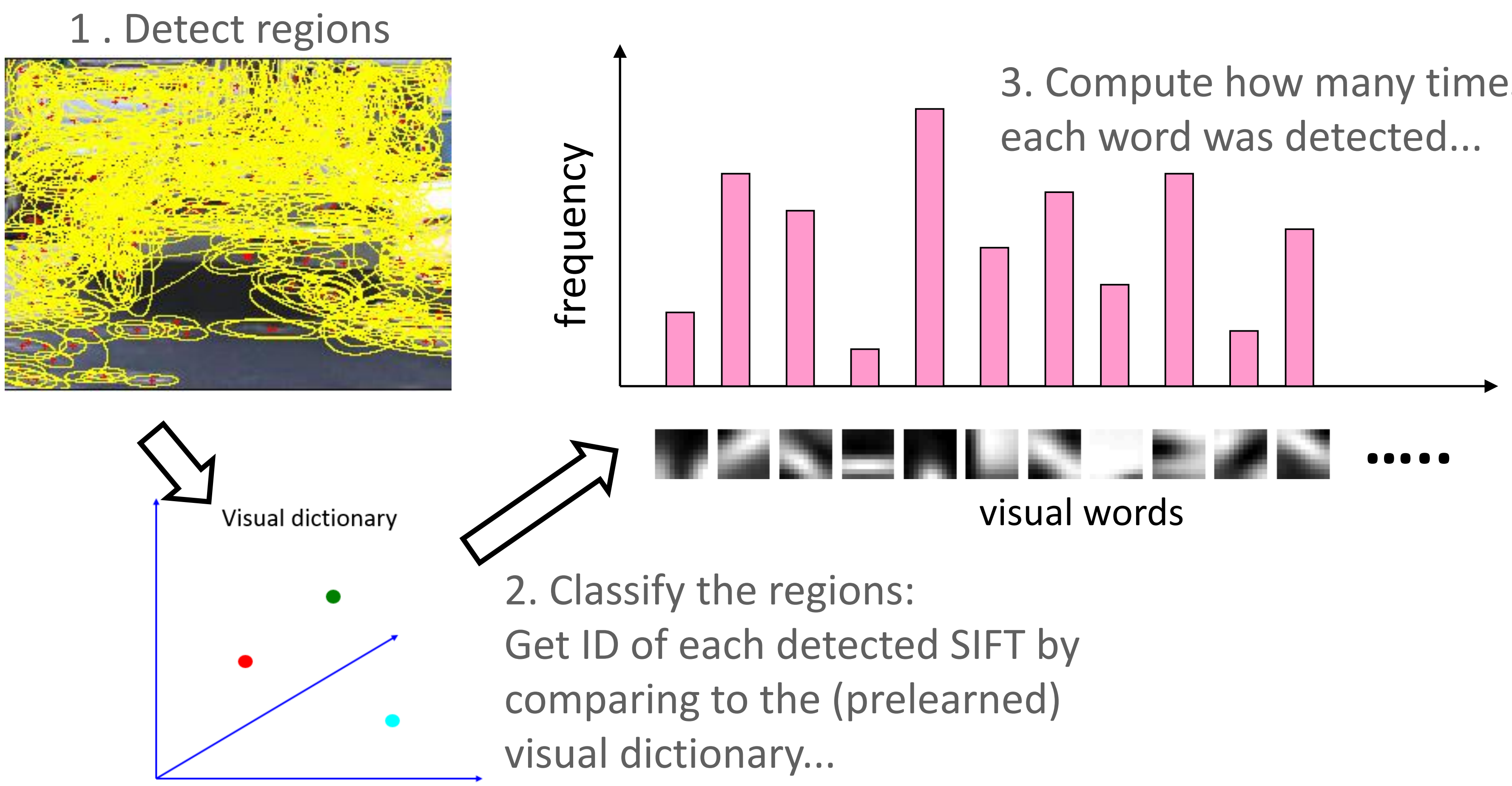
    D. Nistér and H. Stewénius, *“Scalable recognition with a vocabulary tree,”* in Proc. CVPR, 2006



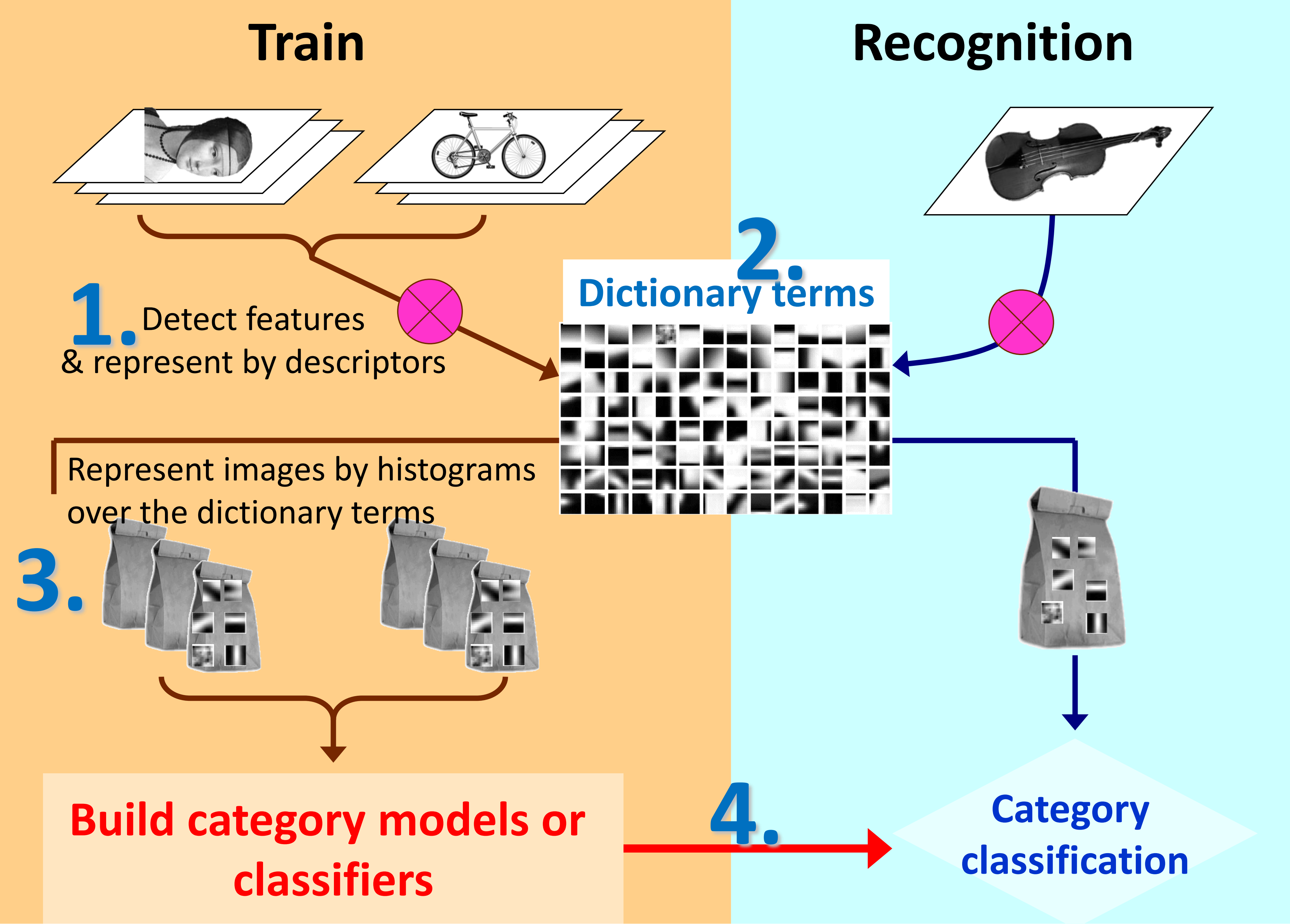


Slide credit: Li Fei-Fei

3. Image representation

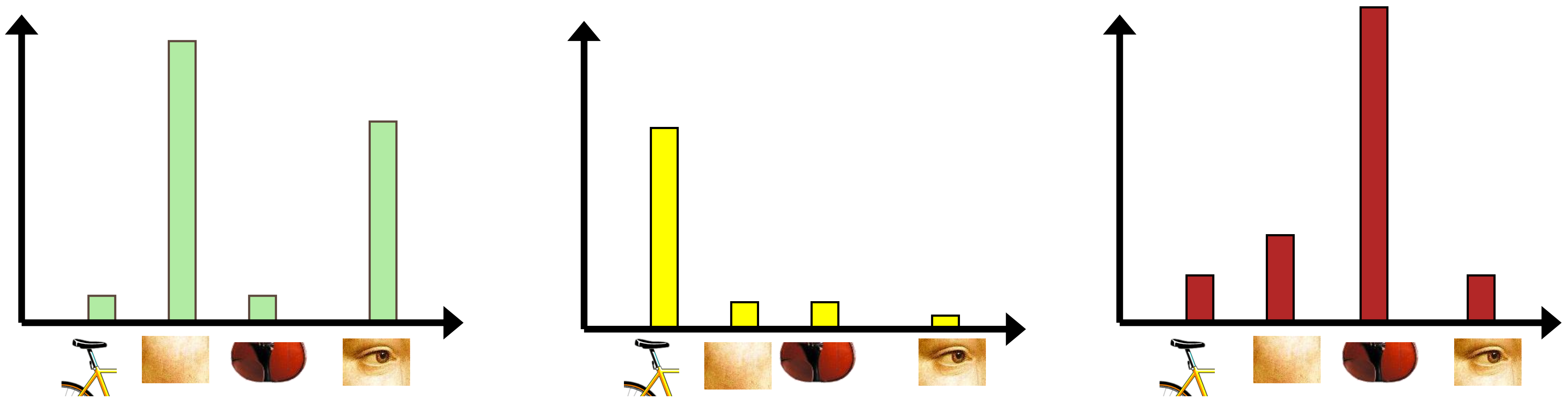


- Each image is represented by a 1000-4000 dimensional histogram, which is then normalized (L1/L2 norm)

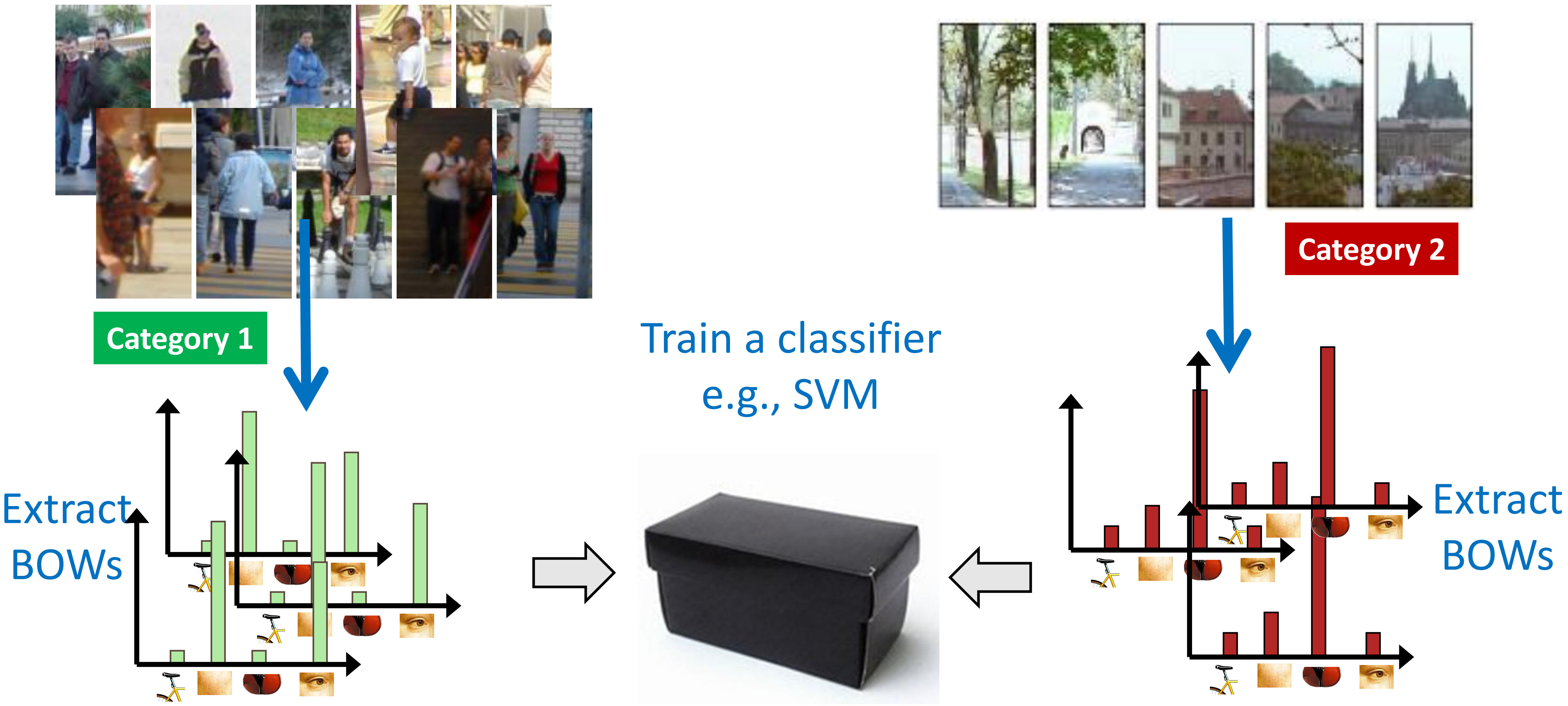

Train
Recognition
1. Detect features
& represent by descriptors
2. Dictionary terms
Represent images by histograms
over the dictionary terms
3.
Build category models or
classifiers
4.
Category
classification

4. Build a classifier

- Using the training set, we have first built a visual vocabulary.
- The vocabulary can be now used to encode any image with the histogram
- As the final stage of learning, we need to train a classifier that will classify images based on the extracted bag of word histograms.

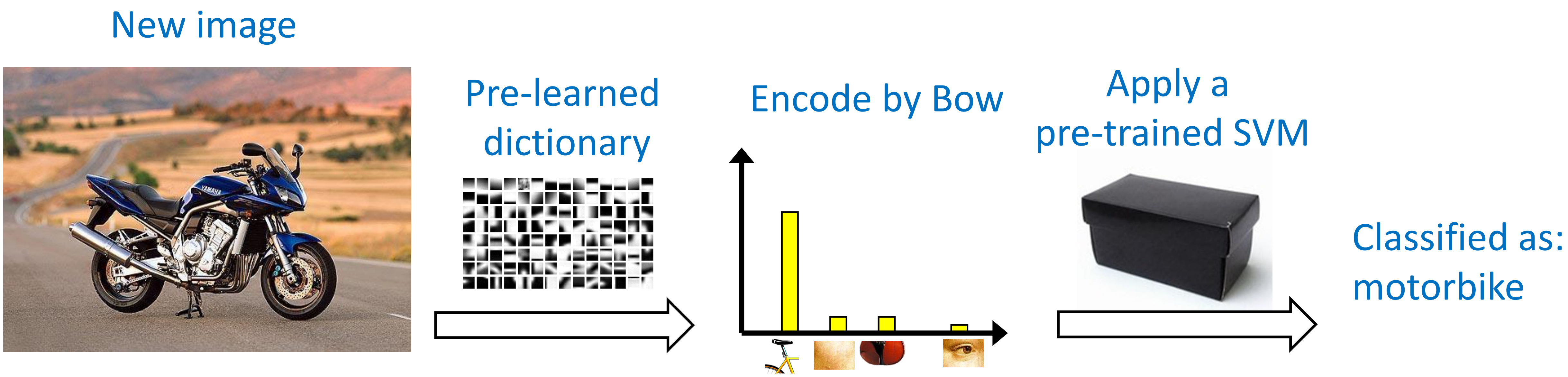


4.1 Build a classifier by SVM



5. Recognition

- How to classify a new image?
- Encode the image with the dictionary learned in the training stage
- Feed to a classifier trained at training stage



6. BoW application in practice

- Performs very well in image classification despite the background clutter...

6.1 Examples of false classification

Books classified as faces and buildings

Buildings classified as faces and trees

Cars classified as buildings and phones

6.2 Bags of words: Summary

- Strengths:
  - Fixed descriptor length.
  - Robust to object position and orientation

- Weaknesses:
  - Does not account for spatial relations among visual words.
  - Does not localize objects in the image.

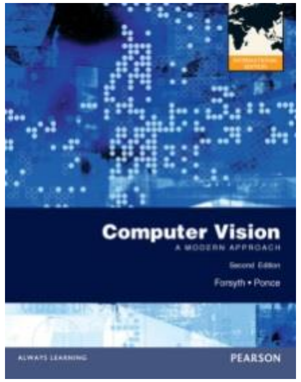
Machine perception

OBJECT DETECTION BY FEATURE CONSTELLATIONS

Detection as a recognition problem

- How to detect an object in arbitrary pose and estimate that pose?
- Brute force sliding windows not always a good option*.

*Actually, modern deep learning detectors can be considered as sliding window operations…



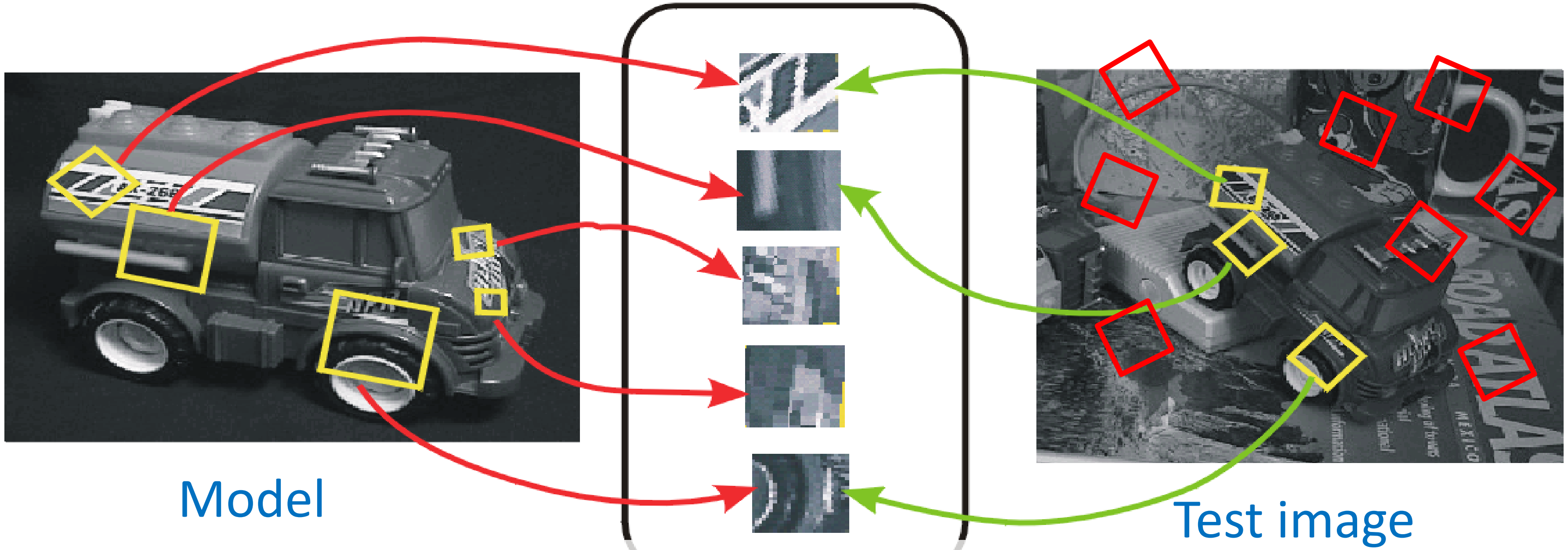
scale

rotation

perspective

Detection as a recognition problem

- Represent target model in terms of small “parts” that can be detected even under an affine deformation
- Detect “parts” in image (detection should be invariant to rotation and scale)
- Verify consistency of geometric configurations



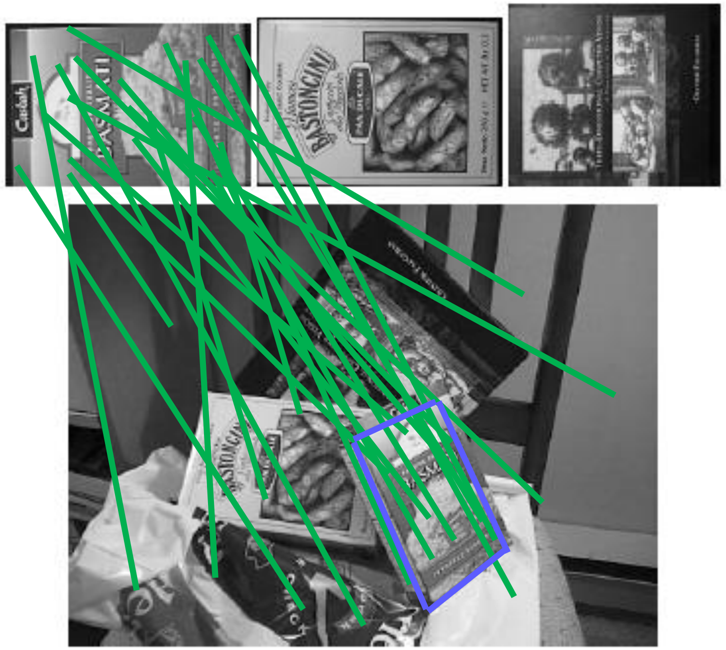
Model

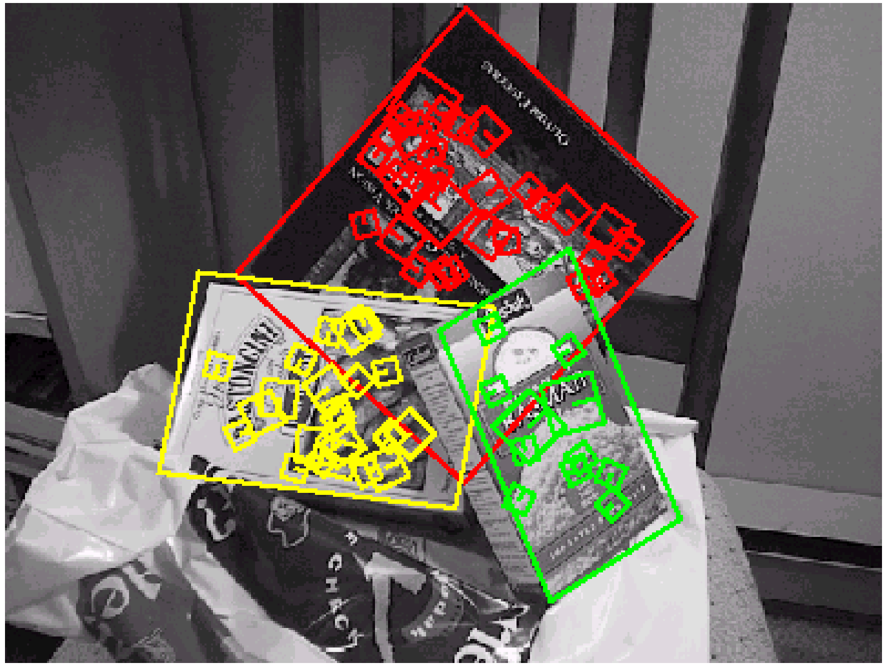
Local descriptors, e.g., SIFT

Test image

Fitting an affine deformation

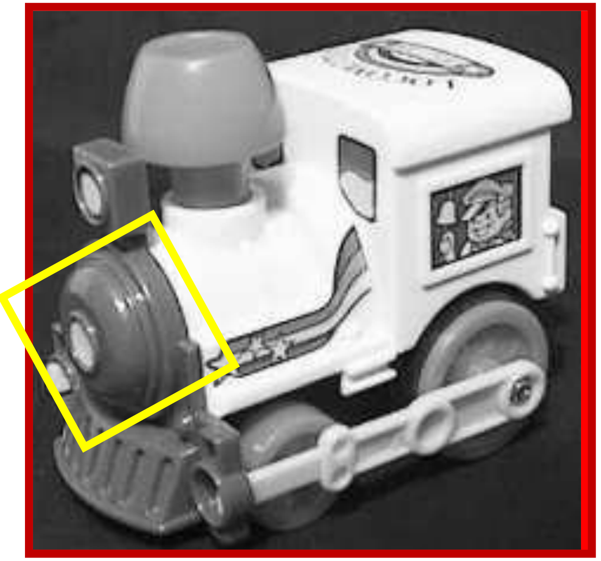
- Affine model approximates perspective transform of planar objects.
- Apply RANSAC to get a globally-valid correspondence.





Detection by Generalized Hough Transform

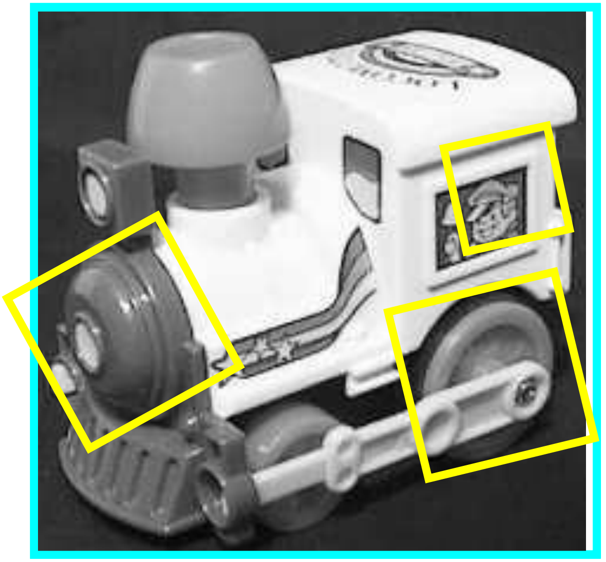
- Assume features are invariant to scale and rotation
  - Then each detected feature becomes a hypothesis of fitting (translation, rotation, scale)
- Each feature casts a vote into the Hough translation/rotation/scale space



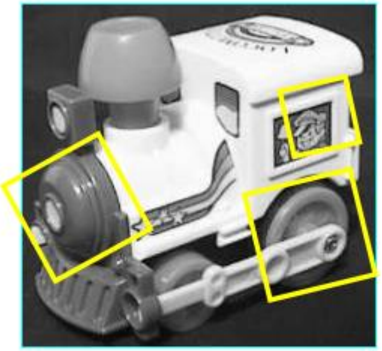
Detection by Generalized Hough Transform

- Assume features are invariant to scale and rotation
  - Then each detected feature becomes a hypothesis of fitting (translation, rotation, scale)
- Each feature casts a vote into the Hough translation/rotation/scale space

model



GHT detection refinement

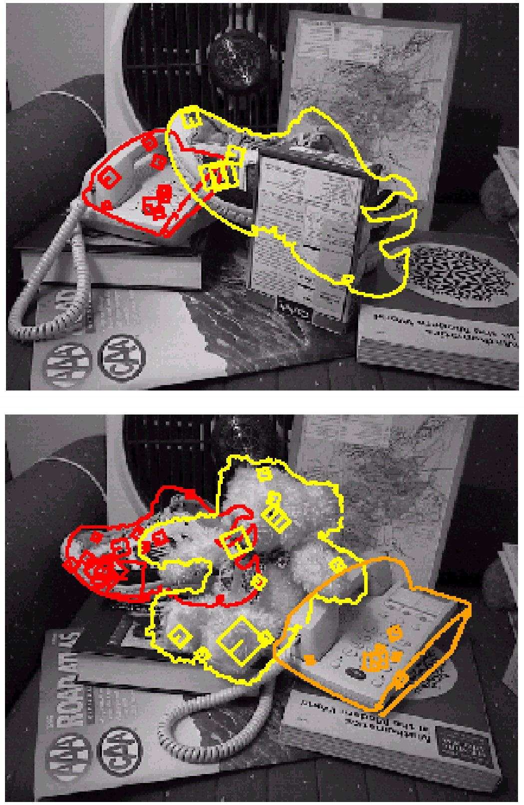


1. Index descriptors
   - Distinctive descriptors reduce the search space
2. Apply a generalized Hough transform (GHT) to obtain approximate detections
   - Key-points associated with local transformation, relative to coordinate frame of the object.
3. Refine each detection by fitting affine transform between the points on the object and the detected points from HGT
   - Fit and verify using features, which vote for the same cell in the Hough space (at least 3 votes)

Detection results

Background subtraction
to remove background clutter
in training phase

Detected objects

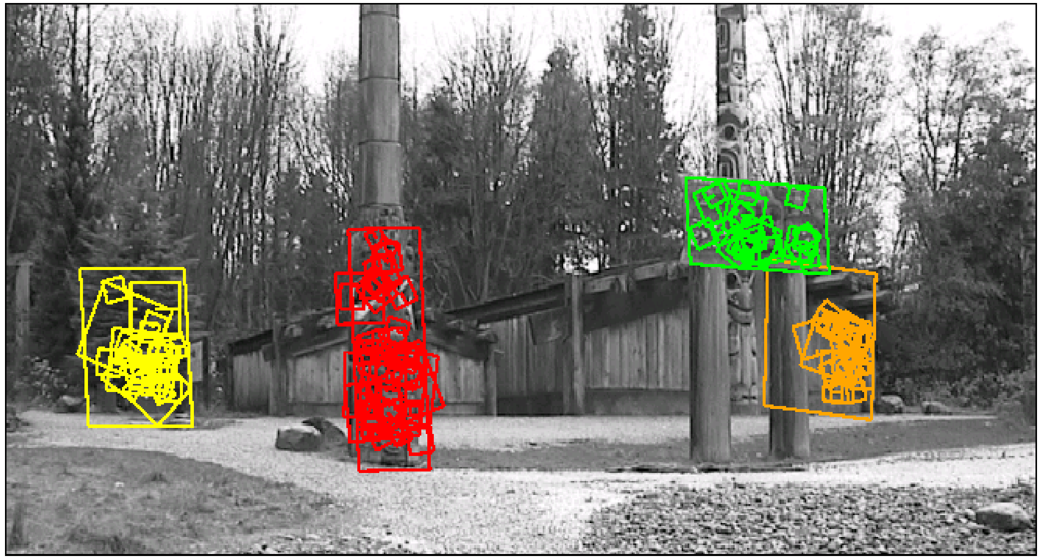


Detection despite
partial occlusion

Lowe, "Distinctive image features from scale-invariant keypoints." *IJCV* 2004.

Location recognition

Training examples of a single location



Lowe, ["Distinctive image features from scale-invariant keypoints."]() *IJCV* 2004.

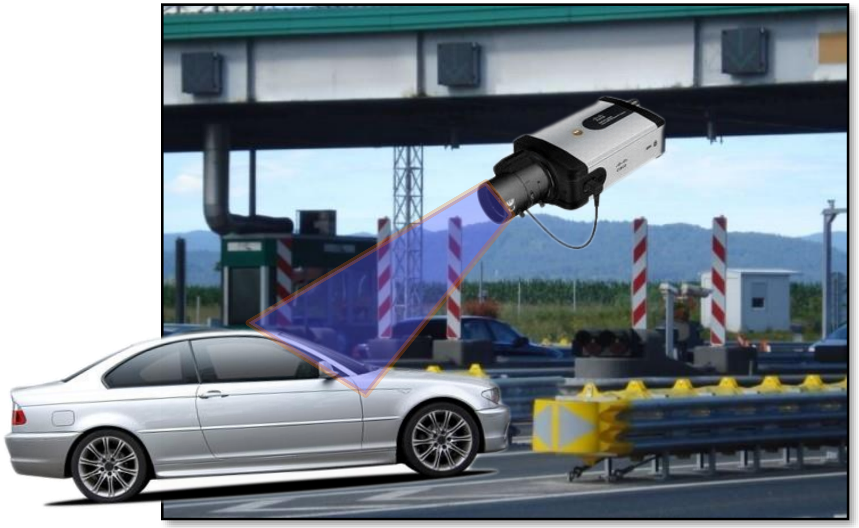
Applications: specific object recognition

- Sony Aibo
  (Evolution Robotics)

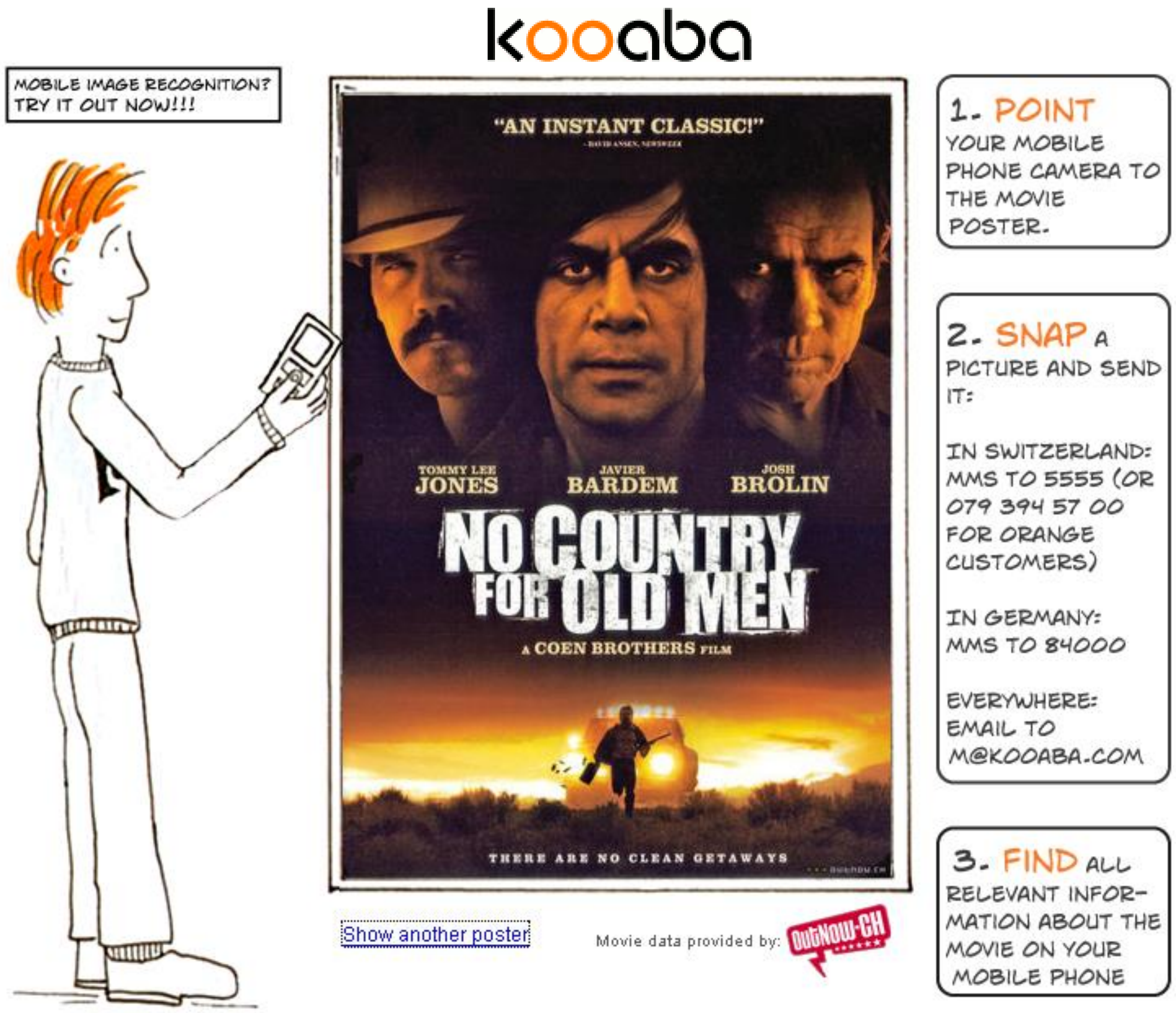
- Application of SIFT
  - Recognition of the charging station
  - Comunication using visual cards

Applications: Highway vignette verification

Highway checkpoint



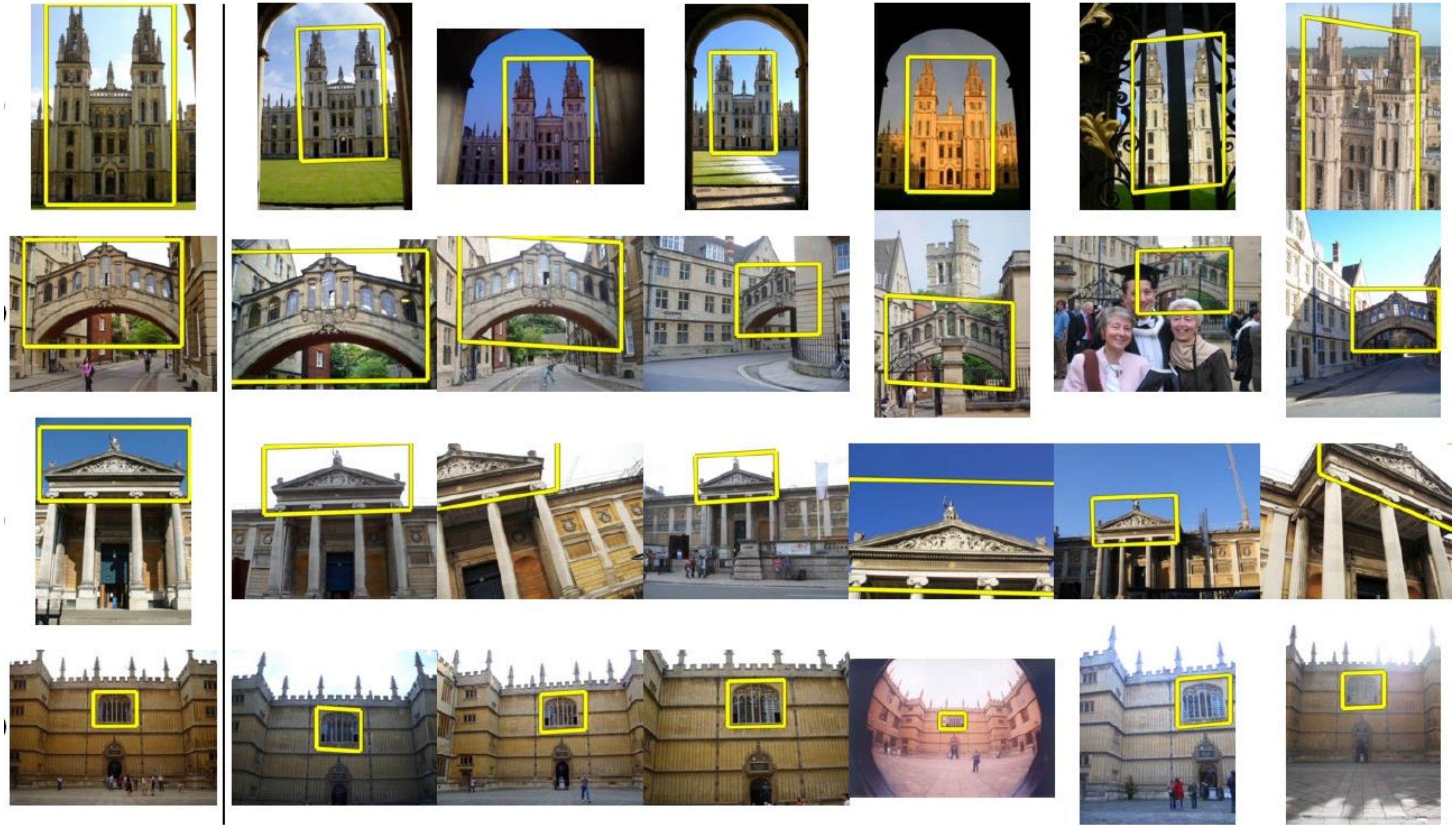
Applications: specific object recognition



http://www.kooaba.com

Applications: retrieval systems

Query Results (http://www.robots.ox.ac.uk/~vgg/research/oxbuildings/index.html)



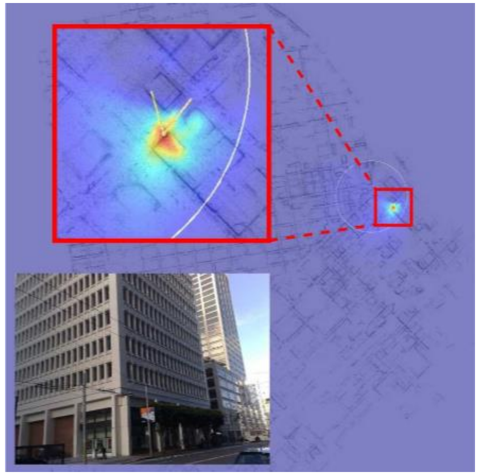
Interesting work in retrieval: Radenovic, Tolias, and Chum: CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples , ECCV 2016

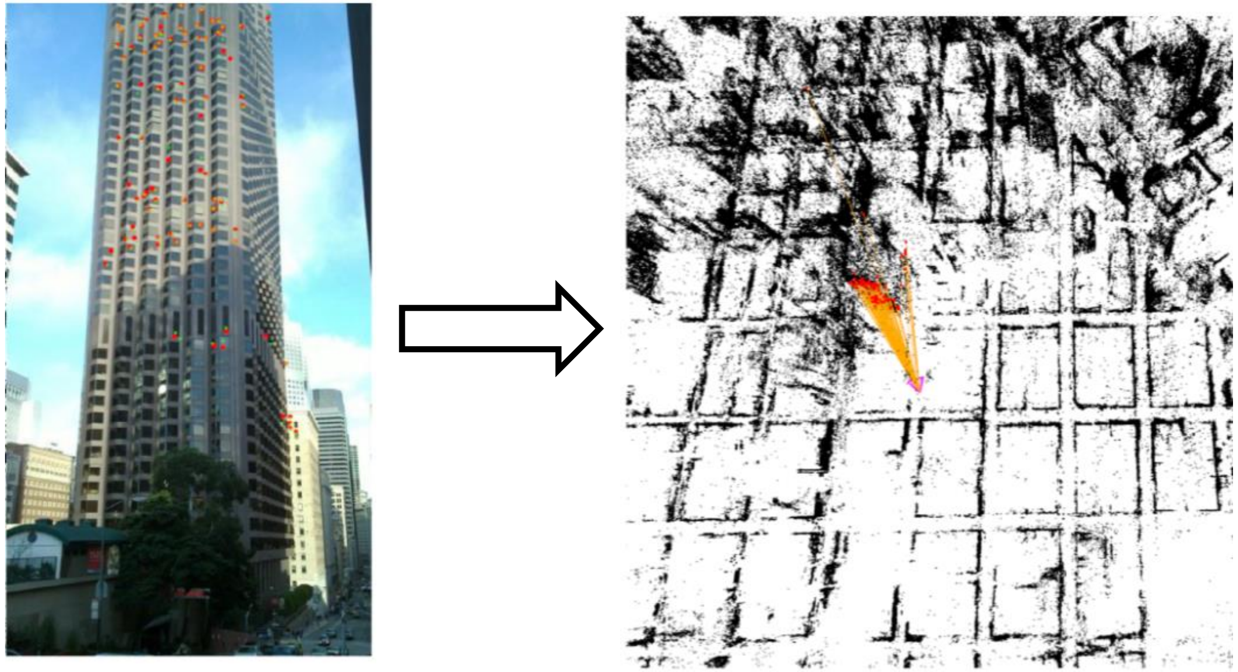
Applications: Augmented reality

- Match flat template keypoints to the scene keypoints
- Estimate camera position
- Project 3D graphic into image

Application: Large-scale pose estimation

- Use a large set of pre-recorded gps-positioned images of a city as training set (e.g., Google street-view).
- From a single image predict camera pose in the city.





Stattler et al., Hyperpoints and Fine Vocabularies for Large-Scale Location Recognition, ICCV2015
Zeisl et al., Camera Pose Voting for Large-Scale Image-Based Localization, ICCV2015

References

- David A. Forsyth, Jean Ponce, Computer Vision: A Modern Approach (2nd Edition), (prva izdaja dostopna na spletu)
- Li Fei-Fei (Stanford), Rob Fergus (NYU), Antonio Torralba (MIT) , Recognizing and Learning Object Categories, (na spletu)
- Cordelia Schmid, Bag-of-features for category classification, lecture
- Lazebnik, Schmid, Ponce, Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, CVPR, 2006
- Lowe, "Distinctive image features from scale-invariant keypoints." *IJCV* 2004

Machine perception

SUMMARY AND OUTLOOK

What did we learn?

- (1,2) Basic image processing
  - Thresholding, Morphology, Region descriptors
  - Linear/nonlinear filter – convolution, Image pyramids.
- (3) Edge detection and image gradients
  - Image derivatives, Canny edge detector, Hough transform
- (4) Fitting models
  - Least-squares fitting (iterative, robust), Normal equations, Homogenous systems, RANSAC
- (5) Key-points and correspondences between images
  - Key-point detection in scale-space, local descriptors, SIFT

What did we learn?

- (6,7) Cameras and stereo systems
  - Pinhole camera model, Calibration, Epipolar geometry, Dense correspondence, Triangulation, Active stereo
- (8a-d) Feature learning for recognition and detection:
  - Natural linear coordinate systems: PCA, LDA (face recognition)
  - Nonlinear hand-crafted transforms: HoG+SVM (pedestrian detection)
  - Feature selection: Adaboost+integral images (face detection)
  - End-to-end feature & classifier learning: Convolutional neural nets (CNNs)

What did we learn?

- (9) Key-point-based recognition
  - Bag-of-words models.
  - Detection/recognition by RANSAC and Generalized Hough transform.

The Next Big Thing on Your List...

- The written exam – Technical details first
- COVID → online form the safest and most fair
- Outline (see e-classroom for details):
  - Zoom channel on your phone (mike&cam on, speakers muted)
  - SEB installed on your computer
  - Exam will start at the given hour SHARP!
  - You'll be ID-ed *during the exam at random.*
  - Answers written in online form, and at the end you take photos of your sketches and submit to the SEB.

The Next Big Thing on Your List...

- The written exam (see studis for dates)
  - Approx. two hours -- Covers entire course
  - Theoretical as well as analytical assignments (see the lab exercises for examples of analytical parts)
- Oral exam potentially required for low scores (X = ~50%-60%)
  - Need to know *all that you got wrong* on written exam
  - + ~2 random questions
- If >X% do not have to come to oral
    - Can if you would like to increase/decrease grade by 1 (or fail?)
- Please fill-out the poll at *studis*
  - Constructive suggestions towards improving the course

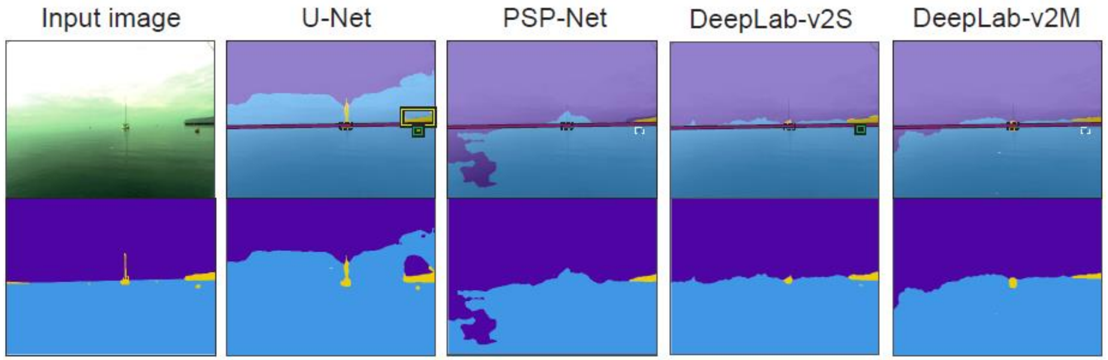
Where to go from here...

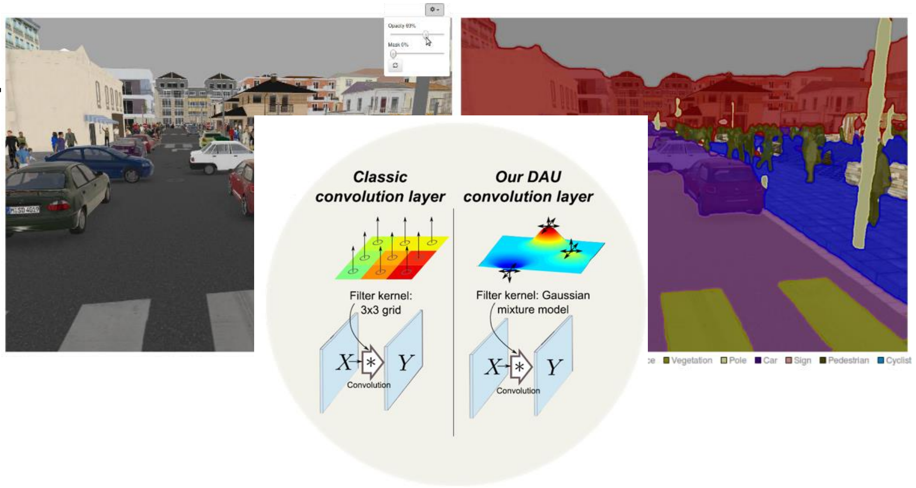
- Check out similar courses at other Universities:
  - Aachen: https://www.vision.rwth-aachen.de/course/6/
  - Stanford: http://vision.stanford.edu/teaching/cs131_fall1617/schedule.html
  - Illinois: http://slazebni.cs.illinois.edu/spring18/
  - … many more can be found on the net

Where to go from here...

- Semantic segmentation

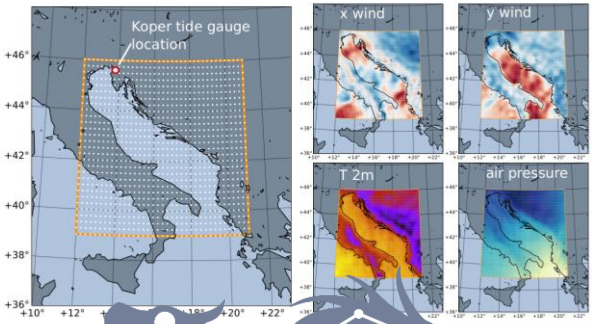
  Borja Bovcon, Matej Kristan. WaSR -- A Water Segmentation and Refinement Maritime Obstacle Detection Network, IEEE Transactions on Cybernetics, 2021





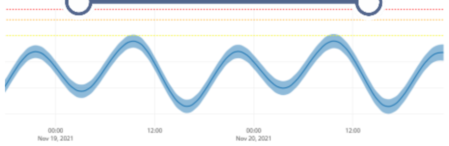
- Improvements of the CNN architectures

  Tabernik, Kristan, Leonardis, [Spatially adaptive units for deep neural networks](), CVPR2018



- Climate time series prediction & reconstruction

  Žust, Fettich, Kristan, Ličer. HIDRA 1.0 : deep-learning-based ensemble sea level forecasting in the northern Adriatic, GMD2021

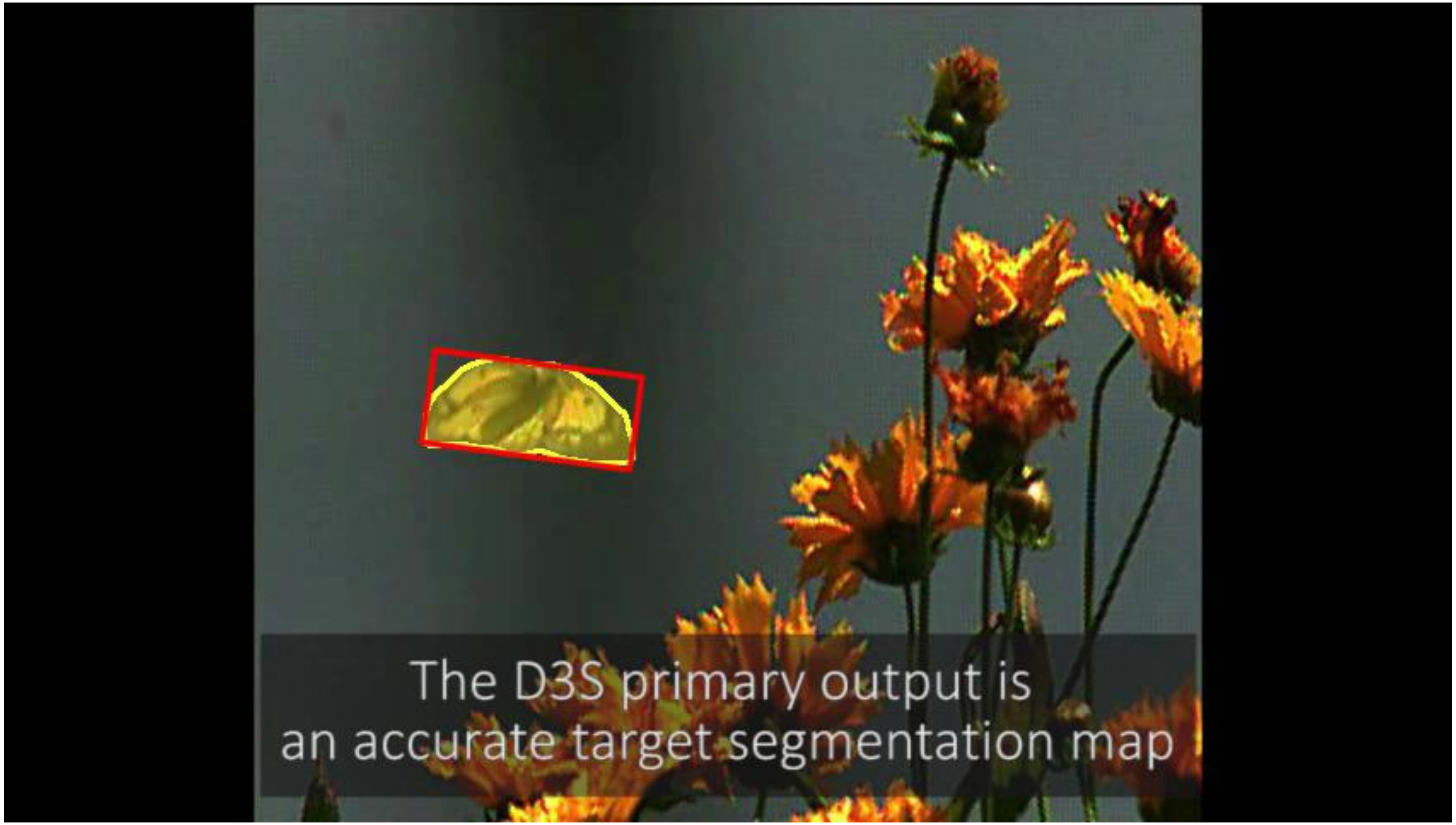


Where to go from here...

- Object tracking

VOT2020 benchmark

The VOT2020 benchmark addresses short-term, long-term, real-time, RGB, RGBT and RGBD trackers. Results were presented at ECCV2020 VOT workshop.



Lukezic, Matas, Kristan, CVPR2020

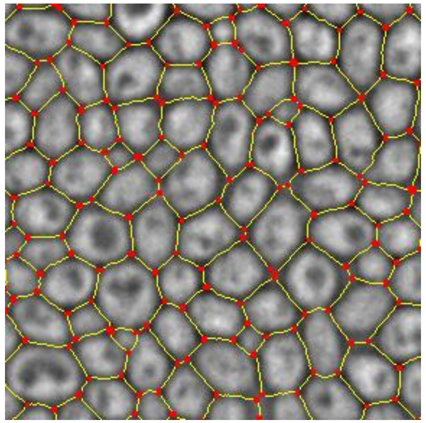
Lots of possibilites:

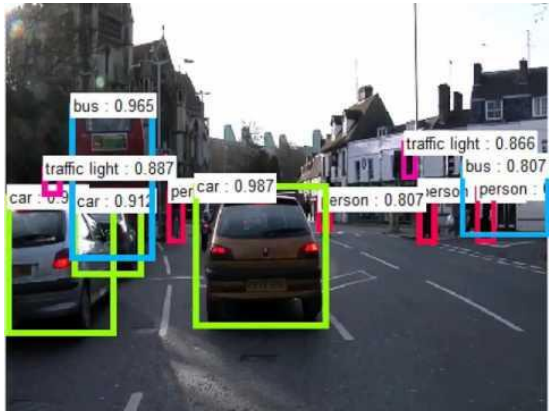
- Fast implementations
- Improvement of existing methods
- Trackers for drones ...

Where to go from here...

- Image style transfer (for domain adaptation)

- Object and category detection
  (e.g., CNN if you're interested)
- Image classification, scene classification
- Machine (industrial) vision





Where to go from here...

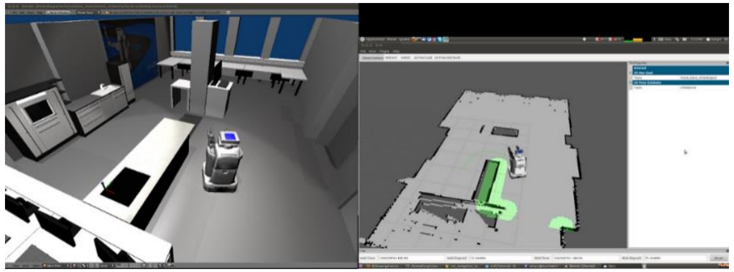
- Look for fun publications at ICCV, ECCV, CVPR – if you like, you can study one of these for your thesis.

- Your own ideas welcome!
- Those doing their thesis at Vicos can publish demo videos at the Vicos student project homepage!

- **Caution:** Historically, students either dropped out on a topic under my supervision or did A LOT of work (and hopefully finished with satisfaction)…

Other Computer-vision-oriented courses at FRI

- Bachelor's level:
  - Multimedia Systems (Luka Čehovin, Vicos)
  - Development of Intelligent Systems (Danijel Skočaj, Vicos)



- Master's level
  - Advanced computer vision methods (Matej Kristan, Vicos)
  - Deep learning (Danijel Skočaj, Vicos)
  - Image-based biometry (Peter Peer)
  - Biomedical Signal and image Processing (Franc Jager)

Thanks!

Good luck with the exam(s)!