CS688: Web-Scale Image Retrieval

Bag-of-Words (BoW) Models for Local Descriptors

Sung-Eui Yoon
(윤성의)

Course URL:
http://sgvr.kaist.ac.kr/~sungeui/IR

KAIST

Class Objectives

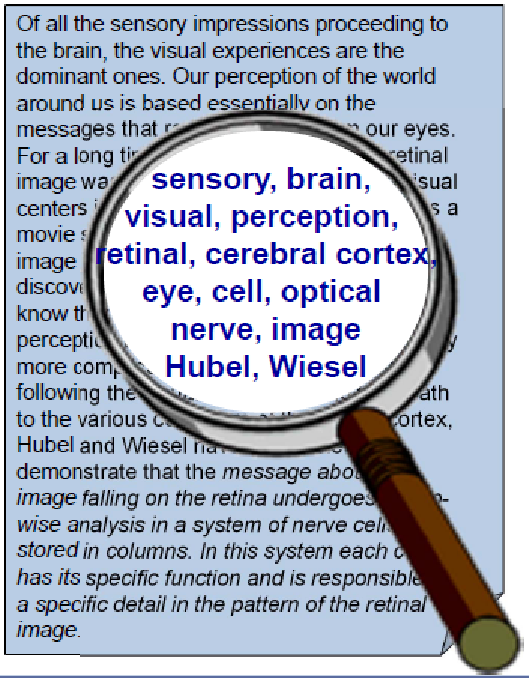
- Bag-of-visual-Word (BoW) model
  - Pooling operation
- Ranking loss for CNN features

- At the prior class:
  - Went over main components of CNNs: local connectivity and pooling

Object → Bag of 'words'

Represent an image with a histogram of words

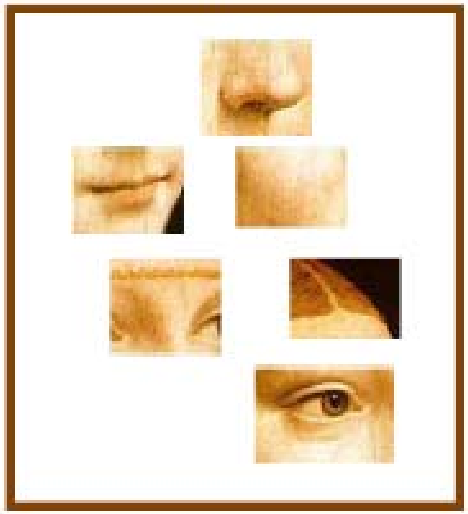
Inspired by text search



definition of "BoW"

– Independent features

face



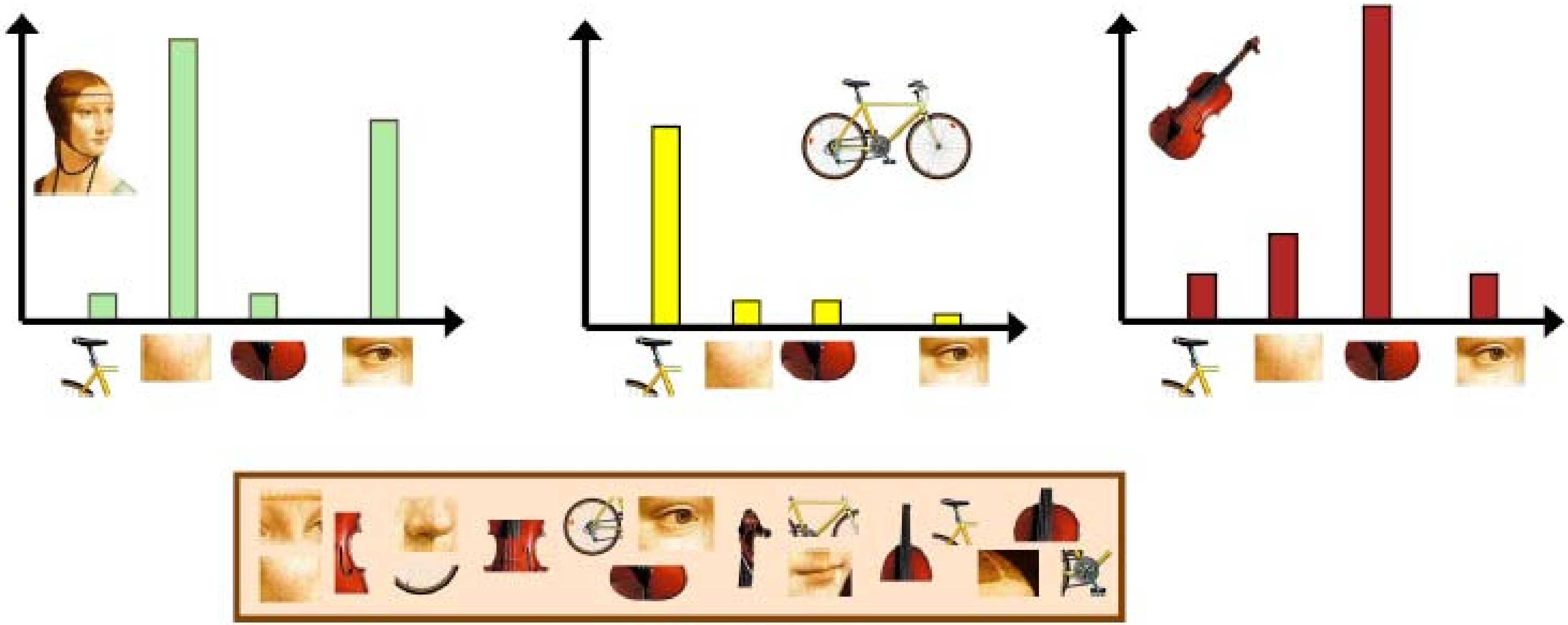
bike

violin

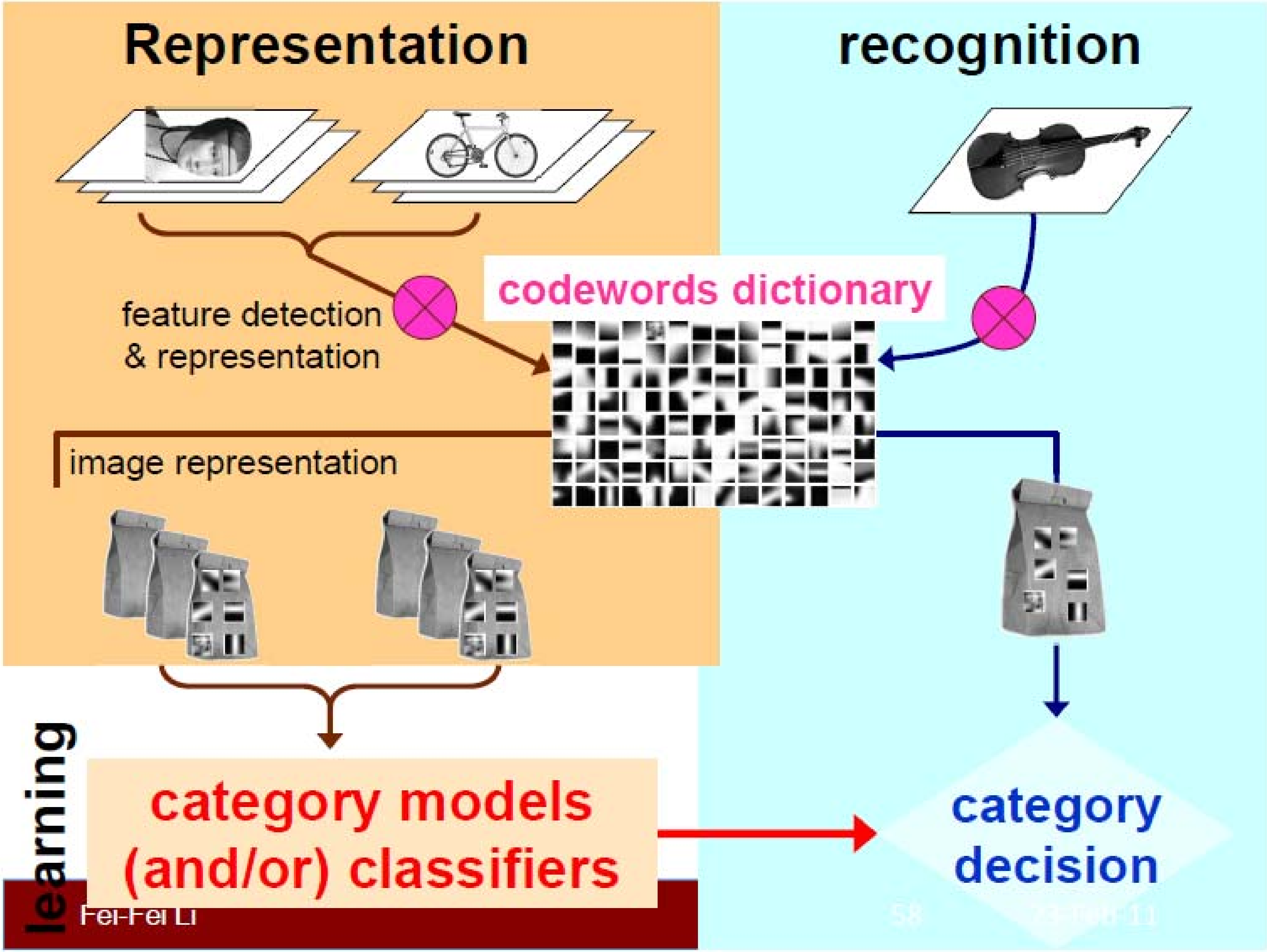
definition of "BoW"

– Independent features

– histogram representation

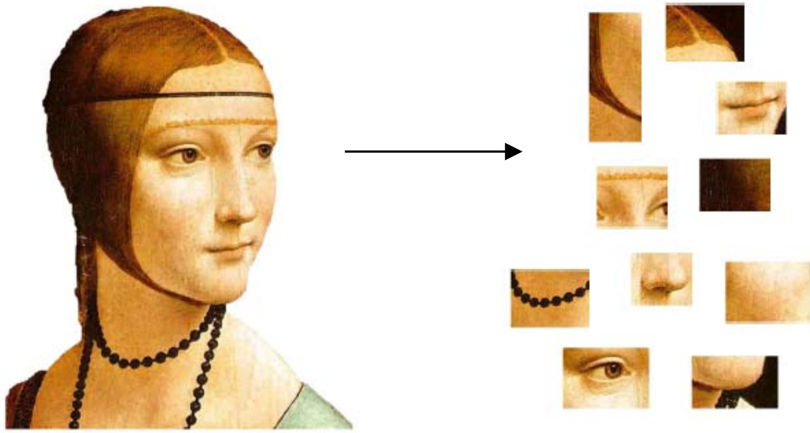


codewords dictionary

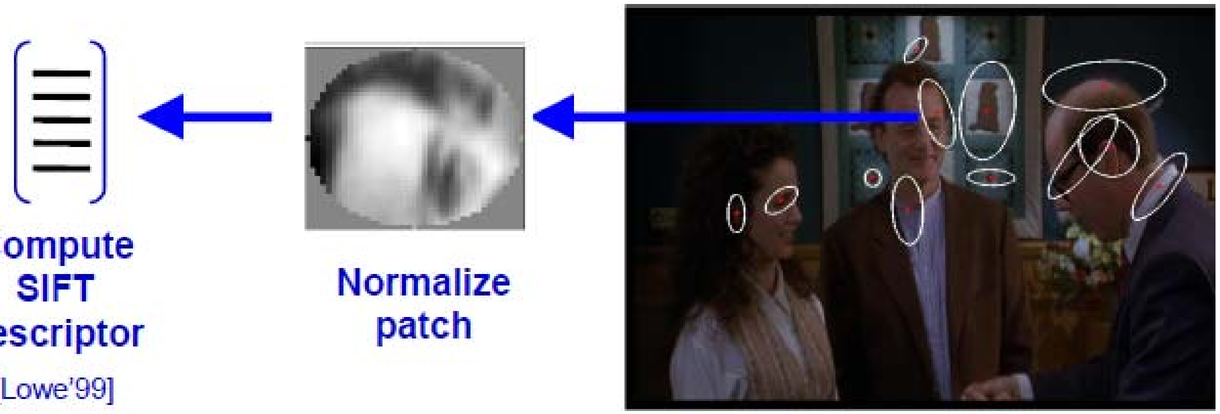

Representation
recognition
codewords dictionary
feature detection
& representation
image representation
learning
category models
(and/or) classifiers
category
decision
Fei-Fei Li
58

1. Feature Detection and Representations

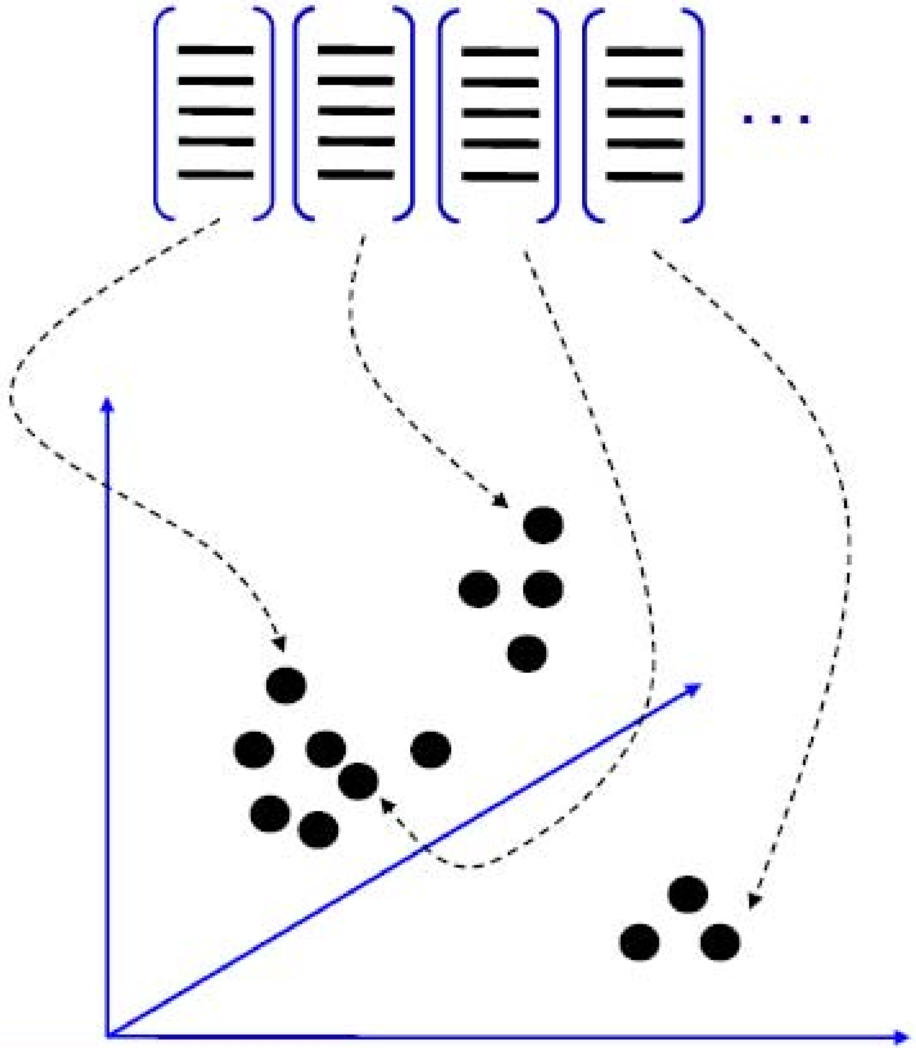
- **Assume many local features as an aggregation model**
  - **Global feature is not used in this context**
- **Densely sampled or sampled only at key points**
  - **Detect patches a extract features from them**



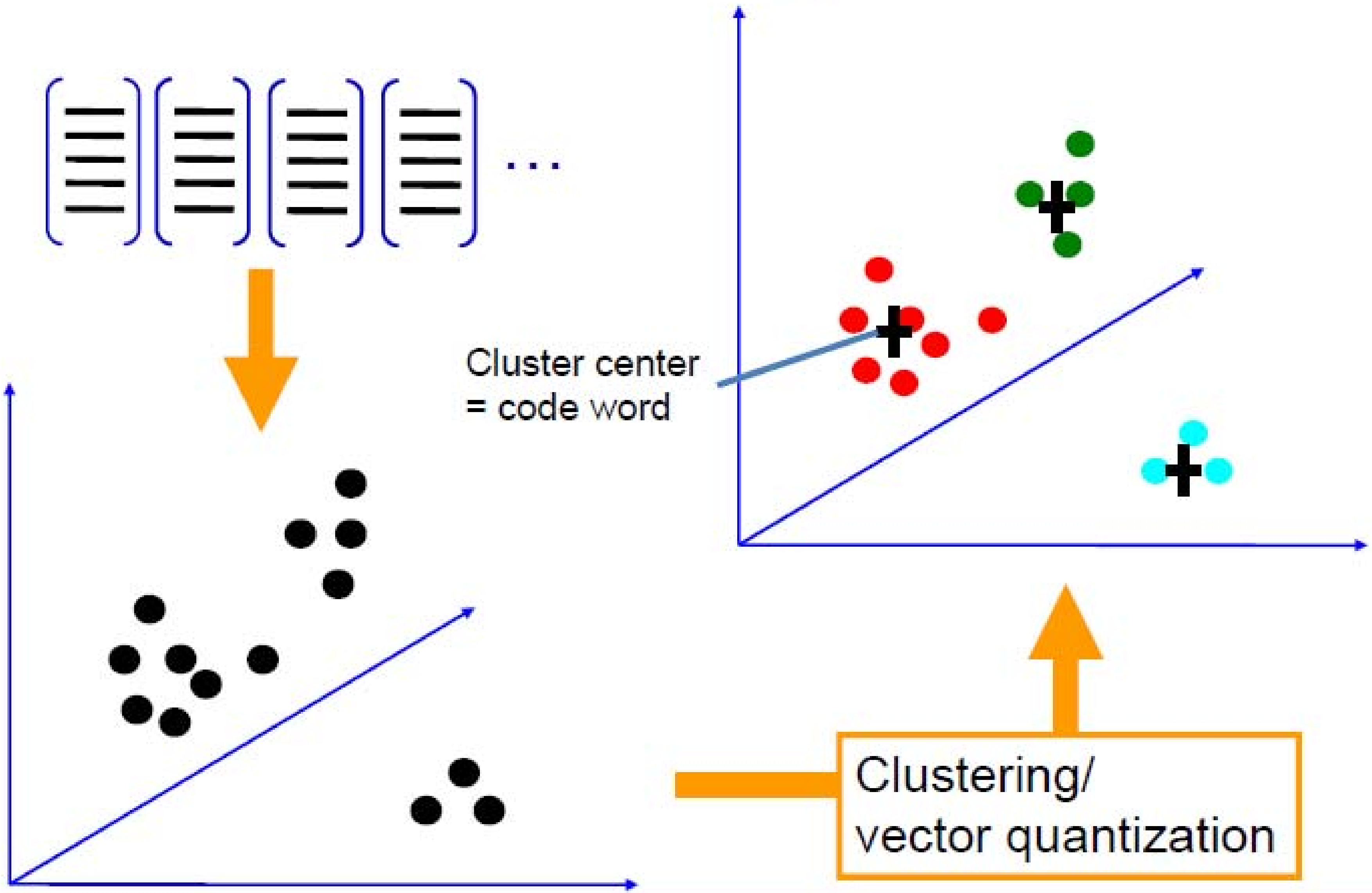
Ack.: Josef Sivic and Li Fei-Fei



2. Codewords dictionary formation



2. Codewords dictionary formation

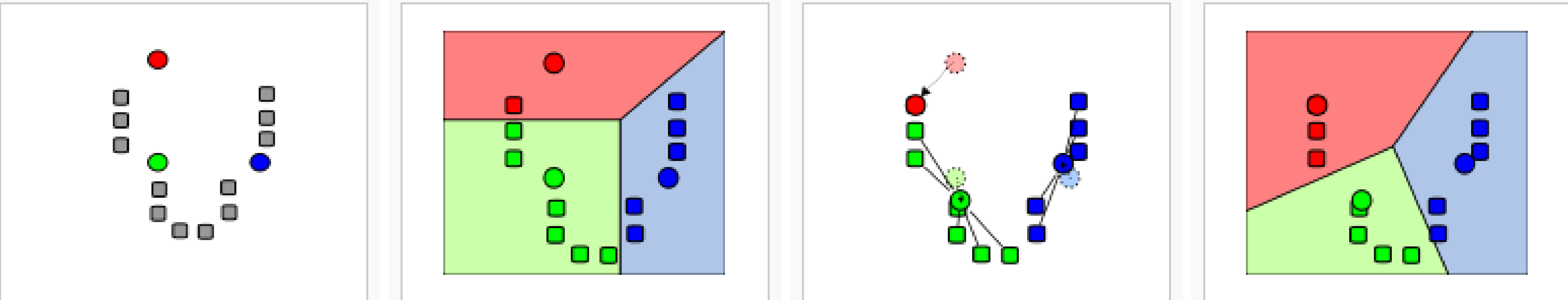


K-Means Clustering

- **An unsupervised learning**
- **Minimize the within-cluster sum of squares**

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \left\| \mathbf{x}_j - \boldsymbol{\mu}_i \right\|^2$$

Demonstration of the standard algorithm



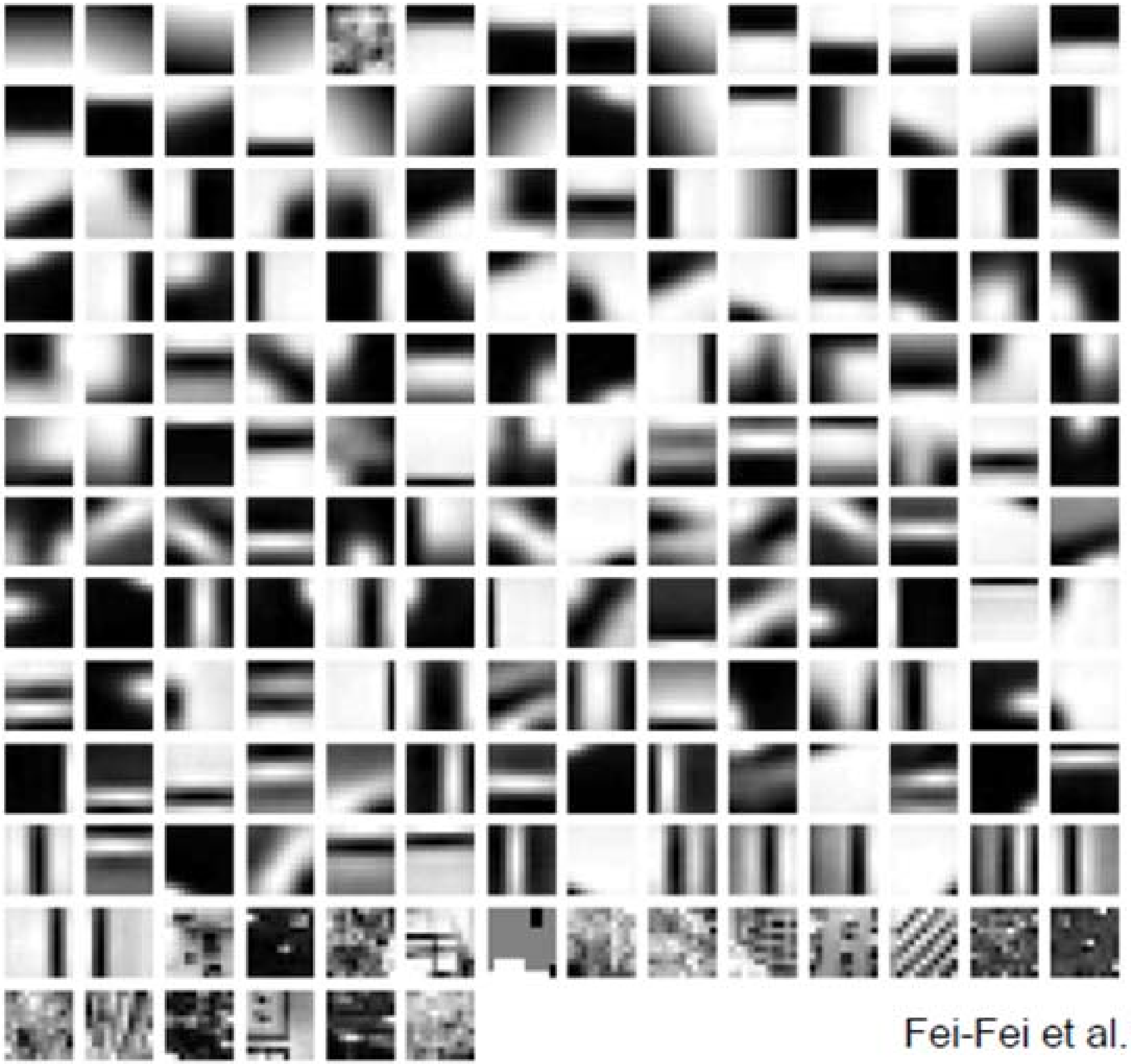
1) *k* initial "means" (in this case *k*=3) are randomly selected from the data set (shown in color).

2) *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3) The centroid of each of the *k* clusters becomes the new means.

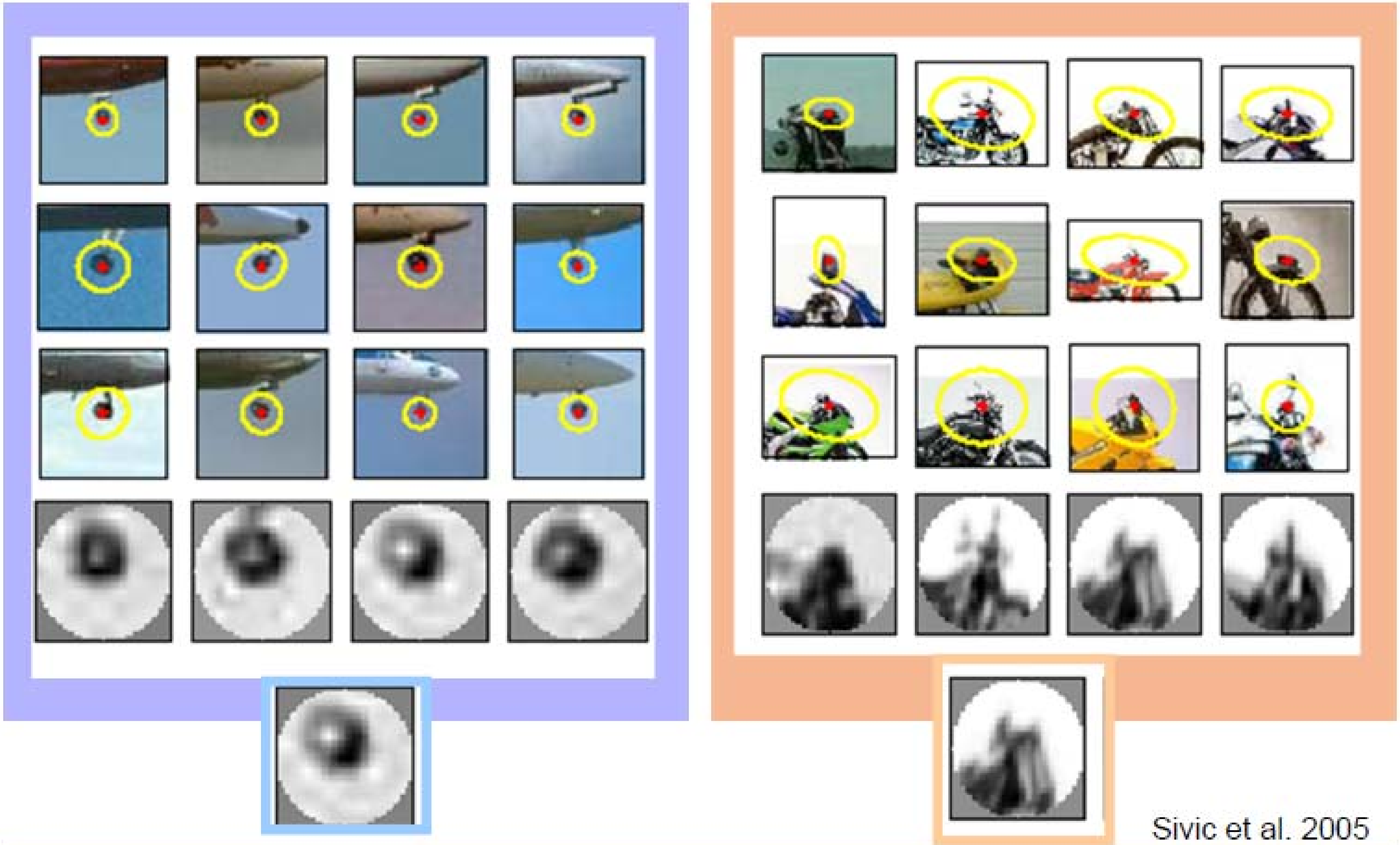
4) Steps 2 and 3 are repeated until convergence has been reached.

Codewords Dictionary Formation



Fei-Fei et al. 2005

Image patch examples of codewords

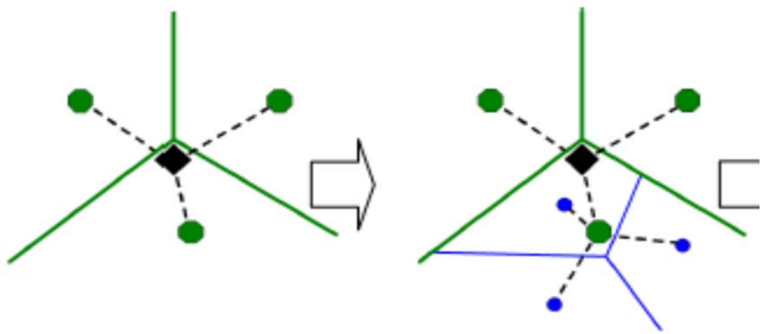


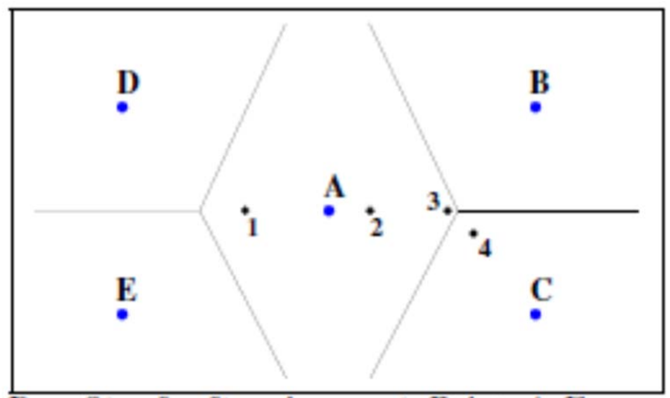
Sivic et al. 2005

Issues of Visual Vocabulary

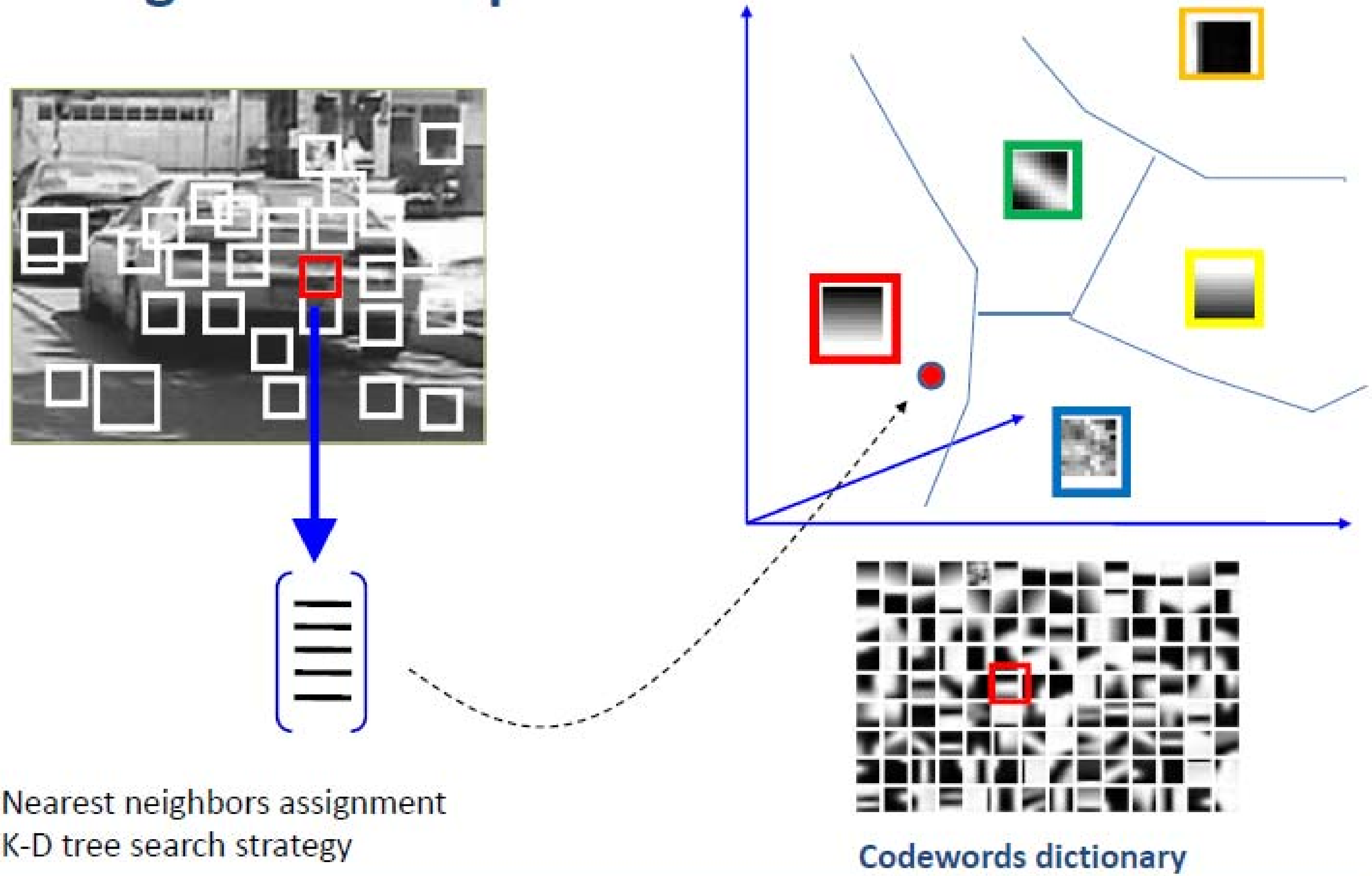
- **Related to quantization**
  - **Too many words: quantization artifacts**
  - **Too small words: not representative**
- **K-means also takes long computation times**

- **Alternatives**
  - **Faster performance: vocabulary tree, Nister et al.**
  - **Low quantization artifacts: soft quantization, Philbin et al.**



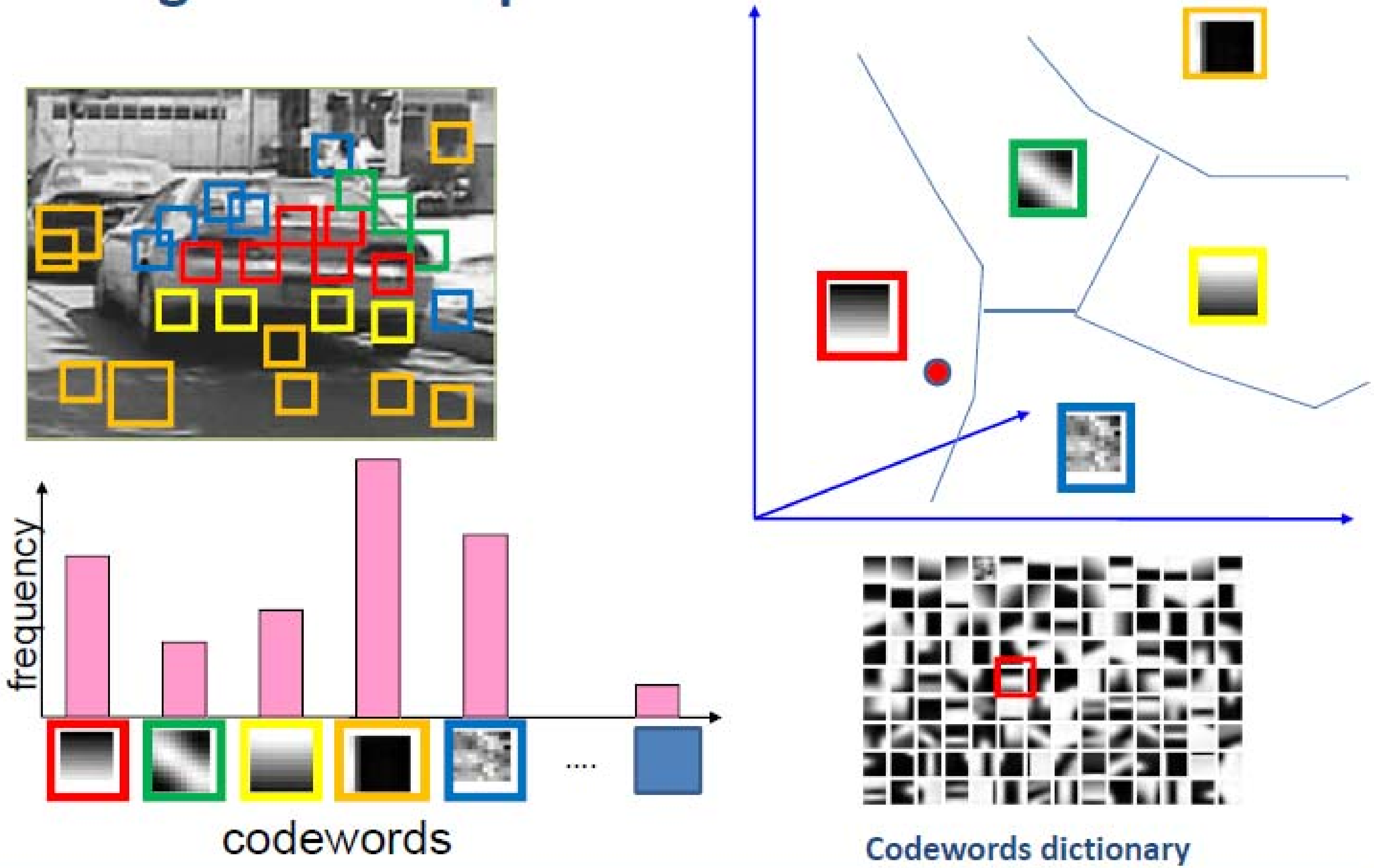


3. Bag of word representation



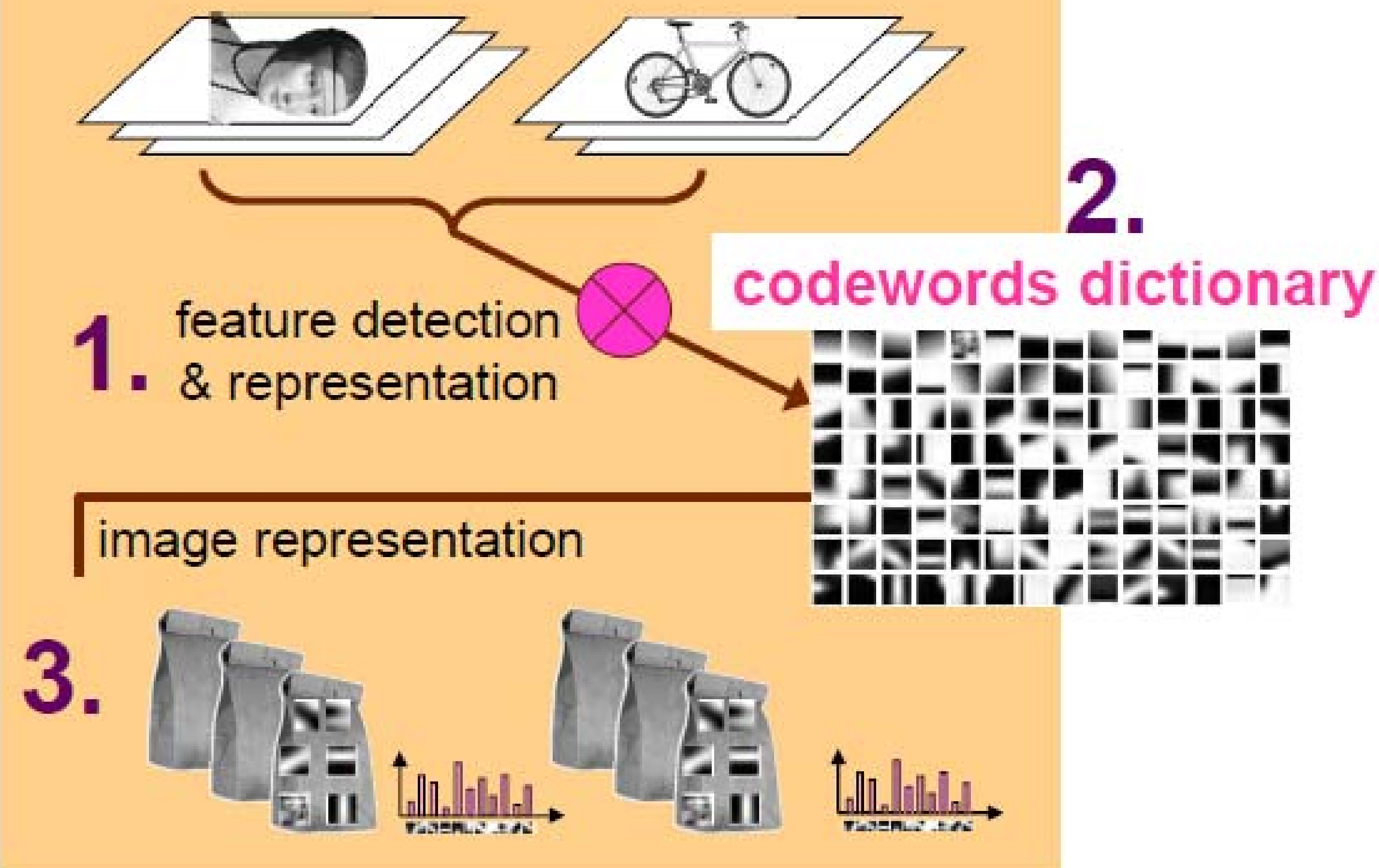
- Nearest neighbors assignment
- K-D tree search strategy

3. Bag of word representation

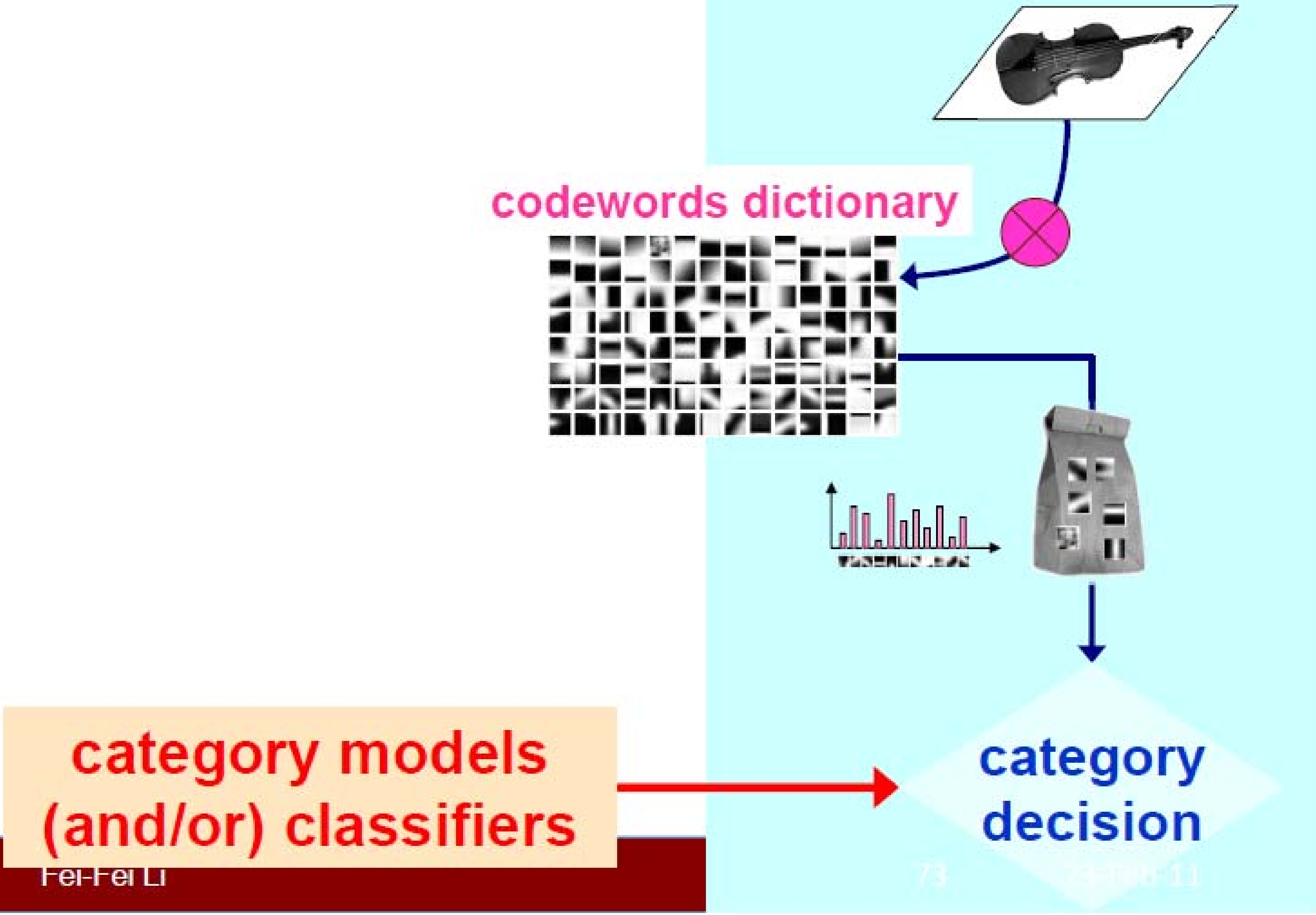


A kind of pooling operations

Representation



Learning and Recognition



TF-IDF

- **Adopted from text search**
  - **A kind of weighting and normalization process**
- **Assume a document to be represented by** $(t_1, ..., t_i, ..., t_k)^\top$
- **Weighted by TF (Term frequency) * log (IDF (Inverse Document Frequency))**

$$t_i = \frac{n_{id}}{n_d} \log \frac{N}{n_i}$$

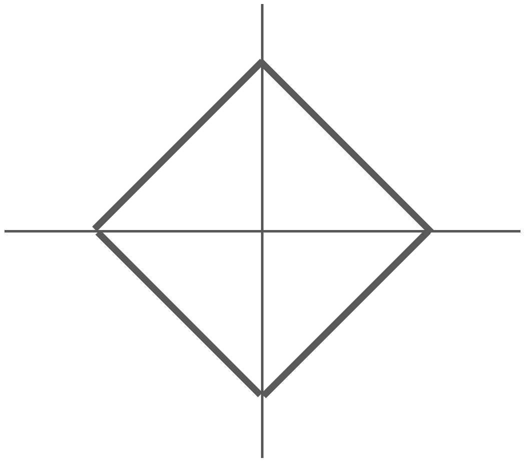
  - **n_{id} : # of occurrences of word i in document d**
  - **n_d : total # of words in the document d**
  - **n_i : # of occurrences of term i in the whole database**
  - **N: # of documents in the whole database**

Similarity and Distance Functions

- **Dot product measuring the angle between two vectors**

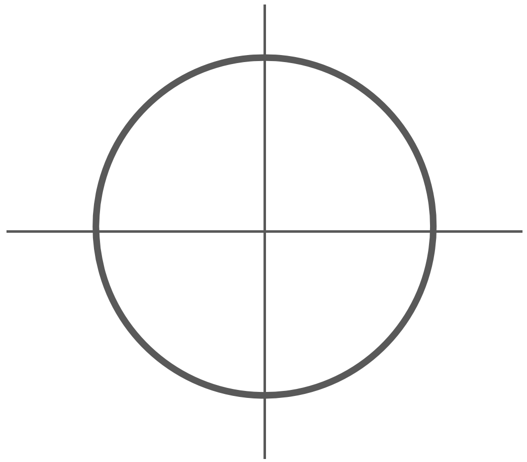
L1 (Manhattan) distance

$$d_1(I_1, I_2) = \sum_p \left| I_1^p - I_2^p \right|$$



L2 (Euclidean) distance

$$d_2(I_1, I_2) = \sqrt{\sum_p \left(I_1^p - I_2^p \right)^2}$$



Mahalanobis Distance

- **Mahalanobis weighs L2 distance between two points, by the standard deviation of the data**

  $$f(x, y) = (x - y)^T \Sigma^{-1} (x - y),$$

  where Σ is the mean-subtracted covariance matrix of all data points.

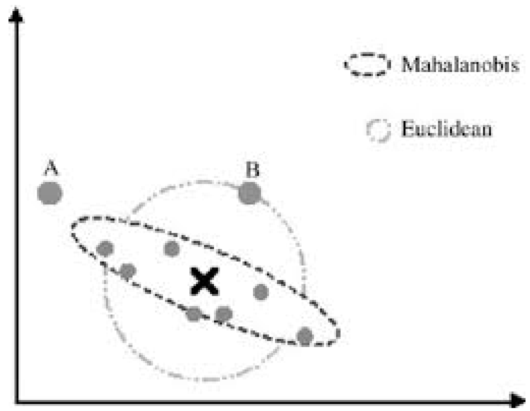
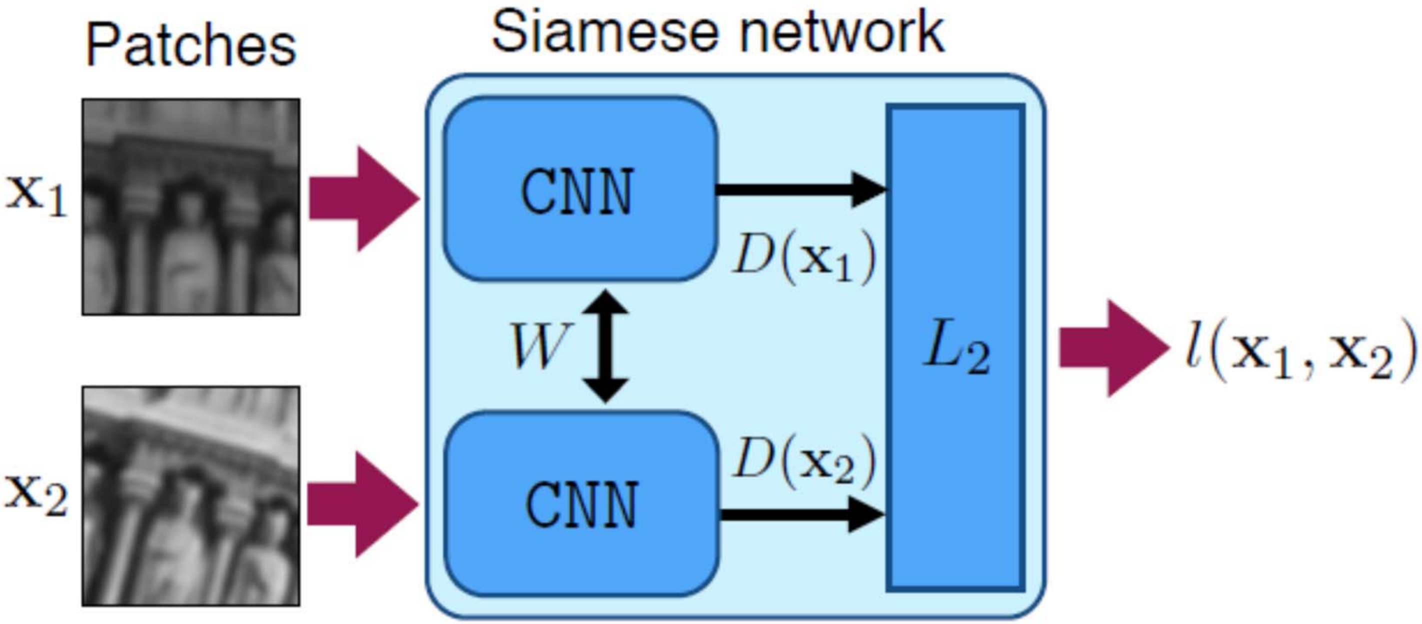


Image Source: Google

Chandra, M.P., 1936. On the generalised distance in statistics. In *Proceedings of the National Institute of Sciences of India* (Vol. 2, No. 1, pp. 49-55).

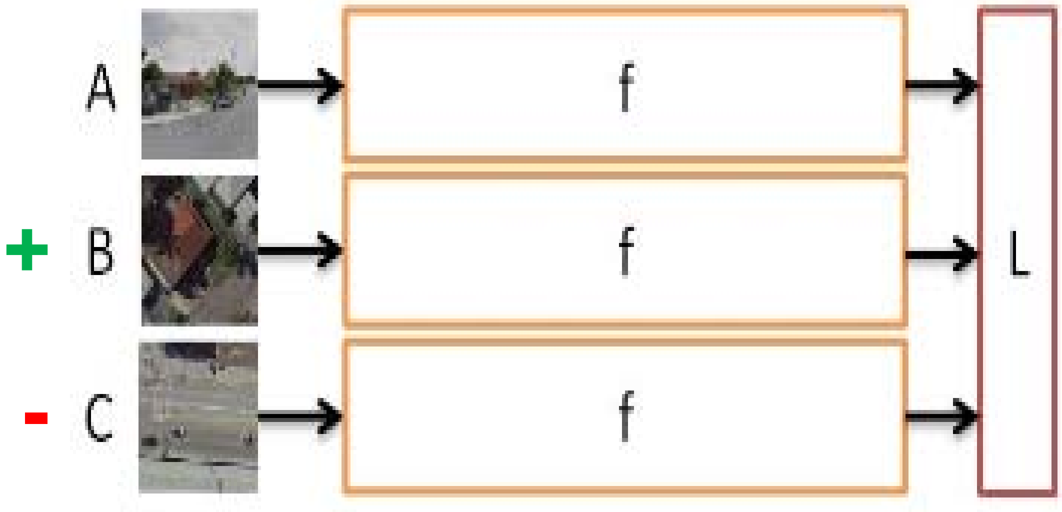
Similarity Learning: Siamese CNN

- **Learn a feature representation mapping the sample patches with the L2 distance**



Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P. and Moreno-Noguer, F., 2015. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 118-126).

Siamese CNN Variants: Triplet Network or Loss



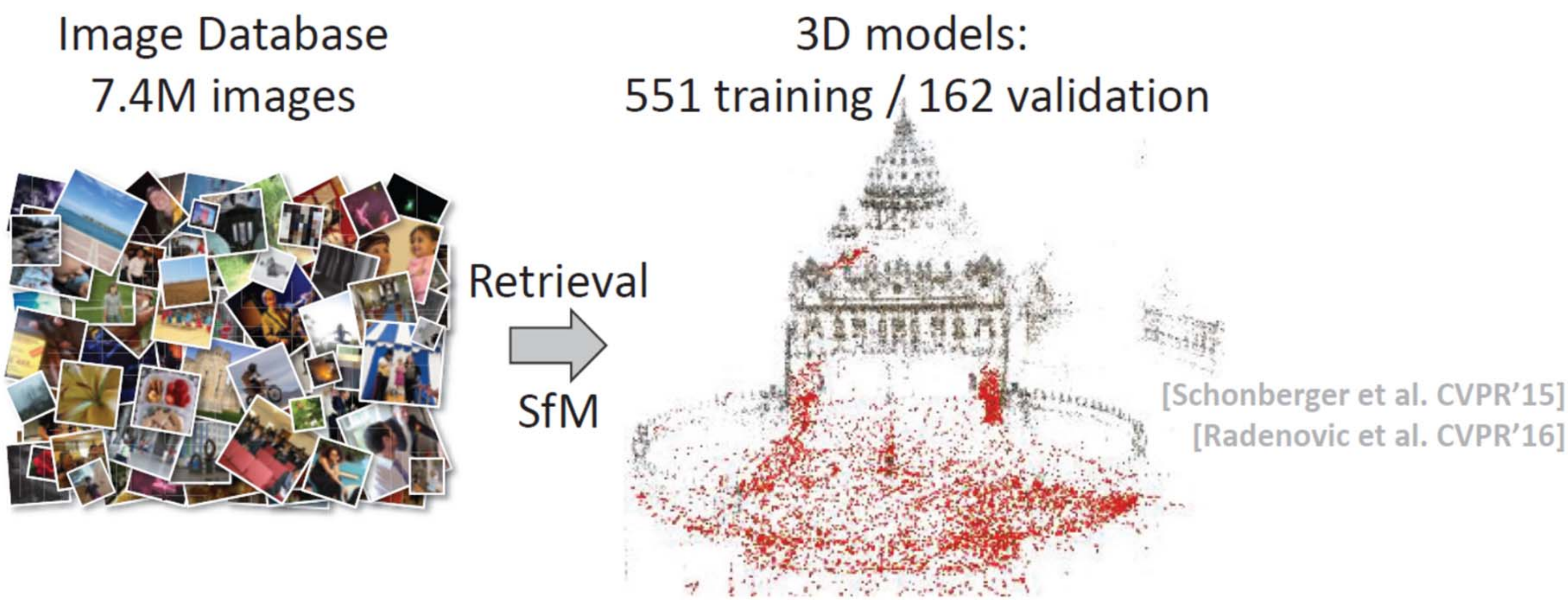
D(f(A), f(B)) < D(f(A), f(C))

- Allows us to learn ranking between samples
  - Known as a ranking loss

Vo, N.N. and Hays, J., 2016, October. Localizing and orienting street views using overhead imagery. In European Conference on Computer Vision (pp. 494-509).

Utilize BoW for CNN Image Retrieval

- **Construct 3D models from BoW based image retrieval**
  - **Refine CNN features by mimicking BoW-based retrieval**
  - **Unsupervised groups of photos with different landmarks**



Given a query, identify its positive (same cluster or city) and its negative image

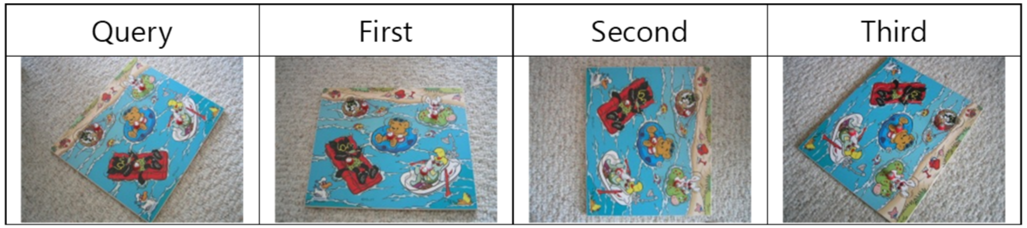
Negative images

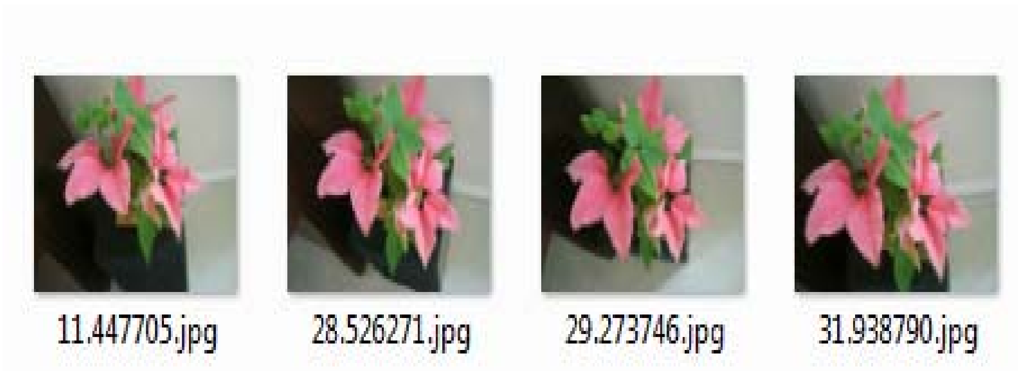
Positive images

CNN Image Retrieval Learns from BoW:
Unsupervised Fine-Tuning with Hard Examples, ECCV

PA1

- **Understand and implement a basic image retrieval system**
- **Use the original UKBenchmark**
- **Measure its accuracy**

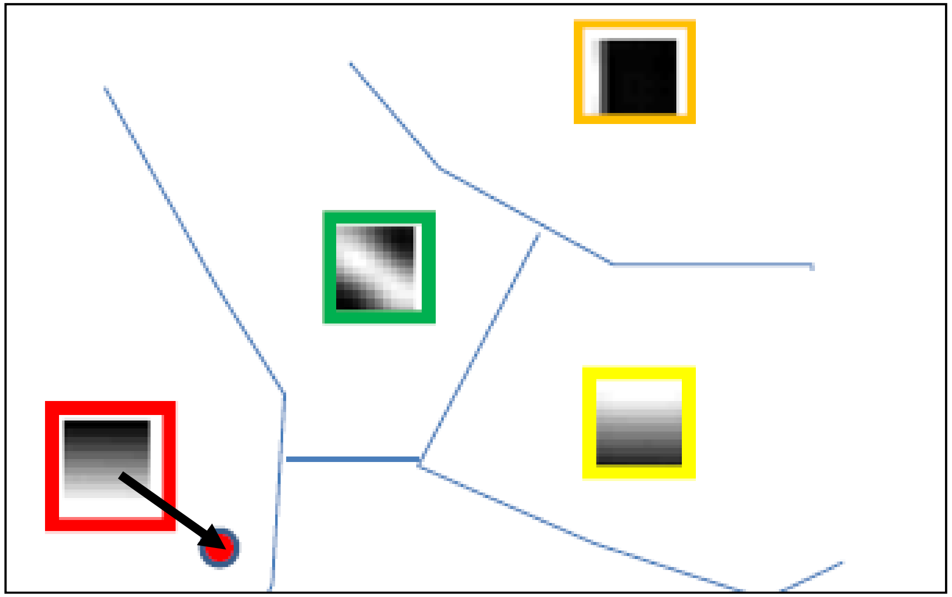




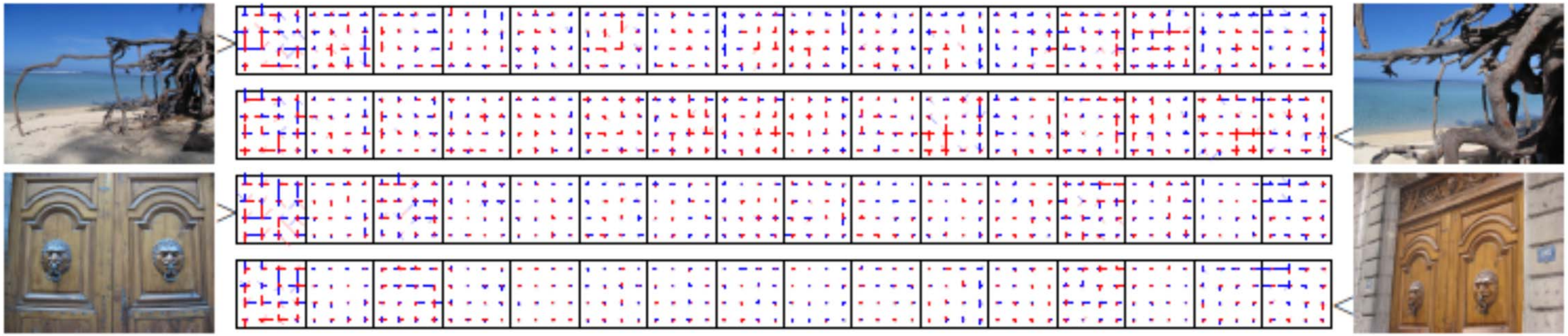
VLAD (Vector of Locally Aggregated Descriptors)

- **BoW**
  - **Count the number of SIFTs assigned to each cluster**
- **VLAD**
  - **Compute the difference between a feature and its cluster center**

$$v_{i,j} = \sum_{x \text{ such that } \mathrm{NN}(x)=c_i} x_j - c_{i,j}$$



VLAD



- **VLAD descriptors w/ 16 clusters**
- **Show better accuracy than BoW**

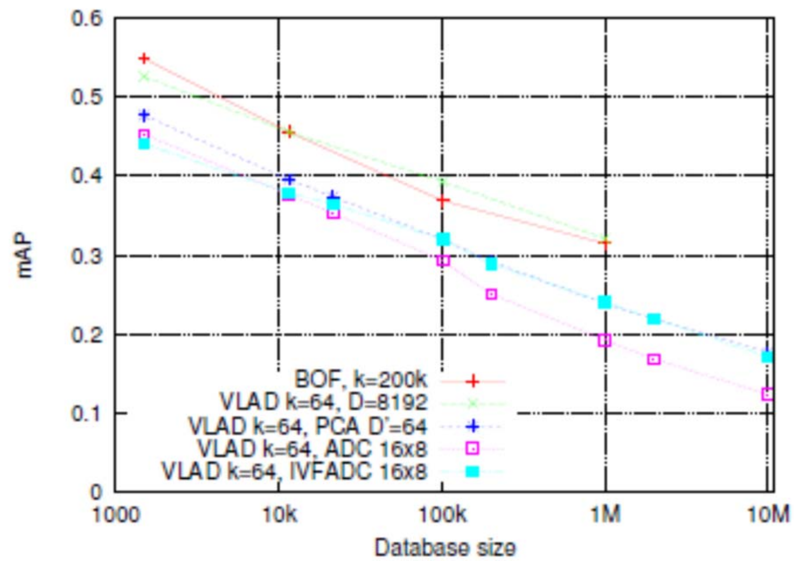


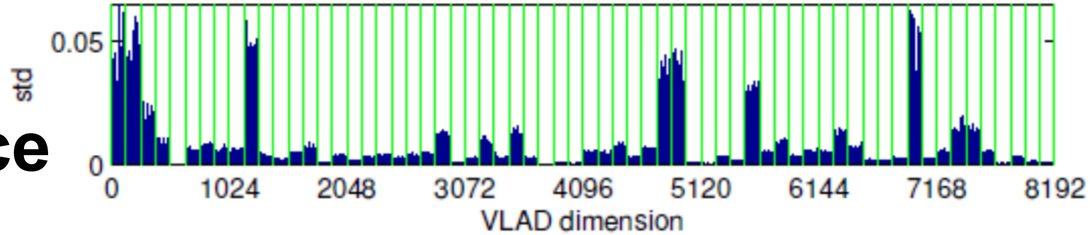
Figure 5. Search accuracy as a function of the database size.

Ack. Aggregating local descriptors ..

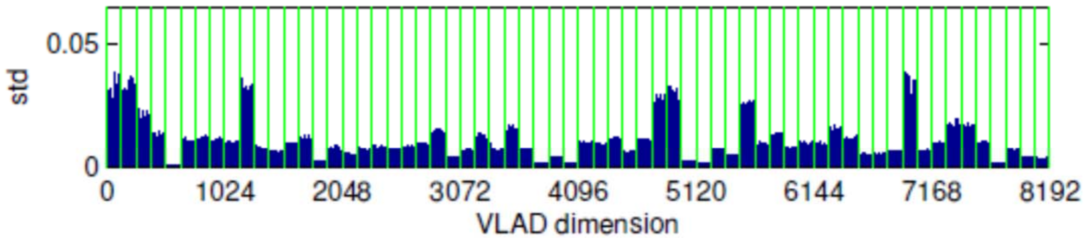
Normalization for VLAD

- **Results in better accuracy**

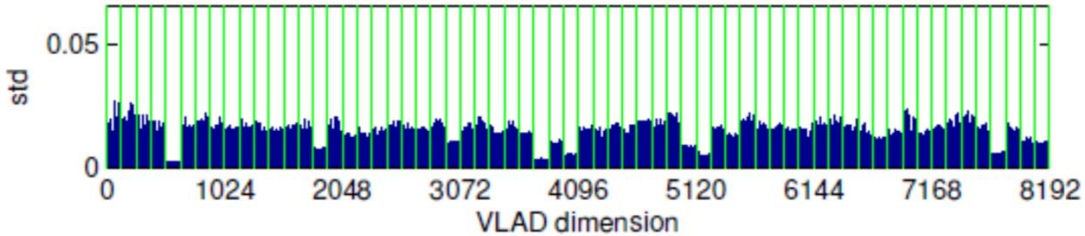
Variance



(a) Original VLAD normalization (L2)



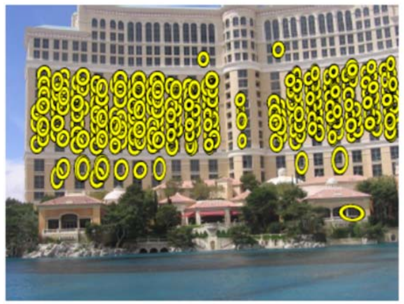
(b) Signed square rooting (SSR) followed by L2



(c) Intra-normalization (innorm) followed by L2

L2 normalization,
i.e., $\frac{v}{|v|^2}$

Square rooting
for burstiness



L2 normalization
within each VLAD
block

Ack. All about VLAD

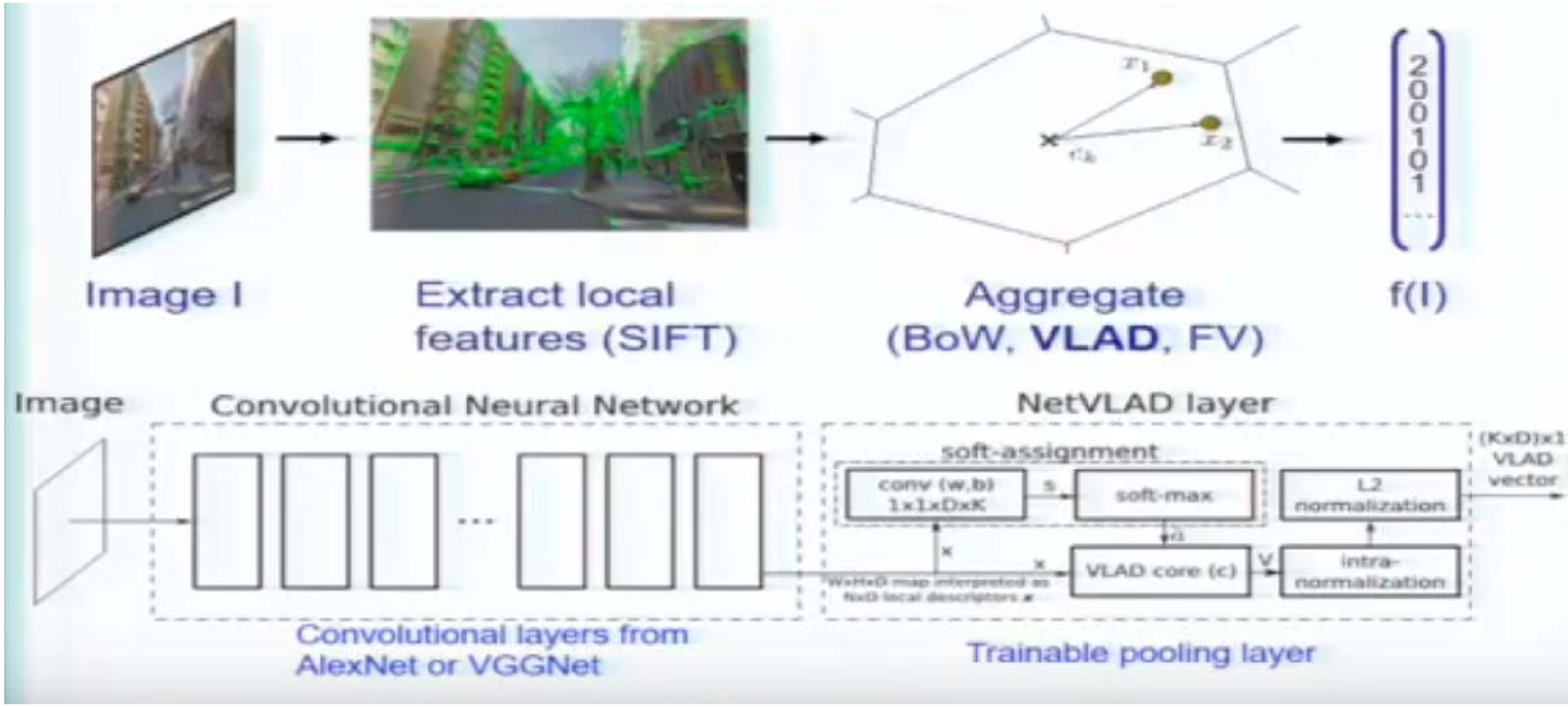
NetVLAD: CNN architecture for weakly supervised place recognition

- **Identify its location given an query image**
  - **Application of place recognition**

From the author talk

Mimic the classical approach

- **Make it end-to-end trainable for achieving better accuracy**

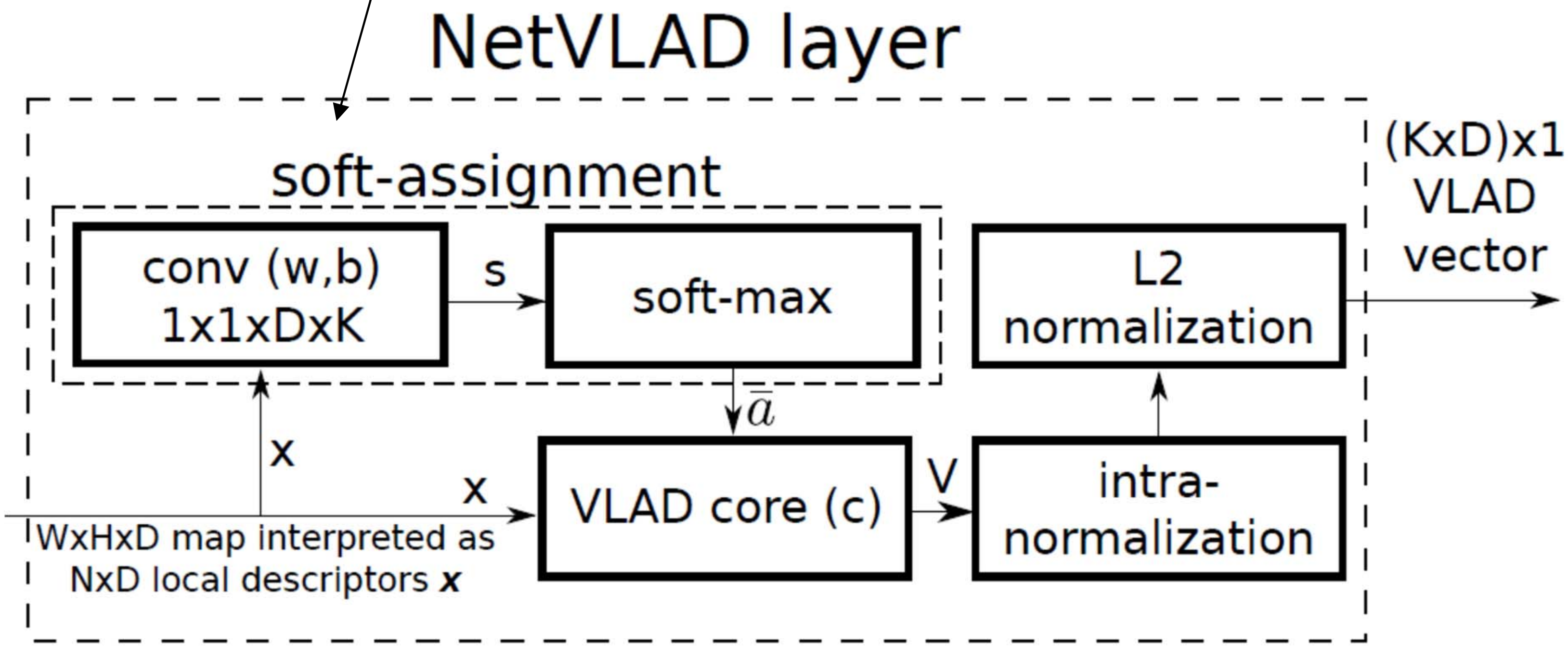


Trainable VLAD

- **Hard assignment to soft assignment using the soft-max, to make it differentiable**

$$V(j,k) = \sum_{i=1}^{N} a_k(\mathbf{x}_i)\,(x_i(j) - c_k(j)),$$

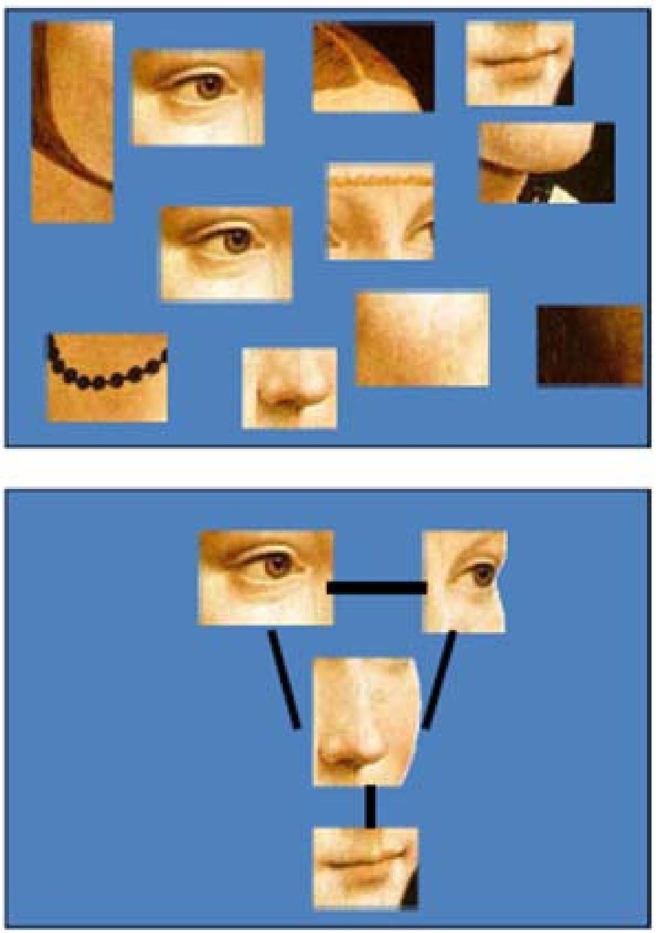
$$\bar{a}_k(\mathbf{x}_i) = \frac{e^{-\alpha\|\mathbf{x}_i - \mathbf{c}_k\|^2}}{\sum_{k'} e^{-\alpha\|\mathbf{x}_i - \mathbf{c}_{k'}\|^2}},$$



Problems of BoW Model

- No spatial relationship between words

- How can we perform segmentation and localization?



Ack.: Fei-Fei Li

Class Objectives were:

- **Bag-of-visual-Word (BoW) model**
  - **Pooling operation**
- **Ranking loss for CNN features**

Next Time…

- **Inverted index**

Homework for Every Class

- Go over the next lecture slides
- Come up with one question on what we have discussed today
  - Write questions three times
- Go over recent papers on image search, and submit their summary before Tue. class

Figs