
CS688: Web-Scale Image Retrieval

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

Sung-Eui Yoon
(윤성익)

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

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 visual image was considered as a movie screen. The image is discovered by the eye, and we know that perception is a more complex process following the path to the various centers of the brain. Hubel and Wiesel have demonstrated that the message about an image falling on the retina undergoes a fine-grained 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

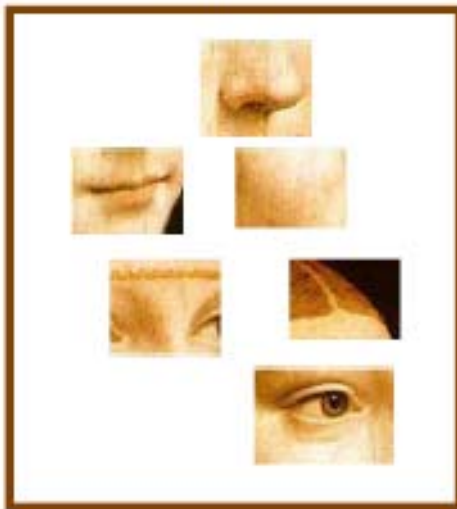
China is forecasting a trade surplus of \$90bn (£51bn) to \$100bn this year, a threefold increase on 2004's \$32bn. The Commerce Ministry said the surplus would be created by a predicted 30% increase in exports to \$750bn, compared with \$570bn in 2004. The \$660bn trade deficit in 2004 annoyed the Chinese government. China's deliberate policy of keeping the yuan undervalued against the dollar has also needed to be addressed. The demand for yuan is growing in the country. China has been permitted 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

definition of “BoW”

– Independent features

face



bike

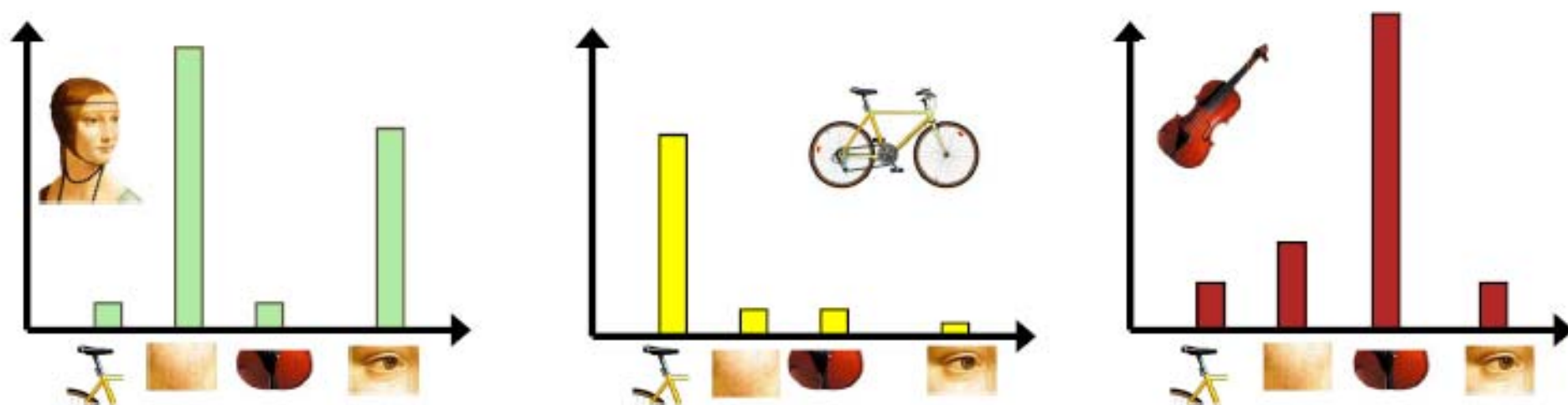


violin



definition of “BoW”

- Independent features
- histogram representation



codewords dictionary

Representation



feature detection & representation



codewords dictionary



image representation



recognition



category decision

learning

category models (and/or) classifiers

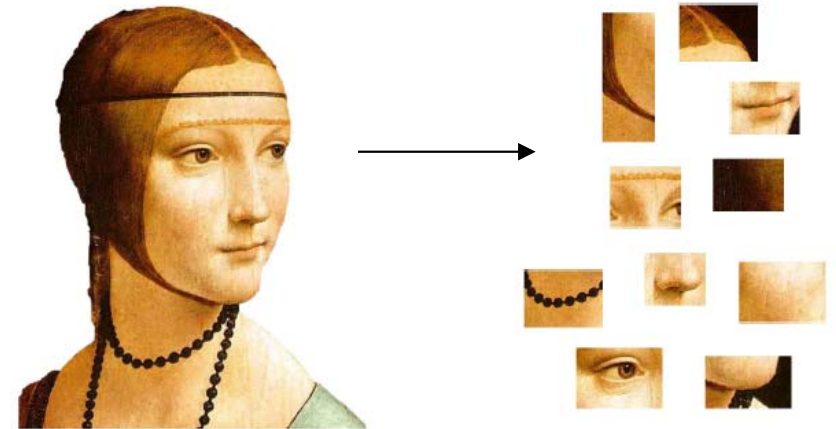
Fer-Fer Li

58

23-108-11

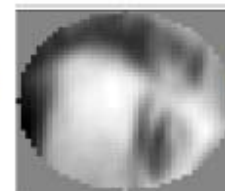
1. Feature Detection and Representations

- 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 and extract features from them

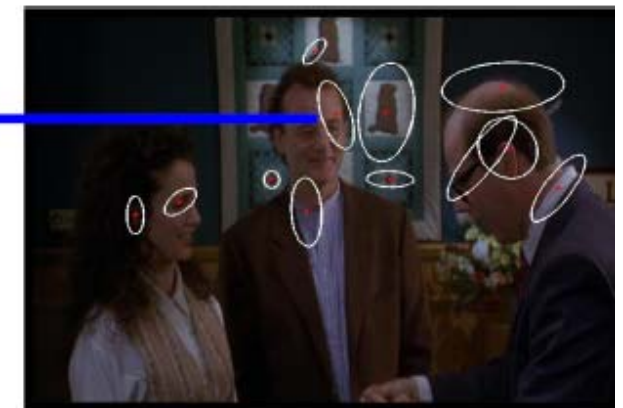


Ack.: Josef Sivic and Li Fei-Fei

Compute
SIFT
descriptor
[Lowe'99]



Normalize
patch

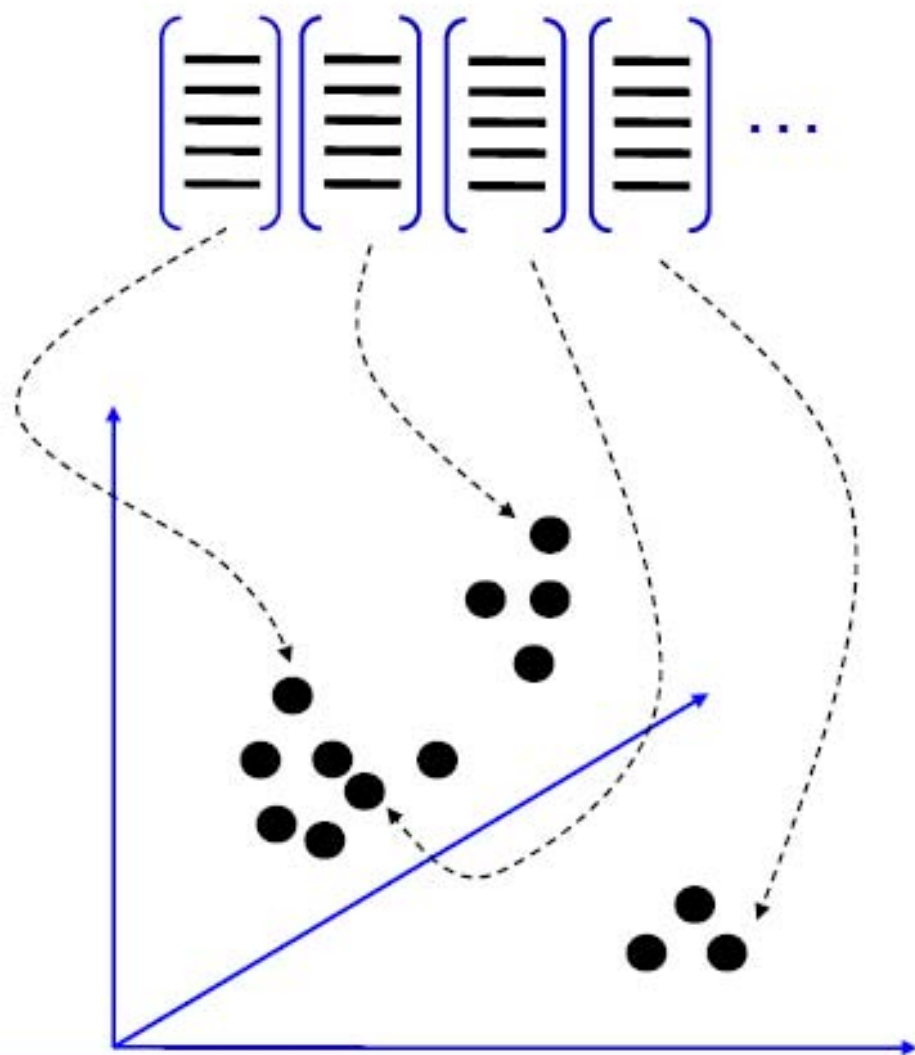


Detect patches

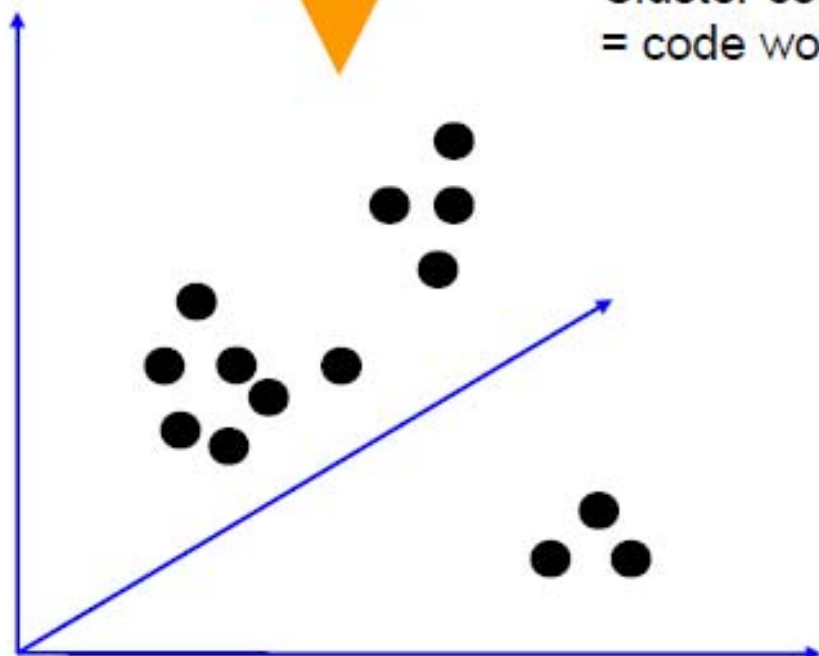
[Mikojczyk and Schmid '02]

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

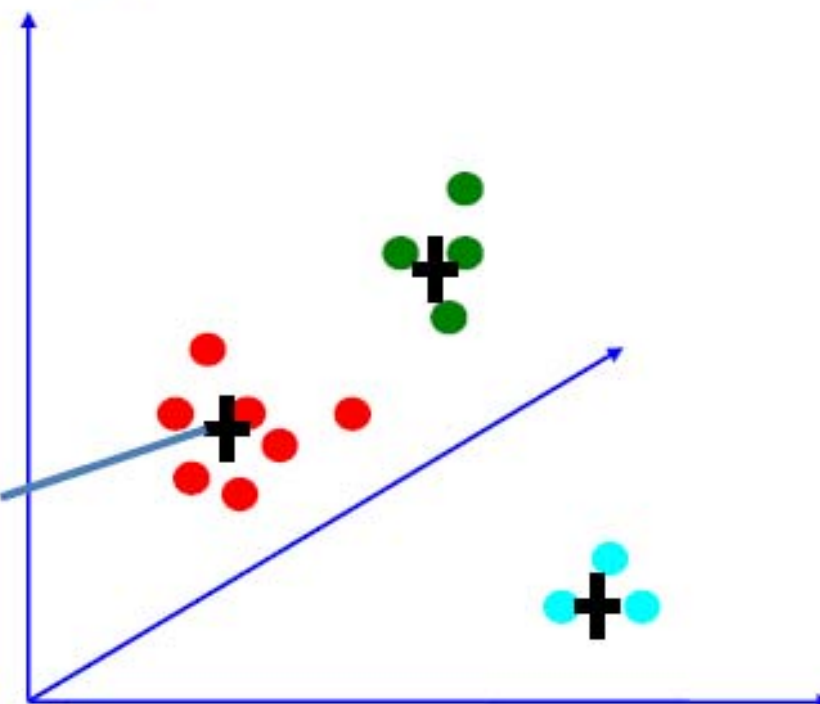
2. Codewords dictionary formation



2. Codewords dictionary formation



Cluster center
= code word



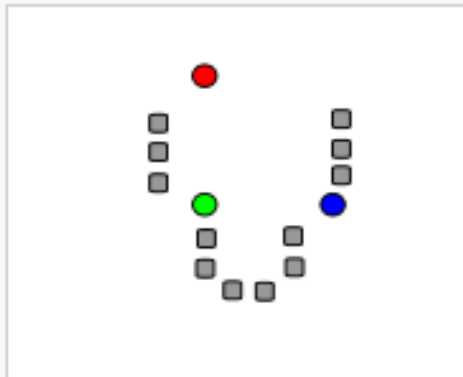
Clustering/
vector quantization

K-Means Clustering

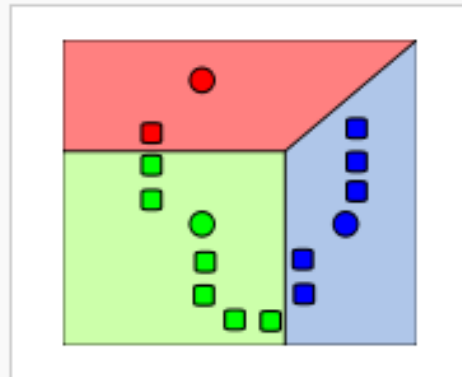
- An unsupervised learning
- Minimize the within-cluster sum of squares

$$\operatorname{argmin}_{\mathcal{S}} \sum_{i=1}^k \sum_{\mathbf{x}_j \in \mathcal{S}_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

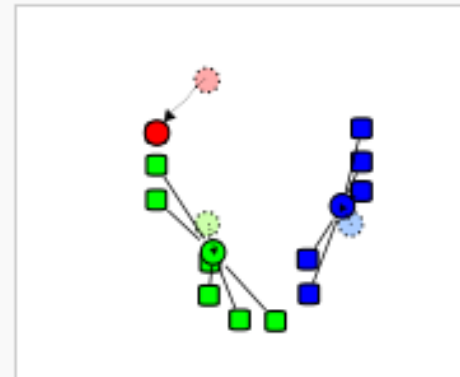
Demonstration of the standard algorithm



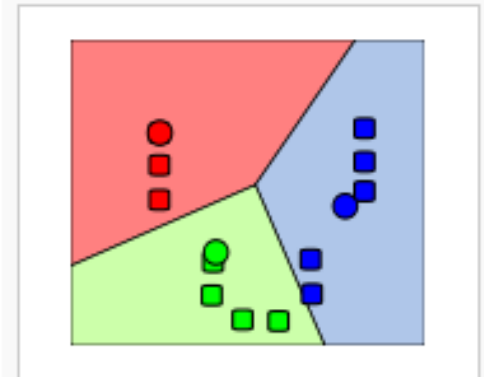
1) k initial "means" (in this case $k=3$) are randomly selected from the data set (shown in color).



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

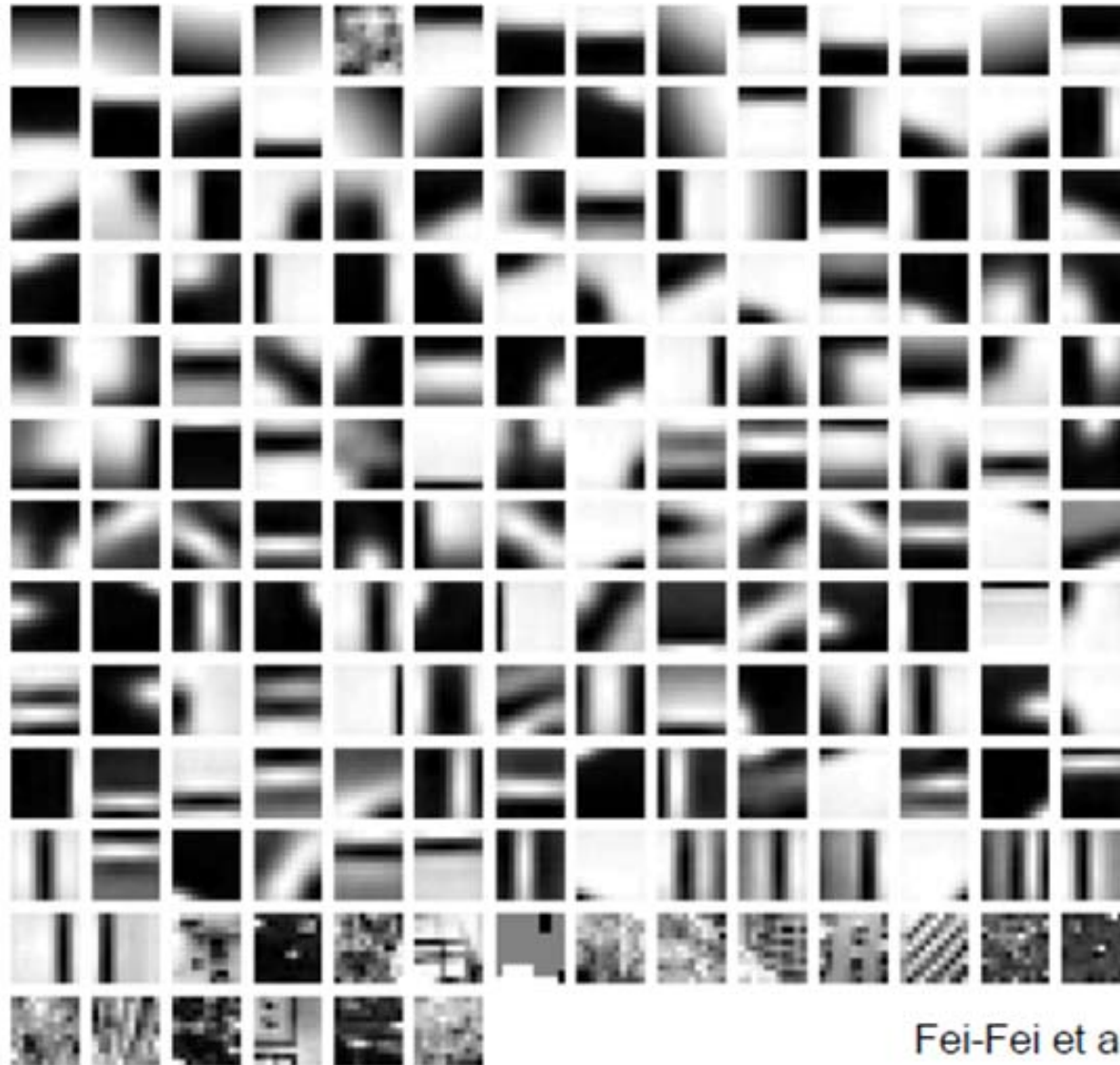


3) The [centroid](#) of each of the k clusters becomes the new means.



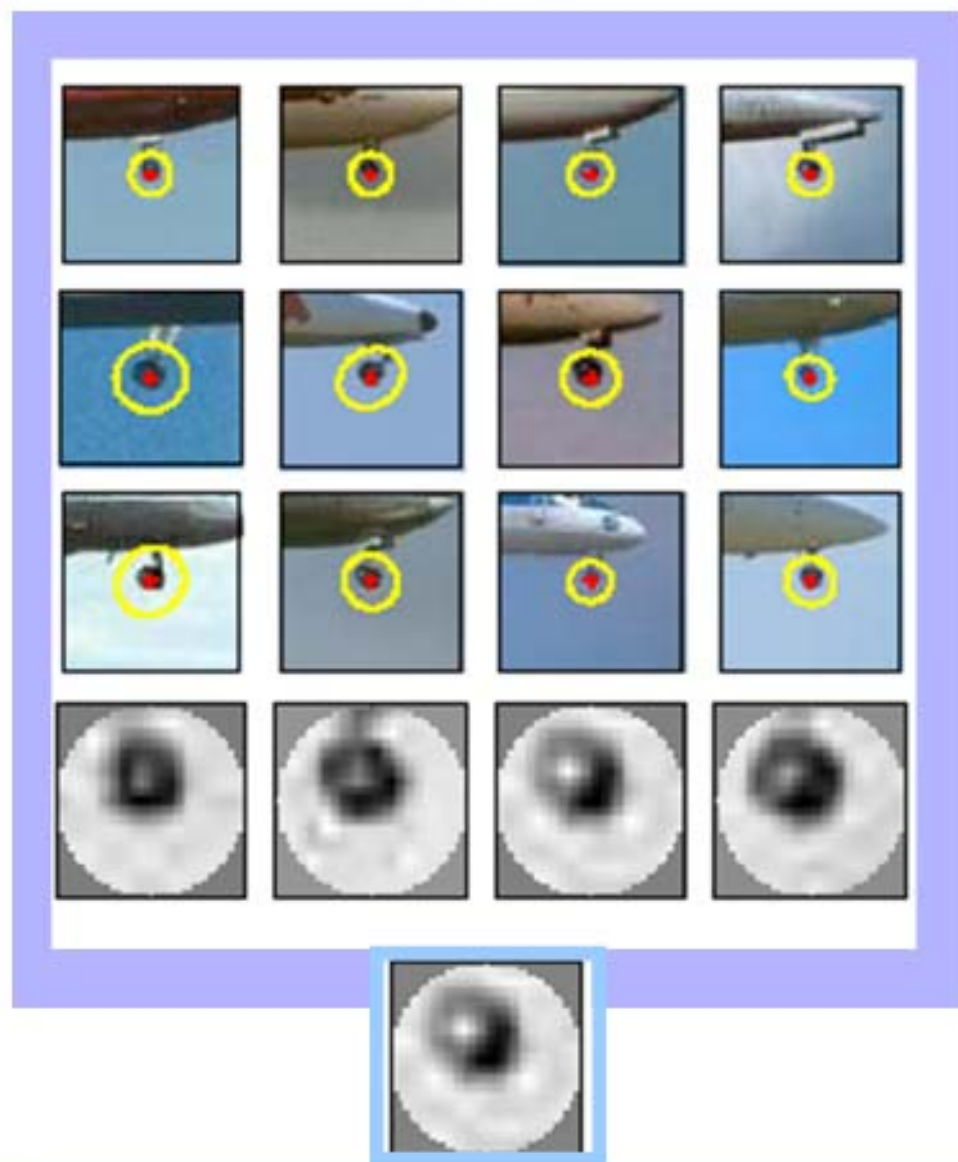
4) Steps 2 and 3 are repeated until convergence has been reached.

Codewords Dictionary Formation



Fei-Fei et al. 2005

Image patch examples of codewords

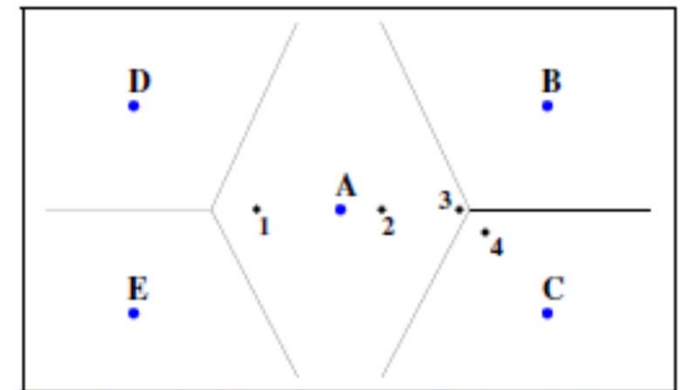
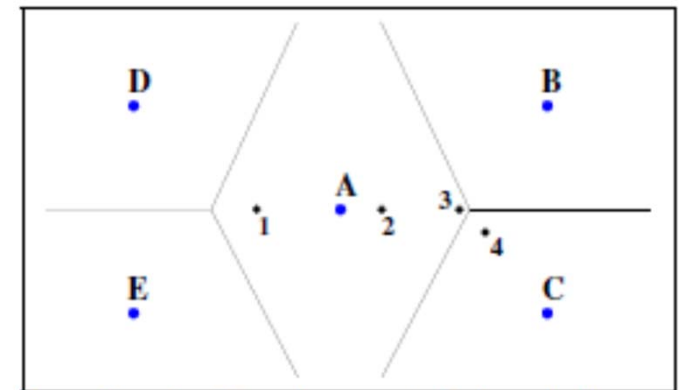
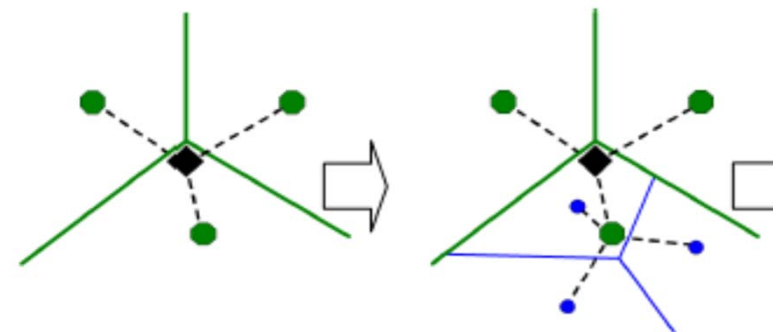


Issues of Visual Vocabulary

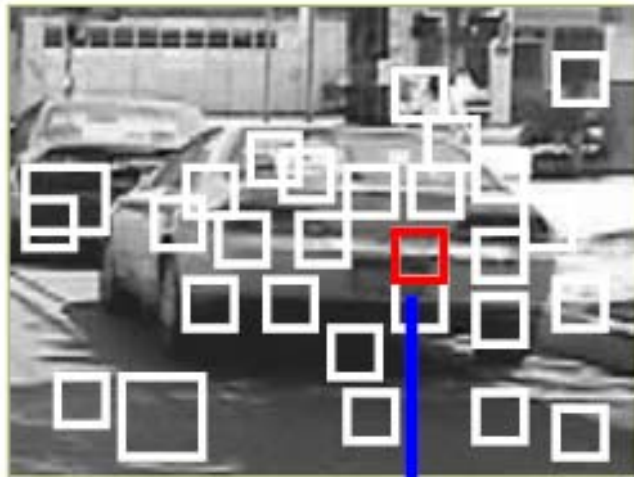
- **Related to quantization**
 - **Too many words: quantization artifacts**
 - **Too small words: not representative**
- **K-means also takes long computation times**

- **Alternatives**

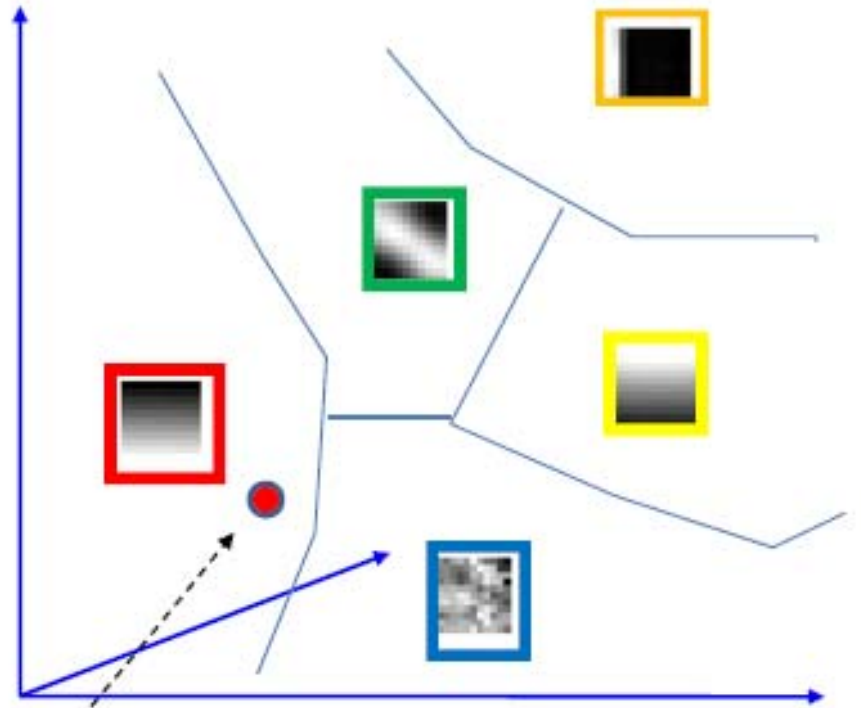
- **Faster performance: vocabulary tree, Nister et al.**
- **Low quantization artifacts: soft quantization, Philbin et al.**



3. Bag of word representation

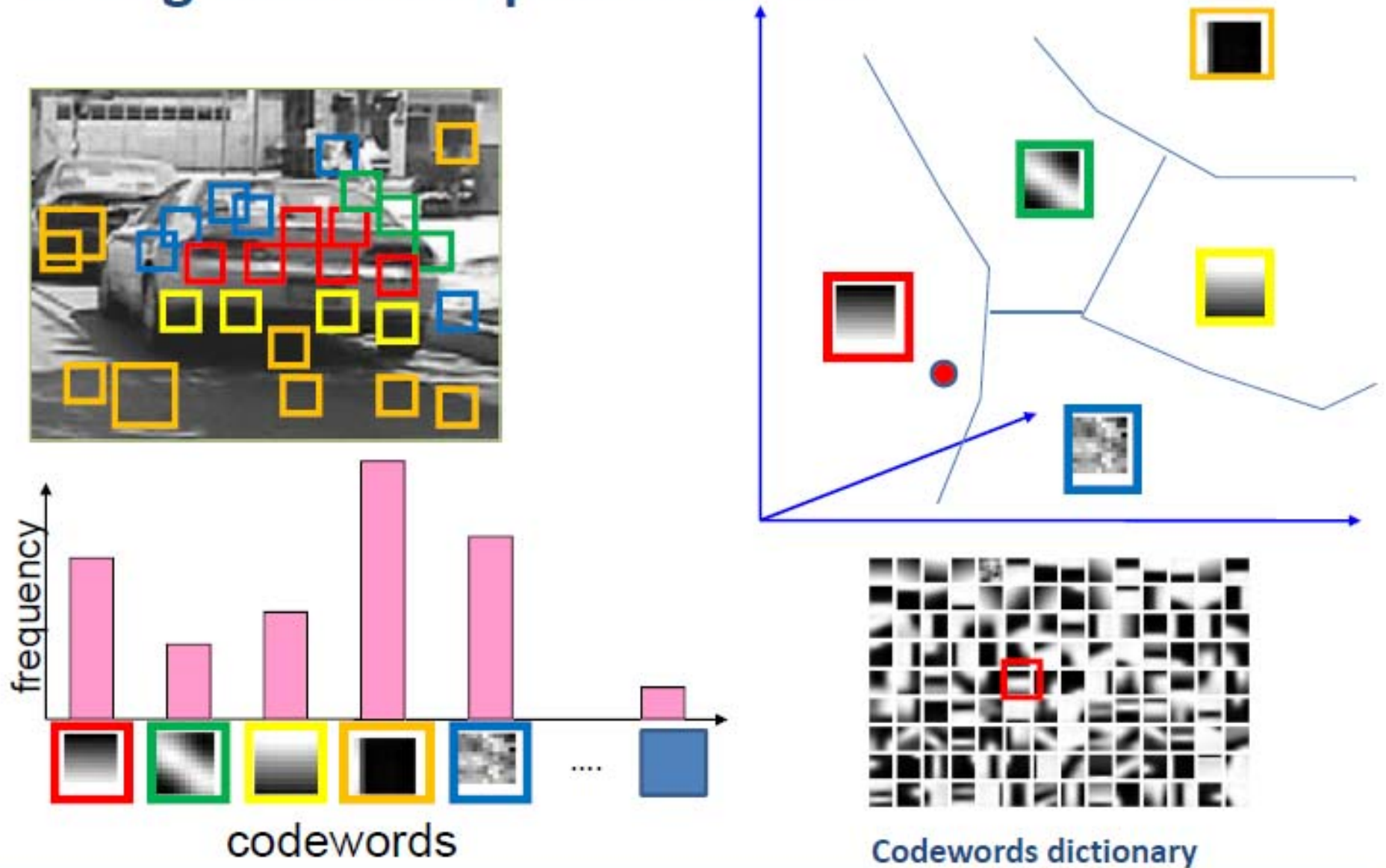


- Nearest neighbors assignment
- K-D tree search strategy



Codewords dictionary

3. Bag of word representation



A kind of pooling operations

Representation



1. feature detection & representation



2. codewords dictionary



image representation

3.



Learning and Recognition



codewords dictionary



category models
(and/or) classifiers

category
decision

TF-IDF

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

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

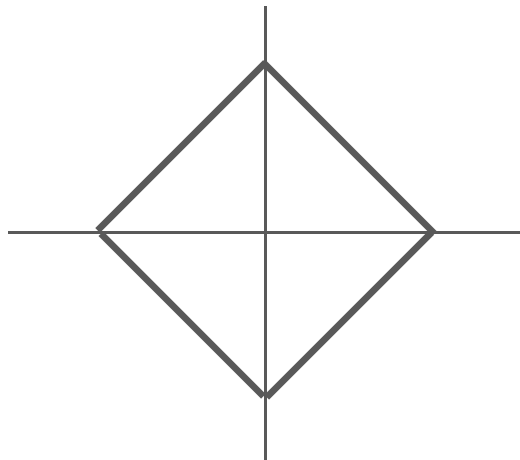
- n_{id} : # of occurrences of word i in document d
- n_d : total # of words in the document d
- n_i : # of occurrences of term i in the whole database
- N : # of documents in the whole database

Similarity and Distance Functions

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

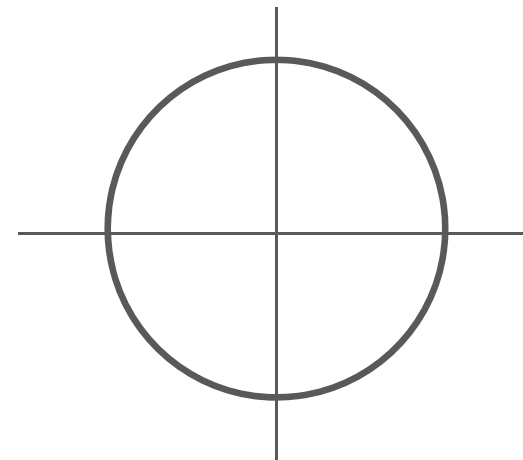
L1 (Manhattan) distance

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



L2 (Euclidean) distance

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



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.

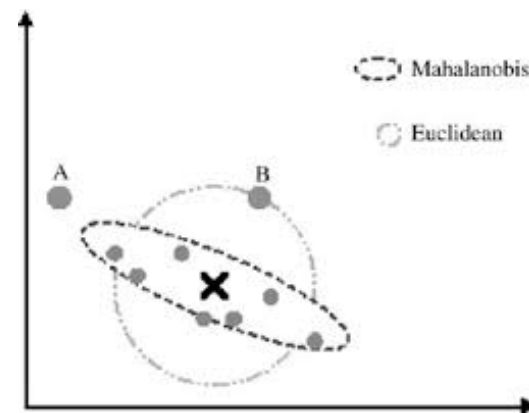
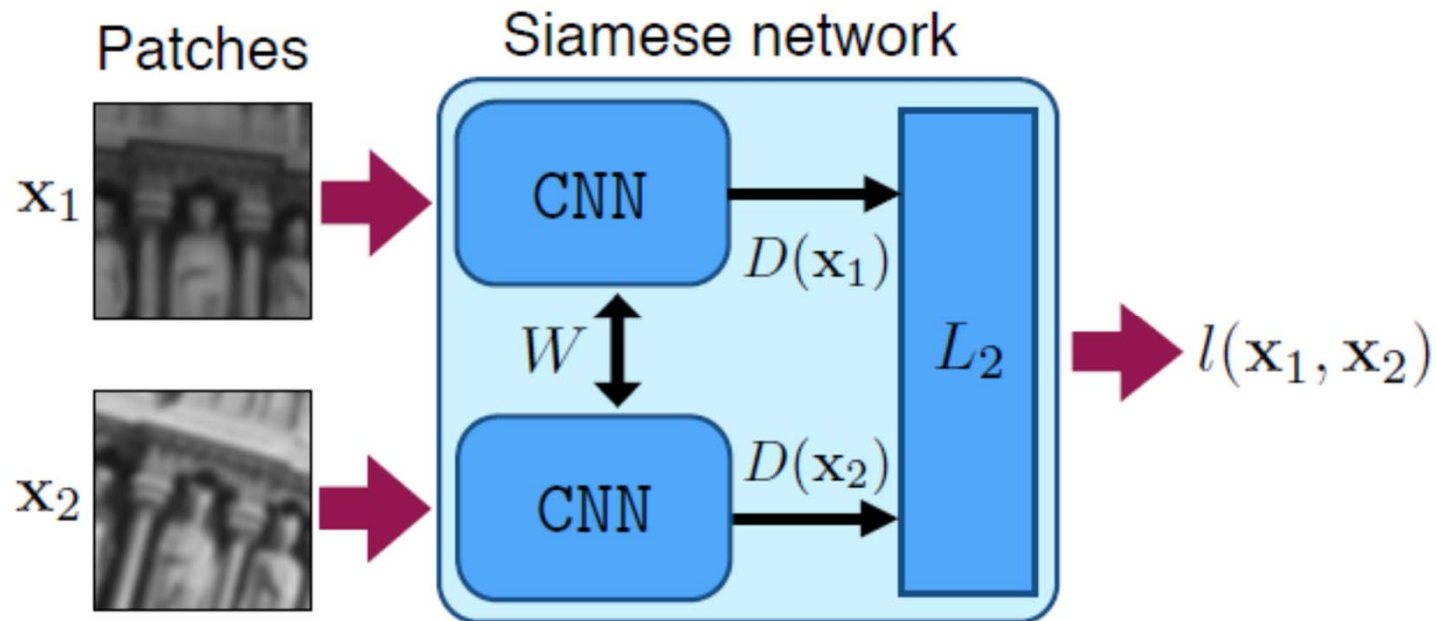


Image Source:
Google

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).

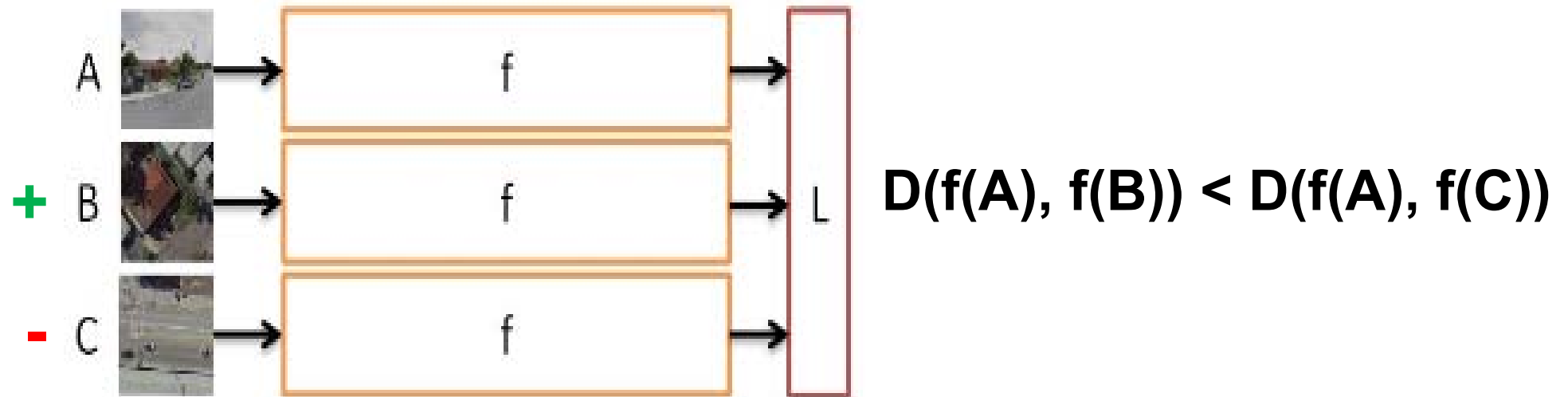
Similarity Learning: Siamese CNN

- 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



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

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**

Image Database
7.4M images



Retrieval
→
SfM

3D models:
551 training / 162 validation



[Schonberger et al. CVPR'15]
[Radenovic et al. CVPR'16]

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

query



diverse hard negatives
top k: one per 3D model



Negative images

query



top 1 by CNN



top 1 by BoW



random from
top k by BoW

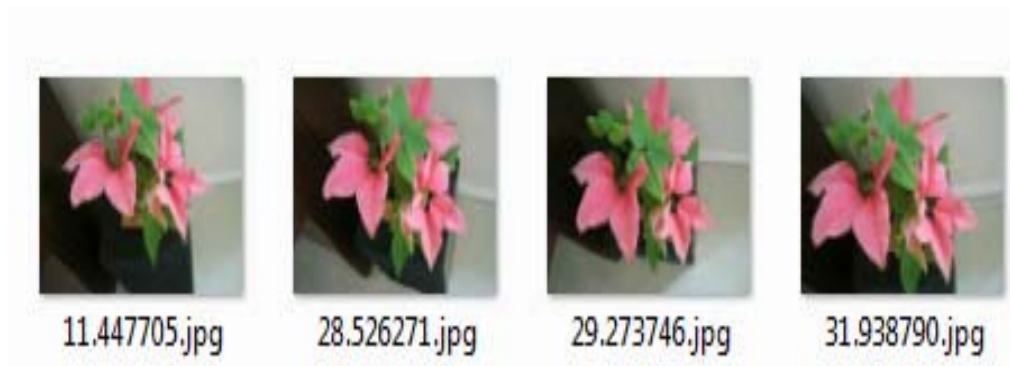
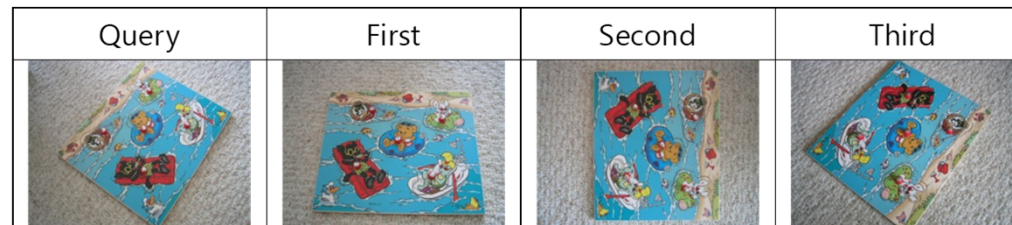


Positive images

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

PA1

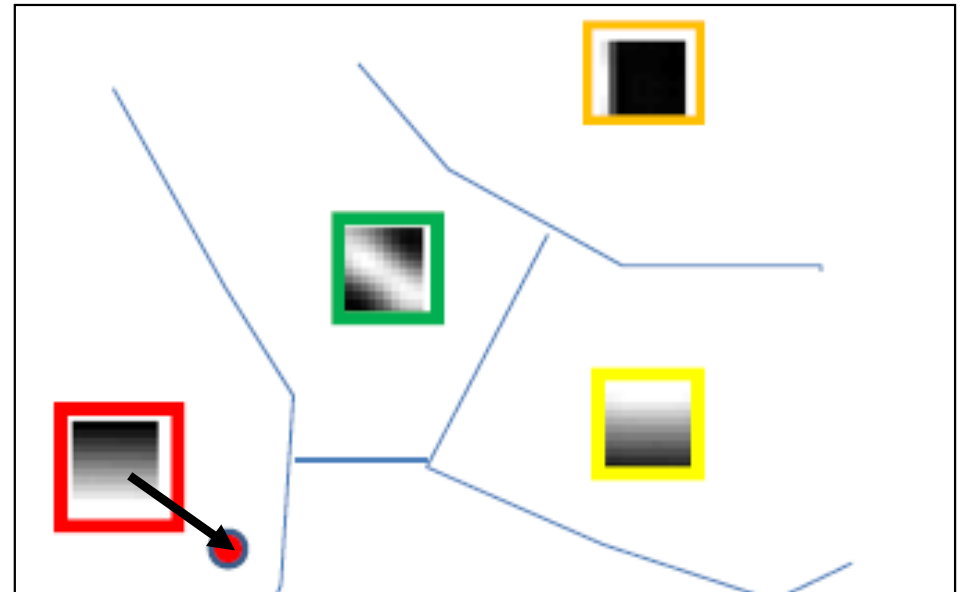
- Understand and implement a basic image retrieval system
- Use the original UKBenchmark
- Measure its accuracy



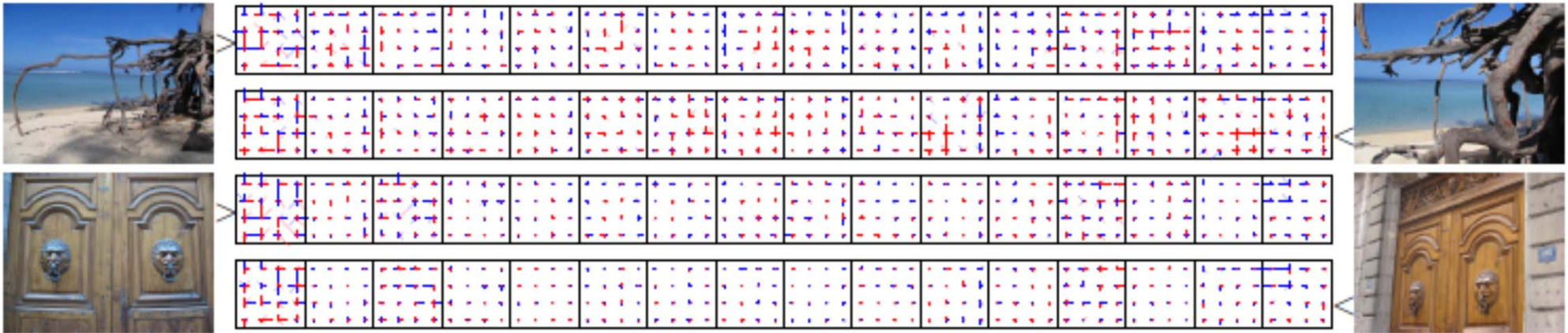
VLAD (Vector of Locally Aggregated Descriptors)

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

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



VLAD



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

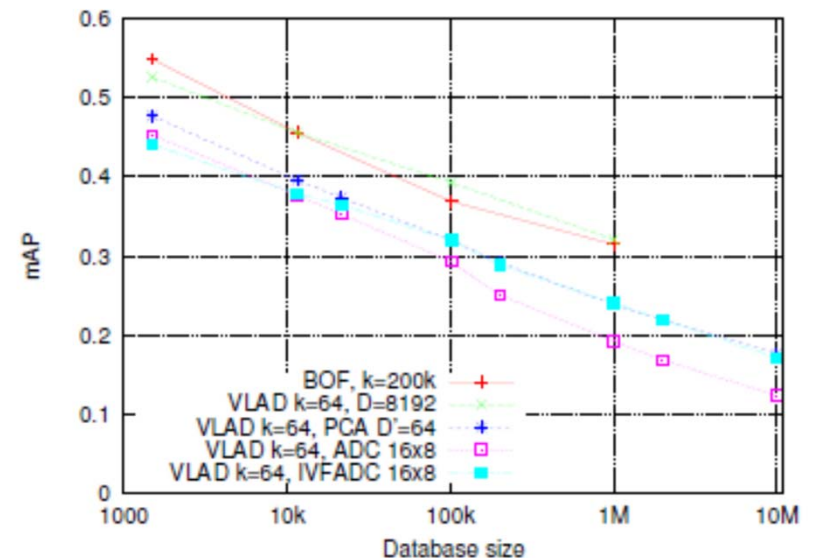
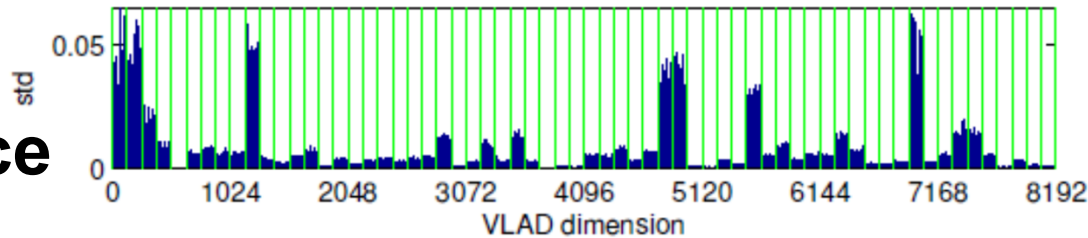


Figure 5. Search accuracy as a function of the database size.

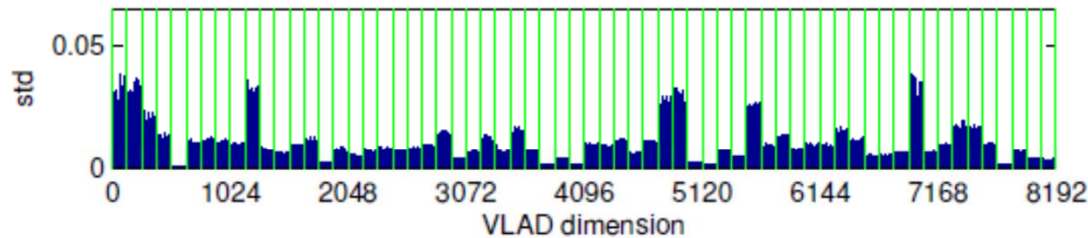
Normalization for VLAD

- Results in better accuracy

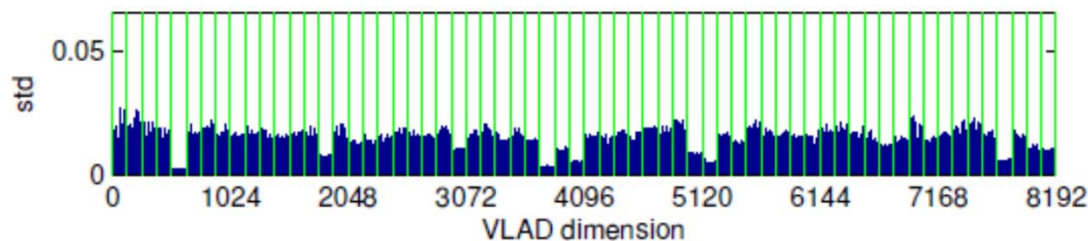
Variance



(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

NetVLAD: CNN architecture for weakly supervised place recognition

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

1. Legacy and historical imagery



2. Improve accuracy of GPS (augmented reality, navigation in robotics)



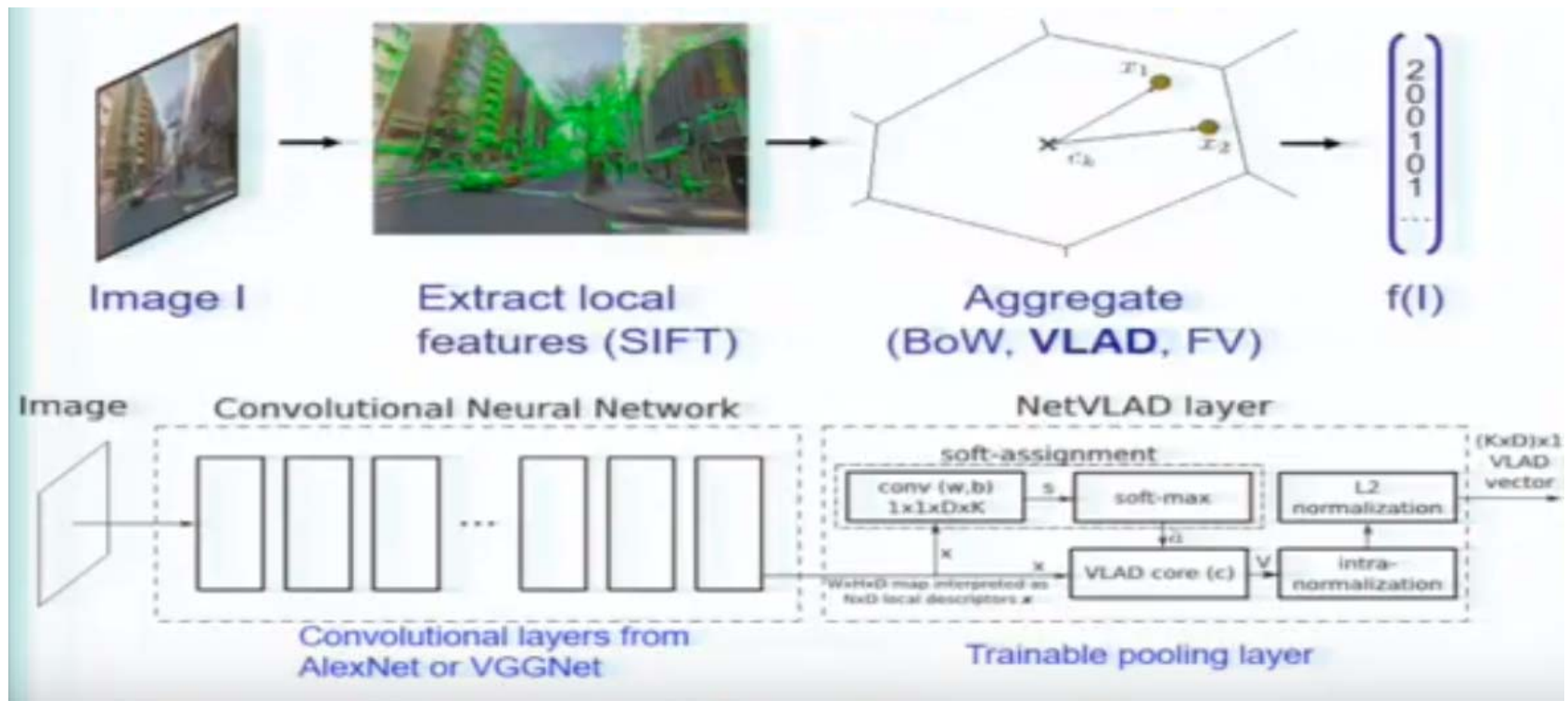
3. Understand personal photo collections



From the author talk

Mimic the classical approach

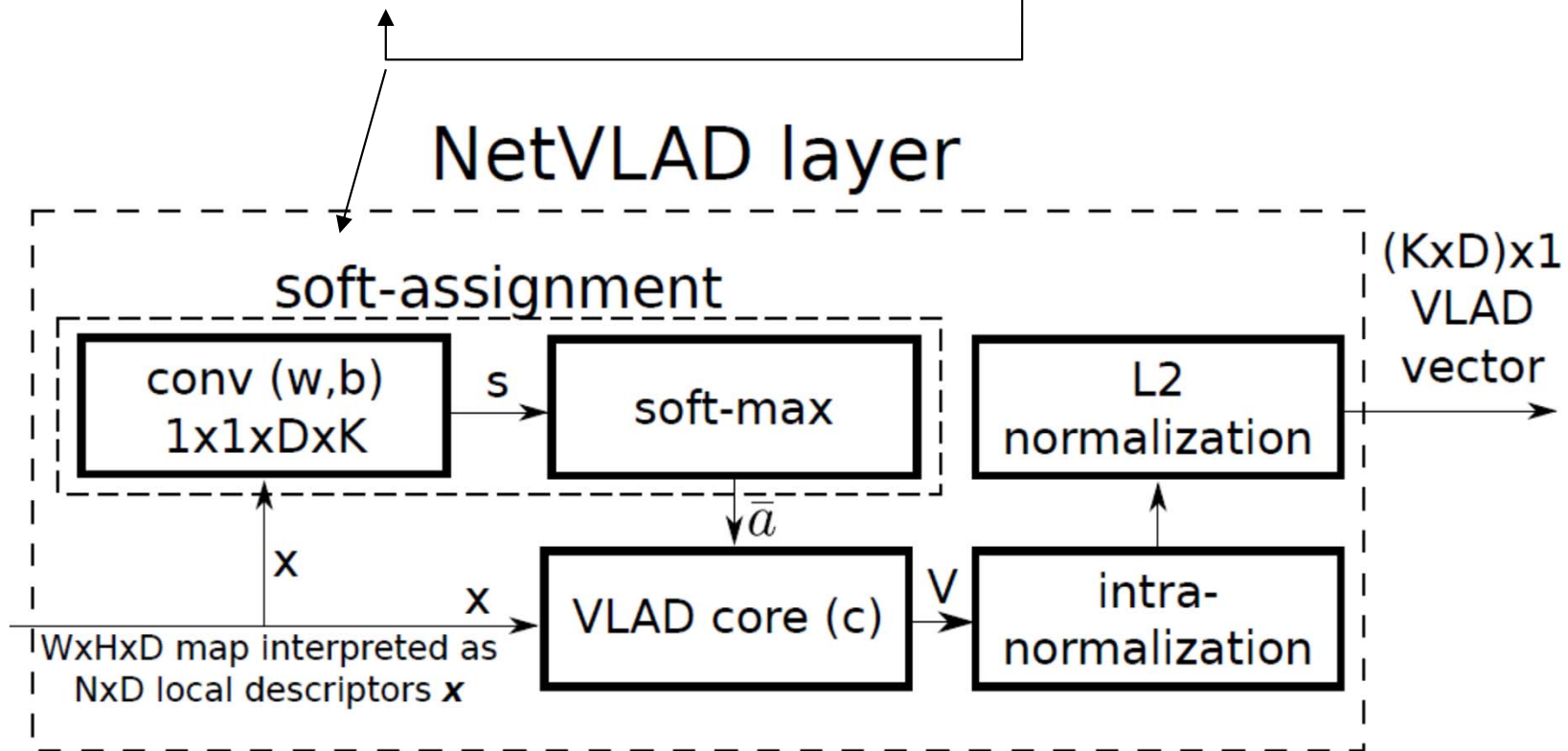
- Make it end-to-end trainable for achieving better accuracy



Trainable VLAD

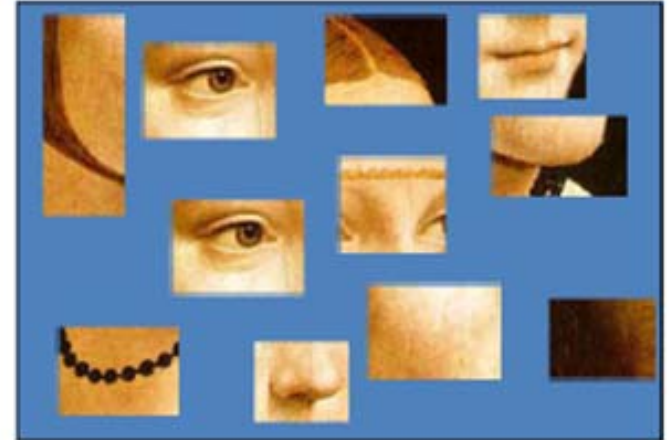
- **Hard assignment to soft assignment using the soft-max, to make it differentiable**

$$V(j, k) = \sum_{i=1}^N a_k(\mathbf{x}_i) (x_i(j) - c_k(j)), \quad \bar{a}_k(\mathbf{x}_i) = \frac{e^{-\alpha \|\mathbf{x}_i - \mathbf{c}_k\|^2}}{\sum_{k'} e^{-\alpha \|\mathbf{x}_i - \mathbf{c}_{k'}\|^2}},$$



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
