ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Thị giác máy tính

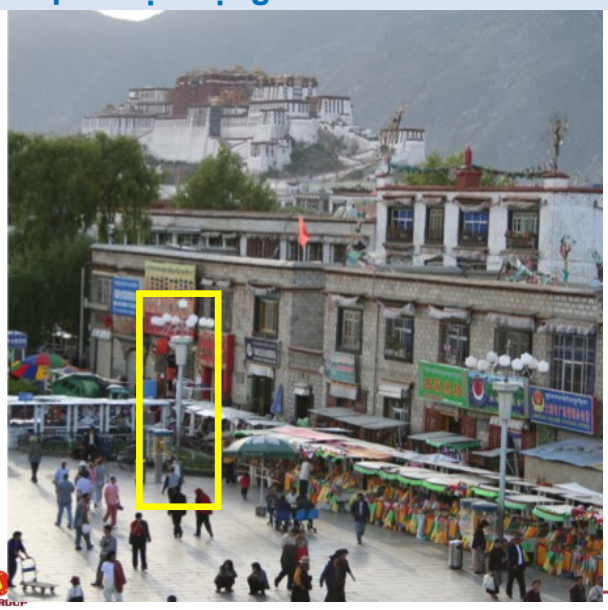
Bài 8: Phân loại ảnh

Nội dung buổi học

- Tổng quan về thị giác ngữ nghĩa
- Bài toán phân lớp / nhận dạng ảnh
- Mô hình Bag-of-words
  - Mô hình
  - Vocabulary tree
- Đánh giá
- Phân lớp và lựa chọn siêu tham số?

Thị giác ngữ nghĩa
Semantic vision

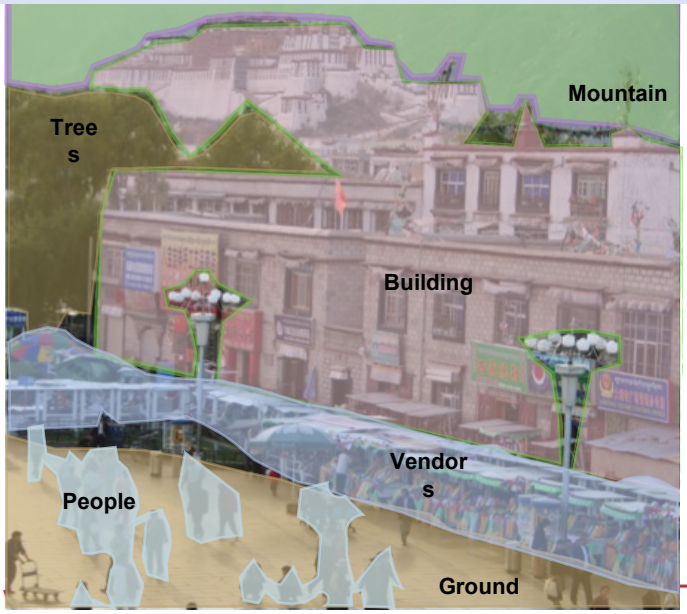
Đây có phải đèn giao thông?
(Phân lớp / nhận dạng – Classification / recognition)



Con người xuất hiện ở đâu trên ảnh?
(Phát hiện - Detection)

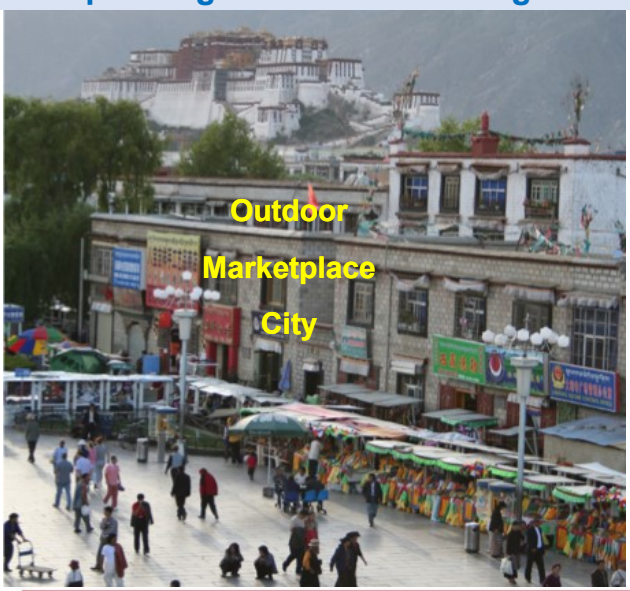
Đây có đúng là địa điểm Potala?
(Định danh - Identification)

Có những gì trong bức ảnh này?
(Phân vùng ngữ nghĩa - semantic segmentation)



7

Đây là loại khung cảnh gì?
(Phân lớp khung cảnh - Scene categorization)



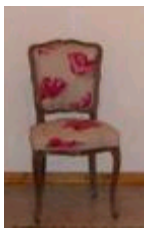
8

Những người này đang làm gì?

Nhận dạng hành động / sự kiện - Activity / Event Recognition

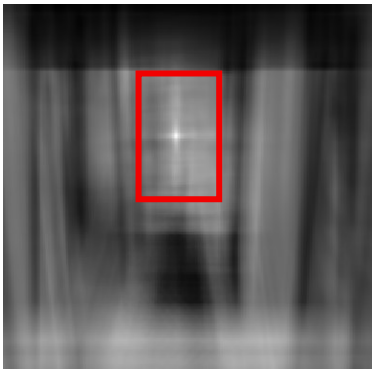
Nhận dạng đối tượng khó không?

Đây là cái ghế



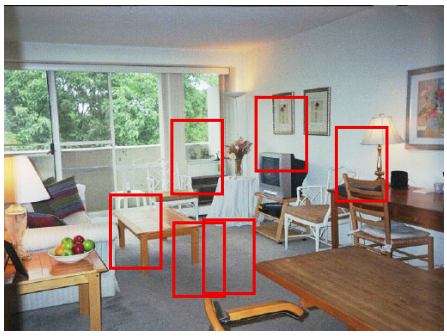
Hãy tìm cái ghế trong ảnh

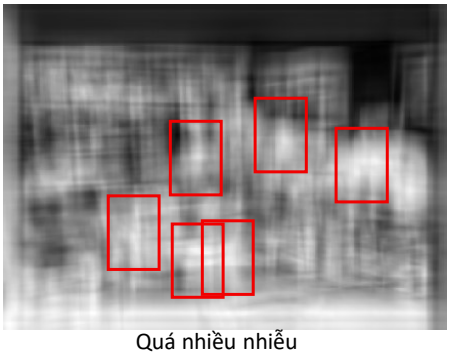
Độ tương quan sau khi chuẩn hoá



Nhận dạng đối tượng khó không?

Tìm cái ghế trong bức ảnh này



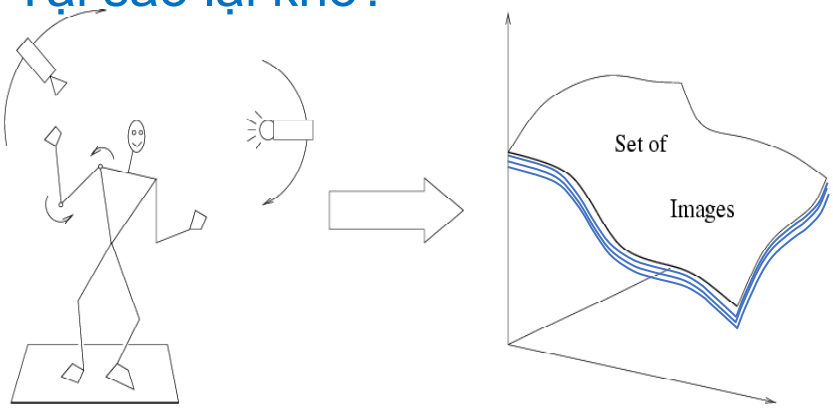


Quá nhiều nhiễu
Kỹ thuật template matching đơn giản không thể giải quyết được

Ví dụ khác …

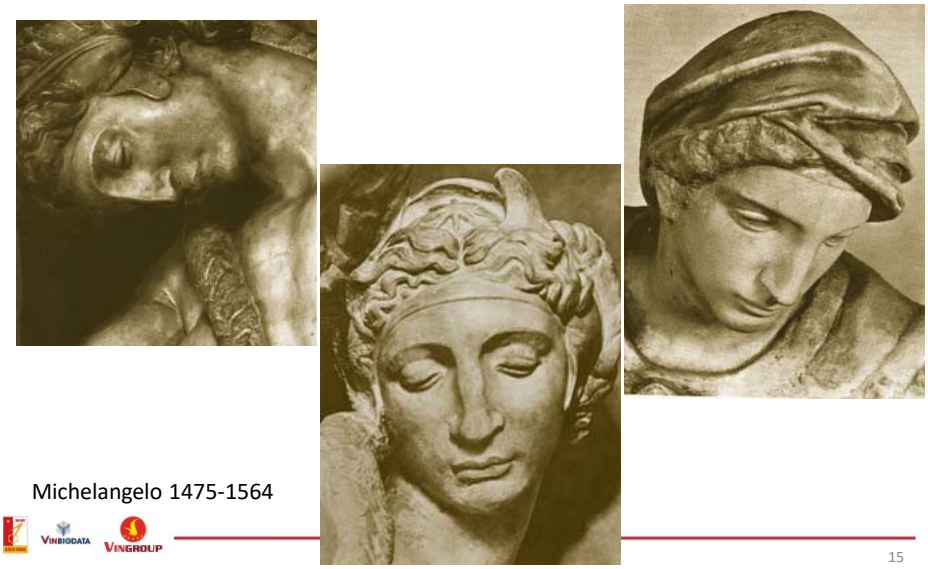
Brady, M. J., & Kersten, D. (2003). Bootstrapped learning of novel objects. J Vis, 3(6), 413-422

Tại sao lại khó?



Đa dạng: Vị trí camera
Ánh sáng
Hình dáng vật thay đổi

Thách thức: các góc chụp khác nhau



Michelangelo 1475-1564

Thách thức: ánh sáng thay đổi

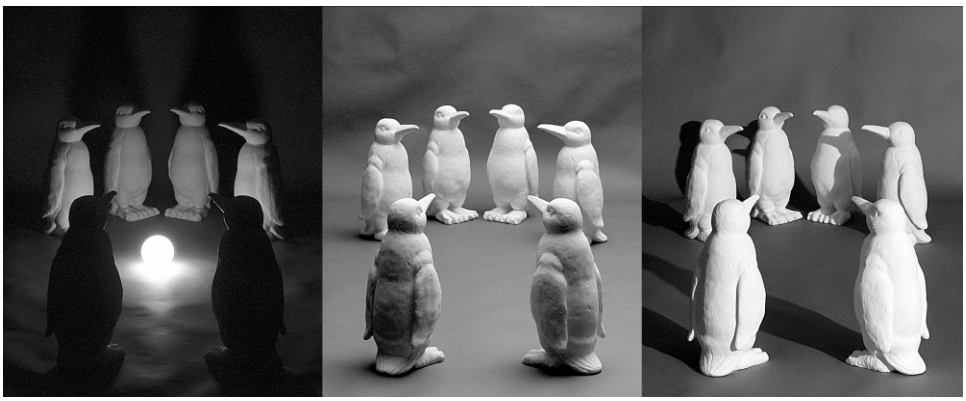


image credit: J. Koenderink

Thách thức: kích thước đa dạng

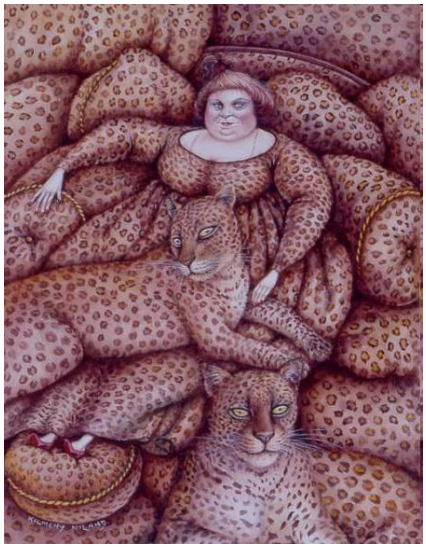
Thách thức: đối tượng biến dạng

Thách thức: che khuất

Magritte, 1957

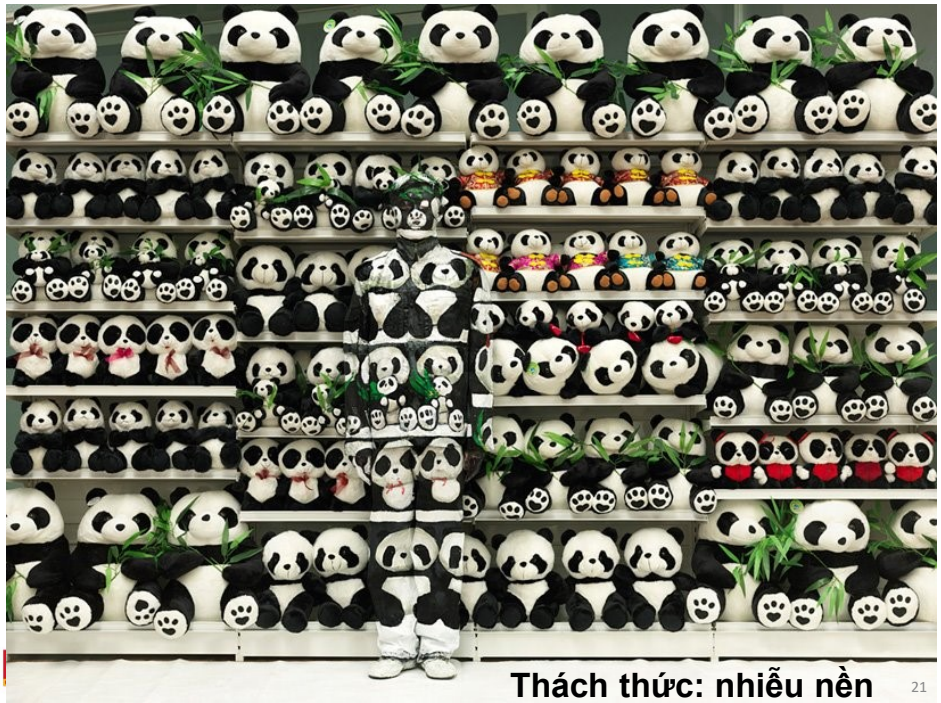
19

Thách thức: nhiễu/ nền



Kilmeny Niland. 1995

20



Thách thức: nhiễu nền

Thách thức: sự đa dạng về đối tượng trong từng lớp





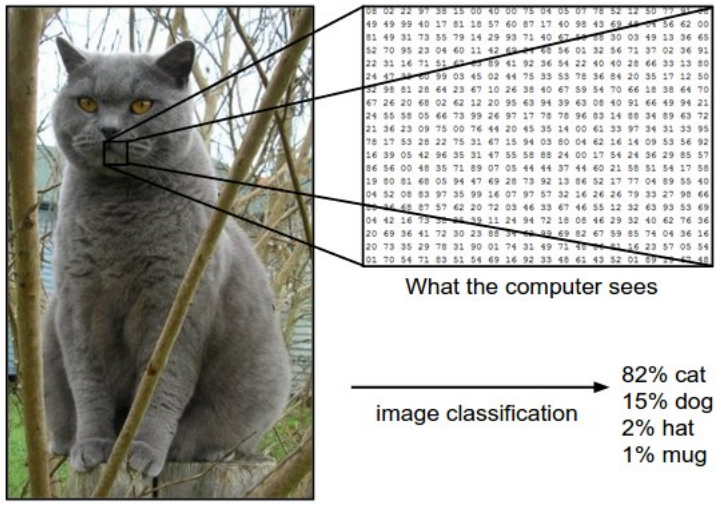

Phân lớp / nhận dạng ảnh

Phân lớp / nhận dạng ảnh

(assume given set of discrete labels)
{dog, cat, truck, plane, ...}

→ cat

Bài toán phân lớp ảnh



Tiếp cận dựa trên dữ liệu

- Thu thập bộ dữ liệu ảnh kèm nhãn các lớp
- Sử dụng ML huấn luyện một bộ phân lớp
- Đánh giá bộ phân lớp trên ảnh test

Các bước huấn luyện

Ảnh huấn luyện

Đặc trưng ảnh

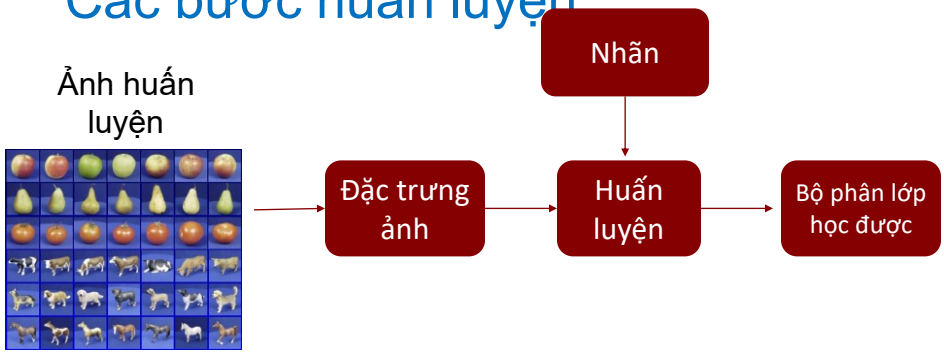
Các bước huấn luyện

Nhãn

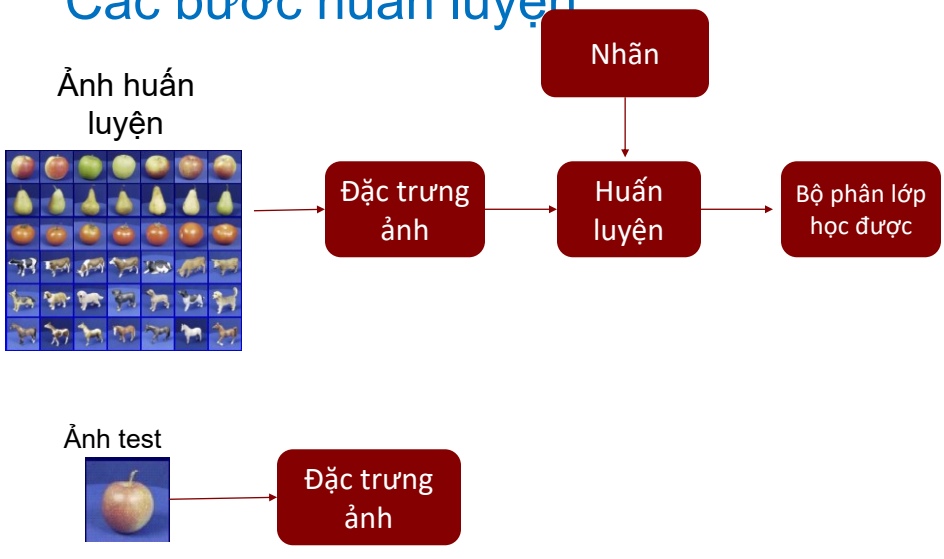
Ảnh huấn luyện

Đặc trưng ảnh → Huấn luyện

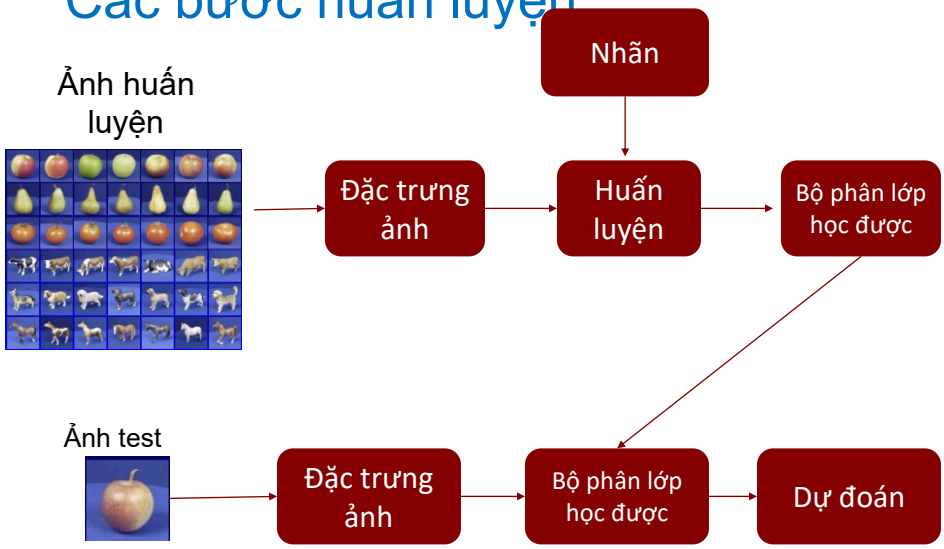
Các bước huấn luyện



Các bước huấn luyện



Các bước huấn luyện

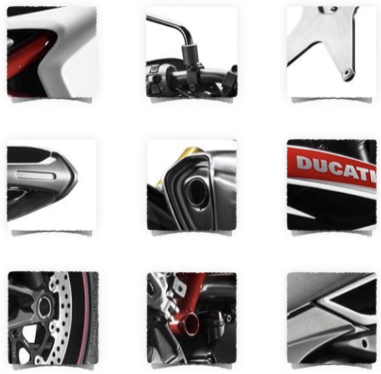


Trích chọn đặc trưng (nhắc lại)

- Hai loại đặc trưng được trích chọn từ ảnh:
  - Đặc trưng cục bộ và toàn cục
- **Đặc trưng toàn cục:**
  - Mô tả toàn bộ ảnh như 1 đối tượng
  - Đặc trưng đường biên, đặc trưng hình dạng, đặc trưng kết cấu
    - Ví dụ: Invariant Moments (Hu, Zernike), Histogram Oriented Gradients (HOG), PHOG, and Co-HOG,...
- **Đặc trưng cục bộ**:
  - Mô tả đặc trưng cục bộ mô tả từng vùng nhỏ trong ảnh, từng vùng cục bộ của đối tượng (điểm đặc trưng trong ảnh).
  - Biển diễn đặc trưng kết cấu/màu sắc trong mỗi vùng cục bộ ảnh
    - Ví dụ: SIFT, SURF, LBP, BRISK, MSER và FREAK, …

Mô hình Bag of words

Một số đặc trưng cục bộ giàu thông tin



Một đối tượng

Tập hợp các đặc trưng cục bộ

(bag-of-features)

- Giải quyết tốt vấn đề che khuất
- Bất biến với kích thước
- Bất biến với phép xoay

CalTech6 dataset

class	bag of features	bag of features	Parts-and-shape model
	Zhang et al. (2005)	Willamowski et al. (2004)	Fergus et al. (2003)
airplanes	**98.8**	97.1	90.2
cars (rear)	98.3	**98.6**	90.3
cars (side)	**95.0**	87.3	88.5
faces	**100**	99.3	96.4
motorbikes	**98.5**	98.0	92.5
spotted cats	**97.0**	—	90.0

BoW có độ chính xác khá tốt cho bài toán phân lớp ảnh

Csurka et al. (2004), Willamowski et al. (2005), Grauman & Darrell (2005), Sivic et al. (2003, 2005)

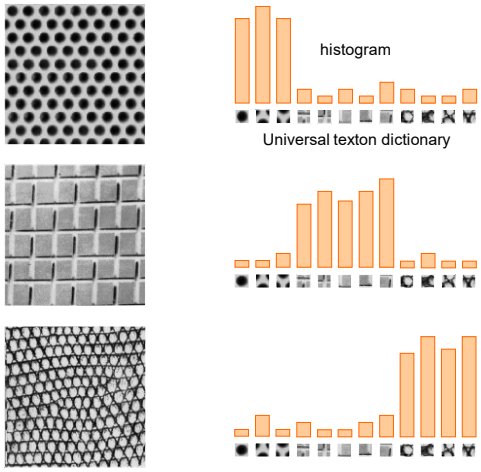
Bag-of-features

Biểu diễn một dữ liệu (văn bản, vân ảnh, hình ảnh) dưới dạng lược đồ (histogram) của các đặc trưng

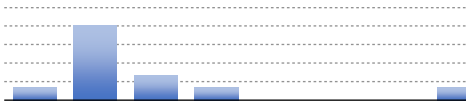
Một ý tưởng cũ

(e.g., nhận dạng vân ảnh và truy vấn thông tin)

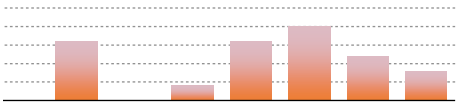
Texture recognition



Vector Space Model



1	6	2	1	0	0	0	1
Tartan	robot	CHIMP	CMU	bio	soft	ankle	sensor



0	4	0	1	4	5	3	2
Tartan	robot	CHIMP	CMU	bio	soft	ankle	sensor

G. Salton. 'Mathematics and Information Retrieval' Journal of Documentation,1979

Một văn bản (datapoint) biểu diễn dưới dạng véc-tơ tần suất xuất hiện các từ (feature)

$$\boldsymbol{v}_d = [n(w_{1,d}) \quad n(w_{2,d}) \quad \cdots \quad n(w_{T,d})]$$

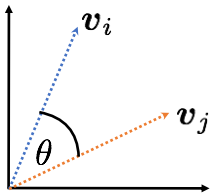
Lược đồ theo các từ

$n(\cdot)$ Tần suất xuất hiện

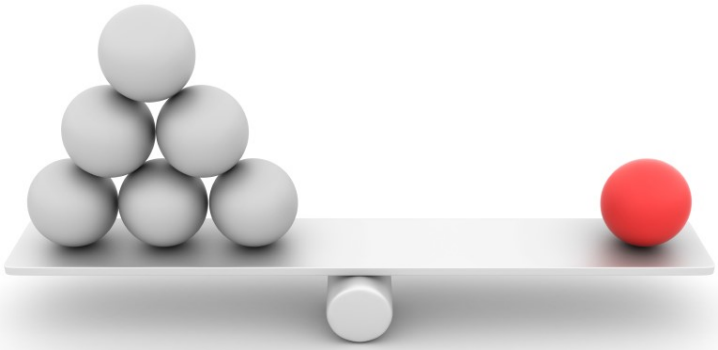
Độ tương đồng giữa hai văn bản?

Có thể dùng độ đo cosine

$$d(\boldsymbol{v}_i, \boldsymbol{v}_j) = \cos\theta = \frac{\boldsymbol{v}_i \cdot \boldsymbol{v}_j}{\|\boldsymbol{v}_i\|\|\boldsymbol{v}_j\|}$$



Nhưng không phải các từ đều quan trọng như nhau



TF-IDF

Term **F**requency **I**nverse **D**ocument **F**requency

$$\boldsymbol{v}_d = [n(w_{1,d}) \quad n(w_{2,d}) \quad \cdots \quad n(w_{T,d})]$$

Đánh trọng số của các từ theo quy tắc sau:

$$\boldsymbol{v}_d = [n(w_{1,d})\alpha_1 \quad n(w_{2,d})\alpha_2 \quad \cdots \quad n(w_{T,d})\alpha_T]$$

term frequency

inverse document frequency

$$n(w_{i,d})\alpha_i = n(w_{i,d}) \log \left\{ \frac{D}{\sum_{d'} \mathbf{1}[w_i \in d']} \right\}$$

(down-weights **common** terms)

BOW cho bài toán phân lớp ảnh

Mô hình túi từ

- Đặc trưng cục bộ ~~ 1 từ
- 1 ảnh ~~ 1 văn bản
- Sử dụng mô hình vector: ảnh = tần suất xuất hiện của các từ trực quan



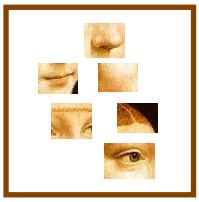

Dictionary Learning:
Học Visual Words bằng phân cụm

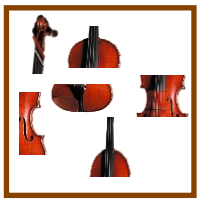
Encode:
Xây dựng Bags-of-Words (BOW) vectors cho từng ảnh

Classify:
Huấn luyện và test dữ liệu mới sử dụng BOWs

Dictionary Learning:
Học Visual Words bằng phân cụm

1. Trích xuất đặc trưng (e.g., SIFT) từ các ảnh



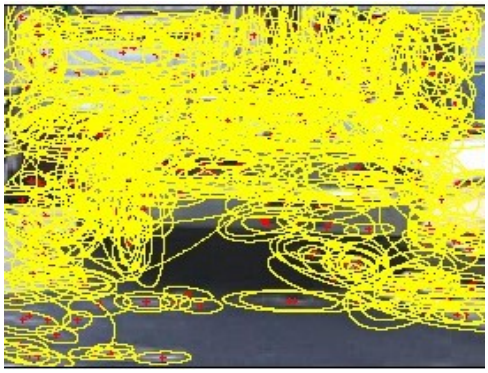


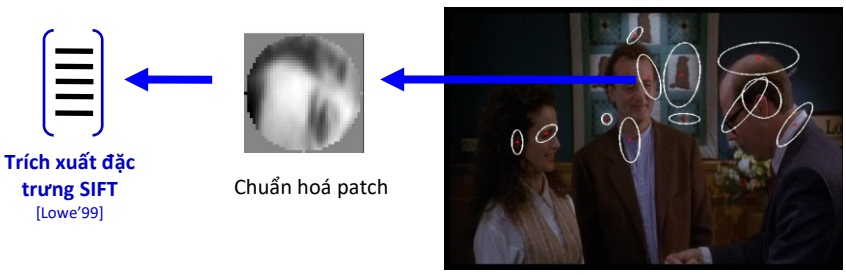
Dictionary Learning:
Học Visual Words bằng phân cụm

2. Học visual dictionary (e.g., phân cụm K-means)

What kinds of features can we extract?

- Regular grid
  - Vogel & Schiele, 2003
  - Fei-Fei & Perona, 2005
- Interest point detector
  - Csurka et al. 2004
  - Fei-Fei & Perona, 2005
  - Sivic et al. 2005
- Other methods
  - Random sampling (Vidal-Naquet & Ullman, 2002)
  - Segmentation-based patches (Barnard et al. 2003)

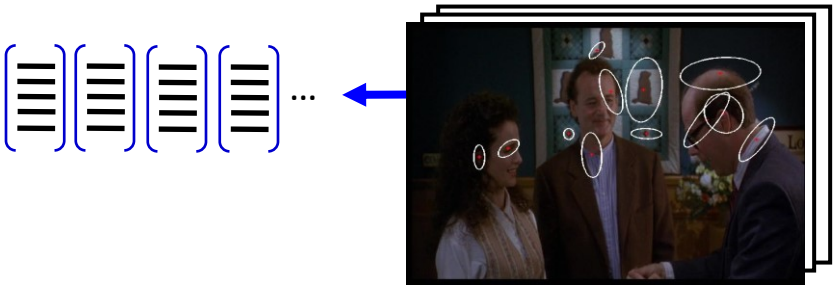


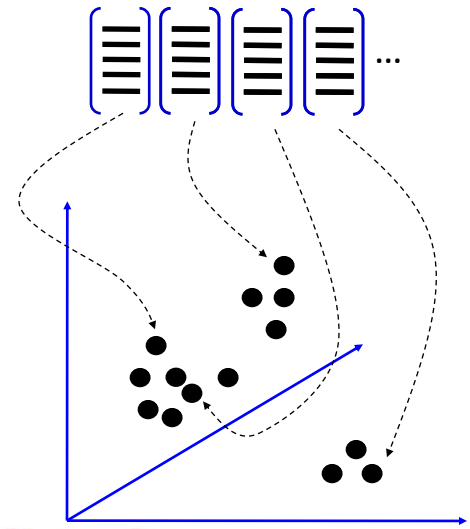


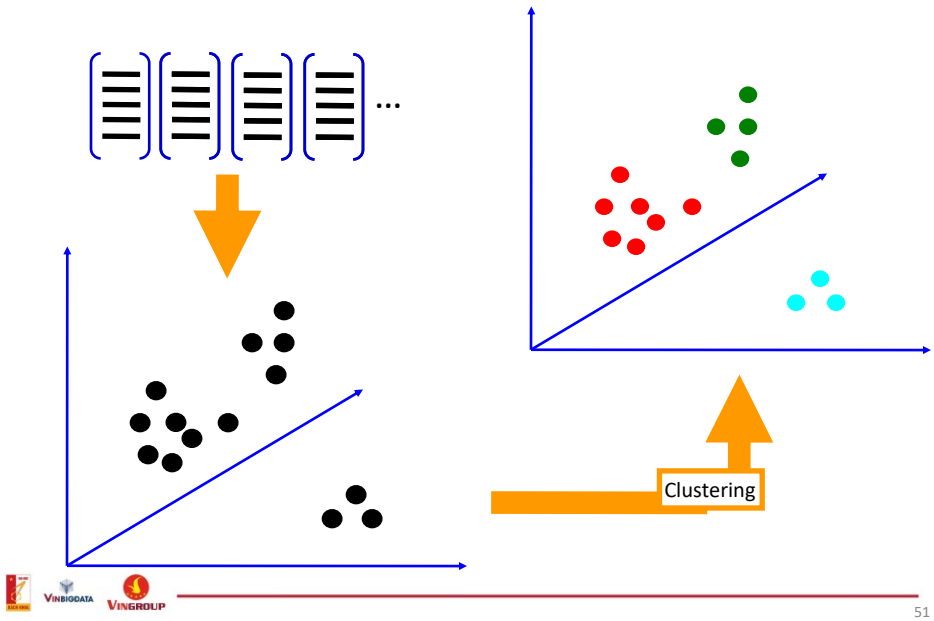
Trích xuất đặc trưng SIFT
[Lowe'99]

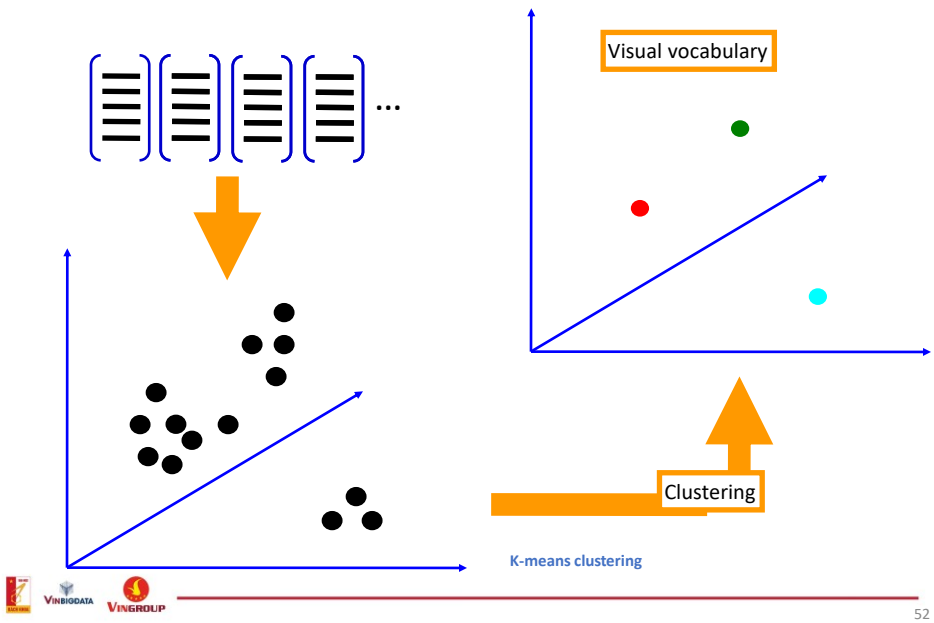
Chuẩn hoá patch

Phát hiện patches
[Mikojaczyk and Schmid '02]
[Mata, Chum, Urban & Pajdla, '02]
[Sivic & Zisserman, '03]

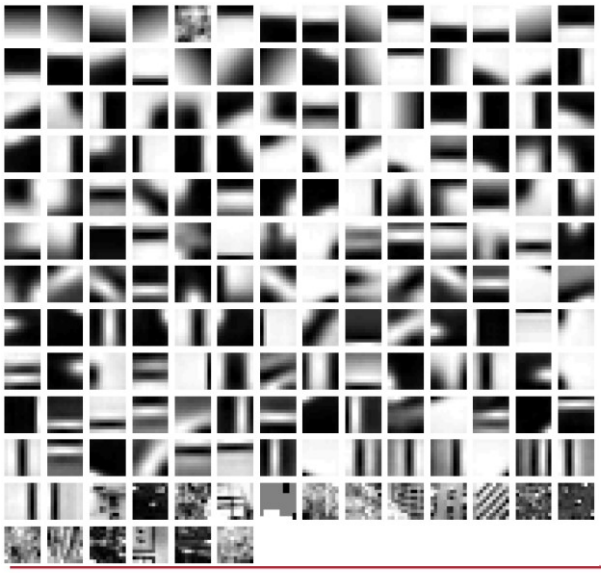




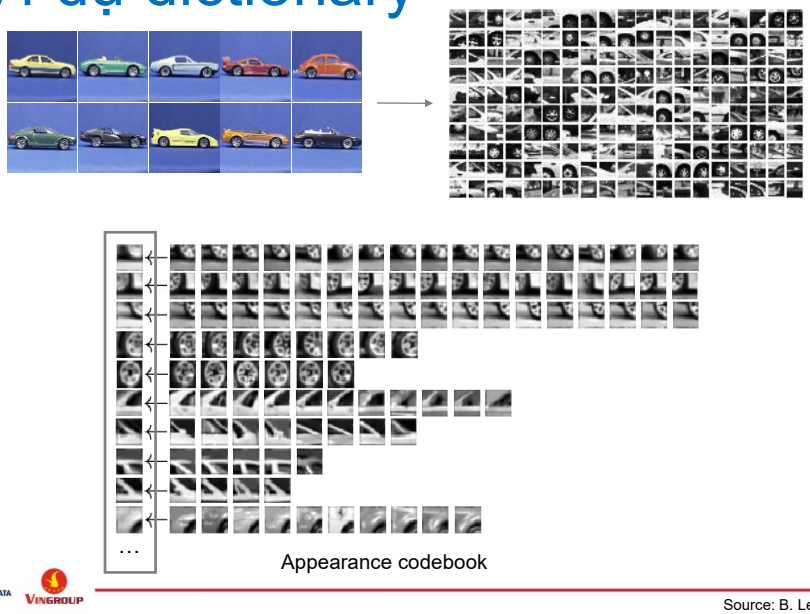

...
Clustering


...
Visual vocabulary
Clustering
K-means clustering

Ví dụ visual dictionary



Ví dụ dictionary



Source: B. Leibe

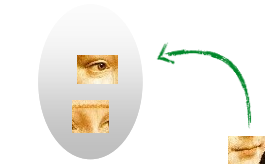
Một ví dụ khác dictionary

Source: B. Leibe

Dictionary Learning:
Học Visual Words bằng phân cụm

Encode:
Xây dựng Bags-of-Words (BOW) vectors cho từng ảnh

Classify:
Huấn luyện và test dữ liệu mới sử dụng BOWs



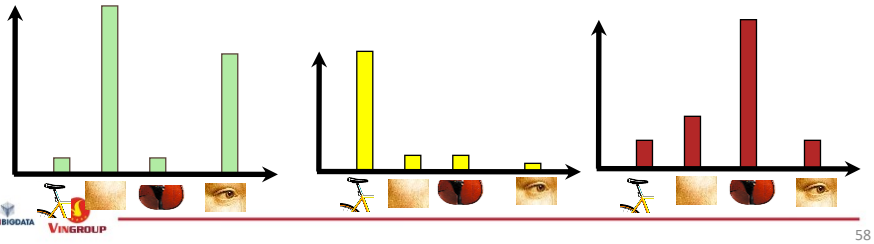
1. Lượng tử hoá: đặc trưng ảnh được gán với một visual word (tâm cụm gần nhất)

Encode:
Xây dựng Bags-of-Words (BOW) vectors cho từng ảnh

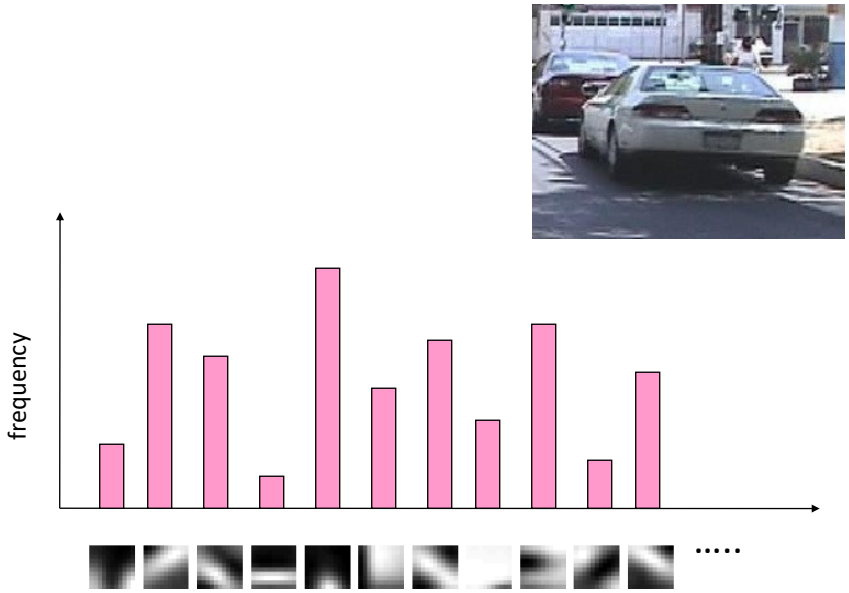
57

Encode:
Xây dựng Bags-of-Words (BOW) vectors cho từng ảnh

2. Lược đồ: đếm tần suất xuất hiện các visual word

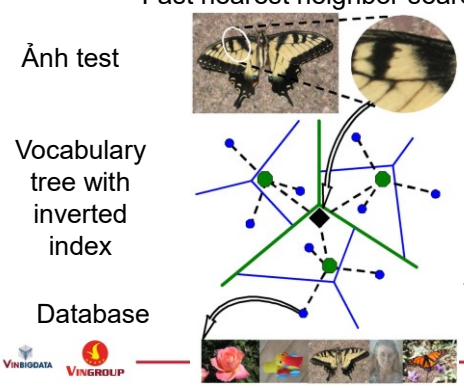


58



Khả năng mở rộng cho CSDL lớn

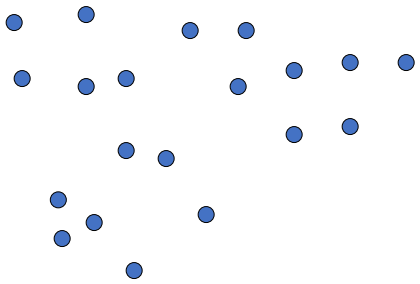
- Làm sao để truy vấn một ảnh trong CSDL hang triệu ảnh?
  - Efficient putative match generation
    - Fast nearest neighbor search, inverted indexes



D. Nistér and H. Stewénius, Scalable Recognition with a Vocabulary Tree, CVPR 2006

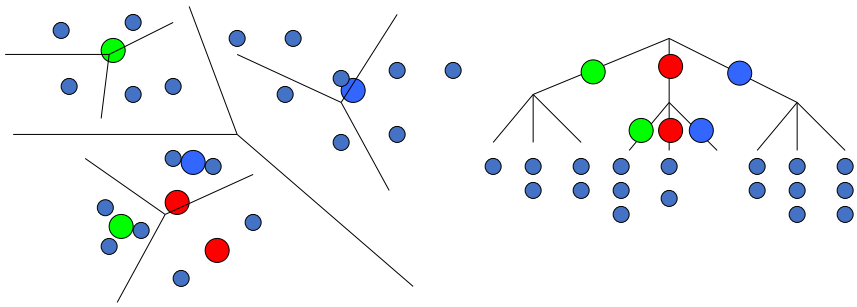
slide: S. Lazebnik

Vocabulary Tree là gì?



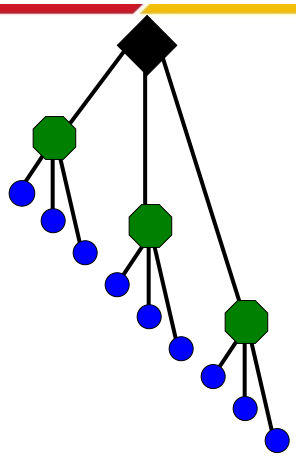
Nister and Stewenius CVPR 2006

Vocabulary Tree là gì?

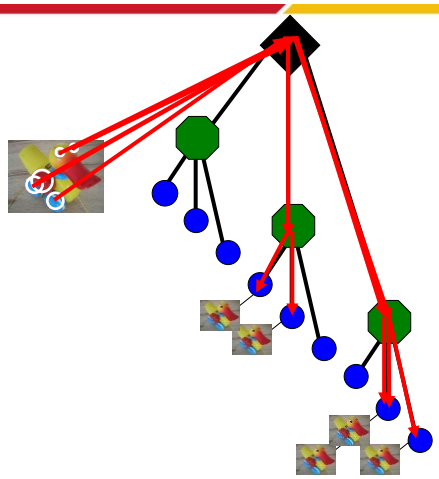


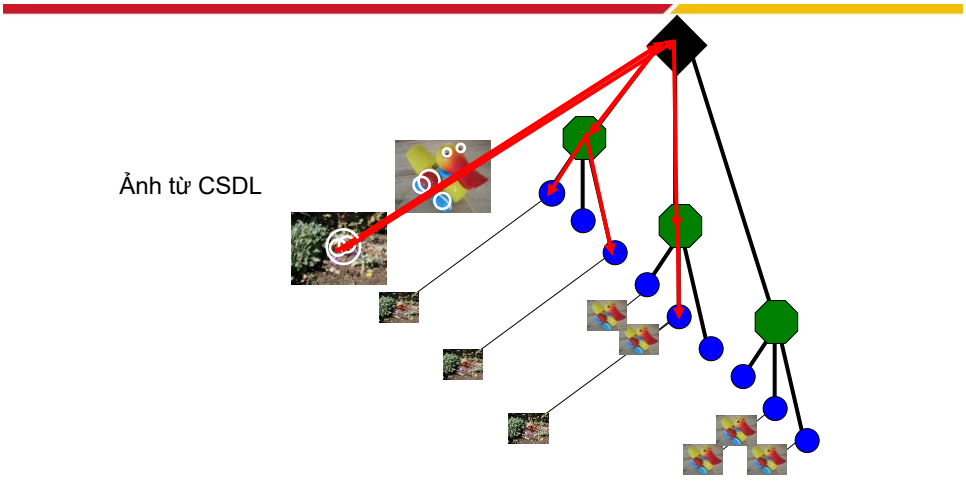
- Nhiều bước K-Means để xây dựng decision tree (offline)
- Truy vấn cây online

Nister and Stewenius CVPR 2006



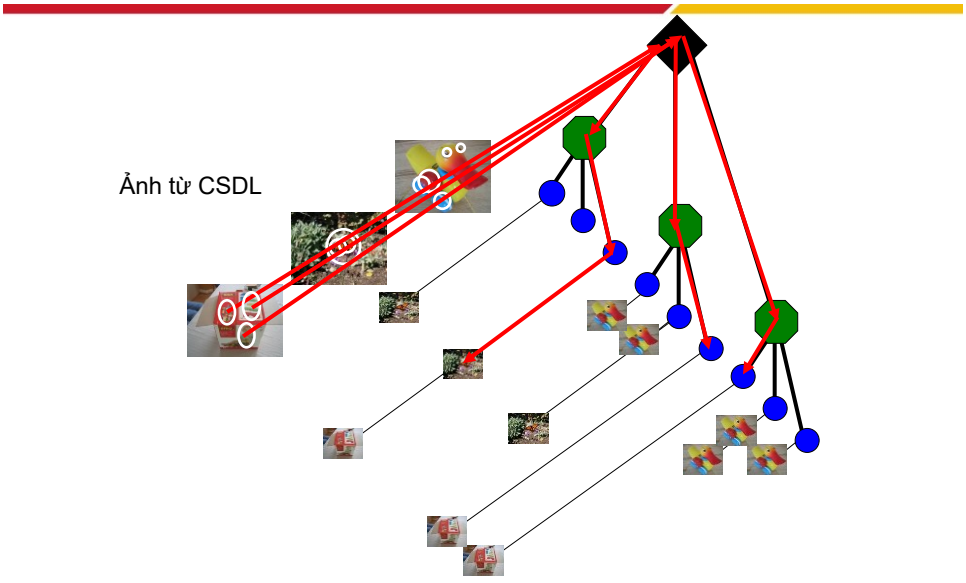
Ảnh từ CSDL





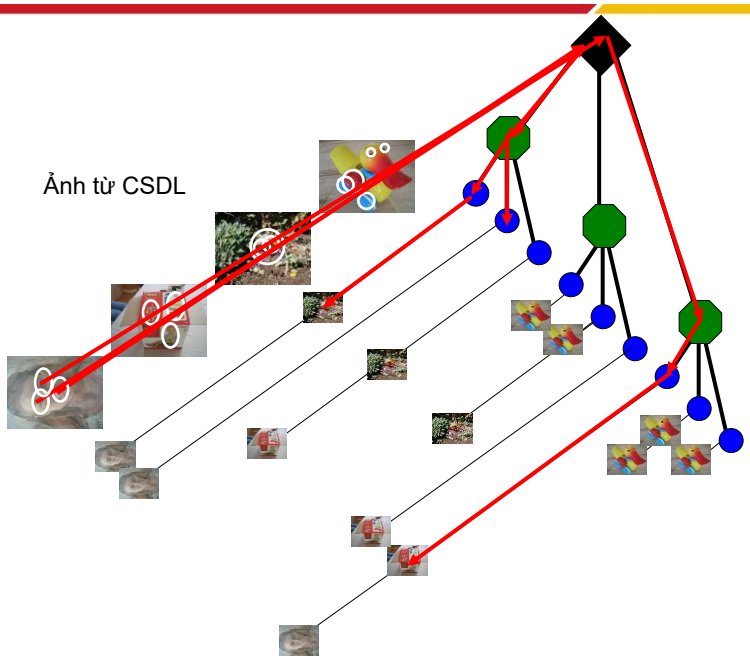
Slide credit: D. Nister

65



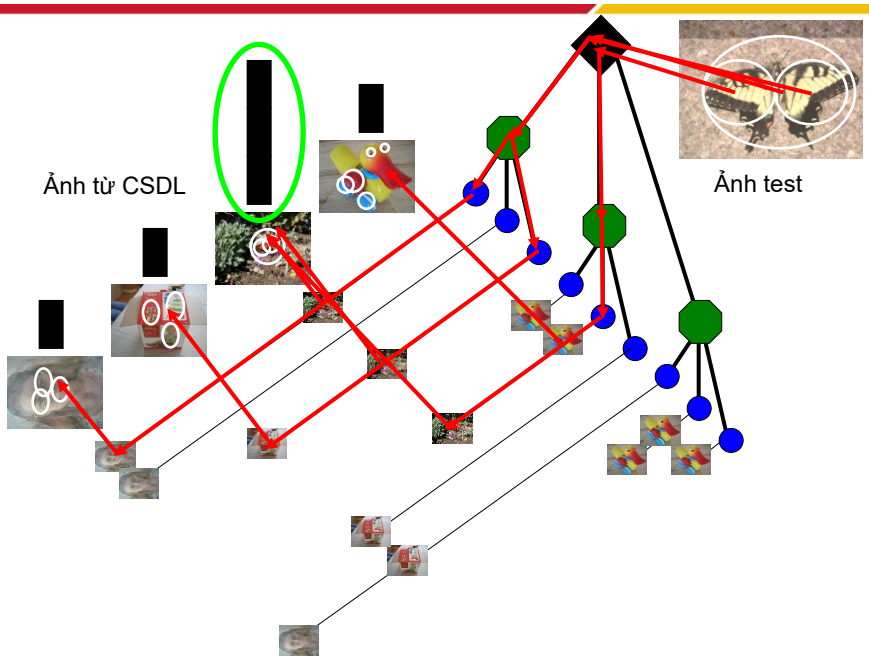
Slide credit: D. Nister

66



Slide credit: D. Nister

67



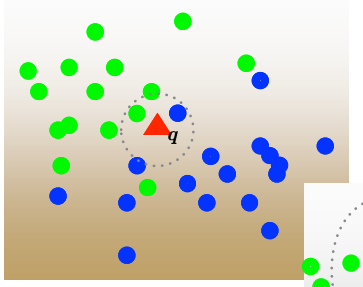
Slide credit: D. Nister

68

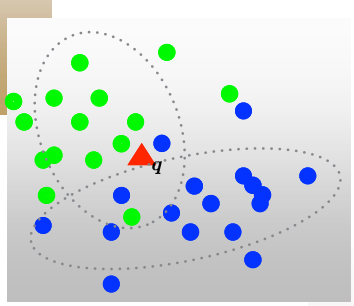
Dictionary Learning:
Học Visual Words bằng phân cụm

Encode:
Xây dựng Bags-of-Words (BOW) vectors
cho từng ảnh

Classify:
Huấn luyện và test dữ liệu mới sử dụng BOWs

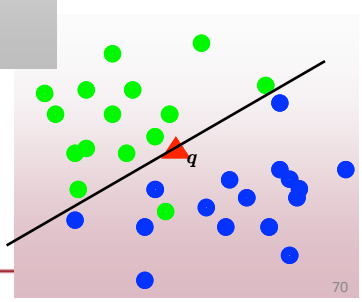


K nearest neighbors



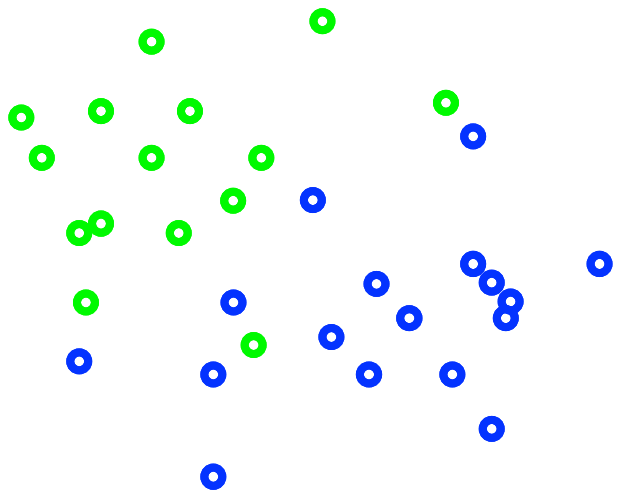
Naïve Bayes

Support Vector Machine

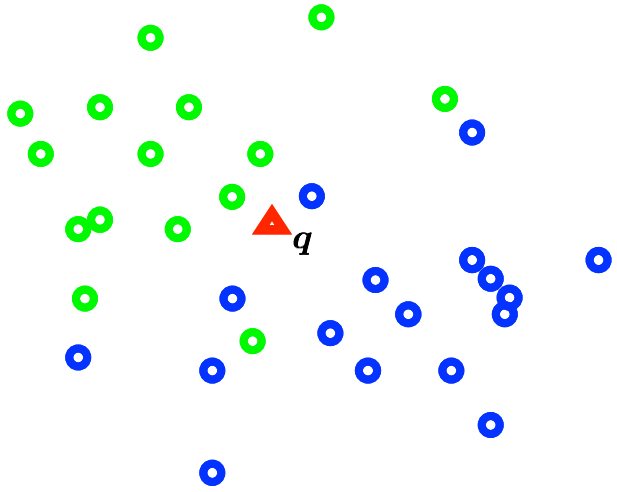


K nearest neighbors

Distribution of data from two classes

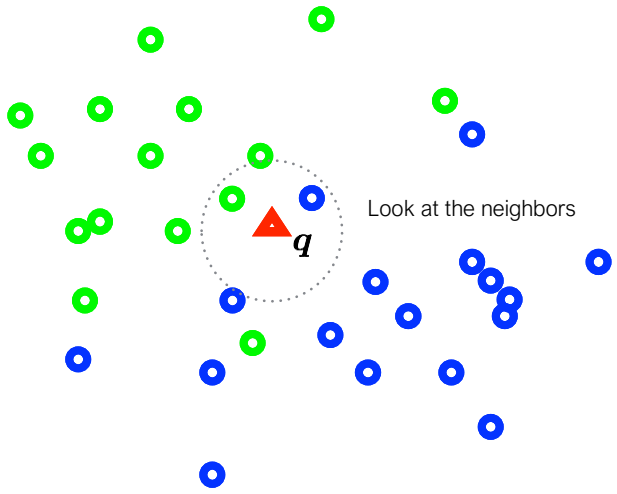


Distribution of data from two classes

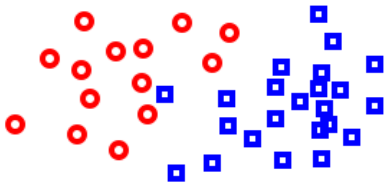


Which class does q belong too?

Distribution of data from two classes



K-Nearest Neighbor (KNN) Classifier

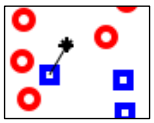


Non-parametric pattern classification approach

Consider a two class problem where each sample consists of two measurements (x,y).

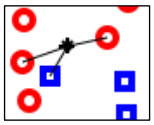
For a given query point q, assign the class of the nearest neighbor

k = 1



Compute the k nearest neighbors and assign the class by majority vote.

k = 3



Nearest Neighbor is competitive

MNIST Digit Recognition

- Handwritten digits
- 28x28 pixel images: d = 784
- 60,000 training samples
- 10,000 test samples

Yann LeCunn

	Test Error Rate (%)
Linear classifier (1-layer NN)	12.0
K-nearest-neighbors, Euclidean	5.0
K-nearest-neighbors, Euclidean, deskewed	2.4
K-NN, Tangent Distance, 16x16	1.1
K-NN, shape context matching	0.67
1000 RBF + linear classifier	3.6
SVM deg 4 polynomial	1.1
2-layer NN, 300 hidden units	4.7
2-layer NN, 300 HU, [deskewing]	1.6
LeNet-5, [distortions]	0.8
Boosted LeNet-4, [distortions]	0.7

What is the best distance metric between data points?

- Typically Euclidean distance
- Locality sensitive distance metrics
- Important to normalize.
  Dimensions have different scales

How many K?

- Typically k=1 is good
- Cross-validation (try different k!)

Distance metrics

$$D(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_N - y_N)^2}$$ Euclidean

$$D(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} = \frac{x_1 y_1 + \cdots + x_N y_N}{\sqrt{\sum_n x_n^2}\sqrt{\sum_n y_n^2}}$$ Cosine

$$D(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\sum_n \frac{(x_n - y_n)^2}{(x_n + y_n)}$$ Chi-squared

Distance metrics

L1 (Manhattan) distance

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

L2 (Euclidean) distance

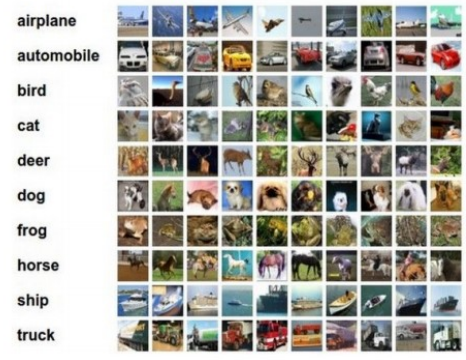
$$d_2(I_1, I_2) = \sqrt{\sum_p \left(I_1^p - I_2^p\right)^2}$$

- Two most commonly used special cases of p-norm

$$||x||_p = \left(|x_1|^p + \cdots + |x_n|^p\right)^{\frac{1}{p}} \qquad p \geq 1, x \in \mathbb{R}^n$$

CIFAR-10 and NN results

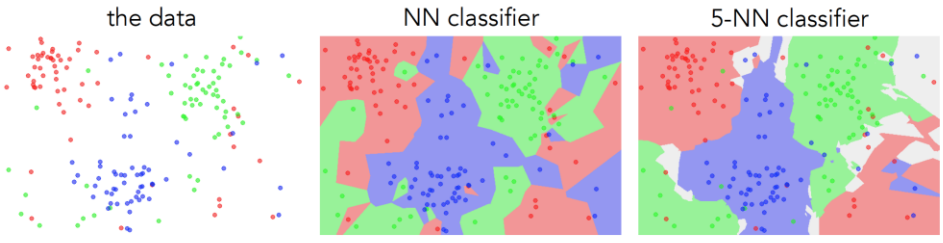
Example dataset: **CIFAR-10**
10 labels
50,000 training images
10,000 test images.



For every test image (first column), examples of nearest neighbors in rows

k-nearest neighbor

- Find the k closest points from training data
- Labels of the ***k points "vote"*** to classify



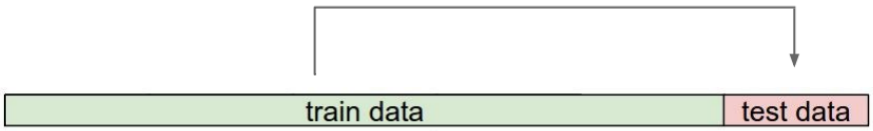
Hyperparameters

- What is the best distance to use?
- What is the best value of k to use?

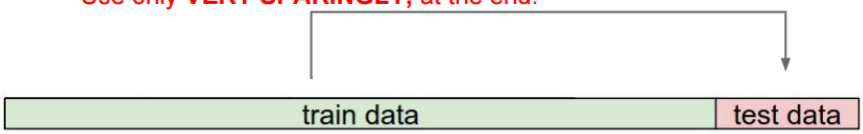
- i.e., how do we set the hyperparameters?

- Very problem-dependent
- Must try them all and see what works best

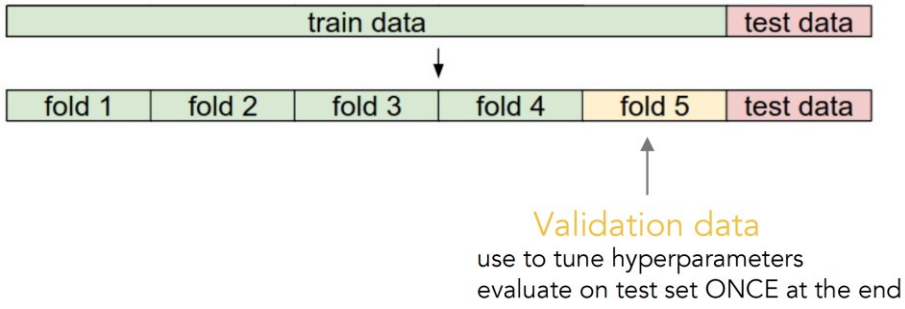
Try out what hyperparameters work best on test set.



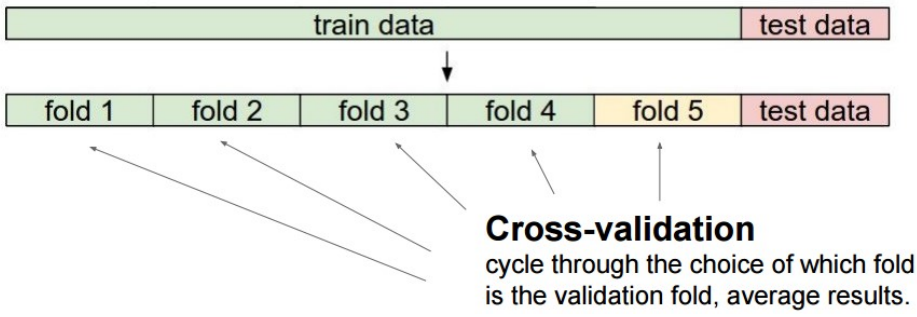
Trying out what hyperparameters work best on test set:
Very bad idea. The test set is a proxy for the generalization performance!
Use only **VERY SPARINGLY,** at the end.

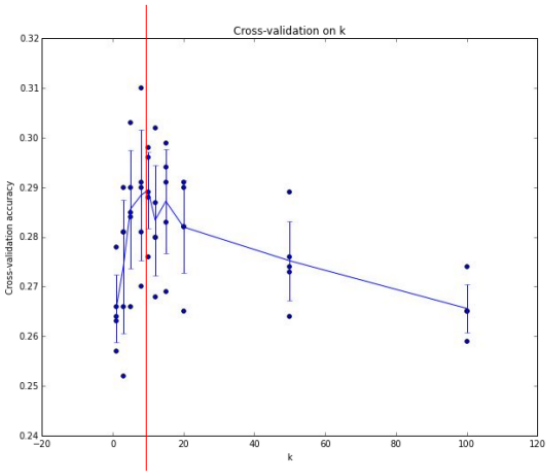


Validation



Cross-validation





Example of
5-fold cross-validation
for the value of **k.**

Each point: single
outcome.

The line goes
through the mean, bars
indicated standard
deviation

(Seems that k ~= 7 works best for this data)

How to pick hyperparameters?

- Methodology
  - Train and test
  - Train, validate, test

- Train for original model
- Validate to find hyperparameters
- Test to understand generalizability

kNN

Pros

- simple yet effective

Cons

- search is expensive (can be sped-up)
- storage requirements
- difficulties with high-dimensional data

kNN -- Complexity and Storage

- N training images, M test images

- Training: O(1)
- Testing: O(MN)

- Hmm…
  - Normally need the opposite
  - Slow training (ok), fast testing (necessary)

Độ đo đánh giá kết quả

Confusion matrix

	Predicted 1	Predicted 0
True 1	true positive	false negative
True 0	false positive	true negative

	Predicted 1	Predicted 0
True 1	TP	FN
True 0	FP	TN

	Predicted 1	Predicted 0
True 1	hits	misses
True 0	false alarms	correct rejections

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

Precision versus recall

- Accuracy
  - correct predictions / total predictions
- Precision:
  - how many of the object detections are correct?
- Recall:
  - how many of the ground truth objects can the model detect?
  - True Positive Rate (TPR)
- F1 score:

  *2*precision*recall / (precision + recall)*

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

Tài liệu tham khảo

- Kristen Grauman (CS 376: Computer Vision, Spring 2018, The University of Texas at Austin)
- Ioannis Yannis, Gkioulekas (16-385 Computer Vision, Spring 2020, CMU)

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attention!

soict.hust.edu.vn/ fb.com/groups/soict