Research Institute for Future Media Computing
未来媒体技术与研究所
Institute of Computer Vision
计算机视觉研究所

多媒体系统导论

Fundamentals of Multimedia System

授课教师：文嘉俊

邮箱：wenjiajun@szu.edu.cn

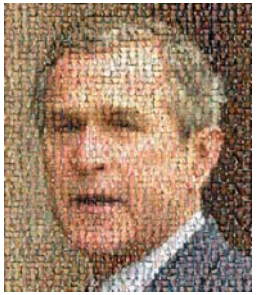
2024年春季课程

Background

- Necessity of **retrieval**
  - *Information is of no use, unless you can actually access it.*

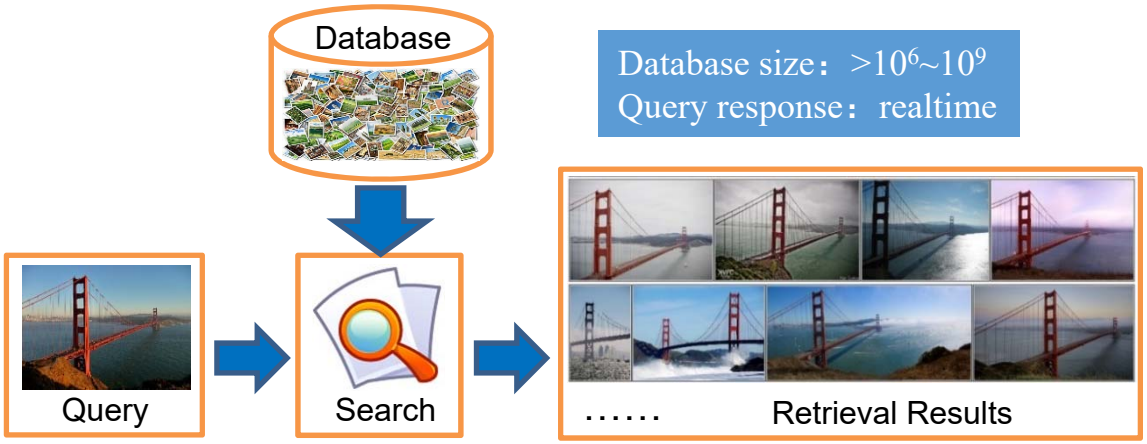
[from the TREC homepage: trec.nist.gov]

- Why do we need image retrieval?
  - "A Picture is worth thousand words"
  - Not everything can be described in text
  - Not everything is described in text



Background

- Content based image retrieval



- Potential applications of content based image retrieval

Image Retrieval

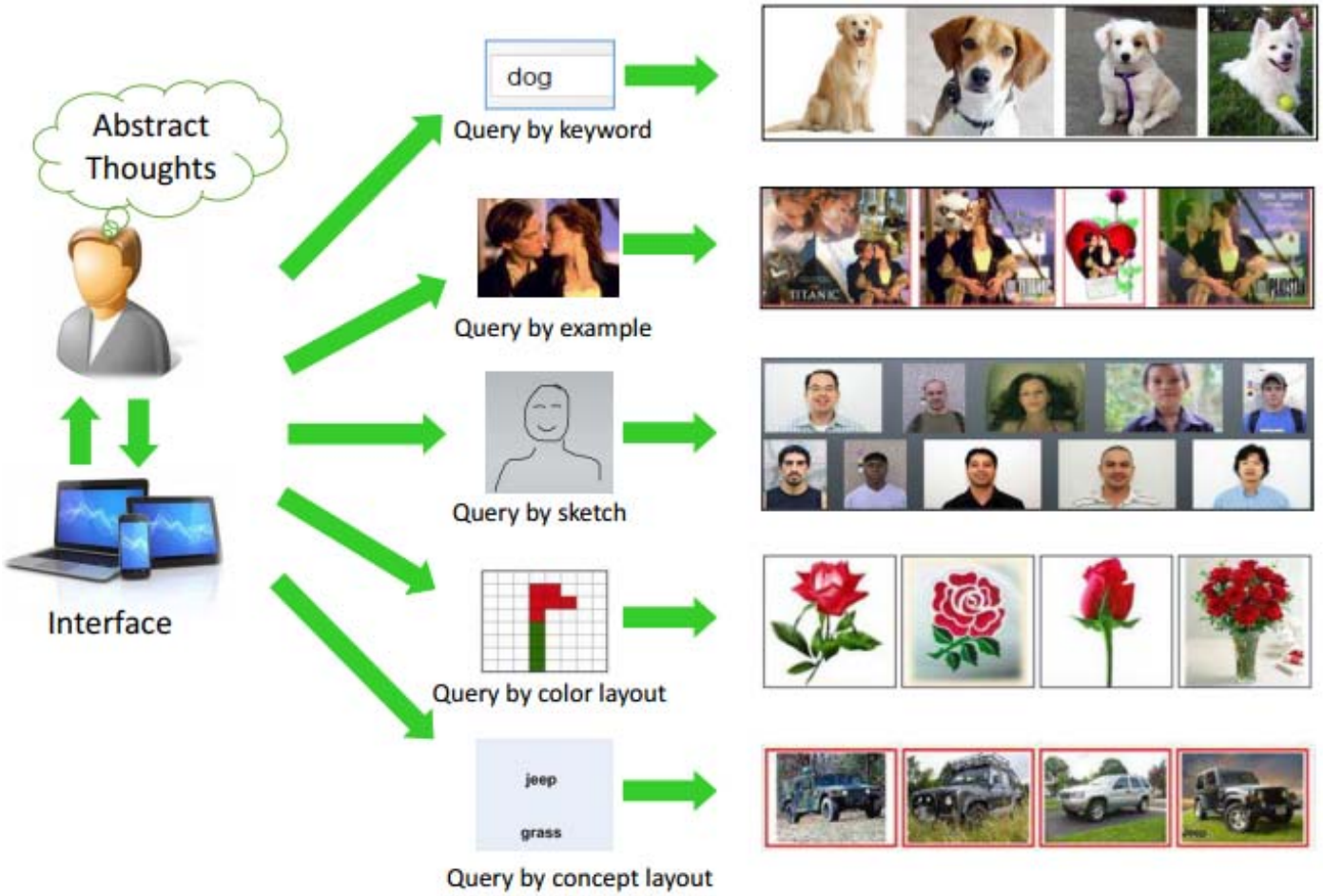
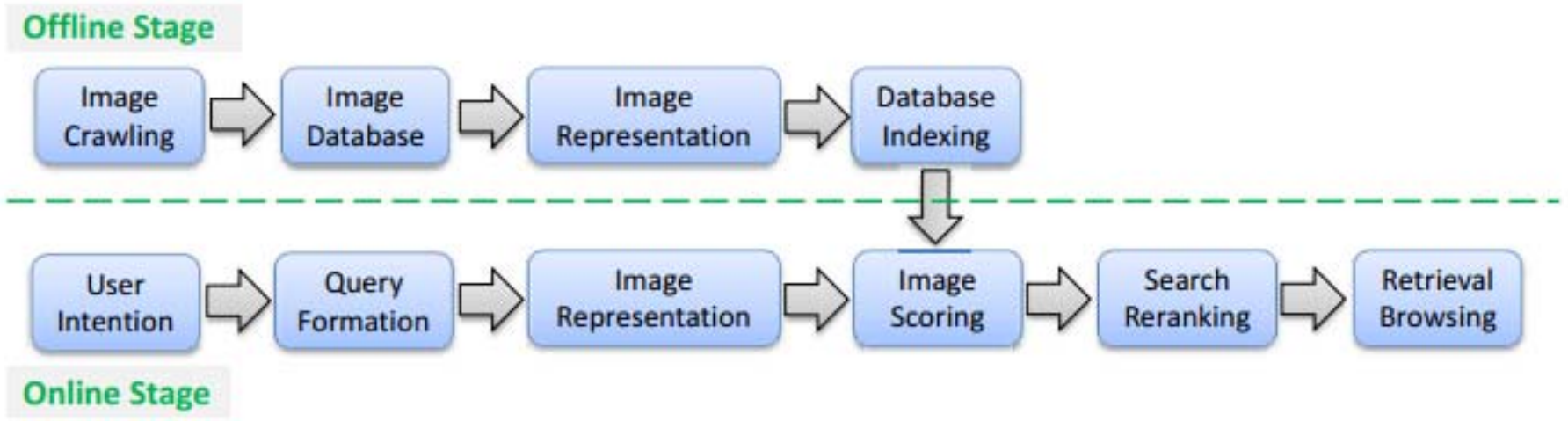


Illustration of different query schemes with the corresponding retrieval results

Why is Image Retrieval Hard ?

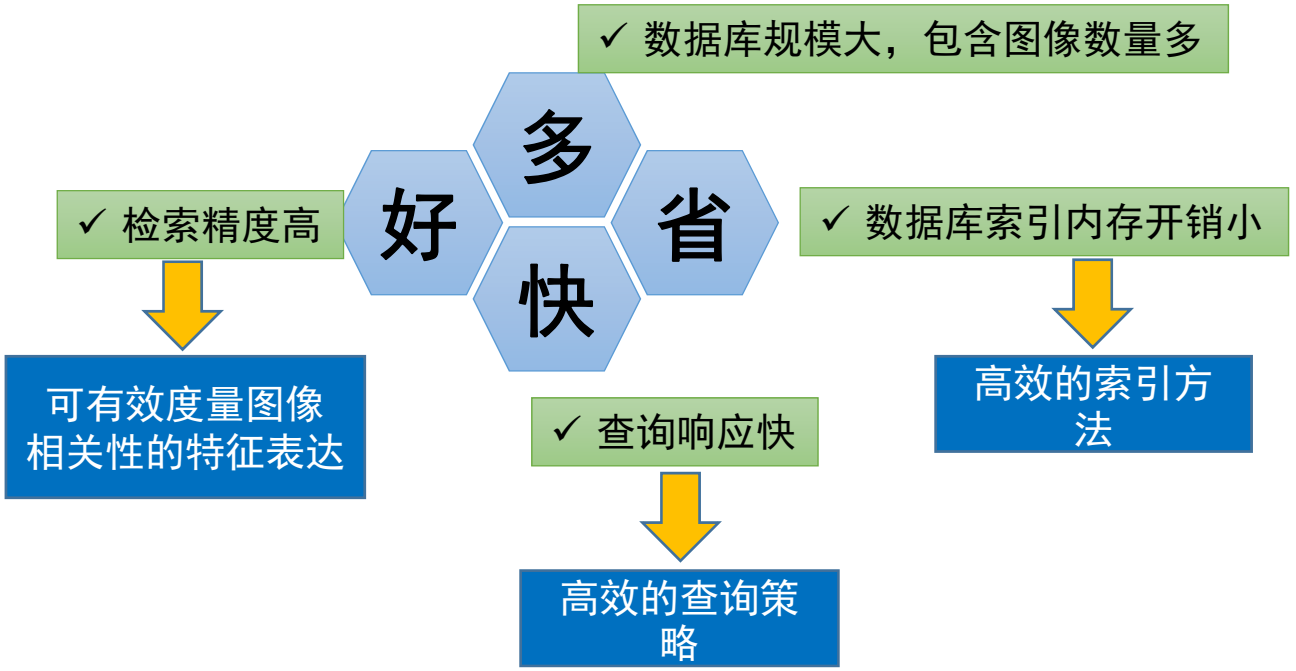
- A picture is worth a thousand words
- The meaning of an image is highly individual and subjective

Framework



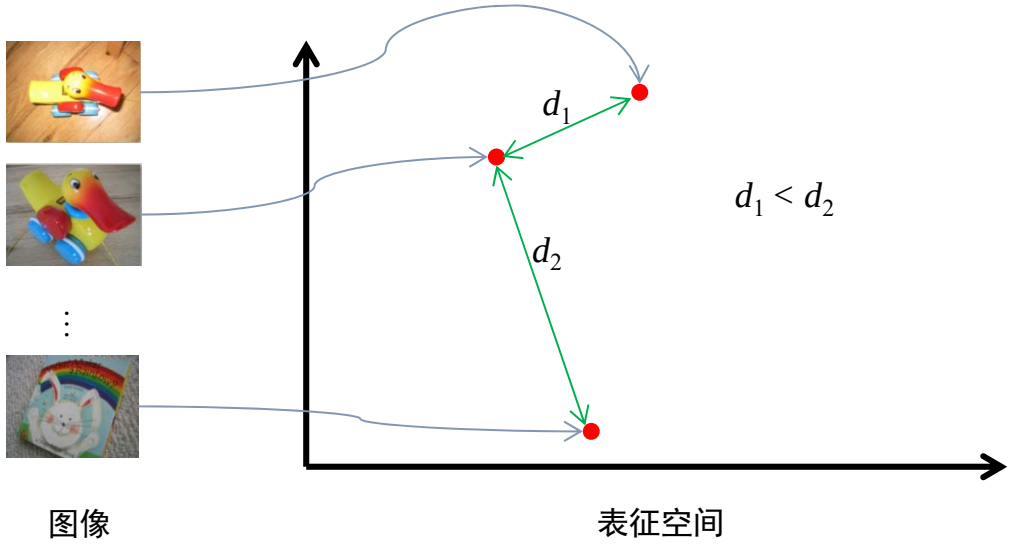
The general framework of content-based image retrieval

Problems with Image Retrieval



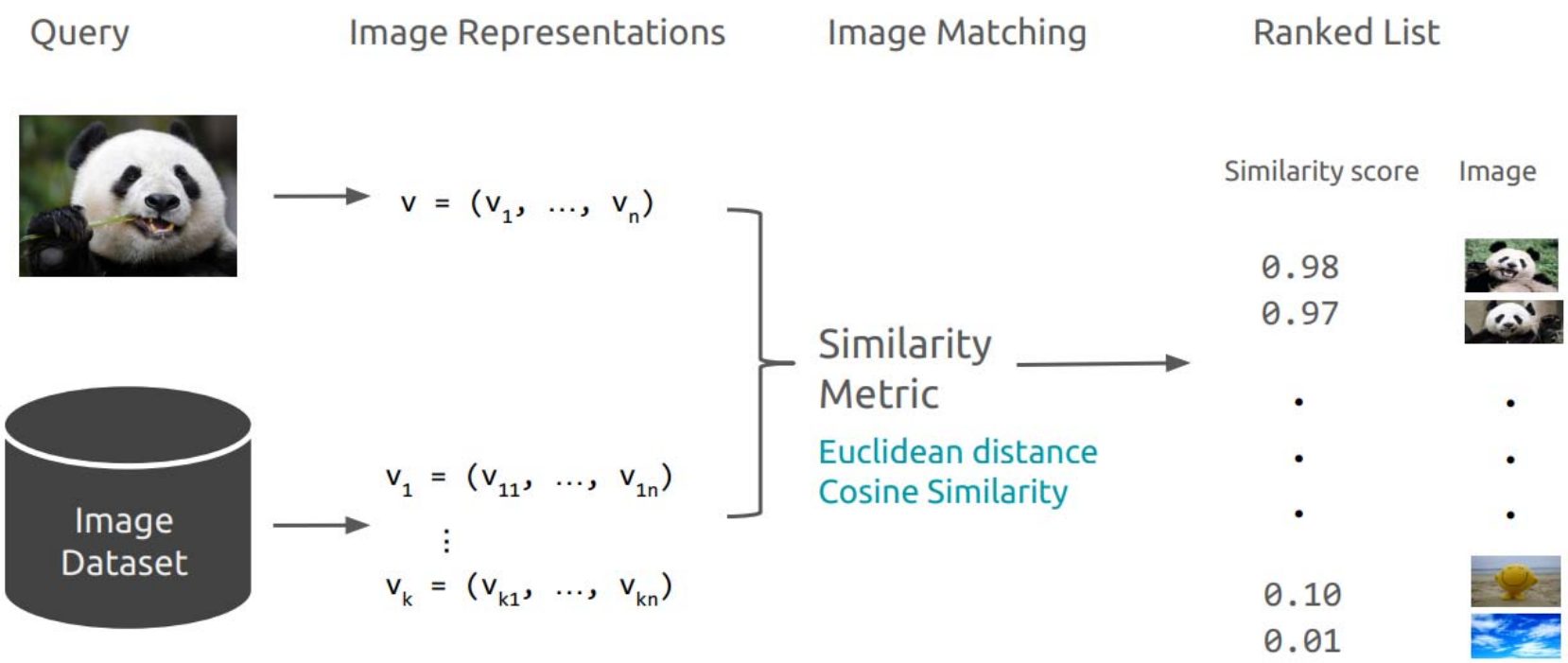
Basic Problems

- 图像检索基本问题之一：如何计算图像间的内容相关性?
  - **图像表征**：非结构化图像数据的结构化表达
    - SIFT + BoW/VLAD/FV
    - Activations of the intermediate layers of CNN model
  - **相似性度量**：基于图像表征的相关性计算



Based on Hand-Crafted Features

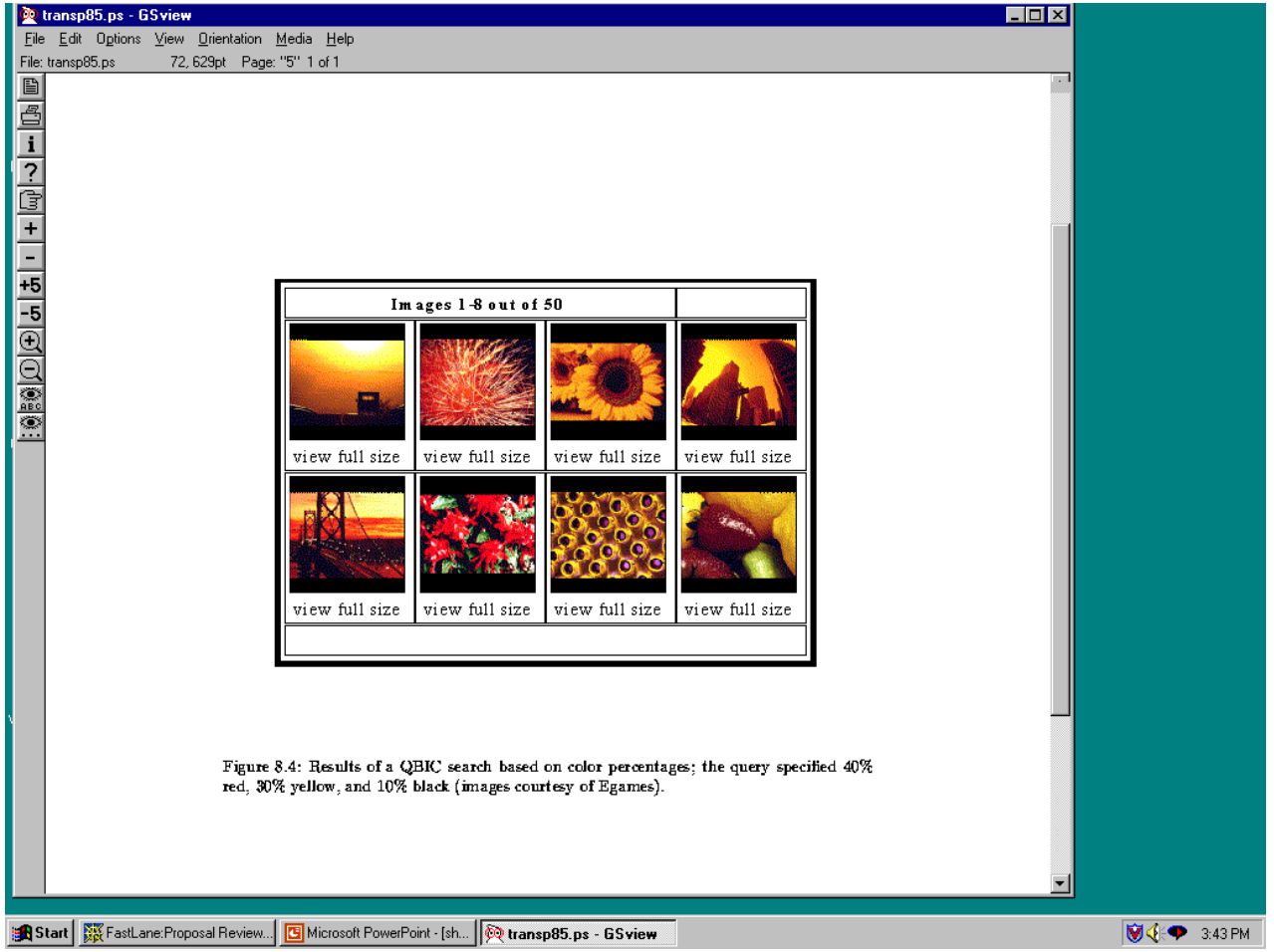
The retrieval pipeline



Global Features

- Color (histograms, wavelets)
- Texture (灰度共生矩阵, Gabor filters, LBP（局部二值模式）)
- Shape (Boundary Matching)
- Objects and their Relationships

Color Histograms



Histogram Similarity

The QBIC color histogram distance is:

$$d_{hist}(I,Q) = (h(I) - h(Q))^T \mathbf{A} (h(I) - h(Q))$$

- h(I) is a K-bin histogram of a database image
- h(Q) is a K-bin histogram of the query image
- A is a K x K similarity matrix

颜色直方图

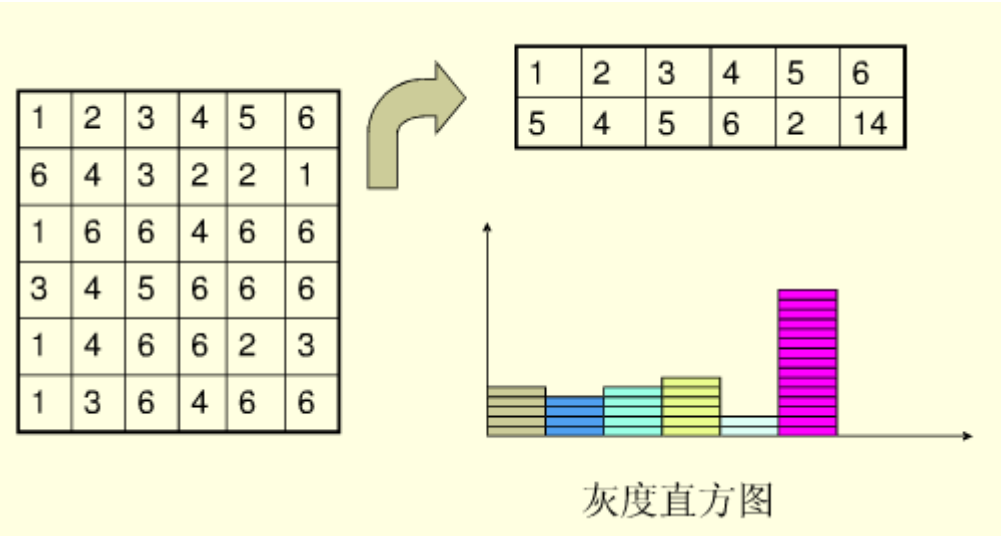
- 一个灰度级在范围[0，L-1]的数字图像的直方图是一个离散函数

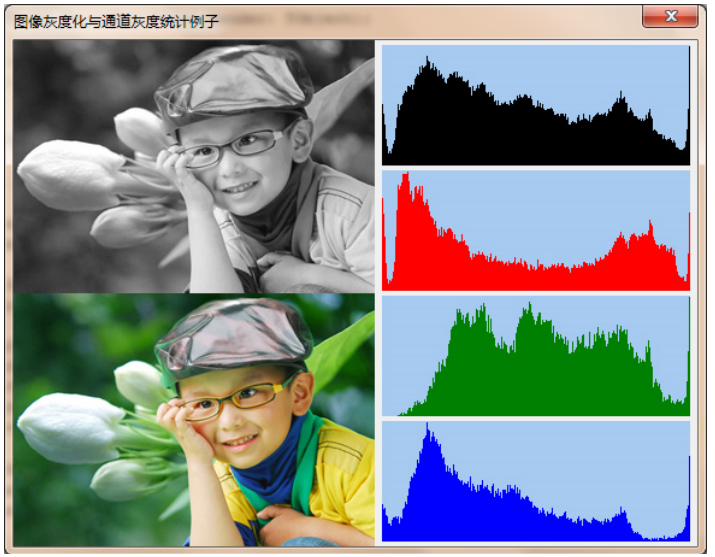
$$p(r_k)= n_k /n$$

n 是图像的像素总数

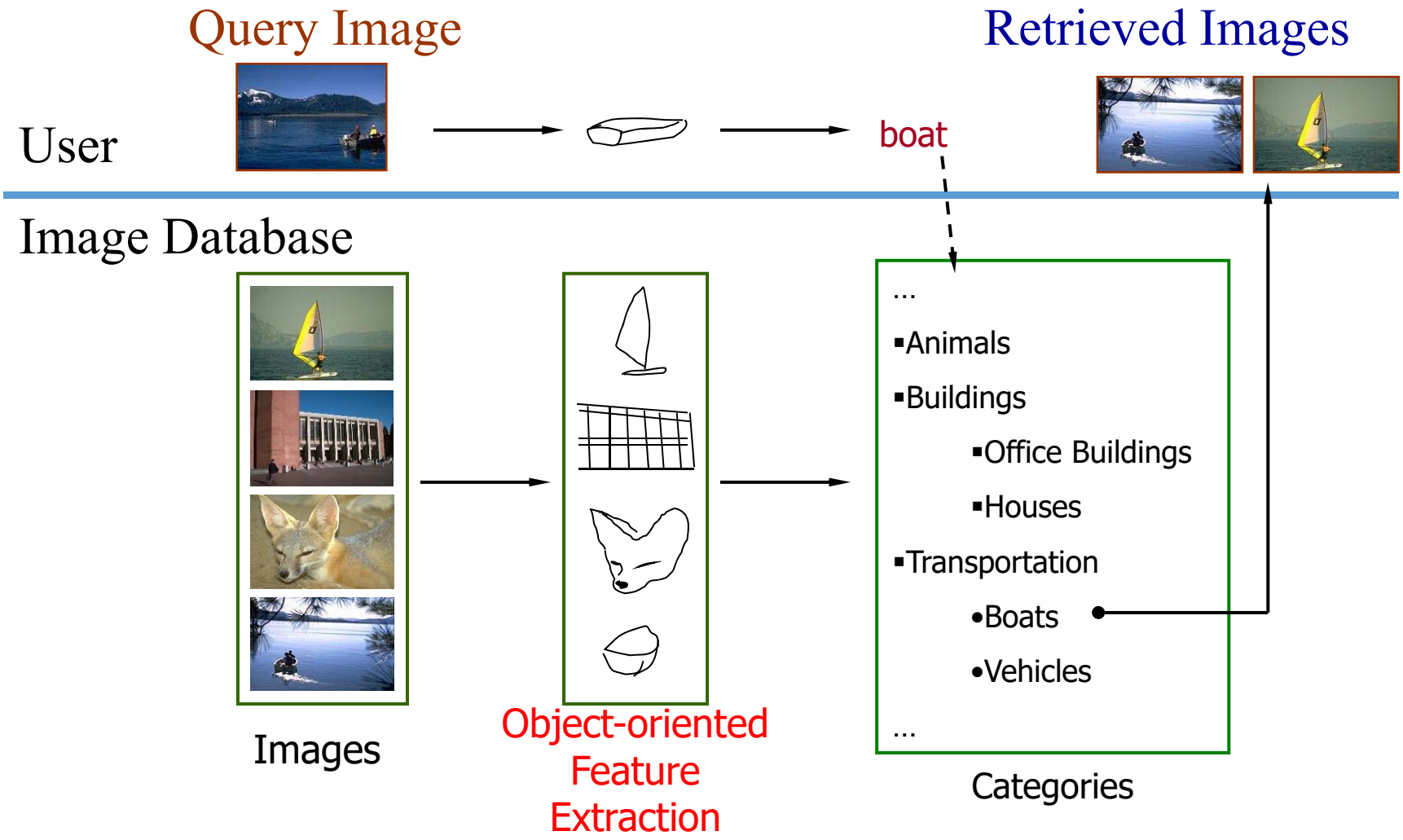
n_k是图像中灰度级为r_k的像素个数

r_k 是第k个灰度级，$k = 0,1,2,\ldots,L$-1





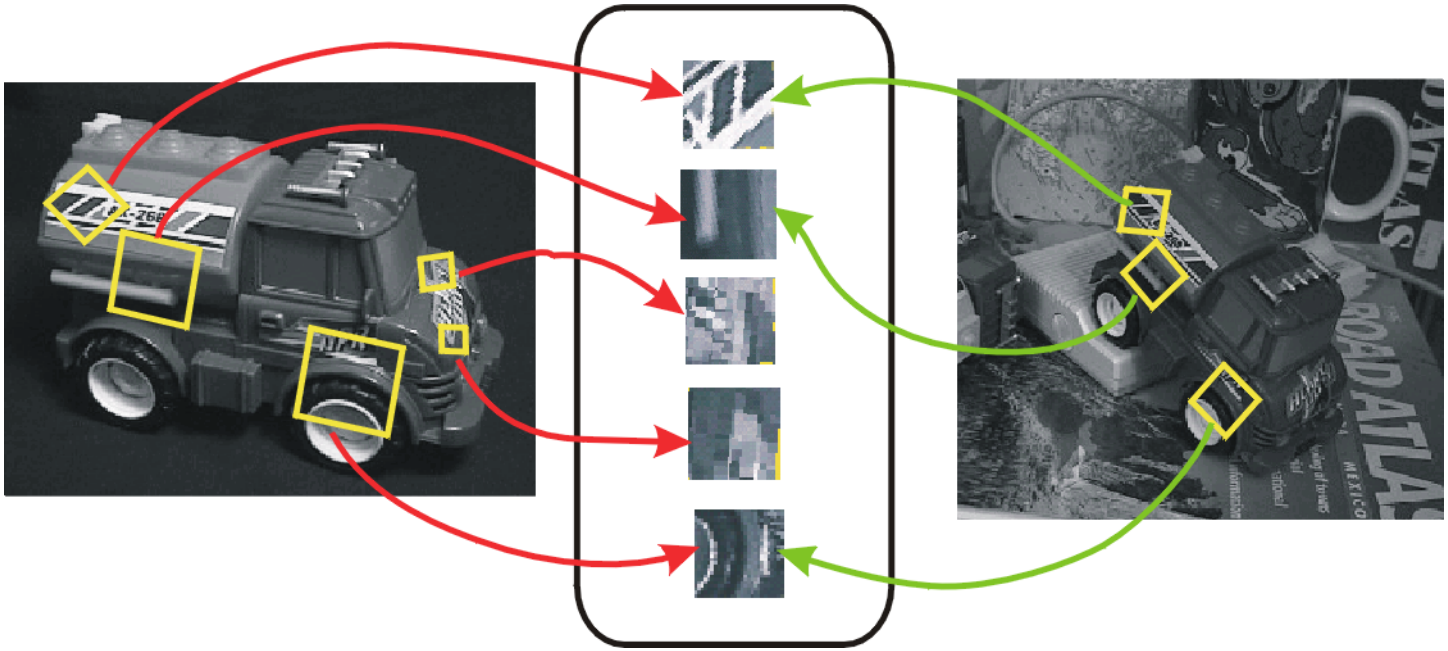
Object-based



Local Features

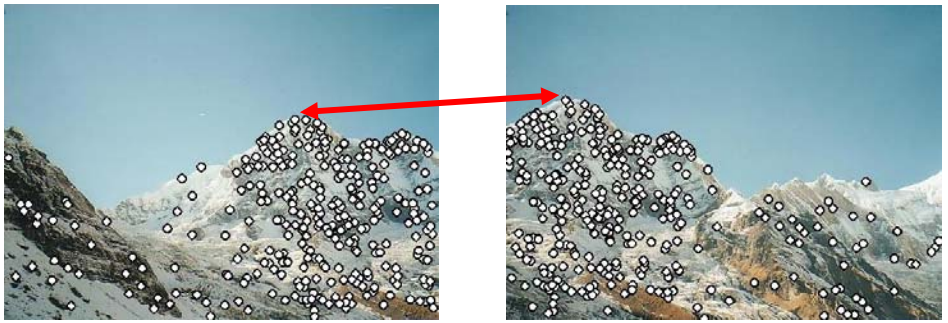
Find features that are invariant to transformations

- geometric invariance: translation, rotation, scale
- photometric invariance: brightness, exposure, …



Local Features

- SIFT
- LBP
- SURF
- BRISK
- And so on



Local feature extraction in Matlab
https://www.mathworks.com/help/vision/ug/local-feature-detection-and-extraction.html

Local Features

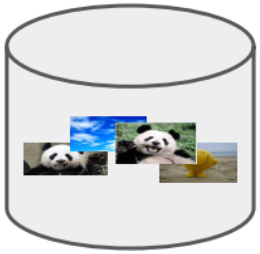
The classic SIFT retrieval pipeline

variable number of
feature vectors per image

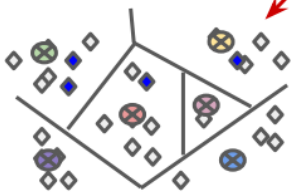
$v_1 = (v_{11}, \ldots, v_{1n})$

$\vdots$

$v_k = (v_{k1}, \ldots, v_{kn})$



N-Dimensional
feature space



Bag of Visual
Words

M visual words
(M clusters)

INVERTED FILE

word	Image ID
1	1, 12,
2	1, 30, 102
3	10, 12
4	2,3
6	10

Large vocabularies (50k-1M)
Very fast!
Typically used with SIFT features

Some important terms

- Global Features: e.g. 颜色直方图
- Local Features: e.g. sift
- BoW[1]: Bag of (Visual) Words
- VLAD[2]: Vector of Aggregate Locally Descriptor
- FV[3]: Fisher Vector
- Invert Index

[1] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in Proc. Int. Conf. Comput. Vis., 2003, Art. no. 1470.

[2] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2010, pp. 3304–3311.

[3] F. Perronnin, J. S anchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in Proc. Eur. Conf. Comput. Vis., 2010, pp. 143–156.

Image Retrieval & Image Classification

Query: This chair

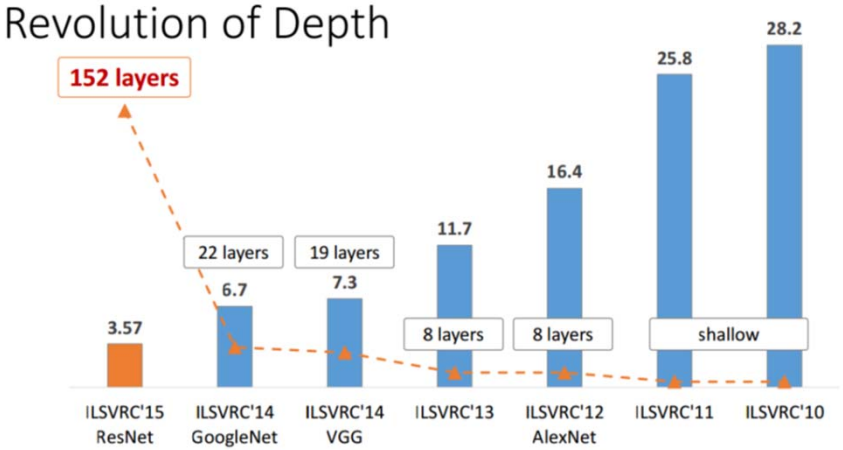
Results from dataset classified as “chair”

Query: This chair

Results from dataset ranked by similarity to the query

Based on Deep Learning

- 深度学习在计算机视觉领域取得巨大成功
  - ImageNet Grand Challenge



- 深度学习：时势造英雄
  - 大规模图像视频数据
  - 强大的计算能力：GPU/TPU

IMAGENET

flickr

60亿张图片，2011年

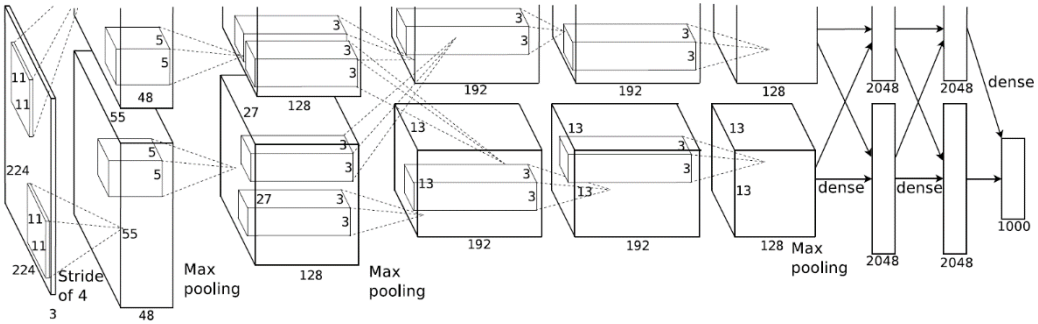
facebook

2500亿张图片，2013年

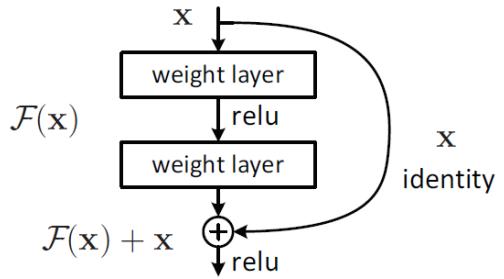
NVIDIA

研究背景：深度学习

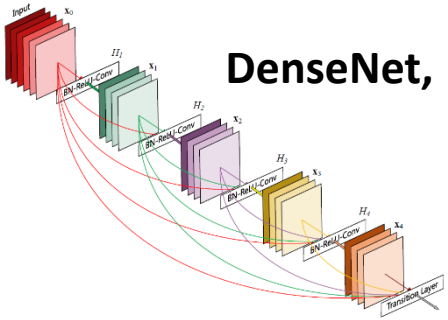
- 面向图像分类的深度学习模型



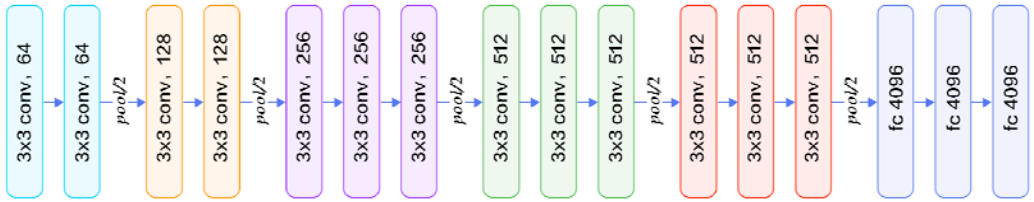
AlexNet, 2012



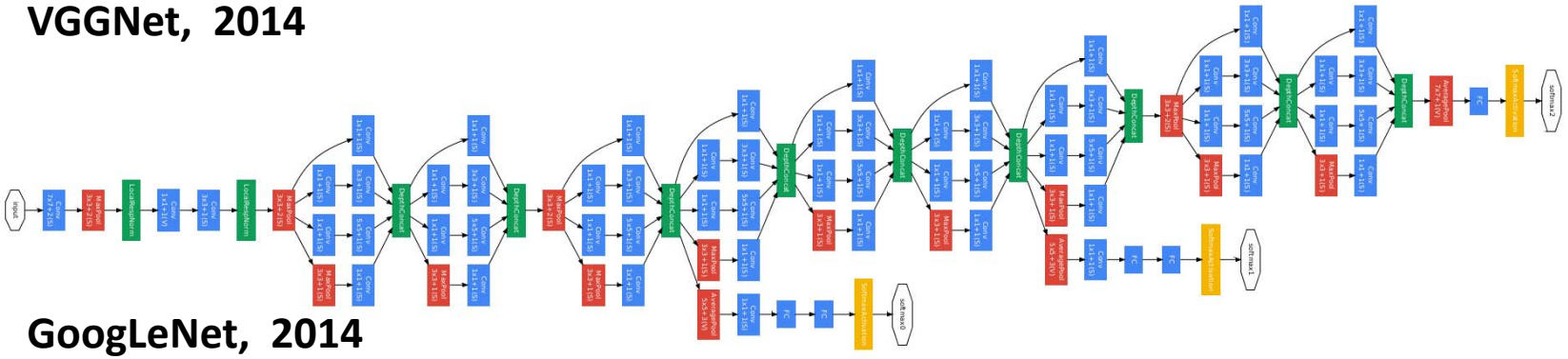
ResNet, 2015



DenseNet, 2017



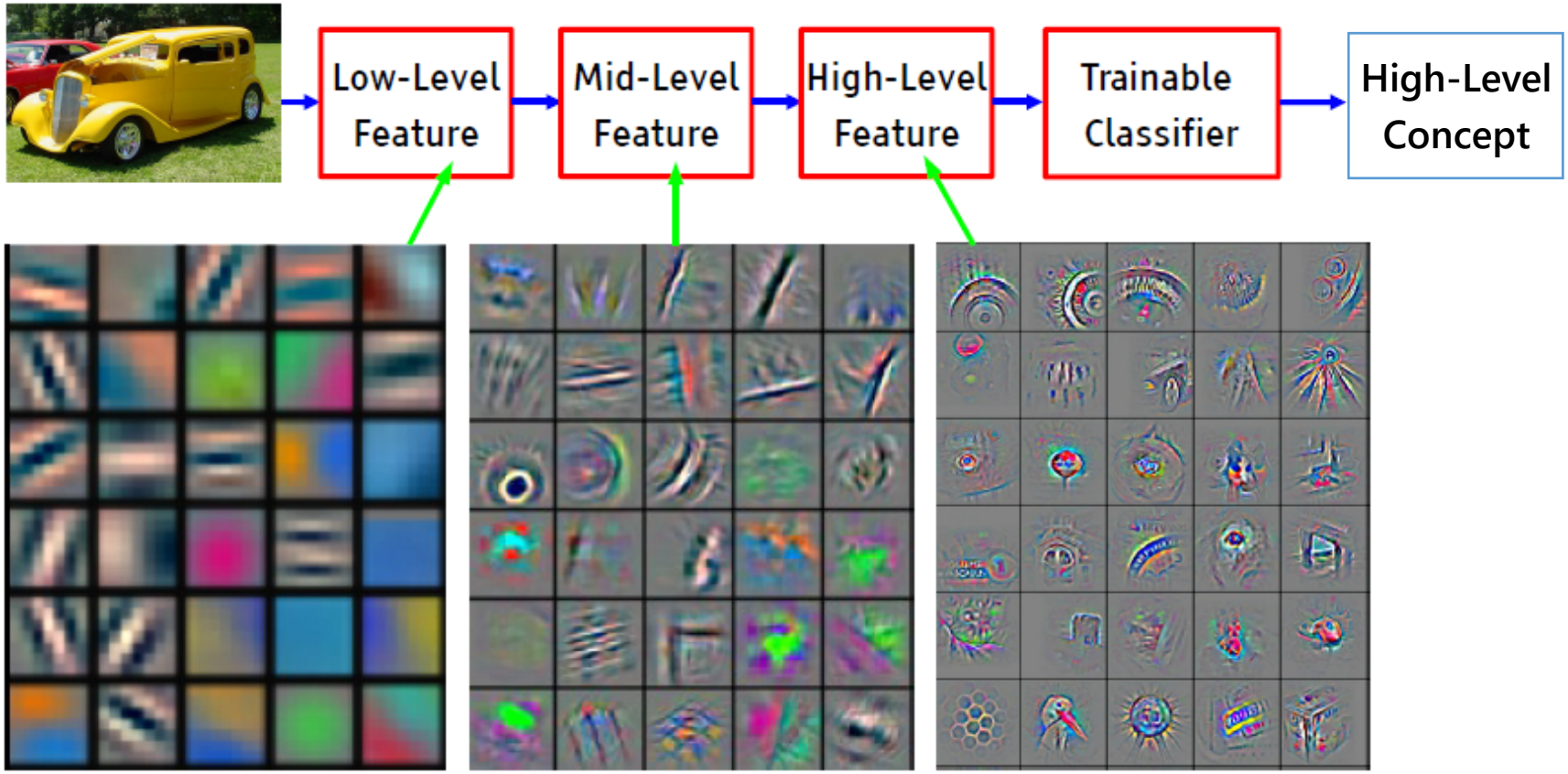
VGGNet, 2014



GoogLeNet, 2014

研究背景：深度学习

- 深度学习本质：层次化的表征学习

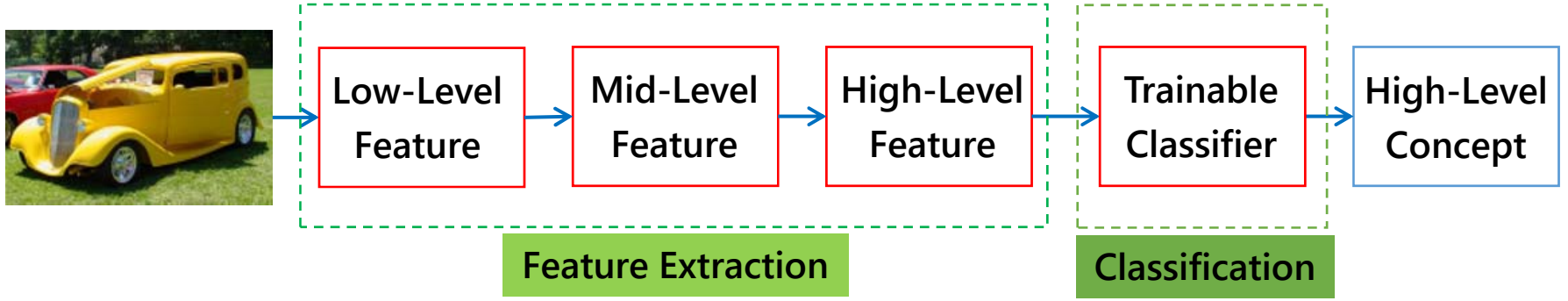


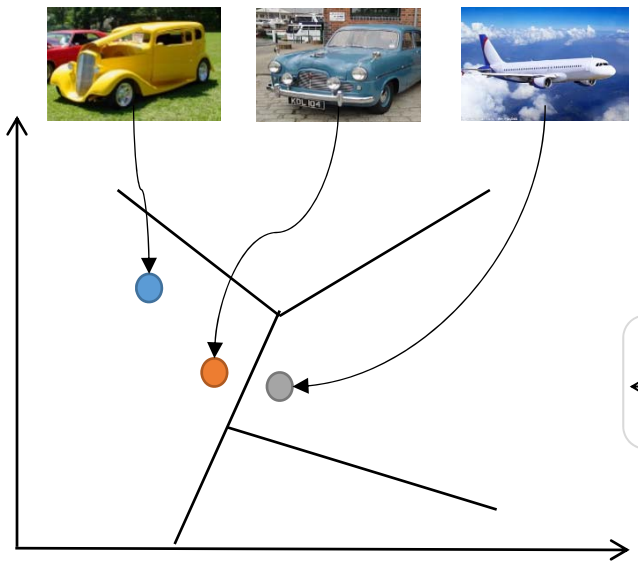
[Courtesy of Yann Le Cun]

Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

研究背景：深度学习

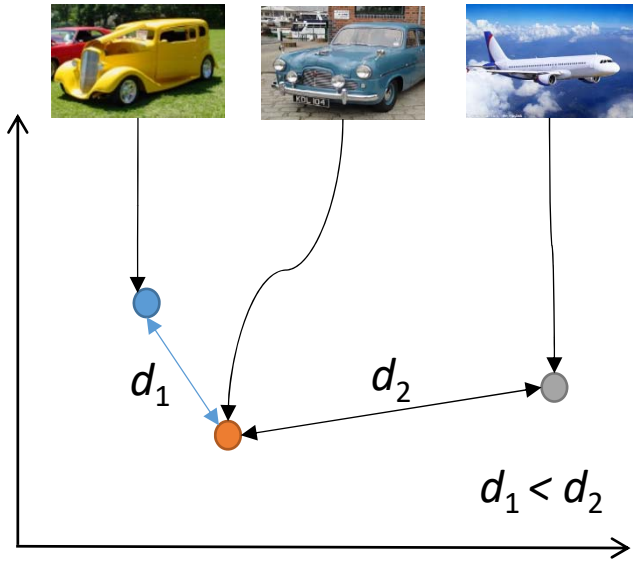
- 图像分类 vs. 图像检索





已知 图像类别 未定义

图像分类：特征空间划分



图像检索：相关度度量

如何将深度学习用于图像检索?

- 关键：大规模标注的训练数据

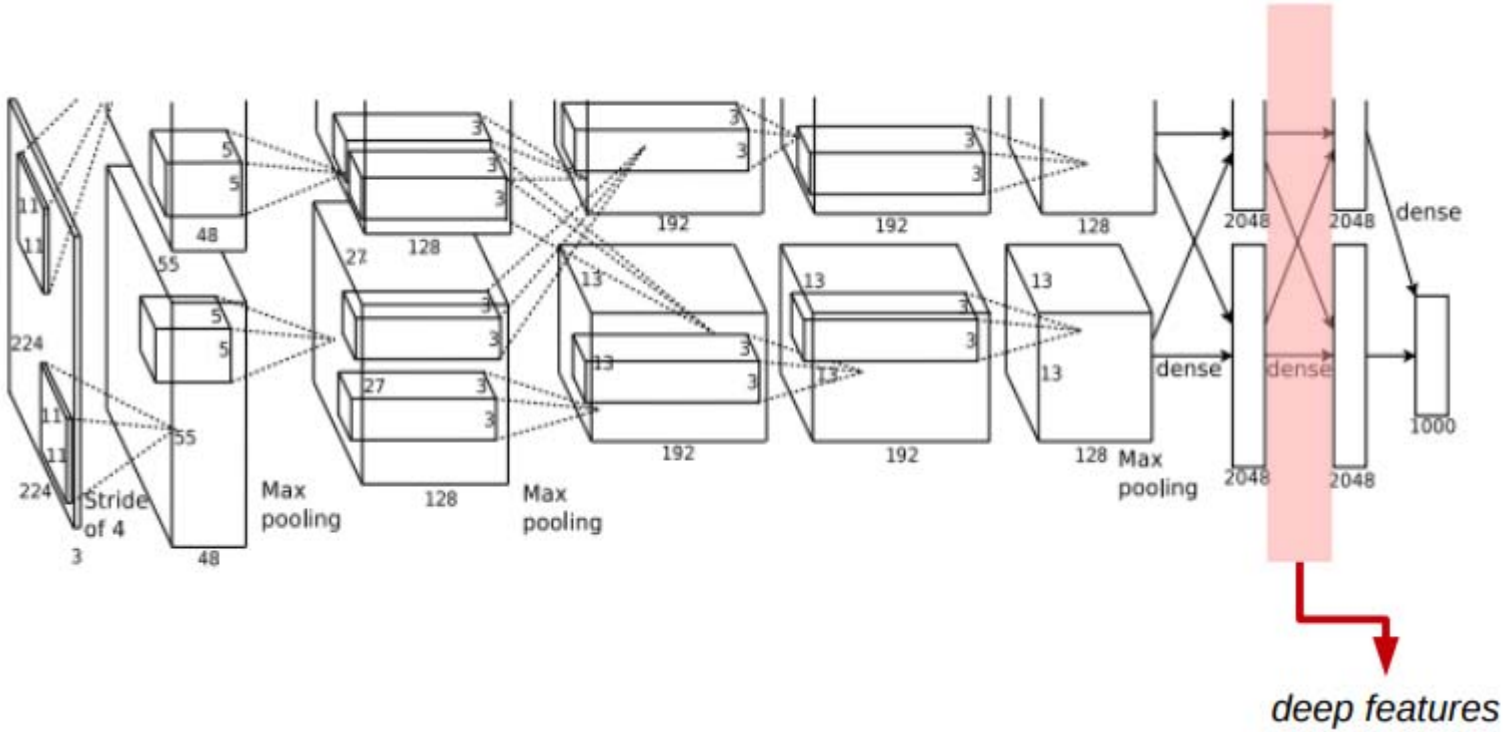
Low-Level Feature → Mid-Level Feature → High-Level Feature → Trainable Classifier → High-Level Concept

Feature Extraction

Classification

Option 1:
- 利用图像分类任务进行预训练
- 基于上述网络提取特征，用于检索

Option 2:
- 面向检索任务，设计监督信息
- fine-tune预训练好的CNN模型



https://www.mathworks.com/help/deeplearning/ug/deep-learning-in-matlab.html

https://www.mathworks.com/help/deeplearning/ref/vgg16.html

https://www.mathworks.com/help/deeplearning/ug/extract-image-features-using-pretrained-network.html

Dataset

name	# images	# queries	content
Holidays [13]	1,491	500	scene
Ukbench [11]	10,200	10,200	common objects
Paris6k [25]	6,412	55	buildings
Oxford5k [12]	5,062	55	buildings
Flickr100k [25]	99,782	-	from Flickr's popular tags

Holidays: http://lear.inrialpes.fr/people/jegou/data.php

Ukbench: http://www.vis.uky.edu/~stewe/ukbench/

Paris6k: http://www.robots.ox.ac.uk/~vgg/data/parisbuildings/

Oxford5k: http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/

Flickr100k: http://www.flickr.com/

Dataset

UKBench dataset It contains 10,200 images from 2,550 categories[9]. In each category, there are four images taken on the same scene or object from different views or illumination conditions. All the 10,200 images are taken as query and their retrieval performances are averaged.

Holidays dataset There are 1,491 images from 500 groups in the Holidays dataset[10]. Images in each group are taken on a scene or an object with various viewpoints. The first image in each group is selected as query for evaluation.

Oxford Building dataset (Oxford-5K) The Oxford Buildings Dataset consists of 5062 images collected from Flickr by searching for particular Oxford landmarks. The collection has been manually annotated to generate a comprehensive ground truth for 11 different landmarks, each represented by 5 possible queries. This gives a set of 55 queries over which an object retrieval system can be evaluated. Some junk images are mixed in it as distractor.

Paris dataset In the Paris dataset13, there are 6,412 images, which are collected from Flickr by searching for 12 text queries of particular Paris landmarks. For this dataset, 500 query images are used for evaluation

常用的评价标准：AveP，MAP

- AveP强调排列在前的图像更重要

$$AveP = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{number\ of\ relevant\ documents}$$

其中$rel(k)$表示第k个文档是否相关，若相关则为1，否则为0，$P(k)$表示前k个文档的准确率。 $AveP$的计算方式可以简单的认为是：

$$AveP = \frac{1}{R} \times \sum_{r=1}^{R} \frac{r}{position(r)}$$

其中R表示相关文档的总个数，$position(r)$表示，结果列表从前往后看，第r个相关文档在列表中的位置。比如，有三个相关文档，位置分别为1、3、6，那么$AveP = \frac{1}{3} \times (\frac{1}{1} + \frac{2}{3} + \frac{3}{6})$。在编程的时候需要注意，位置和第i个相关文档，都是从1开始的，不是从0开始的。

- MAP就是AveP按照query的数量进行平均，得到平均结果

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}$$

思考题

- 假设两个查询图，查询图1有4张相关图，查询图2有5张相关图。某系统对于主题1的4张相关图分别的排位是：1，2，4，7。对于查询图2，只检索出3张图，相关排位是1，3，5。请计算MAP。

答案

- 查询图1：平均准确率为
  (1/1+2/2+3/4+4/7)/4 = 0.83

- 查询图2：平均准确率为
  (1/1+2/3+3/5+0+0)/5 = 0.45

  MAP=(0.83+0.45)/2=0.64