5月21日12:00前，提交文献阅读相关素材
6月3日12:00前，提交实验报告及相关素材

信息检索与数据挖掘

图像分类的算法思想

5月15日，第13章 多媒体信息检索
5月20日，复习
5月22日，同学们文献阅读报告
5月27日，同学们文献阅读报告
6月3日，期末考试【暂定】

课程内容

- 第1章 绪论
- 第2章 布尔检索及倒排索引
- 第3章 词项词典和倒排记录表
- 第4章 索引构建和索引压缩
- 第5章 向量模型及检索系统
- 第6章 检索的评价
- 第7章 相关反馈和查询扩展
- 第8章 概率模型
- 第9章 基于语言建模的检索模型
- 第10章 文本分类
- 第11章 文本聚类
- 补充：概率图及主题模型
- 补充：数据挖掘经典算法概述
- 第12章 Web搜索
- 第13章 多媒体信息检索
- ~~第14章 其他应用简介~~

引言：多媒体检索示例
微软识花：精细物体识别是怎么做到的

微软亚洲研究院
2016年9月28日

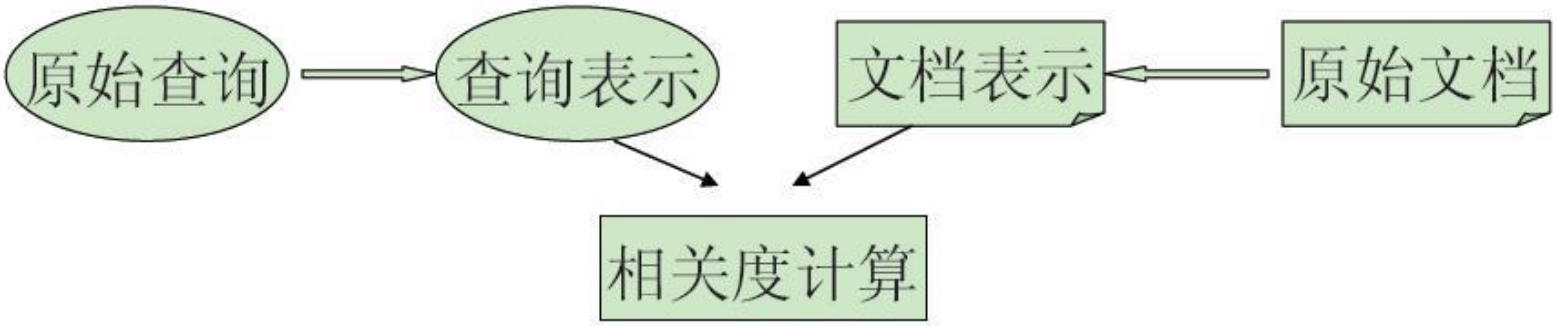
“微软识花 “app的开发是**微软亚洲研究院**和**中国科学院植物研究所**多年来学术合作的成果。

中科院植物所不仅提供了260万张花卉的识别图片，还提供了经过专家鉴定的中国常见花列表。而微软亚洲研究院的研究员们利用先进的技术开发出识别花卉的算法，并把识别结果挑选出来，经植物所专家鉴定。经过了两三次迭代的过程，才得到了最终训练机器识别的样本集合。

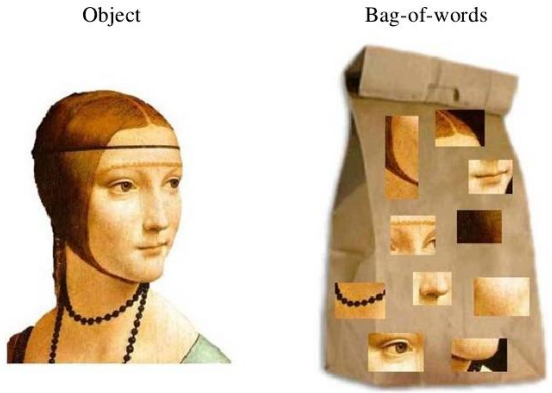
http://www.msra.cn/zh-cn/news/features/flower-recognition-20160928.aspx

引言：从文本检索→图像检索

- Term→Feature: 图像的表示？



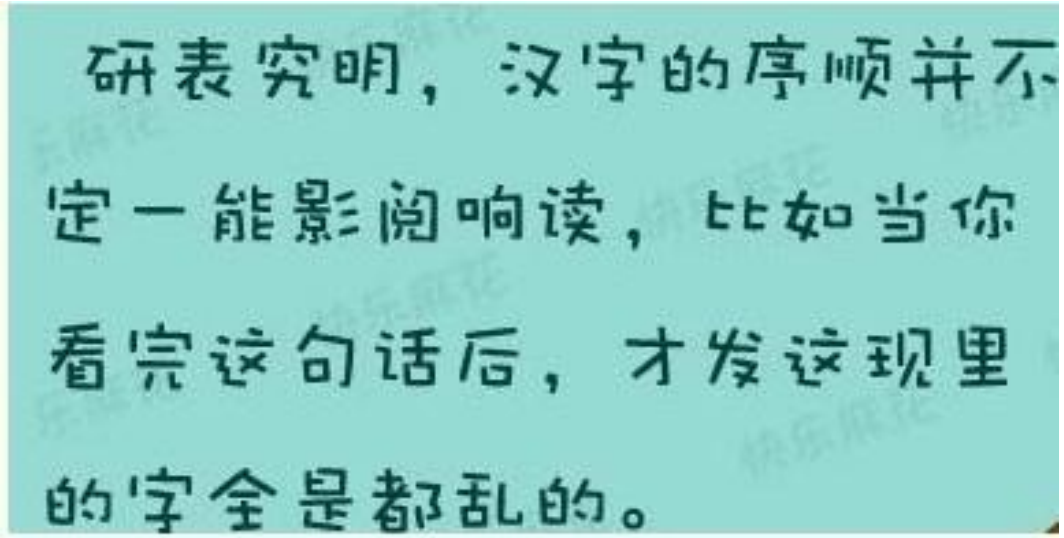
- Bag of Words → Bag of Features



词袋(Bag of words)模型

- 不考虑词在文档中出现的顺序
- John is quicker than Mary 及 Mary is quicker than John are 的表示结果一样
- 在某种意思上说，这种表示方法是一种“倒退”，因为位置索引中能够区分上述两篇文档

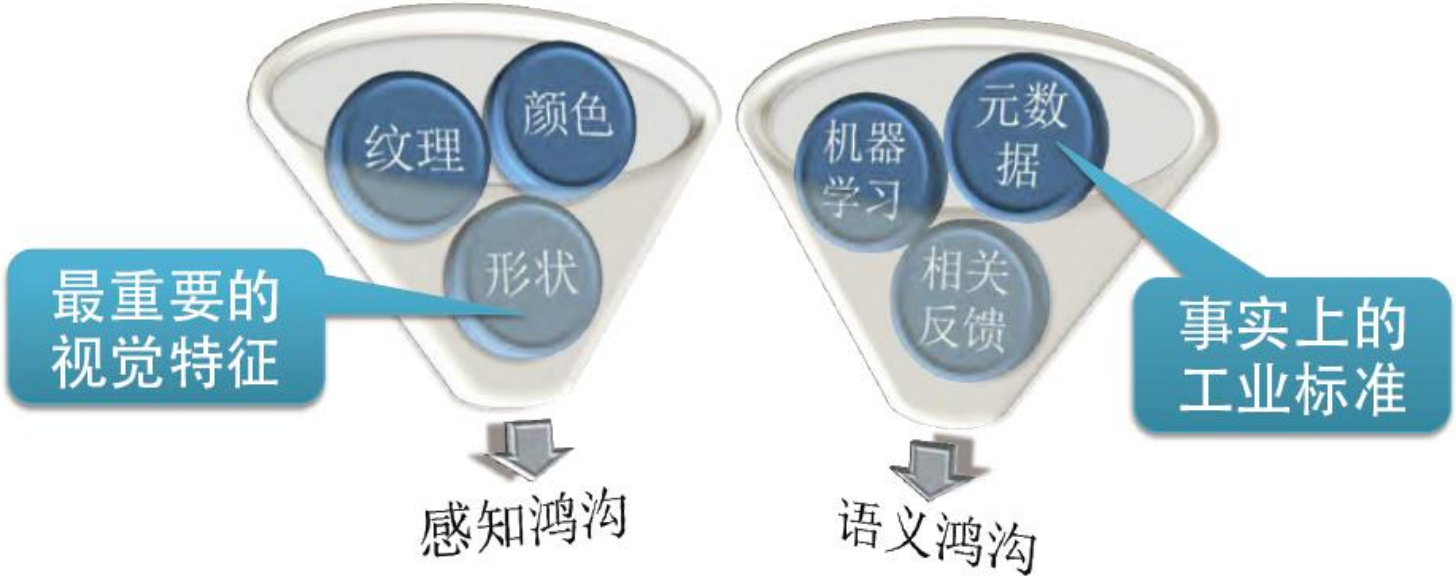
然而，词袋模型是有效的！



引言：

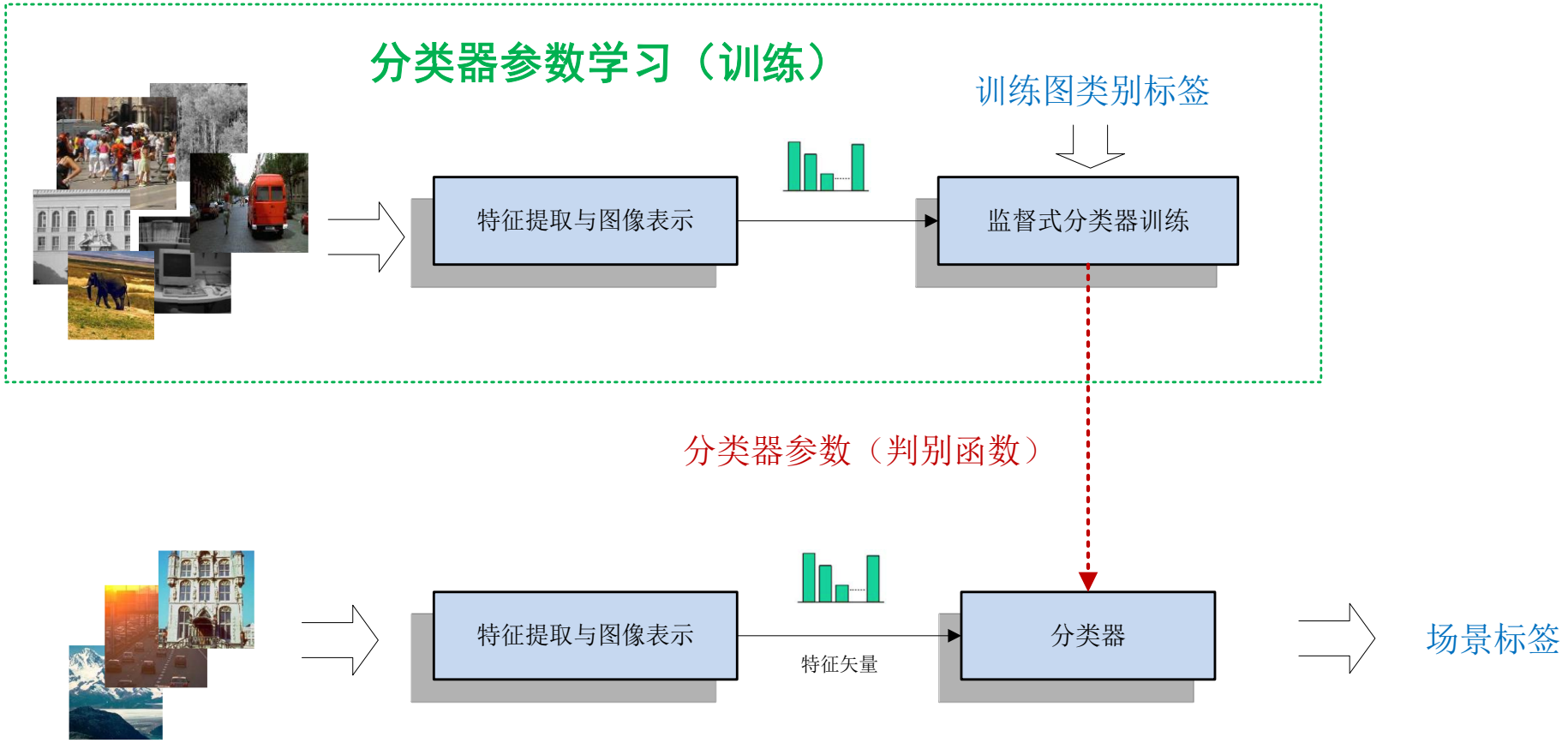
感知鸿沟(Sensory Gap) /语义鸿沟(Sematic Gap)

(1) 感知鸿沟是指真实世界的物体和从该物体场景对应的图像中提取的描述信息之间的鸿沟；
(2) 语义鸿沟是指人们从视觉数据中所能提取到的信息和某个用户在特定情况下对相同数据的描述缺乏一致性。



Saliency and Similarity Detection for Image Scene Analysis
清华大学博士论文《图像内容的显著性与相似性研究》，程明明，2012
http://mmcheng.net/

引言：
传统图像分类与识别系统基本结构



图像分类的算法思想

- 从文本分类→图像分类
  - 如何从图像中获取全局特征？
    - 颜色特征、纹理特征、形状特征
  - 如何从图像中获取局部特征？
    - SIFT: Scale-invariant feature transform
- 图像分类的几个发展阶段
  - Low-level Modelling
  - Semantic Modelling
  - Sparse Coding
  - Deep Learning

传统图像分类方法包含两部分工作：特征提取，分类器设计。关于分类器在课程前面章节已有充分的讨论。

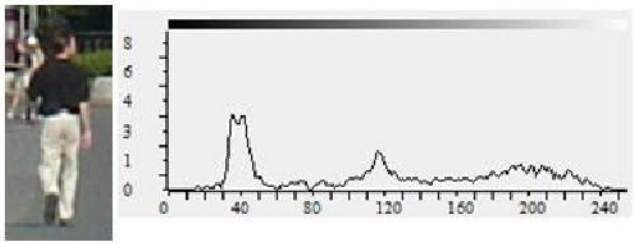
图像分类的算法思想

- 从文本分类→图像分类
  - 如何从图像中获取全局特征？
    - 颜色特征、纹理特征、形状特征
  - 如何从图像中获取局部特征？
    - SIFT: Scale-invariant feature transform
- 图像分类的几个发展阶段
  - Low-level Modelling
  - Semantic Modelling
  - Sparse Coding
  - Deep Learning

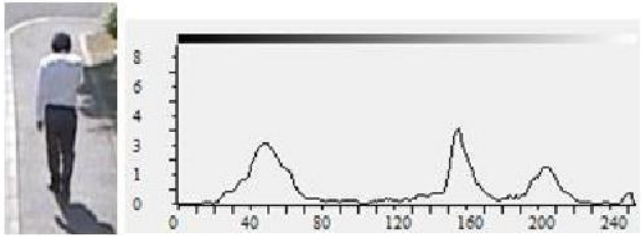
颜色(Color)特征

分块直方图

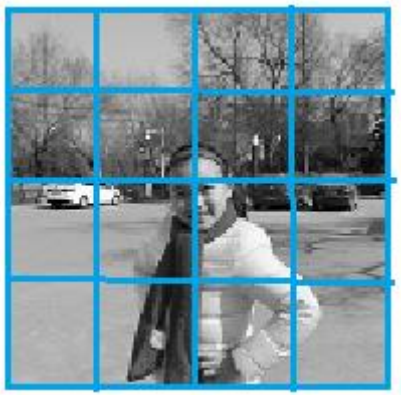
将图像区域做了进一步分割，原来的单一直方图变为了多个小区域直方图的联合

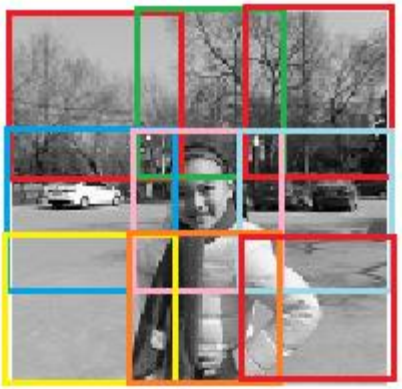


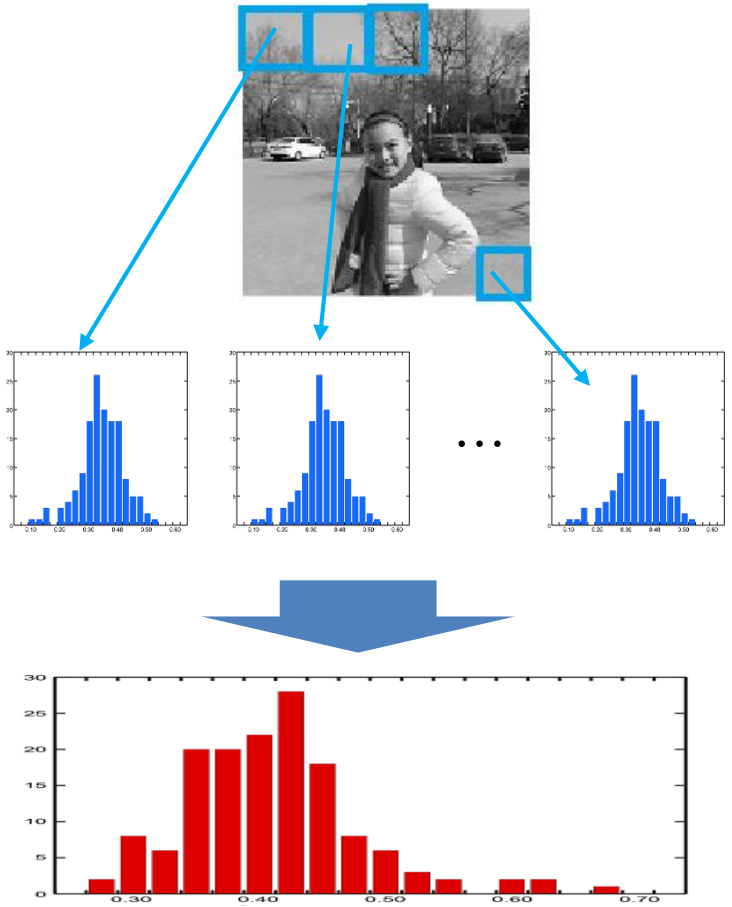
内容不同的图可能具有相似的直方图



将图像分成不重叠（或重叠）的子区域，对每一个子区域单独提取颜色直方图，然后将各子区域的短直方图按固定顺序连接成一个长直方图作为目标特征。

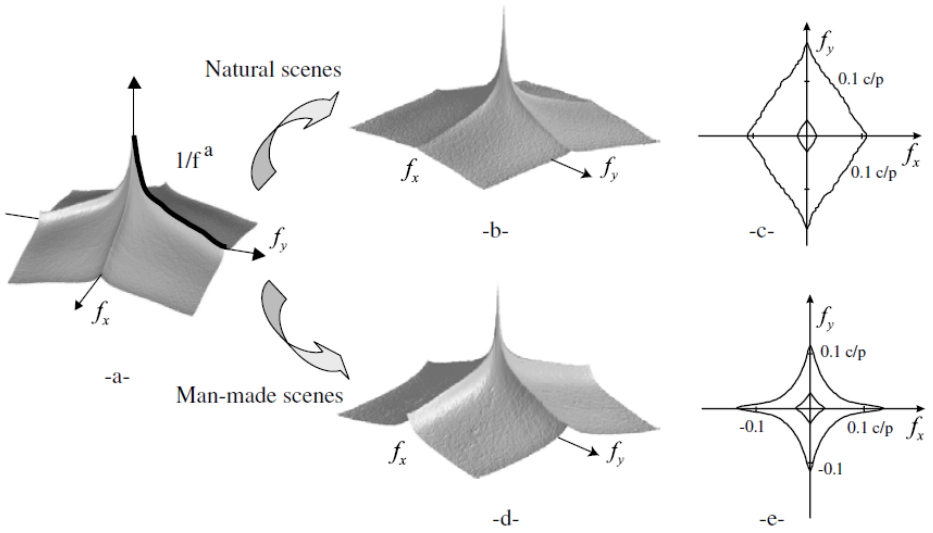






颜色(Color)特征

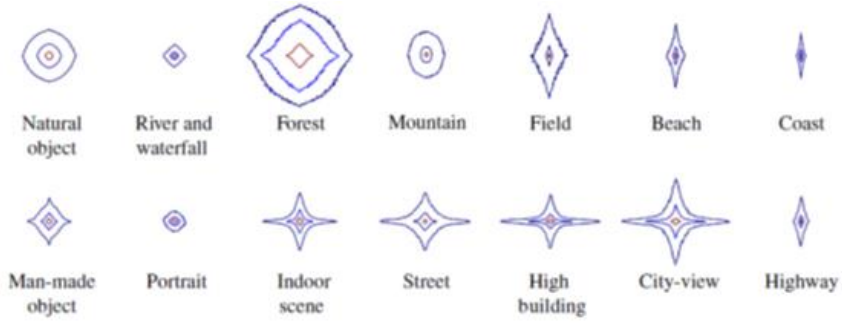
变换域：Oliva和Torralba所提gist图像全局特征



$$E[A(f,\theta)^2 \mid S] \simeq \Gamma_s(\theta)/f^{-\alpha_s(\theta)}$$

极坐标形式的傅立叶变换：不同类别图像显示出明显的差异性

对输入图像进行预滤波，采用离散傅立叶变换和加窗傅立叶变换来提取输入图像的全局特征信息。



Oliva, A. and A. Torralba (2001). "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope." International Journal of Computer Vision **42**(3).
2015.05 Google cited: 3020

颜色(Color)特征

探索新的图像特征：暗原色

暗原色先验（dark channel prior）

汤晓鸥（859校友）1990年毕业于中国科大精密机械与精密仪器系并获学士学位; 1991年于罗切斯特大学获得硕士学位；1996年获得麻省理工学院博士学位；汤博士现任香港中文大学信息工程系教授、工程学院副院长。其主页为：http://mmlab.ie.cuhk.edu.hk/

3. Dark Channel Prior

The dark channel prior is based on the following observation on haze-free outdoor images: in most of the non-sky patches, at least one color channel has very low intensity at some pixels. In other words, the minimum intensity in such a patch should has a very low value. Formally, for an image $\mathbf{J}$, we define

$$J^{dark}(\mathbf{x}) = \min_{c\in\{r,g,b\}} \left(\min_{\mathbf{y}\in\Omega(\mathbf{x})} (J^c(\mathbf{y})) \right), \qquad (5)$$

where J^c is a color channel of $\mathbf{J}$ and $\Omega(\mathbf{x})$ is a local patch centered at $\mathbf{x}$. Our observation says that except for the sky region, the intensity of J^{dark} is low and tends to be zero, if $\mathbf{J}$ is a haze-free outdoor image. We call J^{dark} the *dark channel* of $\mathbf{J}$, and we call the above statistical observation or knowledge the *dark channel prior*.

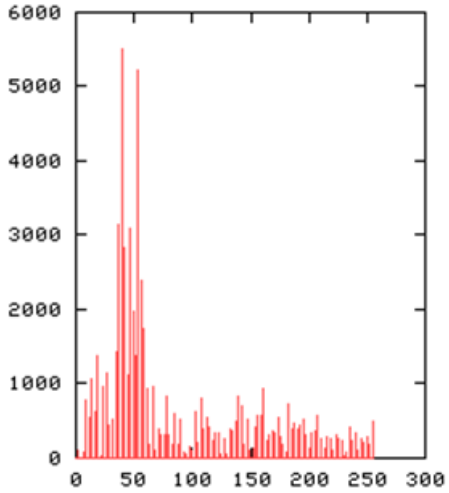
Figure 3. Top: example images in our haze-free image database. Bottom: the corresponding dark channels. Right: a haze image and its dark channel.

有雾的图片其暗原色通道的亮度大于无雾情形

K. He, J. Sun, and X. Tang, " Single Image Haze Removal Using Dark Channel Prior ," CVPR, 2009.

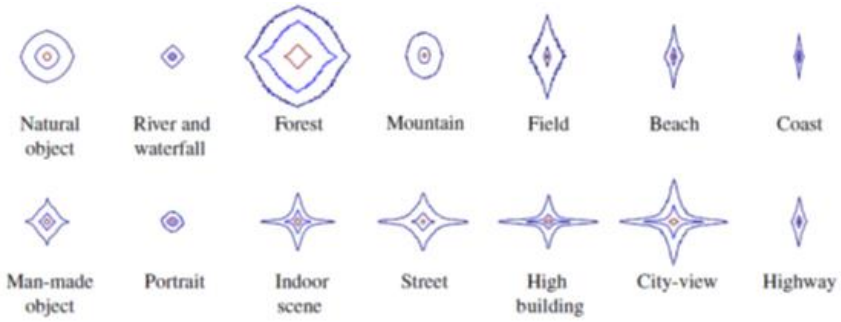
小结：颜色特征与图像分类/检索

- 直方图
  - 分块直方图
  - 梯度图像的直方图
- 直方图的比较



$$D(A,B)=1-\frac{\sum_{i=1}^{n}\min(a_i,b_i)}{\min\left(\sum_{i=1}^{n}a_i,\sum_{i=1}^{n}b_i\right)}$$

- 二维傅里叶变换幅度谱
- 暗原色



$$J^{dark}(\mathbf{x})=\min_{c\in\{r,g,b\}}\left(\min_{\mathbf{y}\in\Omega(\mathbf{x})}\left(J^{c}(\mathbf{y})\right)\right)$$

图像分类的算法思想

- 从文本分类→图像分类
  - 如何从图像中获取全局特征？
    - 颜色特征、纹理特征、形状特征
  - 如何从图像中获取局部特征？
    - SIFT: Scale-invariant feature transform
- 图像分类的几个发展阶段
  - Low-level Modelling
  - Semantic Modelling
  - Sparse Coding
  - Deep Learning

纹理特征

Canny edge detector

Canny 算子实现检测边缘的步骤如下：

（1）用高斯滤波器平滑图像。

$$\mathbf{B} = \frac{1}{159}\begin{bmatrix} 2 & 4 & 5 & 4 & 2 \\ 4 & 9 & 12 & 9 & 4 \\ 5 & 12 & 15 & 12 & 5 \\ 4 & 9 & 12 & 9 & 4 \\ 2 & 4 & 5 & 4 & 2 \end{bmatrix} * \mathbf{A}$$

（2）计算平滑后的图像的梯度幅值和方向。

$$\mathbf{G} = \sqrt{\mathbf{G}_x^{\,2} + \mathbf{G}_y^{\,2}} \qquad \mathbf{\Theta} = \operatorname{atan2}(\mathbf{G}_y, \mathbf{G}_x)$$

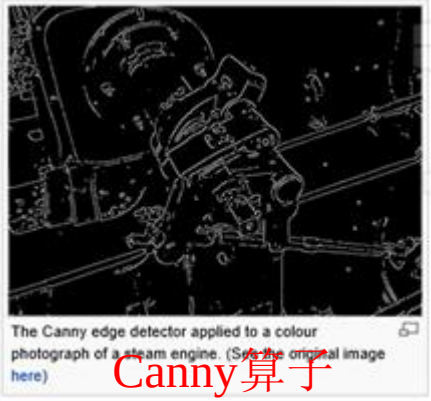
（3）对梯度幅值采用非极大值抑制，其过程为找出图像梯度中的局部极大值点，把其他非极大值点置零而得到细化的边缘。

（4）用双阈值算法检测和连接边缘【高阈值和低阈值是人为来确定的】。


A color picture of a steam engine


The Sobel operator applied to that image

Sobel算子


The Canny edge detector applied to a colour photograph of a steam engine. (See the original image here)

Canny算子

纹理特征

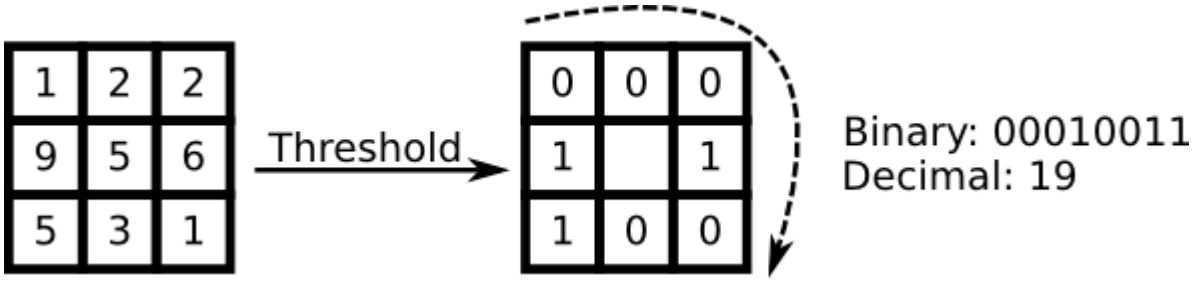
LBP, Local Binary Pattern

局部二值模式

- LBP是一种用来描述图像局部纹理特征的算子；它具有旋转不变性和灰度不变性等显著的优点。

$$LBP(x_c, y_c) = \sum_{p=0}^{P-1} 2^p s(i_p - i_c)$$

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases}$$

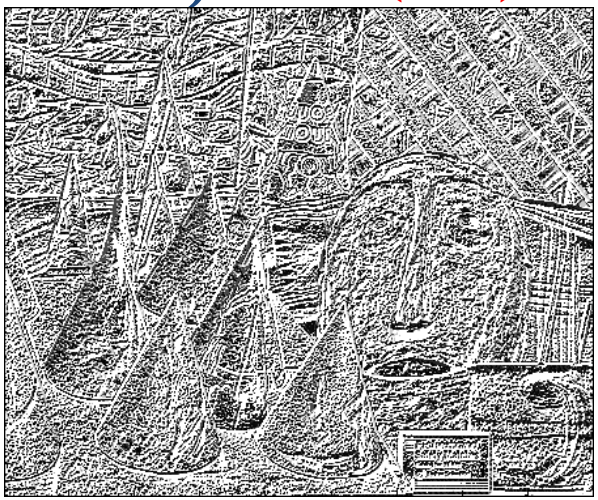


原始的LBP算子[1994年]定义为在3*3的窗口内，以窗口中心像素为阈值，将相邻的8个像素的灰度值与其进行比较，若周围像素值大于中心像素值，则该像素点的位置被标记为1，否则为0。这样，3*3邻域内的8个点经比较可产生8位二进制数（通常转换为十进制数即LBP码，共256种），即得到该窗口中心像素点的LBP值，并用这个值来反映该区域的纹理信息。

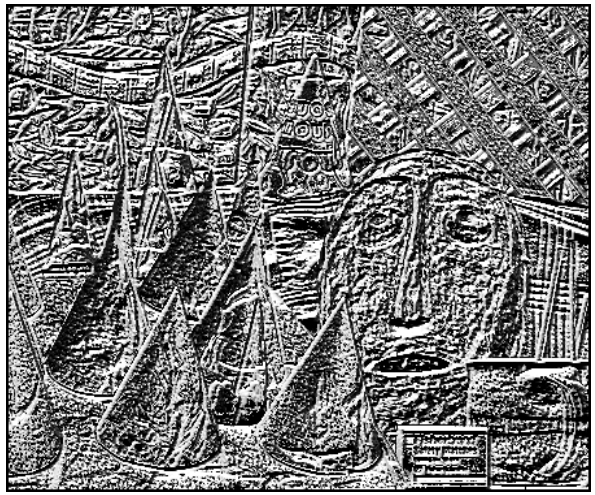
纹理特征

CT (Census Transform)

Census Transform (3×3)



Census Transform (5×5)



CT

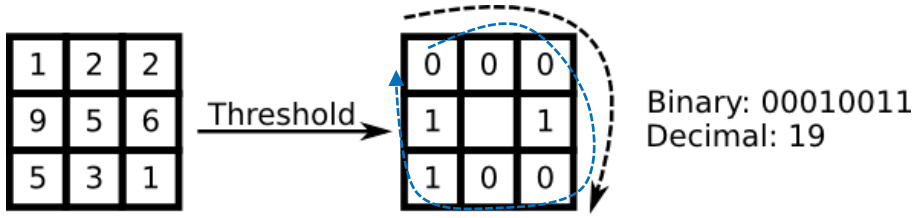
32	64	96
32	64	96
32	32	96

$\Rightarrow$

1	1	0
1	-	0
1	1	0

$\Rightarrow (11010110)_2 \Rightarrow CT = 214$

LBP

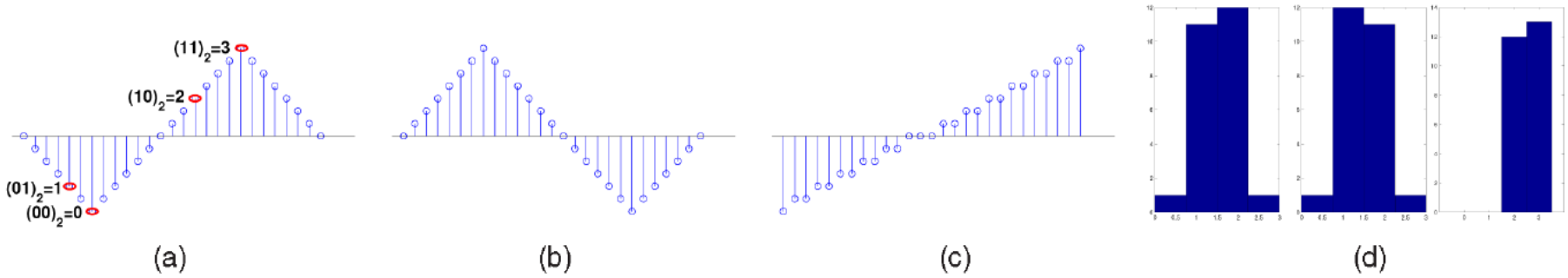


CT 比较一个像素的灰度值与它的 8-邻域的灰度值，如果一个像素的灰度值大于等于它的 8-邻域某个像素的灰度值，则在对应像素的位置赋值为比特 “1”，否则 “0”。然后从左到右、从上到下收集这 8 个比特并转换为一个值在 [0,255] 之间的整数，此整数值就是原中心像素的 CT 值。相比其它基于灰度值比较的非参局部变换而言，CT 对光照变换、γ 变换等不太敏感。

R. Zabih and J. Woodfill, “Non-parametric Local Transforms for Computing Visual Correspondence”, ECCV 1994
2015.05 Google cited :1081

纹理特征

CENTRIST (CENsus TRansform hISTogram)



CENTRIST 比灰度直方图具有更强的辨识能力。例如，在一维情况下，CT 的取值只有 $(00)_2$, $(01)_2$, $(10)_2$, $(11)_2$共 4 种可能，所以图 (a)–(c) 的灰度直方图是相同的，但是它们的 CENTRIST 却是不同的，如图 (d) 所示。

CENTRIST: A Visual Descriptor for Scene Categorization
Jianxin Wu and James M. Rehg 2015.05 Google cited: 278
IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(8), 2011: pp. 1489-1501.

mCENTRIST: A Multi-channel Feature Generation Mechanism for Scene Categorization
Yang Xiao, Jianxin Wu and Junsong Yuan
IEEE Transactions on Image Processing, 23(2), 2014: pp. 823-836.

http://cs.nju.edu.cn/wujx/projects/mCENTRIST/mCENTRIST.html

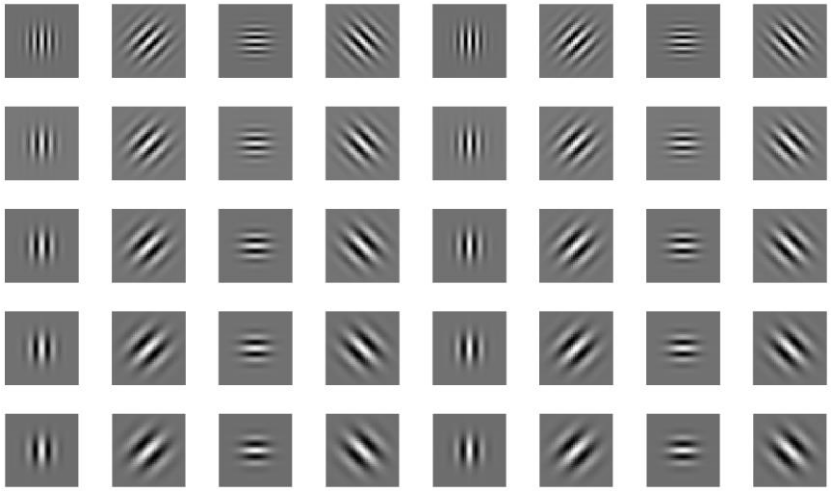
纹理特征

变换域：二维 Gabor 滤波器

为了刻画局部时间轴范围内的频谱特征，1946 年，Dennis Gabor 定义了窗口傅里叶变换（短时傅里叶变换，即 Gabor 变换），对 f (t)的 Gabor变换为：

$$G_f(\omega,\tau)=\int_{-\infty}^{+\infty} e^{j\omega t} f(t) g_a(t-\tau)dt$$

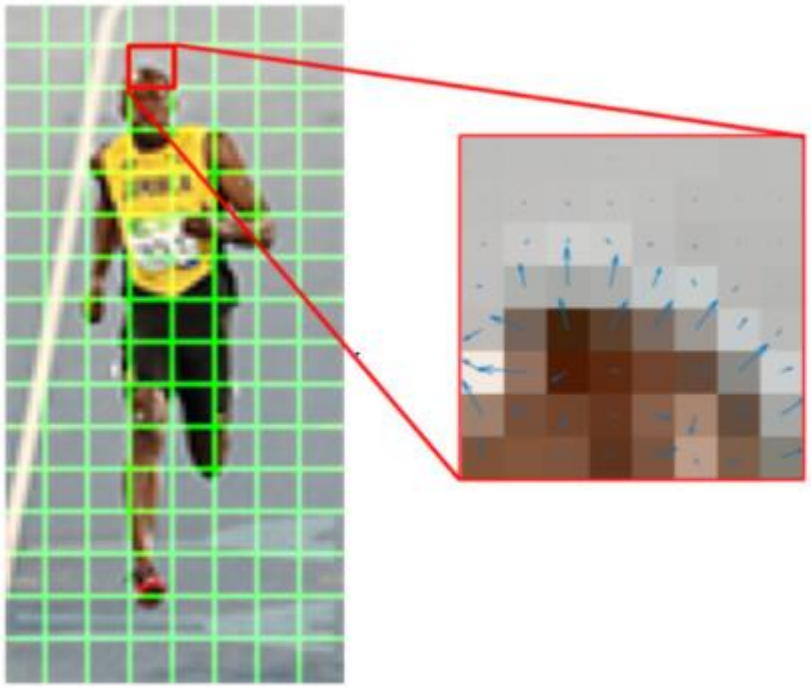
$$g_a(t)=\frac{1}{2\sqrt{\pi a}} e^{\frac{-t^2}{4a}}$$



通过频率参数和高斯函数参数的选取，Gabor变换可以选取很多纹理特征，但是Gabor是非正交的，不同特征分量之间有冗余，所以在对纹理图像的分析中效率不太高。

二维 Gabor 滤波器是在二维 Gabor 函数上构建的。二维 Gabor 滤波器的参数分别定义了滤波器的方向、滤波器的尺度和窗口函数。上图为5个尺度(高斯函数的方差a取值不同)、8个方向上的Gabor滤波器核函数示意图

方向梯度



2	3	4	4	3	4	2	2
5	11	17	13	7	9	3	4
11	21	23	27	22	17	4	6
23	99	165	135	85	32	26	2
91	155	133	136	144	152	57	28
98	196	76	38	26	60	170	51
165	60	60	27	77	85	43	136
71	13	34	23	108	27	48	110

Gradient Magnitude

80	36	5	10	0	64	90	73
37	9	9	179	78	27	169	166
87	136	173	39	102	163	152	176
76	13	1	168	159	22	125	143
120	70	14	150	145	144	145	143
58	86	119	98	100	101	133	113
30	65	157	75	78	165	145	124
11	170	91	4	110	17	133	110

Gradient Direction

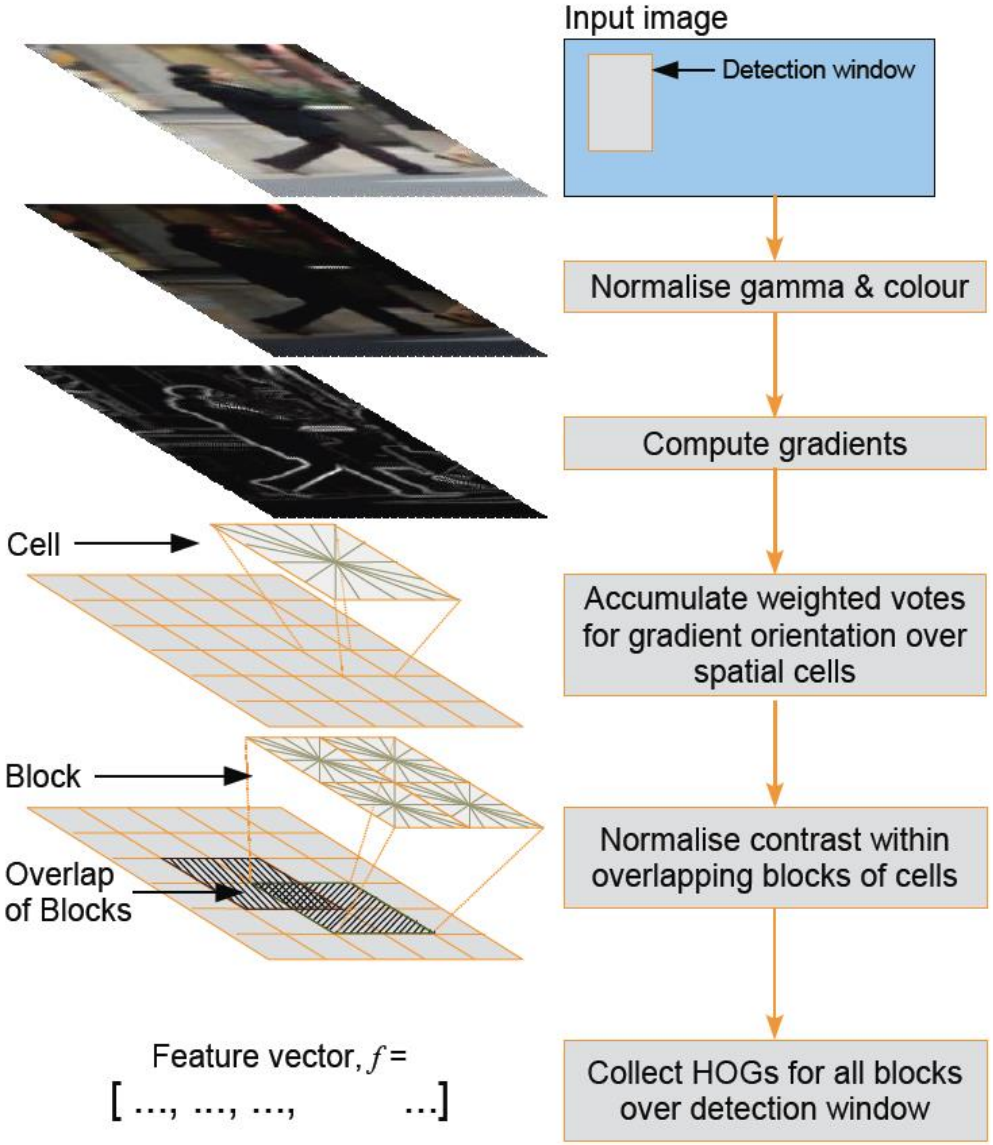
Center : The RGB patch and gradients represented using arrows. Right : The gradients in the same patch represented as numbers

中间图，箭头是梯度的方向，长度是梯度的大小，箭头的指向方向是像素强度变化方向，幅值是强度变化的大小。
右边图，梯度方向矩阵中可以看到角度是0-180度，不是0-360度，这种被称之为"无符号"梯度("unsigned" gradients)，因为一个梯度和它的负数是用同一个数字表示的，也就是说一个梯度的箭头以及它旋转180度之后的箭头方向被认为是一样的。

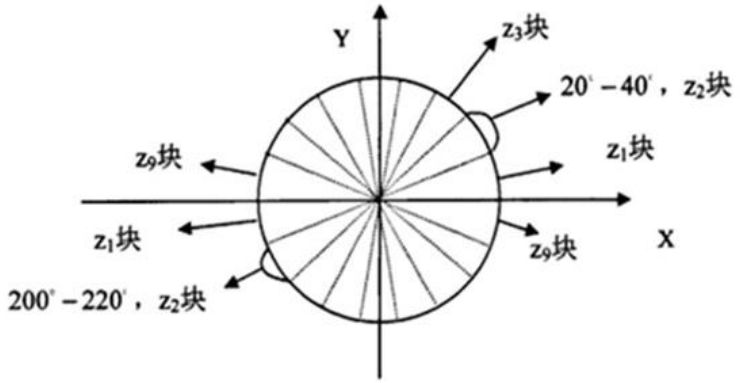
方向梯度直方图

http://tel.archives-ouvertes.fr/docs/00/39/03/03/PDF/NavneetDalalThesis.pdf
Histogram of Oriented Gradient, Navneet Dalal and Bill Triggs, CVPR 2005
http://lear.inrialpes.fr/people/triggs/pubs/Dalal-cvpr05.pdf 2015.05 google cited:11065

HOG: Histogram of Oriented Gradient

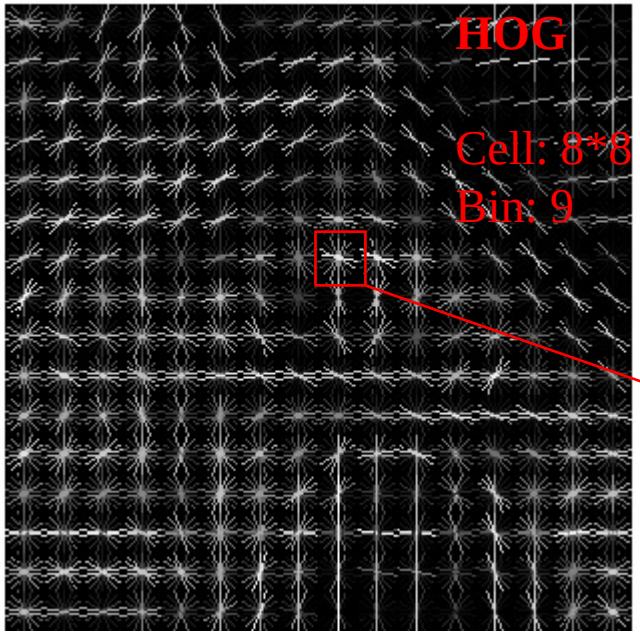


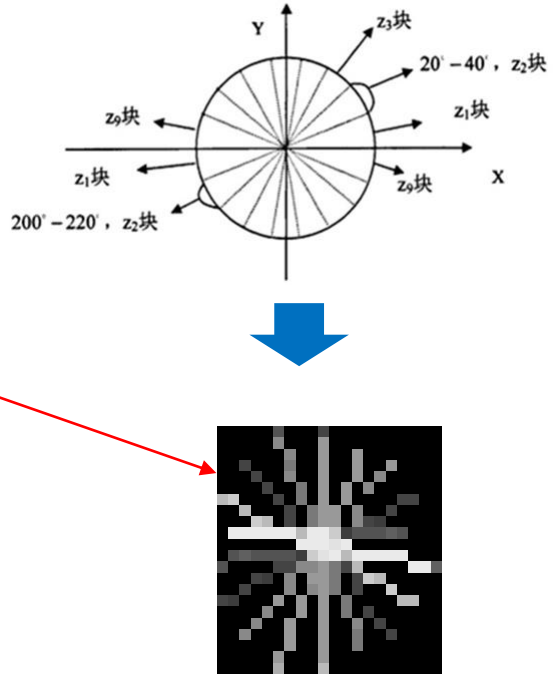
- 将图像分成16*16的cell，用9个bin的直方图来统计cell的梯度信息（9维特征向量）。即可形成每个cell的descriptor；
- 2*2个cell组成一个block，block内所有cell的特征descriptor串联起来便得到该block的HOG特征descriptor。
- 将图像内的所有block的HOG特征descriptor串联起来就可以得到该image的HOG特征descriptor了。



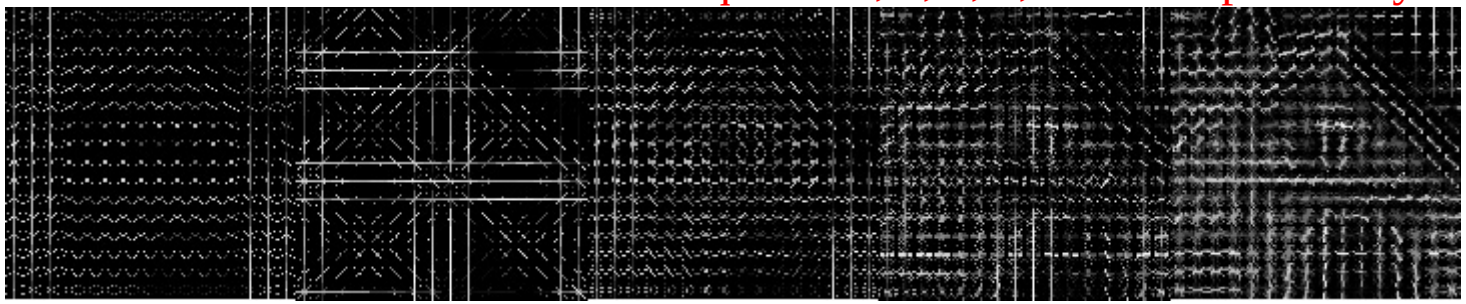
方向梯度直方图 结果示例

Histogram of Oriented Gradient, HOG





HOG features for numOrientations equal to 3, 4, 5, 9, and 21 repsectively.



梯度计算：对于彩色图像，3个颜色通道单独计算，取最大值

http://www.vlfeat.org/overview/hog.html

小结：纹理特征与图像分类/检索

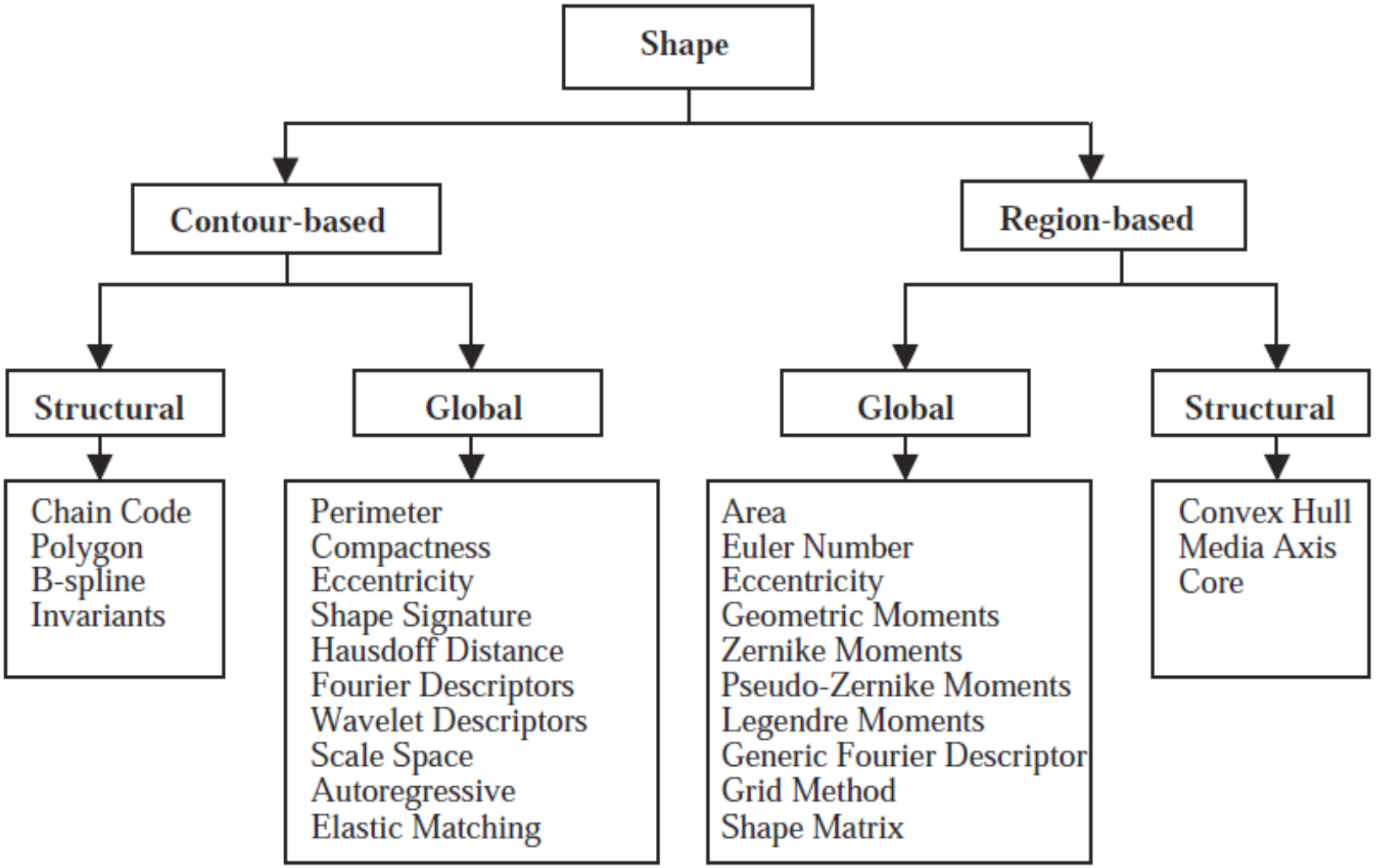
- 基于纹理特征的图像分类、检索常用的方法：
  - 基于梯度的算子
    - 一阶微分算子：Robert 算子，Sobel 算子，Prewitt算子等
    - 二阶微分算子：Laplace 算子，LOG 算子和 Canny 算子等
  - 基于 Gabor 小波的纹理特征提取
  - LBP 纹理统计特征提取
  - HOG: Histogram of Oriented Gradient 2005
  - CENTRIST (CENsus TRansform hISTogram) 2011
  - 基于灰度共生矩阵的纹理分析
  - 基于傅里叶变换的纹理特征提取
  - ……

图像分类的算法思想

- 从文本分类→图像分类
  - 如何从图像中获取全局特征？
    - 颜色特征、纹理特征、形状特征
  - 如何从图像中获取局部特征？
    - SIFT: Scale-invariant feature transform
- 图像分类的几个发展阶段
  - Low-level Modelling
  - Semantic Modelling
  - Sparse Coding
  - Deep Learning

形状特征

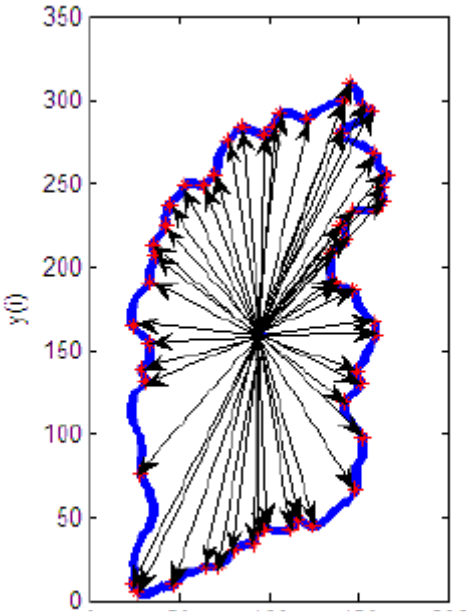
两类不同的形状描述方式



Review of shape representation andd escription techniques, Pattern Recognition, 2004
2015.05 Google cited: 1341

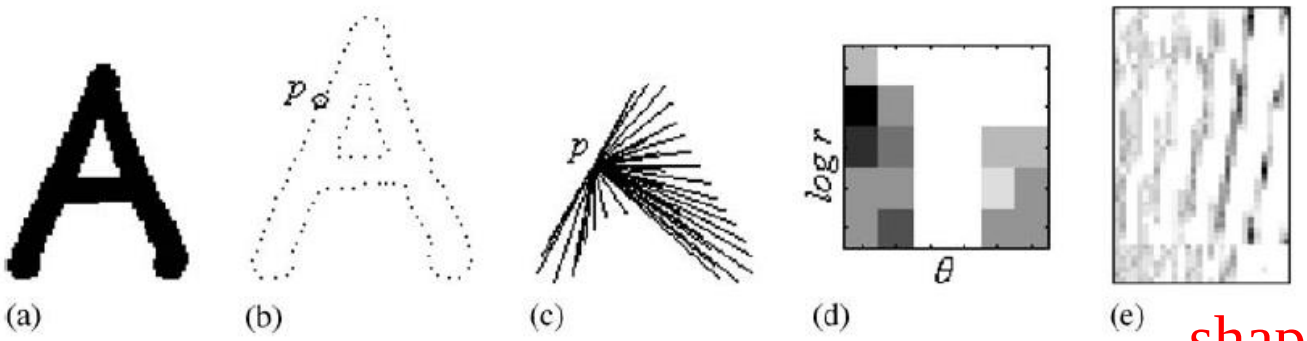
形状特征

基于轮廓的描述方法示例



基于轮廓线的特征描述示例：

- 在轮廓线上提取特征点：如把轮廓线上局部曲率极大值的点定义为尖点。
- 把由特征点向轮廓线质心所引向量定义为特征向量。所有特征向量集合称为特征向量集。



shape context

Fig. 3. Shape context. (a) a character shape; (b) edge image of (a); (c) a point p on shape (a) and all the vectors started from p; (d) the log-polar histogram of the vectors in (c), the histogram is the context of point p; (e) the context map of shape (a), each row of the context map is the flattened histogram of each point context, the number of rows is the number of sampled points. (reprinted from [10]).

形状特征

基于区域的描述方法示例

Fig. 10. (a) An original shape in polar space; (b) polar-raster sampled image plotted in Cartesian space.

通用傅里叶描述符(Generic Fourier Descriptor，GFD)。GFD先采用修正了的平面极坐标傅立叶变换来对图像进行采样，然后将采样的信息重新绘制在笛卡儿直角坐标系下，最后再对该直角坐标下的图像做傅立叶变换。
其中修正了的极坐标傅里叶变换的表示如公式：

$$GF(\rho,\theta)=\sum_{r}\sum_{i} f(r,\theta_i)\bullet \exp\left[j2\pi\left(\frac{r}{R}\rho z+\frac{2\pi i}{T}\phi\right)\right]$$

通用傅里叶描述符不仅计算简单，而且由于特征为纯光谱特征可以在形状的径向和圆形方向上进行多分辨率分析，所以 GDF 具有较好的检索性能。

小结：如何从图像中获取全局特征？

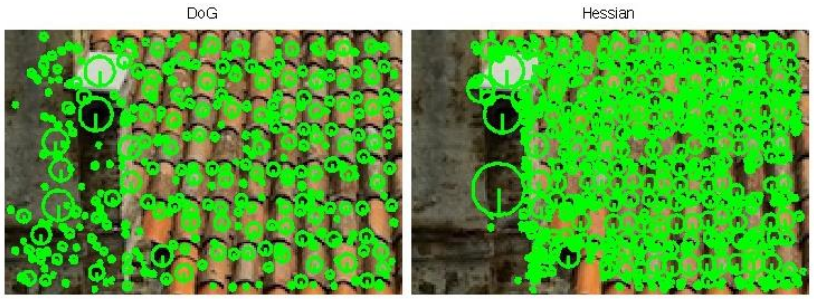
- 颜色特征
  - 直方图、分块直方图、梯度直方图
- 纹理特性
  - 基于梯度的算子：一阶微分、二阶微分
  - 基于 Gabor 小波的纹理特征提取
  - LBP 纹理统计特征提取
  - HOG: Histogram of Oriented Gradient 2005
  - CENTRIST (CENsus TRansform hISTogram) 2011
- 形状特性
  - 基于轮廓、基于区域

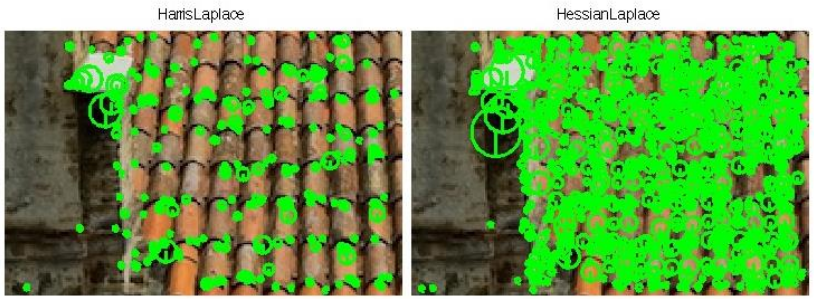
图像分类的算法思想

- 从文本分类→图像分类
  - 如何从图像中获取全局特征？
    - 颜色特征、纹理特征、形状特征
  - 如何从图像中获取局部特征？
    - SIFT: Scale-invariant feature transform
- 图像分类的几个发展阶段
  - Low-level Modelling
  - Semantic Modelling
  - Sparse Coding
  - Deep Learning

图像的局部特征

- 图像中存在一些能够描述图像主要内容的像素点(也称为显著点)，换句话说，图像中各个部分对表达图像内容的重要性是不同的。因此，使用图像的局部特征比上述全局特征能够更好地反映图像的内容。
- 局部特征提取的一般步骤
  - 特征点的检测
  - 特征点的描述





Scale-invariant feature transform (SIFT)

David G. Lowe
Computer Science Department
2366 Main Mall
University of British Columbia
Vancouver, B.C., V6T 1Z4,
Canada
E-mail: lowe@cs.ubc.ca

1999年British Columbia大学大卫.劳伊（David G.Lowe）教授总结了现有的基于不变量技术的特征检测方法，并正式提出了一种基于尺度空间的、对图像缩放、旋转甚至仿射变换保持不变性的图像局部特征描述算子—SIFT（尺度不变特征变换），这种算法在2004年被加以完善。

SIFT
二维图像的尺度空间

the scale space of an image is defined as a function, L(x, y, σ),

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y)$$

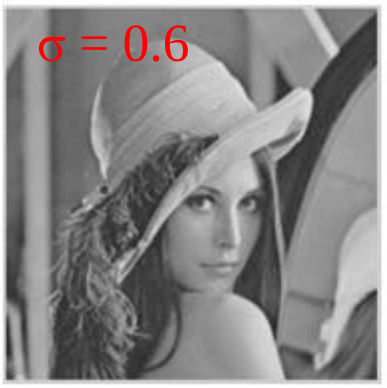
where $*$ is the convolution operation in x and y, and

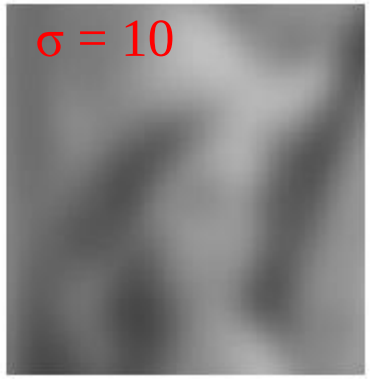
$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$$

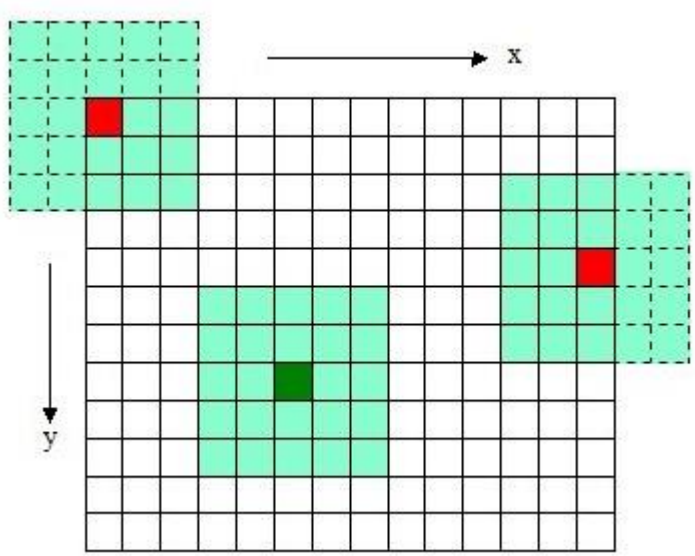
← G(x,y,σ) 是尺度可变高斯函数

尺度空间理论的基本思想是：在图像信息处理模型中引入一个被视为尺度的参数，通过连续变化尺度参数获得多尺度下的尺度空间表示序列，对这些序列进行尺度空间主轮廓的提取，并以该主轮廓作为一种特征向量，实现边缘、角点检测和不同分辨率上的特征提取等。

5*5的高斯模板卷积计算示例







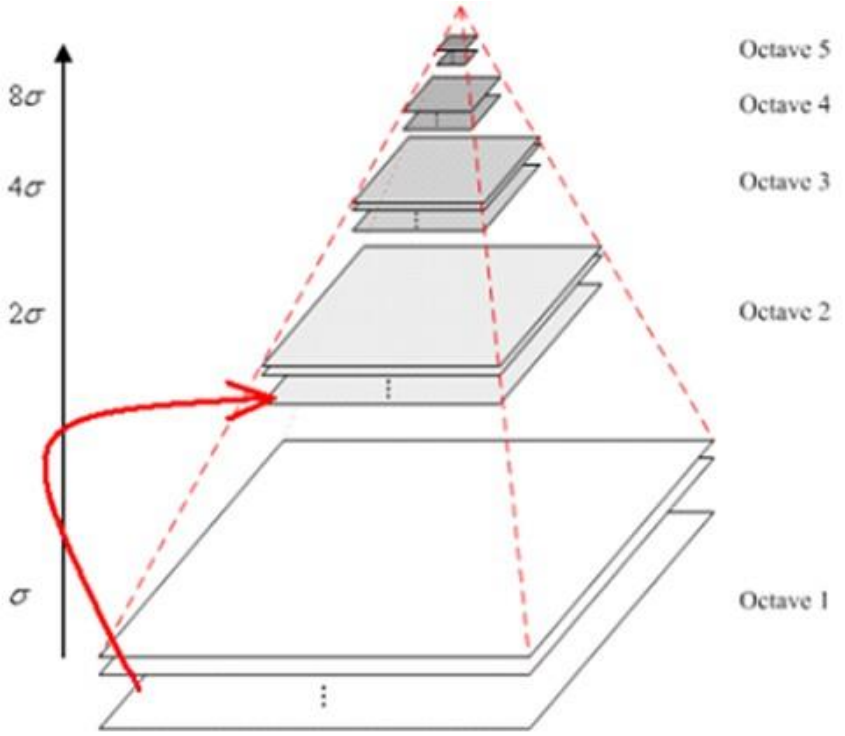
5x5 Gaussian filter, with $\sigma = 1.4$

$$\mathbf{B} = \frac{1}{159}\begin{bmatrix} 2 & 4 & 5 & 4 & 2 \\ 4 & 9 & 12 & 9 & 4 \\ 5 & 12 & 15 & 12 & 5 \\ 4 & 9 & 12 & 9 & 4 \\ 2 & 4 & 5 & 4 & 2 \end{bmatrix} * \mathbf{A}$$

图像的金字塔模型

图像金字塔：将原始图像不断降阶采样，得到一系列大小不一的图像，由大到小，从下到上构成的塔状模型。原图像为金子塔的第一层，每次降采样所得到的新图像为金字塔的一层(每层一张图像)。

为了让尺度体现其连续性，高斯金字塔在简单降采样的基础上加上了高斯滤波。将图像金字塔每层的一张图像使用不同参数做高斯模糊，使得金字塔的每层含有多张高斯模糊图像，将金字塔每层多张图像合称为一组(Octave)



i为塔的层数（即Octave数目），s为每层（Octave）内图像数目。尺度空间的取值为：

$$2^{i-1}(\sigma, k\sigma, k^2\sigma, \cdots k^{n-1}\sigma),\ k = 2^{1/s}$$

塔内每张高斯模糊图像对应一个特定的σ

高斯差分尺度空间（DOG scale-space）

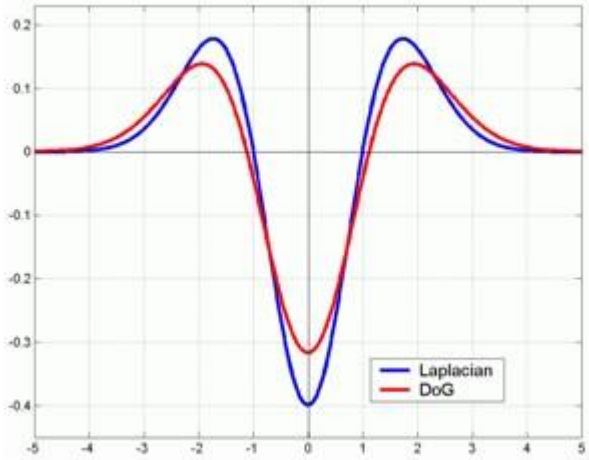
difference-of-Gaussian function convolved with the image, D(x, y, σ)

$$D(x,y,\sigma) = (G(x,y,k\sigma) - G(x,y,\sigma)) * I(x,y)$$
$$= L(x,y,k\sigma) - L(x,y,\sigma).$$

拉普拉斯算子做二阶微分: $\nabla^2 G$

拉普拉斯高斯算子(Laplacian of Gaussian, LOG)进行了滤波和二阶微分

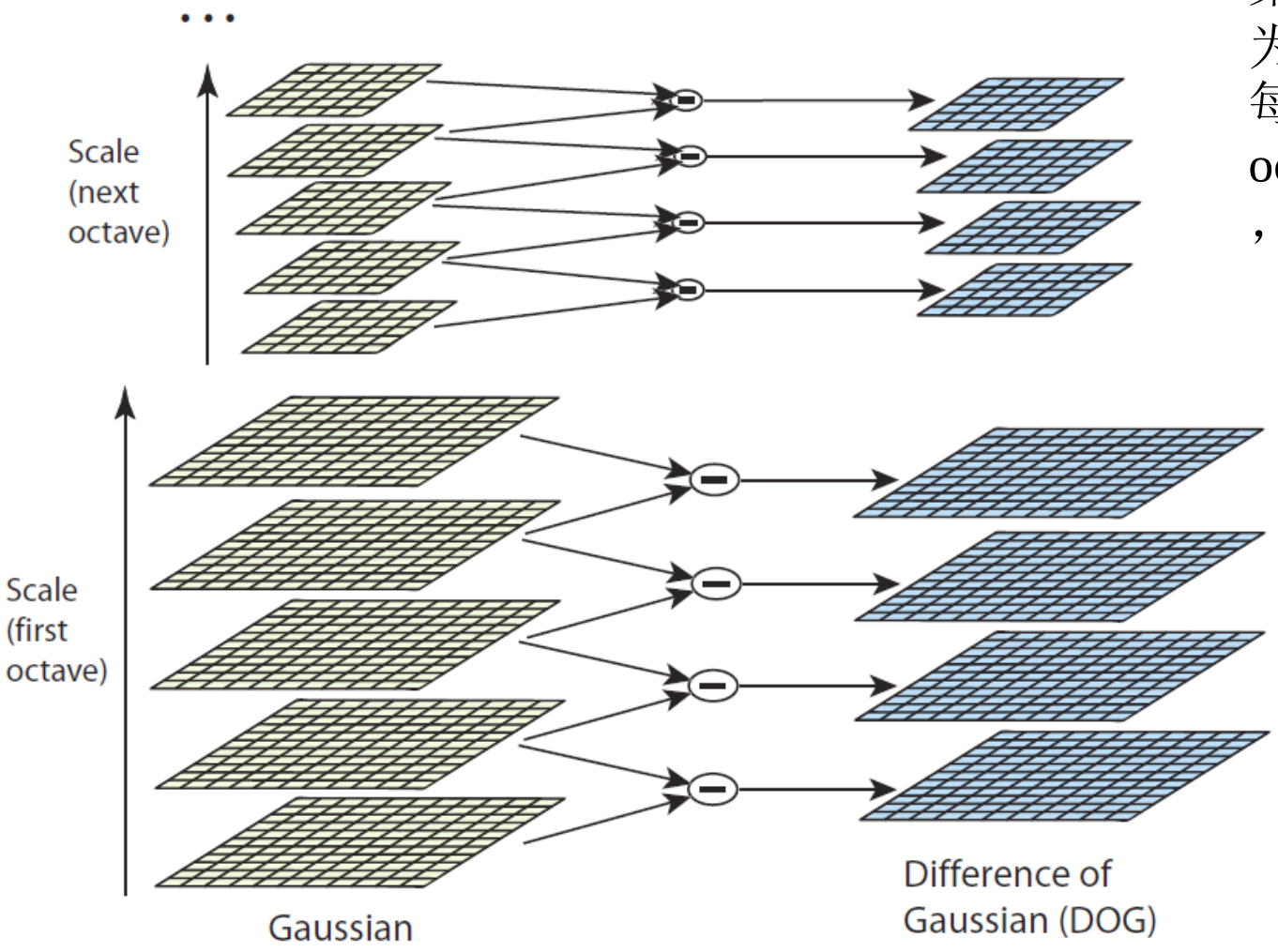
DOG近似是尺度归一化的LOG: $G(x, y, k\sigma) - G(x, y, \sigma) \approx (k-1)\sigma^2 \nabla^2 G$



高斯拉普拉斯和高斯差分的比较

高斯差分金字塔

DoG (difference-of-Gaussian) images构造

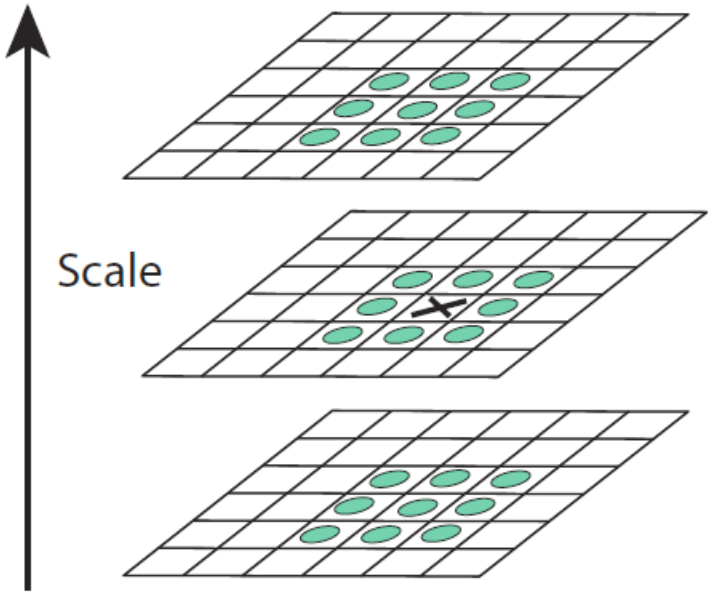


第一个octave的scale为原图大小，后面每个octave为上一个octave降采样的结果，即原图的1/4

在检测极值点前对原始图像的高斯平滑以致图像丢失高频信息，所以Lowe 建议在建立尺度空间前首先对原始图像长宽扩展一倍，以保留原始图像信息，增加特征点数量。

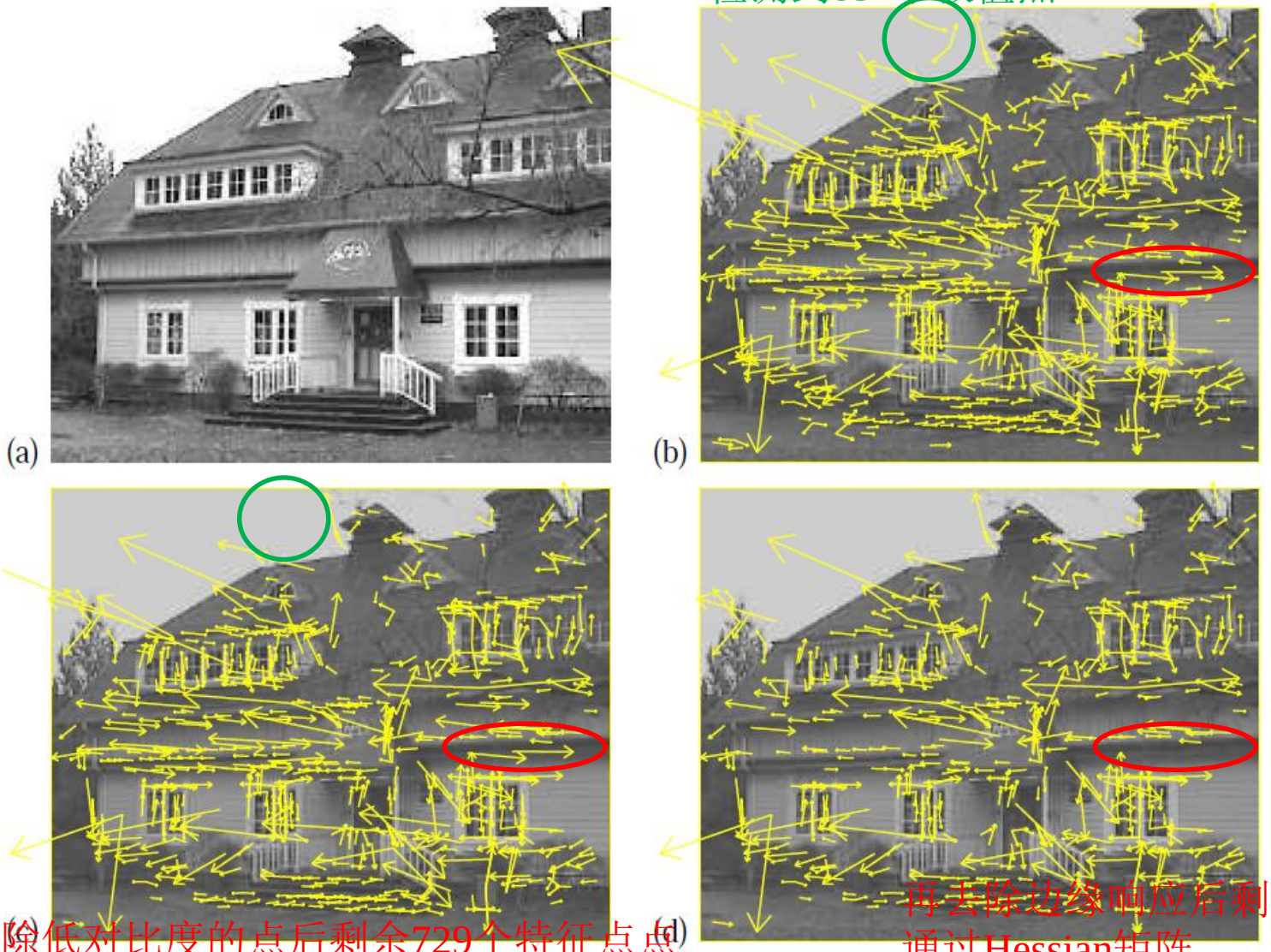
检测D(x, y, σ)的极值点 这些极值点为备选特征点

为了寻找尺度空间的极值点，每一个采样点要和它所有的相邻点比较，看其是否比它的图像域和尺度域的相邻点大或者小。



如图所示，中间的检测点和它同尺度的8个相邻点和上下相邻尺度对应的9×2个点共26个点比较，以确保在尺度空间和二维图像空间都检测到极值点。一个点如果在DOG尺度空间本层以及上下两层的26个领域中是最大或最小值时，就认为该点是图像在该尺度下的一个特征点。

从备选特征点中去除不好的特征点



绿色圈内的点为低对比度的特征点

丢弃|D(x)|<0.03的极值点

红色圈内的点为边缘响应的点

$$\frac{\mathrm{Tr}(\mathbf{H})^2}{\mathrm{Det}(\mathbf{H})} < \frac{(r+1)^2}{r}$$

$$r = 10$$

去除低对比度的点后剩余729个特征点点

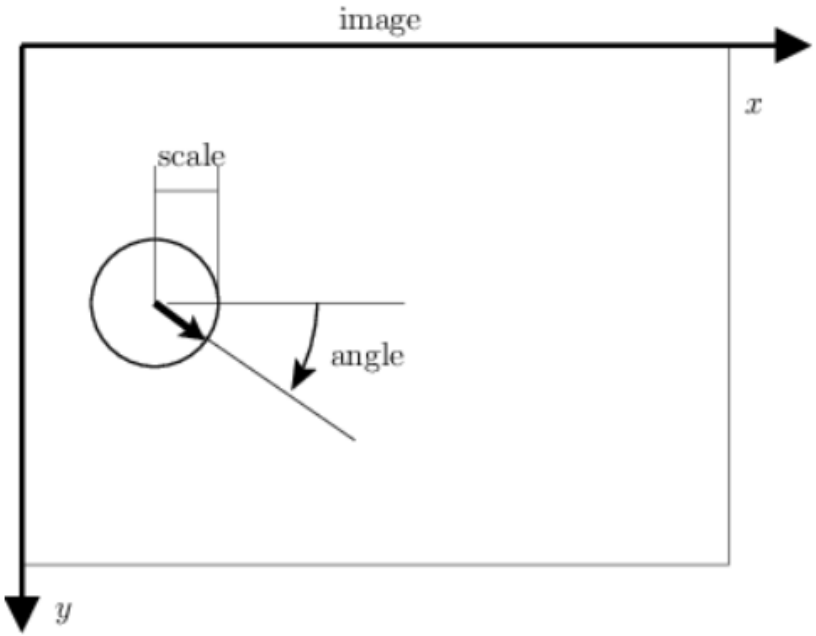
再去除边缘响应后剩余635个特征点

通过Hessian矩阵

计算特征点的主方向
Orientation assignment

前述步骤检测出的特征点的邻域构成一个关键点（Keypoint）
A SIFT keypoint is a circular image region with an orientation.
每个关键点（Keypoint）有三项信息：位置，尺度、方向
It is described by a geometric frame of four parameters: the keypoint center coordinates x and y, its scale (the radius of the region), and its orientation (an angle expressed in radians).

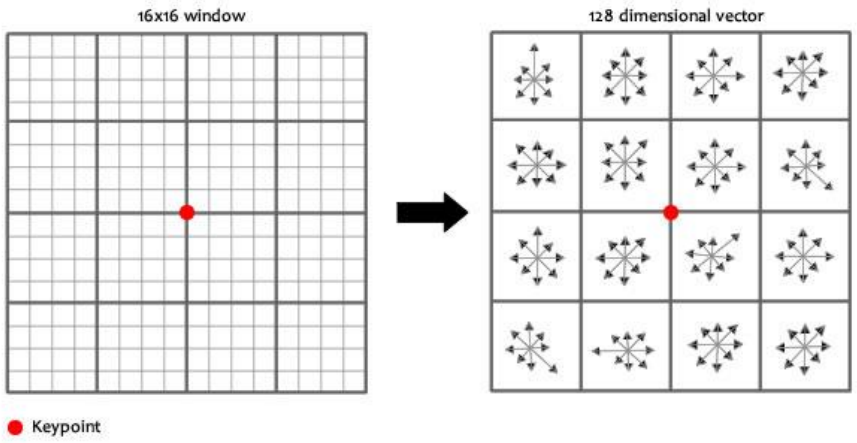
主方向的计算：
以关键点为中心的邻域窗口内采样，并用直方图统计邻域像素的梯度方向。梯度直方图的范围是0～360度，其中每10度一个柱，总共36个柱。直方图的峰值则代表了该关键点处邻域梯度的主方向，即作为该关键点的方向。



SIFT keypoints are circular image regions with an orientation.

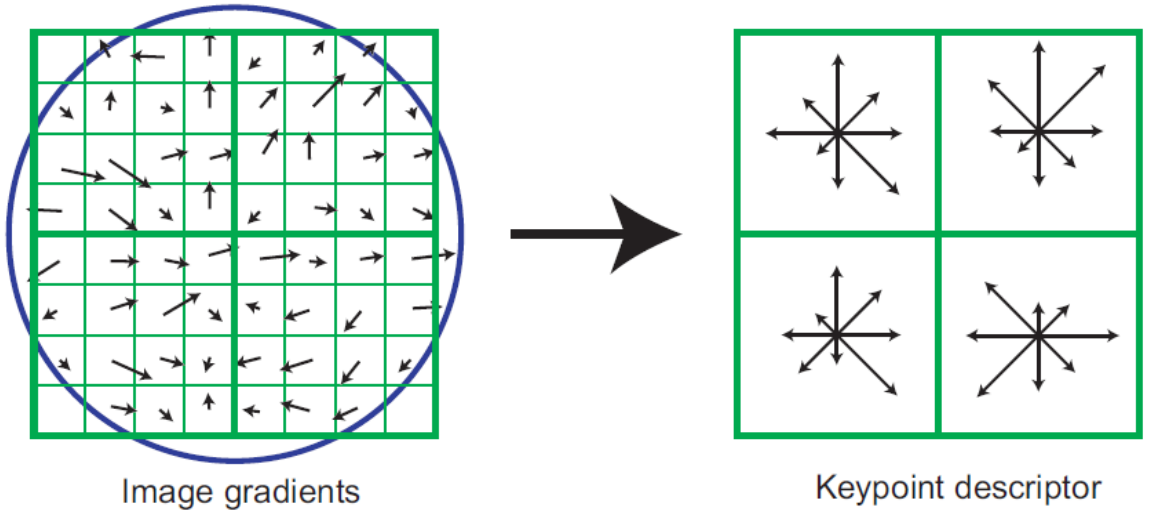
Picture Source http://www.vlfeat.org/api/sift.html

SIFT描述子
SIFT descriptor



接下来需要为每个关键点建立一个描述子（descriptor），用一组向量将这个关键点描述出来，使其不随各种变化而改变，如光照变化、视角变换等。这个描述符不但包括关键点，也包含关键点周围对其有贡献的像素点。

将坐标轴旋转为关键点的方向，Lowe建议描述子使用关键点尺度空间内4*4的窗口中计算的8个方向的梯度信息，共4*4*8=128维向量表征。



SIFT算法的应用示例

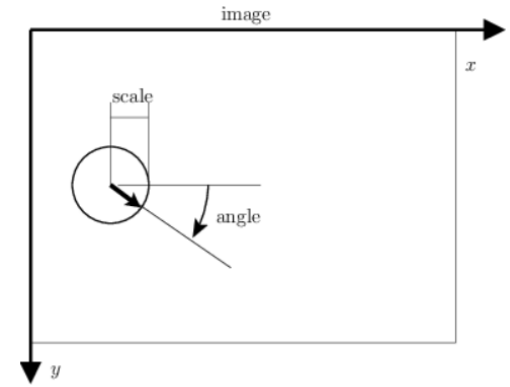
Demo Software: SIFT Keypoint Detector
http://www.cs.ubc.ca/~lowe/keypoints/

小结：SIFT原理

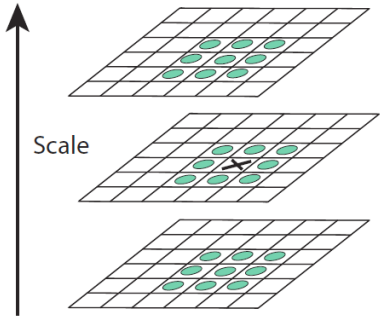
1、Detection of scale-space extreme 构建尺度空间
2、Accurate keypoint localization 关键点检测
3、Orientation assignment 指定方向
4、The local image descriptor 局部图像描述子

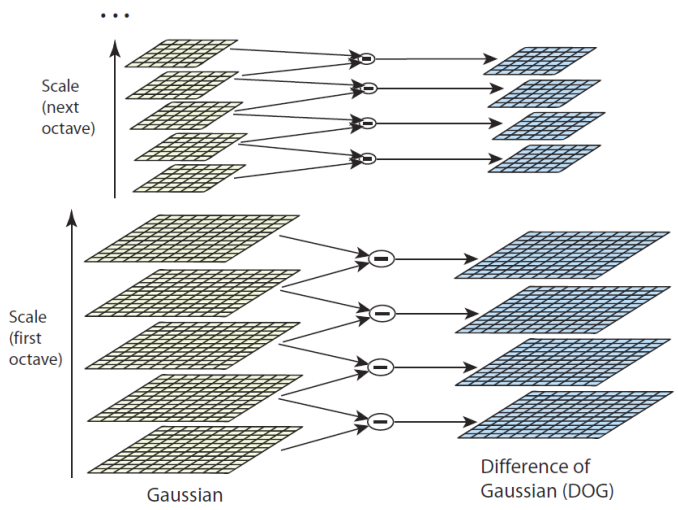
$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y)$$

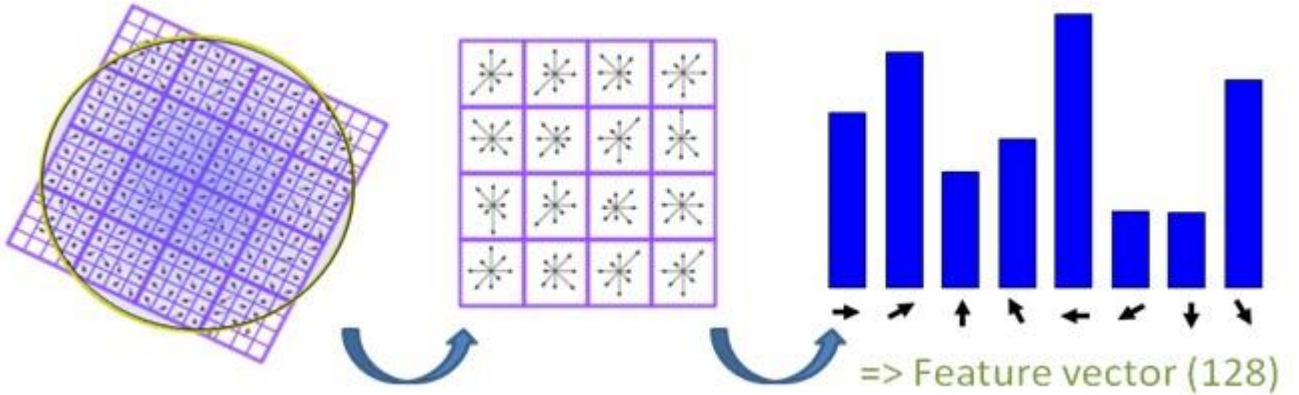
$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma)$$



SIFT keypoints are circular image regions with an orientation.







缺点：实时性不高、有时特征点太少、对边缘光滑的目标无法准确提取特征点

小结：SIFT之后的特征检测算法

主要论文及截止2015年5月Google cited情况

Binary (faster, real time)

- Alexandre Alahi, Raphael Ortiz, Pierre Vandergheynst: 'FREAK: Fast Retina Keypoint', CVPR 2012 Google cited: 462
- Rublee, E., Rabaud, V., Konolige, K., and Bradski, G.: 'ORB: An efficient alternative to SIFT or SURF'. ICCV 2011 Google cited: 1037
- Leutenegger, S., Chli, M., and Siegwart, R.Y.: 'BRISK: Binary Robust Invariant Scalable Keypoints', ICCV, 2011 Google cited: 666
- Calonder, M., Lepetit, V., Strecha, C., and Fua, P.: 'BRIEF: Binary Robust Independent Elementary Features', ECCV 2010 Google cited: 972
- Bay, H., Tuytelaars, T., and Gool, L.V.: 'SURF: Speeded Up Robust Features', ECCV, 2006 Google cited: 5428
- **Lowe, D.G.: 'Distinctive Image Features from Scale-Invariant Keypoints', International Journal of Computer Vision, 2004** Google cited: 29656
- Lowe, D.G.: 'Object recognition from local scale-invariant features'. ICCV 1999 Google cited: 9366

Traditional (slower, accurate)

关于Descriptor的新思路？

http://www.cv-foundation.org/openaccess/CVPR2015.py 14篇

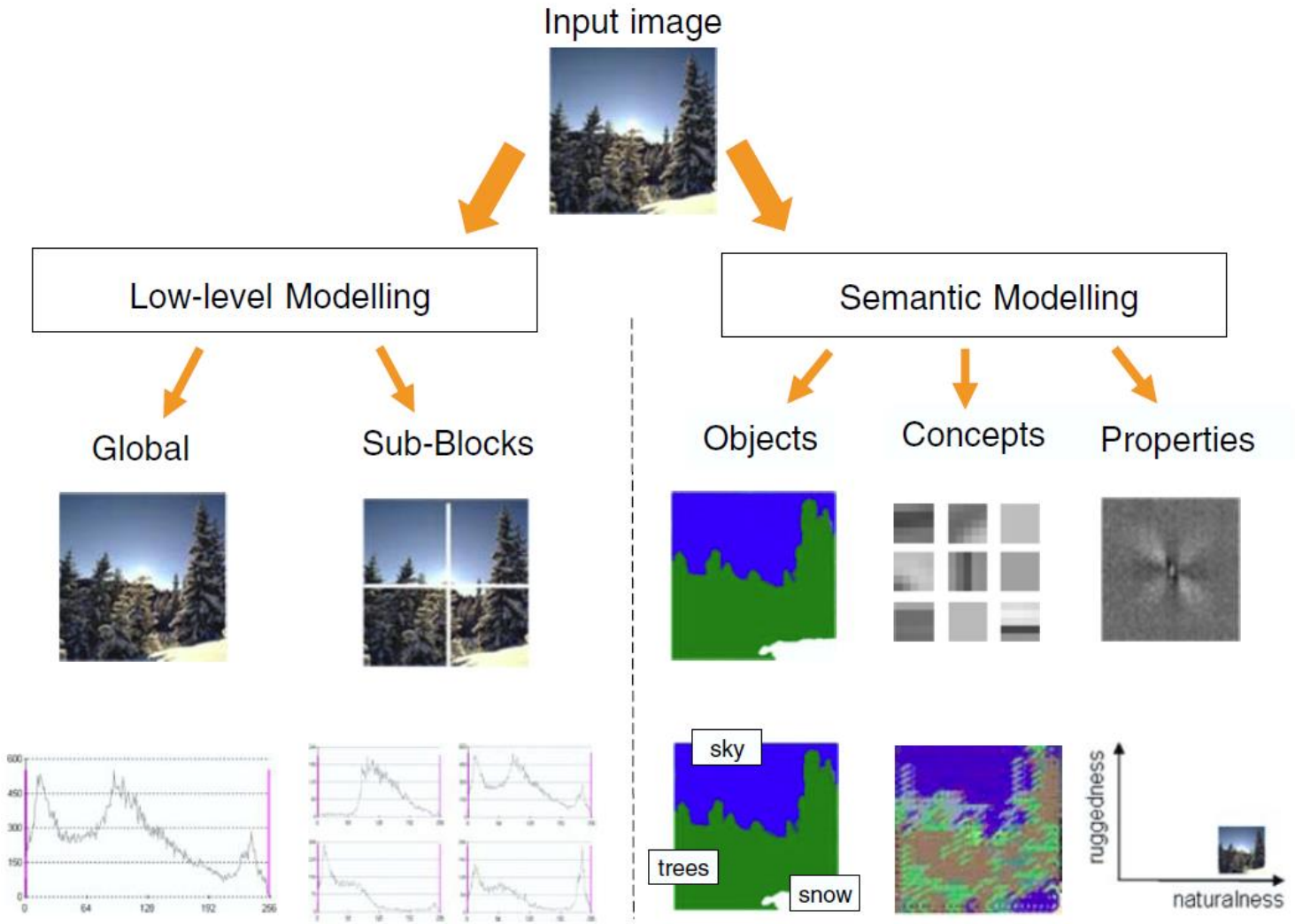
- Heat Diffusion Over Weighted Manifolds: A New Descriptor for Textured 3D Non-Rigid Shapes
- Nested Motion Descriptors
- Supervised Descriptor Learning for Multi-Output Regression
- DeepShape: Deep Learned Shape Descriptor for 3D Shape Matching and Retrieval
- Robust Image Alignment With Multiple Feature Descriptors and Matching-Guided Neighborhoods
- DASC: Dense Adaptive Self-Correlation Descriptor for Multi-Modal and Multi-Spectral Correspondence
- 3D Deep Shape Descriptor
- BOLD - Binary Online Learned Descriptor For Efficient Image Matching
- Learning Descriptors for Object Recognition and 3D Pose Estimation
- Descriptor Free Visual Indoor Localization With Line Segments
- Shape-Tailored Local Descriptors and Their Application to Segmentation and Tracking
- Action Recognition With Trajectory-Pooled Deep-Convolutional Descriptors
- Domain-Size Pooling in Local Descriptors: DSP-SIFT
- A Maximum Entropy Feature Descriptor for Age Invariant Face Recognition

形状特征迄今没有被很好的描述

图像分类的算法思想

- 从文本分类→图像分类
  - 如何从图像中获取全局特征？
    - 颜色特征、纹理特征、形状特征
  - 如何从图像中获取局部特征？
    - SIFT: Scale-invariant feature transform
- 图像分类的几个发展阶段
  - Low-level Modelling
  - Semantic Modelling
  - Sparse Coding
  - Deep learning

图像分类/识别系统的不同思路

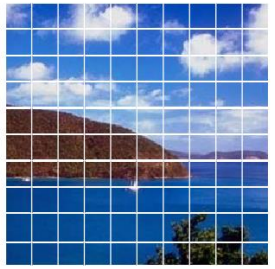


Bosch, A., et al. (2006). "A review: Which is the best way to organize/classify images by content" Image and Vision Computing.

Low-level：基于低层特征的方法

- 颜色直方图、功率谱、纹理共生矩阵、颜色矩、颜色相关图.....
- 分为两类：(1) 全局方法，即提取图像整体的低层特征；(2) 局部方法，即首先按照某种规则把图像划分成块，然后提取块的低层特征。

(a) 原图



(b) 规则网格法特征检测

(c) 原图

(d) Harris 角点检测

语义目标方法
Semantic Objects

- 语义目标：这类方法主要依赖于对图像的初始分割，以处理图像中不同的区域。
- 在图像分割后，采用局部分类器将分割区域标记为已知的目标种类（如人、天空、草地等），最后使用这些局部信息对整个场景进行分类。

语义概念方法
Semantic Concepts

- 语义概念：这类方法利用在关键点周围的局部描述子所携带的中层信息来表示图像的语义类别，通过中层语义表示的引入，可以解决低层特征与高层概念之间的“语义鸿沟”问题。
- 此外，这类方法并不依赖于初始的图像分割，而是通过局部描述子表示场景内容。这类方法与统计文本分析的“词汇袋”（bag-of-words）方法类似，而后被应用于计算机视觉领域并推广为“特征袋”（bag-of-features）方法。

语义属性方法
Semantic Properties

- 语义属性：这类方法与上述两类语义建模方法有着较大区别，语义属性探索的是场景的统计性质，涉及到场景图像的全局结构而不是局部的目标或区域，因此不需要对图像进行分割或局部区域的处理。
- 这类方法以 Oliva 和 Torralba 提出的场景识别模型为代表。Oliva 和 Torralba【Oliva, A. and A. Torralba (2001). "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope." International Journal of Computer Vision 42(3).】提出的场景识别模型不需对图像进行分割，也不需对局部的目标或区域进行处理，整个模型基于一种低维的场景表示，称之为空域包络（Spatial Envelope）。它包括五类感知属性：自然度、开放度、粗糙度、展开度和崎岖度。每类属性对应于空域包络空间中的一维，所有特征维的组合表示了一幅场景的主要空域结构。

图像分类的算法思想

- 从文本分类→图像分类
  - 如何从图像中获取全局特征？
    - 颜色特征、纹理特征、形状特征
  - 如何从图像中获取局部特征？
    - SIFT: Scale-invariant feature transform
- 图像分类的几个发展阶段
  - Low-level Modelling
  - Semantic Modelling
  - Sparse Coding
  - Deep learning

稀疏编码

起源 Sparse coding (Olshausen & Field,1996)

1996年Cornell大学心理学院的Bruno在Nature上发表了一篇题名为：“emergence of simple-cell receptive fieldproperties by learning a sparse code for nature images”的文章，大意是讲哺乳动物的初级视觉的简单细胞的感受野具有空域局部性、方向性和带通性（在不同尺度下，对不同结构具有选择性），和小波变换的基函数具有一定的相似性。

Input: Images $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ (each in $R^{n \times n}$)

Learn: Dictionary of bases $f_1, f_2, \ldots, f_k$ (also $R^{n \times n}$), so that each input x can be approximately decomposed as:

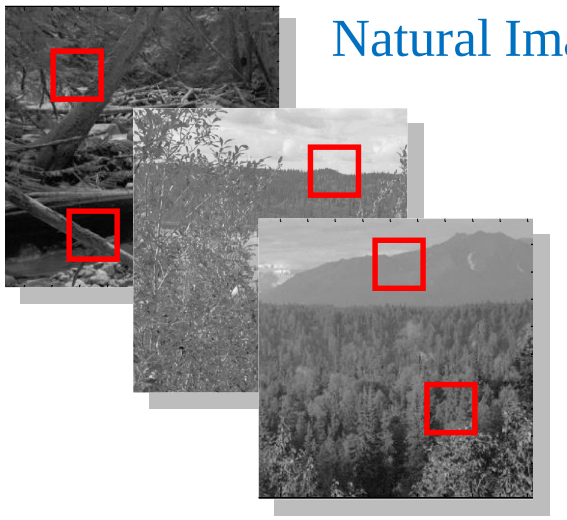
$$x \approx \sum_{j=1}^{k} a_j \emptyset_j$$

s.t. a_j's are mostly zero (“sparse”)

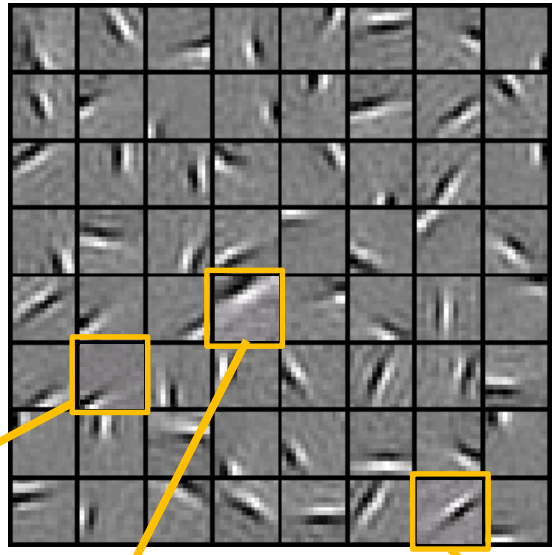
Bruno Olshausen和 David Field 收集了很多黑白风景照片，从这些照片中，提取出400个小碎片，每个照片碎片的尺寸均为 16x16 像素，不妨把这400个碎片标记为 S[i], i = 0,.. 399。接下来，再从这些黑白风景照片中，随机提取另一个碎片，尺寸也是 16x16 像素，不妨把这个碎片标记为 T。他们提出的问题是，如何从这400个碎片中，选取一组碎片，S[k], 通过叠加的办法，合成出一个新的碎片，而这个新的碎片，应当与随机选择的目标碎片 T，尽可能相似，同时，S[k] 的数量尽可能少。

稀疏编码
图像稀疏表示的示例

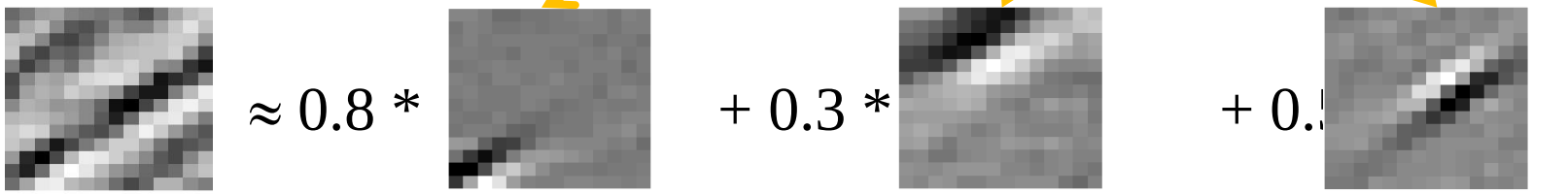
Learned bases (f_1, …, f_{64}): "Edges"



Natural Images



Test example

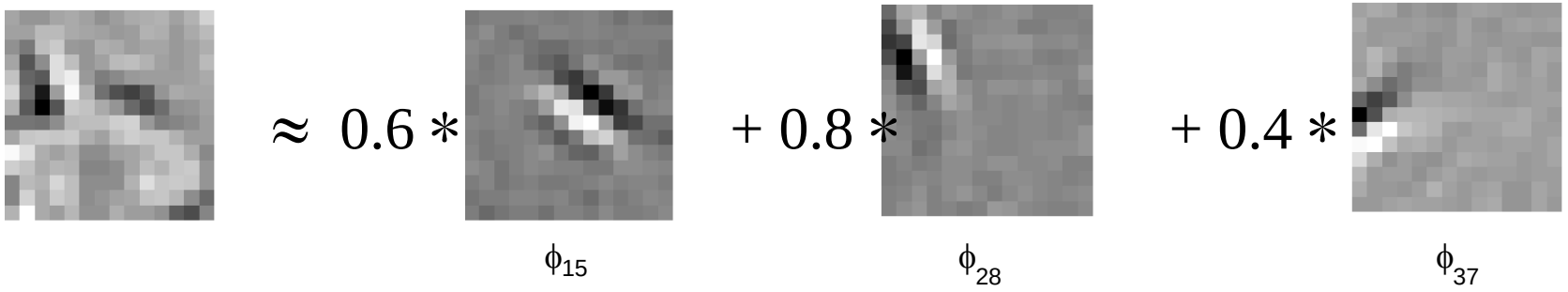


$x \approx 0.8 * \phi_{36} + 0.3 * \phi_{42} + 0.5 * \phi_{63}$

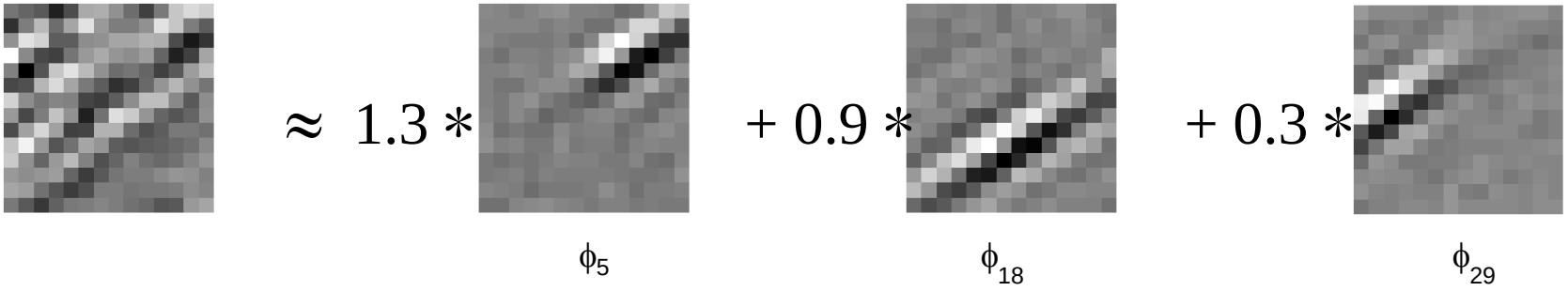
[0, 0, …, 0, **0.8**, 0, …, 0, **0.3**, 0, …, 0, **0.5**, …]
= [a_1, …, a_{64}] (feature representation)

Compact & easily interpretable

稀疏编码
图像稀疏表示的示例



Represent as: [0, 0, …, 0, 0.6, 0, …, 0, 0.8, 0, …, 0, 0.4, …]

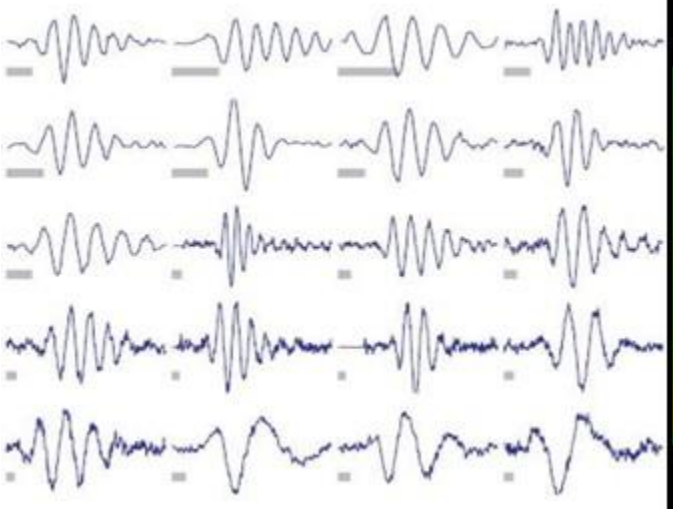


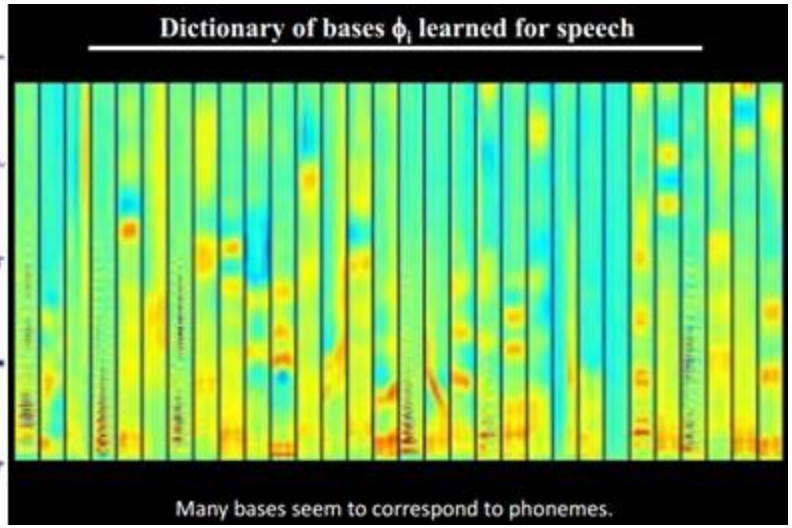
Represent as: [0, 0, …, 0, 1.3, 0, …, 0, 0.9, 0, …, 0, 0.3, …]

- Method hypothesizes that edge-like patches are the most "basic" elements of a scene, and represents an image in terms of the edges that appear in it.
- Use to obtain a more compact, higher-level representation of the scene than pixels.

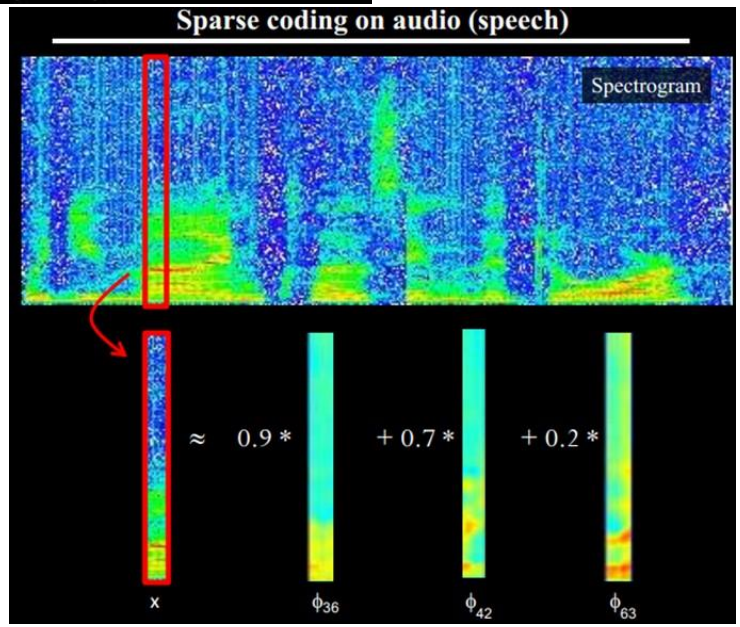
http://ufldl.stanford.edu/eccv10-tutorial/

稀疏编码
音频的稀疏编码



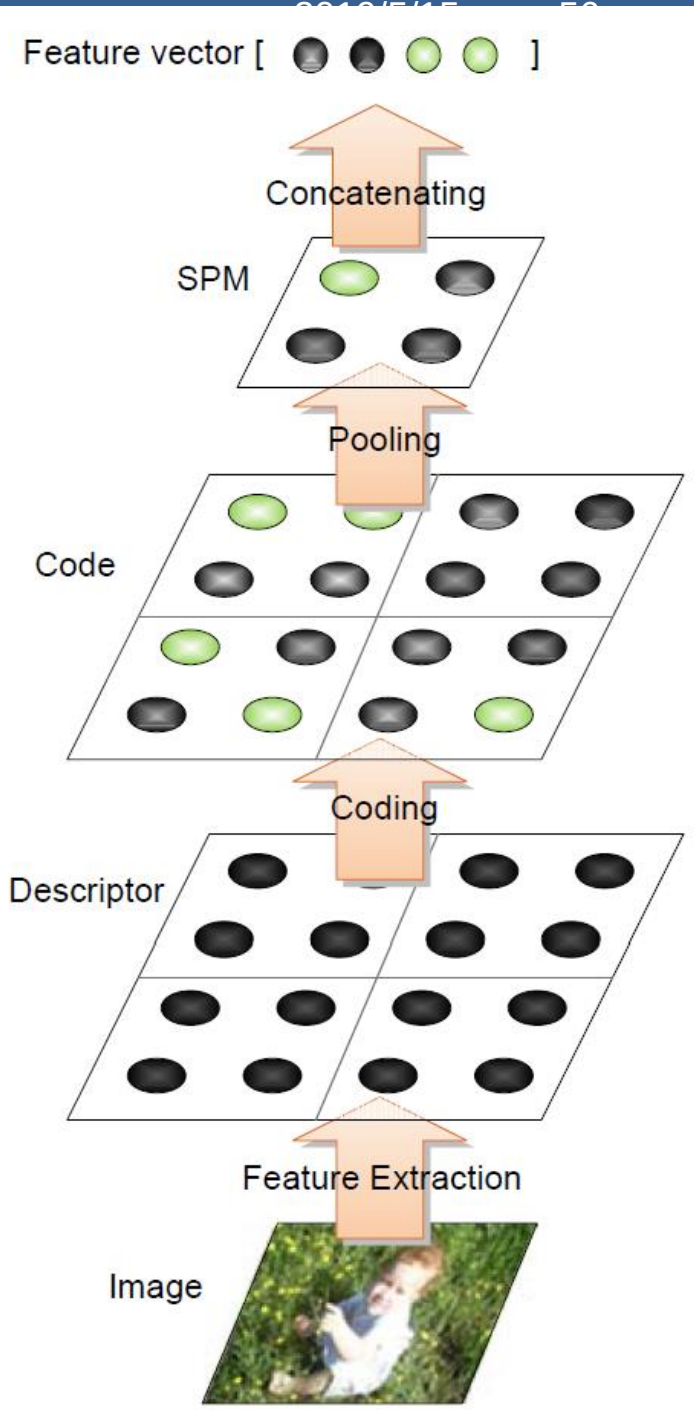


大牛们发现，不仅图像存在这个规律，声音也存在。他们从未标注的声音中发现了20种基本的声音结构，其余的声音可以由这20种基本结构合成。



稀疏编码的图像分类
Sparse coding

- **Feature**
  - Harris detector、salient region detector...
- **Descriptor**
  - SIFT、color moment
- **Code**
  - 一个或多个特征排列起来是一个vector；再聚类得到Codebook；把每个descriptor用码本中的code表示出来



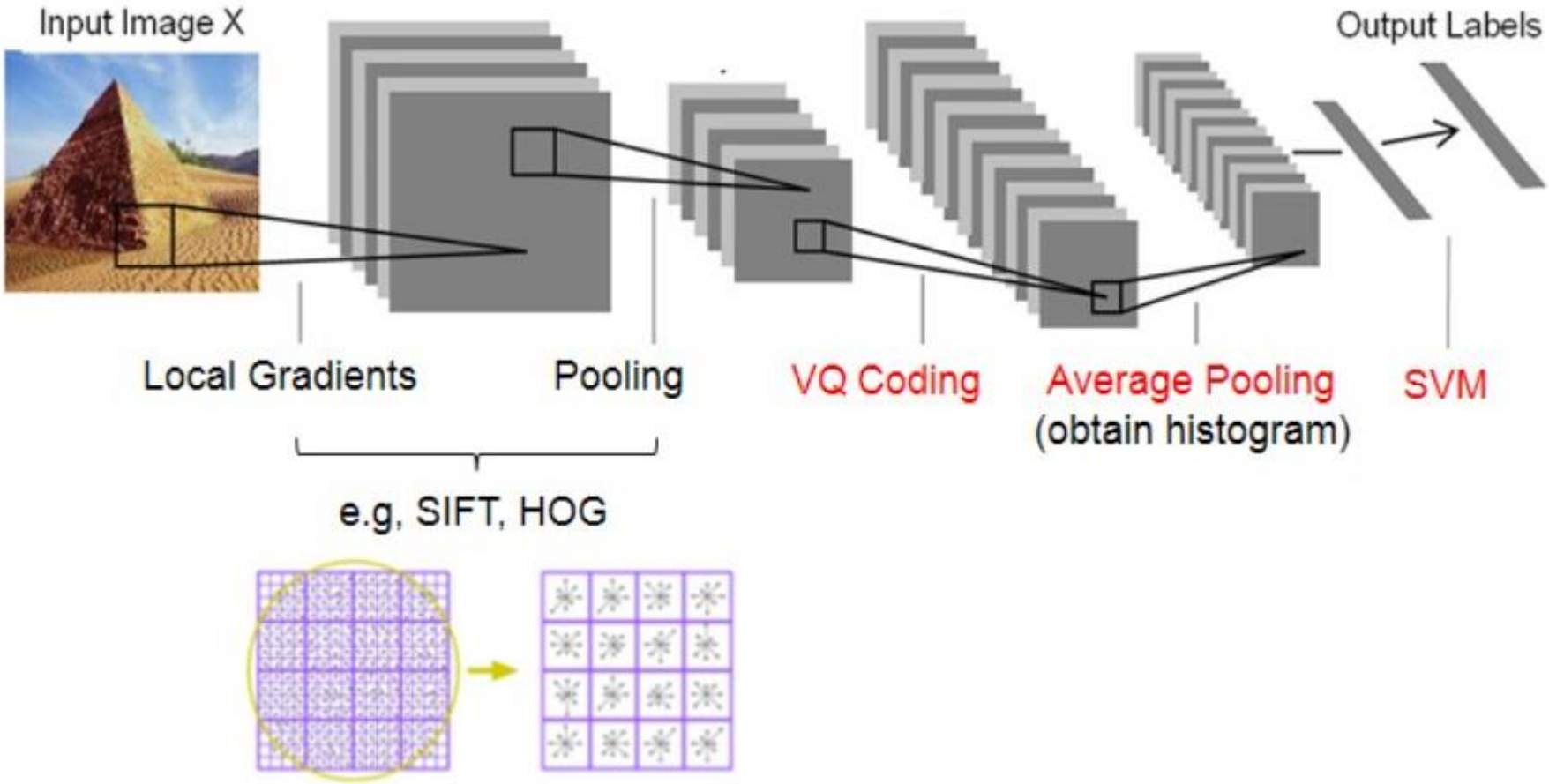
Picture Source:
Locality-constrained Linear Coding for image classification
Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, Yihong Gong
Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on
(Wang, Yang et al. 2010)

Coding & Pooling

两层Coding和Pooling的结构

Coding: nonlinear mapping data into another feature space

Pooling: obtain histogram



第一层coding+pooling

什么是PM(pyramid matching)?

K. Grauman and T. Darrell. Pyramid match kernels: Discriminative classification with sets of image features. In *Proc. ICCV*, 2005.

设X、Y为*d*维特征空间中的点集（ sets of vectors）。将特征空间划分为不同的尺度*0*, . . . , *L*，在尺度*l*下特征空间的每一维划出2^l个cells，，那么*d*维的特征空间就能划出2^{dl}个bins。令H_X^l 和 H_Y^l为尺度*l*下X和Y的直方图，则$H_X^l(i)$和$H_Y^l(i)$为X、Y 中点落入第i个cell的数目。那么在尺度*l*上匹配点的个数I^l可通过直方图相交来求：

$$\mathcal{I}^\ell = \mathcal{I}(H_X^\ell, H_Y^\ell) = \sum_{i=1}^{D} \min\left(H_X^\ell(i), H_Y^\ell(i)\right)$$

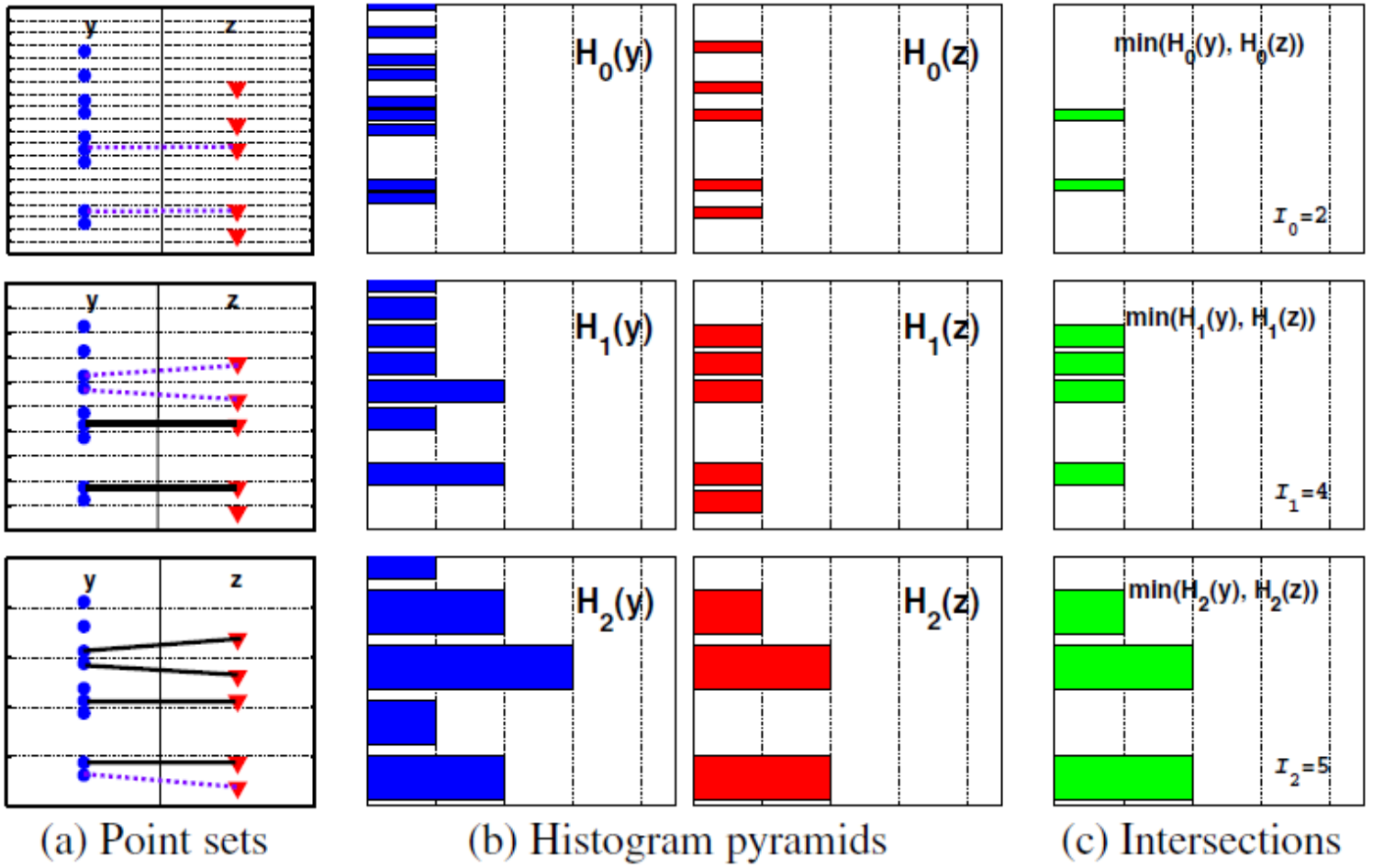
由于细粒度的bin被大粒度的bin所包含，为了不重复计算，每个尺度的有效Match定义为match的增量 $I^l - I^{l+1}$, $l=0,\ldots,L-1$。而不同的尺度下的match应赋予不同权重，显然大尺度的权重小，而小尺度的权重大，因此定义权重为$1/(2^{L-l})$。最终，两点集匹配的程度定义为（即pyramid match kernel）：

$$\begin{aligned}\kappa^L(X,Y) &= \mathcal{I}^L + \sum_{\ell=0}^{L-1} \frac{1}{2^{L-\ell}}\left(\mathcal{I}^\ell - \mathcal{I}^{\ell+1}\right) \\ &= \frac{1}{2^L}\mathcal{I}^0 + \sum_{\ell=1}^{L} \frac{1}{2^{L-\ell+1}}\mathcal{I}^\ell\end{aligned}$$

←不同尺度直方图交的加权和

we can implement K^L as a single histogram intersection of “long” vectors formed by concatenating the appropriately weighted histograms of all channels at all resolutions

第一层coding+pooling PM(pyramid matching)
不同尺度下特征点匹配示意图



K. Grauman and T. Darrell. Pyramid match kernels: Discriminative classification with sets of image features. In *Proc. ICCV*, 2005.

Coding & Pooling

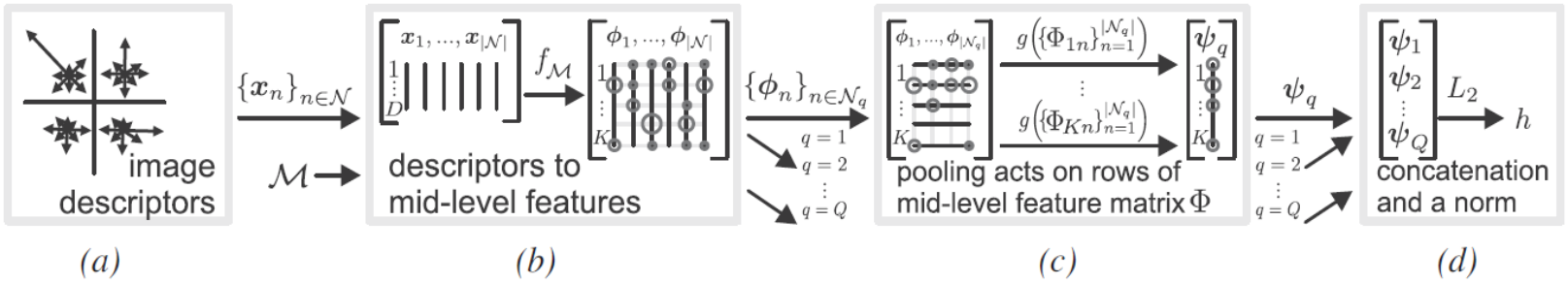
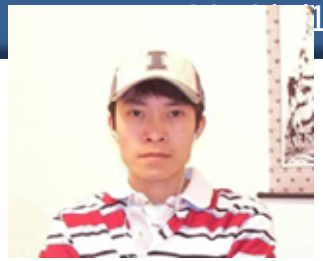


Fig. 1. Overview of Bag-of-Words showing mid-level coding and pooling steps. (a) $|\mathcal{N}|$ local descriptors of dimension D are extracted from an image. (b) Mid-level coding embeds the descriptors into the visual vocabulary space using K visual words from dictionary $\mathcal{M}$. Circles of various sizes illustrate values of mid-level coefficients. (c) Mid-level features of partition q are stacked. Next, pooling aggregates the values along rows and forms a single vector per spatial partition. (d) Vectors from all partitions are concatenated and normalised to form signature h.

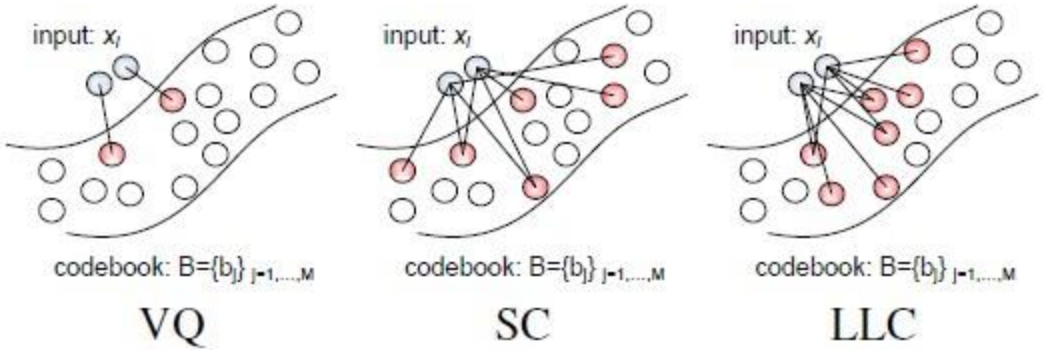
- mid-level coding: embed local descriptors into the visual vocabulary space.
- Pooling: aggregate mid-level features into vectors.
  - pooling目的是为了保持某种不变性（旋转、平移、伸缩等），常用的有mean-pooling，max-pooling和Stochastic-pooling三种。
  - pooling的结果是使得特征减少，参数减少。

Comparison of Mid-Level Feature Coding Approaches And Pooling Strategies in Visual Concept Detection
P. Koniusz, F. Yan, K. Mikolajczyk
CVIU, 117(5):479-492, 2013


Coding：思路的发展

- Hard voting（或称vector quantization, VQ）
  - each descriptor is represented by its nearest code
    - Lazebnik, S., C. Schmid and J. Ponce (2006). "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories." CVPR.
- soft voting
  - a descriptor is represented by multiple codes
    - Yang, J., K. Yu, Y. Gong and T. Huang (2009). "Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification." CVPR.
- Reconstruction based method
  - sparse coding: chooses a group of codes to reconstruct a descriptors plus a constraint to the number of codes.
    - Wang, J., J. Yang, K. Yu, F. Lv, T. Huang and Y. Gong (2010). "Locality-constrained Linear Coding for Image Classification."



Coding：三种思路的具体实现

X: *D*-dimensional local descriptor

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N] \in \mathbb{R}^{D \times N}$$

B: a codebook with M entries

$$\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_M] \in \mathbb{R}^{D \times M}$$

$$V = \arg\min ||x - VB^T||_2 + \lambda \sum_i ||v_i||_1 ||x - b_i||_2$$

$$s.t. \sum_i^M v = 1$$

邻近性
locality constraint

$$V = \arg\min ||x - VB^T||_2 + \lambda ||V||_1$$

$$s.t. \sum_i^M v_i = 1$$

- 在Codebook中找出一个Code
- 用Codebook中的Code线性组合
- 所选Codebook中用于线性组合的Code满足邻近性

$$v_i = \begin{cases} 1, & \text{if } i = \arg\min_j(||x - b_j||_2) \\ 0, & otherwise \end{cases}$$

Coding：不同实现方法的差异(不考虑Pooling的差异)

Hard voting →

Hard voting →

locality →

Table 1. Image classification results on Caltech-101 dataset

training images	5	10	15	20	25	30
Zhang [25]	46.6	55.8	59.1	62.0	-	66.20
Lazebnik [15]	-	-	56.40	-	-	64.60
Griffin [11]	44.2	54.5	59.0	63.3	65.8	67.60
Boiman [2]	-	-	65.00	-	-	70.40
Jain [12]	-	-	61.00	-	-	69.10
Gemert [8]	-	-	-	-	-	64.16
Yang [22]	-	-	**67.00**	-	-	73.20
Ours	**51.15**	**59.77**	65.43	**67.74**	**70.16**	**73.44**

Hard voting →

locality →

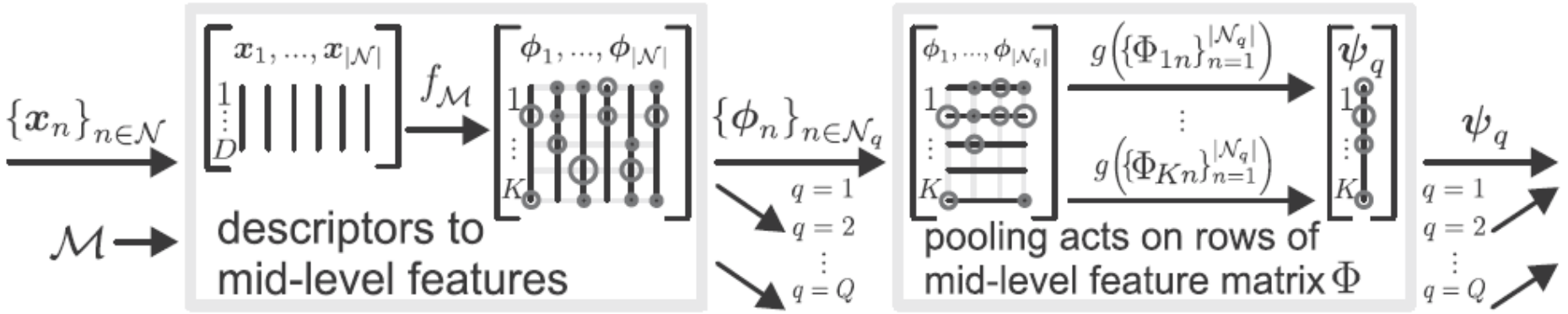
Table 2. Image classification results using Caltech-256 dataset

training images	15	30	45	60
Griffin [11]	28.30	34.10	-	-
Gemert [8]	-	27.17	-	-
Yang [22]	27.73	34.02	37.46	40.14
Ours	**34.36**	**41.19**	**45.31**	**47.68**

Locality-constrained Linear Coding for image classification
Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, Yihong Gong
Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on

Pooling的常用方法

Comparison of Mid-Level Feature Coding Approaches And Pooling Strategies in Visual Concept Detection
P. Koniusz, F. Yan, K. Mikolajczyk CVIU, 117(5):479-492, 2013



$$\psi_k = \frac{1}{|\mathcal{N}|}\sum_{n\in\mathcal{N}} \Phi_{kn}$$

$$\psi_k = \max\left(\{\Phi_{kn}\}_{n\in\mathcal{N}}\right)$$

$$\psi_k = \left(\frac{1}{|\mathcal{N}|}\sum_{n\in\mathcal{N}} |\Phi_{kn}|^p\right)^{1/p}$$

$$\psi_k = \left(\frac{1}{|\mathcal{N}|}\sum_{n\in\mathcal{N}} \Phi_{kn}\right)^{\gamma}$$

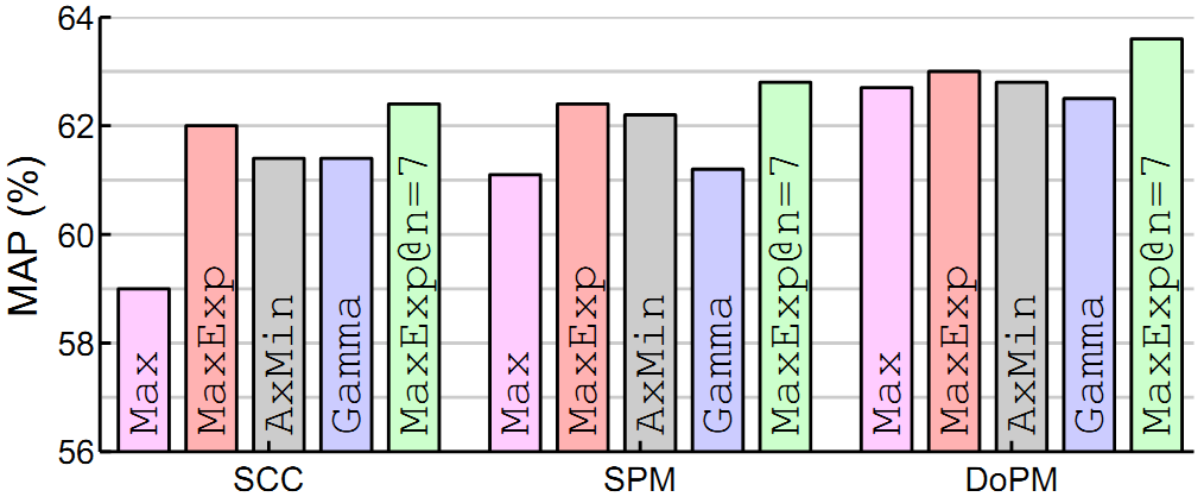


Figure 2: Evaluation of SCC, SPM, and DoPM schemes on the PascalVOC07 set given Max-pooling, MaxExp, AxMin, Gamma, and MaxExp@$n = 7$. The dictionary sizes are 40000, 32000, and 24000 atoms for SCC, SPM, and DoPM.

图像分类的算法思想

- 从文本分类→图像分类
  - 如何从图像中获取全局特征？
    - 颜色特征、纹理特征、形状特征
  - 如何从图像中获取局部特征？
    - SIFT: Scale-invariant feature transform
- 图像分类的几个发展阶段
  - Low-level Modelling
  - Semantic Modelling
  - Sparse Coding
  - Deep learning

图像识别领域的突破

72%, 2010

74%, 2011

85%, 2012

89%, 2013

93%, 2014

95%, 2015

ImageNet Challenge

http://www.image-net.org/ Imagenet Large Scale Visual Recognition Challenge (ILSVRC)

This challenge evaluates algorithms for object detection and image classification at large scale.

2017: Object localization for 1000 categories.

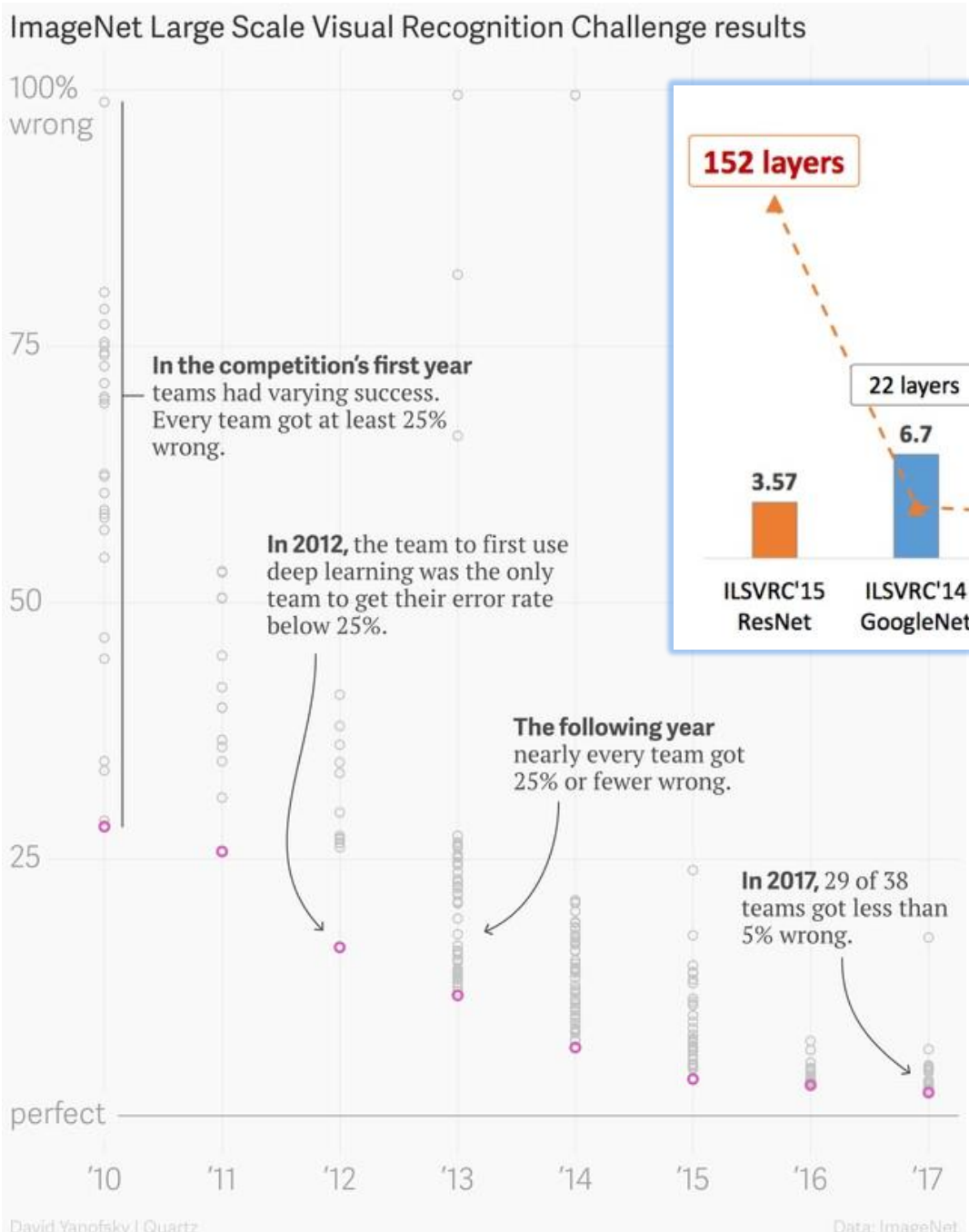
Object detection for 200 fully labeled categories.

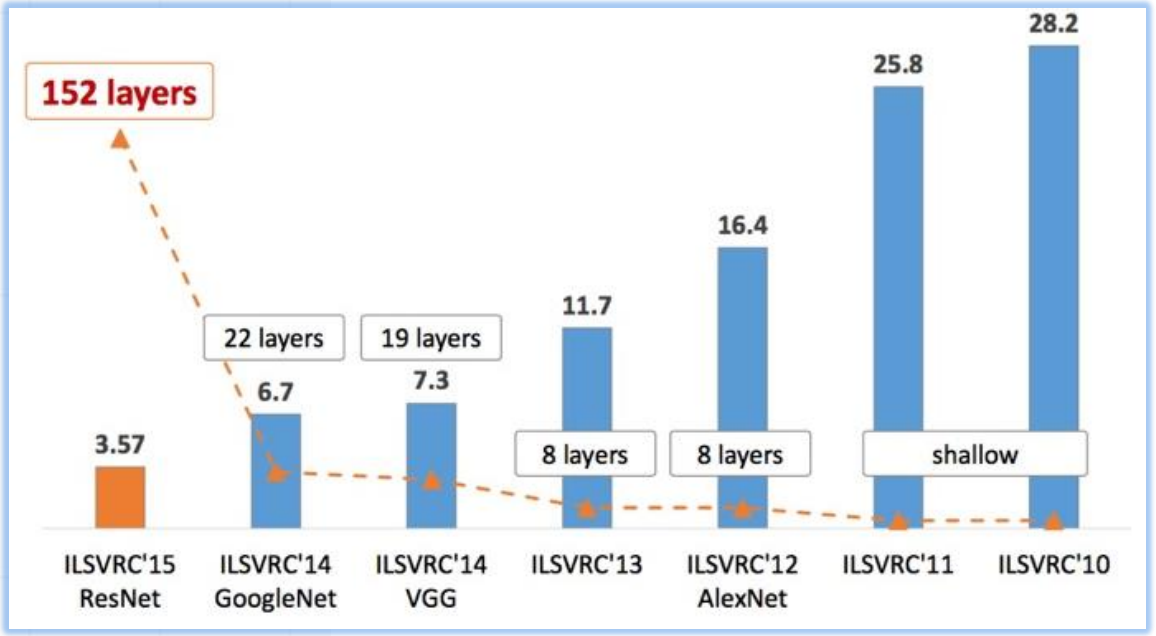
Object detection from video for 30 fully labeled categories.

ILSVRC(ImageNet Large Scale Visual Recognition Challenge)

ImageNet历年竞赛题目

- 2017
  I. Object localization for 1000 categories.
  II. Object detection for 200 fully labeled categories.
  III. Object detection from video for 30 fully labeled categories.
- 2016
  I. Object localization for 1000 categories.
  II. Object detection for 200 fully labeled categories.
  III. Object detection from video for 30 fully labeled categories.
  IV. Scene classification for 365 scene categories
  V. Scene parsing[New] for 150 stuff and discrete object categories
- 2015
  - Main competitions: I: Object detection II: Object localization
  - Taster competitions: I: Object detection from video II: Scene classification
- 2014, 2013, 2012, 2011, 2010
  - Task 1: Detection
  - Task 2: Classification and localization
  - Task 3: Fine-grained classification 【仅2012】





ILSVRC历届冠军论文笔记
https://blog.csdn.net/kangroger/article/details/56522132
#Deep Learning回顾#之LeNet、AlexNet、GoogLeNet、VGG、ResNet
https://zhuanlan.zhihu.com/p/22094600
https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/

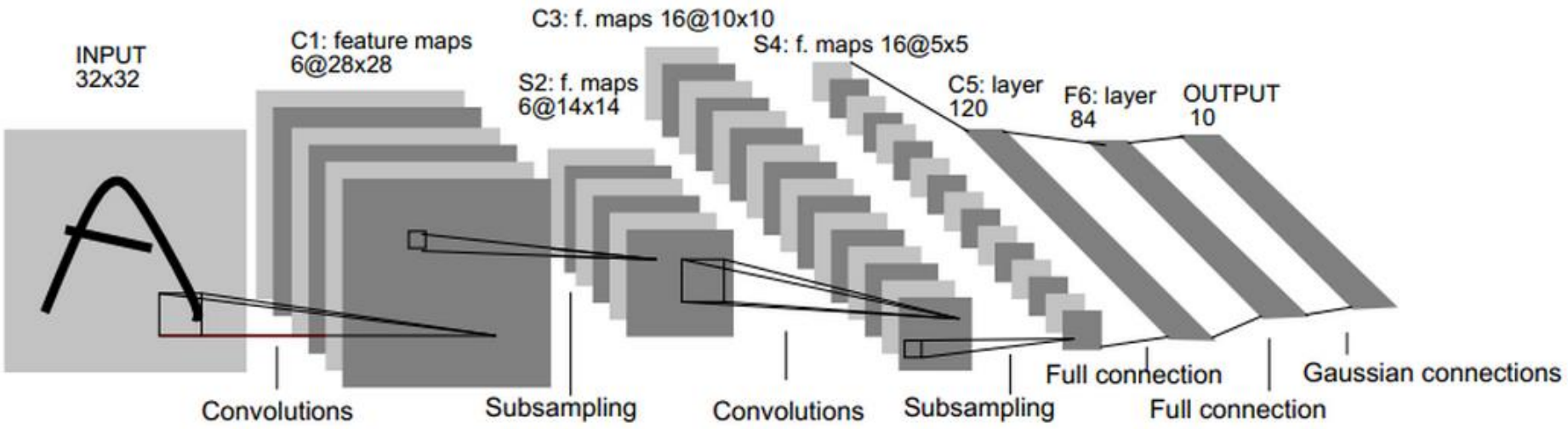
AlexNet、VGG、GoogLeNet、ResNet对比

模型名	AlexNet	VGG	GoogLeNet	ResNet
初入江湖	2012	2014	2014	2015
层数	8	19	22	152
Top-5错误	16.4%	7.3%	6.7%	3.57%
Data Augmentation	+	+	+	+
Inception(NIN)	–	–	+	–
卷积层数	5	16	21	151
卷积核大小	11,5,3	3	7,1,3,5	7,1,3,5
全连接层数	3	3	1	1
全连接层大小	4096,4096,1000	4096,4096,1000	1000	1000
Dropout	+	+	+	+
Local Response Normalization	+	–	+	–
Batch Normalization	–	–	–	+

卷积神经网络 Convolutional Neural Networks (CNN)

First proposed by Fukushima in 1980
Improved by LeCun, Bottou, Bengio and Haffner in 1998

A Convolutional Neural Network (CNN) is comprised of one or more convolutional layers (often with a subsampling step) and then followed by one or more fully connected layers as in a standard multilayer neural network.



LeNet-5系统结构：用于文字识别的7层卷积网络

20世纪60年代，Hubel和Wiesel在研究猫脑皮层中用于局部敏感和方向选择的神经元时发现其独特的网络结构可以有效地降低反馈神经网络的复杂性，继而提出了CNN。

Gradient Based Learning Applied to Document Recognition， Yann LeCun等, 1998 2015.05 Google cited: 3313

ILSVRC 2012 AlexNet

Imagenet classification with deep convolutional neural networks

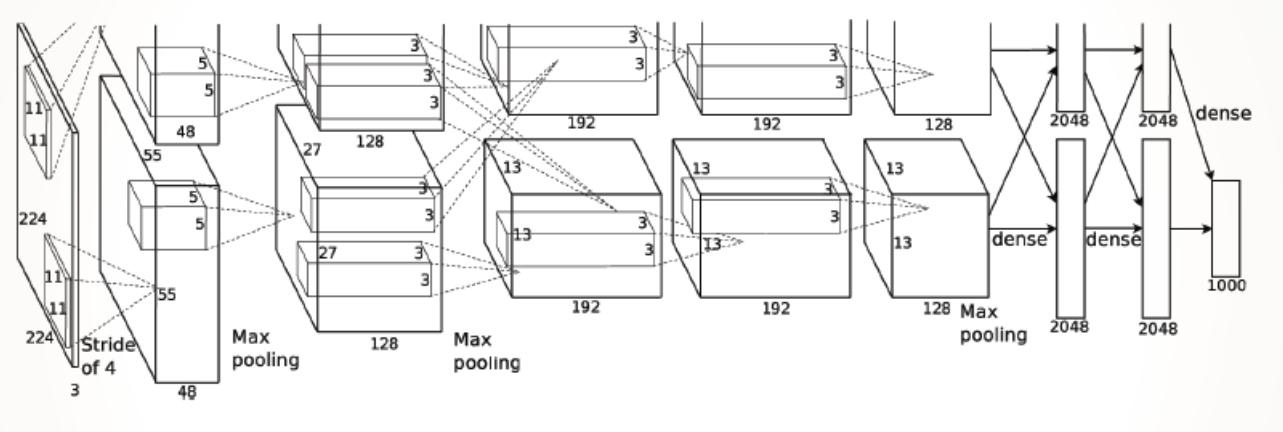
A Krizhevsky, I Sutskever, GE Hinton - Advances in neural ..., 2012 - papers.nips.cc

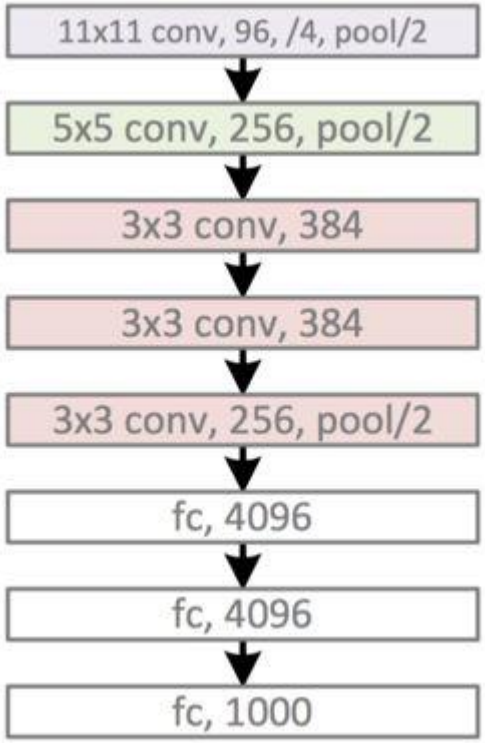
We trained a large, deep convolutional neural network to classify the 1.3 million high-resolution images in the LSVRC-2010 ImageNet training set into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 39.7\% and 18.9\% which is considerably better than the previous state-of-the-art results. The neural network, which has 60 million parameters and 500,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and two globally connected layers with a final ...

☆ 99 被引用次数：22259 相关文章 所有 89 个版本 »

Retrieved: 2018-04-23

(1) Data Augmentation
(2) Dropout
(3) ReLU激活函数
(4) Local Response Normalization
(5) Overlapping Pooling
(6) 多GPU并行

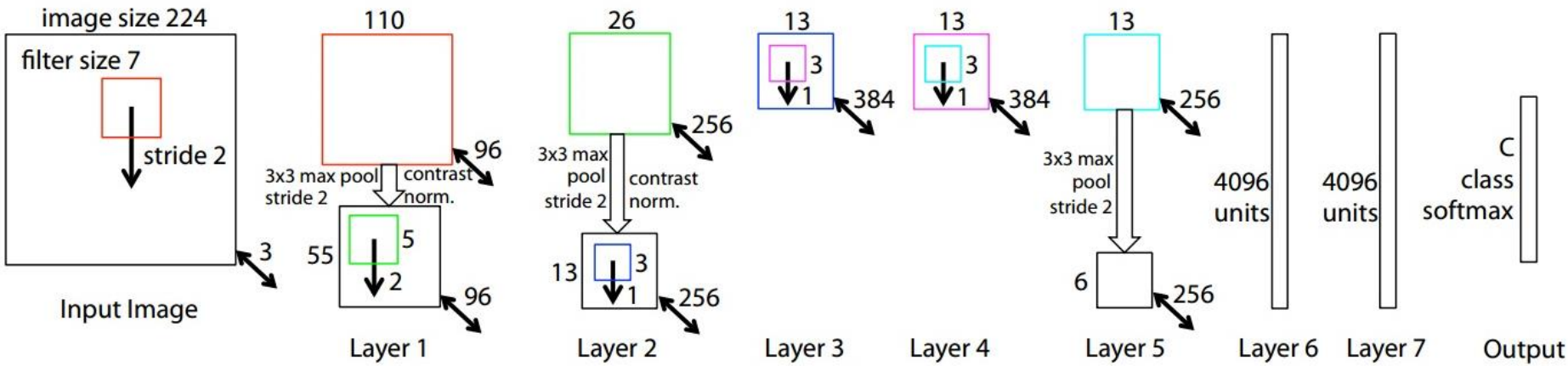




#Deep Learning回顾#之LeNet、AlexNet、GoogLeNet、VGG、ResNet
https://zhuanlan.zhihu.com/p/22094600
https://www.cnblogs.com/52machinelearning/p/5821591.html

ILSVRC 2013
ZFNet

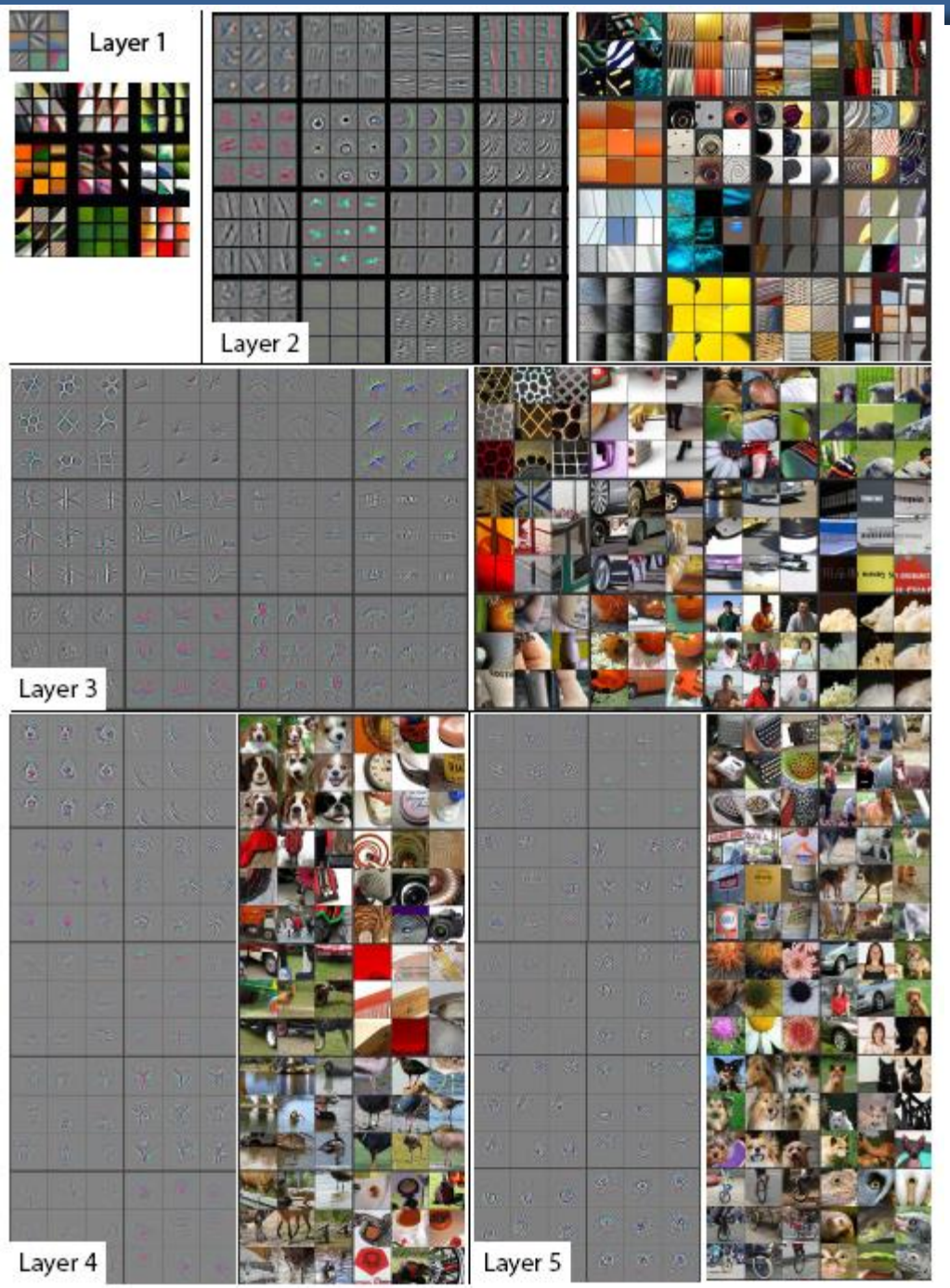
ZFNet的网络结构，是在AlexNet上进行了微调



ZFNet解释了为什么CNNs有效、怎么提高CNN性能。其主要贡献在于：

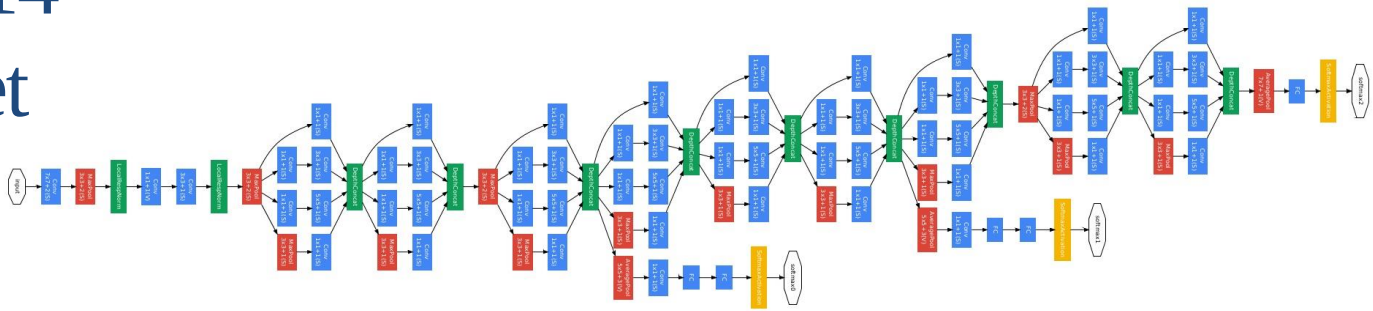
- 使用了反卷积，可视化feature map。通过feature map可以看出，前面的层学习的是边缘、颜色、纹理等，后面的层学习的是类别相关的抽象特征。
- 相比AlexNet，前面的层使用了更小的卷积核和更小的步长，保留了更多特征。
- 通过遮挡，找出了决定图像类别的关键部位。
- 通过实验，说明了深度增加时，网络可以学习到更好的特征。

ZFNet



Visualizing and Understanding Convolutional Networks
https://arxiv.org/pdf/1311.2901.pdf

ILSVRC 2014 GoogleLeNet

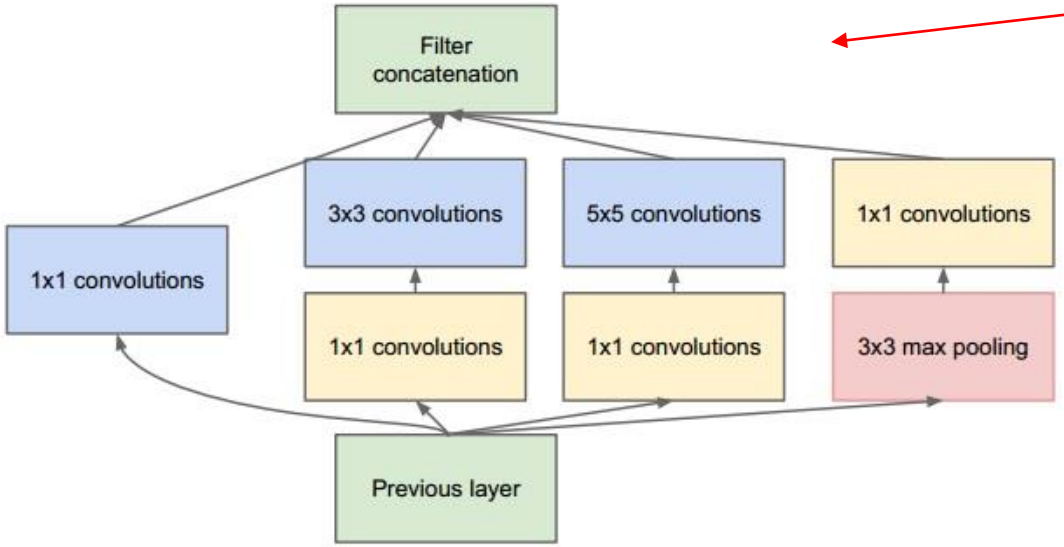


GoogLeNet, Top5 的错误率降低到6.67%. 一个22层的深度网络。

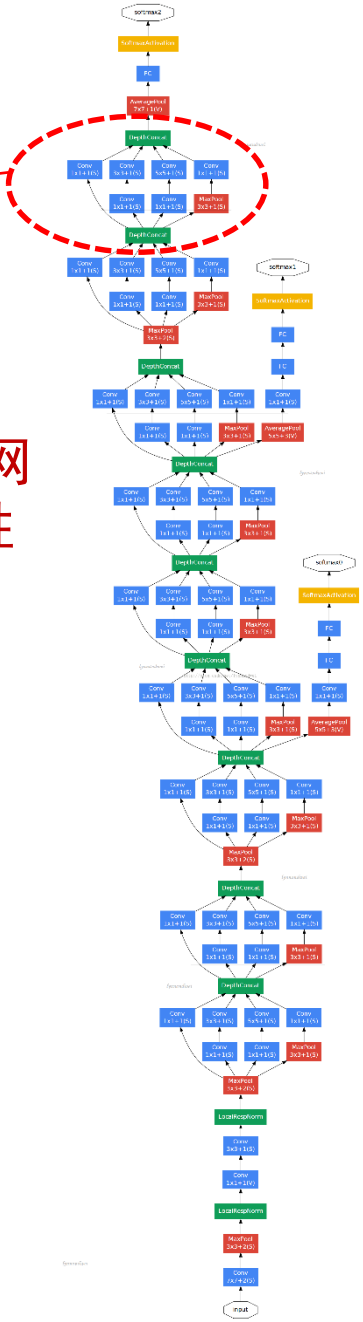
type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

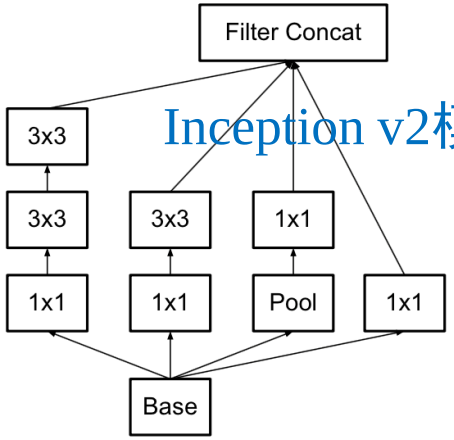
Going deeper with convolutions http://arxiv.org/pdf/1409.4842v1.pdf

Inception
NIN(Network in Network)

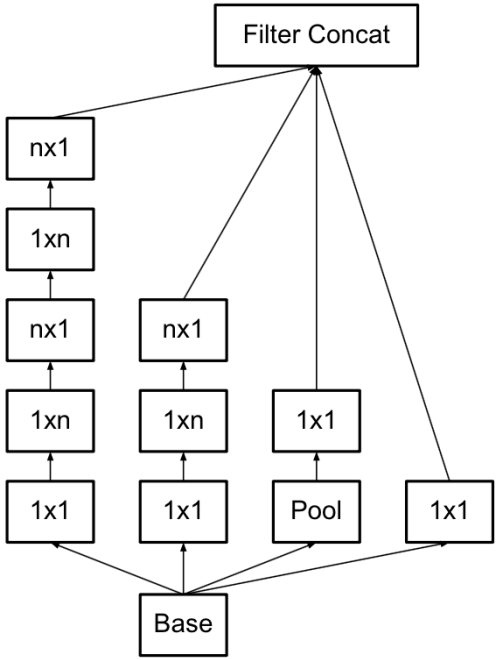


Inception v1模型：既增加了网络的width，又增加了网络对尺度的适应性





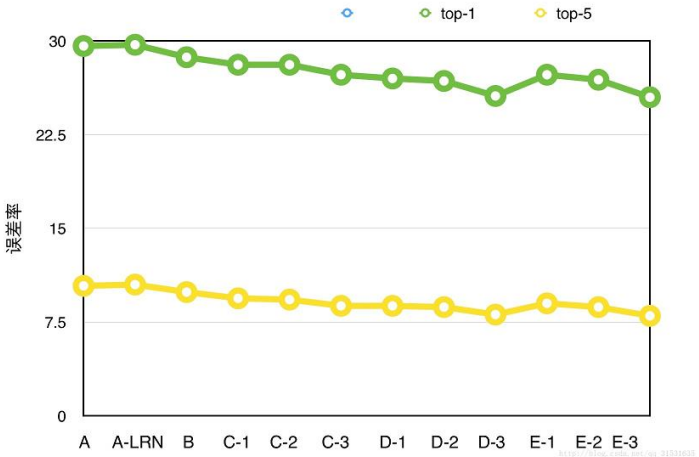
Inception v2模型



深入浅出——网络模型中Inception的作用与结构全解析
https://blog.csdn.net/u010402786/article/details/52433324

ILSVRC 2014 VGGNet

VGGNet是Oxford大学提出的，目的是研究深度对卷积网络的影响。使用简单的3x3卷积，不断重复卷积层，最后经过全连接、池化、softmax，得到输出类别概率。



VGGNET共有6种不同类型配置，命名为A-E，深度从11到19；每个卷积层的depth，从一开始的64到最后的512。

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

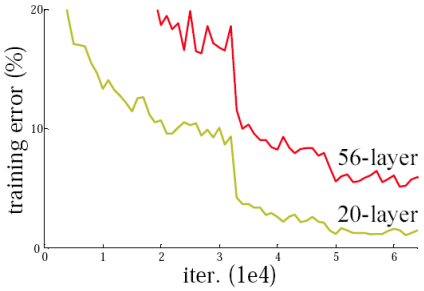
Table 2: **Number of parameters** (in millions).

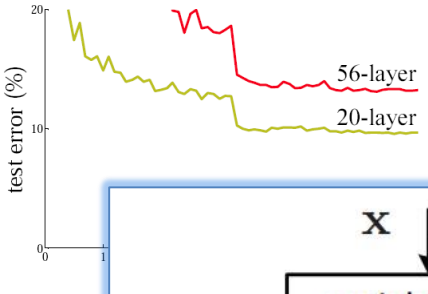
Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

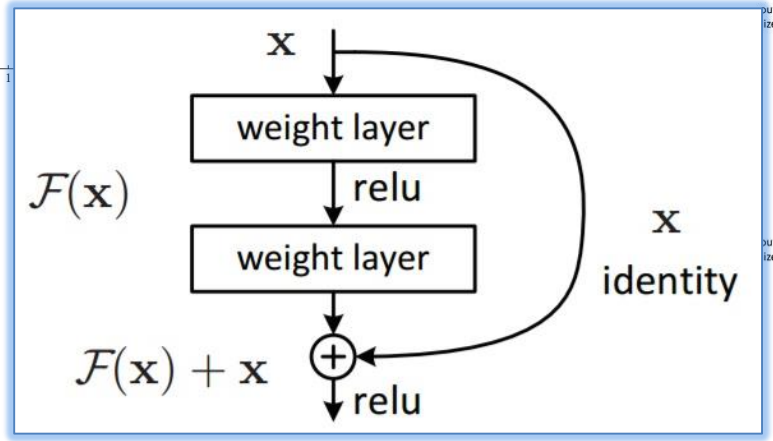
ILSVRC 2015

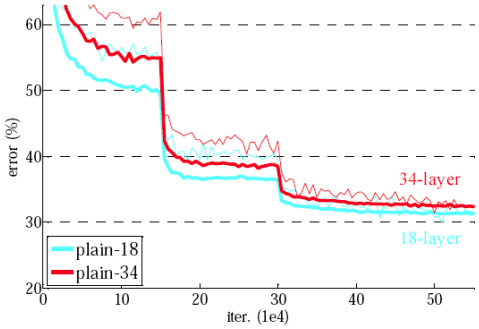
ResNet

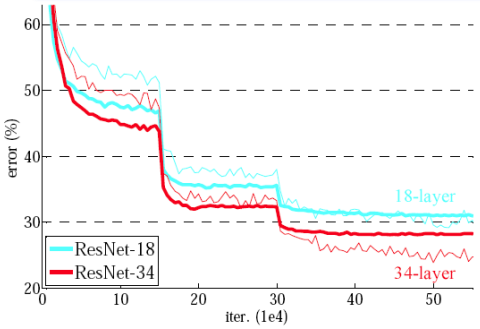
“退化”问题：当模型的层次加深时，错误率却提高了

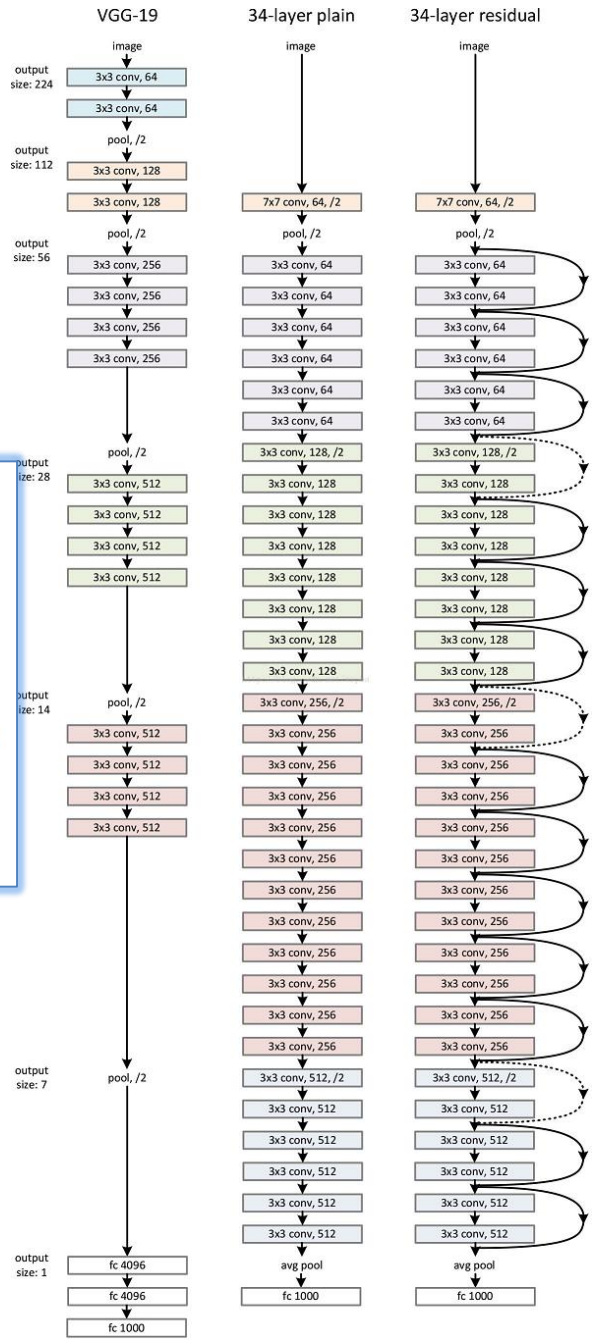








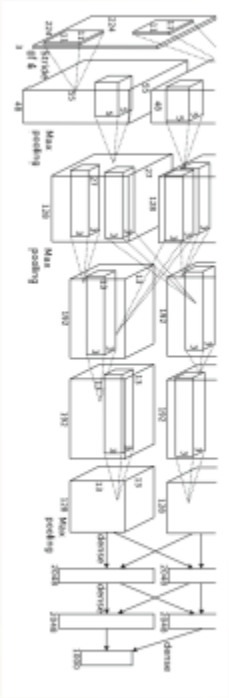




Deep Residual Learning for Image Recognition
http://arxiv.org/pdf/1512.03385.pdf

越来越深？越宽？

“AlexNet”



[Krizhevsky et al. NIPS 2012]

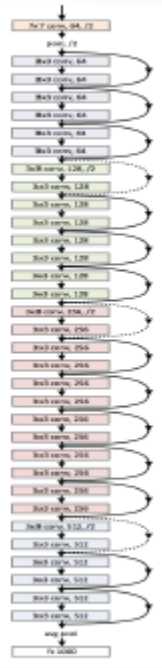
“GoogLeNet”

[Szegedy et al. CVPR 2015]

“VGG Net”

[Simonyan & Zisserman, ICLR 2015]

“ResNet”



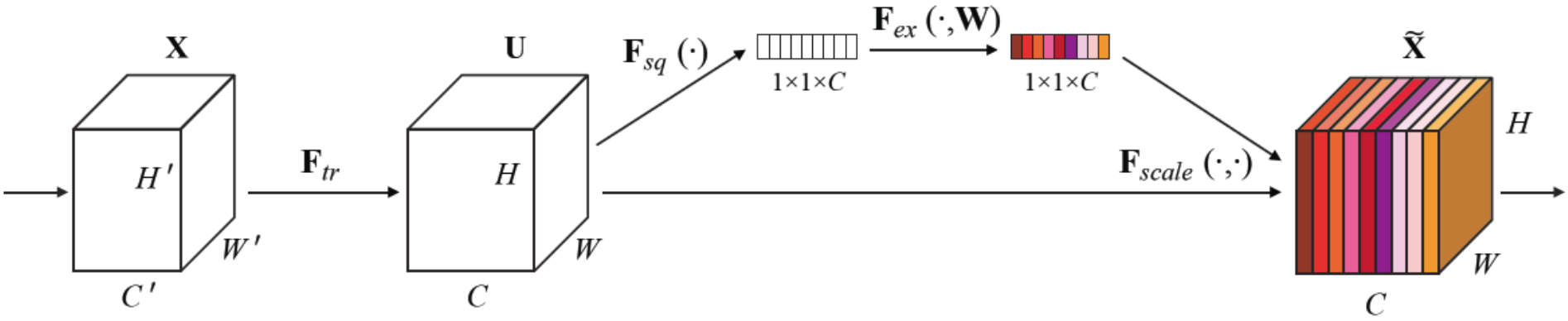
[He et al. CVPR 2016]

http://image-net.org/challenges/LSVRC/2016/results

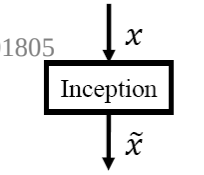
ILSVRC 2017

SENet Momenta & Oxford

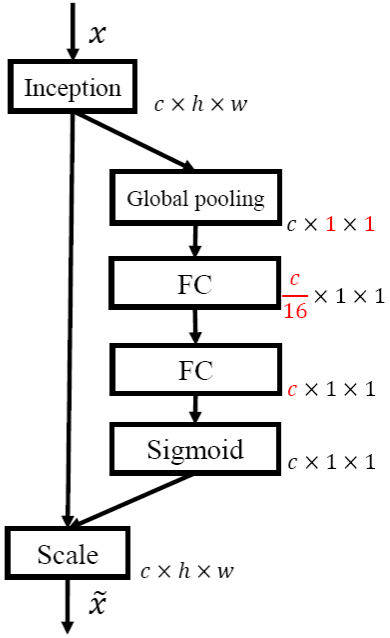
Squeeze-and-Excitation模块



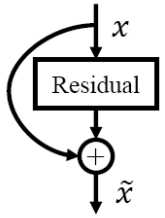
Momenta详解ImageNet 2017夺冠架构SENet
www.sohu.com/a/161633191_465975 cited:201805
Squeeze-and-Excitation Networks
https://arxiv.org/pdf/1709.01507.pdf



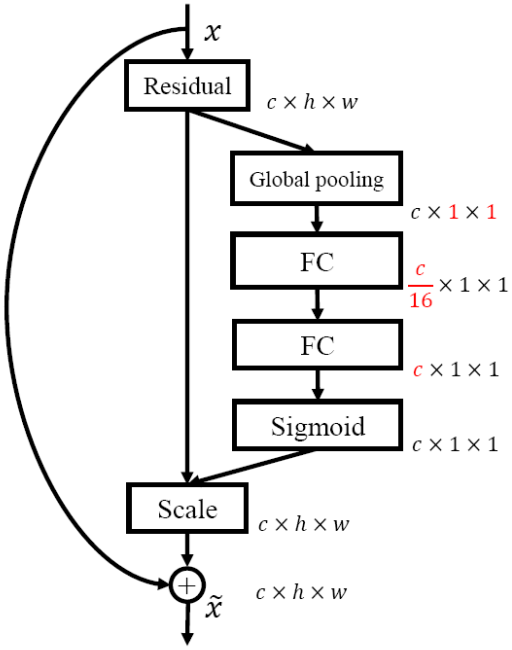
Inception Module



SE-Inception Module



ResNet Module

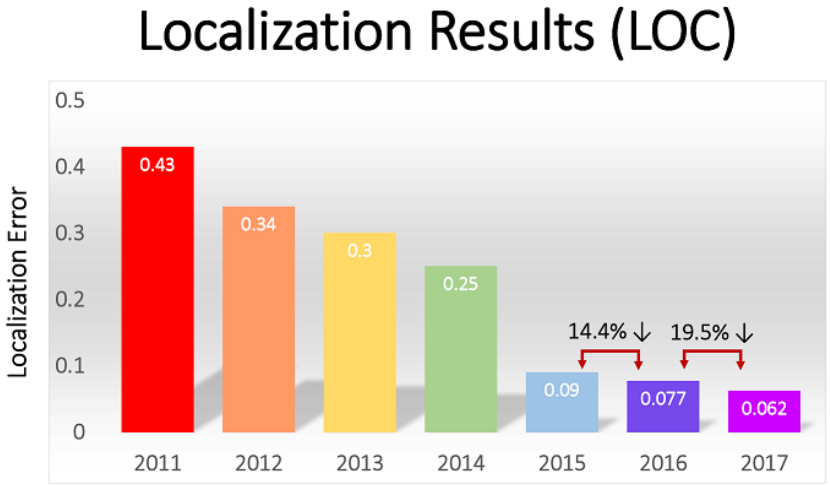


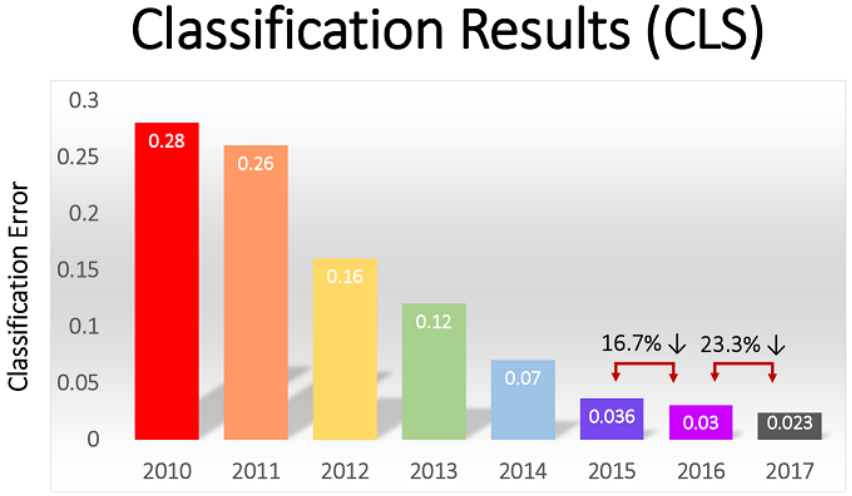
SE-ResNet Module

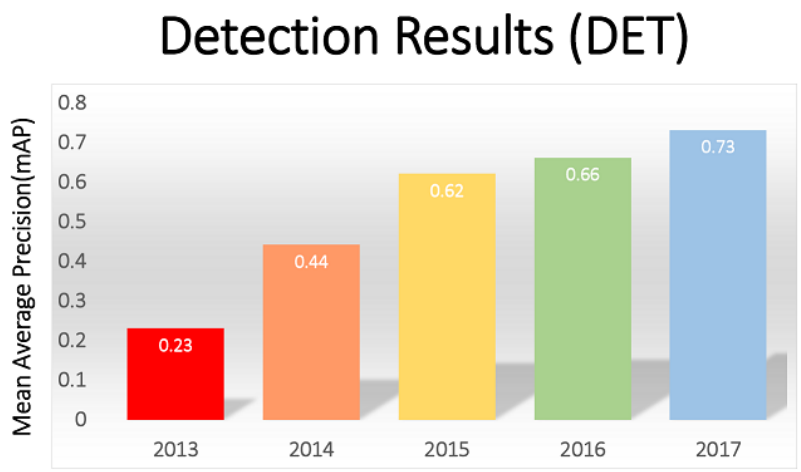
ILSVRC 2017 Classification Task

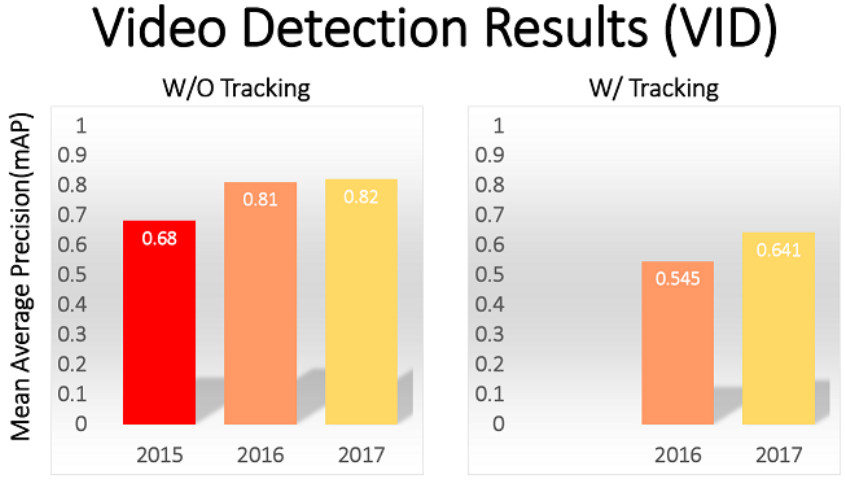
Team	Top-5 error (%)
WMW	2.251
Trimps-Soushen	2.481
NUS-Qihoo-DPNs	2.740
BDAT	2.962
ILSVRC 2016 Winner	2.991

8年来：ILSVRC









http://image-net.org/challenges/talks_2017/ILSVRC2017_overview.pdf

8年来：ILSVRC

- 2012: AlexNet
- 2013: ZFNet
- 2014: GoogleNet, VGGNet
- 2015: ResNet
- 2016
  - CUImage（商汤科技和港中文）：标检测第一；
  - Trimps-Soushen（公安部三所）：目标定位第一；
  - CUvideo（商汤和港中文）：视频中物体检测子项目第一；
  - NUIST（南京信息工程大学）：视频中的物体探测两个子项目第一；
  - HikVision（海康威视）：场景分类第一；
  - SenseCUSceneParsing（商汤和港中文）：场景分析第一。
- 2017
  - 在物体检测（object detection）、物体定位、视频物体检测三个大类中，南京信息工程大学和帝国理工学院组成的 BDAT 团队、新加坡国立大学与奇虎360合作团队、伦敦帝国理工学院和悉尼大学合作的团队分别拿下冠军。

Kaggle是安东尼·高德布卢姆（Anthony Goldbloom）2010年在墨尔本创立的，主要为开发商和数据科学家提供举办机器学习竞赛、托管数据库、编写和分享代码的平台。2017年3月8日，谷歌云计算开发者大会上，谷歌云首席科学家李飞飞宣布了一则重大消息：**谷歌收购 Kaggle**。

COCO
Common Objects in Context

Home People Dataset Tas

Tasks Evalua

- Detection 2018
- Keypoints 2018
- Stuff 2018
- Panoptic 2018
- Detection 2017
- Keypoints 2017
- Stuff 2017
- Detection 2016
- Keypoints 2016
- Detection 2015
- Captioning 2015

News

- We are pleased to announce the COCO 2018 Detection and Keypoint Ch
- This year we will also be hosting a new COCO Panoptic Segmentation C
- Results to be announced at the Joint COCO and Mapillary Recognition E
- This website is now hosted on Github, which provides page source and h

What is COCO?

COCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:

- Object segmentation
- Recognition in context
- Superpixel stuff segmentation
- 330K images (>200K labeled)
- 1.5 million object instances
- 80 object categories
- 91 stuff categories
- 5 captions per image
- 250,000 people with keypoints

COCO Challenge 关注整体图像的 ImageNet 图像分类任务相比，COCO 中的物体检测任务更关注的是场景理解中的物体识别，需对图像中出现的每个物体的个体（如各种小物体，各种遮挡物体）进行识别，因此要求算法对图像细节有更好的理解。

Larry Zitnick FAIR

Piotr Dollár FAIR

http://cocodataset.org/

COCO 2018 Object Detection Task

COCO 2018 Keypoint Detection Task

COCO 2018 Stuff Segmentation Task

COCO 2015 Image Captioning Task

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

COCO 2018 Panoptic Segmentation Task

COCO 2018 Object Detection Task
COCO 2018 Keypoint Detection Task
COCO 2018 Stuff Segmentation Task
COCO 2018 Panoptic Segmentation Task
COCO 2015 Image Captioning Task

COCO2017 Challenge Winners

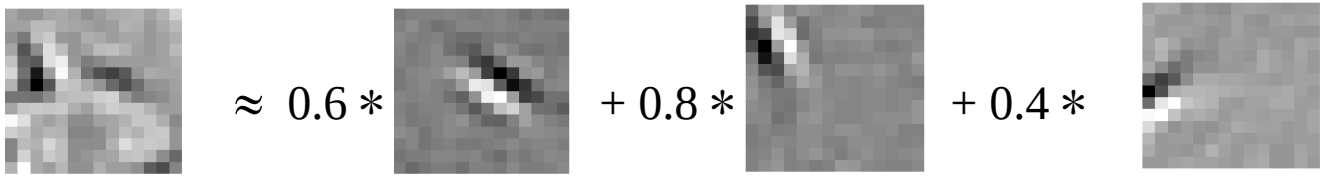
https://places-coco2017.github.io/#winners

	1st place	2nd place	3rd place	4th place
COCO Detection: Bounding Box	Megvii	UCenter	MSRA	FAIR
COCO Detection: Segmentation	UCenter	Megvii	FAIR	MSRA
COCO Keypoints	Megvii	Oks	Bangbangren	—
COCO Stuff	FAIR	G-RMI	Oxford	—
Places Instance Segmentation	Megvii	G-RMI	BlueSky	—
Places Scene Parsing	CASIA_IVA_JD	WinterIsComing	xdliang	—

MSRA：微软亚洲研究院
Megvii：旷世
UCenter：香港中文大学在读博士生、商汤科技研究员组成的团队
FAIR：Facebook AI Research
G-RMI：谷歌研究院

小结：图像分类的几个发展阶段

- Low-level Modelling
- Semantic Modelling
  - 语义目标、语义概念、语义属性
- Sparse Coding
  - Coding、Pooling



Represent as: [0, 0, …, 0, 0.6, 0, …, 0, 0.8, 0, …, 0, 0.4, …]

- Deep Learning