

基于内容的信息处理-一个新的挑战

张钹

清华信息科学与技术国家实验室
智能技术与系统国家重点实验室
北京清华大学信息科学与技术学院
计算机科学与技术系



目的

- 互联网时代，文本、图像与语音处理遇到怎样的共性科学问题？
- 基于这个科学问题，我们对信息处理会有何种新的理解？
- 从这个理解出发，信息处理又面临怎样的新的挑战？



信息

何谓信息

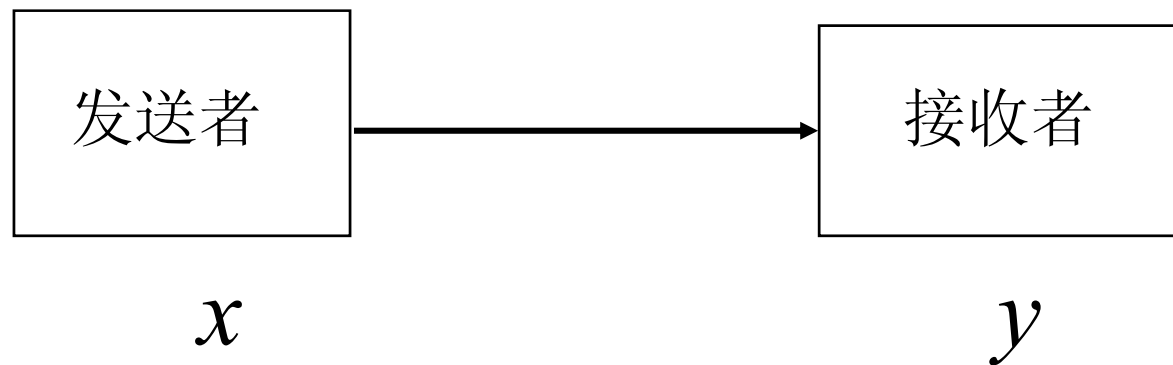
- 一组记录或传递消息的有序符号序列
- 文本、图像、语音以及任何事件的抽象形式

信息理论

是一门交叉学科，它涉及信息的分析、收集、分类、操作、存储、检索与传播

香农的通信理论 (1948)

通信模型



$$P(x) \rightarrow P(y|x) \rightarrow P(x|y)$$

(马尔科夫) 随机过程

-C. E. Shannon



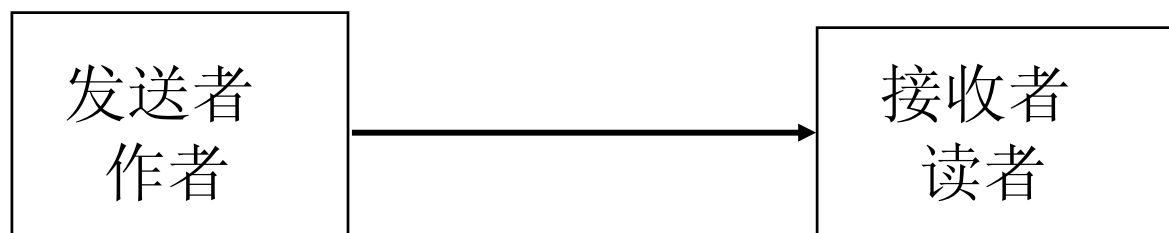
经典的信息论

- 通信的语义方面与（我们讨论的）工程问题无关
-C. E. Shannon
即经典信息理论不考虑信息内容（语义）
- 信息处理指：观察者以任意方法可检测到的信息变化

C. E. Shannon, A mathematical theory of communication, Bell System Technical Journal, vol. 27, pp.379-423, July, pp.623-656, October 1948



经典信息处理模型



$$x \rightarrow y \quad (y = x + e)$$

$$p(x) \rightarrow p(x|y)$$

一阶不确定性（错误、畸变、噪声）



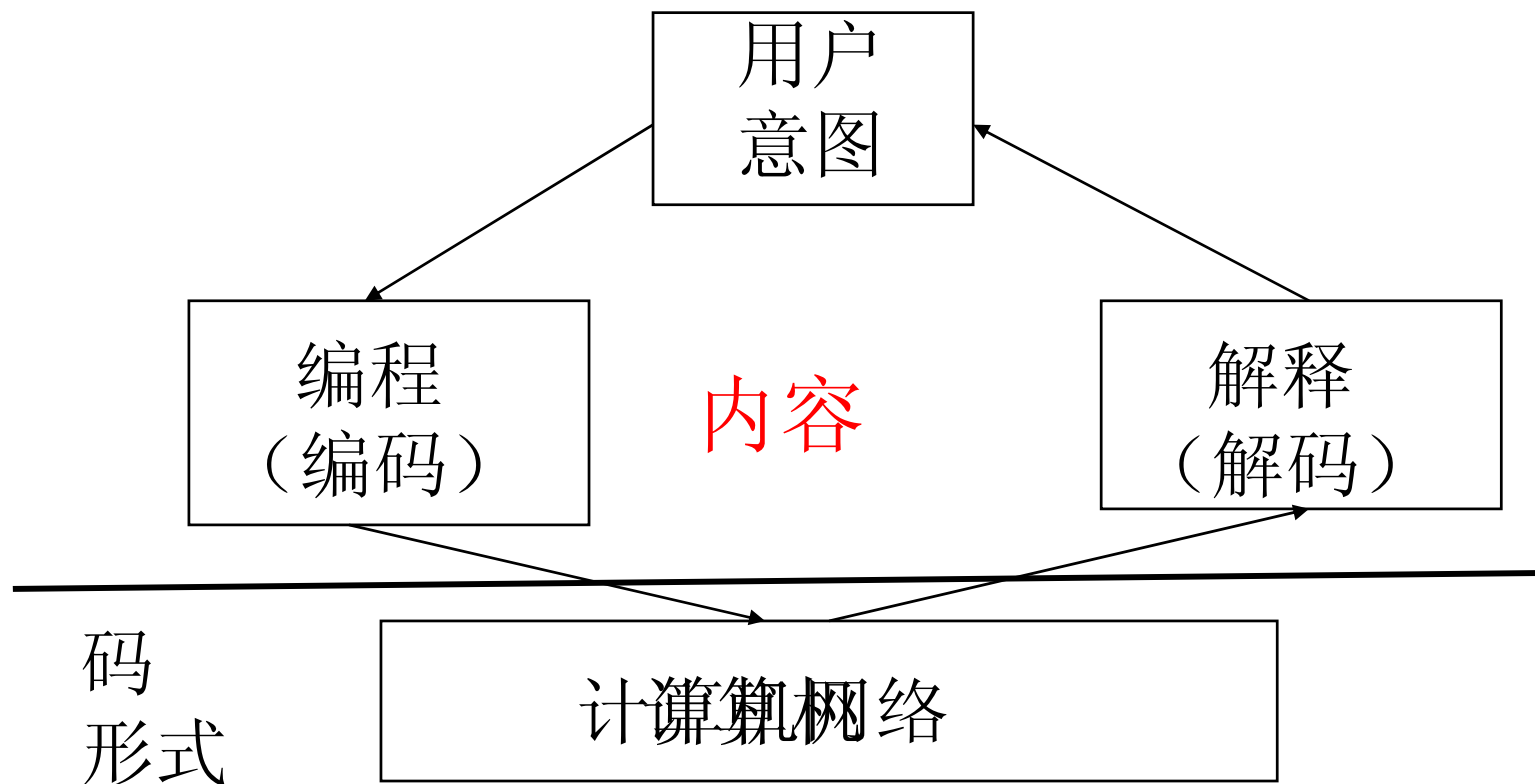
经典信息论在文本、图像处理中的应用

经典信息论的核心是解决信息在噪声信道传输中的工程问题

- 通信： 编码、数据压缩、噪声信道编码等
- 文本： 编辑、压缩、拼写与文法纠正等
- 图像： 编辑、无损压缩、噪声抑制、图像增强等

从形式到与语义相关的信息处理

自然的人机交互





基于内容（语义）信息处理

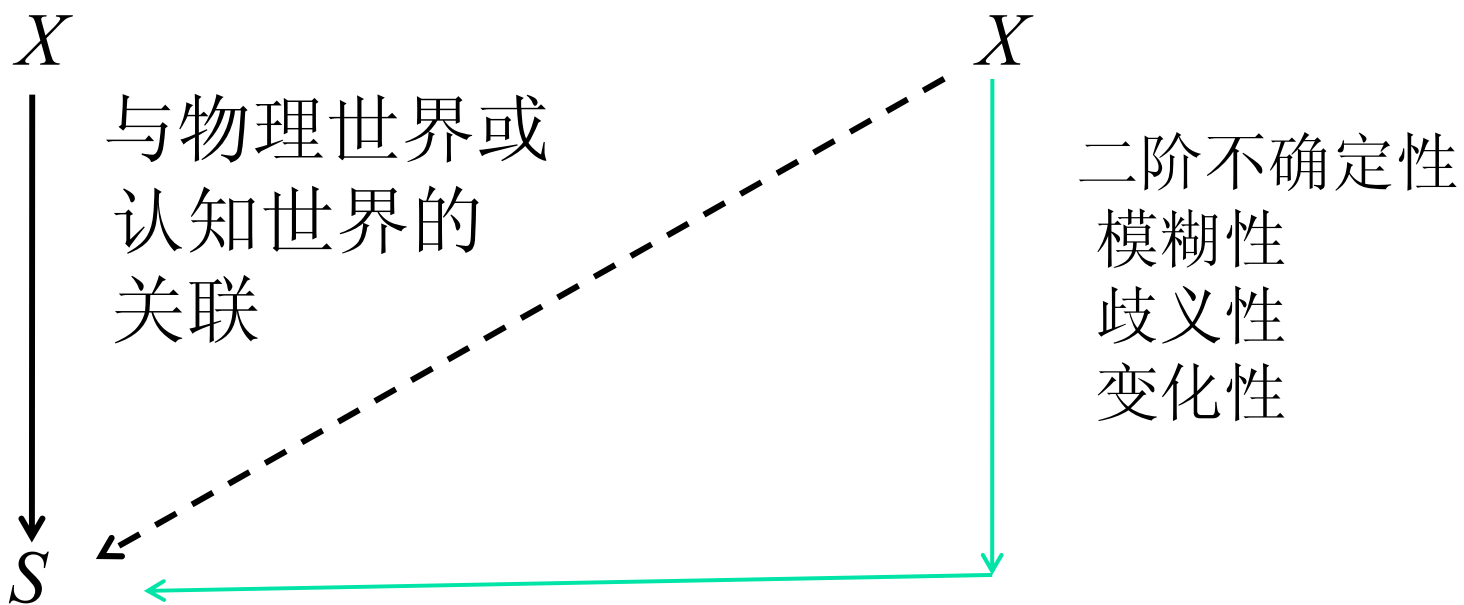
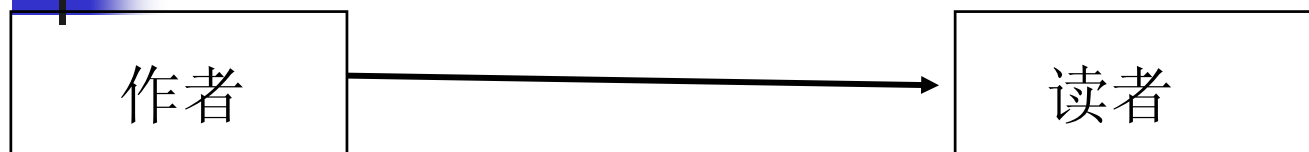
互联网时代

- 复杂（多样化）信息：文本、图像、语音、视频
- 海量数据
- 人机交互

基于内容（语义）信息处理

信息检索、分类、摘要、识别、理解， ..

1. 人工智能方法-基于规则的表达



基于规则的表达



领域背景知识表示

知识库（陈述性知识）

If a, \dots 症状（模糊）

—————→ CF: 置信度

Then b 某种功能失调（模糊）

基于特定知识的推理



基于知识方法的适用范围

适用领域

深思熟虑行为（AI 专家系统：决策、诊断、设计，...）

特定领域、较小规模

不适用领域

感知、常识、自然语言，..

推广能力差！

感知：知识（规则）表示的困难！

-句法与结构分析

检测子

具有语义的特征：

边界、形状等

部件之间没有明显边界

图像分割难题

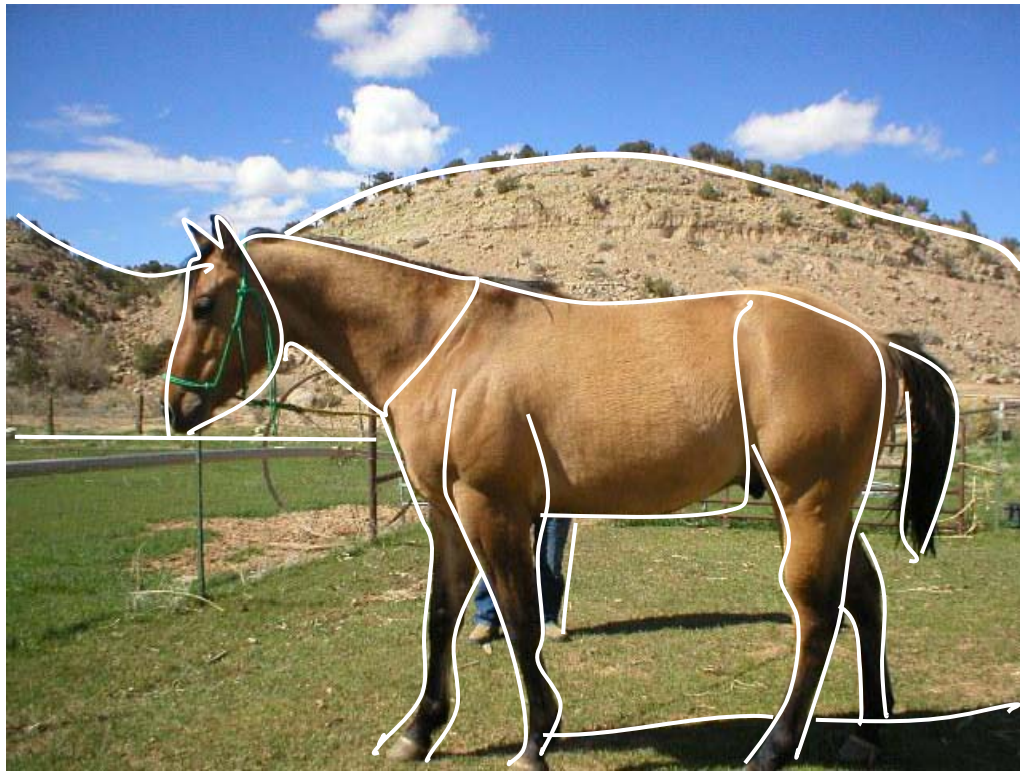
描述子

结构的不确定性



K. S. Fu, Syntactic pattern recognition, New York: Prentice-Hall, 1974. D. Marr, Vision, New York: Freeman, 1982

定位与识别-图像分割

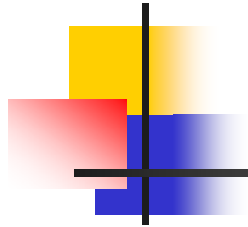


物体在哪里？



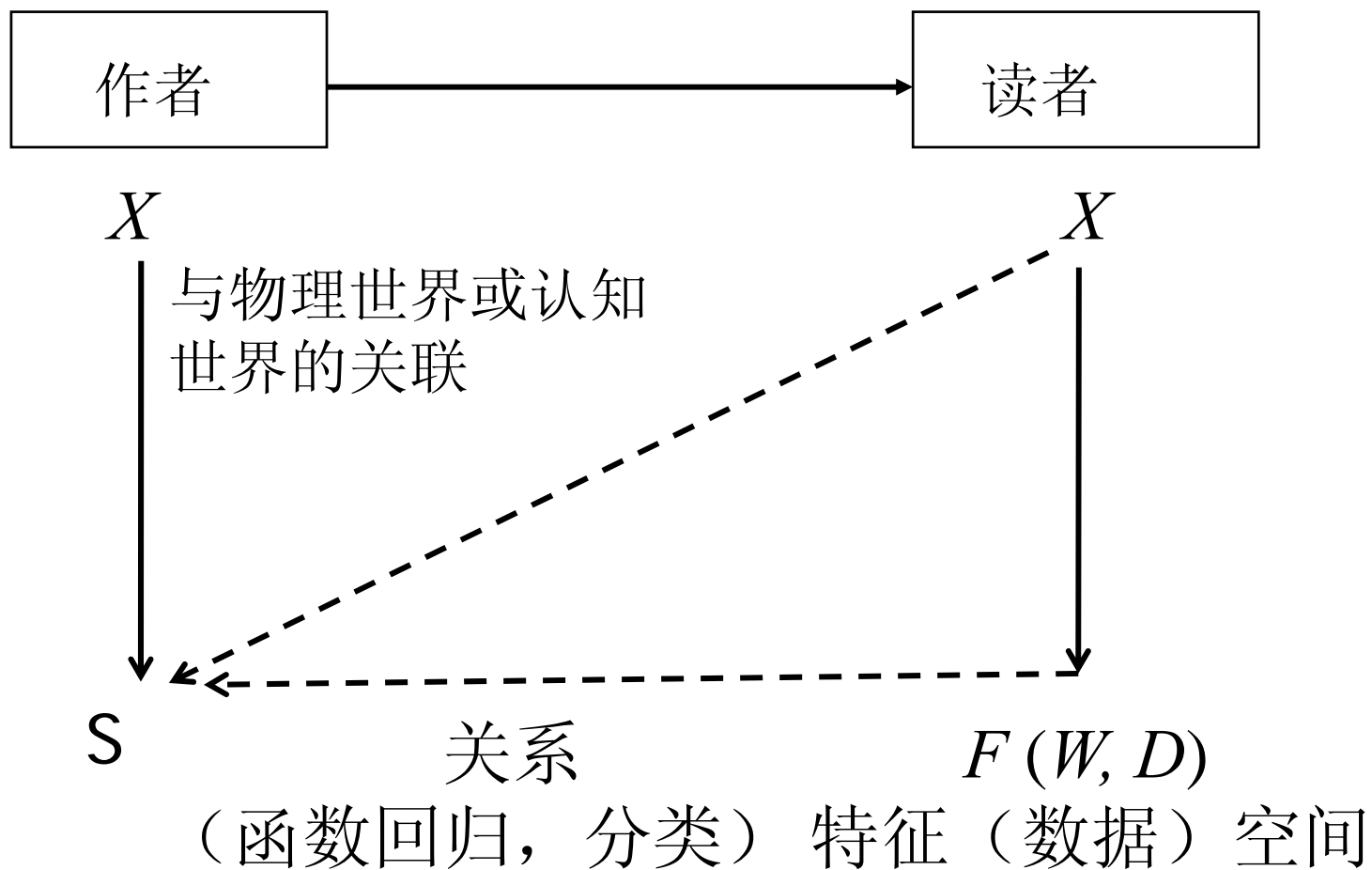
物体是什么？

鸡与蛋问题



如何处理感知问题？

2. 统计方法





经典统计理论

函数空间的大数定理

经典统计理论下的机器学习问题

基本假设：函数类型已知

$$ax^2 + bx + c$$



机器学习理论的三个里程碑

泛函空间的大数定理

学习过程

- 一致收敛的充要条件
- 快速收敛的充分条件
- 在任意概率测度下快速收敛的充要条件

现代统计论下

$$F(x, y) = F(y|x)F(x), \quad y = f(x)$$

数据

函数
语义

$(\omega_1, x_1), \dots, (\omega_l, x_l)$ 优化 $F(x, y)$ or $y = f(x)$

$$R(\alpha) = \int L(\omega, \phi(x, \alpha)) dF(\omega, x)$$

$\phi(x, \alpha), \alpha \in \Lambda$, α : 一组标量、向量、或任何抽象元素

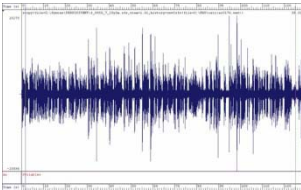
若 $F(x, y)$ 或 $f(x)$ 存在

则在概率意义下, 可由数据推断出函数

实际问题



数字视频编码技术发展至今已有半个世纪的历史，已取得很大的进展。从五十年代的差分预测编码，到七十年代的变换编码、基于块的运动预测编码，直到如今兴起的分布式编码、立体视编码、多视编码、视觉编码等等



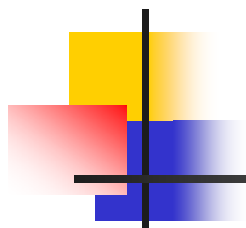
X -表示
(数据)

映射 ?



概念
语义

S -语义



映射

$$F(W, D) \rightarrow S$$

- 特征空间表示与其语义之间的映射
存在？
- 对于一定的表示只存在于
一定的“数据集”中
(文件、图像、语音, ...) !

规格化、对齐的正面人脸库

Extended Yale B

2414 正面人脸

具有不同光照

38 个体

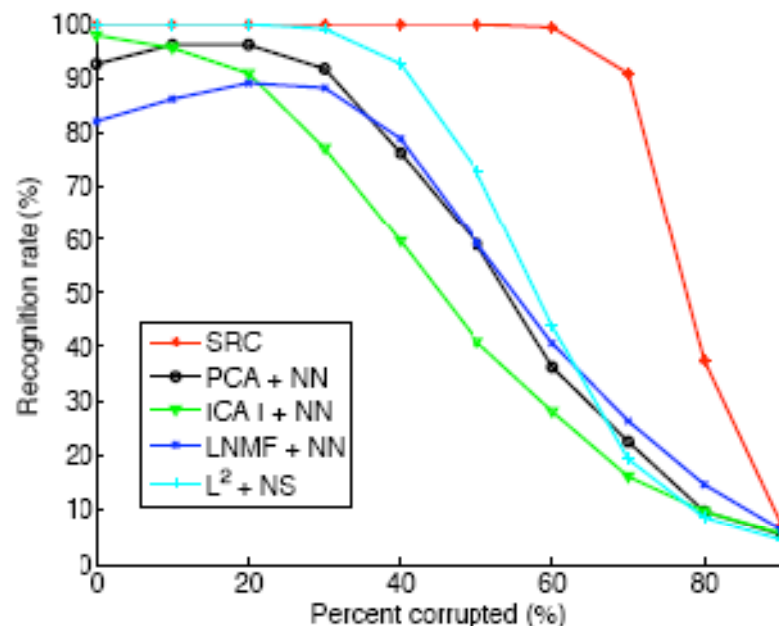
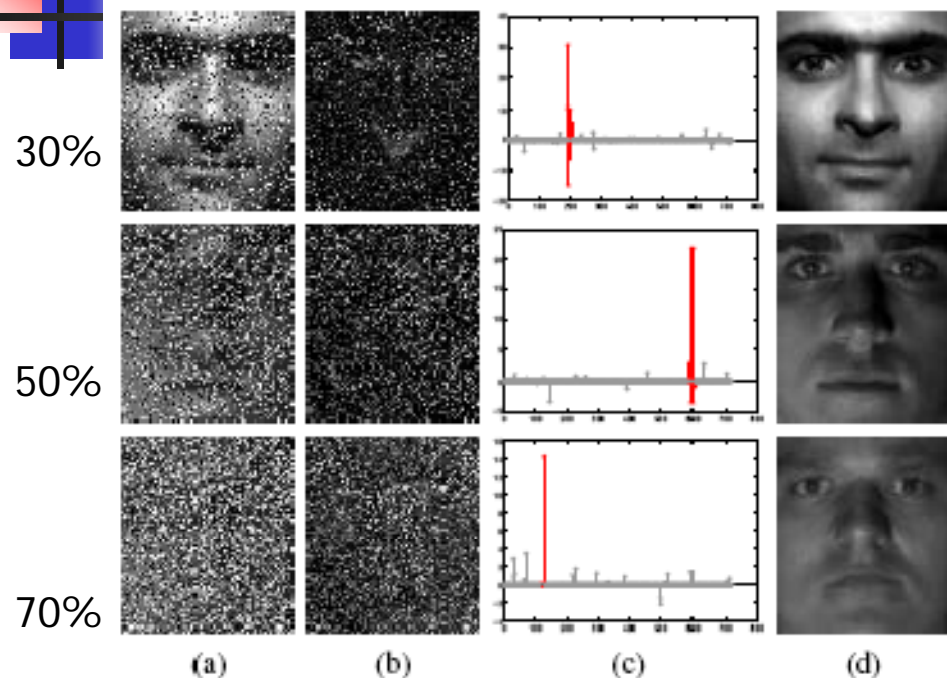
192×168 像素图像

特征空间维度：30, 56,
120, 504



J. Wright, et al. Robust face recognition via sparse representation, IEEE PAMI 09, 31(2):210-227

抗噪声能力



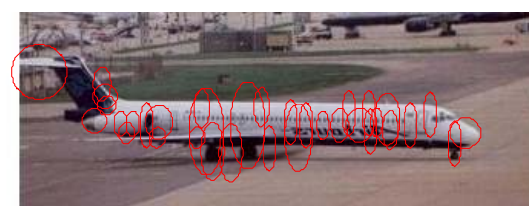
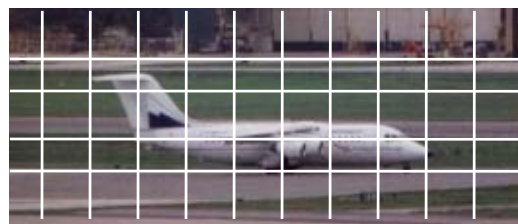
Percent corrupted	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%
Recognition rate	100%	100%	100%	100%	100%	100%	99.3%	90.7%	37.5%	7.1%

SRC-基于稀疏表示的分类, NN-最近邻法,

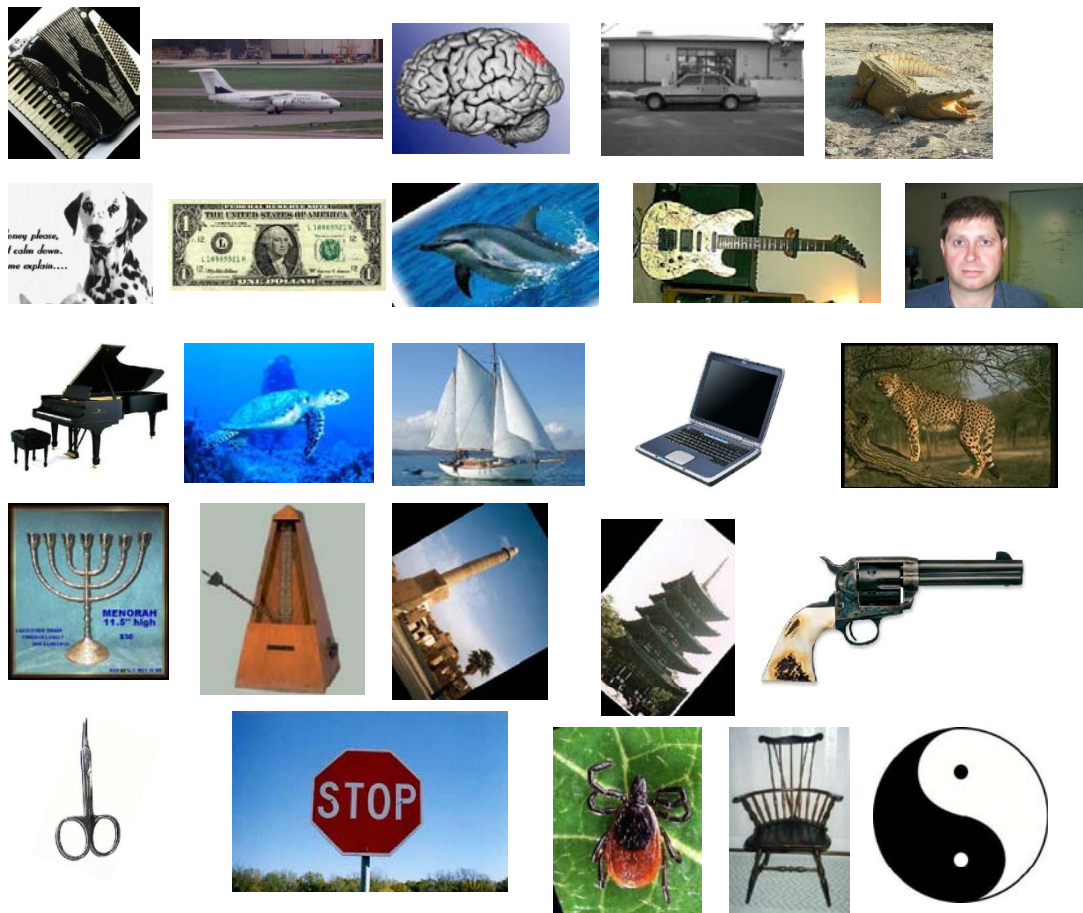
NS-最近子空间法, Extended Yale B. (images)

(视觉) 词袋法

- 定义在图像块上 (2005-06)
- 兴趣点集上提取的描述子 (2002-2004)
- 边界轮廓 (2005-06)
- 区域 (2005-06)



图像库

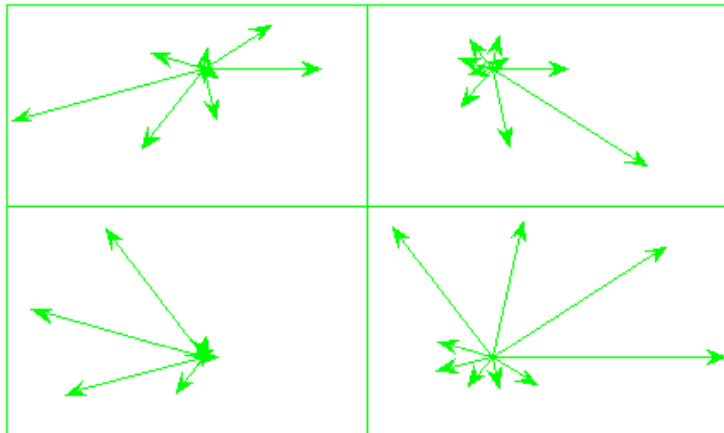


Caltech 101 (25类, 30幅/类)

检测子与描述子



Kadir 显著区域
(点)

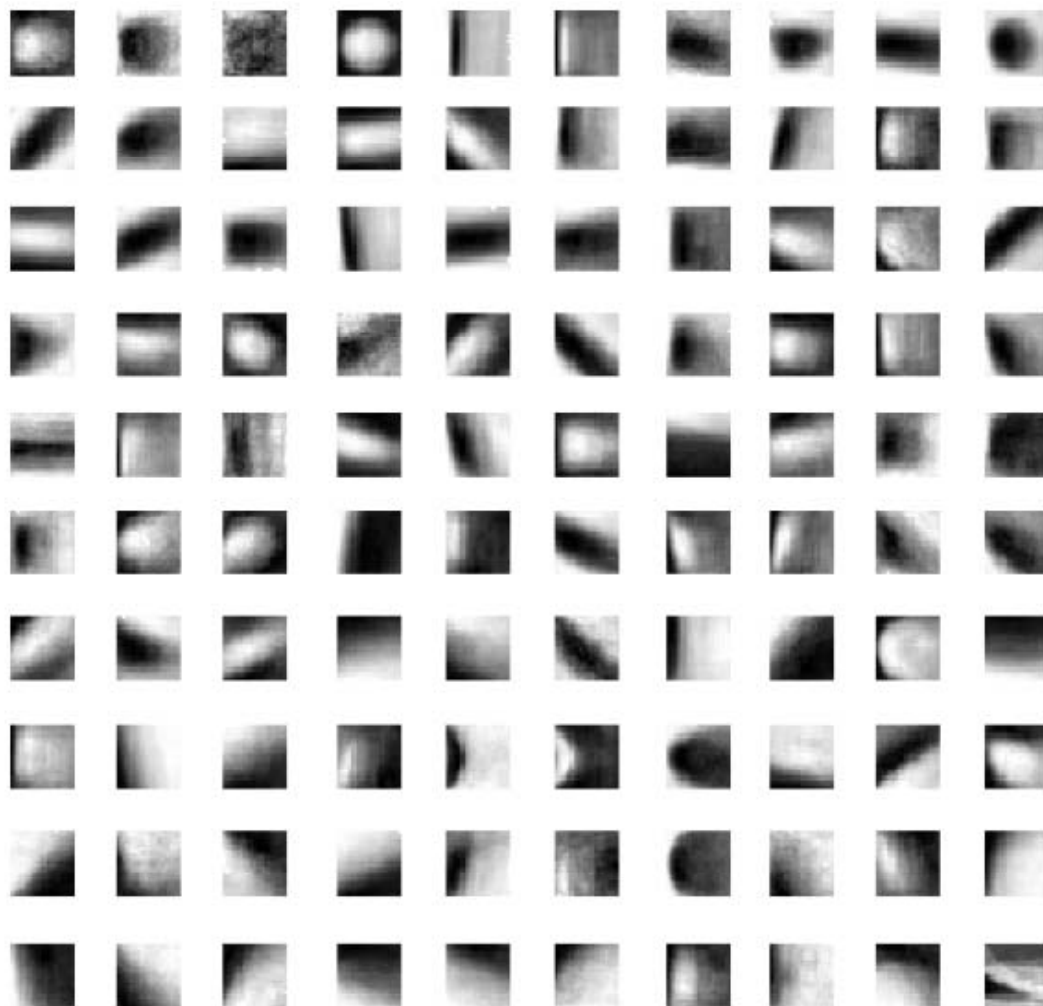


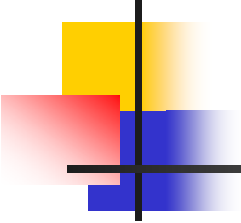
方向梯度的直方图
(HOG)

-72 维

Zuo Yuanyuan, Bo Zhang (2010-)

低层与局部视觉词 (100)




$$CB(w) = \frac{1}{n} \sum_{i=1}^n \begin{cases} 1 & \text{if } w = \arg \min_{v \in V} (D(v, r_i)) \\ 0 & \text{otherwise} \end{cases}$$

CB -码本（直方图）

n -一个图像中的区域数目

r_i -图像区域 i

$D(w, r_i)$ -码字 w 与区域 r_i 之间的距离

V -词汇

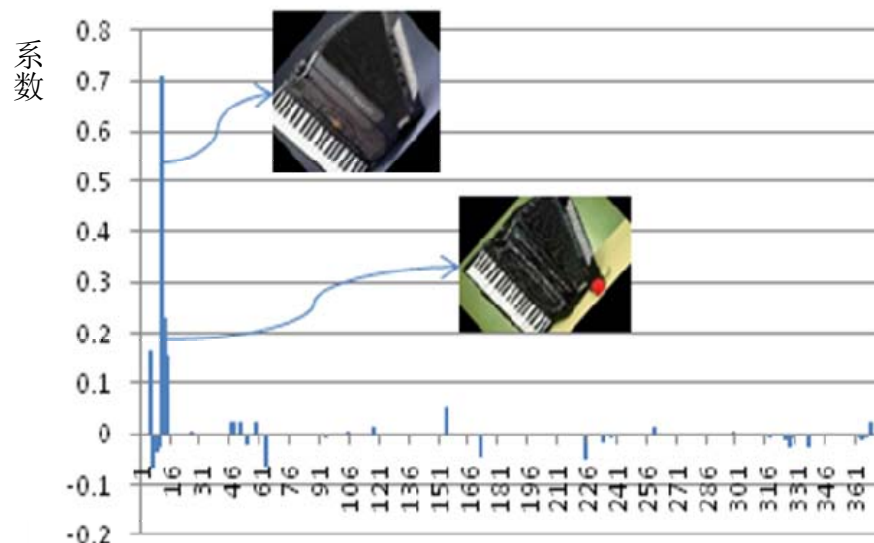
优化

$$\hat{x}_1 = \arg \min \|x\|_1 \quad \text{subject to} \quad Ax = y$$

$$\hat{x}_1 = \arg \min \|x\|_1 \quad \text{subject to} \quad \|Ax - y\|_2 \leq \varepsilon$$

样本空间的稀疏表示

L_1 范式

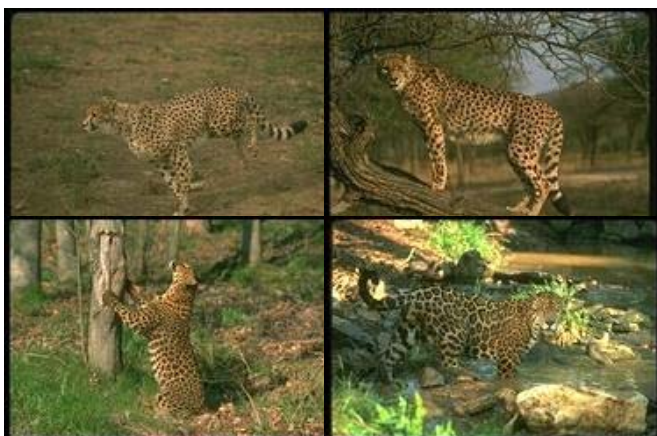
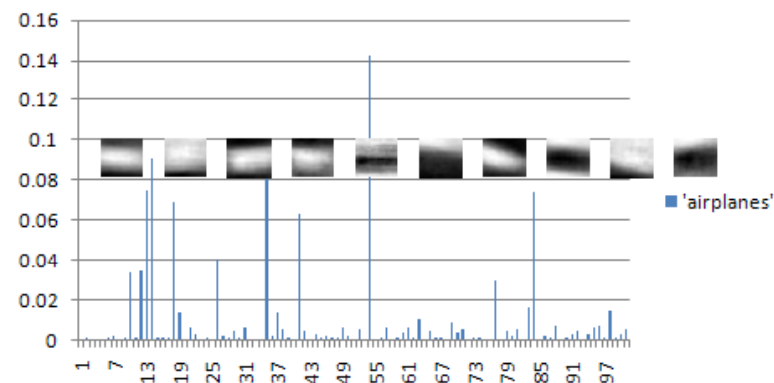


训练样本

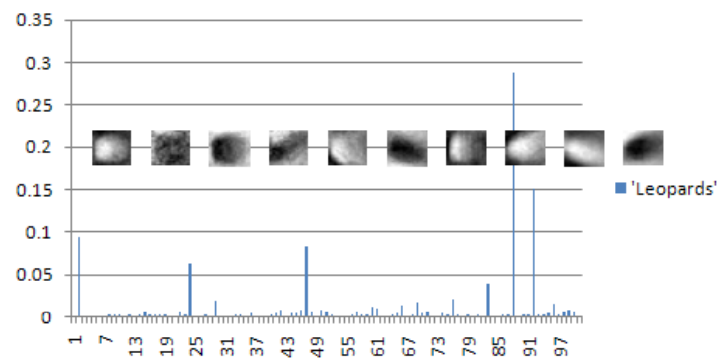
高维空间中的稀疏结构



'airplanes'

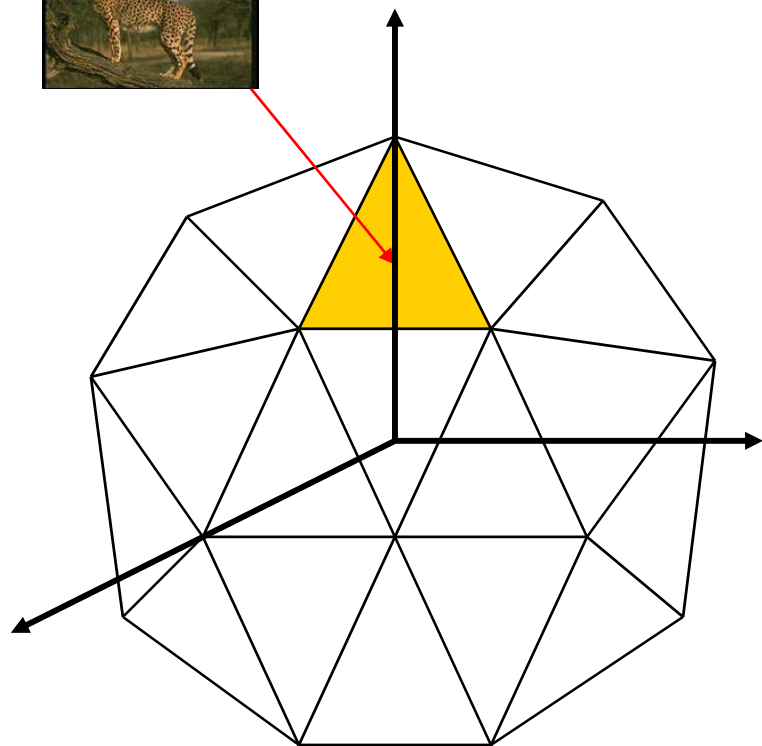


'Leopards'



视觉词分布图

高维空间中的低维结构

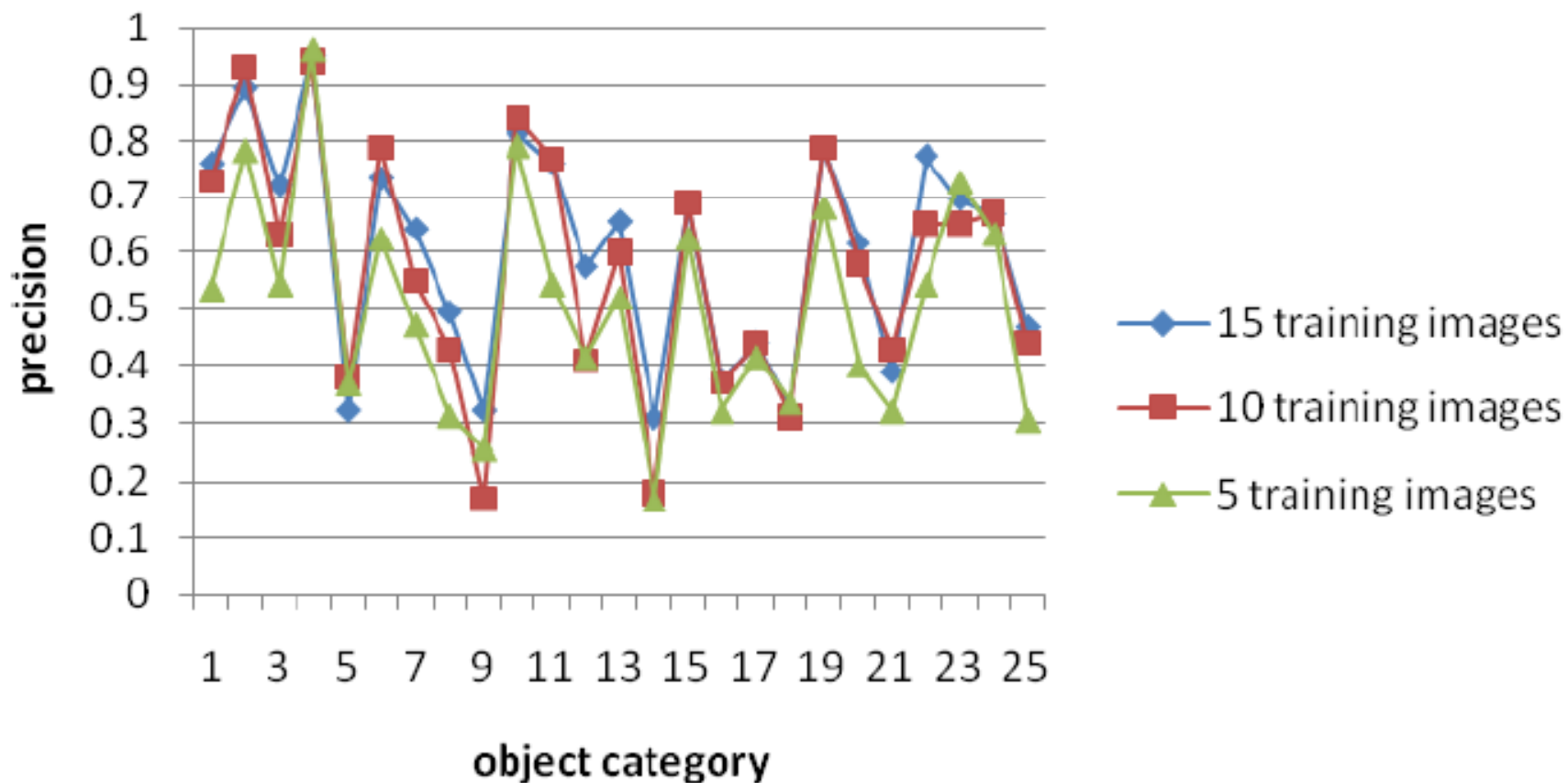


数据结构

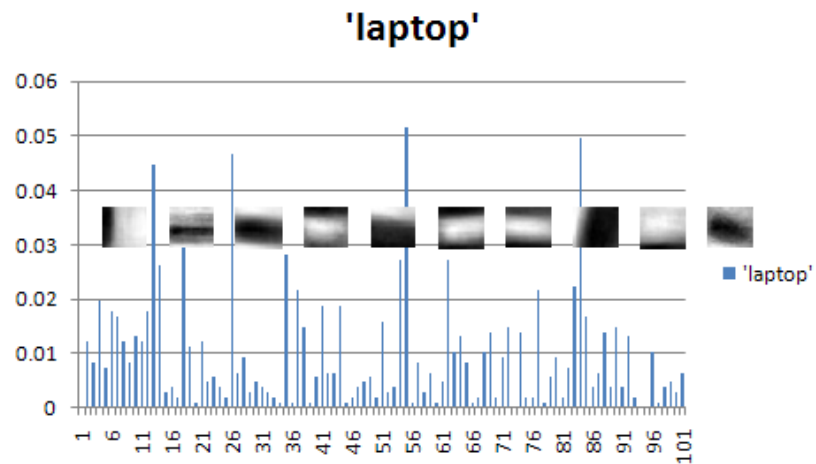
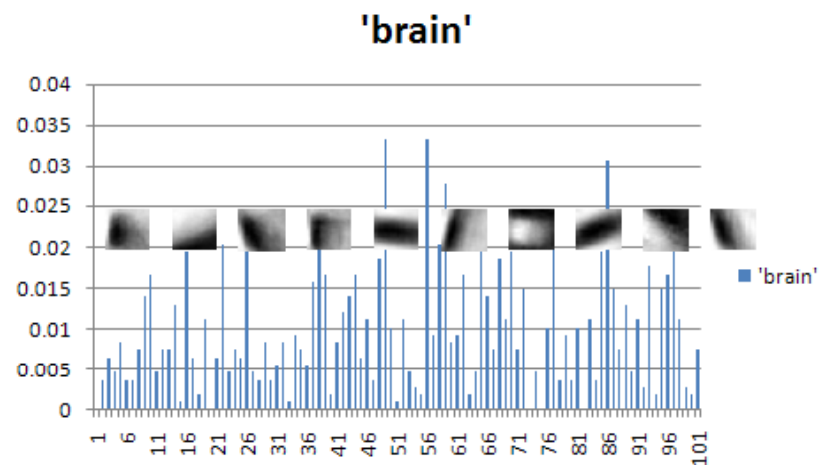
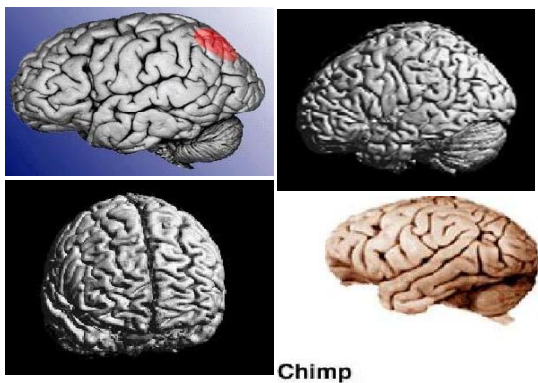
物体	准确度 (SRC)
飞机	0.947
汽车侧面	0.987
斑点狗	0.733
人脸	0.760
美洲豹	0.827
宝塔	0.760
停车标志牌	0.800
温莎椅	0.787

实验结果

SVM, 100 visual words



非稀疏性



语义鸿沟

—统计方法的弱点

- 底层局部特征与高层全部概念之间的鸿沟

具有较少语义的特征：颜色或它的分布（直方图），灰度或它的分布，视觉词（从兴趣点集得到的描述子），图像块，图像区域，图像边界，...

- 缺乏结构知识

泛化能力

无理解的信息处理



语义：意图，解释

语义学：含义的研究

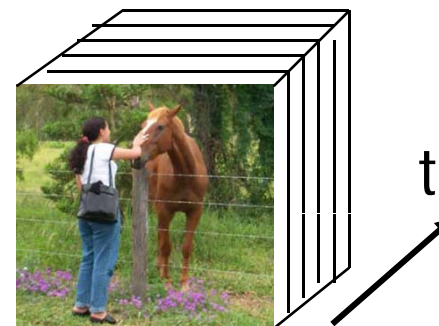
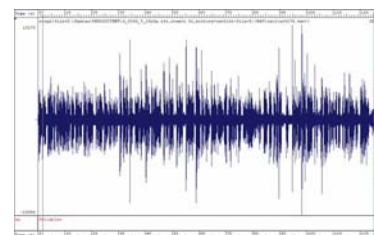
- 含义（语言学）是从当前上下文中推断出来的作者或者演讲者在语用学意义下的意图
- 含义（非语言学）用于描述人们对于世事的解释（客观化的语义）

信息结构

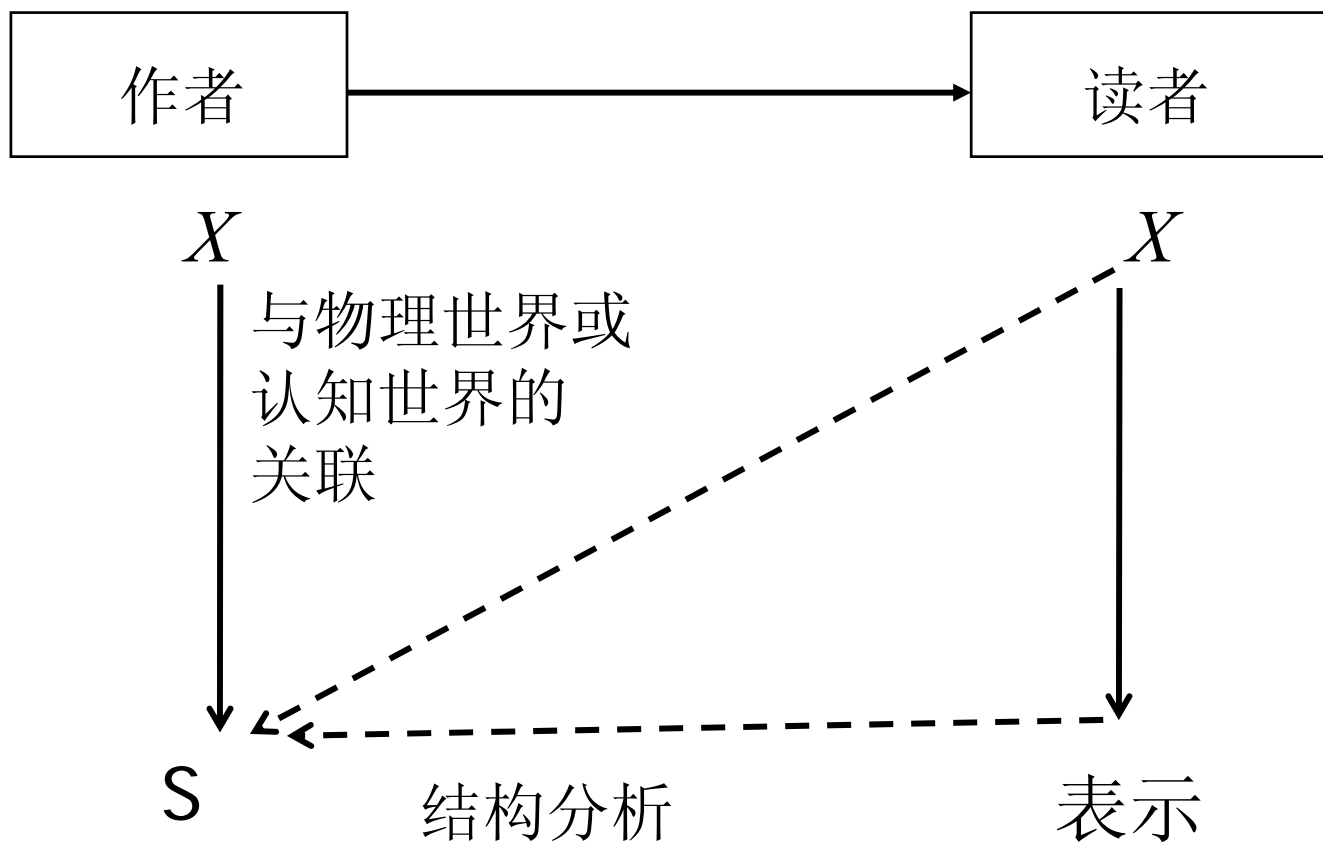
- 文本：上下文结构
- 图像：空间结构
- 语音：时间结构
- 视频：时空结构

信息（环境）结构的获取
与预测

数字视频编码技术发展至今已有半个世纪的历史，已取得很大的进展。从五十年代的差分预测编码，到七十年代的变换编码、基于块的运动预测编码，直到如今兴起的分布式编码、立体视编码、多视编码、视觉编码等等



3. 结构分析模型





上下文（语境）分析

马尔科夫（语言）模型 -C. Shannon

n -gram（项目：音素、音节、字母、词等级别）

$$P(x_i | x_{i-1}, \dots, x_{i-n})$$

$n=0$, 一元语法（单项出现频率）

$n=1$, 二元语法

$n=2$, 三元语法

2D 扁平结构模型

-图像区域标注：马、天空、山，草地



Yuan Jinhui, Bo Zhang (2008-)

结构化的概率模型-概率图模型

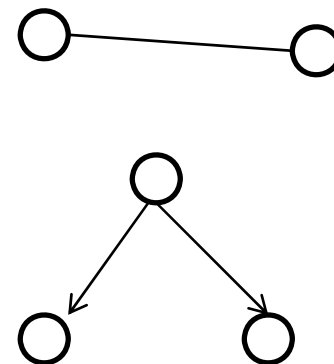
- 马尔科夫网络-无向图
- 贝叶斯网络-有向无环图

陈述性表示

推理-优化

学习

感知采用同样机制！



区域-自适应的方格划分

$$z^* = \arg \max_z p(z|I) = \arg \max_z p(I|z)p(z)$$

$$p(I, z) = \frac{1}{Z} \exp \left\{ \sum_i \alpha_i f(I(x_i), z_i) + \sum_{i,j} \beta_{i,j} g(z_i, z_j) \right\}$$

$$p(I, z) = p(I|z)p(z)$$

I -数据

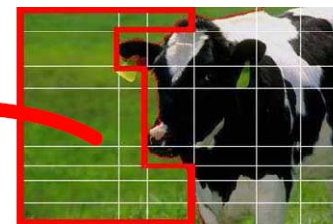
x_i -图像位置

z -状态（特征）

$g(z_i, z_j)$ -概率约束
(同现)



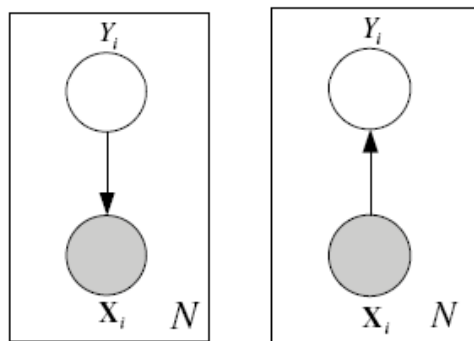
Region segmentation by JSEG



Region-adaptive grid partition

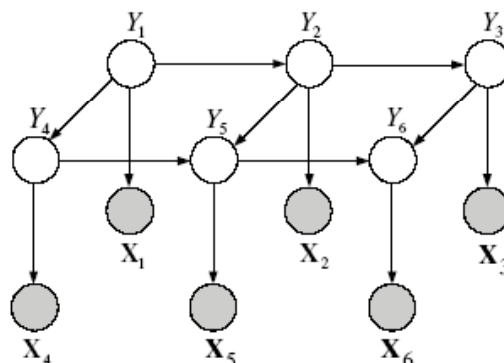
基于区域

图模型

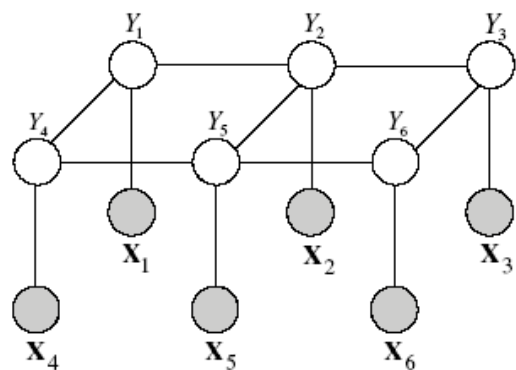


(a)

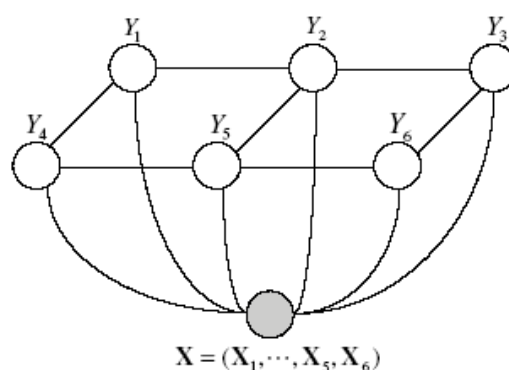
(b)



(c)



(d)



(e)

- (a) 产生式模型
- (b) 区分性模型
- (c) 2D-隐马尔科夫模型 (HMM)
- (d) 马尔科夫随机场(MRF)
- (e) 条件随机场 (CRF)



模型学习

n 图像数, $m_i = H \times V$ 每个图像方格数目

$$(x, y) = \{(x_i, y_i), i = 1, 2, \dots, n\}$$

$$(x_i, y_i) = \{(x_i^j, y_i^j), j = 1, 2, \dots, m_i\}$$

- (a) i.i.d 产生式模型
- (b) i.i.d. 区分式模型
- (c) 2-D 隐马尔科夫模型 (2D HMM)
- (d) 马尔科夫随机场模型 (MRF)
- (e) 条件随机场模型 (CRF)



标记组合

$$\{(x^i, y^i), i = 1, 2, \dots, N\}$$

给定的训练数据

MAP (最大后验) : 标记组合

$$y^* = \arg \max P(y^{1:m} | x^{1:m})$$

对于 2D HMM, MRF, CRF 模型

运用路经受限的 Viterbi 算法



马尔科夫随机场—MRF

概率分布 P

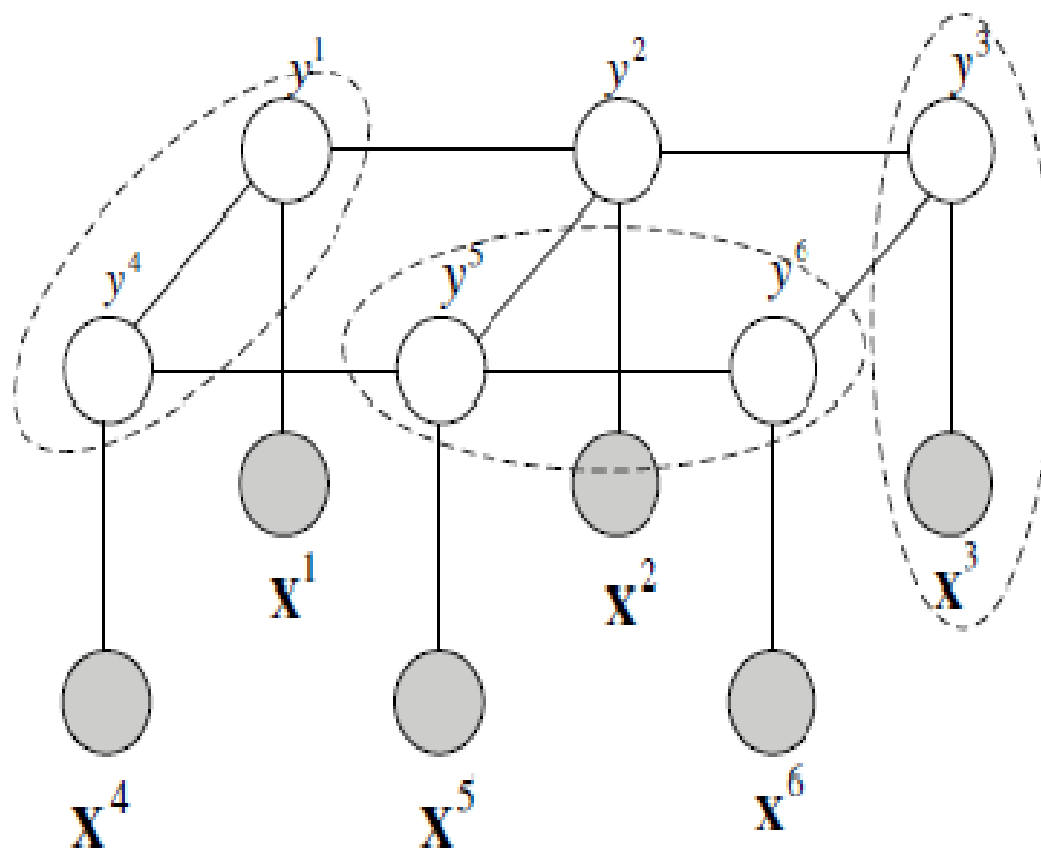
C^s : 标记团 C^0 : 标记与特征团

y^* 最优标记组合

$$P(x^{1:m}, y^{1:m}) = \frac{1}{Z} \prod_{(y^i, y^j \in C^s)} \Psi(y^i, y^j) \prod_{(y^k, x^k \in C^0)} \Theta(y^k, x^k)$$

$$y^* = \arg \max_{y^{1:m}} \prod_{(y^i, y^j \in C^s)} \Psi(y^i, y^j) \prod_{(y^k, x^k \in C^0)} \Theta(y^k, x^k)$$

马尔科夫随机场-MRF





结构预测学习

学习规则	分类	结构预测
最大联合似然估计	朴素贝叶斯网络	隐马尔科夫模型 (1966) ¹
最大条件似然估计	逻辑回归	条件随机场 (2001) ²
最大间隔学习	SVM	最大间隔马尔科夫网 (2003) ³
最大熵判别式学习	最大熵判别式模型	最大熵判别式马尔科夫网 (2008) ⁴



相关论文

- [1] L. E. Baum and T. Petrie. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, Vol. 37, No. 6, pp.1554-1563, 1966
- [2] J. Lafferty et al. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. of International Conference on Machine Learning (ICML)*, 2001
- [3] B. Taskar et al., Max-Margin Markov Networks. *Advances in Neural Information Processing Systems (NIPS)*, 2003
- [4] J. Zhu et al., Laplace Maximum Margin Markov Networks, In *Proc. of International Conference on Machine Learning (ICML)*, 2008



实验设置

- 4002 Corel 图像 (384×256 or 256×384)
- 11 基本 (区域) 概念
- 特征: 颜色矩+小波
- 5 种模型: 2 个无结构知识
(GMM, SVM)
3 个有结构知识
(HMM*, RMF*, CRF*)

The Categories of Image Region Annotations



语义概念	定义及描述
天空 (Sky)	包含空气, 云, 烟, 雾等
水面 (Water)	包括河流, 大海, 湖泊, 喷泉, 瀑布等
山脉 (Mountain)	只含山脉的远景
草 (Grass)	除树木和花朵外的自然植被
树木 (Tree)	包含树干, 树叶等
花 (Flower)	各种色彩的花朵
岩石 (Rock)	较近观察的石头, 注意与“山脉”的区分
土壤 (Earth)	自然裸露的地面
地面 (Ground)	人加工过的地表, 例如道路, 广场等, 注意与“土壤”区分
建筑 (Building)	人建造的结构, 例如房屋, 桥梁等
动物 (Animal)	动物皮毛, 例如老虎, 狮子, 大象等



不同图模型

GMM: 混合高斯模型 (30个成分)

SVM: 支持向量机

高斯核, 一对一

HMM: 隐马尔科夫模型

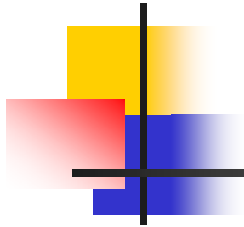
RMF: 马尔科夫随机场

CRF: 条件随机场

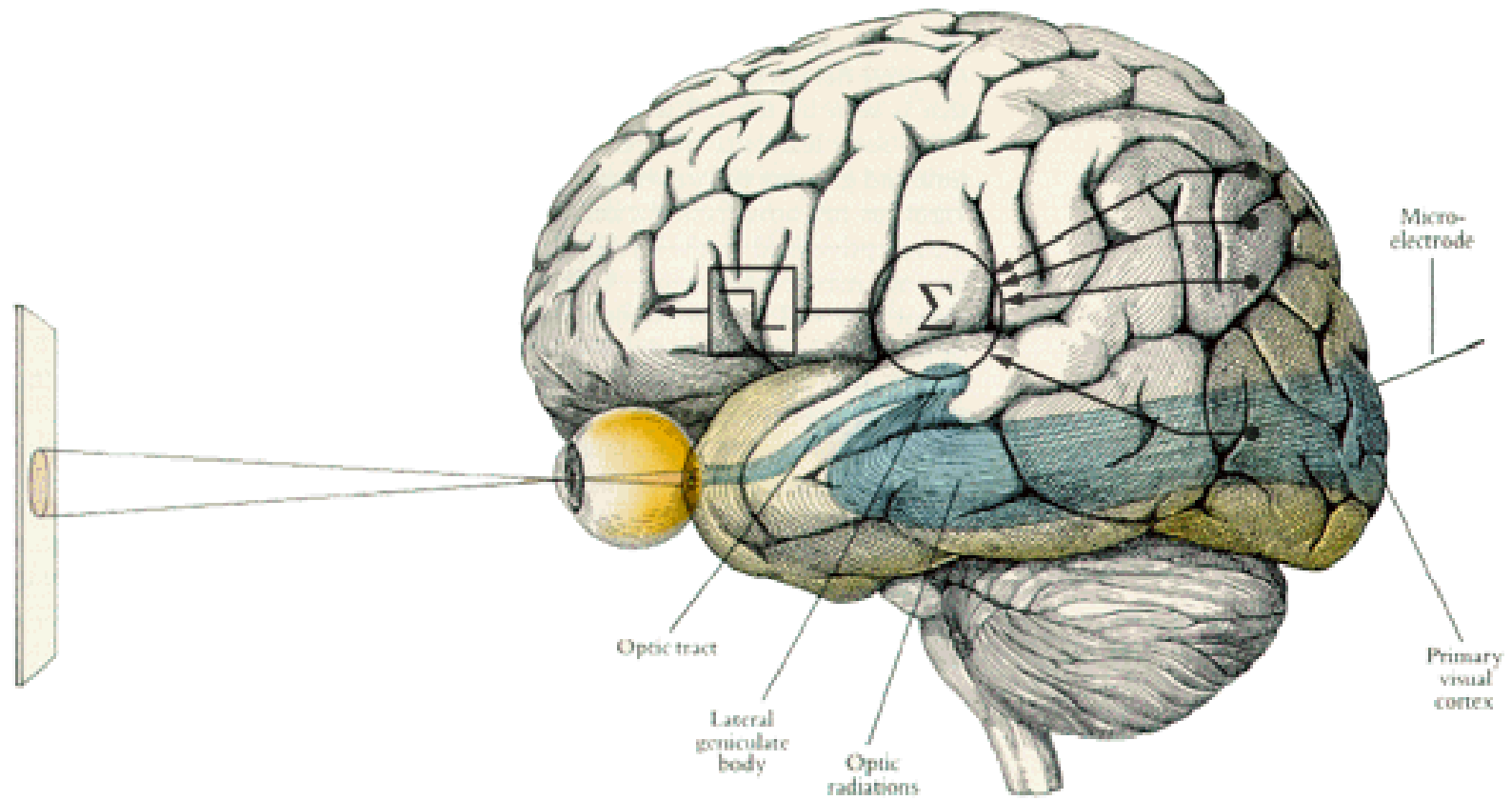
路径 Viterbi 算法

实验结果

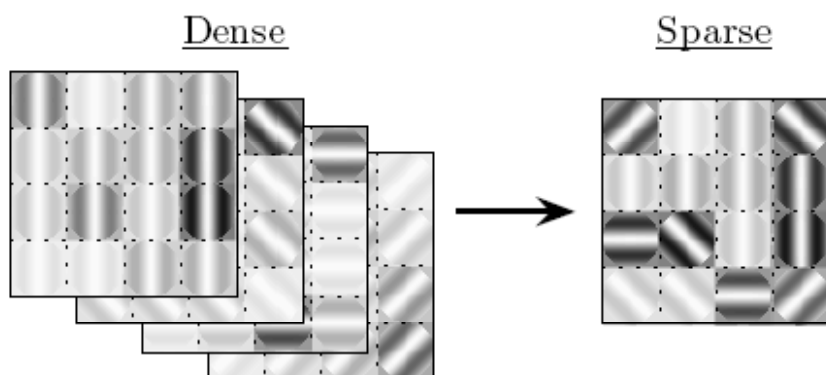
	Precision					Recall					F-score				
	gmm	svm	hmm	mrf	crf	gmm	svm	hmm	mrf	crf	gmm	svm	hmm	mrf	crf
sky	.937	.961	.933	.914	.955	.832	.899	.889	.871	.909	.882	.929	.911	.892	.931
wat.	.410	.583	.531	.390	.598	.449	.588	.489	.496	.639	.429	.585	.509	.437	.618
mmt.	.134	.269	.215	.299	.395	.282	.392	.315	.313	.435	.182	.319	.255	.306	.401
grs.	.616	.655	.651	.591	.661	.652	.757	.679	.696	.780	.633	.702	.665	.639	.715
tre.	.709	.765	.611	.615	.755	.481	.538	.556	.532	.570	.573	.632	.582	.571	.650
flr.	.475	.591	.584	.561	.615	.513	.694	.447	.411	.695	.494	.639	.507	.474	.653
rck.	.033	.088	.198	.242	.220	.132	.281	.216	.189	.341	.050	.134	.207	.212	.268
ert.	.230	.386	.337	.184	.294	.397	.497	.328	.372	.539	.291	.434	.332	.246	.445
grd.	.099	.208	.433	.461	.379	.316	.569	.220	.187	.509	.151	.305	.292	.265	.424
bld.	.610	.730	.484	.481	.625	.437	.569	.582	.550	.645	.509	.640	.529	.513	.687
anl.	.096	.297	.295	.312	.480	.294	.573	.285	.204	.540	.144	.392	.290	.247	.508
avg.	.395	.503	.479	.459	.560	.435	.578	.455	.438	.600	.414	.538	.467	.448	.579



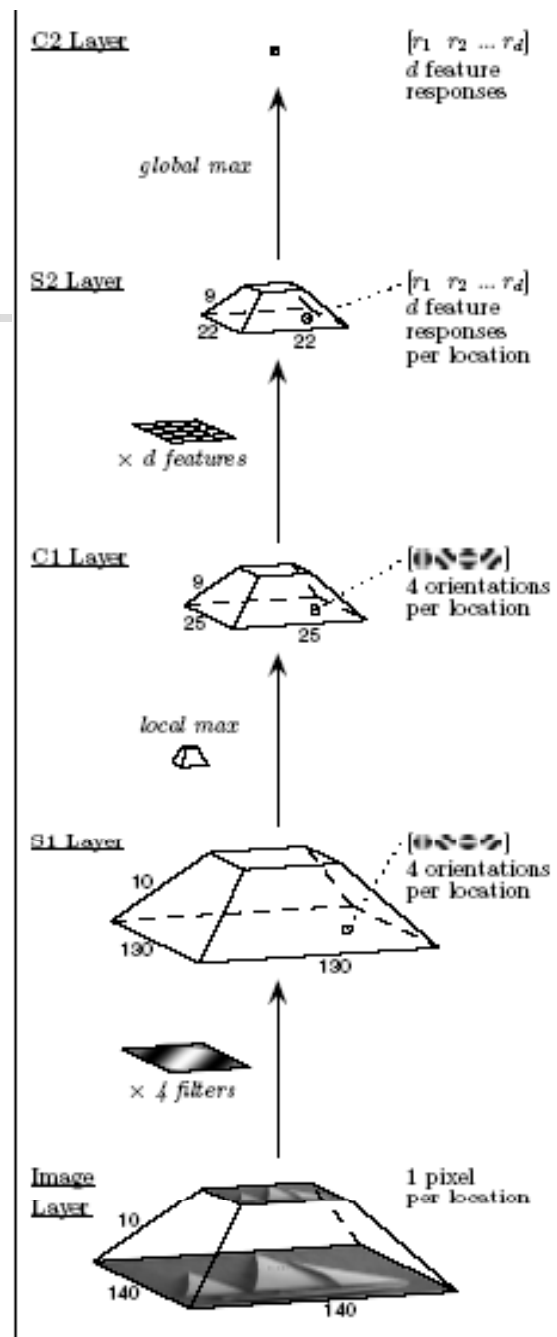
向人脑学习



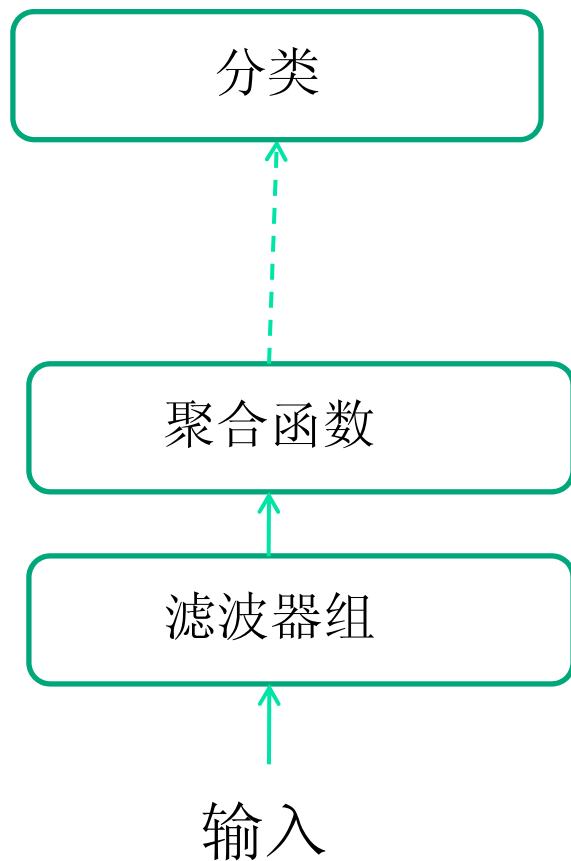
物体识别



MIT-CSAIL-TR-2006-028 T. Serre



理论框架



核函数

Name	Expression
Inner product	$K(x, y) = \langle x, y \rangle$
Normalized inner product	$K(x, y) = \frac{\langle x, y \rangle}{\ x\ \ y\ }$
Gaussian	$K(x, y) = e^{-\gamma \ x-y\ ^2}$

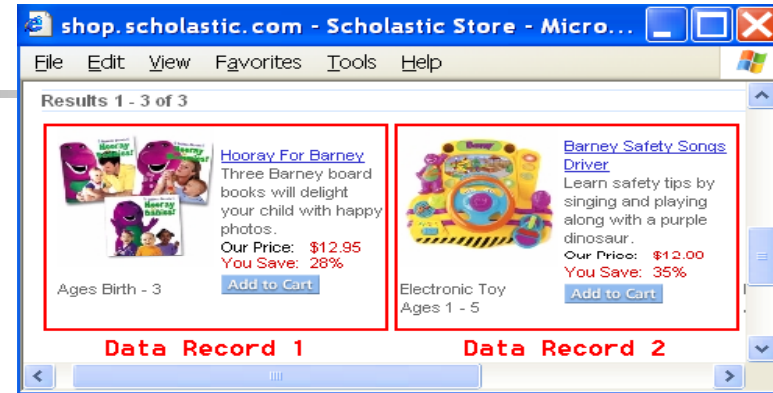
聚合函数

Name	Expression
Average	$\psi(\alpha(h)) = \frac{1}{ H } \sum_{h \in H} \alpha(h)$
l^1 -norm	$\psi(\alpha(h)) = \sum_{h \in H} \alpha(h) $
Max	$\psi(\alpha(h)) = \max_{h \in H} \alpha(h)$
l^∞ -norm	$\psi(\alpha(h)) = \max_{h \in H} \alpha(h) $

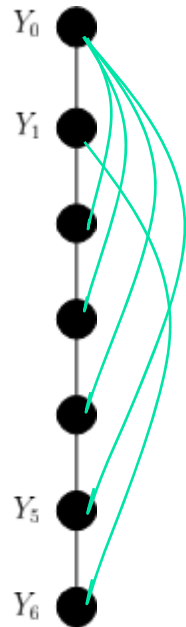
隐分层递阶模型

网页内容提取

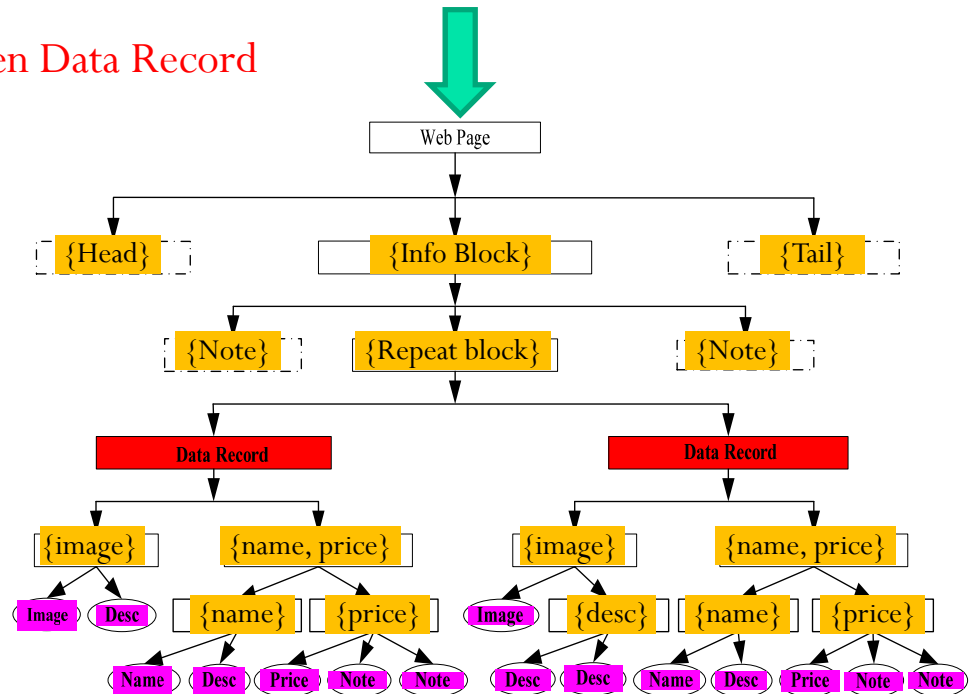
Name, Image, Price, Description, etc.



Given Data Record



- Hierarchy
 - ✓ Computational efficiency
 - ✓ Long-range dependency
 - ✓ Joint extraction



实验设置

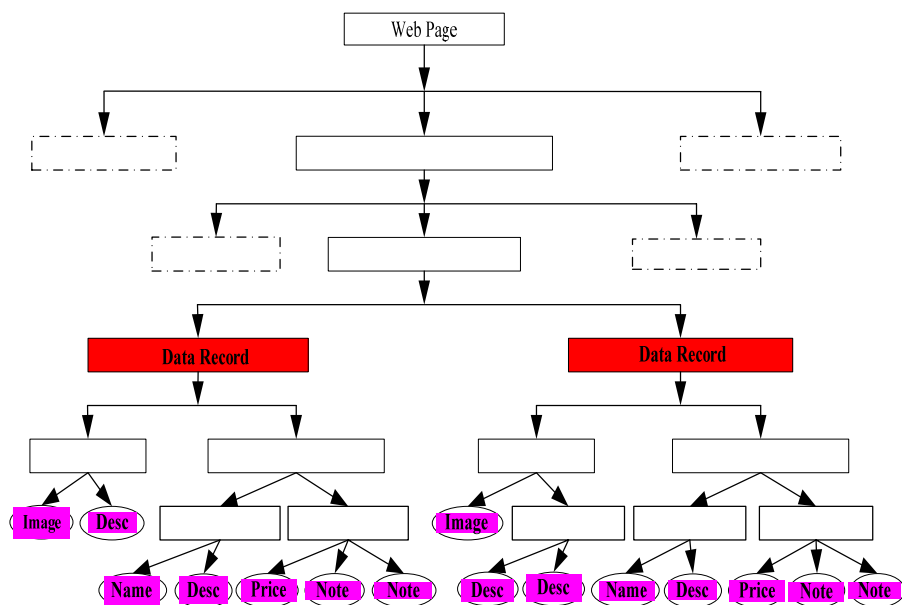
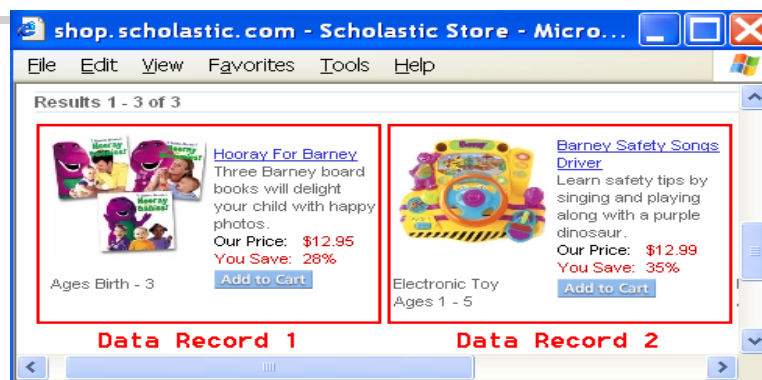
- 网页提取
- *Name, Image, Price, Description*
- 模型

Multi-layer CRFs, Multi-layer M^3N , PoMEN, Partially observed HCRFs

数据集: 37 模板

训练: 185 (5/per template) pages, or 1585 data records

测试: 370 (10/per template) pages, or 3391 data records

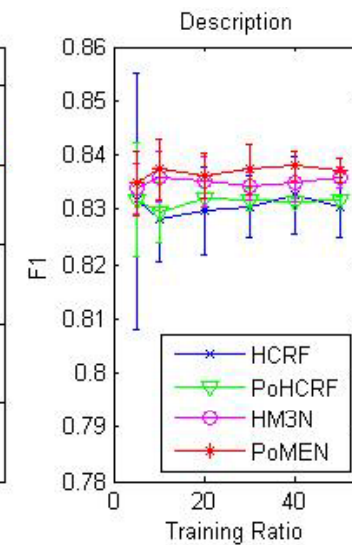
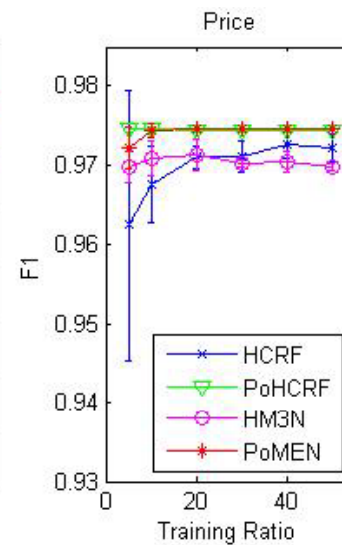
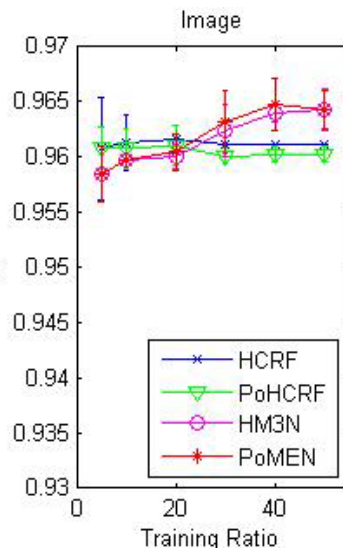
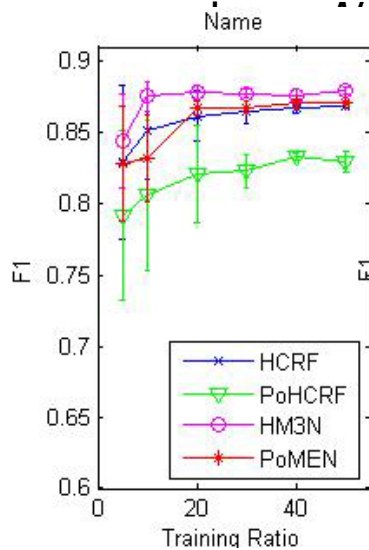
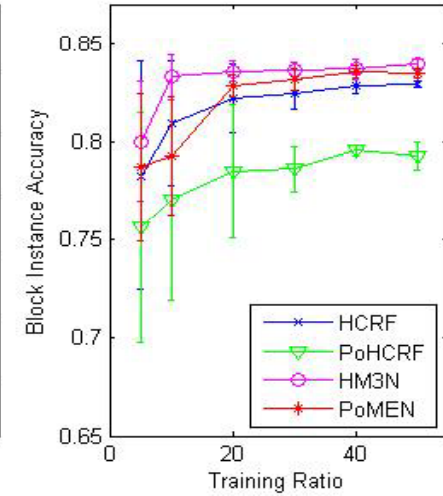
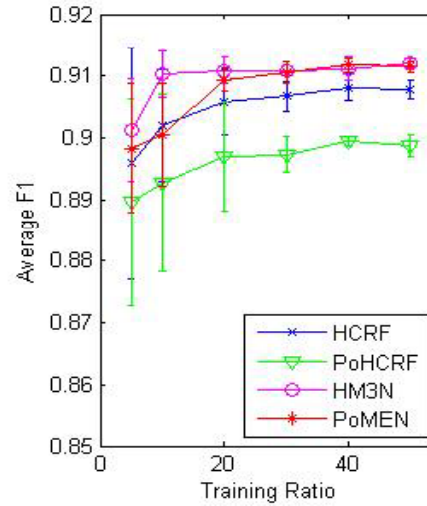


实验结果

性能:

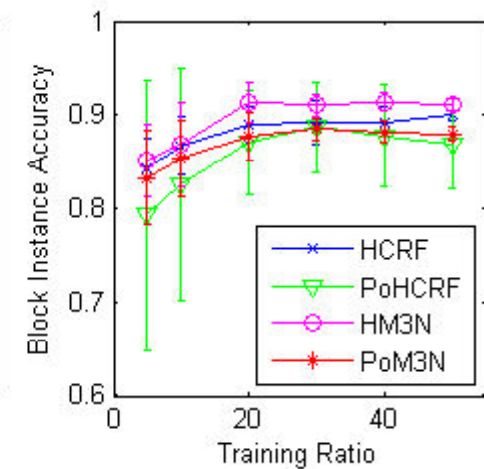
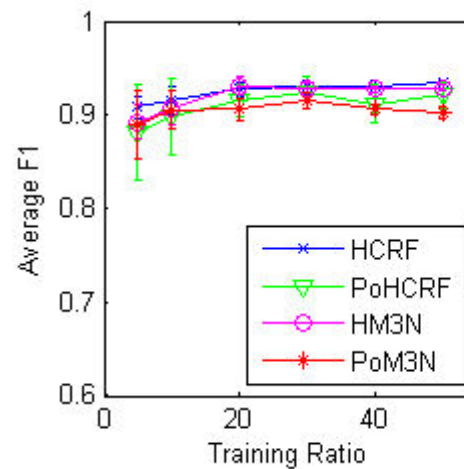
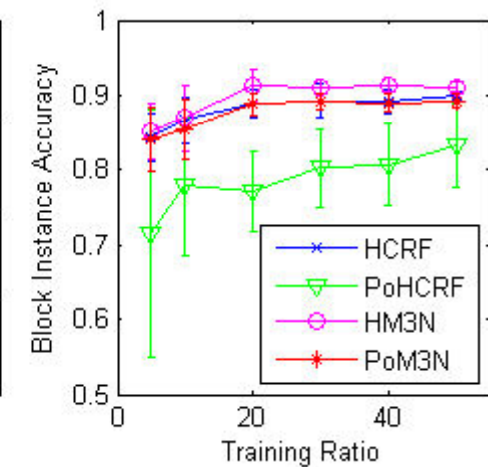
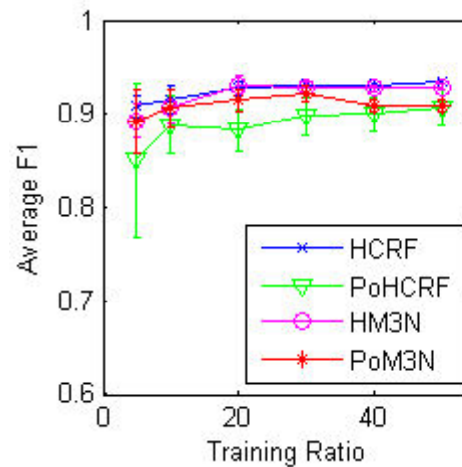
$F1=2 \cdot (\text{precision} \cdot \text{Recall}) / (\text{precision} + \text{recall})$

- Avg F1:
 - avg F1 over all attributes
- Block instance accuracy:
 - % of records



页面级的评价

- Supervision Level 1:
 - Leaf nodes and data record nodes are labeled
- Supervision Level 2:
 - Level 1 + the nodes above data record nodes



相关的论文与专利

[1] Jun Zhu, Zaiqing Nie, Ji-Rong Wen, Bo Zhang, and Wei-Ying Ma.
2D Conditional Random Fields for Web Information Extraction.

Published in ICML'05,

United States Patent 7529761. (Citations: 68)

[2] Jun Zhu, Zaiqing Nie, Ji-Rong Wen, Bo Zhang, and Wei-Ying Ma.
Simultaneous Record Detection and Attribute Labeling in Web Data
Extraction.

Published in SIGKDD'06,

United States Patent 7720830. (Citations: 83)

[3] Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang, and Ji-Rong Wen.
StatSnowball: a Statistical Approach to Extracting Entity
Relationships.

Published in WWW'09, Pending, MS1-4960US.

(Citations: 38)

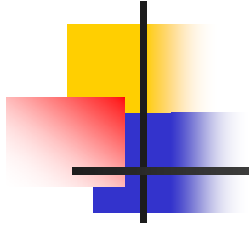
Techniques transferred to Microsoft 's search products:



基于内容信息处理的新趋势

要解决与信息内容相关的处理

- 统计学-处理不确定性
- 灵活的结构化的背景知识表示
- 算法：学习、推理等
- 向人类（大脑）学习



谢谢！