A New Challenge of Information Processing in the Web Age

Bo Zhang
Tsinghua National Lab for Information Science & Technology
State Key Lab of Intelligent Technology & Systems
Information Science & Technology College
Computer Science & Technology Department
Tsinghua University, Beijing China

Classical Information Theory & Processing

- Meaning independent assumption

  -R. Hartley, C. E. Shannon

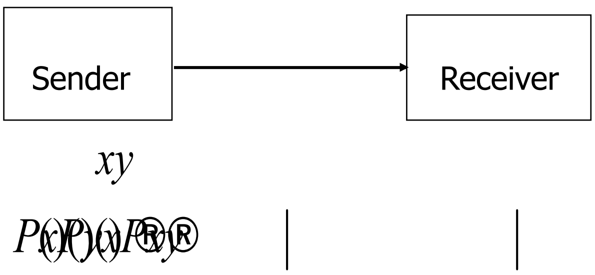
- Probability (Mathematics)-based Information Processing

R. V. L. Hartley, Transmission of information, Bell System Technical Journal, July 1928, pp.535-563

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

Shannon's Communication Theory (1948)

A Communication Model



xy

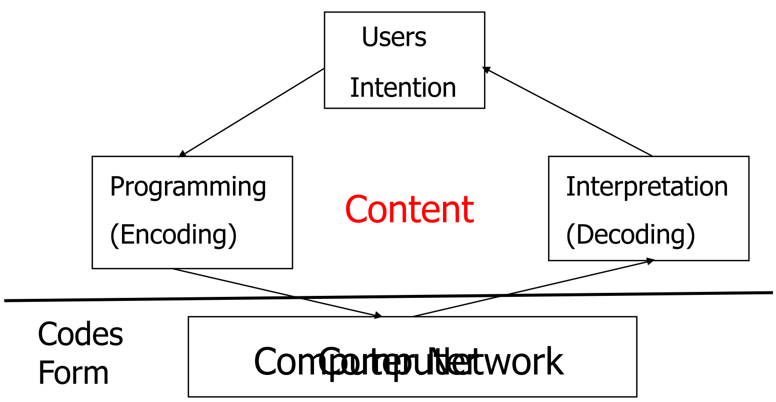
P(x)P(y)(x)P(y)®®

Anstraction: (Markov) Stochastic Process

-C. E. Shannon

From Computer Age to Web Age

Man-Machine Interaction



1. Cognition-based Methods

Cognition-based methods

(Perception-based methods)

refer to a faculty for the processing of information, applying knowledge, and changing preferences.

Symbolic Artificial Intelligence Methods

Production Systems

-A Human-oriented Method

- A Specific Domain Knowledge Representation
- A Common Inference Engine

The Symbolic Domain Knowledge Representations

Rule-based Representation
(declarative knowledge):

Human disease diagnose
Production Rules:

If $a,\dots$ symptoms (fuzzy)

$\longrightarrow$ CF: certainty factors

Then b function disorder (fuzzy)

The Advantages & Disadvantages

- Advantages:

Human-based, Understandable

- Disadvantages

(1) Domain Dependent
(2) Deliberative behaviors except common sense, natural language understanding,…
Perception is not applicable

[1] L A Zadeh, Fuzzy Sets, Information & Control 8, 338-353 (1965)
[2] Shafer, Glenn; *A Mathematical Theory of Evidence*, Princeton University Press, 1976

Syntactic Analysis
- The symbolic representation of images

Detector

Semantically meaningful features:

Boundary, shape,..

There is no clear boundary among parts

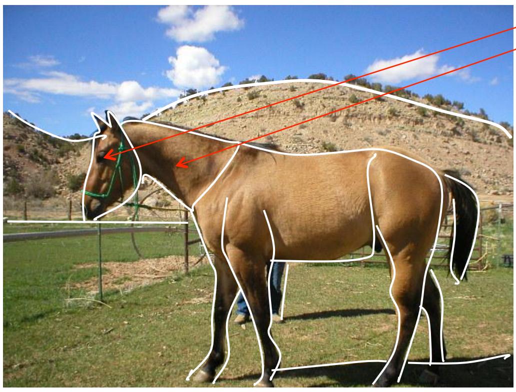
Segmentation Problem

Descriptor

Structural uncertainty

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

The Symbolic Representation -Image Segmentation



If a b,...
then
o (horse)

Where is the object ?

↕

What is the object ?

Chicken or Egg ?

How does the classical information theory to deal with the semantic aspect of image processing ?

2. Statistical Methods -Machine-oriented Methods

Computer representation → Human perception

$F\,(W, D)$ ⟷ S

Data Space
Feature Space

Semantics

Function Regression
Classification

Classical Statistics Theory -Non-deterministic solution

Law of large numbers in function spaces

Parametric Statistics

Assumption: a known function with a few unknown parameters

$$ax^2 + bx + c$$

Three Milestones in Machine Learning Theory

Law of large numbers in functional spaces

Learning process

- The necessary and sufficient conditions for uniform convergence
- The sufficient condition for fast rate of convergence
- The necessary and sufficient conditions for fast convergence for any probability measures

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

Data → Function Semantics

$(\omega_1, x_1), \ldots, (\omega_l, x_l)$ optimization $F(x,y)$ *or* $y = f(x)$

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

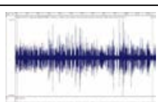
$\phi(x,\alpha), \alpha \in \Lambda, \ \alpha:$ a set of scalar quantities, vectors, abstract elements

If $F(x,y)$ or $f(x)$ exists

Then, data can infer the function in the probabilistic sense 15

The Real Problems

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



Mapping ? → Concepts Semantics

X-representation (Data)

S-semantics

Mapping

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

- Does the mapping (function) between the feature space representation and its meaning exist ?
- Only exists in a certain "dataset" for a certain representation (document, image, speech,...) !

17

A Normalized Alignment Frontal Face Data Base

Extended Yale B

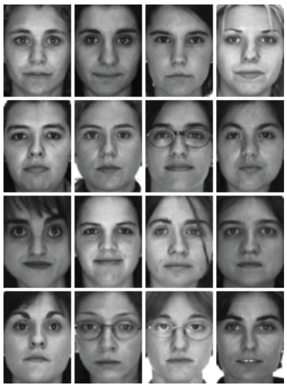
2414 frontal-face

with different lighting

38 individuals

192×168 pixel image

Feature space dimensions: 30, 56, 120, 504



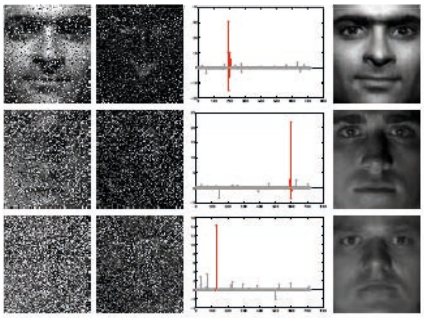
J. Wright, et al. Robust face recognition via sparse representation, IEEE PAMI 09, 31(2):210-227

Anti-Corruption

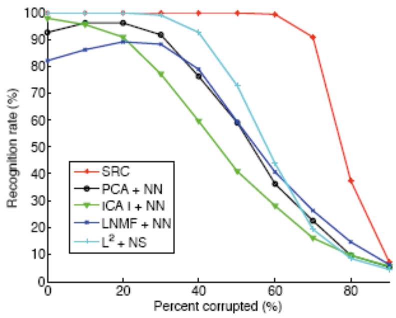
30%

50%

70%



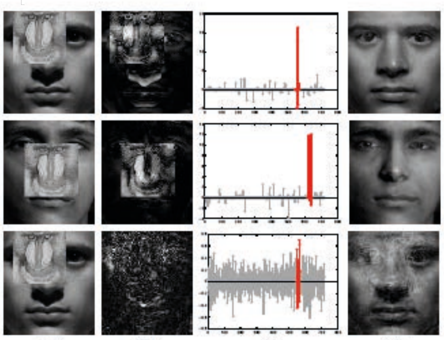
(a) (b) (c) (d)



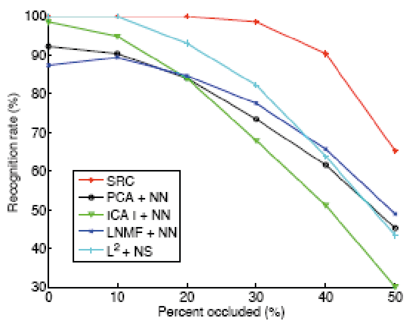
Percent corrupted	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%
Recognition rate	100%	100%	100%	100%	100%	100%	99.3%	90.7%	37.5%	7.1%

SRC-Sparse Representation-based Classification, NN-nearest neighbor, NS-nearest subspace, Extended Yale B. (images)

Anti-Occlusion



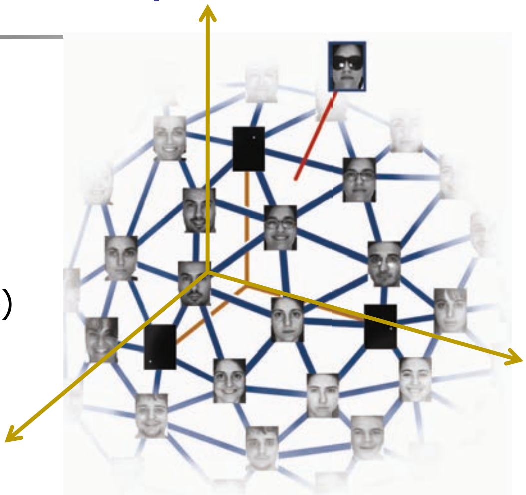
(a) (b) (c) (d)



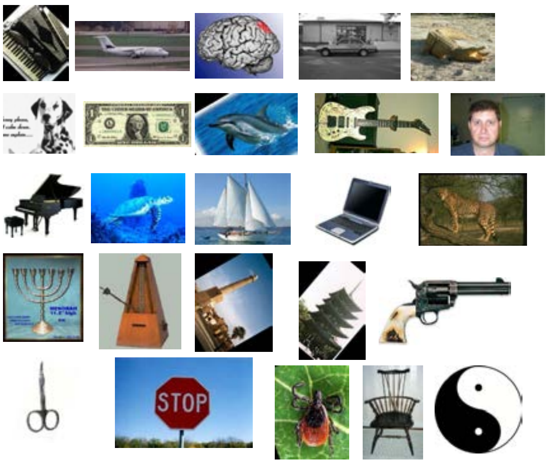
Percent occluded	0%	10%	20%	30%	40%	50%
Recognition rate	100%	100%	99.8%	98.5%	90.3%	65.3%

Low-dimensional Subspace in High-dimensional Space

- The intentionally selective face data base
- The Structure of the sample Space(data space)
- Pixel-based coordinates



The Extended Image Base

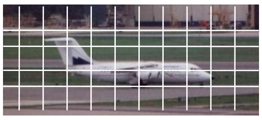
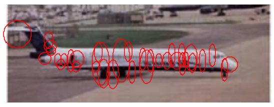


Caltech101 (25 categories, 30 images/category)

Bag of (Visual) Words

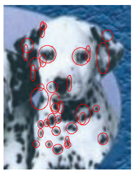
- Defined in image patches (2005-06)
- Descriptors extracted around interest points (2002-2004)
- Edge contours (2005-06)
- Regions (2005-06)



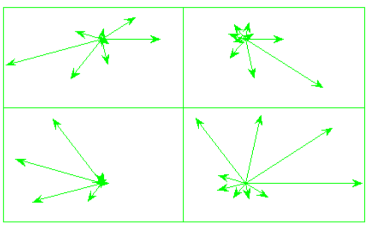



Detector & Descriptor





Kadir salience region (points)

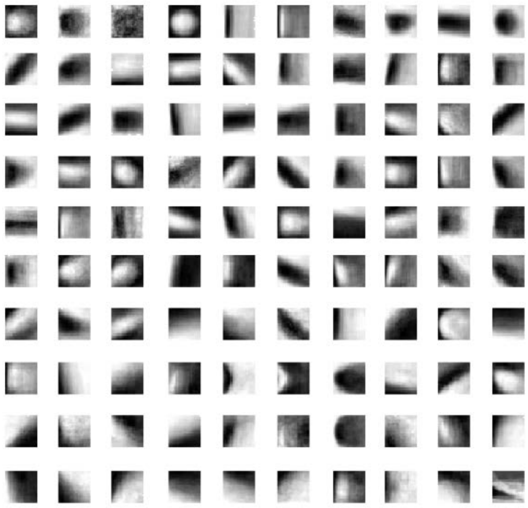


Histograms of Oriented Gradients (HOG)

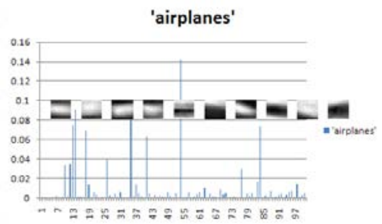
-72 dimension

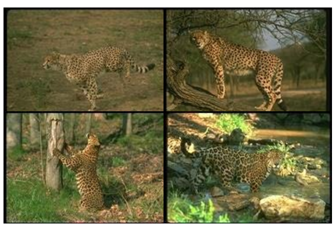
Zuo Yuanyuan, Bo Zhang (2010-)

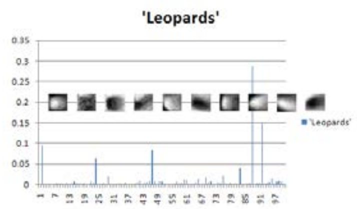
Low-level and Local Visual Words (100)



The Sparse Structure in High-dimensional Data Space

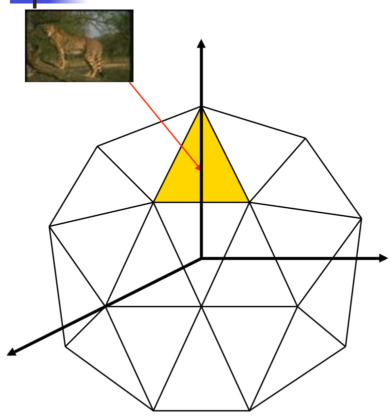






Visual word distribution

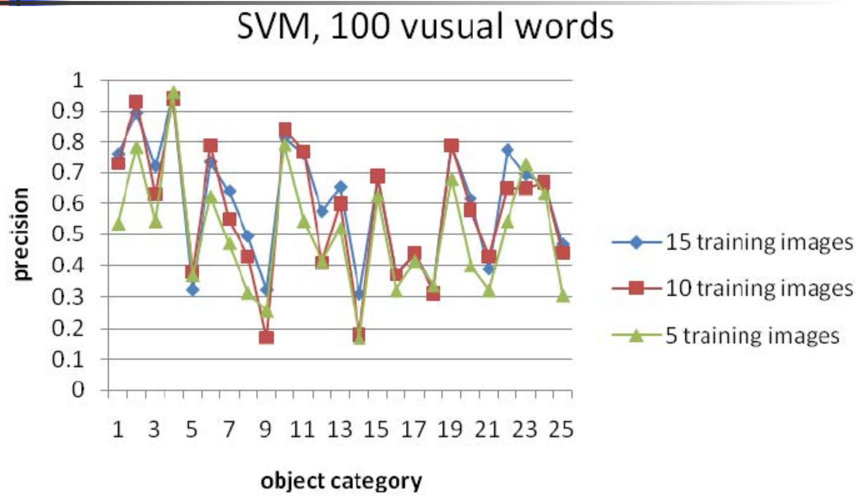
Low-dimensional Structure in High-dimensional Space



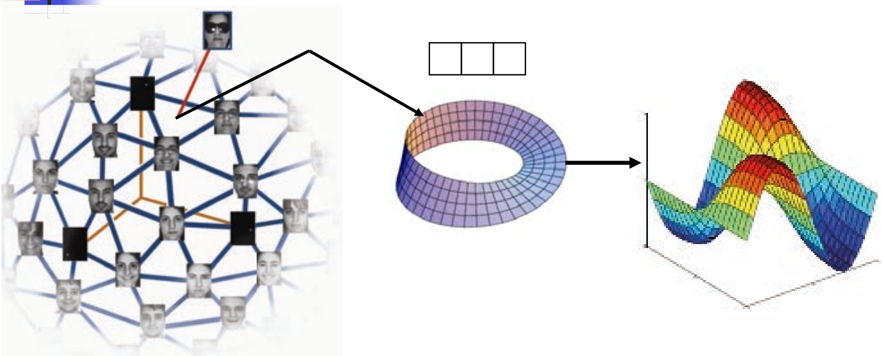
Data Structure

Objects	Precision (SRC)
Airplane	0.947
Car-side	0.987
Dalmatian	0.733
Faces	0.760
Leopard	0.827
Pagoda	0.760
Stop-sign	0.800
Windsor-chair	0.787

Experimental Results



Manifold Learning



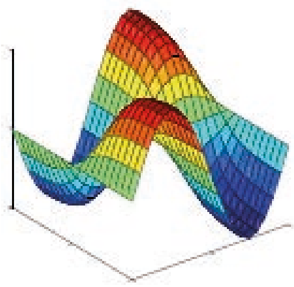
Increase the number of training samples

H. S. Seung, D. Lee, The manifold ways of perception, Vol.290, *Science*, 2000, 2268-2269

Scalability ?

Complex Landscape !

Generalization Capacity ↓



The Semantic Gap
—The weakness of statistical methods

- The semantic gap between low-level local features and high-level global concepts

Less semantically meaningful features: colors or their distribution (histogram), gray-values or their distribution, visual words (descriptors from interest points), image patches, image regions, edge, …

- Lack of structural knowledge

Generalization capacity

Information processing without understanding

The Comparison of Two Frameworks

	Understandable (Man-machine Interaction)	**Knowledge**	**Uncertainty Management**	**Scale**	**Limitation**
Cognition-based	Good	Y	Poor	Small	Domain dependent
Probability -based	Poor	N	Good	Large	Data dependent

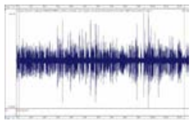
Contextual Knowledge

(Structural Knowledge)

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

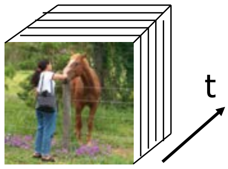
- Text: Contextual structures
- Image: Spatial structure
- Speech: Temporal structure
- Video:

  Temporal-Spatial structure



Information (Contextual)

Structure Acquisition and Prediction



3. The Combination of Two Frameworks

Statistical Inference

over

An Abstract Structured

Knowledge Representation

Contextual Analysis

Markov (language) models -C. Shannon
n-gram (item: phonemes, syllables, letters, words, etc. levels)

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

n=0, unigrams (item occurrence frequency)
n=1, bigrams
n=2, trigrams

2D Flat Structural Models

-Image Region Annotation: horse, sky, mountain, grass, tree

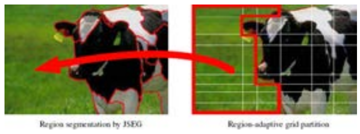
Yuan Jinhui, Bo Zhang (2008-)

Region-adaptive Grid Partition

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

$$p(z) = \frac{1}{Z}\exp\left(\sum_{i,j \in \Omega} g(z_i, z_j)\right)$$

$$p(I|z) = \prod_i p(x_i|z_i)$$



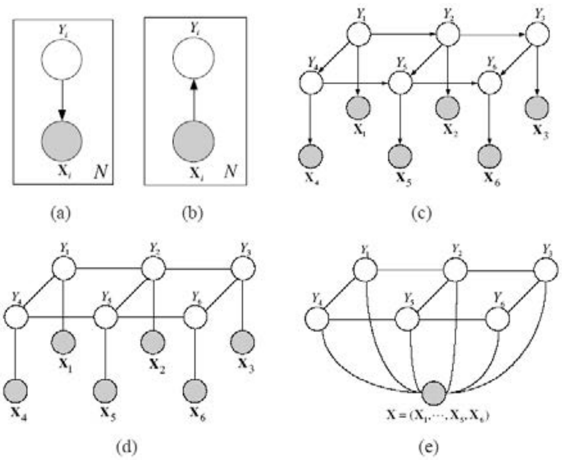
I-data

x_i-image position

z-state (feature)

$g(z_i,z_j)$-the probabilistic region-based constraints (co-occurrence)

The Graphical Models



(a) Generative model
(b) Discriminative model
(c) 2D-Hidden Markov model (HMM)
(d) Markov Random Field (MRF)
(e) Conditional Random Field (CRF)

Experimental Setting

- 4002 Corel images (384×256 or 256×384)
- 11 basic (region) concepts
- Features: color moment + wavelet
- 5 models: 2 without structural knowledge
  (GMM, SVM)
  3 with structural knowledge
  (HMM*, RMF*, CRF*)

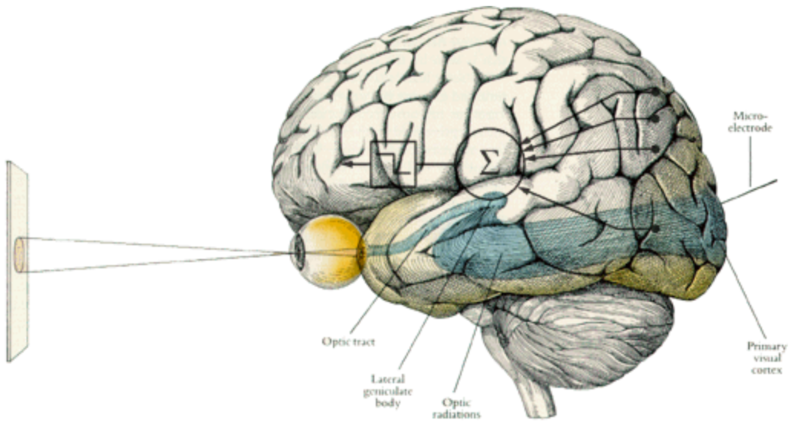
The Categories of Image Region Annotations

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（Moutain）	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

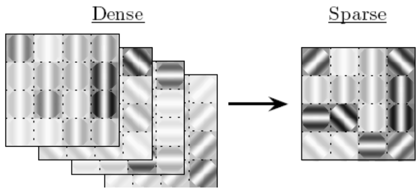
Experimental Results

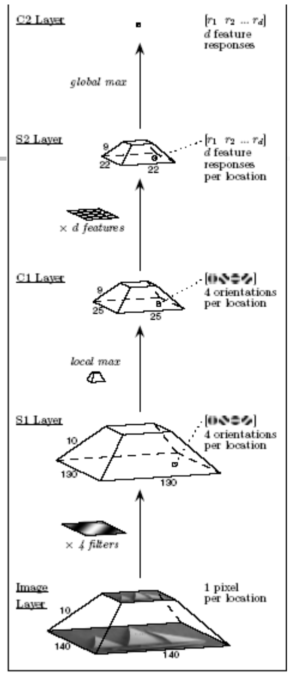
	Precision					Recall					F-score				
	gmm	svm	hmm	mrf	crf	gmm	svm	hmm	mrf	crf	gmm	svm	hmm	mrf	crf
sky	.937	**.961**	.933	.914	.955	.832	.899	.889	.871	**.909**	.882	.929	.911	.892	**.931**
wat.	.410	.583	.531	.390	**.598**	.449	.588	.489	.496	**.639**	.429	.585	.509	.437	**.618**
mnt.	.134	.269	.215	.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**

Learning From Human's Brain

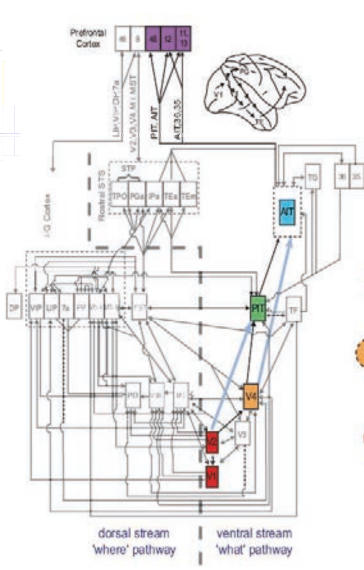


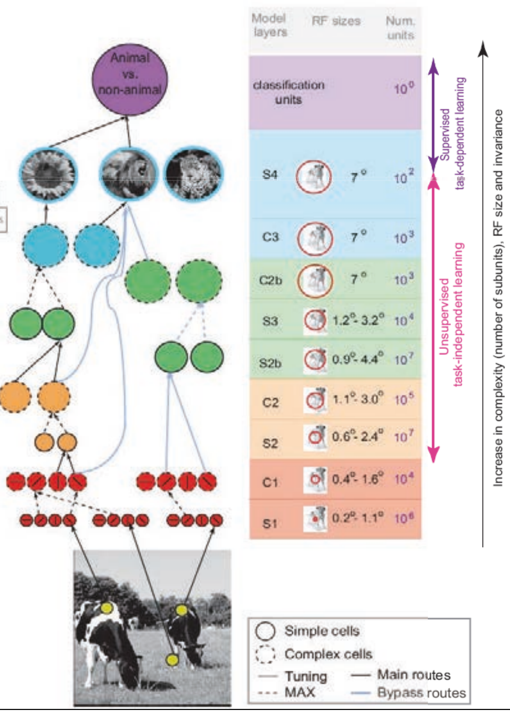
Object Recognition





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





HMAX-sum + max

The Theoretical Framework

Classification

Pooling functions

A set of filters

Input

Kernel functions

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

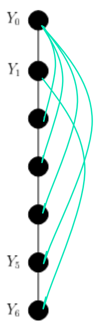
Pooling functions

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	$\psi(\alpha(h)) = \max_{h\in H} \|\alpha(h)\|$

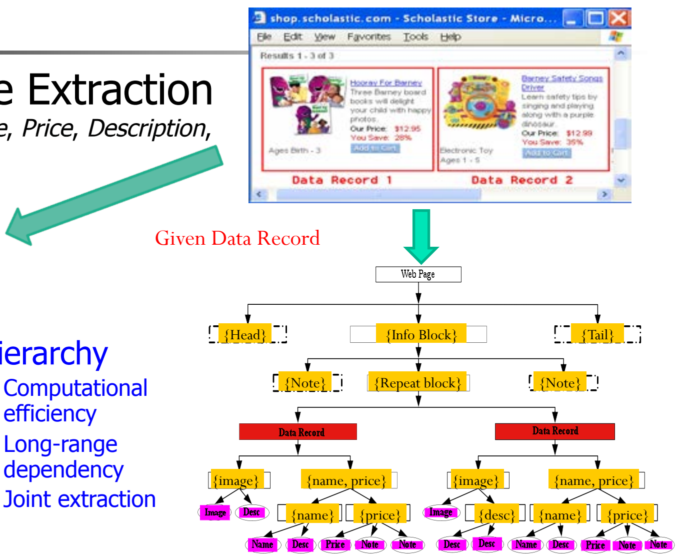
Latent Hierarchical Model

Web Page Extraction

Name, Image, Price, Description, etc.



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



Experimental Setting

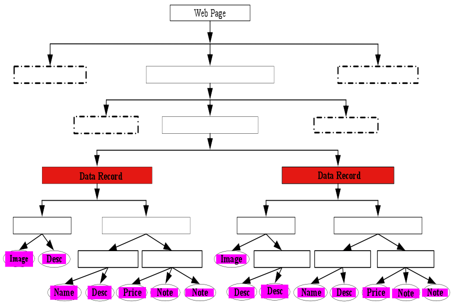
- Web page extraction
- *Name, Image, Price , Description*
- Models

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

Data set: 37 Template

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

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



Structural Prediction Learning

Learning Rules	Classification	Structural Prediction
Maximal Joint Likelihood Estimation	Naïve Bayesian Network	Hidden Markov Model (1966)[1]
Maximal Conditional Likelihood Estimation	Logistic Regression	Conditional Random Field (2001)[2]
Maximal Margin Learning	SVM	Maximal Margin Markov Net (2003)[3]
Maximal Entropy Discrimination Learning	Maximal Entropy Discrimination Model	Maximal Entropy Discrimination Markov Net (2008)[4]

The Related Publications

[1] L. E. Baum and T. Petrie. Statistical Inference for Probabilistic Functions of Finte State Markov Chains. *The Annals of Mathematical Statistics*, Vol. 37, No. 6, pp.1554-1563, 1966

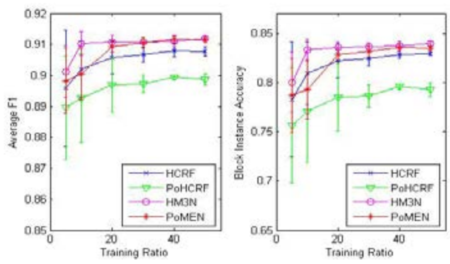
[2] J. Lafferty et al. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. of International Conference on Machine Learning (ICML)*, 2001

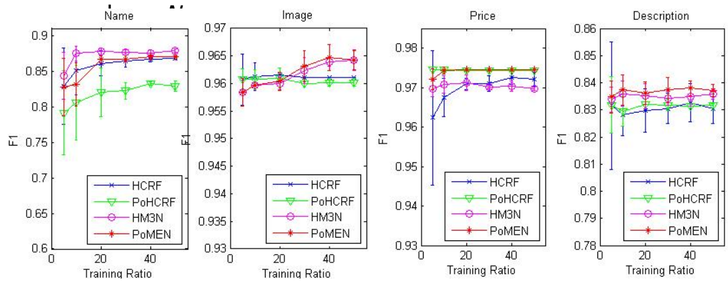
[3] B. Taskar et al., Max-Margin Markov Networks. *Advances in Neural Information Processing Systems (NIPS)*, 2003

[4] J. Zhu et al., Laplace Maximum Margin Markov Networks, In *Proc. of International Conference on Machine Learning (ICML)*, 2008

Experimental Results

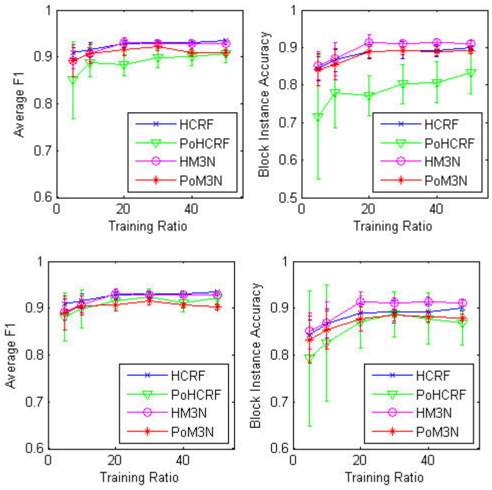
- Performances:
- F1=2 (precision . Recall) / (precision + recall)
  - Avg F1:
    - avg F1 over all attributes
  - Block instance accuracy:
    - % of records





Page-Level Evaluation

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



Related Publications & Patents

[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:

Ohters:

Theoretical foundation (JMLR 2009, ICML 2008, ICML 2009a);
Latent factor models (NIPS 2008, ICML 2009b, ICML 2010, NIPS 2010a,b, JMLR 2011, PAMI 2011, UAI2011
Non-parametric Bayesian (ICML 2011).
Structure learning of Markov networks (NIPS 2010c, SIGKDD 2010);
Multi-task learning (NIPS 2010d);
Sparse topical coding (UAI 2011, SIGKDD 2011, ICCV 2011).

Multi-disciplinary Research at Tsinghua University

Center for Cognitive and Neural Computation

Tsinghua University, Beijing

- Computational Neuroscience
- System Neuroscience
- Intelligent Technology and Systems
- Neural Information and Brain-computer interface
- Learning and Memory
- Cognitive Psychology

Thank You !