
Object Categorization

--- Modeling, Learning and Recognition by Stochastic Image Grammar

Song-Chun Zhu

University of California, Los Angeles
Lotus Hill Research Institute, China

Ref: S.C. Zhu and D. Mumford, "A Stochastic Grammar of Images", *Foundations and Trends in Computer Graphics and Vision*, Vol.2, No.4, pp 259-362, 2006.
downloadable from my website.

Plan of this talk

Objective: to handle large intra-category structural variability.

1, Conceptualization:

How do we define the concept of a category, i.e. the set of all valid instances?

2, Modeling

A grammar is embodied in an [And-Or graph](#).

Define a probabilistic model on the And-Or graph to account for the natural statistics.

3, Image annotation and ground truth

Constructing a large human annotated database

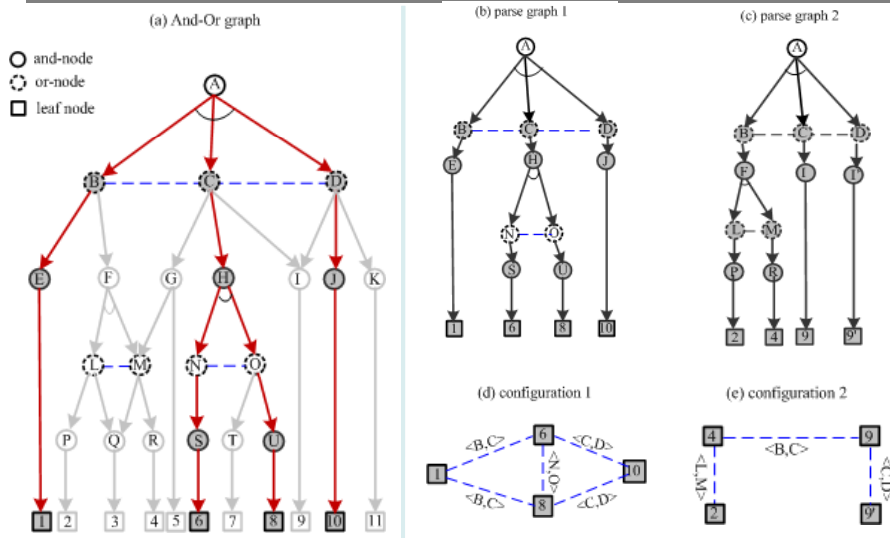
4, Learning

Learning from a relatively small data set and generalizing by MCMC sampling.

5, Computing and parsing

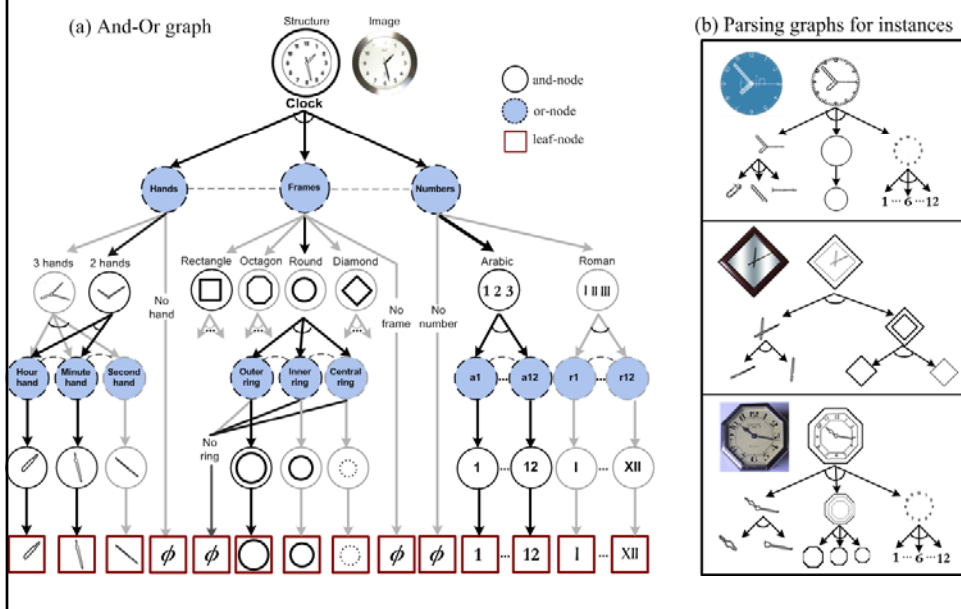
Recursive bottom-up / top-down inference.

Define: And-Or graph, parse graphs, and configurations

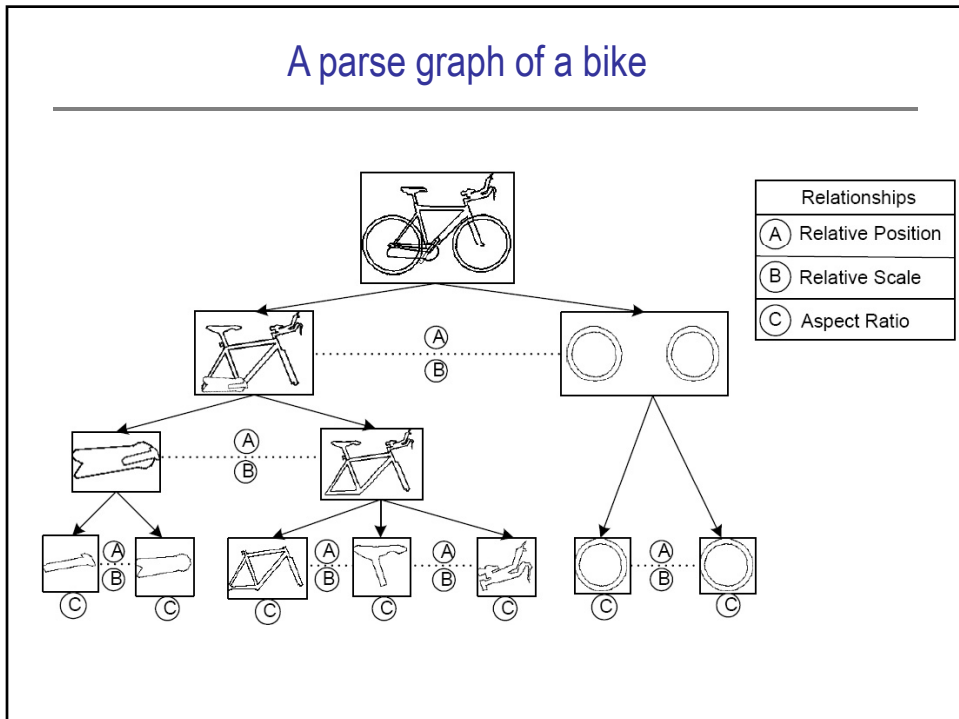


Each category is conceptualized to a grammar whose language defines a set or "equivalence class" for all the valid configurations of the each category.

An example: the clock category



A parse graph of a bike



A relation is like a non-linear filter in low level vision

Some examples

Position	Scale	Orientation	Contained	Hinged	Attached	Butting	Concentric
Low Level Relationships				High Level Relationships			

A **binary relation** is set of links between selected nodes.

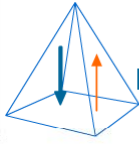
It is applied to selected sites and returns a value (scalar or binary).

Suppose A is a vector of attributes for all nodes

$$A = (a_1, a_1, \dots, a_n)$$

$$r_{ij} = f(a_i, a_j)$$

A large scale human annotation project at Lotus Hill, China

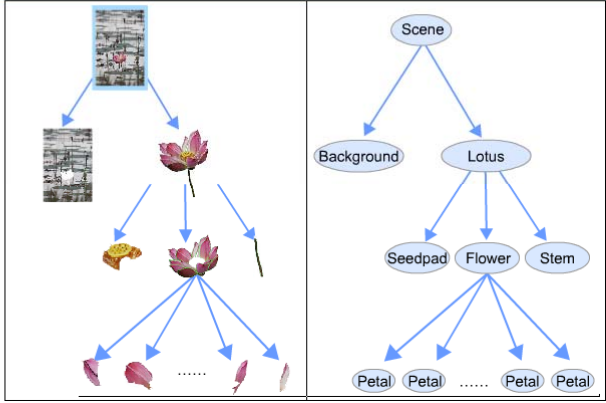


ImageParsing.com

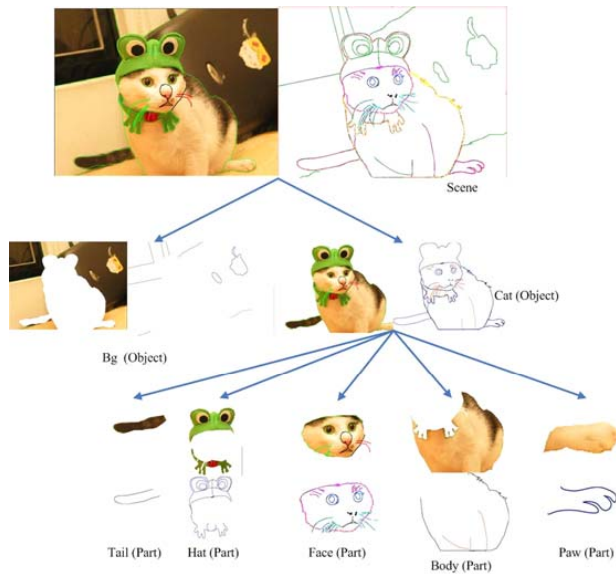
Tel: +86-711-3876688, +86-711-3867183
Fax: +86-711-3876699
Contact person:
Julia Xia, wenhuaxia@gmail.com
Michael Yang, xyang.lhi@gmail.com

About us ⓘ

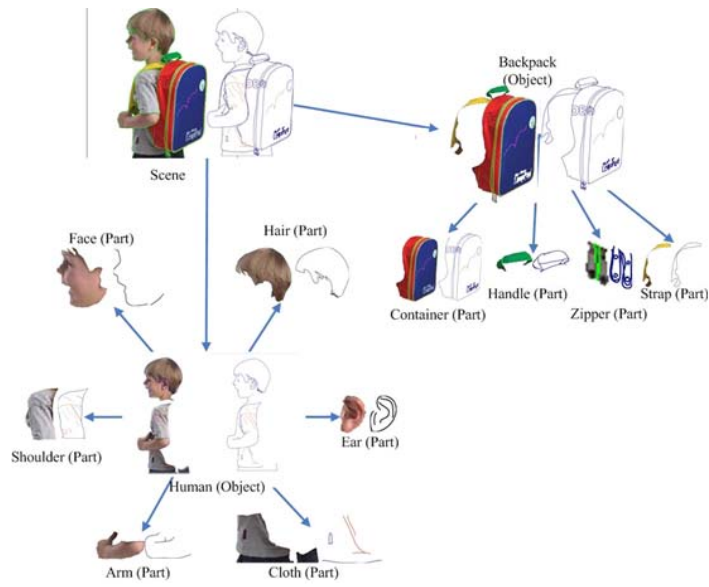
- Home
- About Us
- Image Parsing
- Related Publications
- Data Examples
- Services
- Download Free
- Client Comments
- Acknowledgments



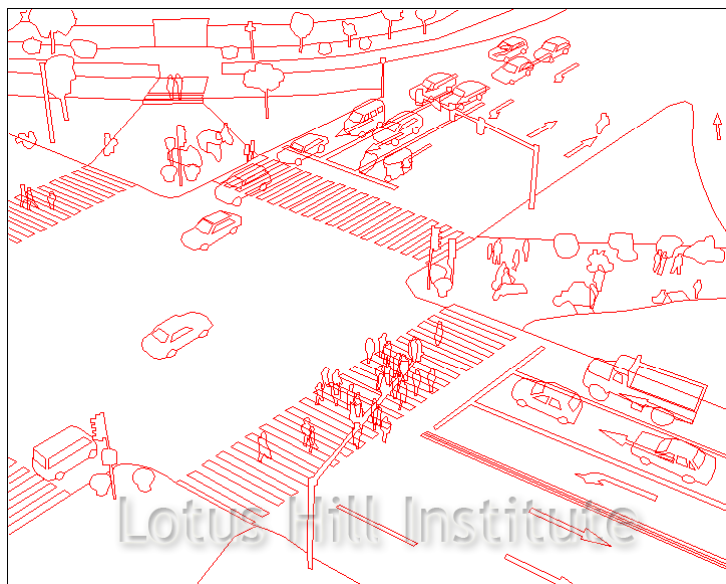
An example: parse graph of a cat



An example: parse graph of a boy with bag

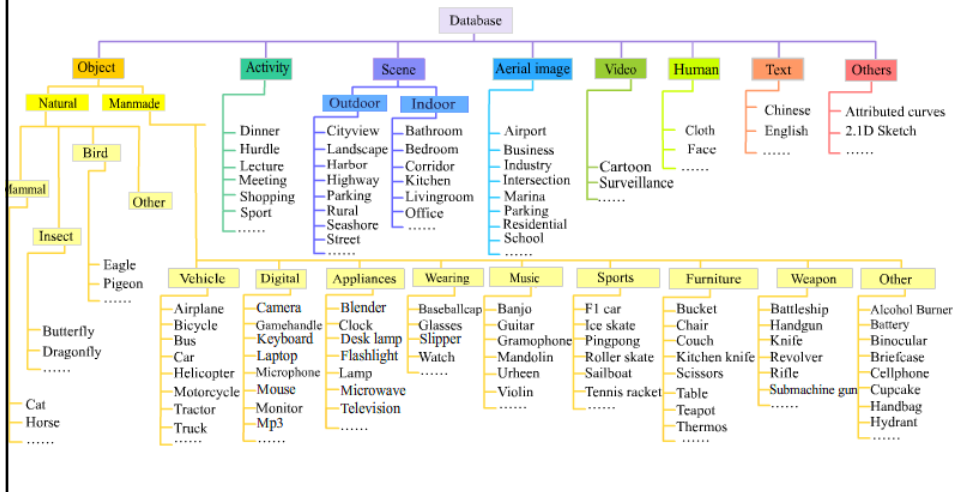


An example: clip for surveillance video



It includes many datasets

280 object categories, 20 scene categories, video, text, segmentation, grouping with ~3,000,000 nodes.



The Probability model on the And-Or graph

Denote:

G ---- a parse graph,

$U(G)$ ---- the set of Or-nodes in G ,

$V(G)$ ---- the set of the And-nodes + leaf nodes in G

$R(G)$ ---- the set of relational links between nodes in G .

The probability model is defined as

$$p(G; \Delta, R, \theta) = \frac{1}{Z} \exp \left\{ - \sum_{u \in U(G)} \lambda(u) - \sum_{v \in V(G)} \varphi(v) - \sum_{r_{ij} \in R(G)} \psi(r_{ij}) \right\}$$

The first term alone stands for a SCFG.

The second and third terms are Markov potentials.

For a context sensitive attribute grammar:

the hard relations / constraints affect the frequency at the or-nodes.

By analogy to texture modeling

observed synthesized

pursuing the relations by minimax entropy

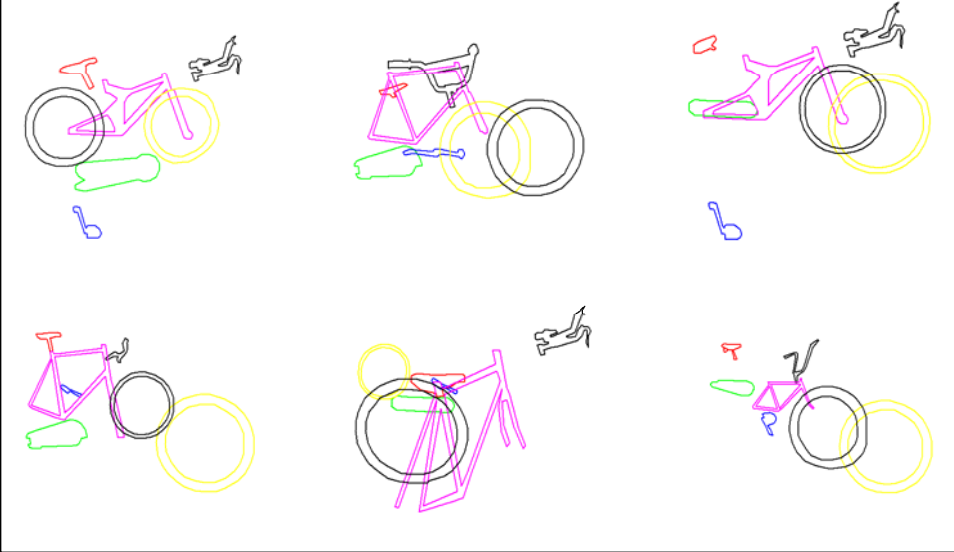
(a) (b) (c) (d) (e)

Sampling Clock: we keep evolving $O(100)$ samples

Poway, Yao, and Zhu, 2006-07

Examples of sampling bicycles

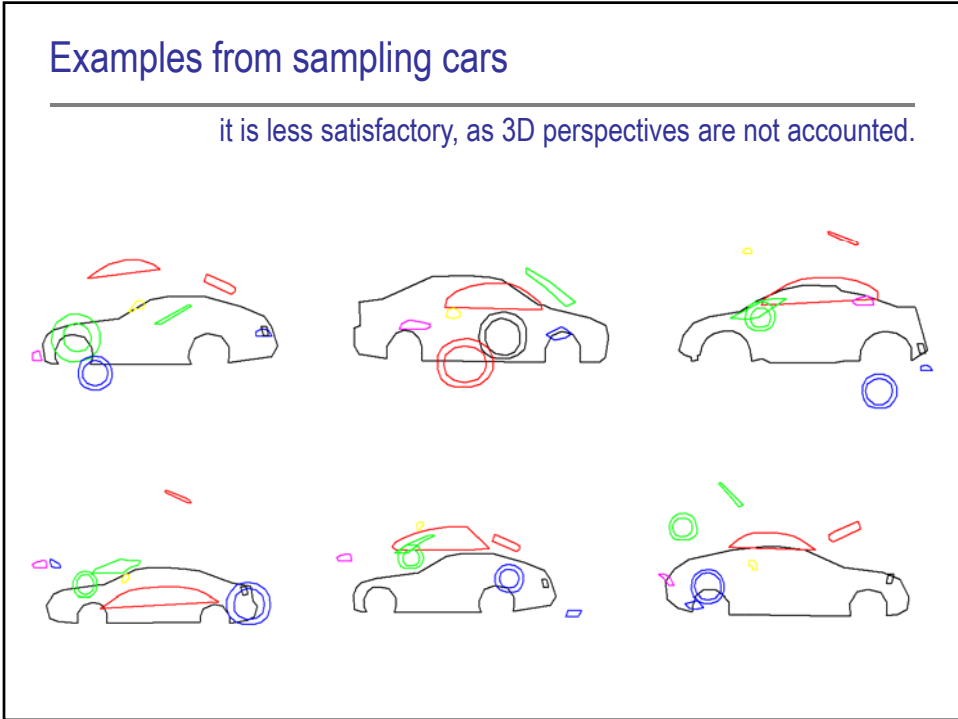
Poway, Yao, and Zhu, 2006-07



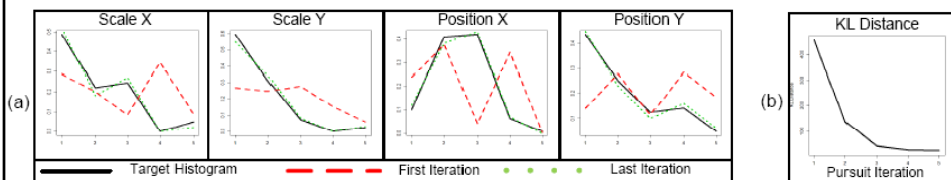
Learning and sampling a bike model	(a)	A collection of individual bicycle components scattered across the frame, including two wheels, a frame, a seat, handlebars, and pedals.
	(b)	A collection of individual bicycle components, similar to (a), but with different color assignments for the parts.
	(c)	Three complete bicycles shown from a side perspective, each with a different color scheme for its frame and wheels.
	(d)	Three complete bicycles shown from a side perspective, similar to (c), but with more varied color schemes for the frames and wheels.

Examples from sampling cars

it is less satisfactory, as 3D perspectives are not accounted.



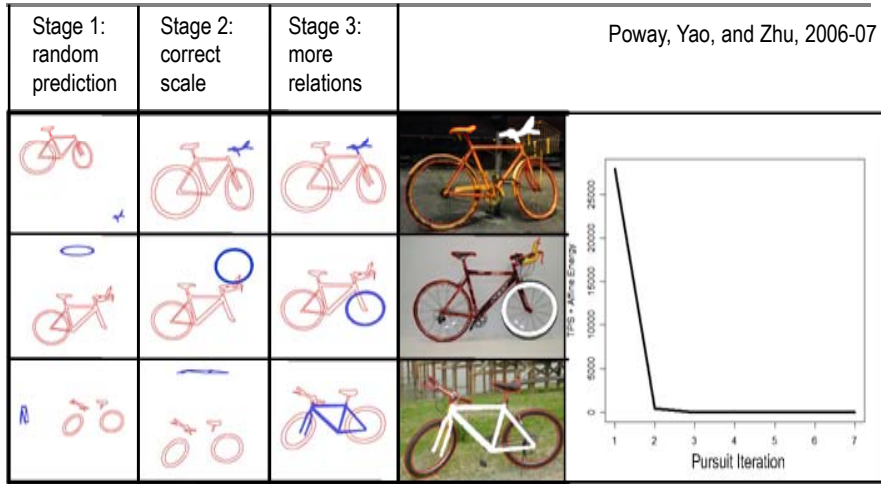
Iterative learning to match the statistics (histogram)



Results of the learning procedure.

- Histograms for four pairwise relationships at different iterations. The last iteration matches the observed histogram quite closely.
- The KL divergence between the current and target model as the relationship pursuit is performed.

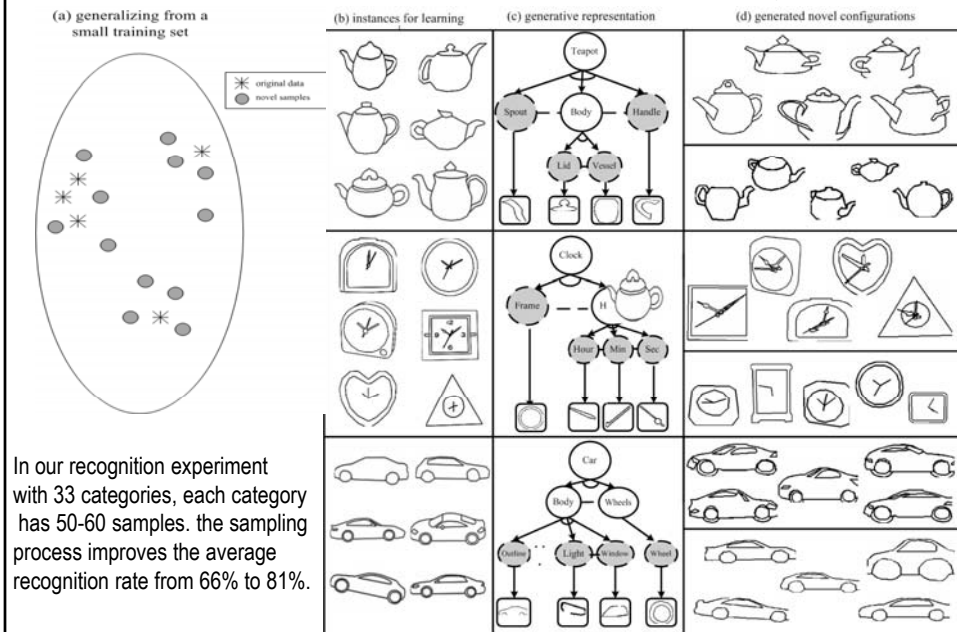
Top-down prediction by sampling the missing part



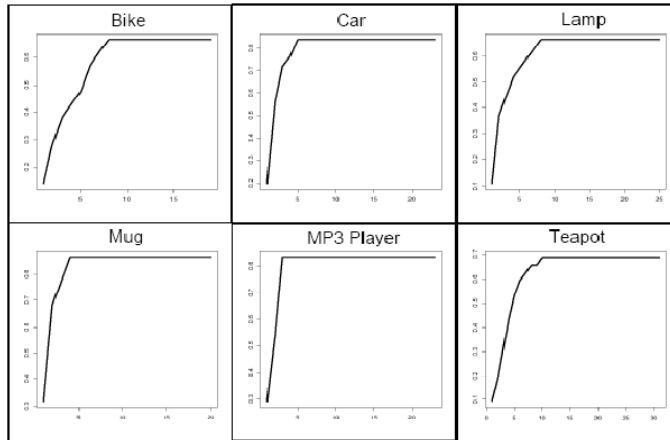
The blue parts are predicted by the learned models at various learning stages

Learning from a small training set & generalization by sampling

Poway, Yao, and Zhu, 2006-07



What is the smallest sample set for training?



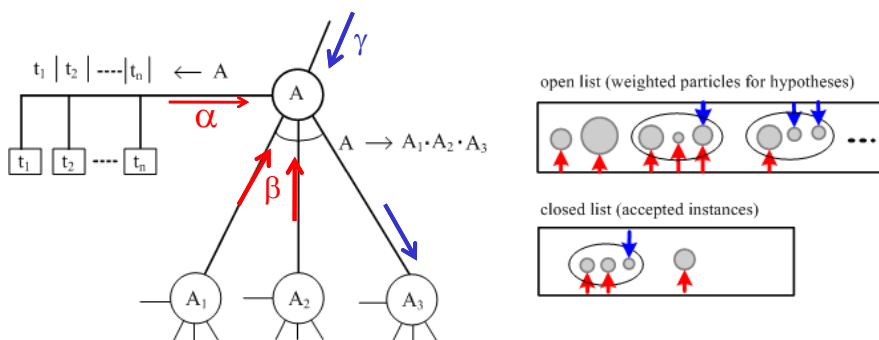
Coverage results for 6 categories. Obviously, as training size increases we cover more of the test set. However, we usually only need a small fraction of the training set to maximally cover the testing set.

Recursive computing and parsing

The And-Or graph is a recursive structure.

we only need to consider a single node A .

- 1, any node A terminate to leaf nodes at a coarse scale.
- 2, any node A is connected to the root.



Compositional boosting, T.F. Wu et al, CVPR 07

Top-down / Bottom-up Inference at all levels

Starting the $\alpha/\beta/\gamma$ channels when they are applicable ---an optimal scheduling problem

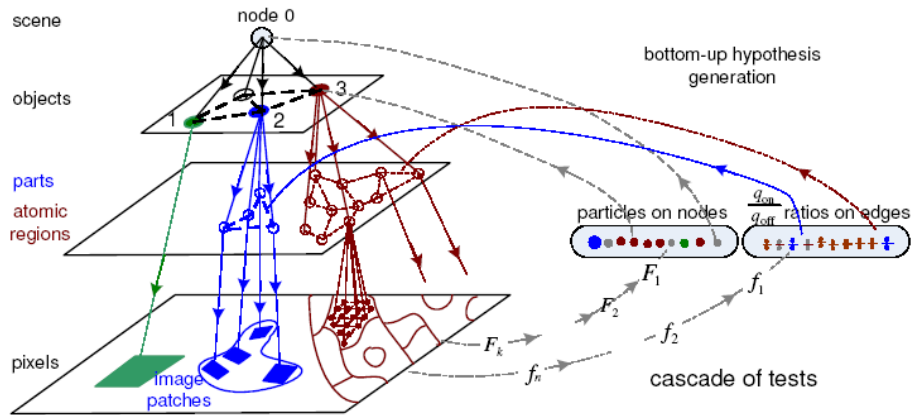
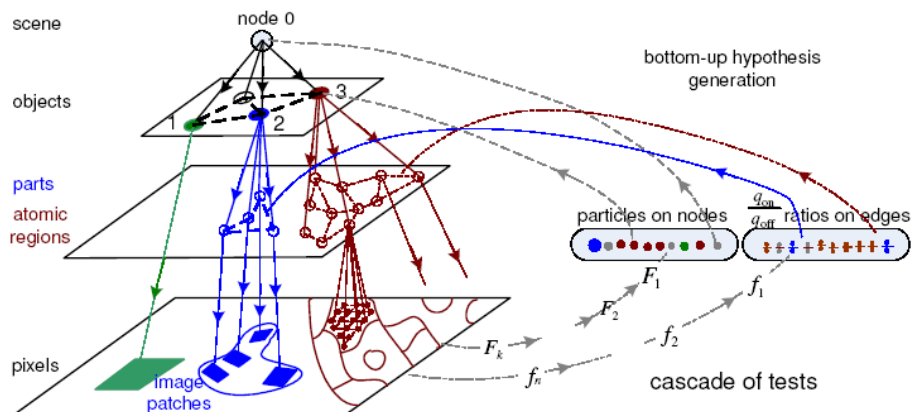


Image parsing by Tu et al, 2002-05

Top-down / Bottom-up Inference

Integrating generative and discriminative methods



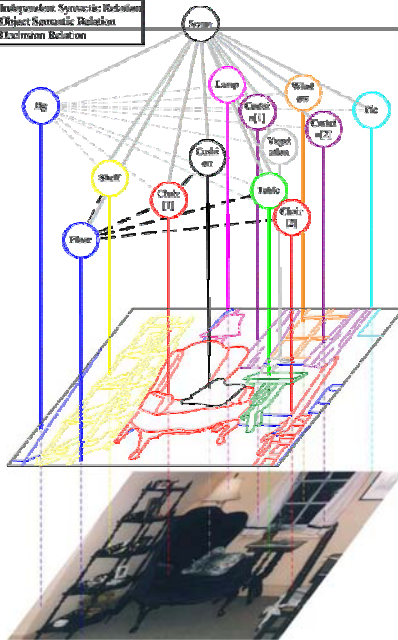
Previous work on image parsing

输入图像



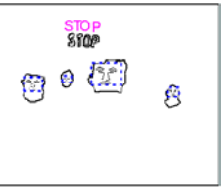


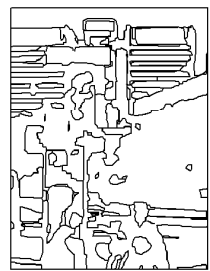




Parse graph with horizontal relations

- Independent Synonymic Relations
 - Object Semantic Relations
 - Extension Relation


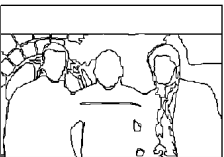



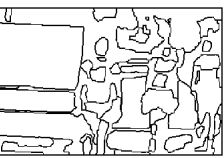



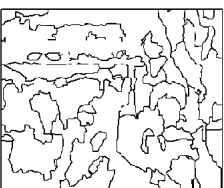




Examples of Image Parsing

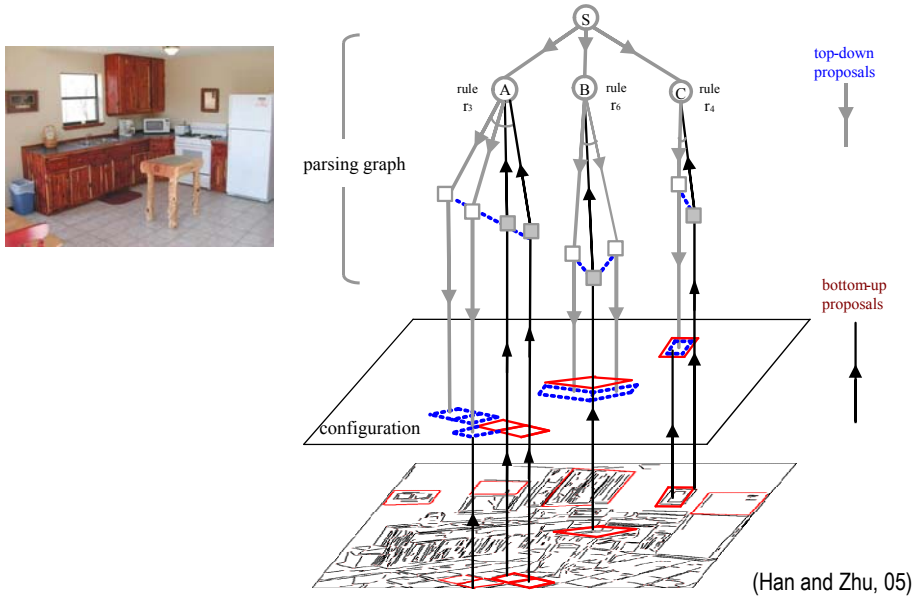
Input	Regions	Objects	Synthesis
			
			

Tu, Chen, Yuille, and Zhu, iccv2003

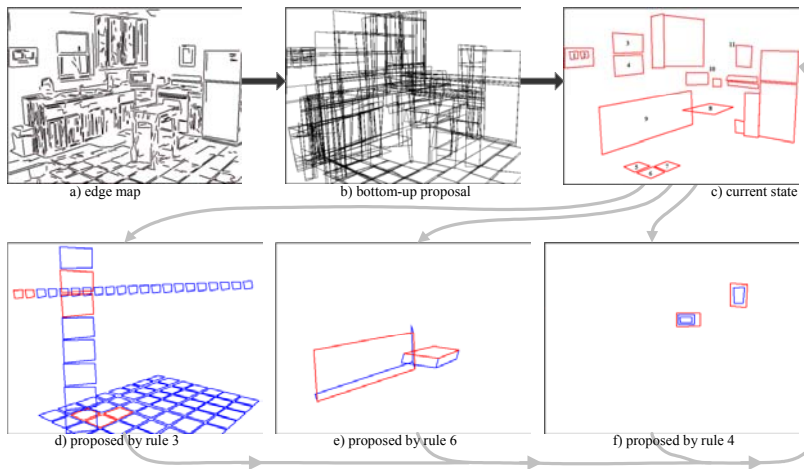
Image Parsing Results

Input	Regions	Objects	Synthesis
			
			
			

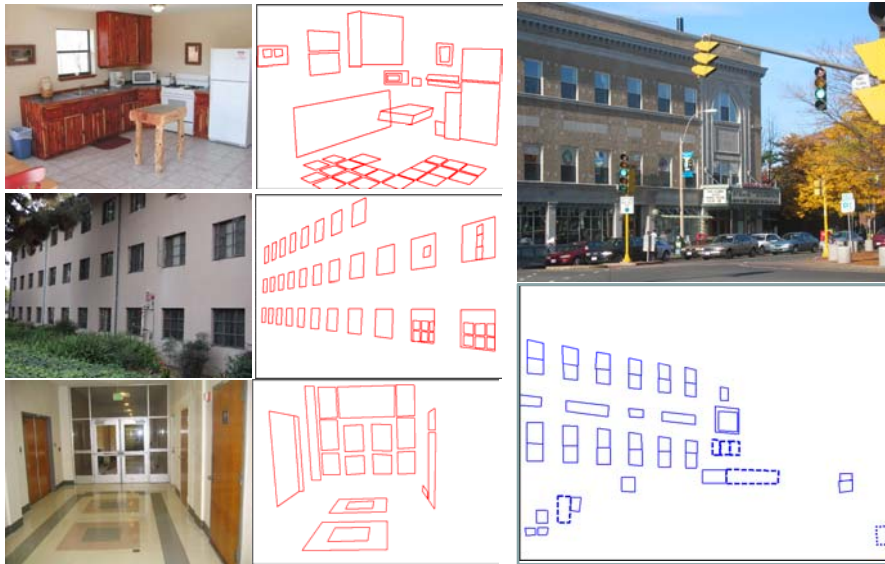
Example: Bottom-up / Top-Down Inference of Rectangular Scenes



A slap shot of the inference algorithm



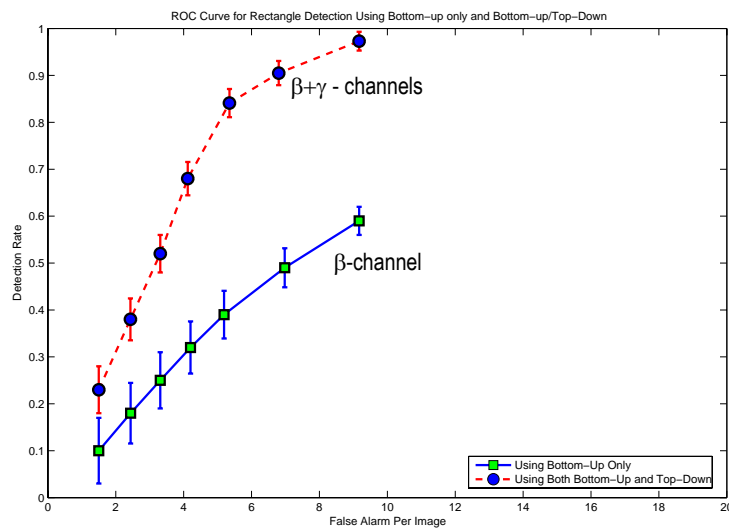
Case study I: parsing rectangular scans by grammar



How much does top-down improve bottom-up?

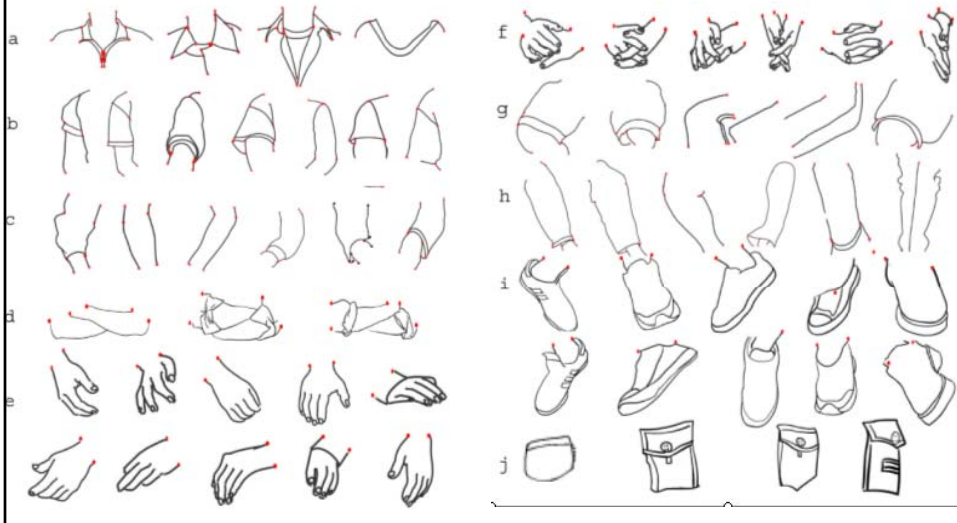
In the rectangle experiments:

Han and Zhu, 2005-07

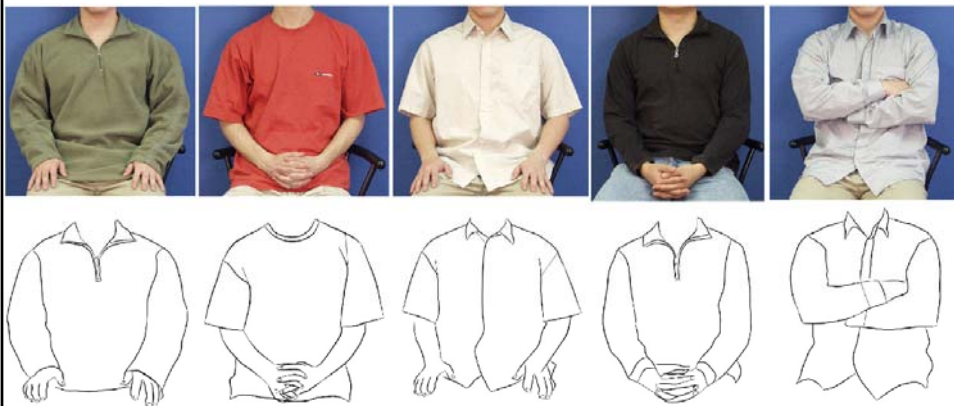


Case study II: parsing human upper-clothes

Elements in the dictionary of human figure



Case study II: parsing human upper-clothes



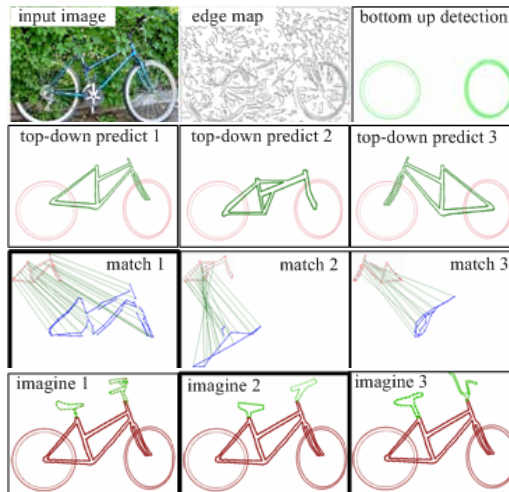
Case study II: parsing human upper-clothes

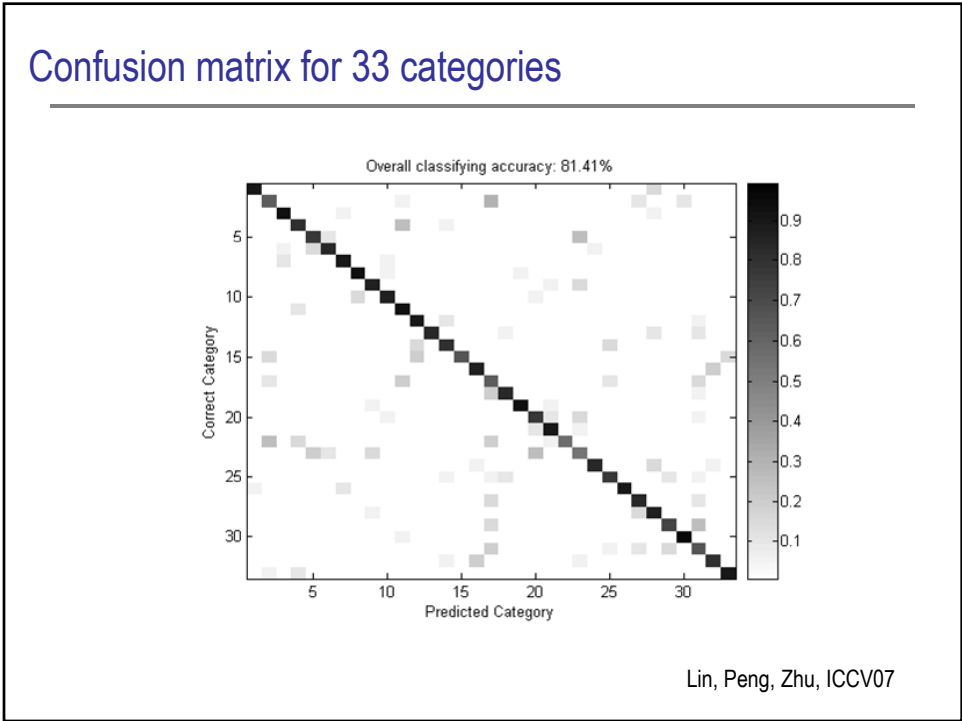
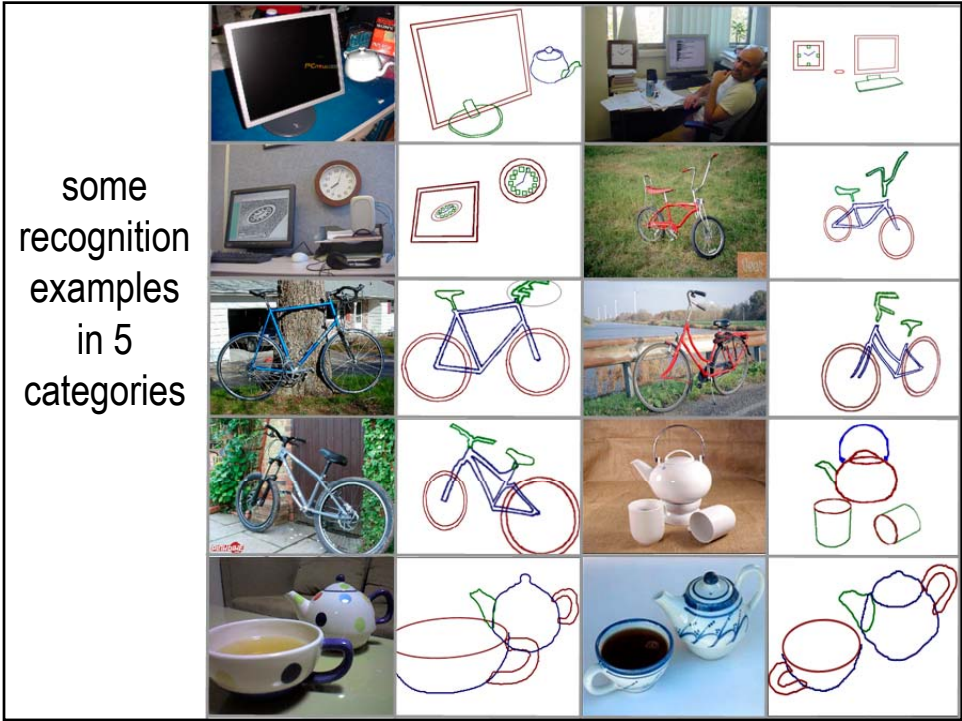


Chen, Xu, and Zhu, CVPR, 2006

Case study III: object category recognition

Example of bottom-up proposing and top-down prediction / hallucination





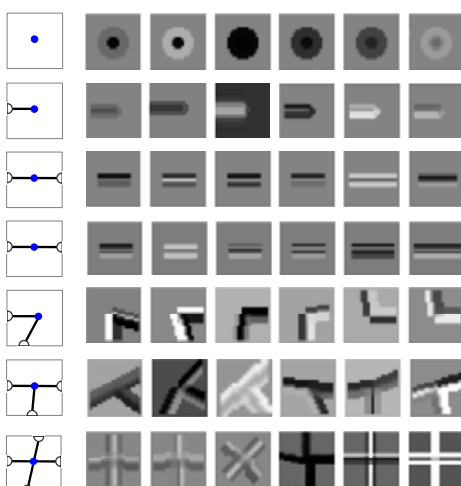
Message to students: three personal opinions

- 1, Object categorization is a finite problem,
---- We need a Google mentality !

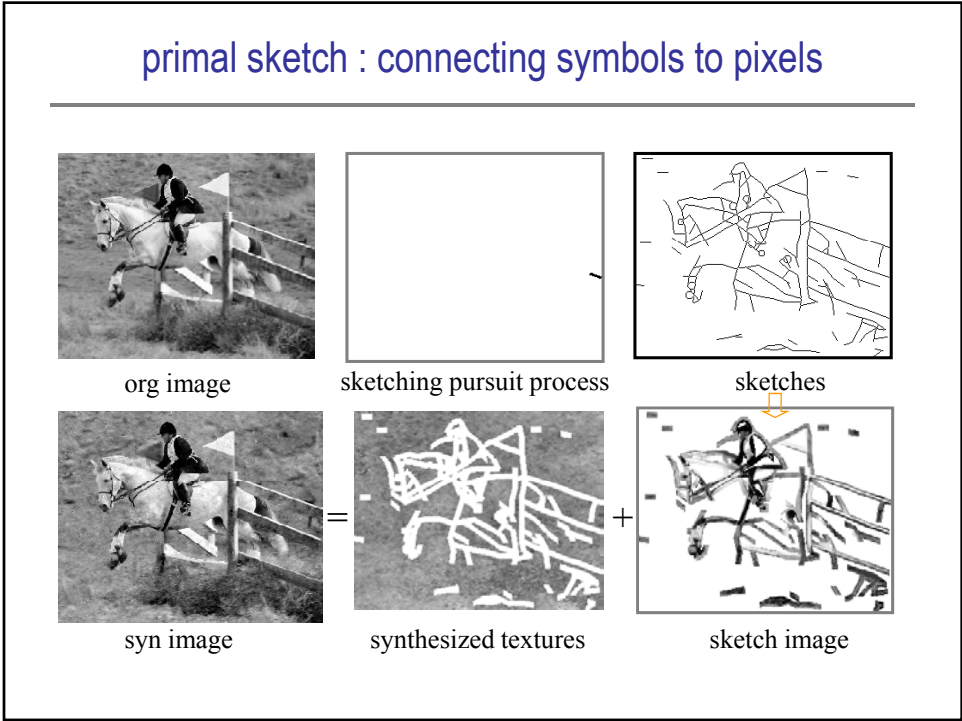
- 2, High level vision needs structures and supervised learning.
---- You thou not feel ashamed for using your hands.
Let's play basketball, go beyond soccer !

- 3, The pendulum swings from statistics to computer science,
---- Study grammar, parsing, compiler, architecture !

Examples of the dictionary of image primitives



primal sketch : connecting symbols to pixels



primal sketch is a 2-layer MRF model

