SIG-12: Tutorial on Stochastic Image Grammar

for Object, Scene and Event Understanding

Song-Chun Zhu, Sinisa Todorovic, and Ales Leonardis

At CVPR, Providence, Rhode Island

June 16, 2012

Layout of topics and lectures

Two axes:

- *Theoretical foundations:* a unified representation (spatial, temporal and causal and-Or Graphs), inference and learning.
- *Vision problems:* parsing objects, scenes, and events; answering what, who, where, when, and why.

		Vision problems		
		Objects	Scenes	Events
Theoretical foundations	Spatial Temporal Causal			
	Inference			
	Learning			

Lecture 1: Introduction and Overview

Song-Chun Zhu
Center for Vision, Cognition, Learning and Arts
University of California, Los Angeles

At CVPR, Providence, Rhode Island
June 16, 2012

Scope of Stochastic Image Grammar

1, Representation -- defining probability models on a set of graphs

Syntactic pattern recognition,
Hierarchical models,
Compositional models,
Reconfigurable models,
Context free/sensitive grammars,
Attributed grammars,
Probabilistic logic,
Sum-Max logic network.

Many names for different aspects of the same thing!

2, Inference

Scheduling top-down / bottom-up computing processes,
Goal guided and cost sensitive computing

3, Learning

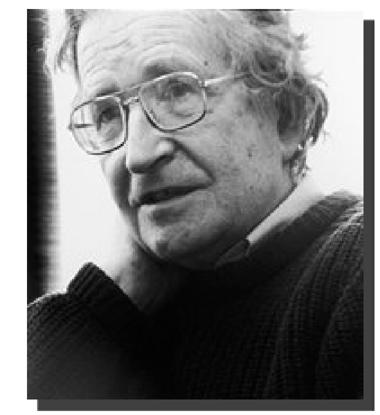
Structure and parameter learning,
Deep learning* (in some community),
Learning rate (PAC, transfer, curriculum),
Regimes of models.

History of Grammar

First recorded grammar originated in India 6th c. BC

"Art of Grammar" -- treatise on Ancient Greek, 2nd c. BC

Transformational-generative grammar -- Chomsky 1957



Derive structured objects in a formal language.
Predict if any utterance is a valid sentence

Syntactic pattern recognition --- K.S. Fu in the 1970s

Grammar of patterns

Example of grammar from K.S. Fu

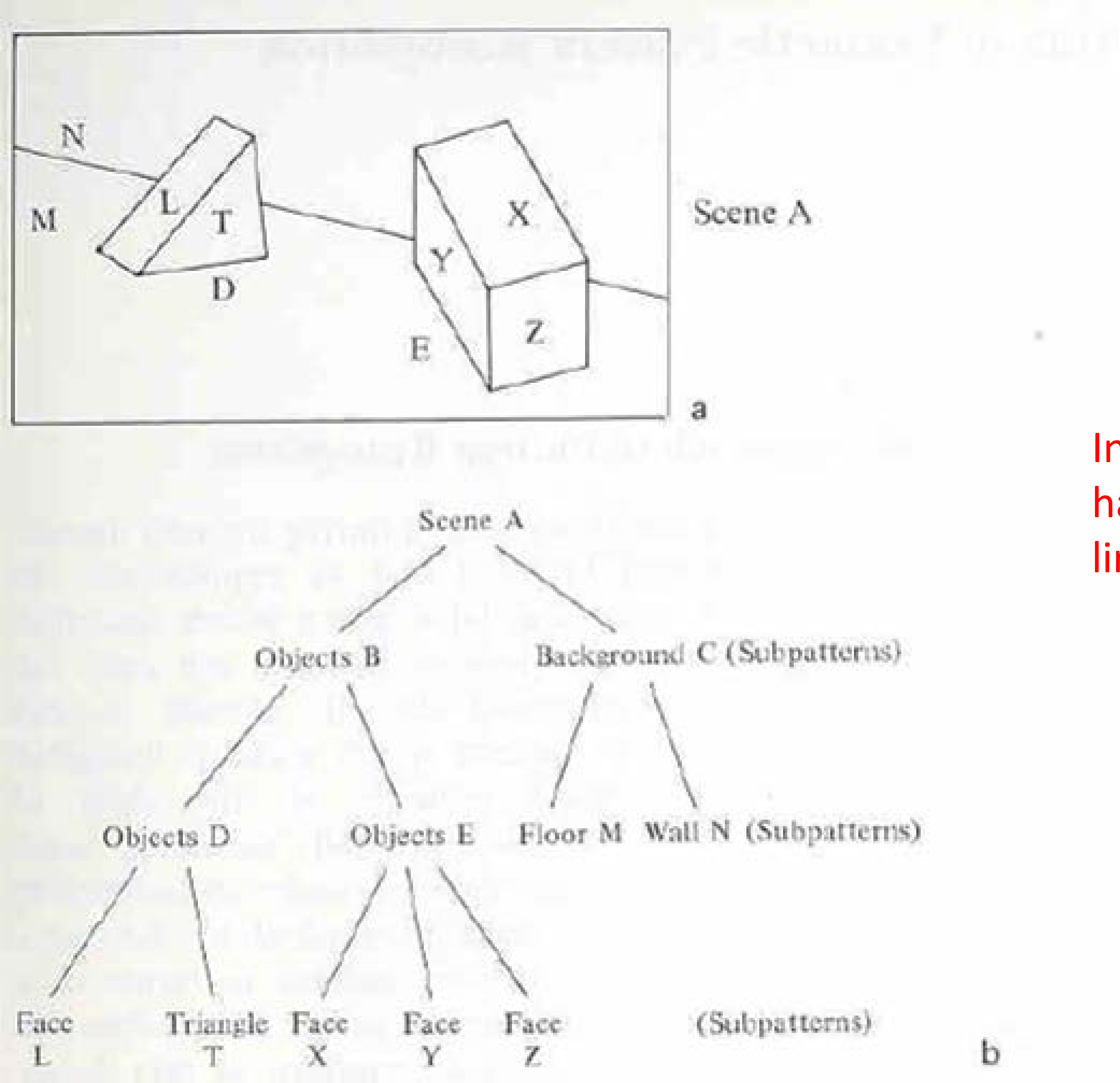
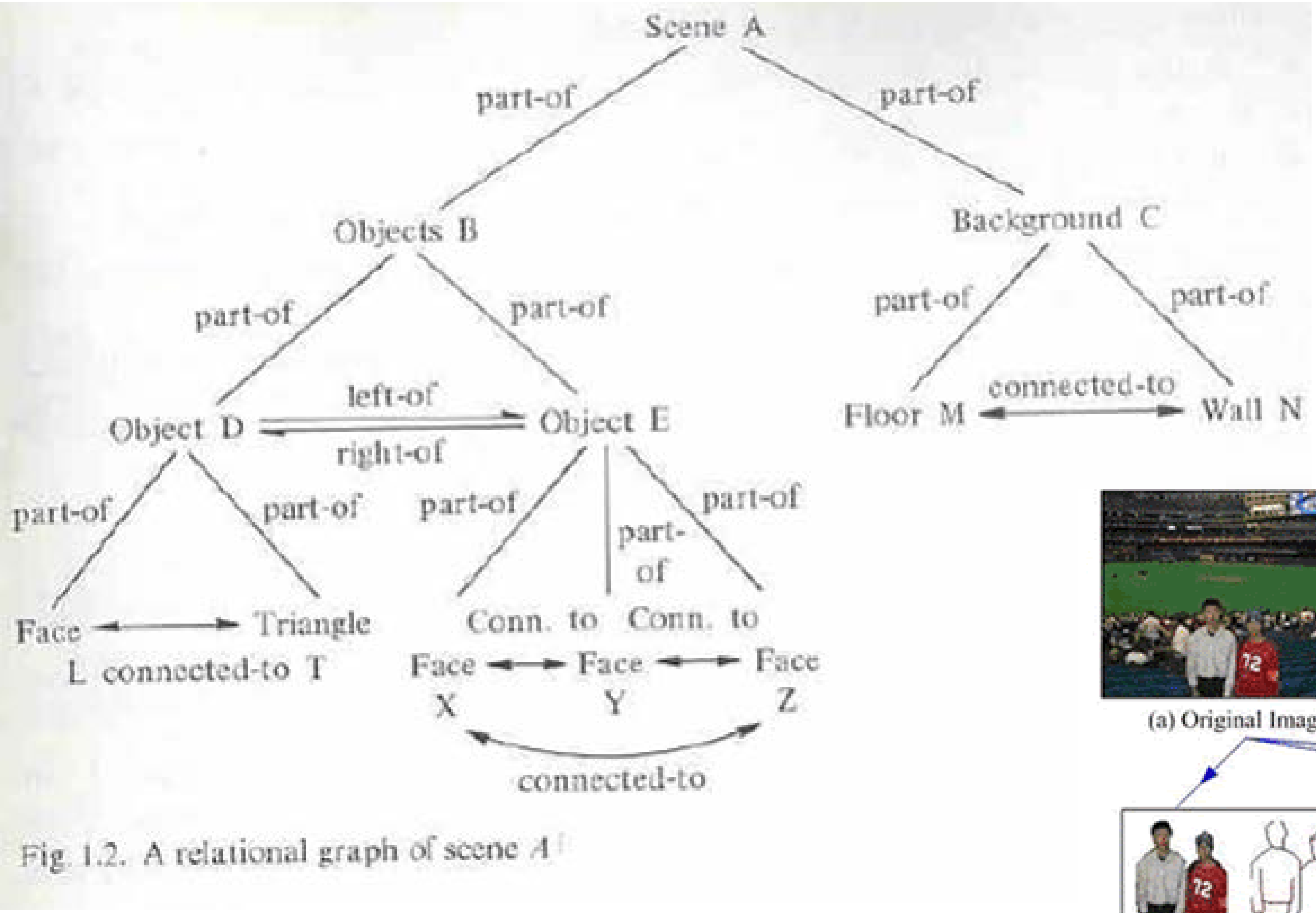
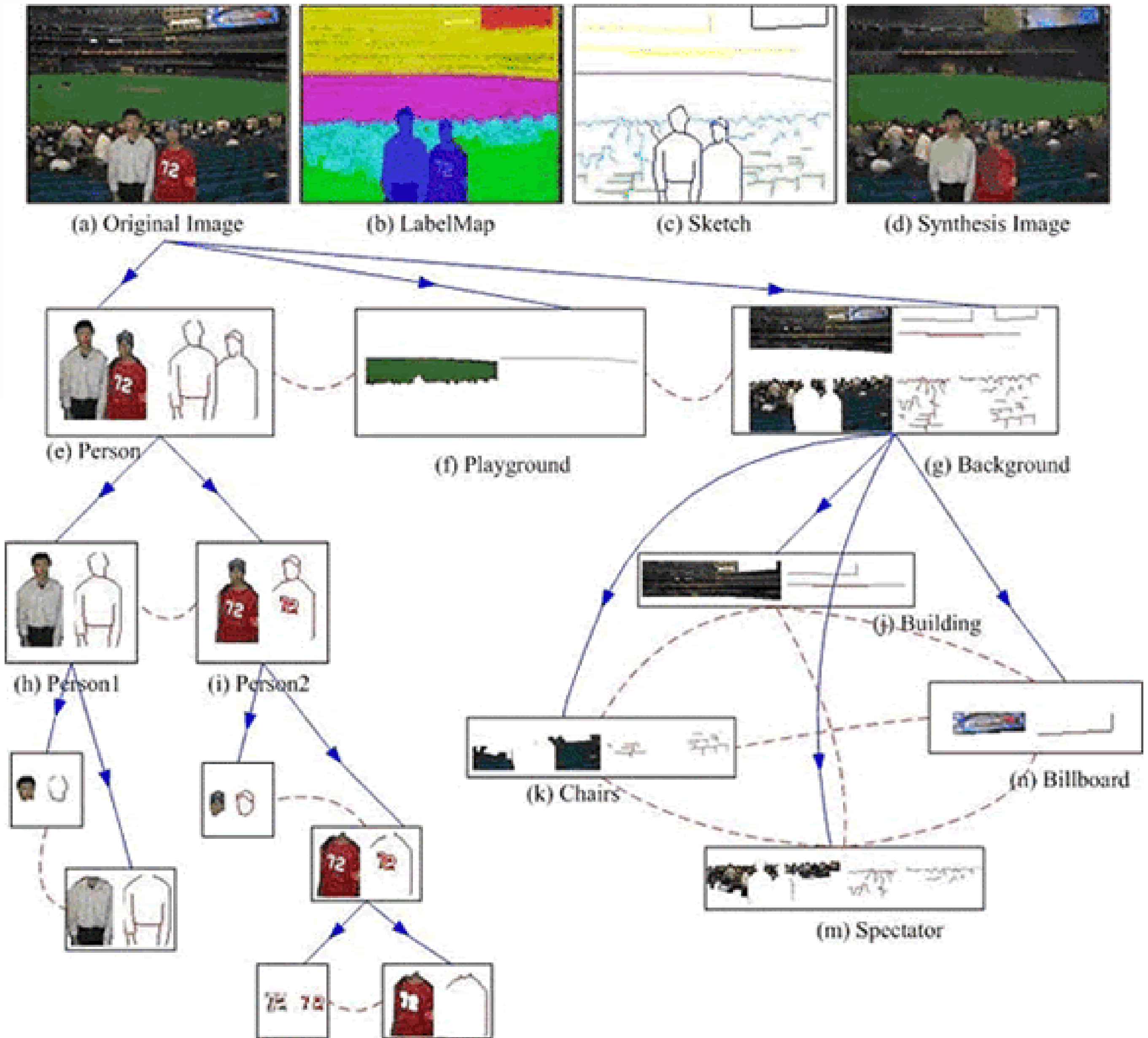


Fig. 1.1a and b. The pictorial pattern A and its hierarchical structural descriptions

In the 1970s, computers typically have 640KB memory, people are limited to line drawings.

Example of scene parsing by grammar from K.S. Fu





The ideas of image parsing was rejuvenated by Tu et al at ICCV 2003.

Stochastic Grammar – the basics

A grammar is often a quadruple *$G=<V_N, V_T, R, s>$*

- *a set of non-terminal nodes V_N,*
- *terminal nodes V_T,*
- *production rules R, and*
- *a initial node s.*

A sentence or string *w* is *valid* if it can be derived by the production rules in finite steps.

$$s \xrightarrow{R^*} w, \quad w \in V_T^*$$

The * sign means multiples: $V_T^* = \cup_{0 \le n < \infty} V_T^n, \quad V_T^n = \underbrace{V_T \times \cdots \times V_T}_{n}$

The set of all valid sentences or strings is called the *language* of the grammar G.

$$\boxed{L(G) = \{w : s \xrightarrow{R^*} w, \quad w \in V_T^*\}}$$

Stochastic Grammar – the basics

In formal language formulation, there are four types of grammars (for text) according to the generality of their production rules:

Type 3 Finite-State grammar (finite state automation).

Its rules are of the following form, and each time it only expands one terminal node.
One typical example is the Hidden Markov Model, the hidden state follows a *Markov chain*.

$$\boxed{A \to aB \quad or \quad A \to b \qquad\qquad a, b \in V_T,\ A, B \in V_N}$$

Type 2 Context-free grammar.

Its rules are of the following form, and each time it expands a non-terminal node independent of context.
This leads to the Markov tree models and is also called the branching process (in continuous form).

$$\boxed{A \to \beta \qquad\qquad \beta \in (V_T \cup V_N)^+, \quad A \in V_N}$$

Type 1 Context-sensitive grammar.

Its rules are of the following form, and each time it expands a non-terminal node with its context.

$$\boxed{\xi_L A \xi_R \to \xi_L \beta \xi_R \qquad \beta \in (V_T \cup V_N)^+, \quad A \in V_N}$$

Type 0 General grammar with no limitations on its production rules.

It is believed that natural languages belong to type 0.

Stochastic Grammar – the basics

A stochastic grammar is a grammar whose production rules are associated with probabilities.

A ::= bB | a with p_1 | p_2 sum to 1 for each A.

Each sentence in the language is associated with a probability.

$$L(G) = \{ (w, p(w)) : \ s \xrightarrow{R^*} w, w \in (V_N \cup V_T)^* \}$$

A sentence w may have multiple ways to parse, each is a series of production rules.
Let's denote by the set of possible parses by

$$\Omega(w) = \{ \ ps_i = (r_{i,1}, r_{i,2}, \ldots, r_{i,n(i)}) : s \xrightarrow{r_{i,1} \cdot r_{i,2} \cdots r_{i,n(i)}} w, \quad i = 1, 2, \ldots, m \ \}$$

The probability for the sentence is summed over all parse tree probabilities.

$$p(w) = \sum_{ps_i \in \Omega(w)} p(r_{i,1}) \cdot p(r_{i,2}) \cdots p(r_{i,n(i)})$$

Stochastic Grammar – the basics

A stochastic grammar *GR* is said to be *consistent* if its total probability sums to one.

$$\sum_{w \in L(G)} p(w) = 1$$

This is not trivial, because some probabilities may lose to infinity. One example is a SCFG

$$\begin{cases} A \to AA & p \\ A \to a & 1-p \end{cases} \qquad a \in V_T, A \in V_N$$

Then the total probability is $\sum_{w \in L(G)} p(w) = \min(1, \frac{1-p}{p})$

It is consistent iff p<1/2. I.e. it must terminate the non-terminals A at a speed faster than it creates new non-terminals.

Stochastic Grammar – the basics

Stochastic *attribute grammar*. Each node A (terminal or non-terminal) has a number of attributes x(A). In vision, e.g. a node could be a person, then its attributes are gender, age, race and so on.

E.g. for a production rule

$$A ::= bB \mid a \qquad \text{with } p_1 \mid p_2$$

We can use the attributes in two ways.

1, **Controlling the branching frequency by the attribute of A**

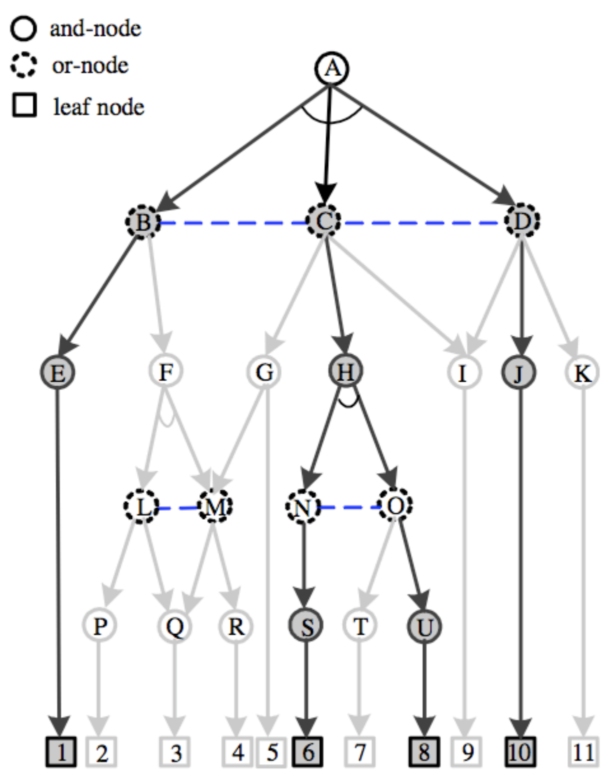
$$p_1 = p_1(x(A)), \quad p_2 = p_2(x(A)), \qquad p_1 + p_2 = 1 \quad \text{for any } x(A).$$

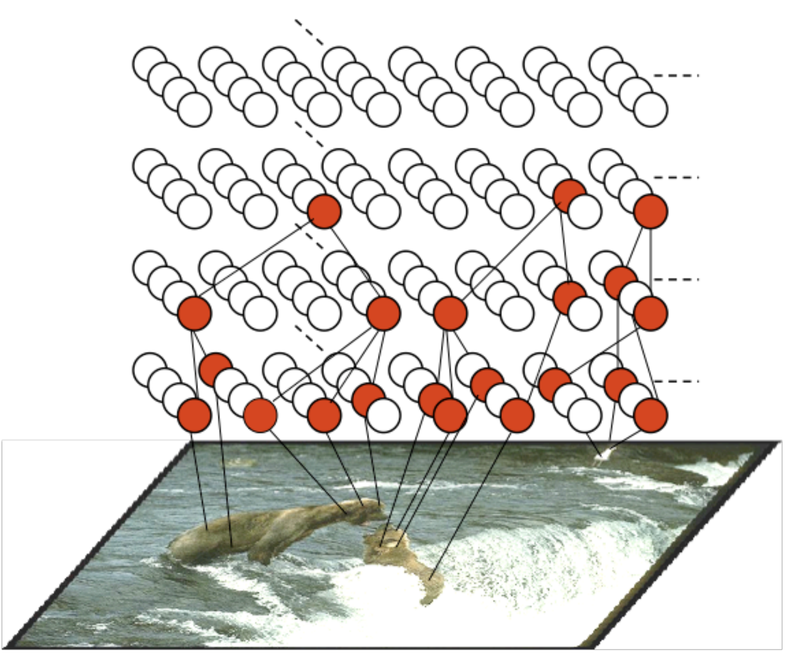
2, Imposing constraints between nodes or passing attributes between parent and child.

$f(x(A)) = g(x(b), x(B))$; $f()$, $g()$ are functions of the attributes.

$X(A) = X(a)$

A partial list of vision groups using various grammar





- Europe: Buhmann, Leonardis
- Brown: Geman, Mumford, Kimia, Felzenszwalb
- MIT: Kaelbling, Poggio, Savova, Tenenbaum
- Oregon State: Todorovic
- Purdue: Bouman, Pollak, Siskind
- SUNY Buffalo: Corso
- U Arizona: Barnard
- UCLA: Yuille, Zhu
- Weizmann: Sharon, Ullman
- U Maryland: Chellappa
- USC: Nevatia
- UCF: Shah
- Georgia Tech: Bobic

Grammar have been more frequentl
used in event understanding.

Now, I show a demo -- what grammar can do for you:

1, Spatial, Temporal, Causal inference for parsing object, scene and events.

2, Answering user queries about what, who, where, when, and why.

stc(xvid).avi

query_demo_04.avi

The demo is made by Dr. Mingtian Zhao at UCLA, Dr. Mun Wai Lee at IAI et al.