Lecture 2: Spatial And-Or graph

for representing the scene-object-part-primitive hierarchy

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

At CVPR, Providence, Rhode Island
June 16, 2012

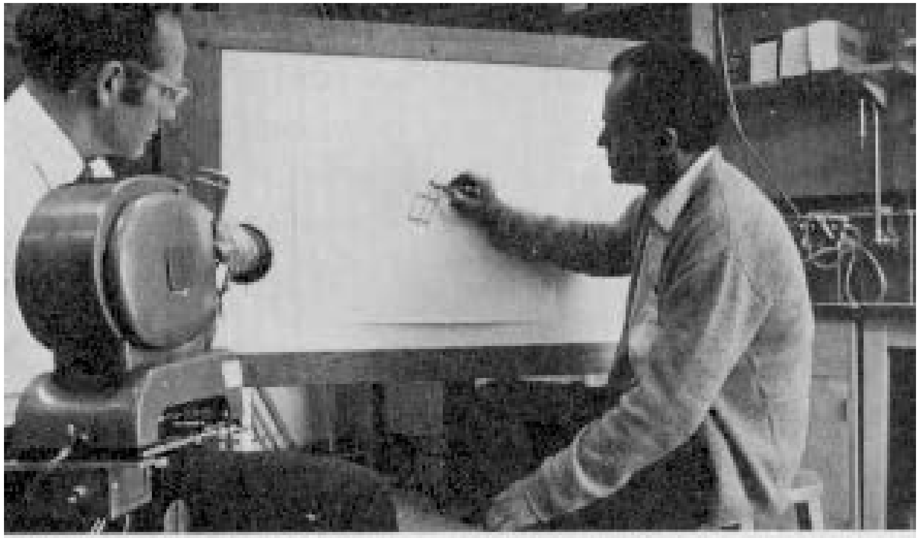
Evidences for hierarchical representations of image patterns are abundant in neuroscience and psychophysics.   But,

1, There are different theories of interpretations, e.g. each unit (neuron) can be viewed as a

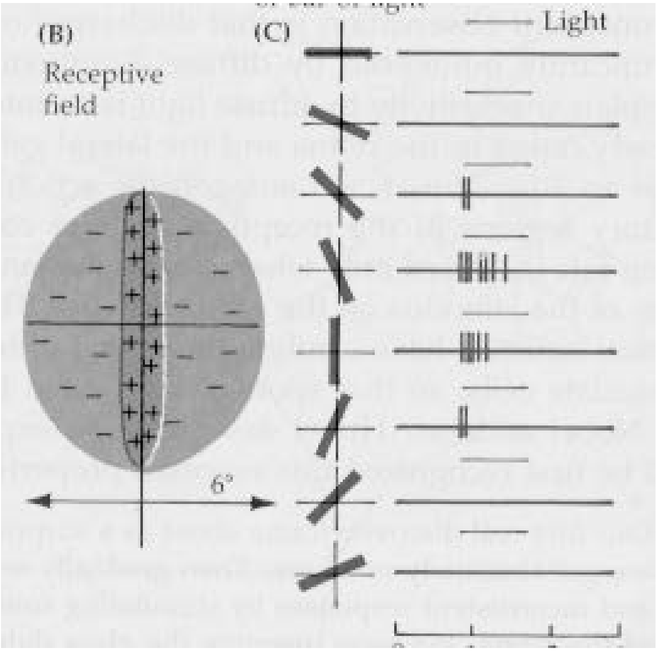
- filter --- MRF/CRF
- base function --- Wavelet, sparse coding
- classifier (decision makers) --- machine learning

2, Unsupervised learning of the hierarchy lags far behind in computer vision and machine learning.

Early evidence in Neuroscience: what cells in V1 see?



Huber and Weissel 1960s on cat experiments



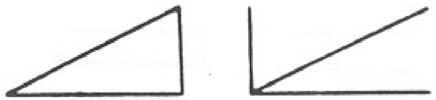
Single neuron recording
in the V1 area in cats and monkeys.

This leads to many explanations: edge detection, texton, and Gabor filters in the 1980s.
It also inspired wavelets in thinking in the 1980s-90s.
By the 2000s, people began to view it as classifiers.

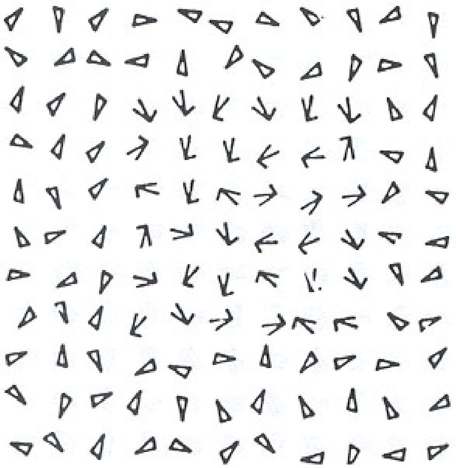
Psychophysics evidence for textons

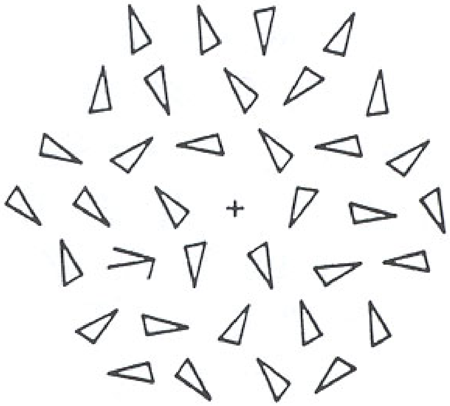
(1) textures vs textons (Julesz, 60-70s)

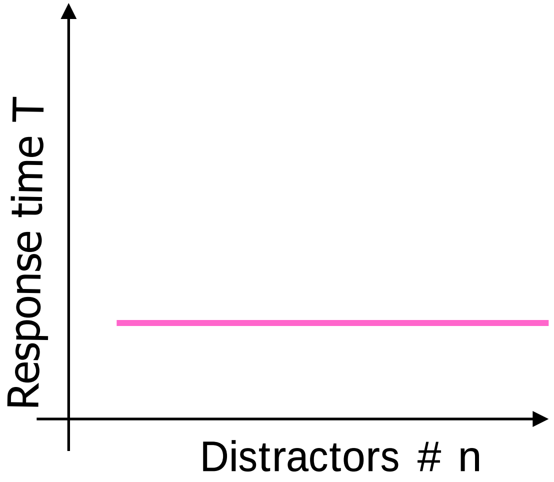
textons



(a)



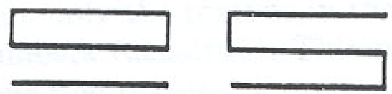




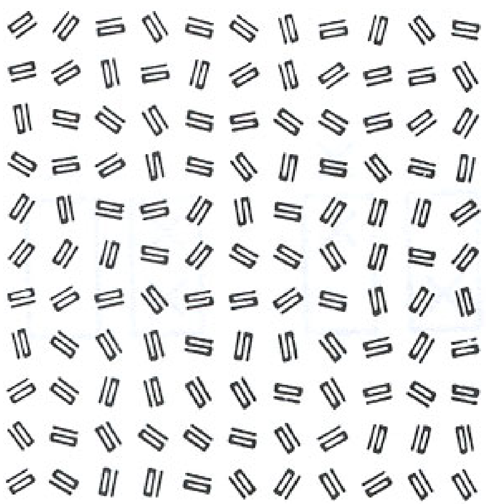
The subject is told to detect a target element in a number of background elements.
In this example, the detection time is independent of the number of background elements.

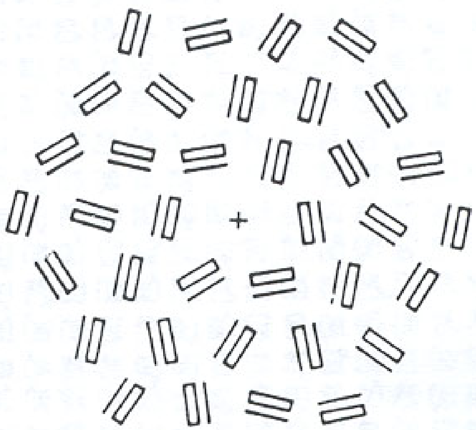
Psychophysics evidence for textons

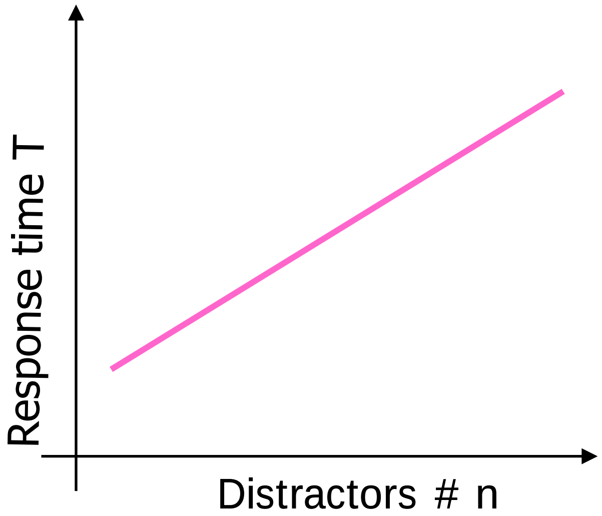
textures



(a)



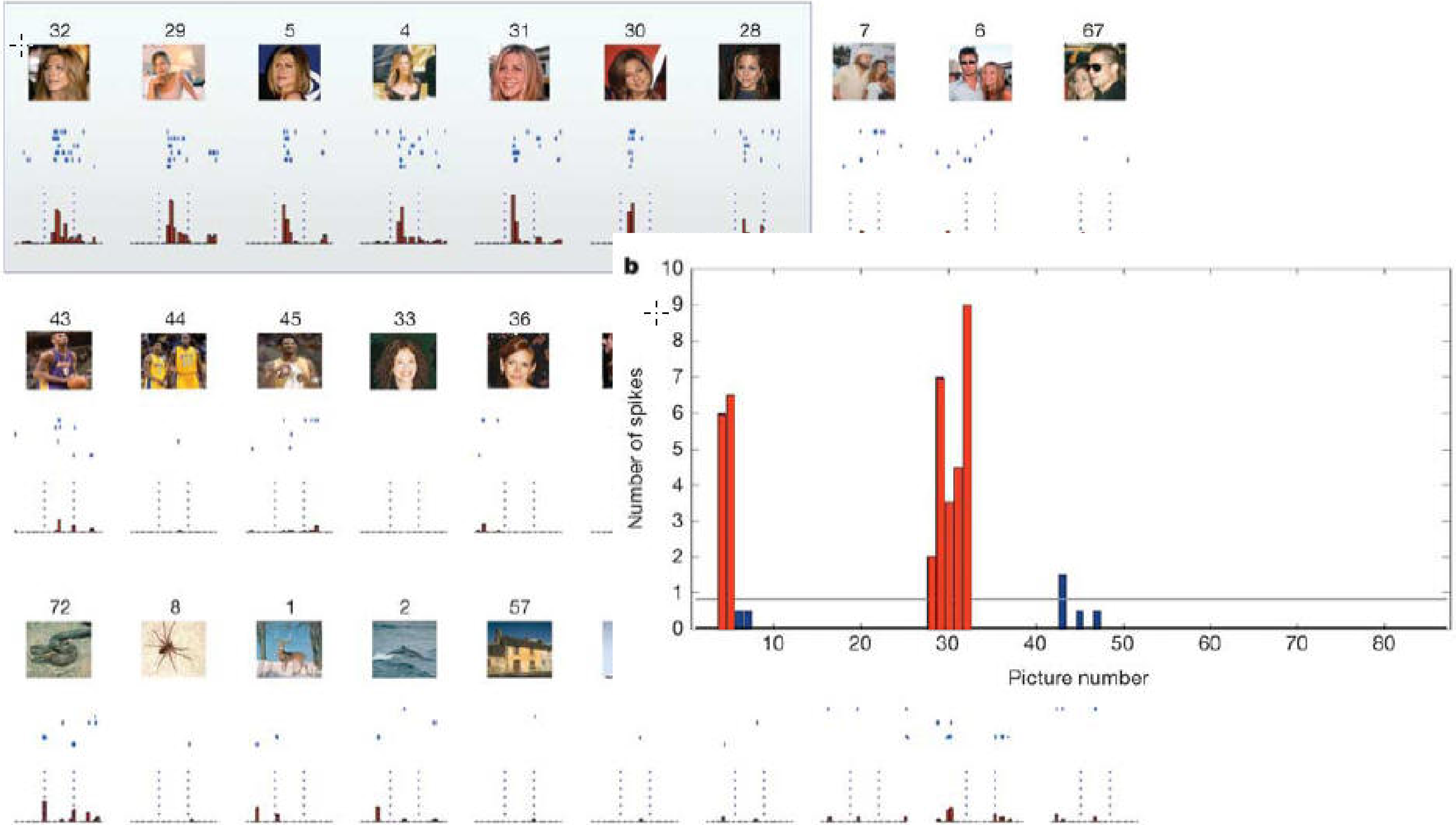




Neurons in the late stage of the visual pathway

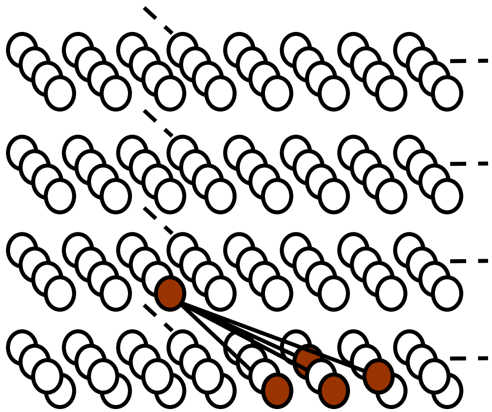
A Jennifer Aniston neuron was found

UCLA-Caltech labs [Fried, and Koch] record neurons in the human medial temporal lobe (MTL).

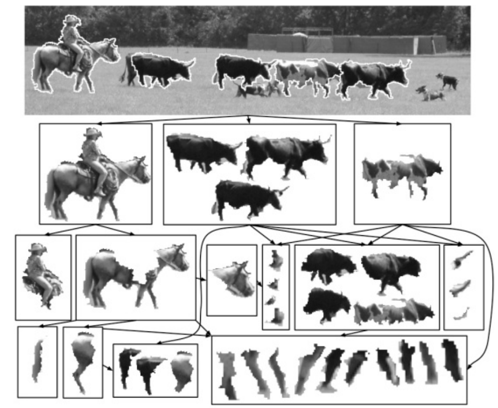


[1] R. Quian Quiroga et al. Invariant visual representation by single neurons in the human brain", Nature, 2005.
[2] M. Cerf et al "On-line, voluntary control of human temporal lobe neurons", Nature, 2010.

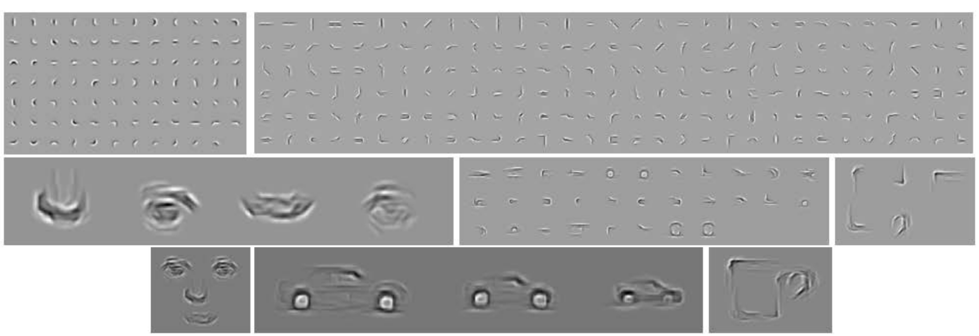
Proposed compositional hierarchies in computer vision



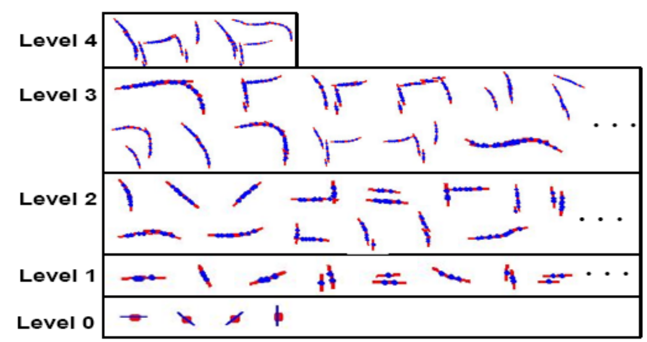
Y. Jin and S. Geman. Context and hierarchy in a probabilistic image model. In CVPR, 2006.



S. Todorovic and N. Ahuja. Unsupervised category modeling, recognition, and segmentation in images. TPAMI, 2008.

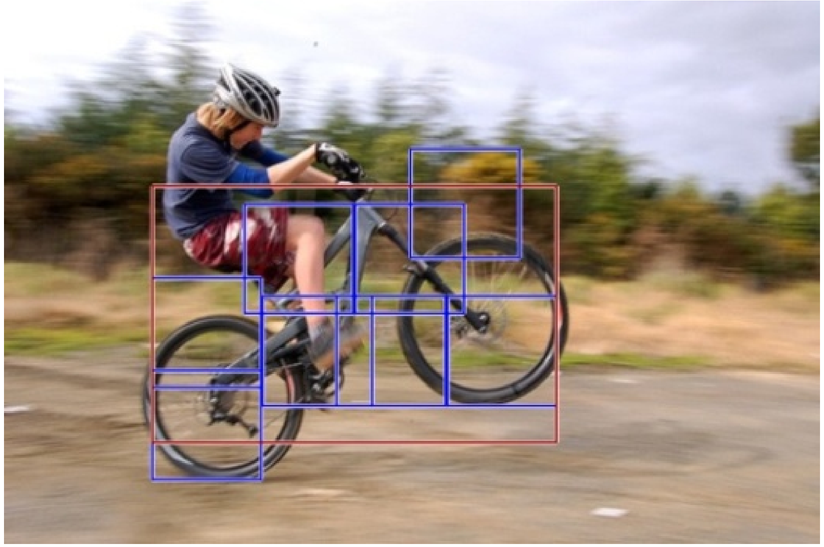


S. Fidler and A. Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In CVPR, 2007.



L. Zhu, Y. Chen, and A. Yuille. Learning a hierarchical deformable template for rapid deformable object parsing. TPAMI, 2009.

Recent well-engineered model: deformable part templates



Parts + HoG + s-SVM

P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. “Object detection with discriminatively trained part-based models,” TPAMI 2010.

1, These models are mostly And-structures in a hierarchy, without mixing with Or-structures.
2, These work do not address learning reconfigurable structures in an unsupervised way.
3, What are the principles for unsupervised learning?

The current literature on hierarchical/compositional models has been very confusing to new students.

Outline of this lecture:

1, Three ways to represent visual concepts

- Ensembles by statistical physics and MRF models;
- Low dimensional subspace or manifold in Sparse coding;
- Language by grammar

2, And-Or graph as a unifying representation

- And-Or graph
- Parse graph
- Configurations
- Probabilistic models

3, Case studies for spatial-AOG

- Object grammars: human parsing
- Scene grammars: 3D scene parsing

1, Visual Concepts – the units of visual knowledge

In Mathematics and logic, concepts are equal to deterministic sets, e.g. Cantor, Boole, or spaces in continuous domain, and their compositions through the "and", "or", and "negation" operators.

$$(A \cap \bar{B} \cap D) \cup (B \cap \bar{C})$$

Visual concepts:

e.g. noun concepts: human face, vehicle, chair?
verbal concept: opening door, making coffee?

The world to us is fundamentally stochastic.
We have three ways to define stochastic sets for visual concepts.

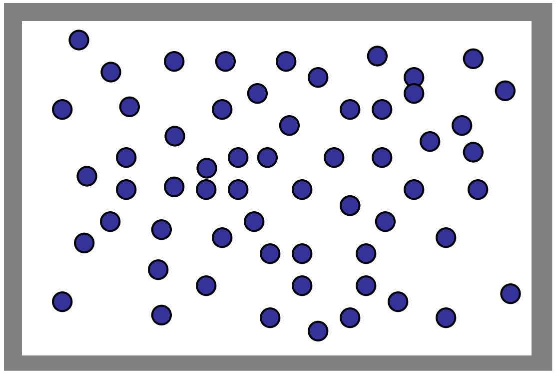
Ref. [1] D. Mumford. *The Dawning of the Age of Stochasticity*. 2000.
[2] E. Jaynes. *Probability Theory: the Logic of Science*. Cambridge University Press, 2003.

Method 1, Stochastic set in statistical physics

Statistical physics studies macroscopic properties of systems that consist of massive elements with microscopic interactions.

e.g.: a tank of insulated gas or ferro-magnetic material

$N = 10^{23}$



Micro-canonical Ensemble

A state of the system is specified by the position of the N elements X^N and their momenta p^N

$$S = (x^N, \ p^N)$$

But we only care about some global properties

Energy E, Volume V, Pressure, ….

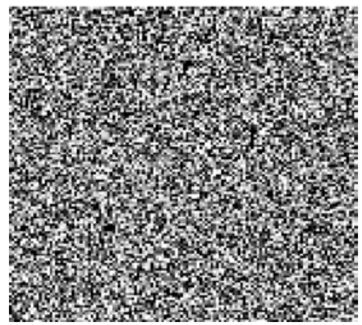
$$\text{Micro-canonical Ensemble} = \Omega(N, E, V) = \{ s : \ h(S) = (N, E, V) \}$$

It took 30-years to make this theory work in vision

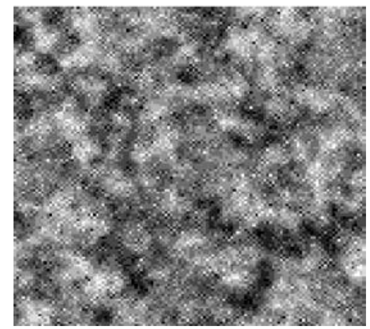
$$\text{a texture} = \Omega(h_c) = \{\, I : \; h_i(I) = h_{c,i} \;\; , i = 1,2,\ldots,K \,\}$$

h_c are histograms of Gabor filters

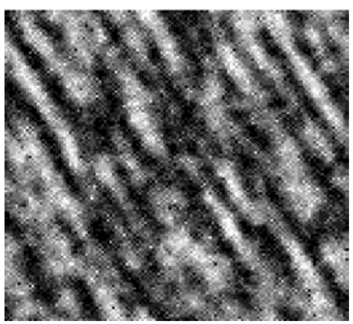
I^{obs}



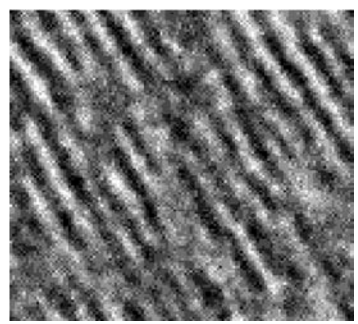
$I^{syn} \sim \Omega(h)$ k=0



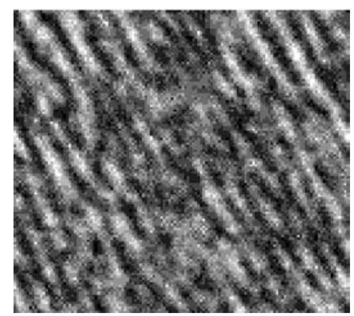
$I^{syn} \sim \Omega(h)$ k=1



$I^{syn} \sim \Omega(h)$ k=3



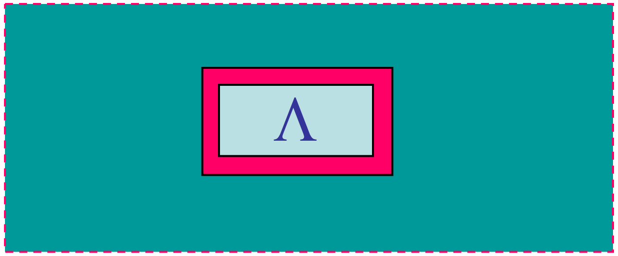
$I^{syn} \sim \Omega(h)$ k=4



$I^{syn} \sim \Omega(h)$ k=7

Ref: Zhu, Wu, Mumford, "FRAME: Filters, Random field And Maximum Entropy: --- Towards a Unified Theory for Texture Modeling," IJCV, 1997.

Equivalence of deterministic set and probabilistic models



Gibbs 1902,
Wu and Zhu, 2000

Theorem 1

For a very large image from the texture ensemble $\mathrm{I} \sim f(\mathrm{I}; \mathrm{h}_c)$ any local patch of the image I_Λ given its neighborhood follows a conditional distribution specified by a FRAME/MRF model $p(\mathrm{I}_\Lambda \mid \mathrm{I}_{\partial\Lambda} : \beta)$

Theorem 2

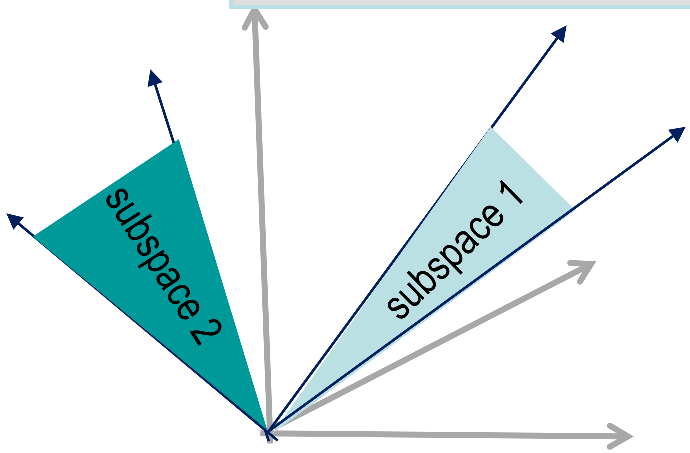
As the image lattice goes to infinity, $f(\mathrm{I}; \mathrm{h}_c)$ is the limit of the FRAME model $p(\mathrm{I}_\Lambda \mid \mathrm{I}_{\partial\Lambda} : \beta)$, in the absence of phase transition.

$$\boxed{p(\mathrm{I}_\Lambda \mid \mathrm{I}_{\partial\Lambda} ; \beta) = \frac{1}{Z(\beta)} \exp\{-\sum_{\mathrm{j}=1}^{\mathrm{k}} \beta_{\mathrm{j}} h_{\mathrm{j}}(\mathrm{I}_\Lambda \mid \mathrm{I}_{\partial\Lambda})\}}$$

Ref. Y. N. Wu, S. C. Zhu, "Equivalence of Julesz Ensemble and FRAME models," *Int'l J. Computer Vision*, 38(3), 247-265, 2000.

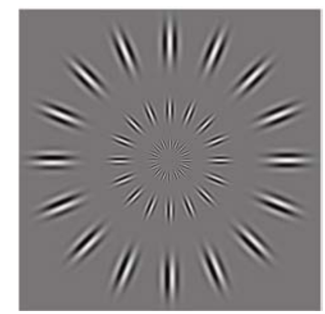
Method 2, Lower dimensional sets or subspaces

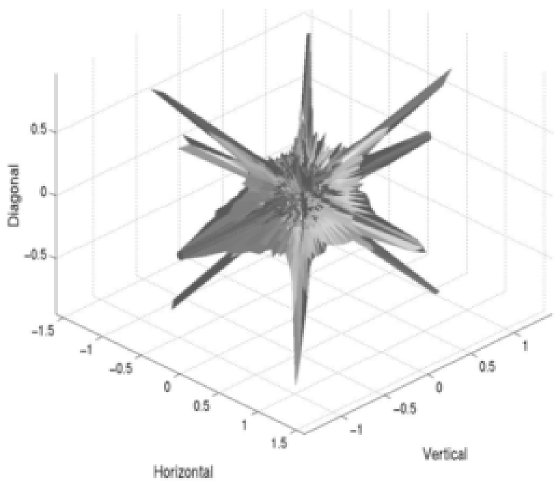
$$\text{a texton} = \Omega(\mathrm{h_c}) = \{\, \mathrm{I}: \ \ \mathrm{I} = \sum_{\mathrm{i}} \alpha_{\mathrm{i}} \psi_i, \ \ \| \alpha \|_0 < \mathrm{k} \,\}$$



K is far smaller than the dimension of the image space.

φ is a basis function from a dictionary.





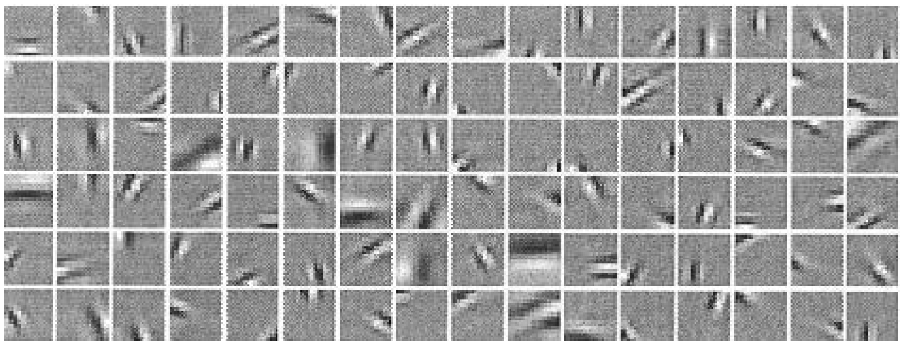
Here is an example of how real world data can be truly complex – non-Gaussian and highly kurtotic. This is an iso-density contour for a 3D histogram of log(range) images (2x2 patches minus their means) (Brown range image database, thesis of James Huang)

Sparsity and harmonic analysis

Stochastic sets from sparse coding

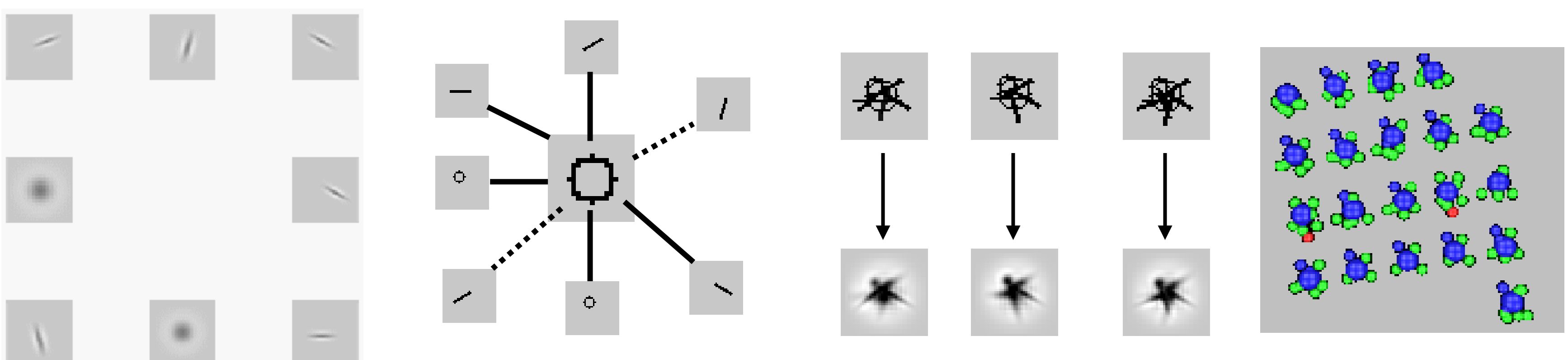
Learning an over-complete image basis from natural images

$$I = \Sigma_i \alpha_i \psi_i + n$$



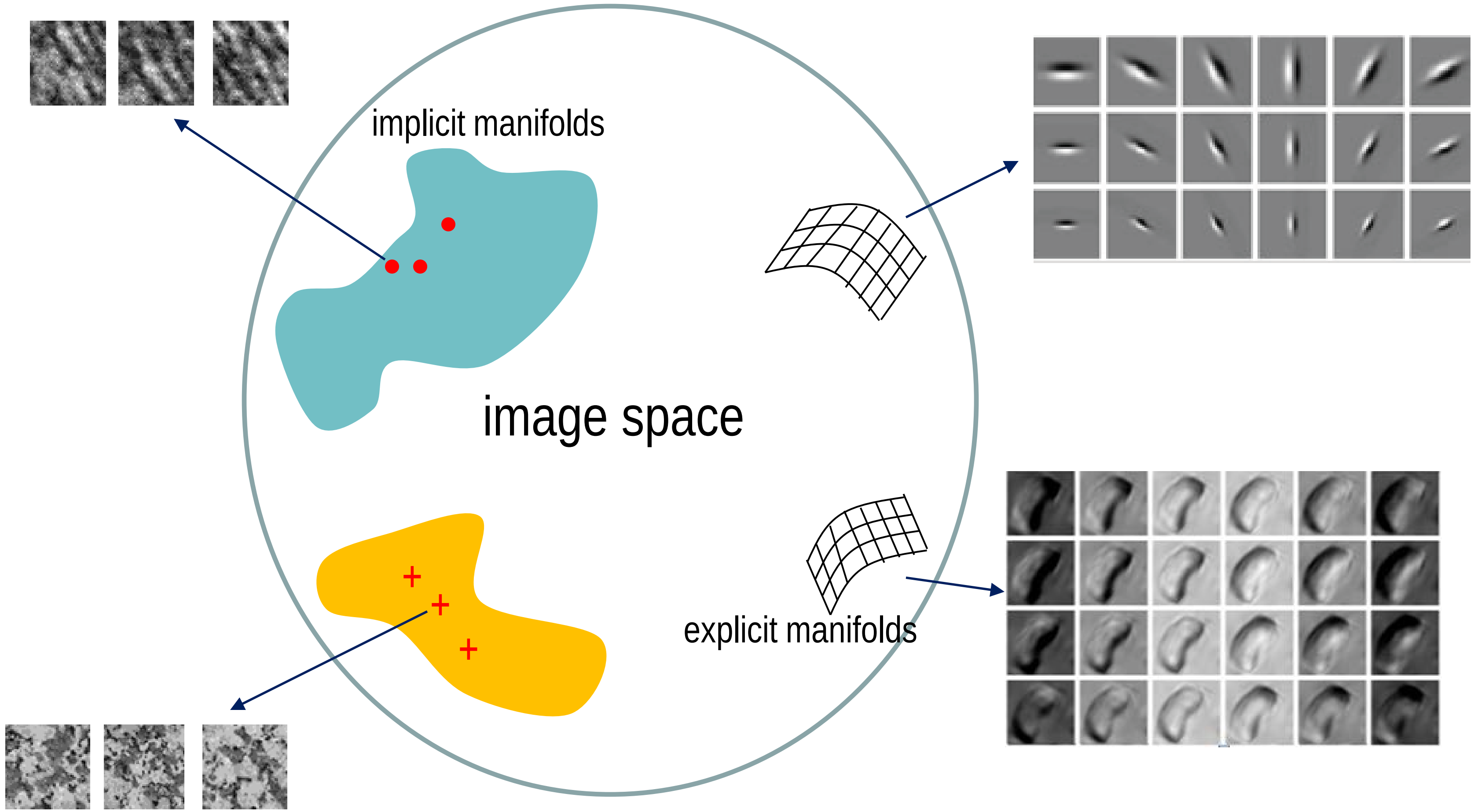
(Olshausen and Fields, 1995-97)

Textons



B. Olshausen and D. Fields, "Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1?" *Vision Research, 37*: 3311-25, 1997.
S.C. Zhu, C. E. Guo, Y.Z. Wang, and Z.J. Xu, **"What are Textons?"** *Int'l J. of Computer Vision,* vol.62(1/2), 121-143, 2005.

Look at the space of image patches



Two regimes of stochastic sets

Sets defined by *implicit* vs. *explicit* functions

$$\Omega = \{\mathrm{I}:\ \ h(\mathrm{I}) = h_o\ \}$$

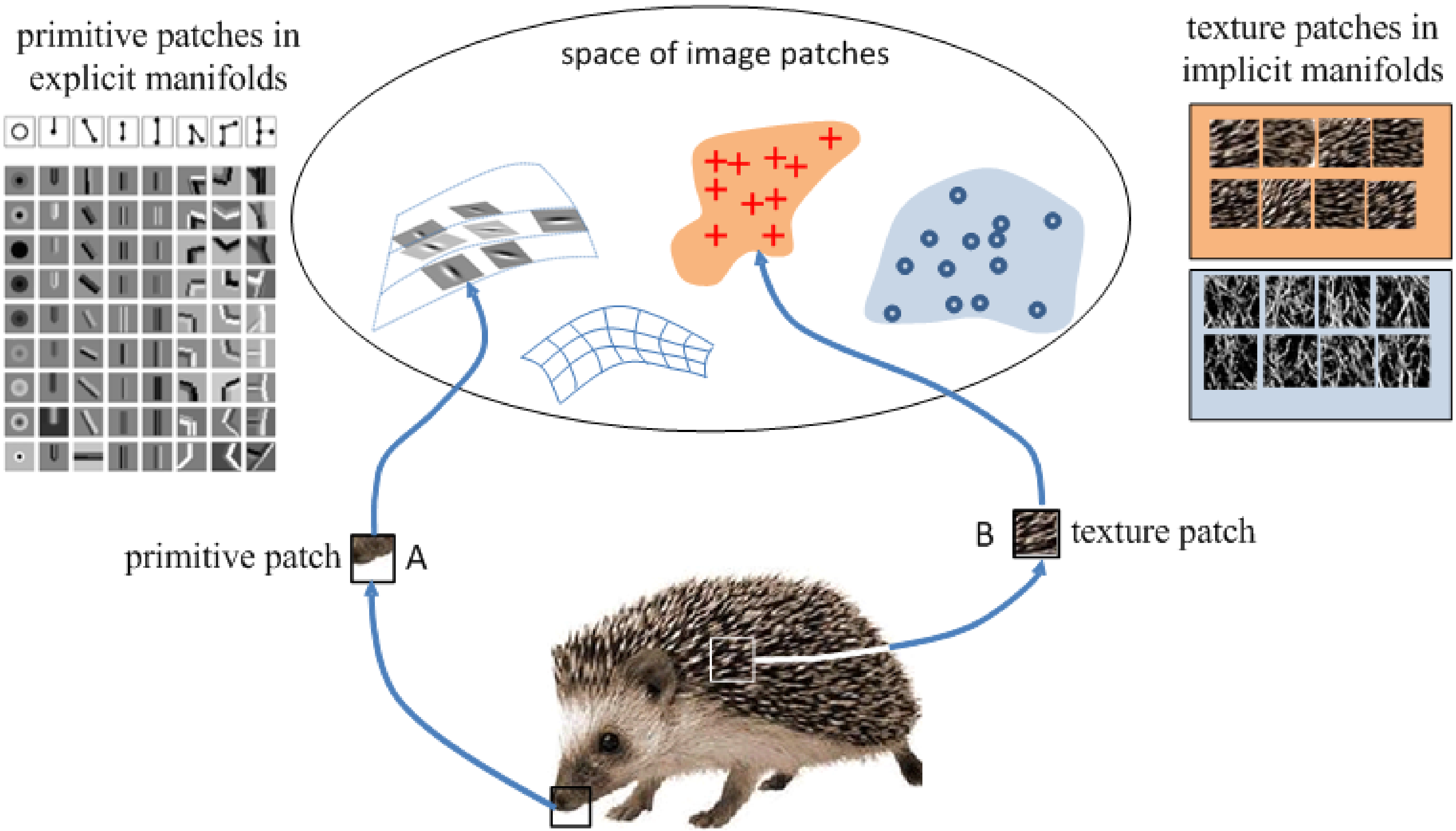
h(I) is some image feature/statistics

$$\Omega = \{\ \mathrm{I}:\ \mathrm{I} = g(w; \Delta)\ \}$$

g is a generation function,
w is intrinsic dimension
Δ is a dictionary

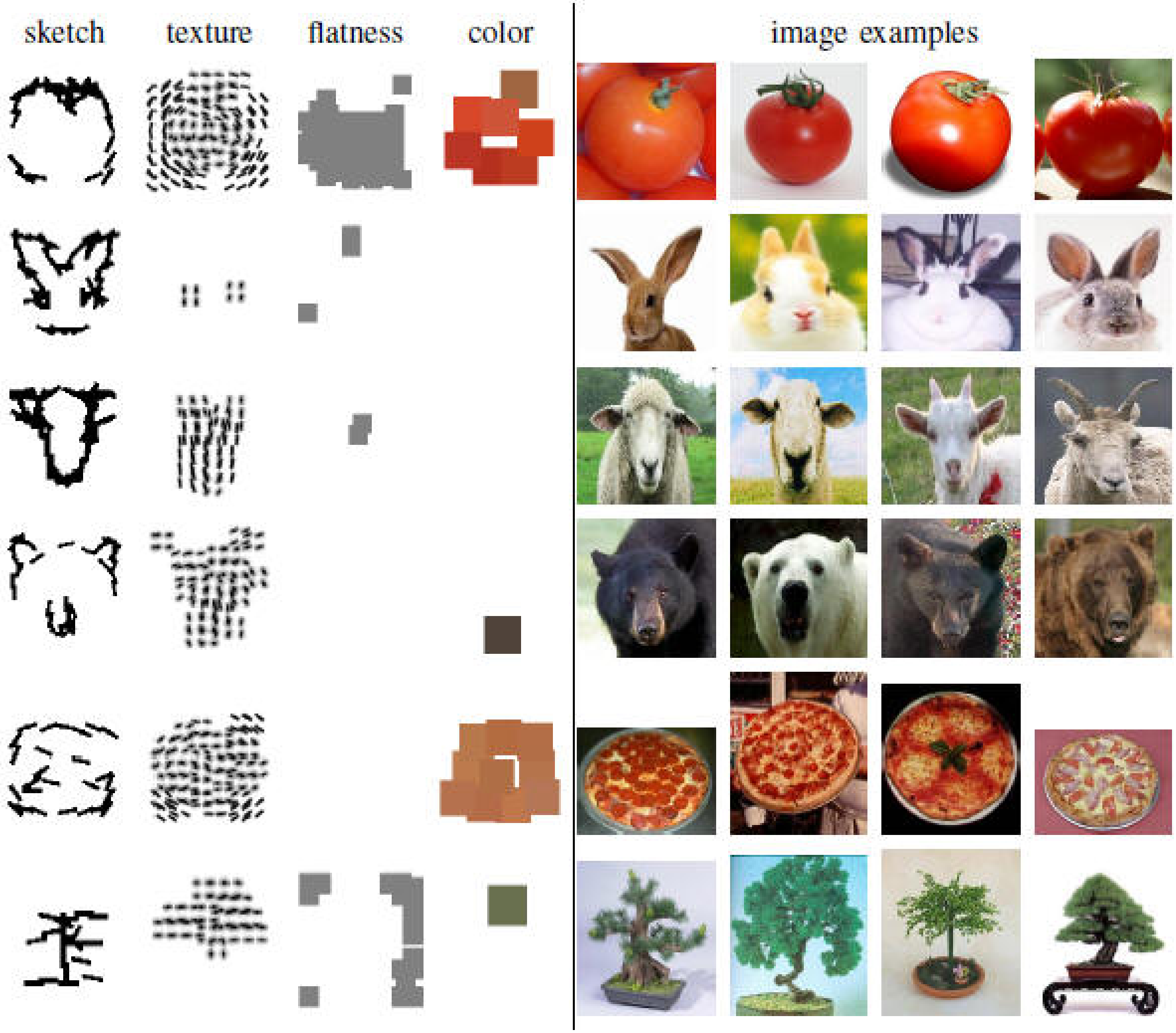
Ref. Zhu et al. “learning implicit and explicit manifold by information projection, PRL, 2009.

Flat object template: mixing the textures and textons



Z. Si and S.C. Zhu Hybrid Image Templates, CVPR09, TPAMI 2012.

Unsupervised Learning of Hybrid Object Templates

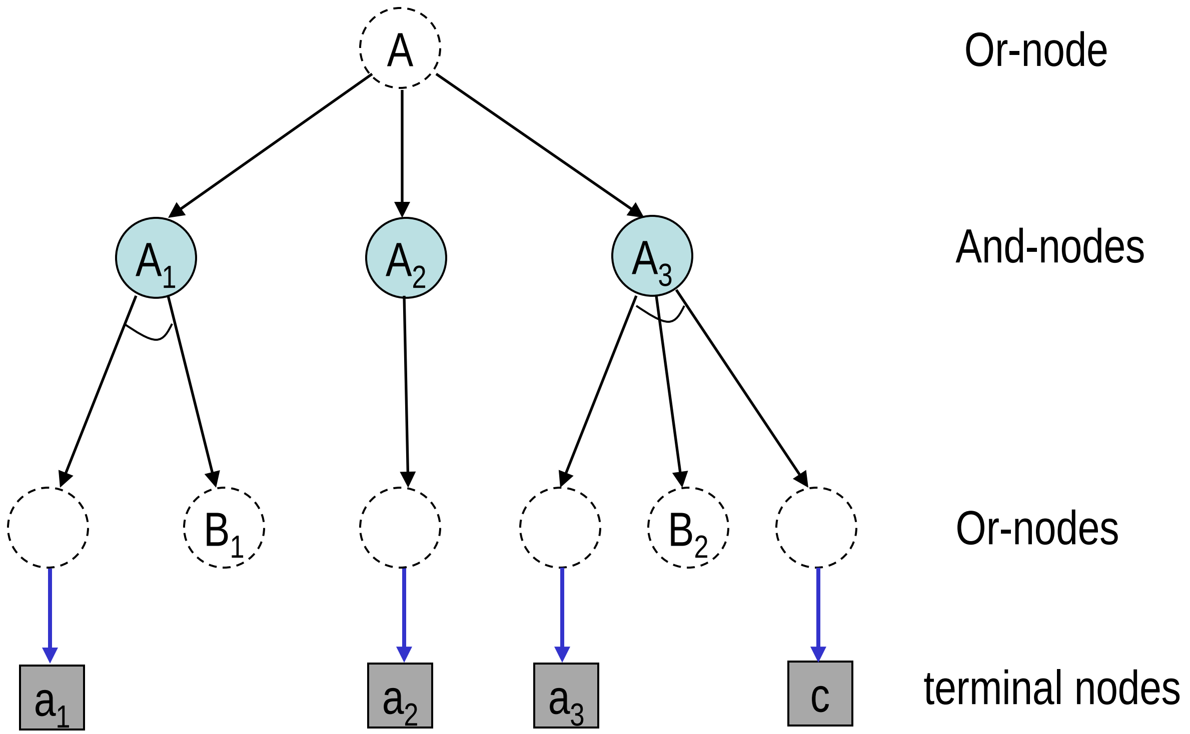


Z. Si and S.C. Zhu, CVPR09, TPAMI 12

Method 3, Stochastic sets by Grammar

A ::= aB | a | aBc

A production rule
can be represented by
an And-Or tree



The language is ***the set of all valid configurations*** derived from a note A.

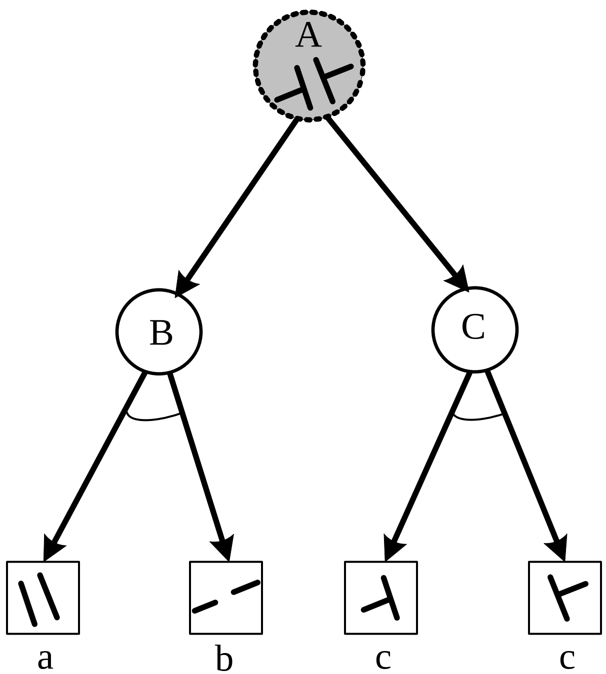
$$L(A) = \{ (\omega, p(\omega)) : \ A \xrightarrow{R^*} \omega \}$$

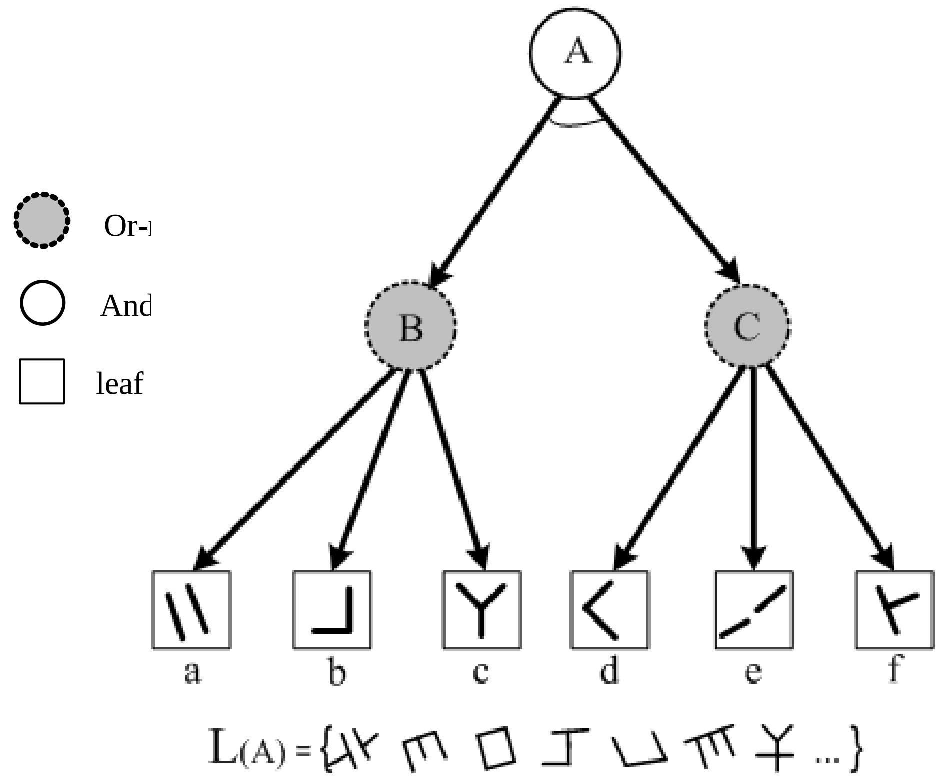
Athe elements in this set has varying configurations and dimensions.

2, And-Or graph as a unifying representation

A grammar production rule can be converted in an And-or tree fragment:

$$A \to ab \,|\, cc$$



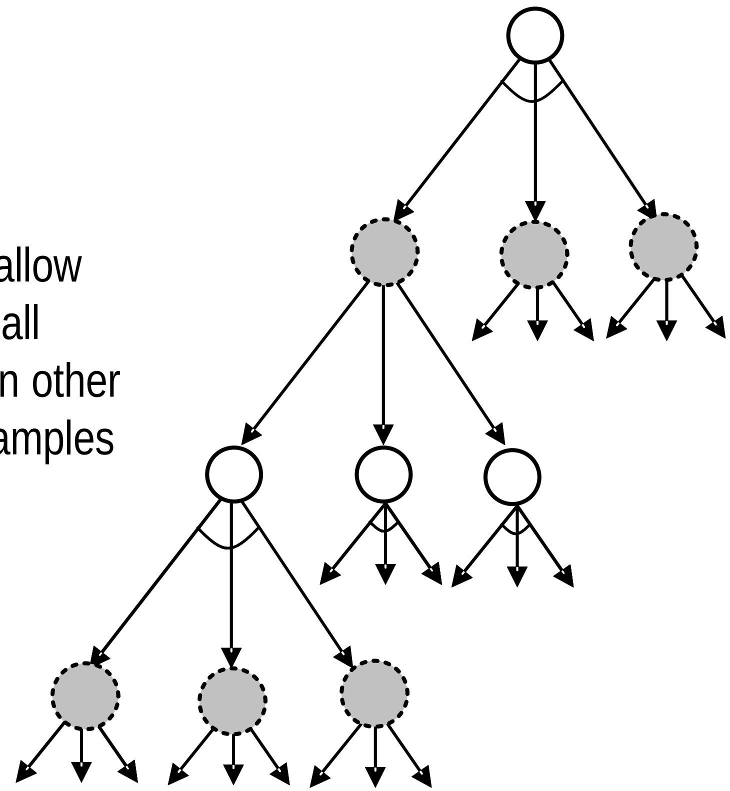


The language of a node A is the set of all valid configurations

The expressive power of and-or trees

Consider And-Or tree with
branching factor = 3 and depth =2.

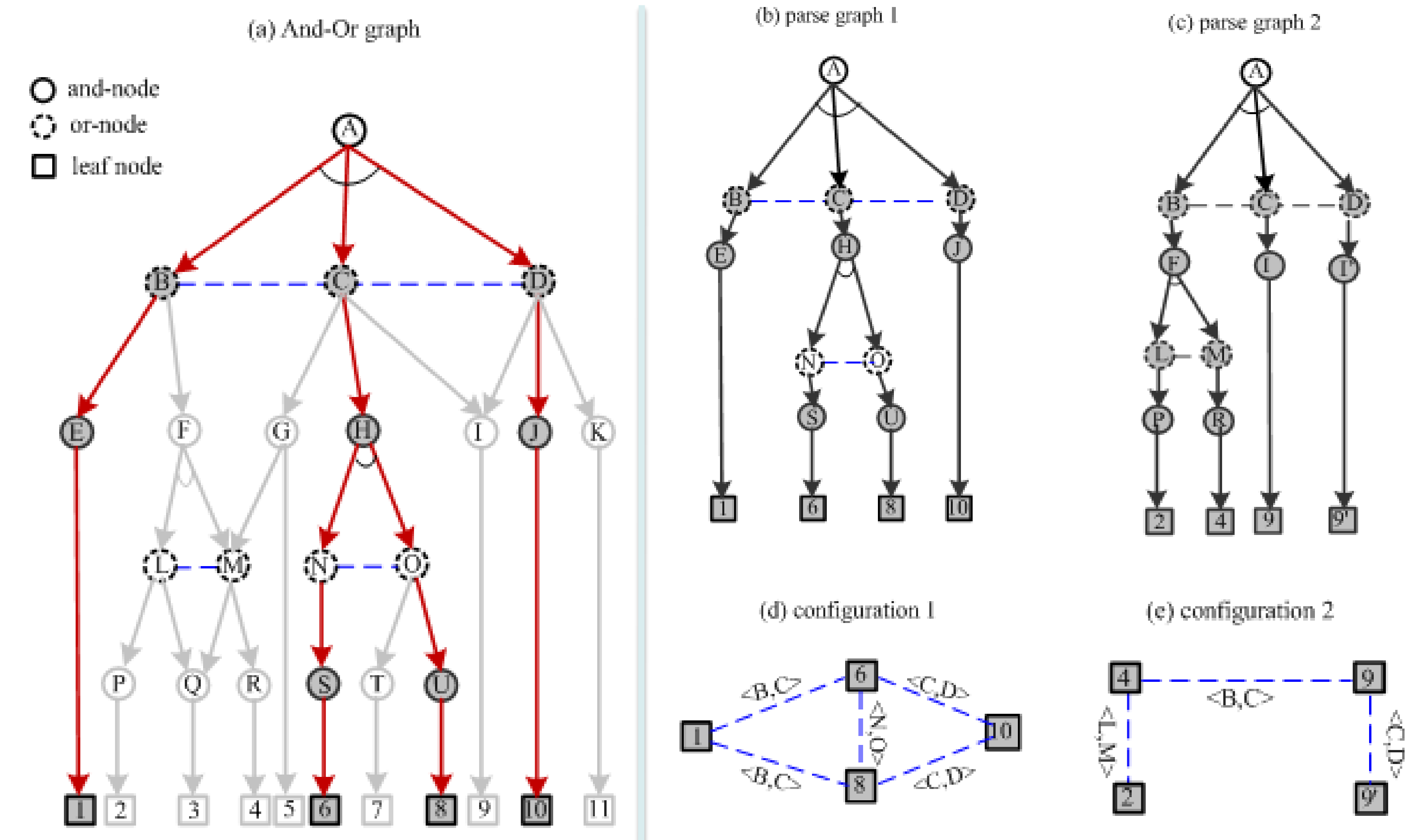
Object grammars are short and does not allow infinite recursion. Therefore the space of all Object grammars has smaller capacity than other Logic formulas and needs less training examples to learn (See lecture 10 for PAC learning).



Total: 1+3+9+27 = 30 nodes with 81 leaves

$(3\text{x}3^3)^3 = 3^{12}$ = 531,441configurations

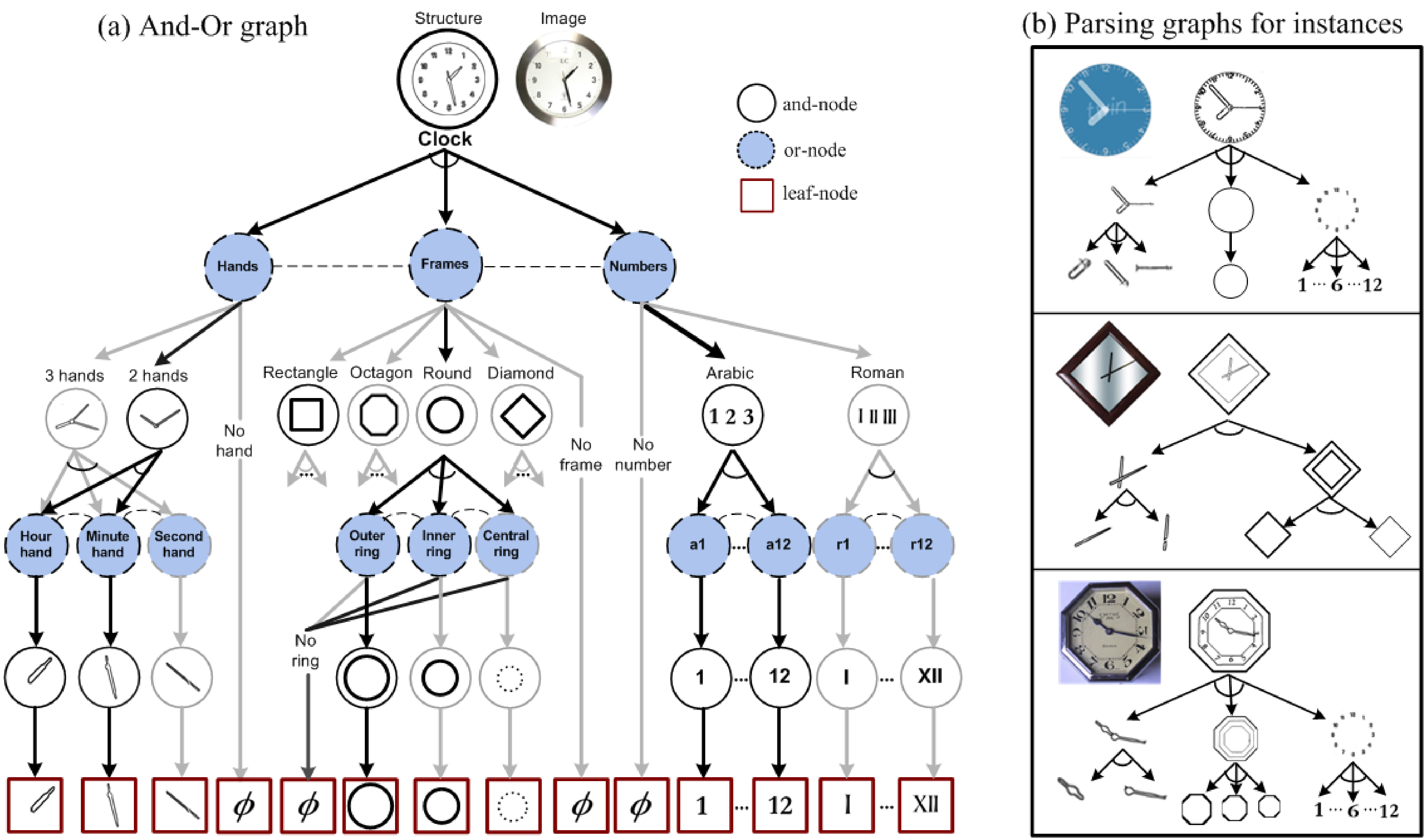
And-Or graph, parse graphs, and configurations



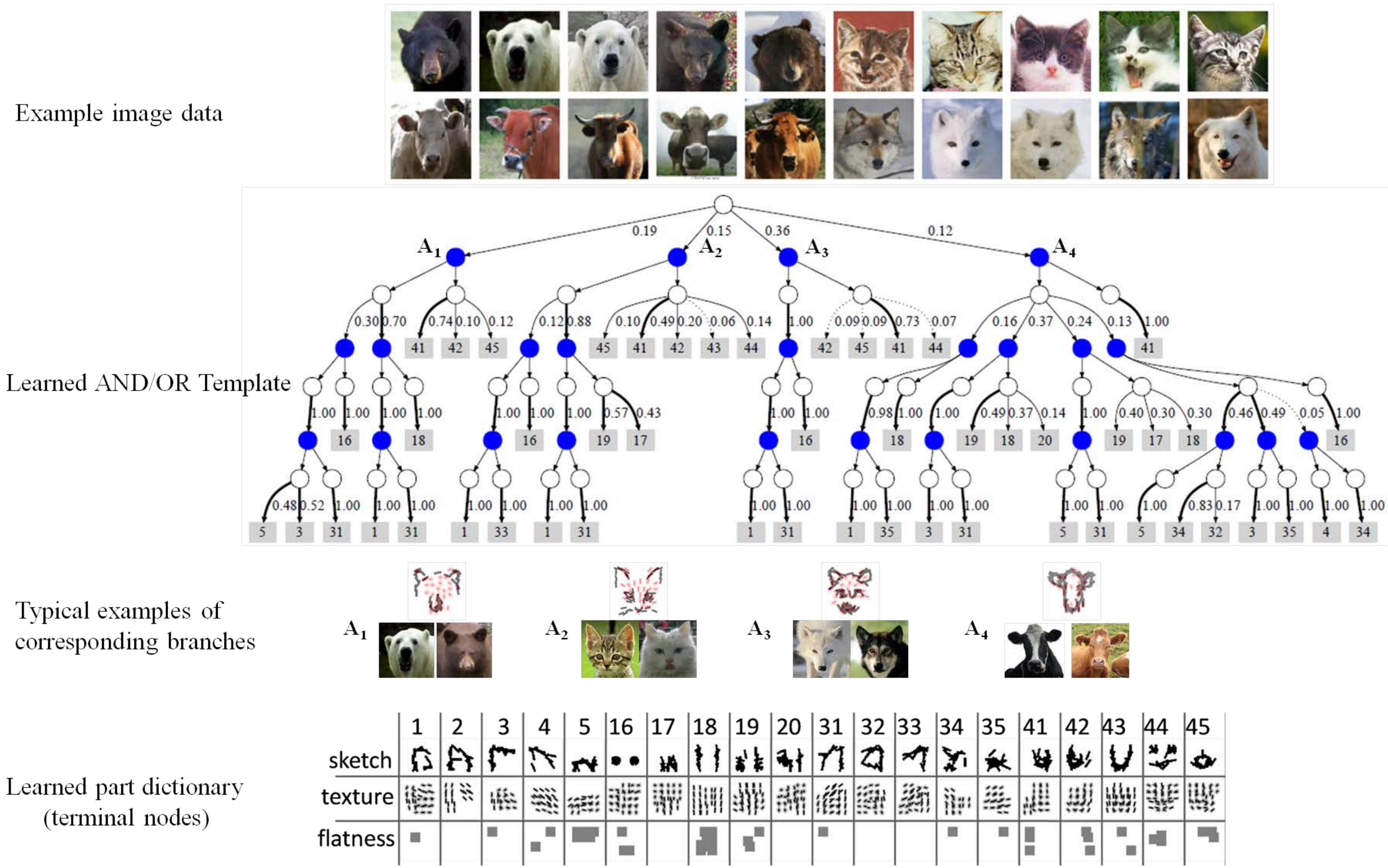
Each category is conceptualized to a grammar whose language defines a set or "equivalence class" for all the valid configurations of the each category.

An example: the clock category

To design a vision algorithm that reads clocks, we need grammar. This cannot be achieved by The flat HoG + SVM paradigm.

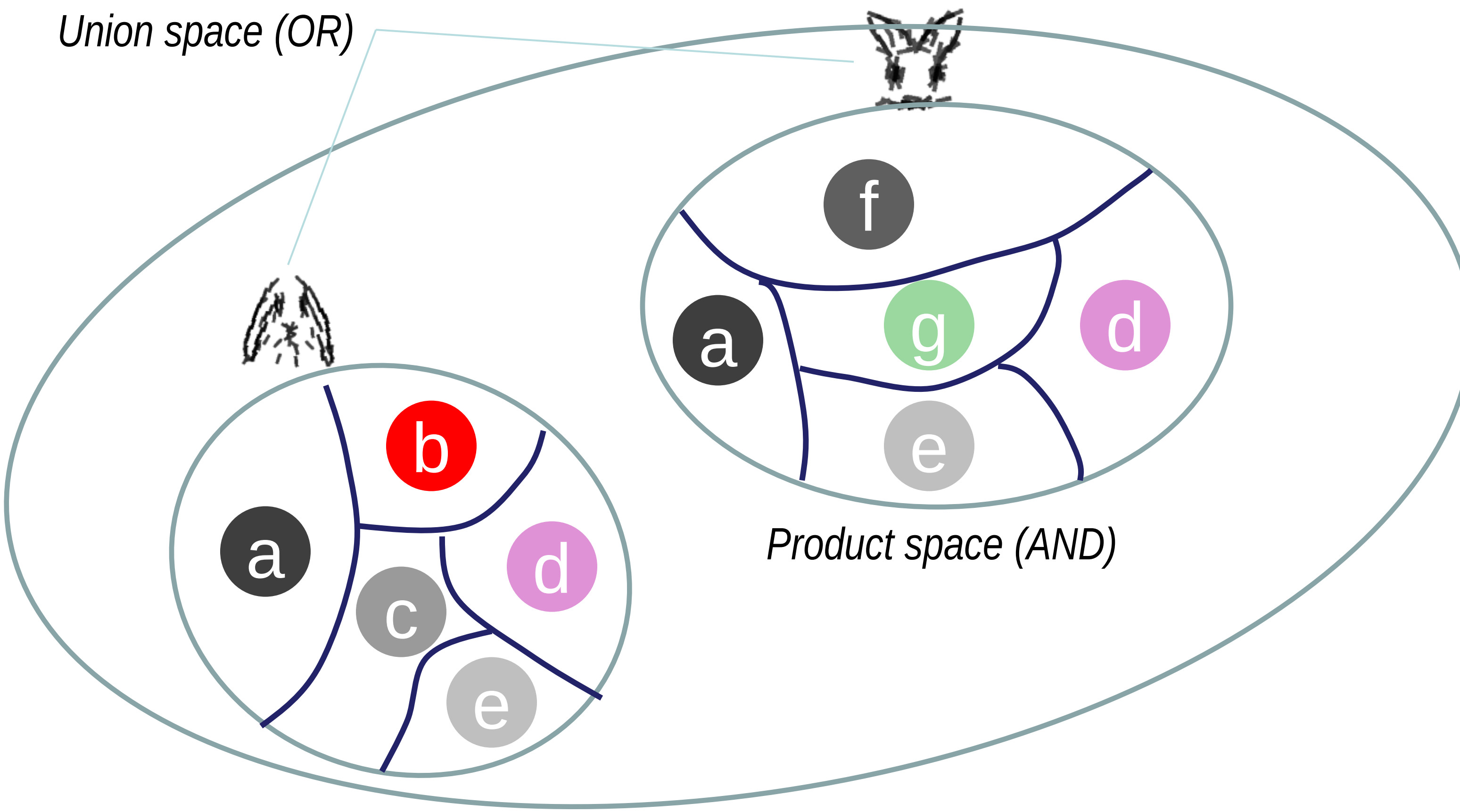


And-Or templates grounded on textures and textons



This is learned unsupervisedly by an information projection principle. See lecture 10.

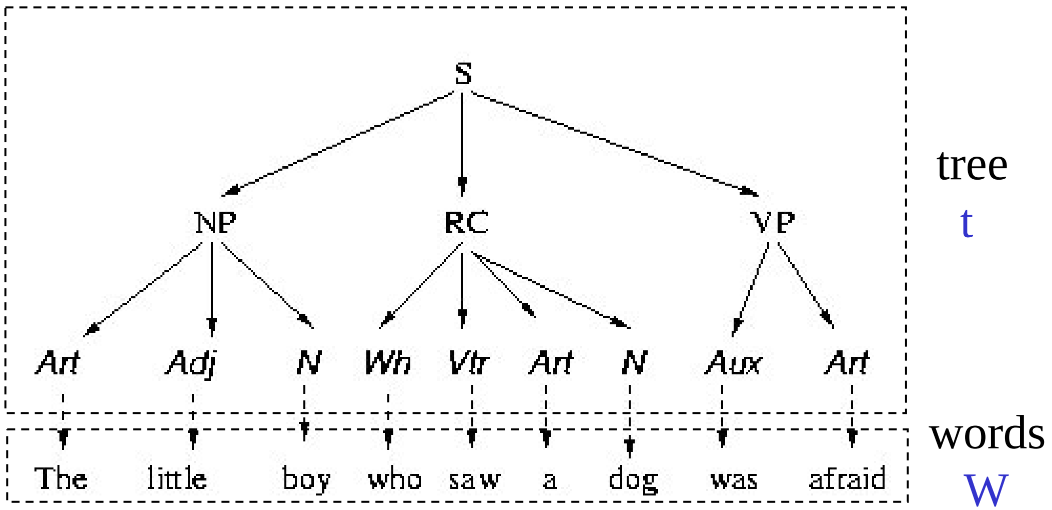
How does a set for an object looks like in the image space?



Defining a probability on the AoG

Take language as a simple example. A sentence $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$

The context free probability of w and a parse tree t is



Deriving the probability model for a context sensitive grammar

The probability for a *parse tree* $t = (r_1, r_2, ..., r_K)$

$$p(t) = p(r_1)p(r_2)\cdots p(r_K) = \frac{1}{Z}\exp\{-\sum_{r_i \in t} \lambda(\omega(r_i))\}$$

the probability for a bi-gram *Markov chain (or field)* of words W

$$p(W) = p(w_0)\prod_{i=1}^{n} p(w_i|w_{i-1}) = \frac{1}{Z}\exp\{-\sum_i \phi(w_i) - \sum_{<i,j>} \psi(w_i, w_j))\}$$

which are derived from maximum entropy with statistical constraints

$$E[\phi(w_i)] = \mu_i, \quad \forall i.$$
$$E[\psi(w_i, w_j)] = \mu_{ij}, \quad \forall < i, j >$$

We pursue a model that should observe the statistical constraints above (for contexts) while minimizing a KL-divergence to the tree model (for hierarchic structures).

$$p(g; \mathcal{G})^* = \arg\min KL(p(g; \mathcal{G}) \,||\, p(t))$$

Formulation : the Probability model

By solving this constrained optimization problem, one obtains a joint probability on the parsing graph G.

$$p(\mathbf{pt}(\omega)) = \frac{1}{Z} h^*(\omega_1) \prod_{i=1}^{n-1} h^*(\omega_{i+1}, \omega_i) \cdot \prod_{j=1}^{n(\omega)} p(\gamma_j).$$

Rewrite it as

$$p(\mathbf{pt}(\omega); \Theta) = \frac{1}{Z} \exp\left\{ - \sum_{j=1}^{n(\omega)} \lambda(\gamma_j) - \sum_{i=1}^{n-1} \lambda(\omega_{i+1}, \omega_i) \right\}$$

The first term alone stands for a SCFG. The second term is the potentials (energy terms) on the context (chain).

This can be easily transferred to the image domain

The probability is defined on the parse graph *pg*

$$p(\mathbf{pg}; \Theta, \mathcal{R}, \Delta) = \frac{1}{Z(\Theta)} \exp\{-\mathcal{E}(\mathbf{pg})\},$$

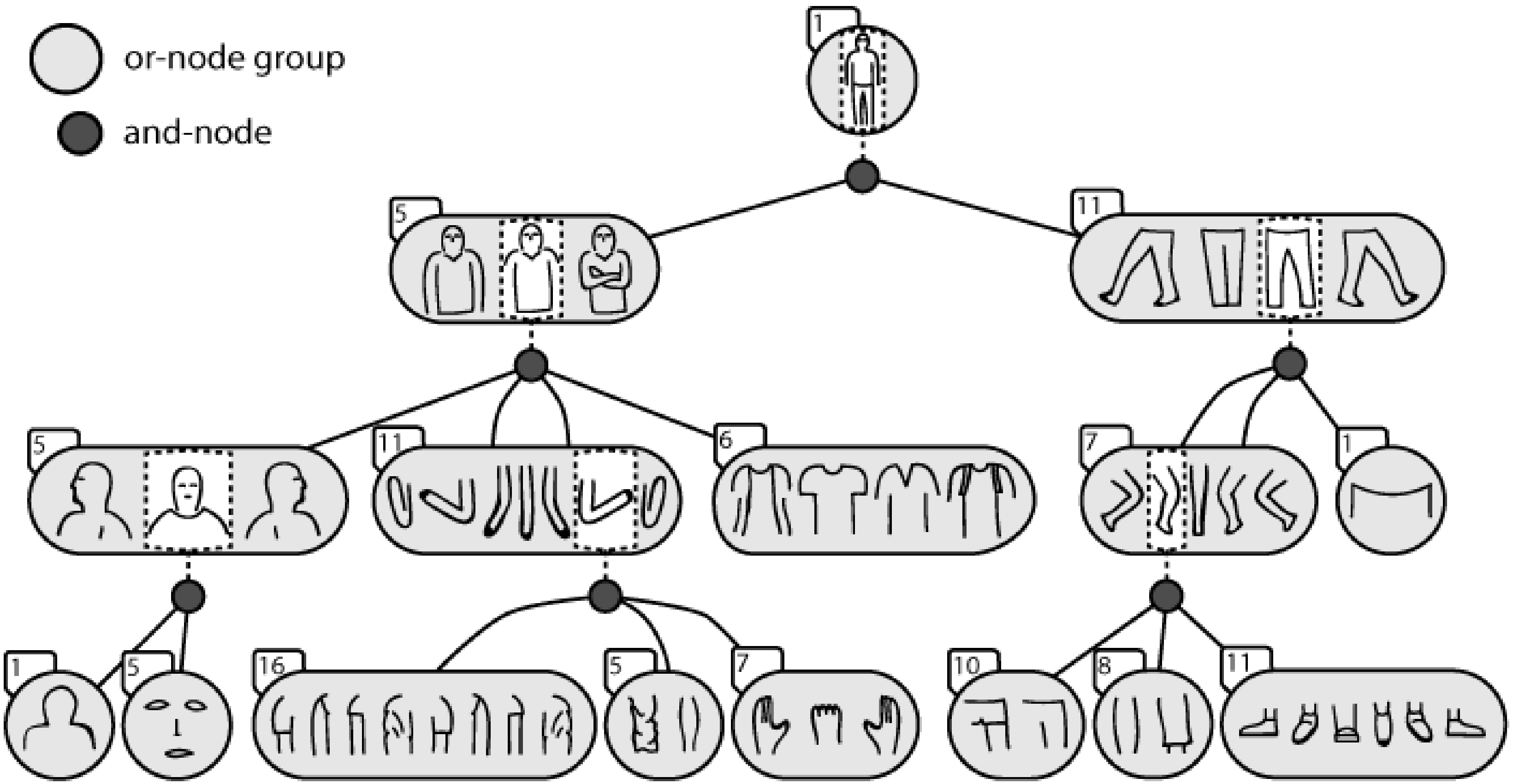
where $\mathcal{E}(\mathbf{pg})$ is the total energy,

$$\mathcal{E}(\mathbf{pg}) = \sum_{v \in V^{\mathrm{or}}(\mathbf{pg})} \lambda_v(\omega(v)) + \sum_{t \in T(\mathbf{pg}) \cup V^{\mathrm{and}}(\mathbf{pg})} \lambda_t(\alpha(t)) + \sum_{(i,j) \in E(\mathbf{pg})} \lambda_{ij}(v_i, v_j, \gamma_{ij}, \rho_{ij}).$$

$$Z = Z(\Theta) = \sum_{\mathbf{pg}} \exp\{-\mathcal{E}(\mathbf{pg})\}.$$

A comprehensive review reference: Zhu and Mumford, “A Stochastic Grammar of Images,” Foundations and Trends of Graphics and Vision, 2006.

3, Case studies: Spatial-AoG for human parsing

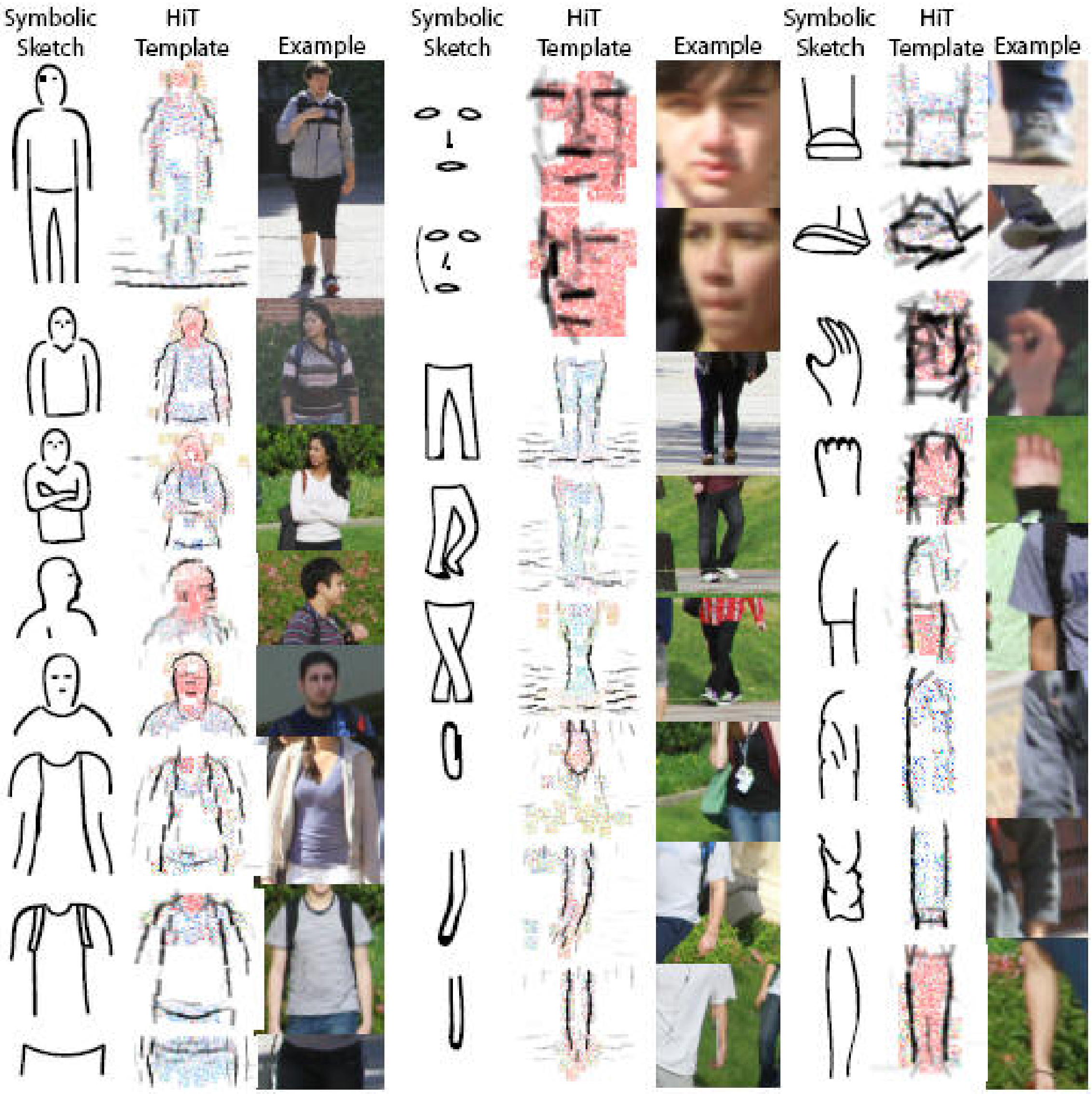


Rothrock and Zhu, 2011

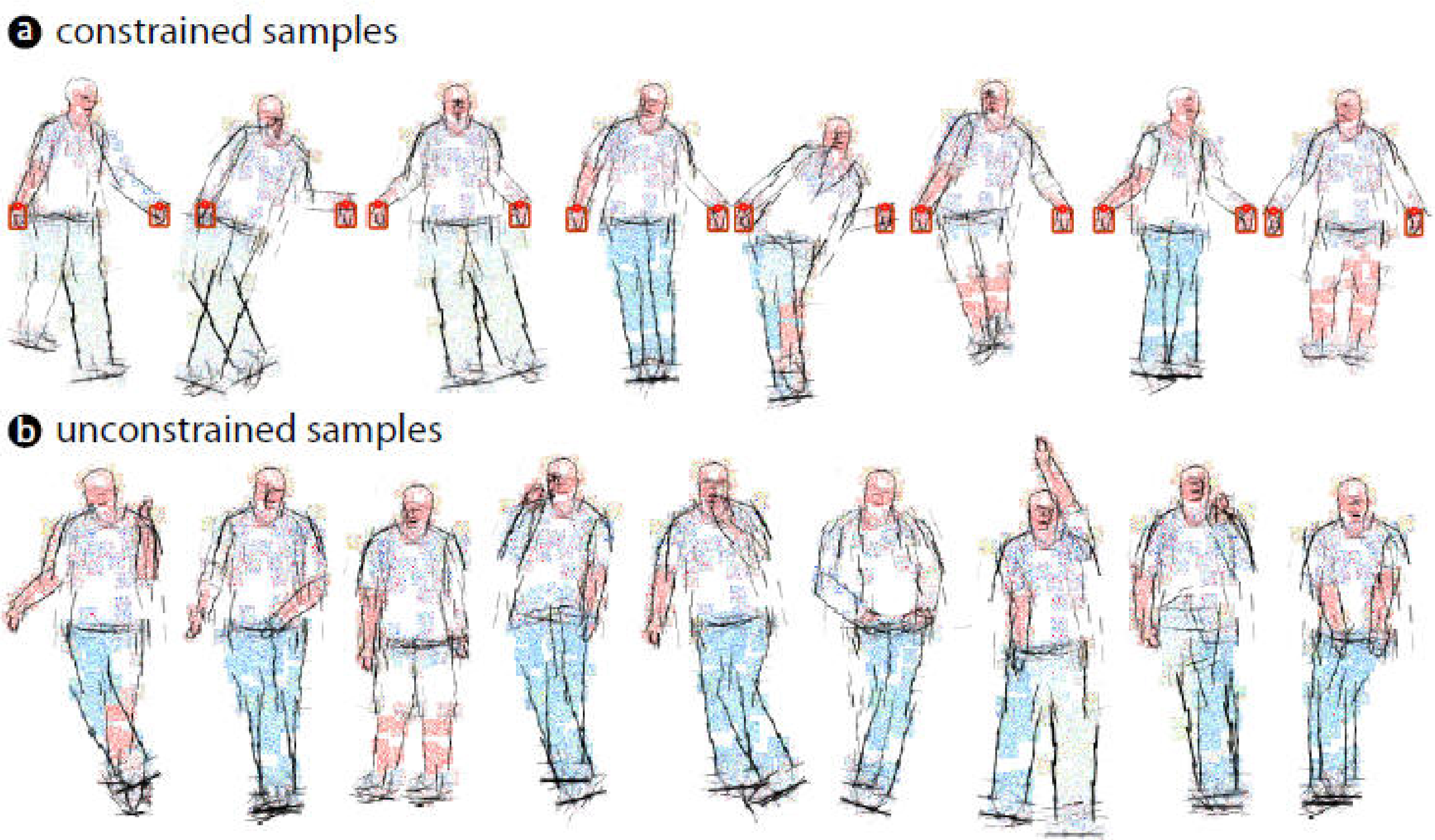
Appearance model for terminals

learned from images

Grounding the symbols



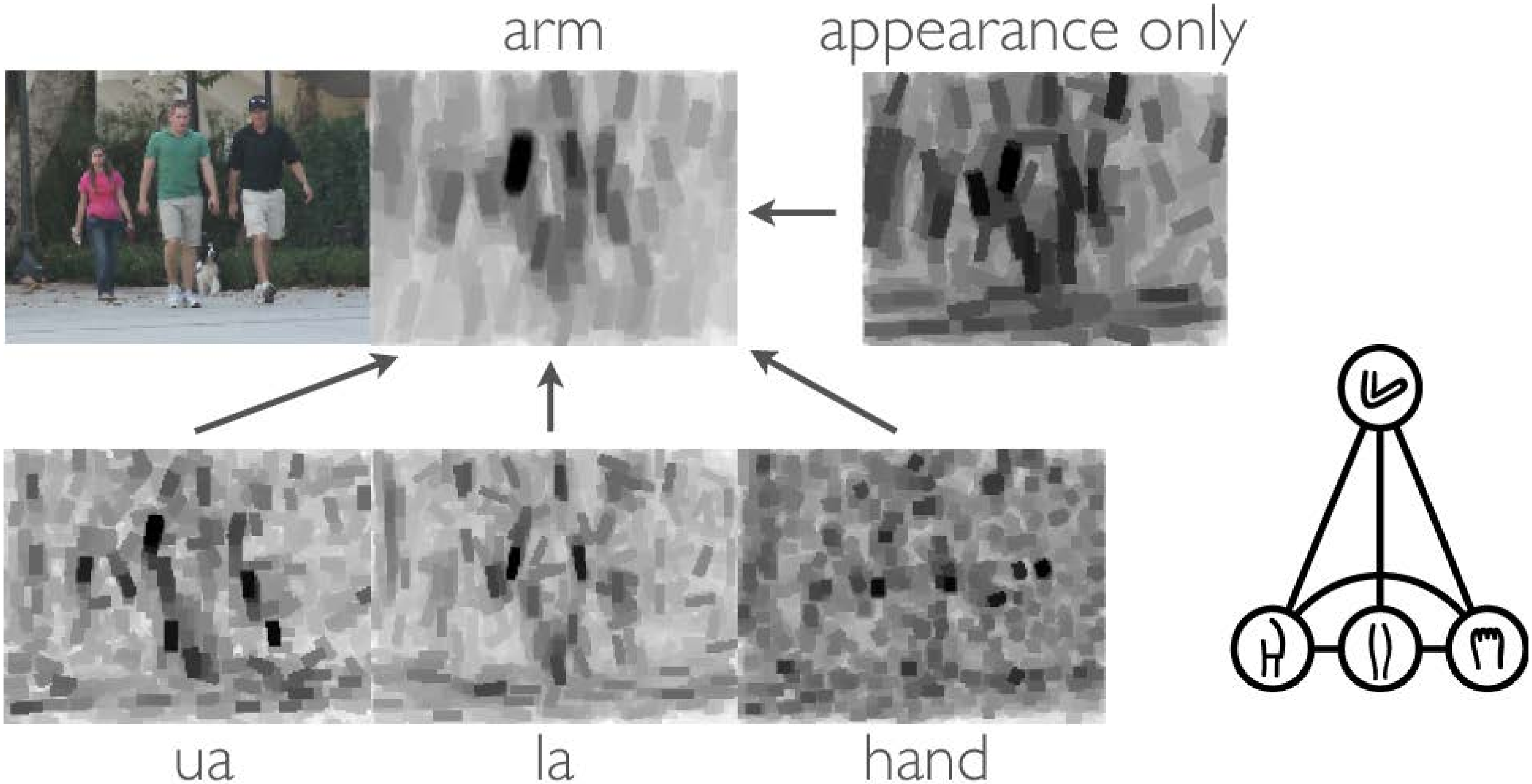
Synthesis (Computer Dream) by sampling the S-AoG



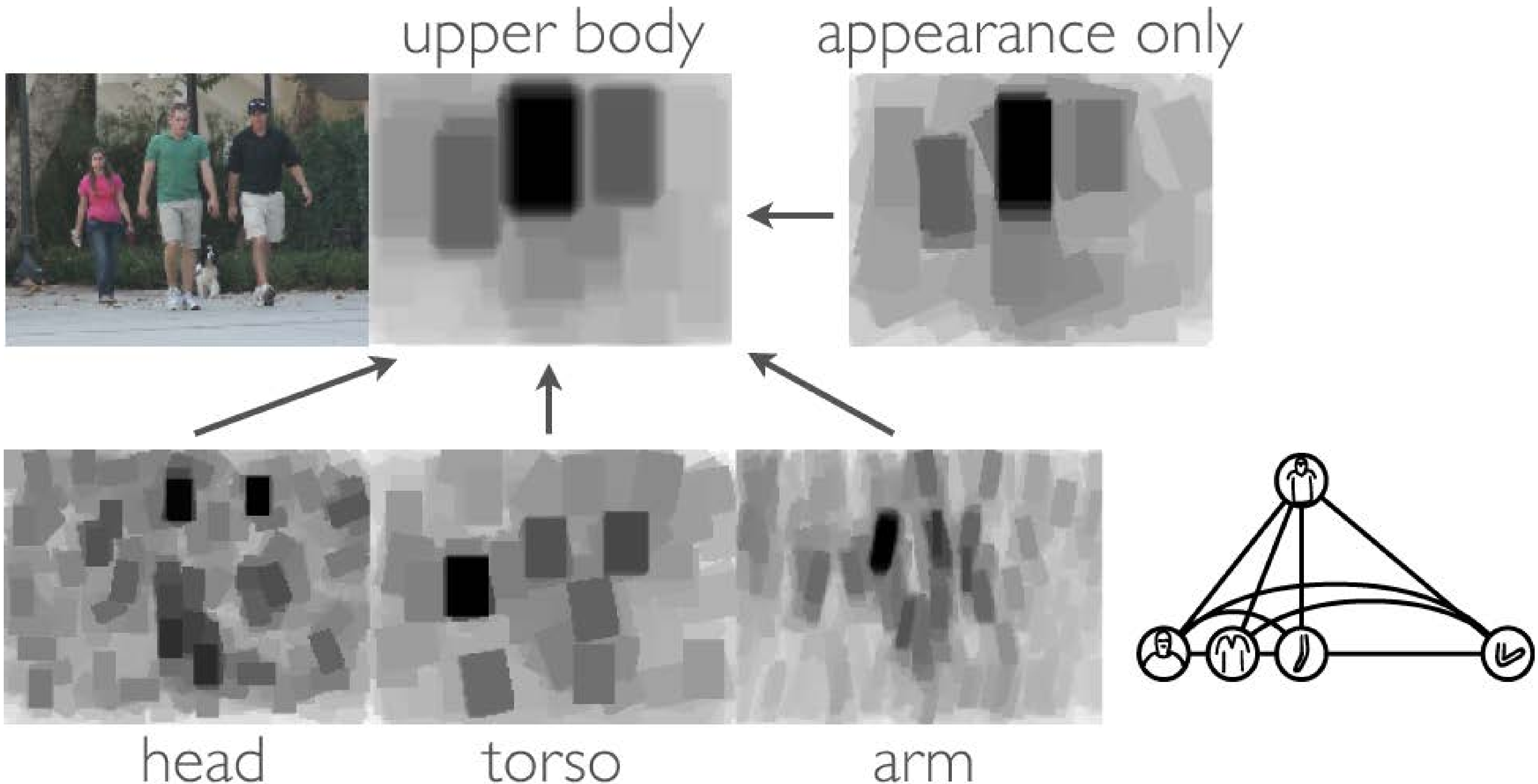
Rothrock and Zhu, 2011

Local computation is hugely ambiguous

Dynamic programming and re-ranking

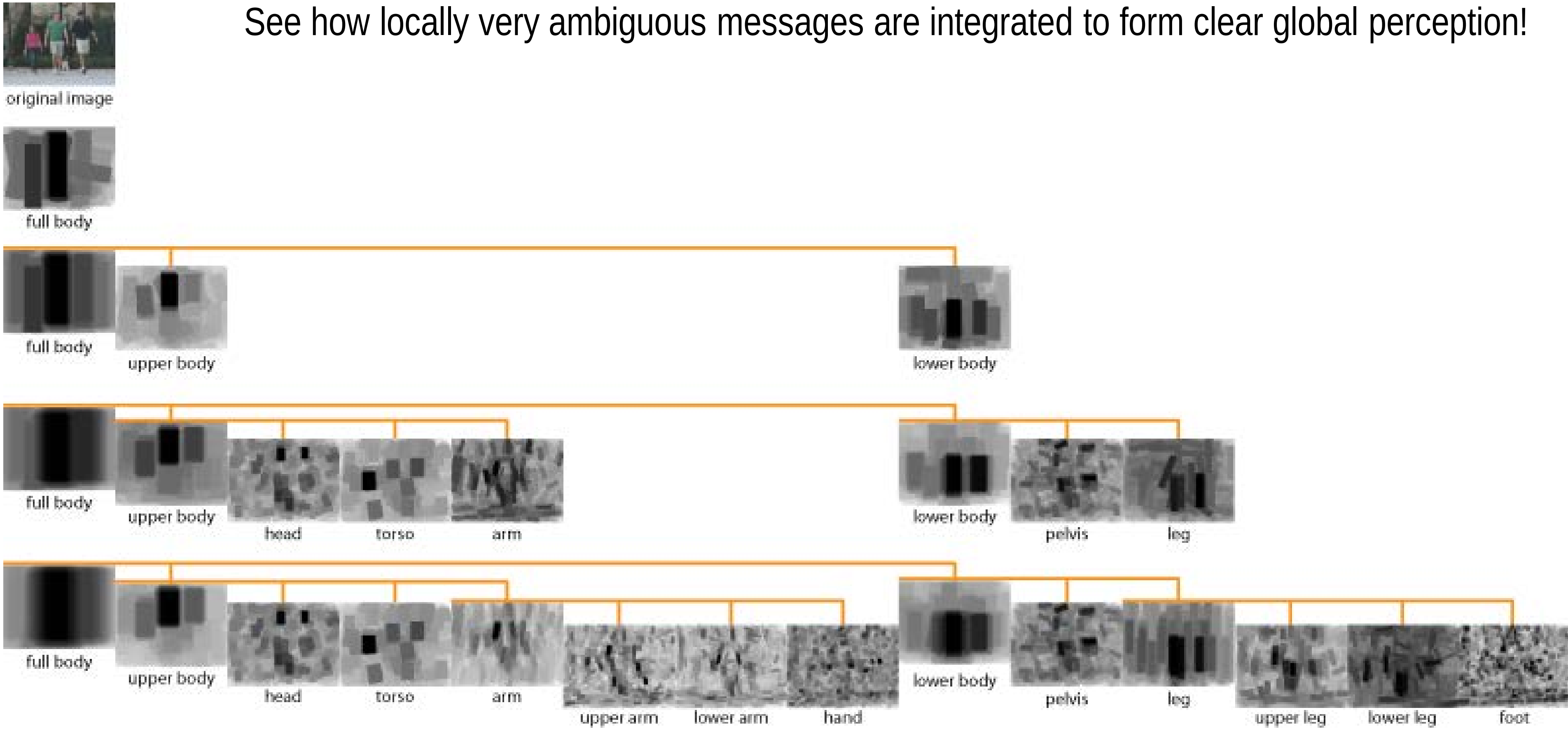


Composing Upper Body

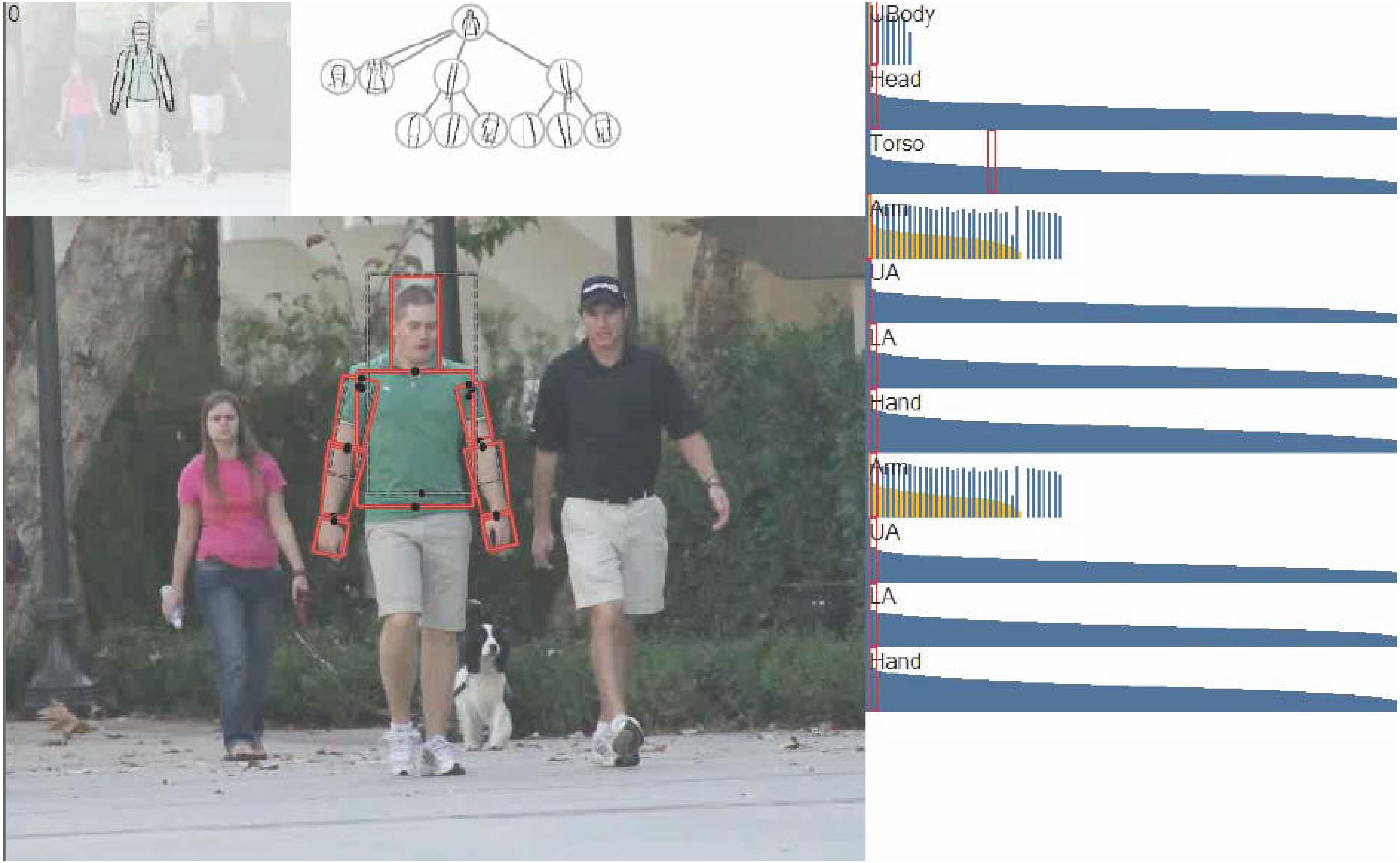


Composing parts in the hierarchy

See how locally very ambiguous messages are integrated to form clear global perception!

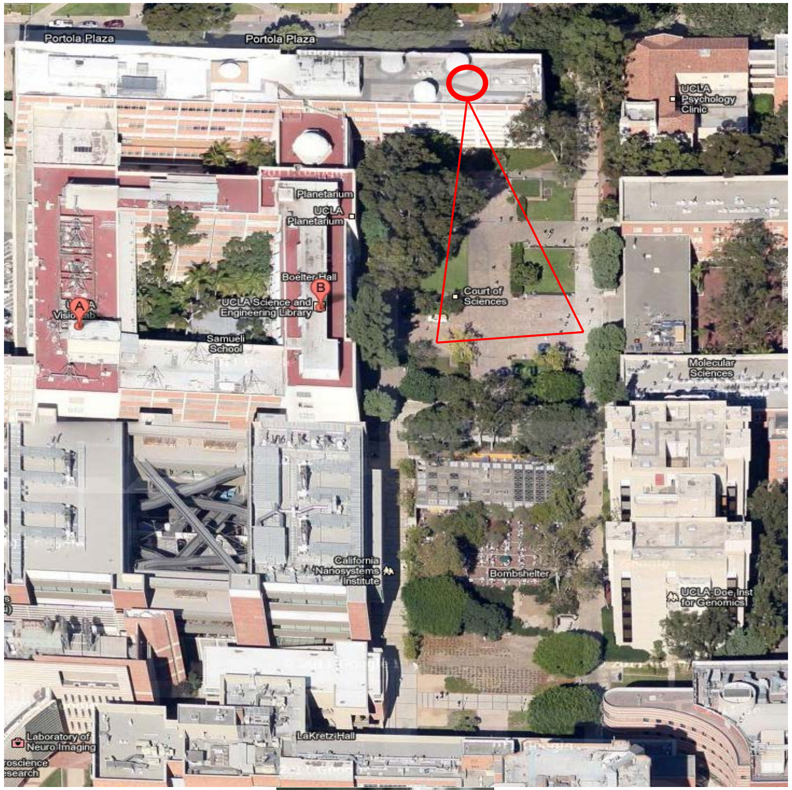


Top-down / bottom-up inference



Demo: Top-down / bottom-up inference for human parsing

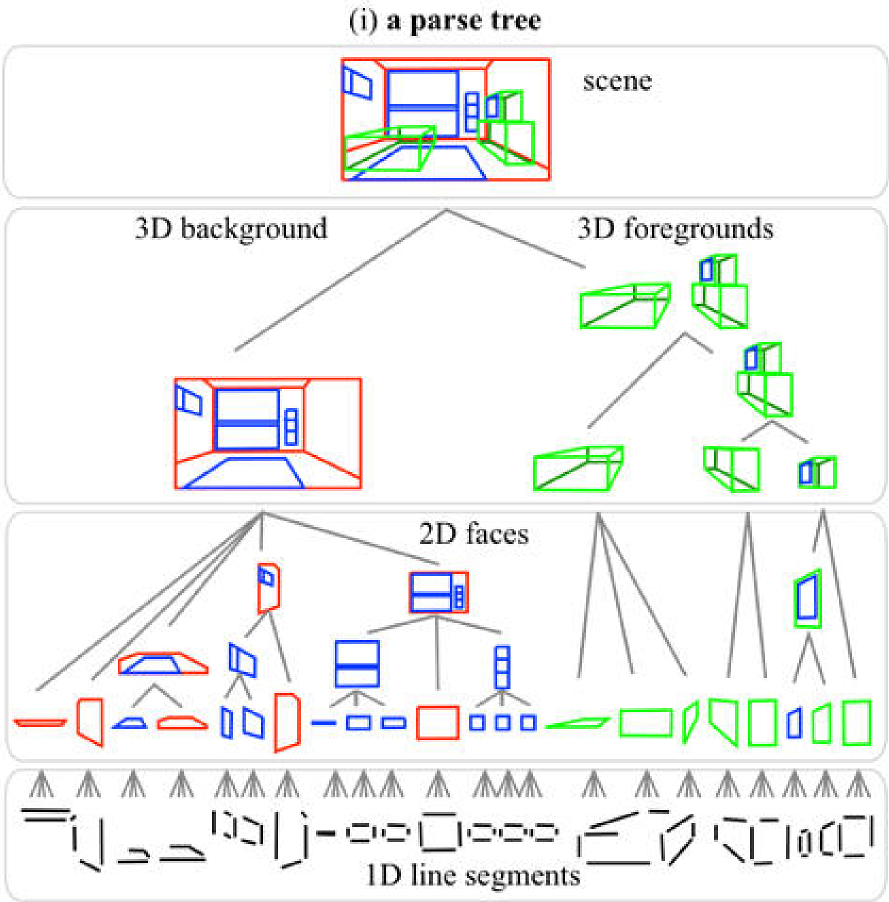
Rerank.avi



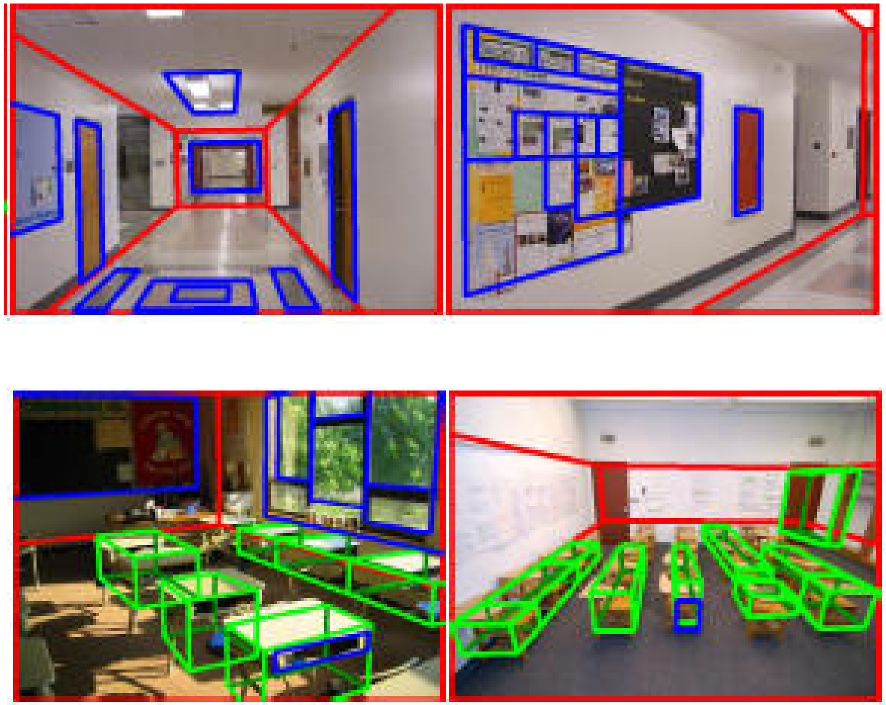
Brandon_parsing_human.avi

by Brandon Rothrock and   Tianfu Wu,   2011.

Case studies: Spatial-AoG for scene parsing

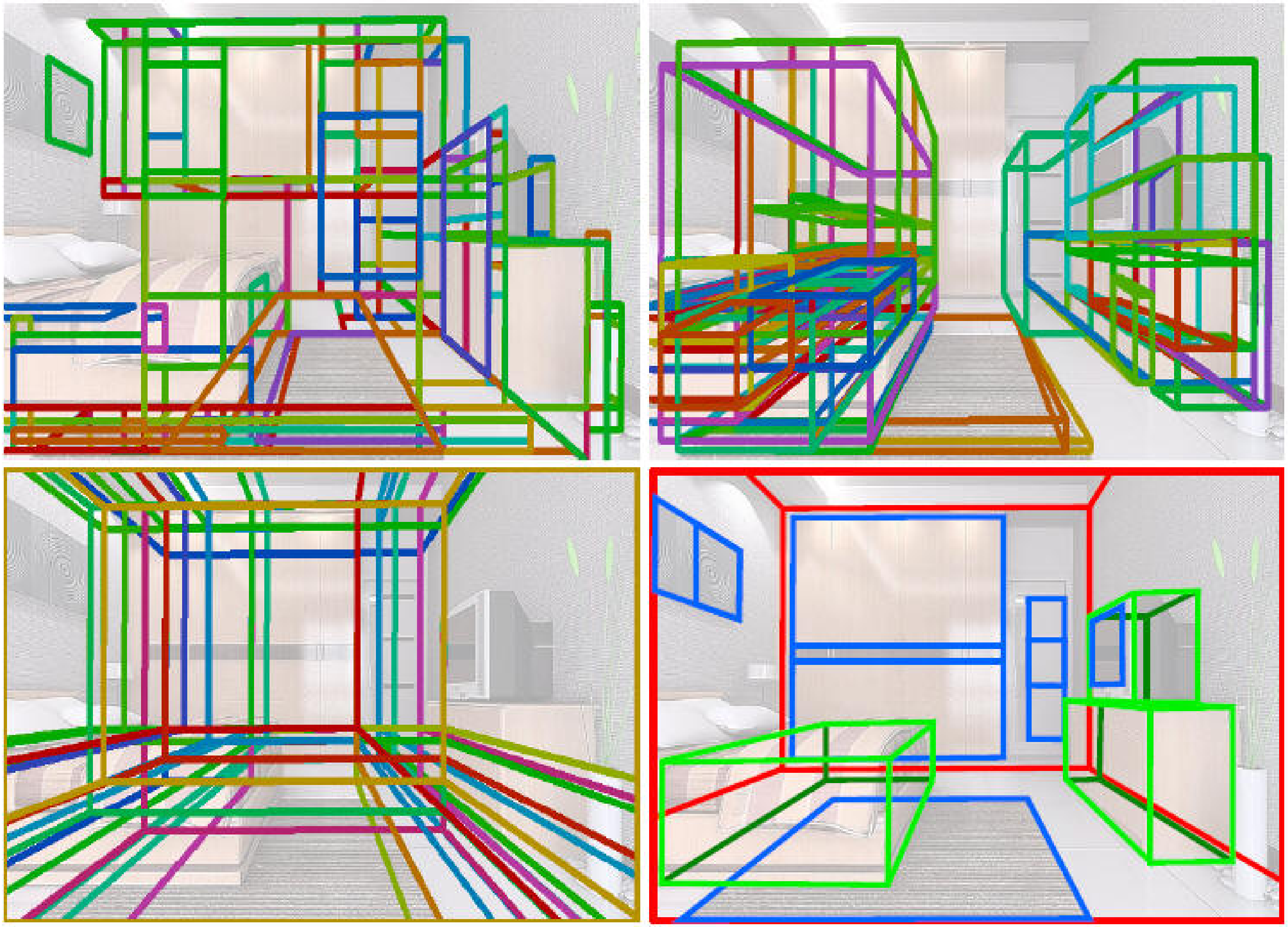


Results on the UCLA dataset

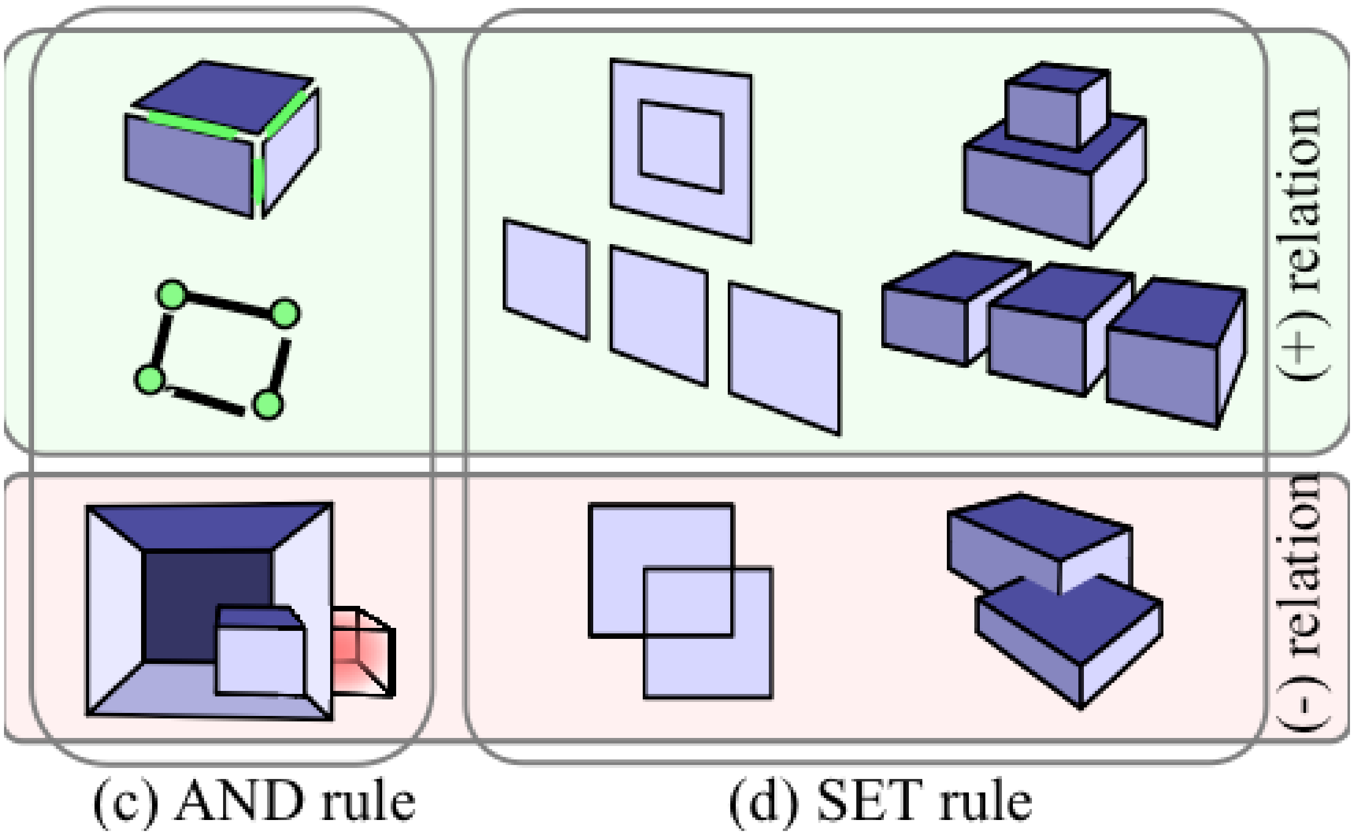


Ref: Y. Zhao and S.C. Zhu, NIPS, 2011.

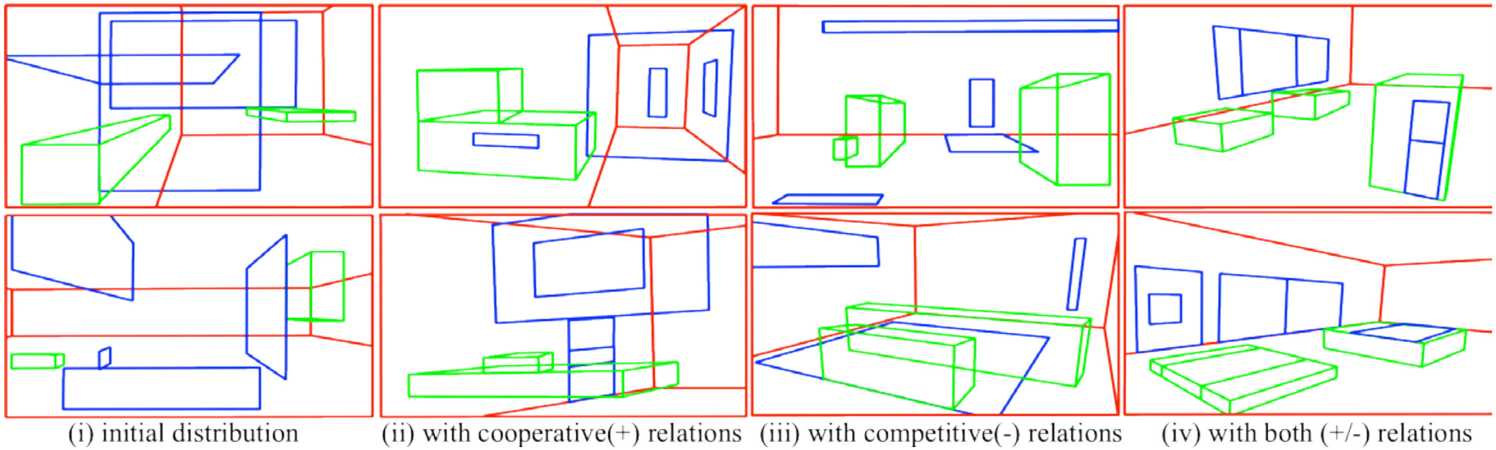
Bottom-up line detection



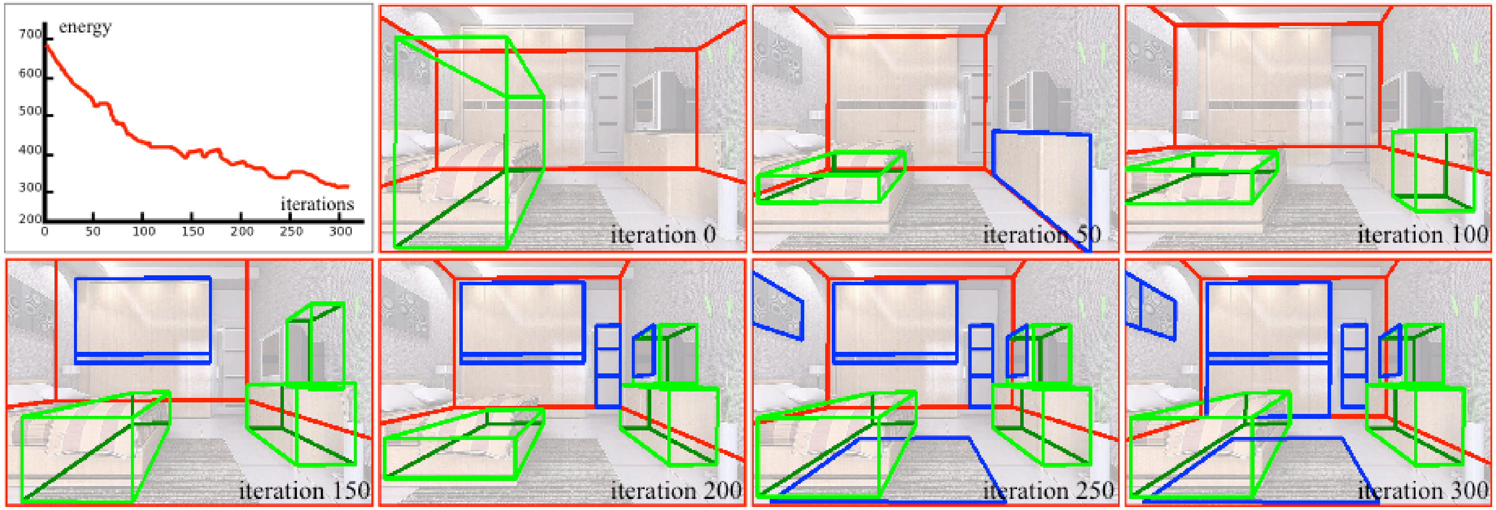
Contextual relations



Hierarchical cluster sampling from grammar model

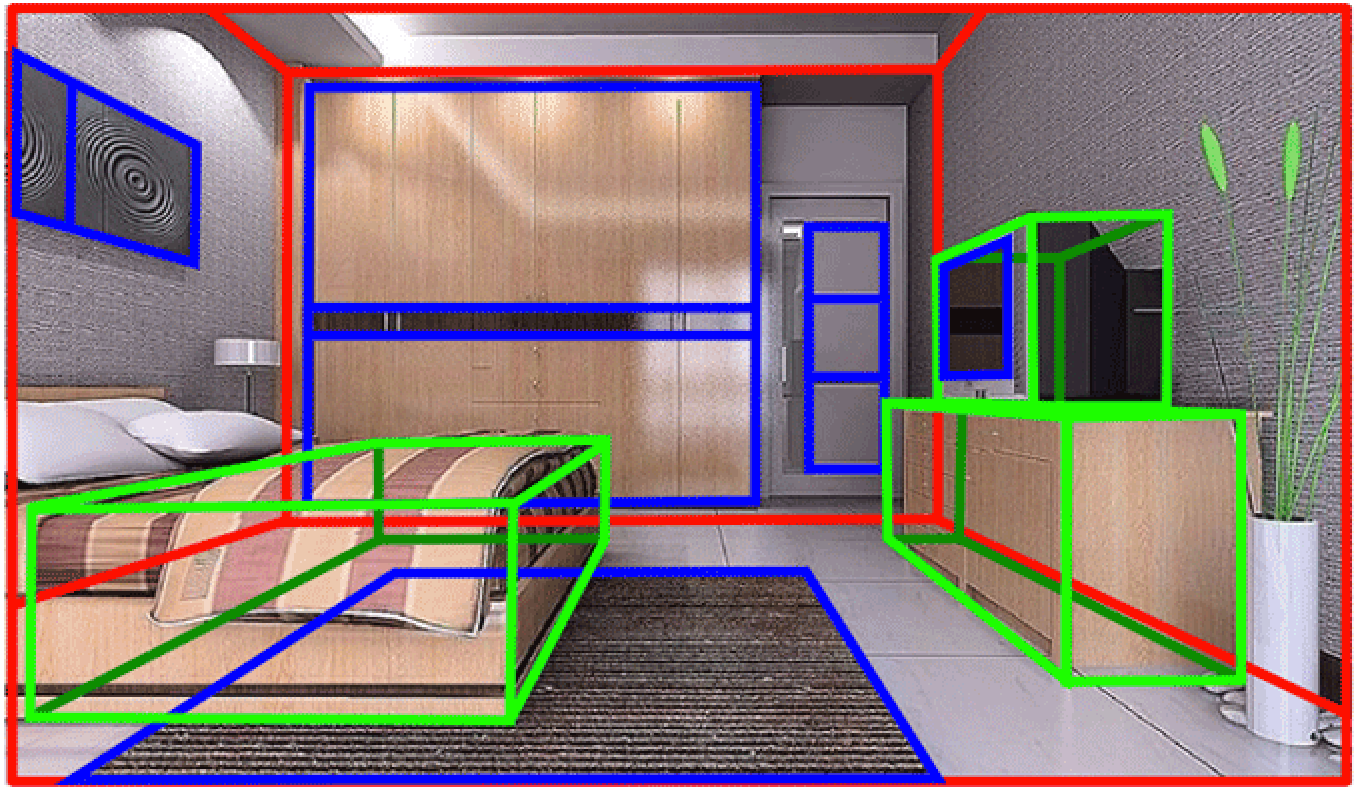


The prior sampling from Stochastic Scene Grammar with/without contextual relations



The hierarchical cluster sampling process.

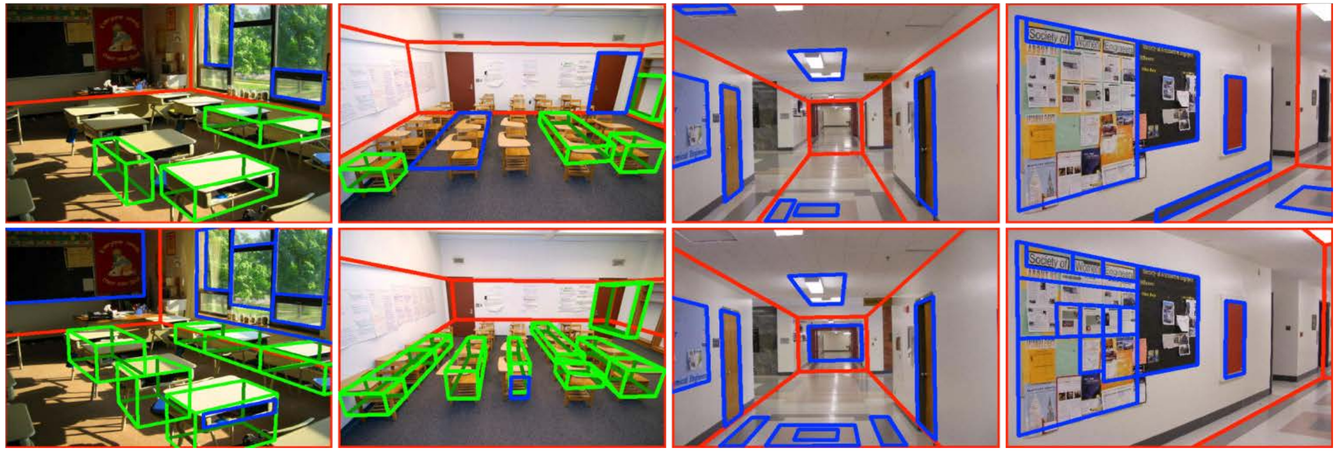
3D scene reconstruction from a single image



Parsing results on UCLA dataset

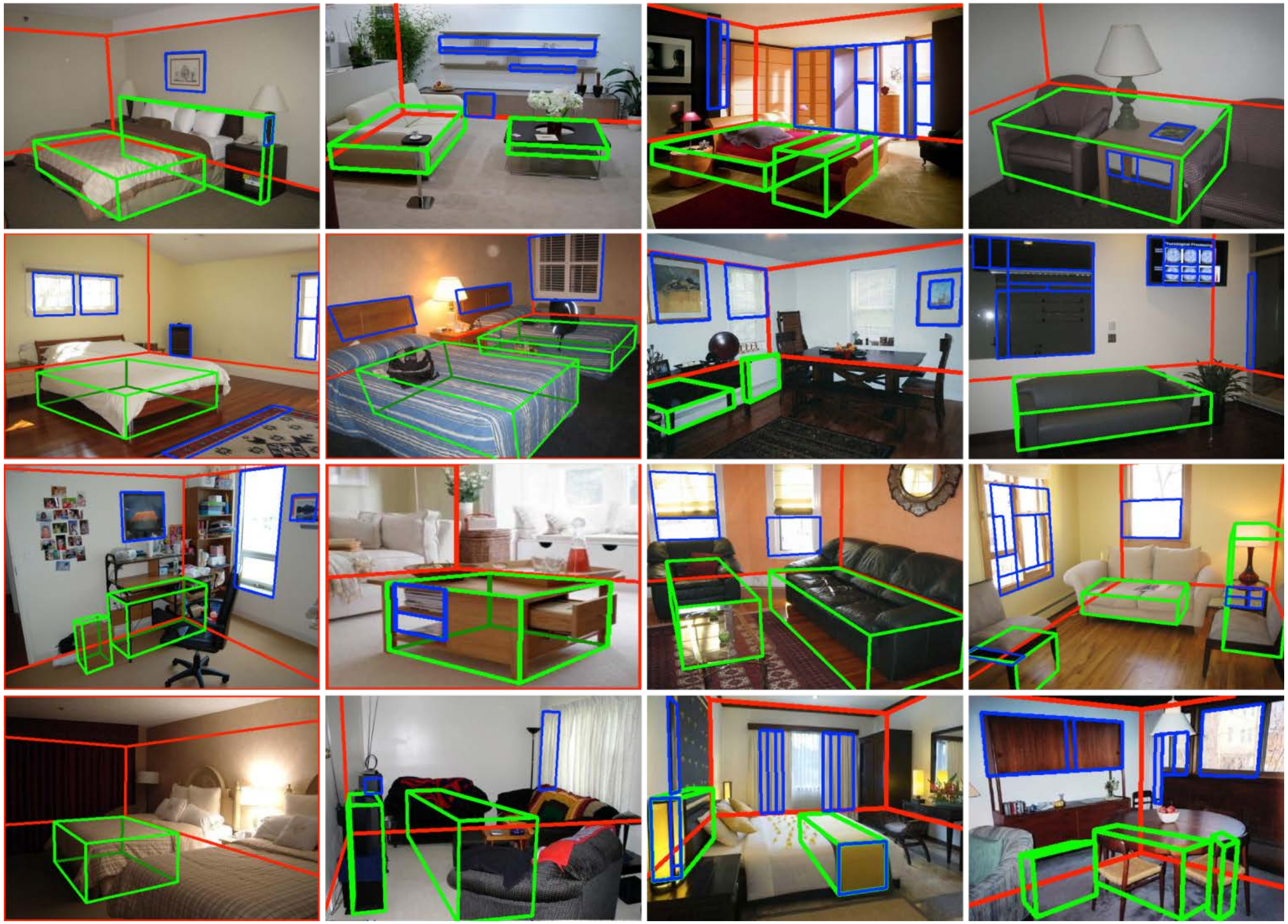
1.with AND/OR rule

2.with AND/OR/SET rule



Parsing results on UIUC dataset

More Results of Indoor Parsing



Demo of Spatial-AoG for scene parsing

Indoor_scene_parsing_demo.avi

This demo is made by Yibiao Zhao at UCLA