Vision as Bayesian Inference: Analysis by Synthesis.

Alan Yuille (Dept. Statistics. UCLA)

Joint with Dept. Computer Science and Dept. Psychology.

Bayes as Unified Framework for Cognition

- Probablistic Models of Cognition: Probabilistic Inference on Structured Representations.
- Organizers: Josh Tenenbaum (MIT) and Alan Yuille (UCLA).
- Summer School. July 2008. IPAM.
- Videos/PDF's available for download.

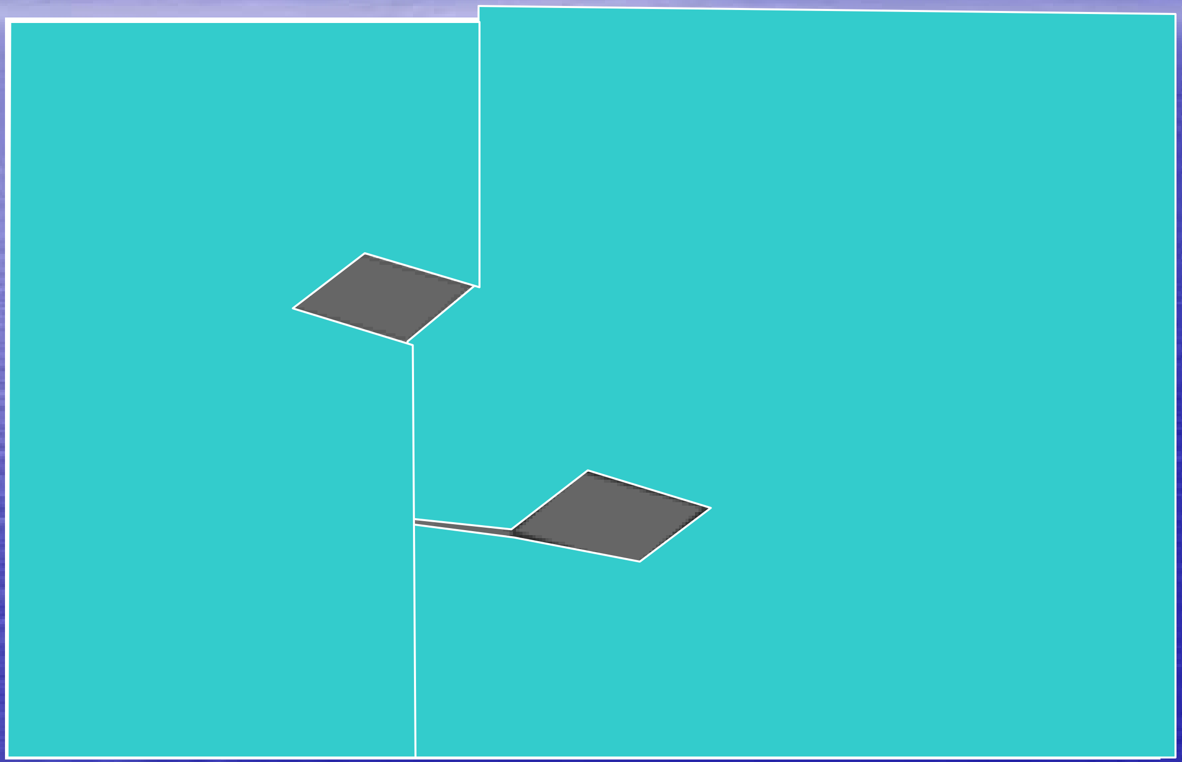
Difficulty of Vision

- *Vision is extremely difficult.*
- 50% cortex involved in vision.

- *Difficulty of vision* is due to the high-dimensionality of the data.
- More 10x10 images than seen by humans over all history.

- *Images are complex and ambiguous*.
- Vision is an act of creation.

Brightness of Patterns: Ted Adelson (MIT)



The Challenge of Vision.

- We have to come to terms with the complexity of real images. SC Zhu's "image genome" project.

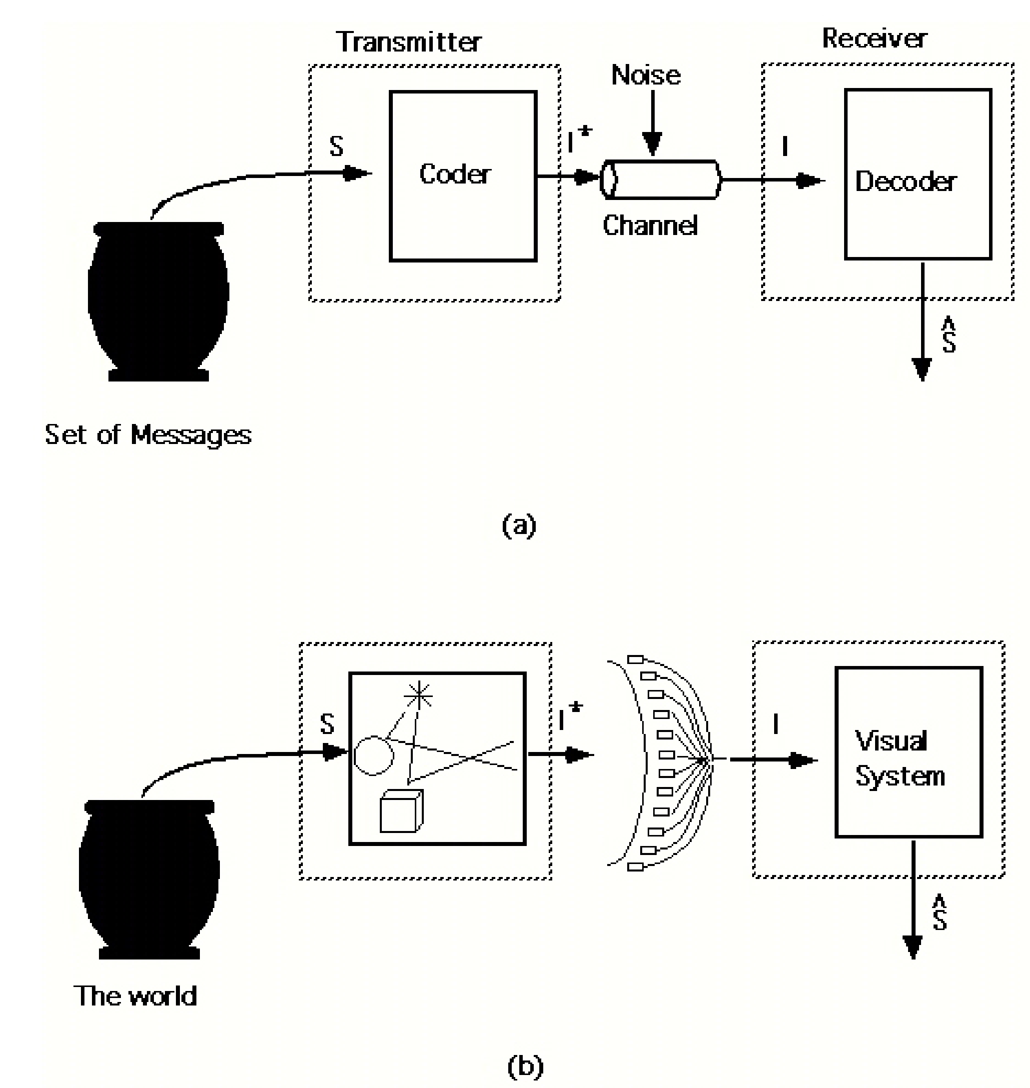
- *Attempts to understand the phenomenology of vision from artificial stimuli, though useful as a starting point, risk leading to faulty generalizations.*

- It is well known to computer vision researchers, that algorithms that work on artificial stimuli almost never generalize to natural images. (Julesz random dot stereograms).

Vision as Bayesian Inference: Analysis by Synthesis

- First formulated by Ulf Grenander in the 1970's.
- David Mumford (1992) speculated on how it could relate to the feedforward and feedback connections in the brain.
- It can be used to model a range of psychophysical phenomena – see reviews (Kersten, Mamassian, Yuille 2006, Geisler and Kersten 2004,…). Bayesian Ideal Observers.
- Multi-cell recordings in V1, V2 by Lee (Lee & Mumford, Lee & Yuille). fMRI studies (Kersten lab.)

Vision: Decoding Images

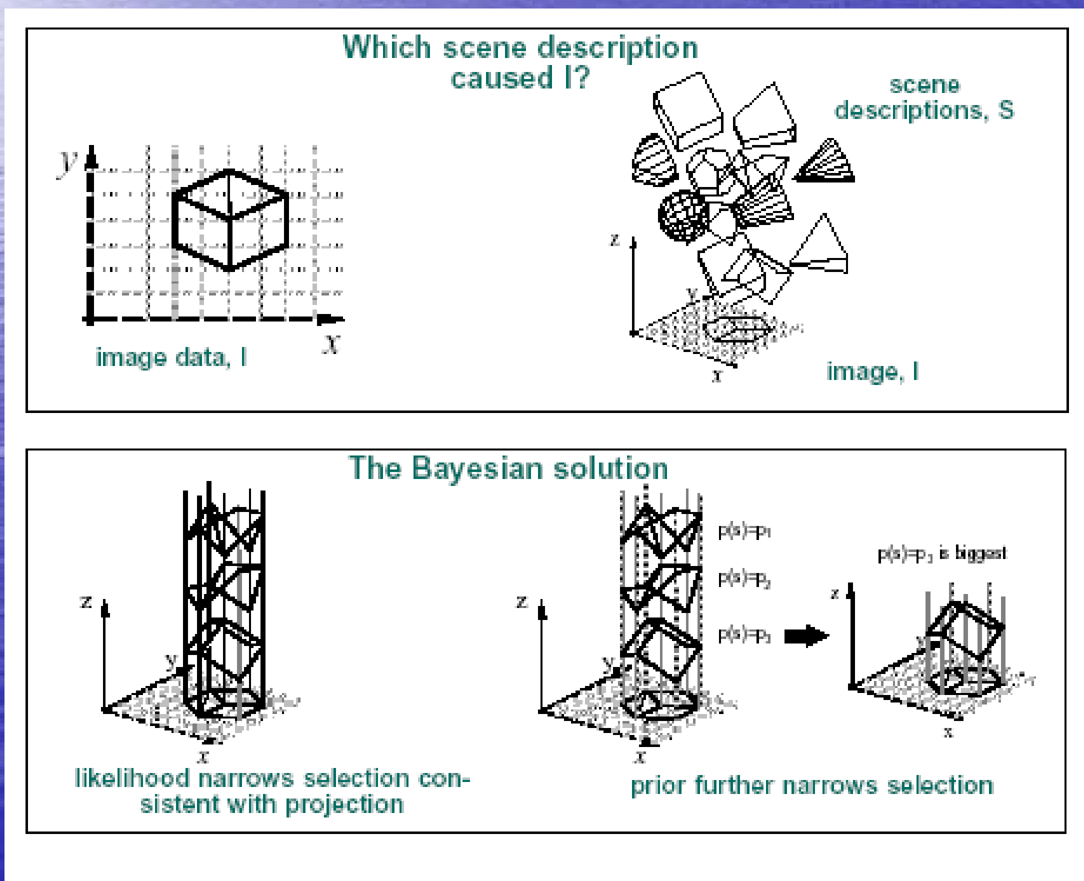


An Inverse Problem: Apply Bayes Theorem

Task: estimate S from I

Bayes to Infer S from I

- $P(S|I) = P(I|S)\, P(S) / P(I)$



Pavan Sinha (MIT)

Image Parsing.

- (I) Image are composed of visual patterns:
- (II) Parse an image by decomposing it into patterns.

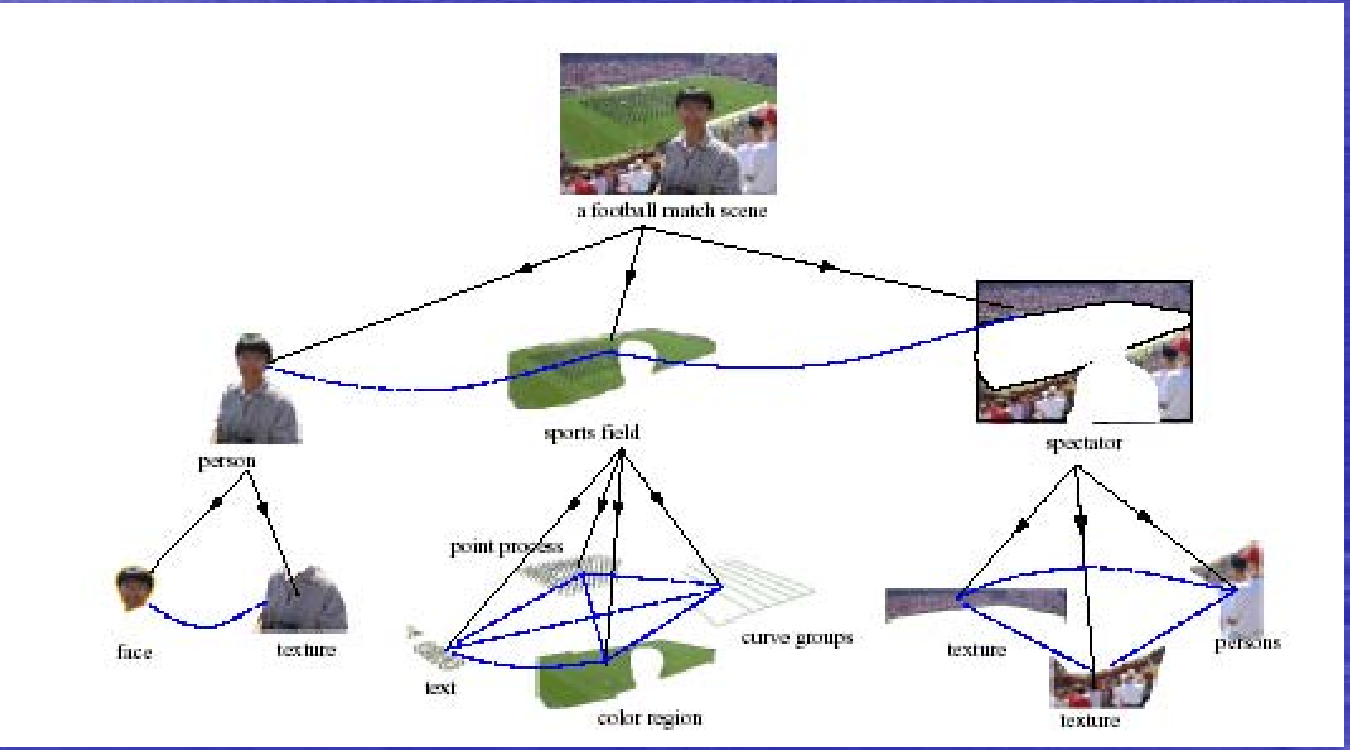


Image Parsing. (Tu et al 2003/2005)

- Stochastic models for generating images in terms of ***visual patterns***.
- Visual patterns can be ***generic*** (texture/shading) or ***objects (***faces and text***)***.

Parsing Graph.

- Nodes represent visual patterns. Child nodes to image pixels.

Stochastic Grammars:
Manning & Schultz.

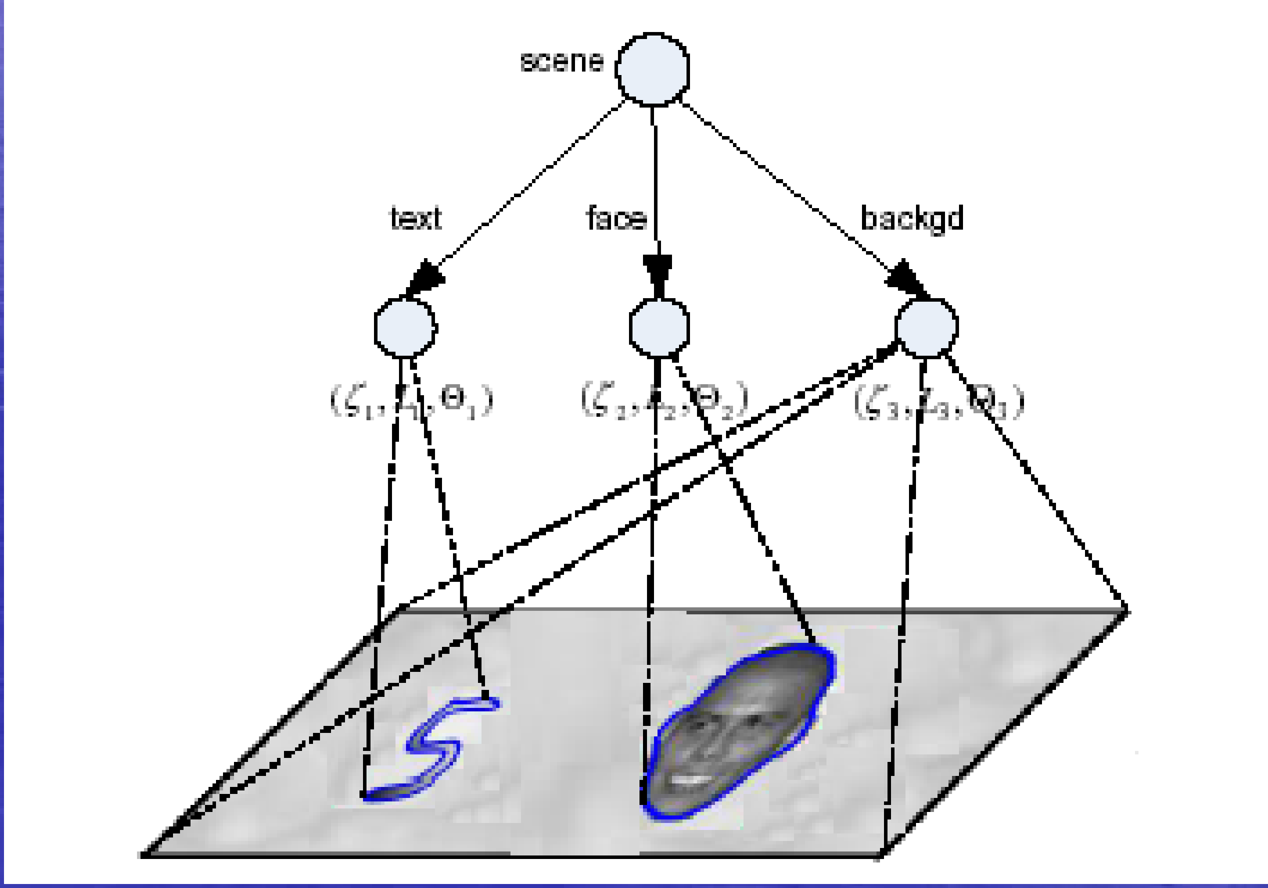


Image Patterns.

- Node attributes: $\zeta_i, L_i, \Theta_i.$
- ***Zeta***: Pattern Type – 66

  (I) Gaussian, (II) Texture/Clutter, (III) Shading. (IV) Faces, (V– LXVI) Text Characters.
- ***L*** – shape descriptor (image region modeled).
- ***Theta***: Model parameters.

$$W = (K, \{(\zeta_i, L_i, \Theta_i) : i = 1, 2, ..., K\}).$$

Generative Model:

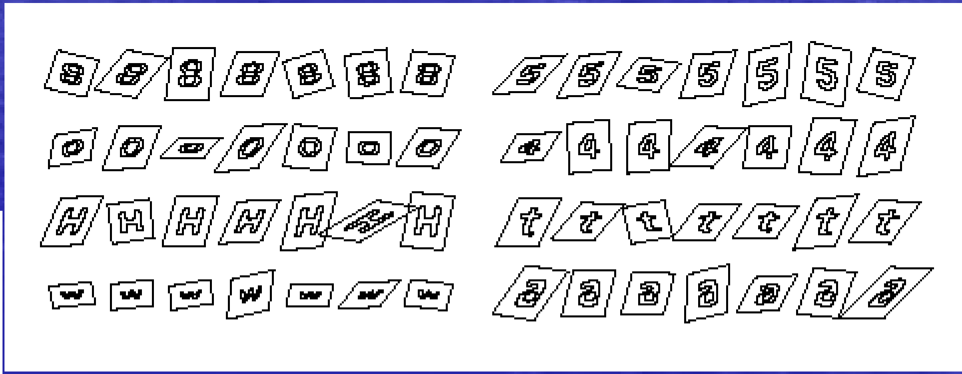
- Likelihood:

$$p(\mathbf{I}|W) = \prod_{i=1}^{K} p(\mathbf{I}_{R(L_i)}|\zeta_i, L_i, \Theta_i).$$

- Prior:

$$p(W) = p(K)\prod_{i=1}^{K} p(L_i)p(\zeta_i|L_i)p(\Theta_i|\zeta_i).$$

- Samples:

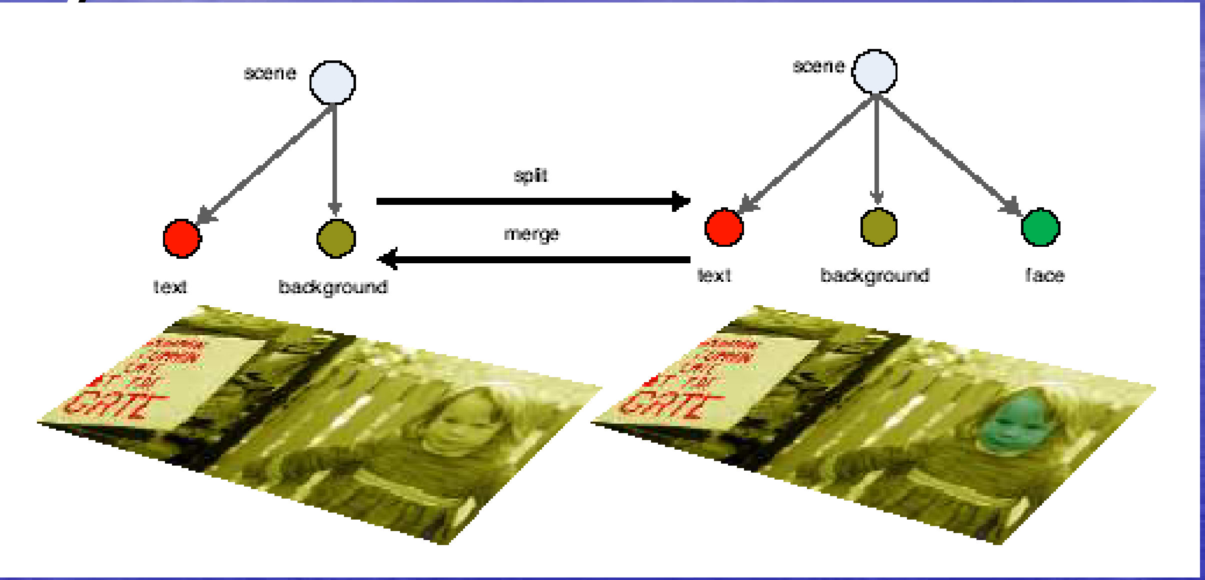


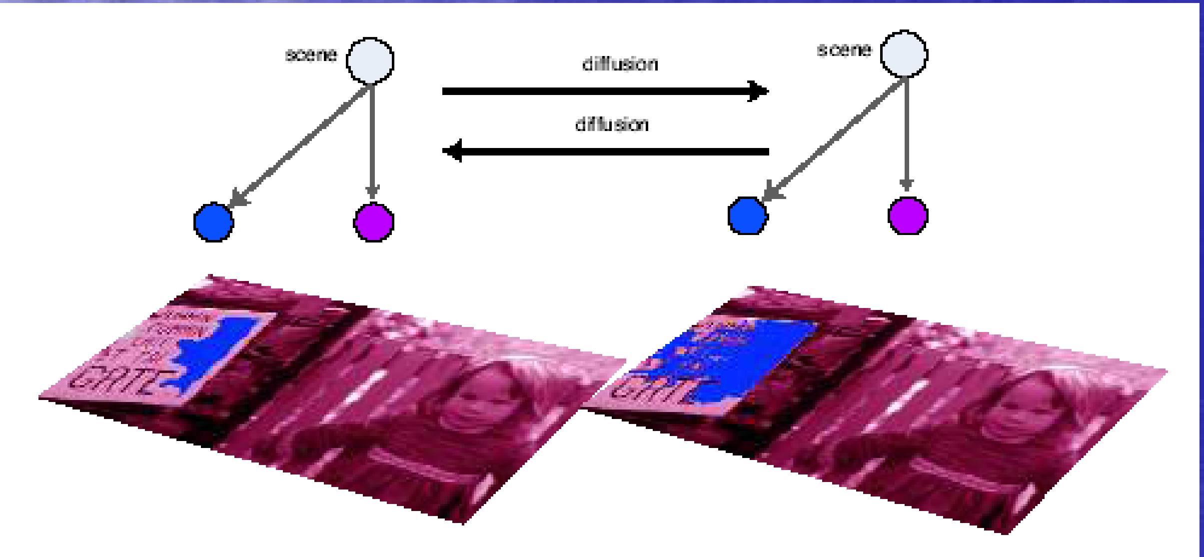
Inference Algorithm

- Want to sample from P(W|I)
- Data-Driven Markov Chain Monte Carlo (DDMCMC).
- Interpreting an image corresponds to constructing a ***parse graph***.
- Set of ***moves*** for constructing the parse graph.
- Dynamics for moves use bottom-up & top-down visual processing.

Inference Dynamics

Moves:





Data Driven Markov Chain Monte Carlo.

- Design a Markov Chain (MC) with transition kernel

$$K(W'|W : \mathbf{I})$$

- Satisfies Detailed Balance. $p(W|\mathbf{I})K_a(W'|W : \mathbf{I}) = p(W'|\mathbf{I})K_a(W|W' : \mathbf{I}).$

- Then repeated sampling from the MC will converge to samples from the posterior P(W|I).

Moves & Sub-kernels.

- Implement each move by a transition sub-kernel:

$$K_a(W'|W:\mathbf{I})$$

- Combines moves by a full kernel:

$$K(W,W') = \sum_i \alpha_i(I) K_i(W,W'), \quad \sum_i \alpha_i(I) = 1$$

- At each time-step – choose a type of move, then apply it to the graph.
- Kernels obey:

$$\sum_W K(W,W') P(W|I) = P(W'|I)$$

Data Driven Proposals.

- Use data-driven proposals to make the Markov Chain efficient.
- Metropolis-Hastings design:

$$K_i(W, W') = Q_i(W, W'|Tst_i(\mathbf{I})) \min\{1, \frac{P(W'|\mathbf{I})}{P(W|\mathbf{I})} \frac{Q_i(W, W'|Tst_i(\mathbf{I}))}{Q_i(W, W'|Tst_i(\mathbf{I})}\}.$$

- Proposal probabilities are based on discriminative cues.

$$Q_i(W, W'|\mathbf{I})?$$

Proposals from Discriminative Cues

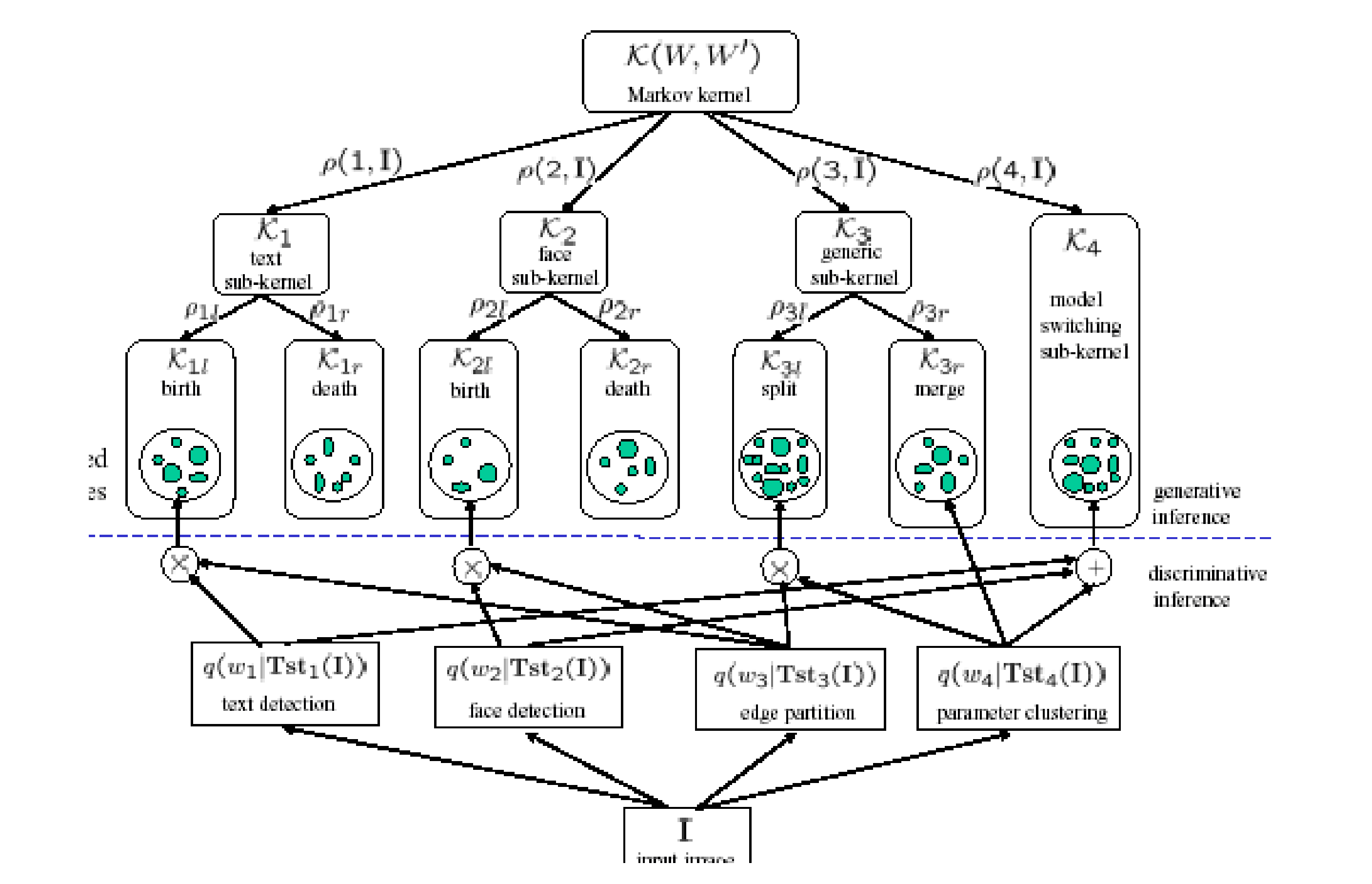
- Proposals Q(.|.) are obtained from machine learning.
- For example, AdaBoost gives proposals for the presence/absence of faces and text.

Illustration for finding text.

- Text Detection and Binarization.

Full Strategy:

- Integration:

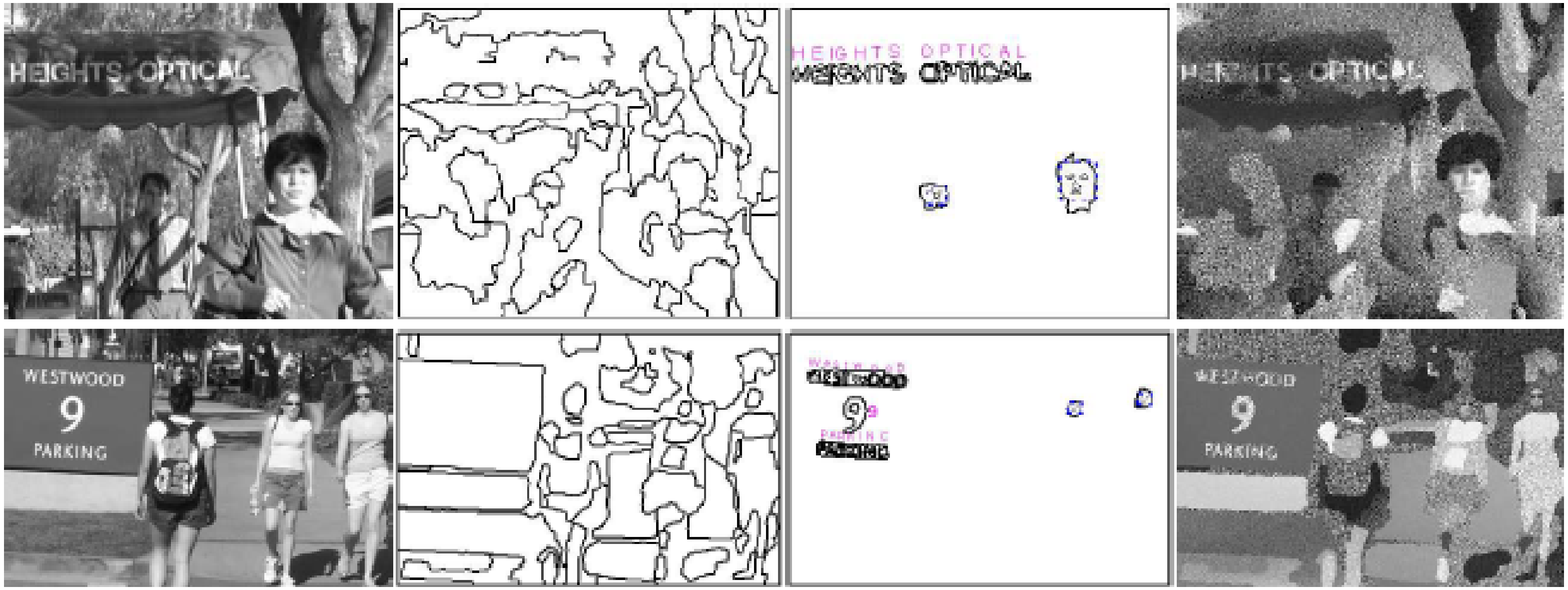


Bottom-Up Proposals.

- Bottom-up proposals for faces and text. False positives and false negatives.

High-Level Models validate bottom-up cues and resolve ambiguities:

- Competition & Cooperation.

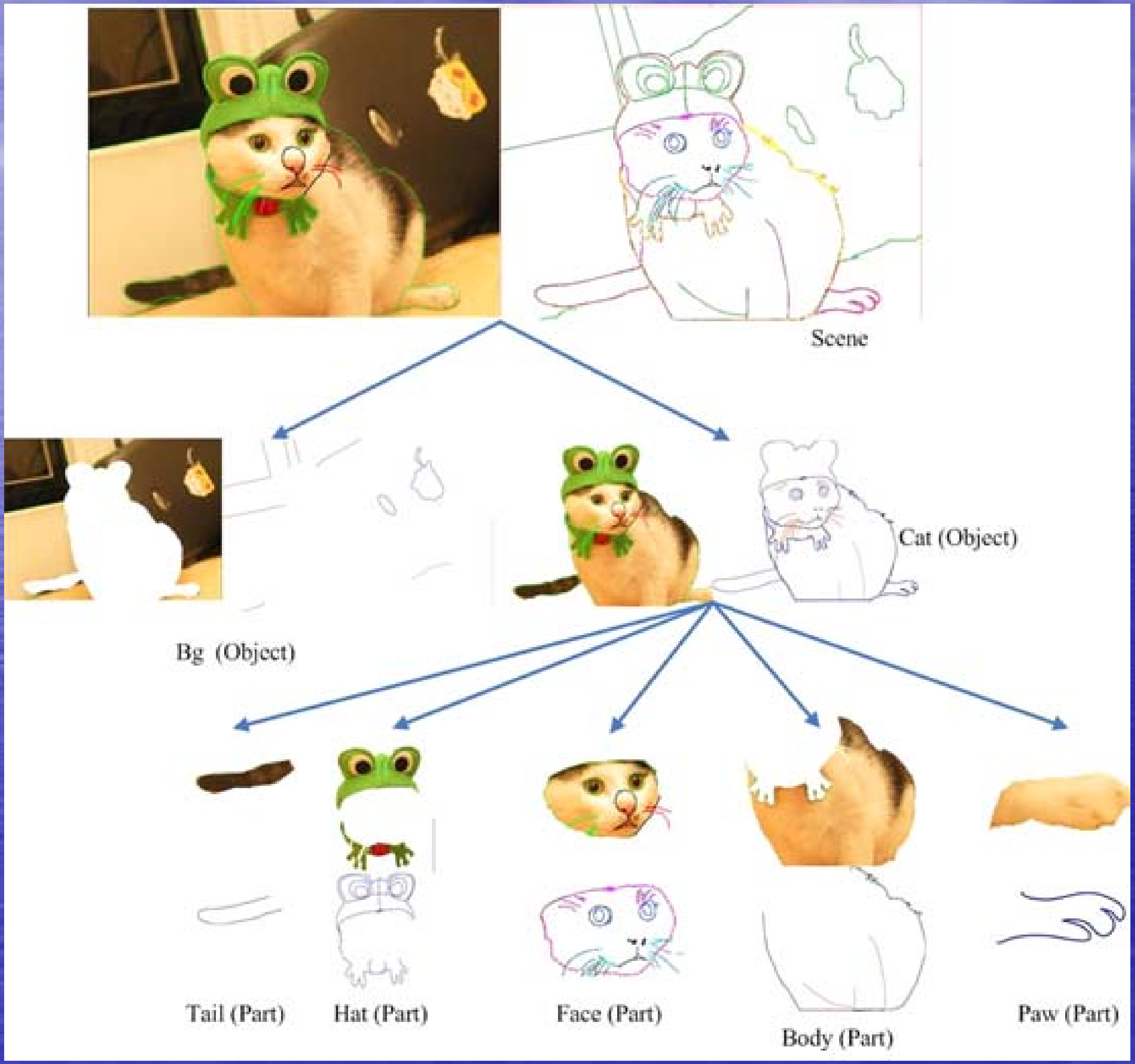


a. Input image b. Segmentation c. Object recognition d. Synthesized image

Image Parsing (2008)

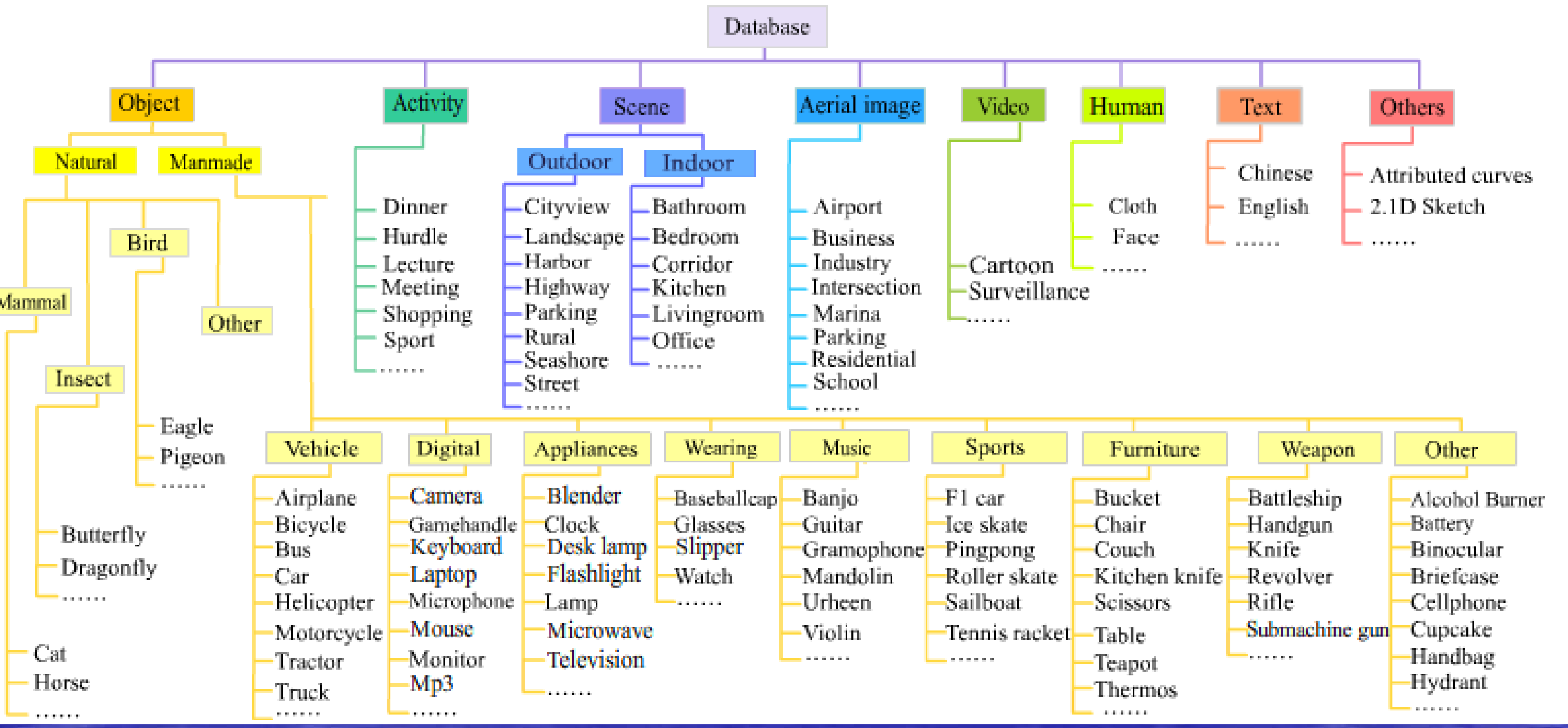
- Current work (Zhu et al) extends parsing to include far more patterns.
- This is part of his “image genome” project at the Lotus Hill Institute (China).

An example: parse graph of a cat



Over 1,000,000 hand-parsed images

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

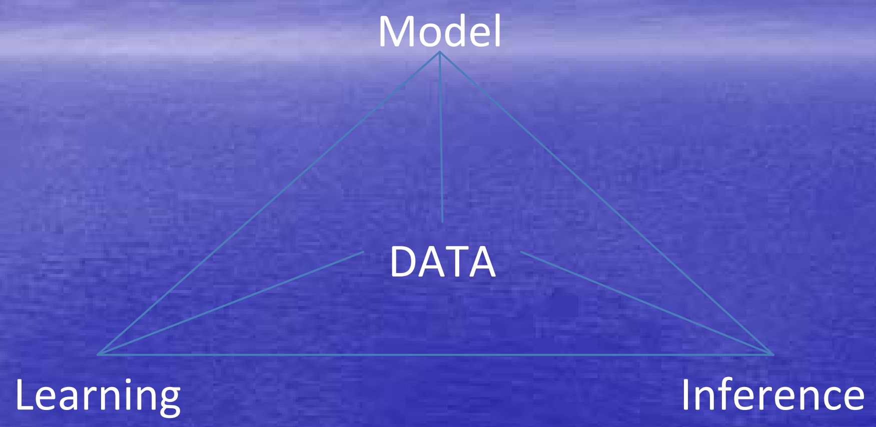


Structure Learning of Hierarchical Object Models

Leo Zhu and Alan Yuille
Department of Statistics
University of California Los Angeles
May. 2008

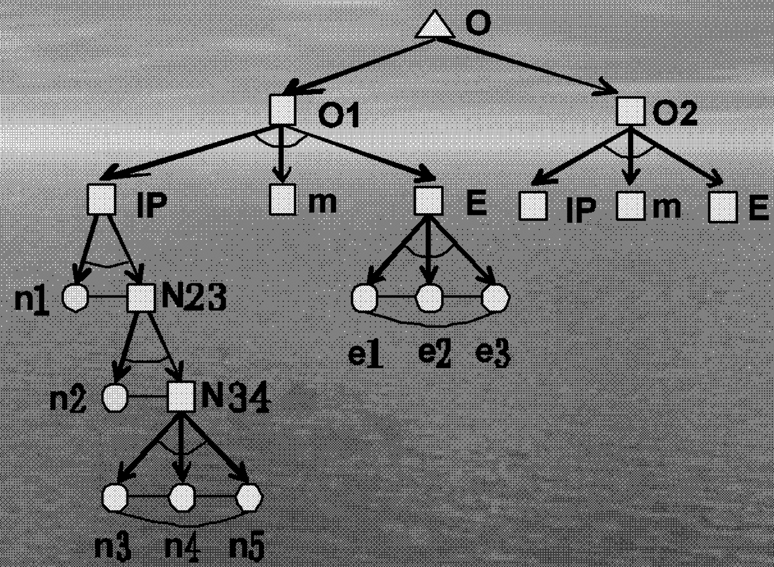
Research Program on Objects

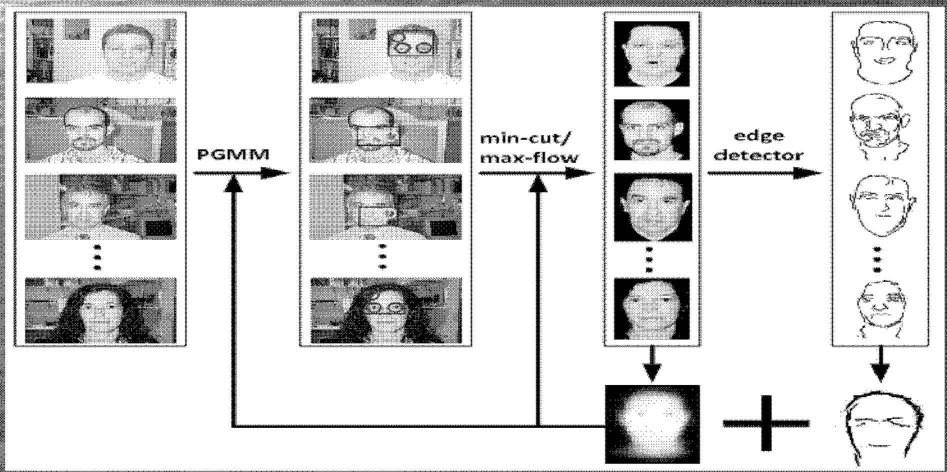
- Model
- Inference
- Learning

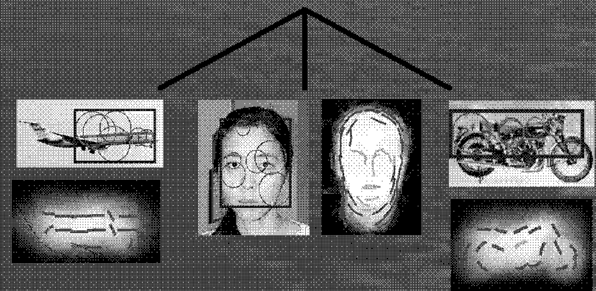


Unsupervised Learning of Probabilistic Object Models

- Represent object classes by a mixture model
- Cues come from different sources
- Efficient Learning from spare interest points to dense region statistics

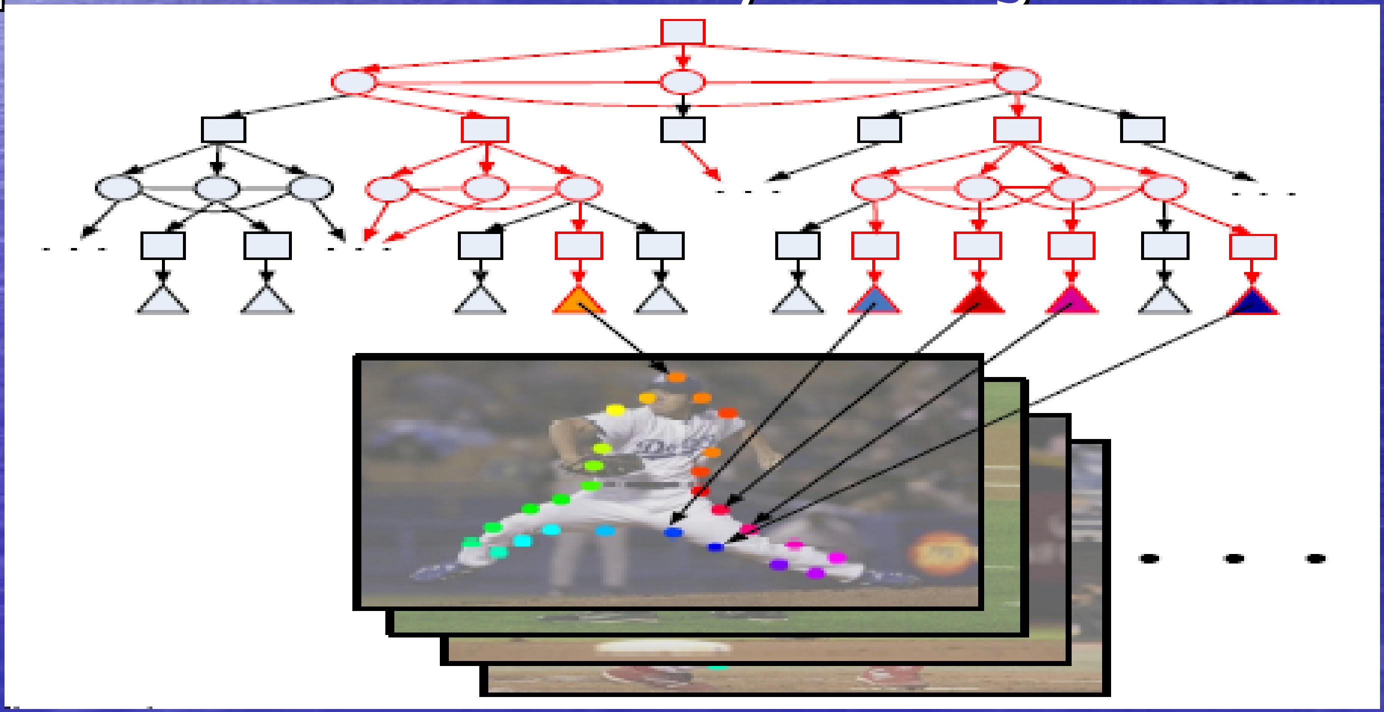




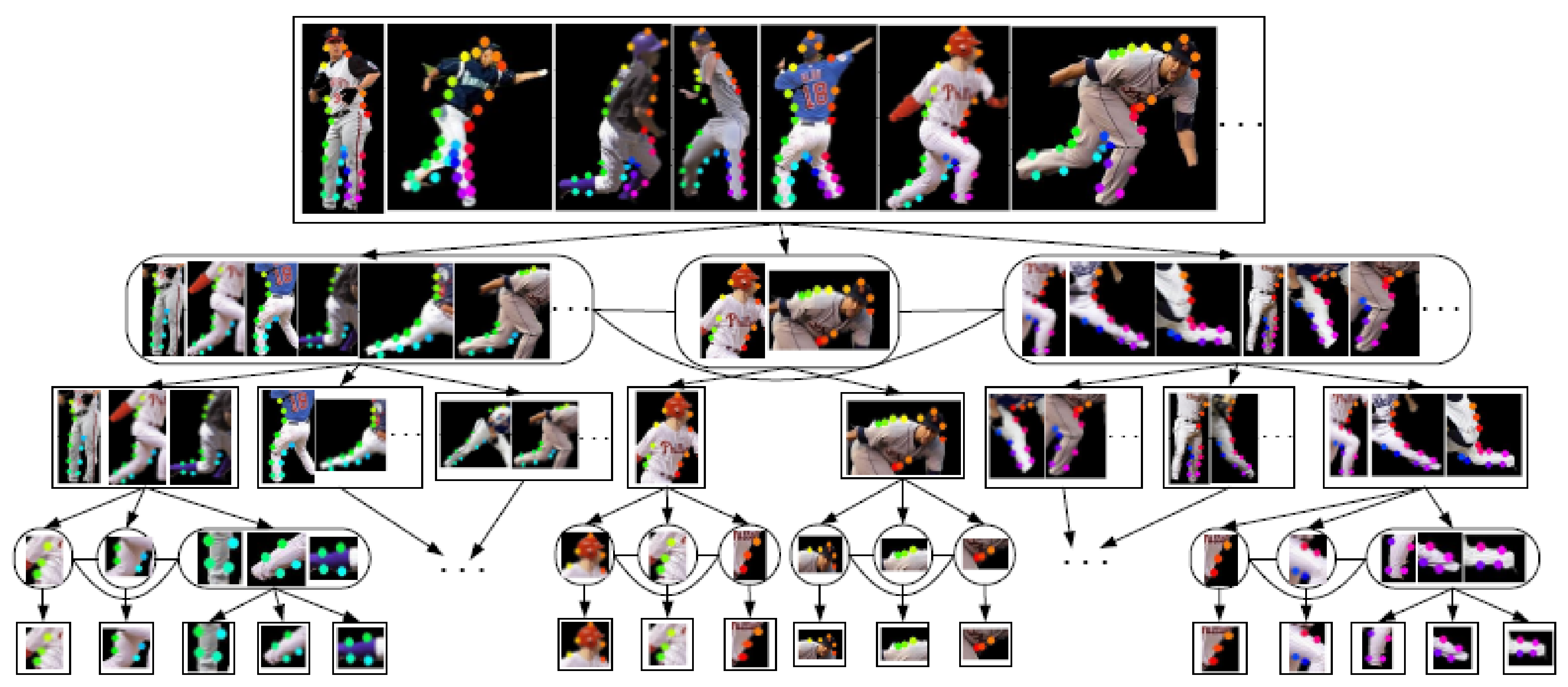


AND/OR Graph Learning

- A novel AND/OR graph is proposed to model enormous poses.
- Learning is performed in a supervised manner.
- Applications: Human Body Parsing

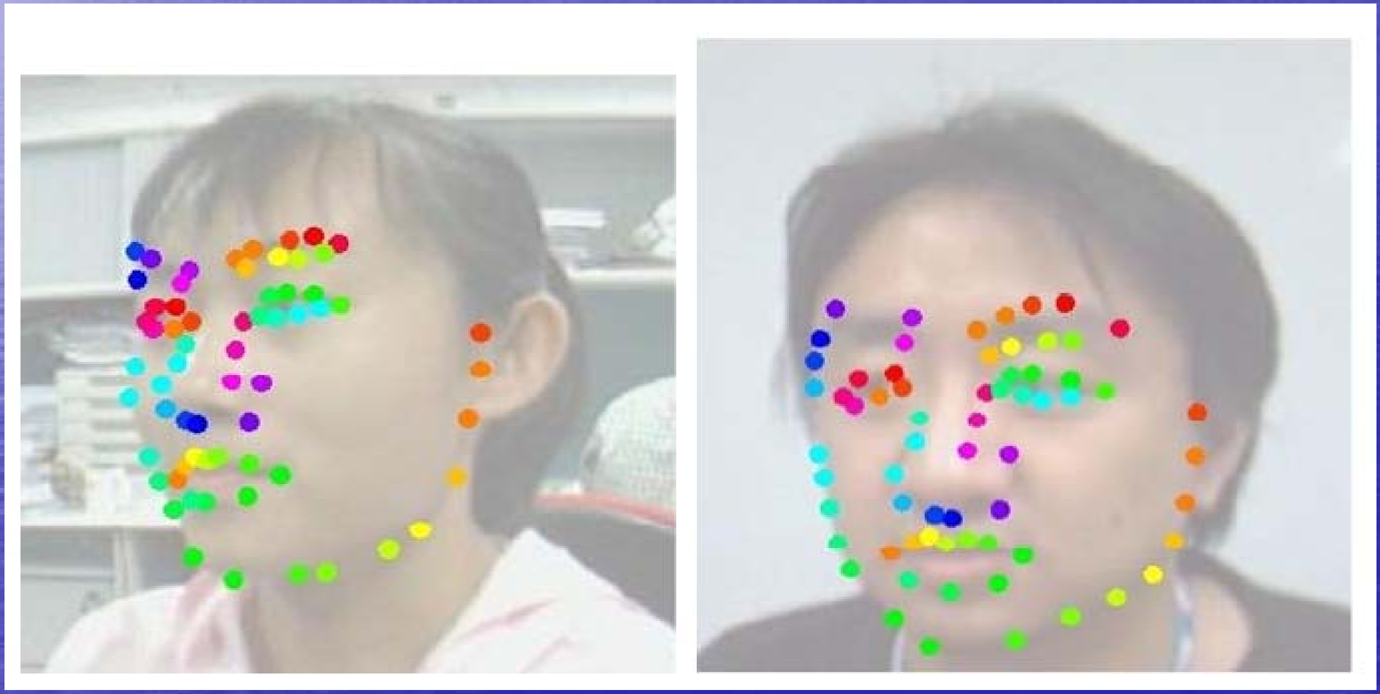


AND/OR Graph for the Human Body



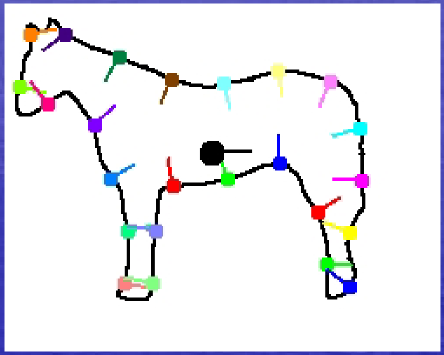
Human Body Parsing

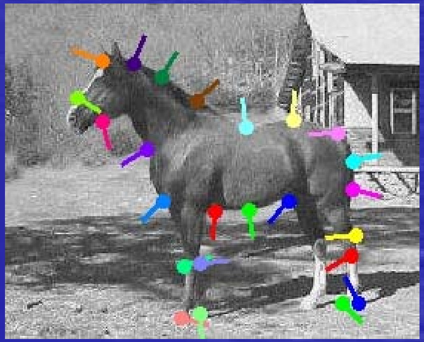
Multi-view Face Alignment



Deformable Object Modeling, Inference and Unsupervised Learning

- Task: Deformable Object Parsing
- Difficulties
  - Large shape and appearance variations.
  - Cluttered Background
  - Occlusion, lighting, etc.





Hierarchical Composition Model

- Formulation: $$P(z,d;w)=\frac{1}{Z}\exp\left\{-\underbrace{E_L(z,d)}_{Pixel-Level}-\underbrace{E_S(z)}_{Multi-Level-Shape}-\underbrace{E_V(z)}_{Vertical}\right\}$$
- Image: d States: $z_v=(x_v,y_v,s_v,\theta_v)$
- Parameters: w
- Image Features is defined between the leaf nodes and image pixel.

$$E_L(z,d)=\langle w_L, f_L(z,d)\rangle$$

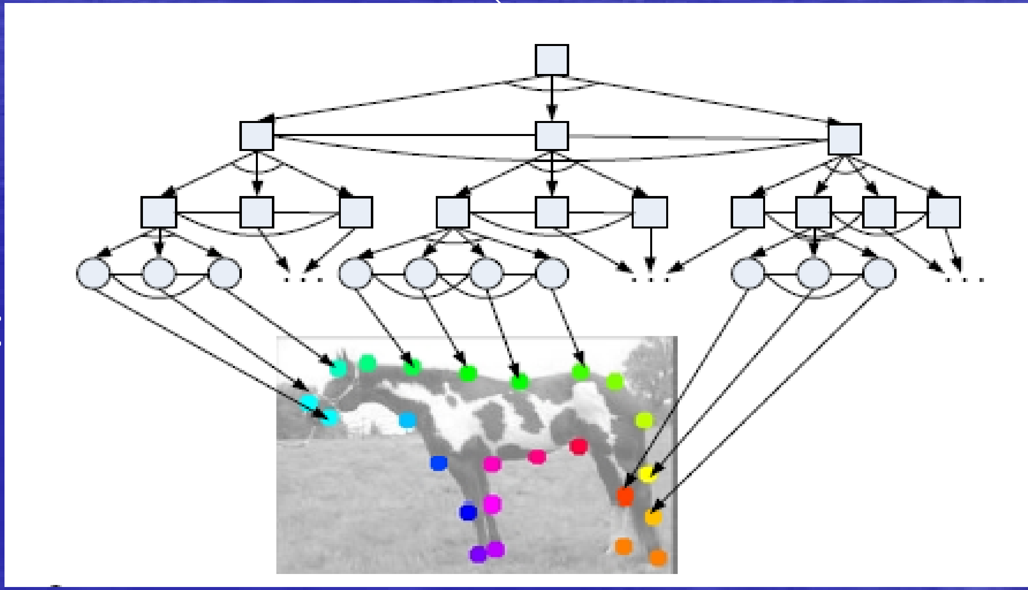
- Horizontal Shape Priors at multiple levels

$$E_S(z)=\langle w_S, g(z_\mu,z_\rho,z_\gamma)\rangle$$

- Vertical constraints

$$E_V(z)=\langle w_V, h(z_v,z_\mu,z_\rho,z_\gamma)\rangle$$

$z_v=(x_v,y_v,s_v,\theta_v)$

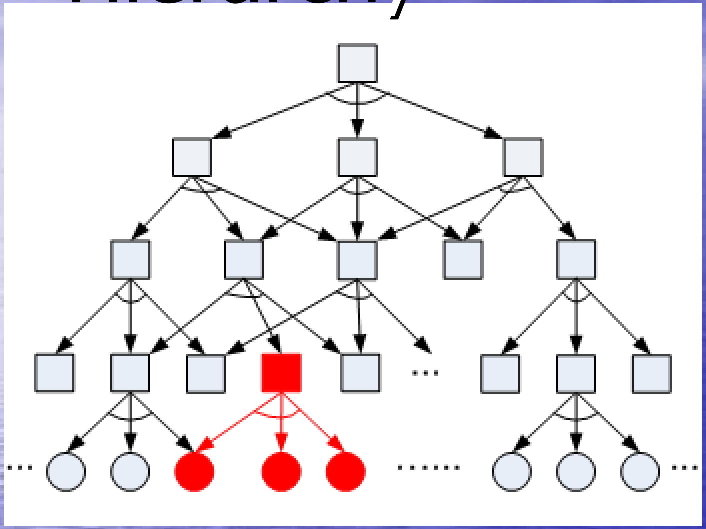


Bottom-Up Inference

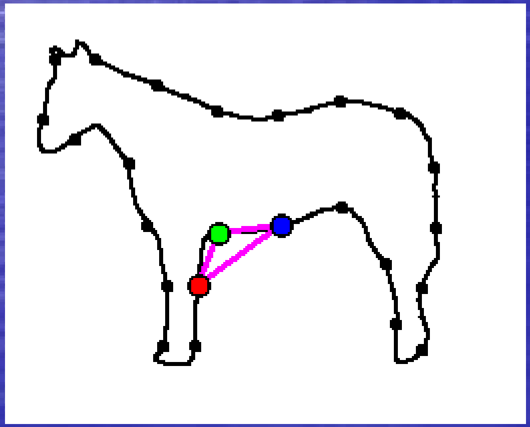
- From the Bottom Level to the Top Level
  1. Composition
  2. Pruning
  3. Surround Suppression
- Complexity: empirically linear in the size of image and ranges of scale and orientation.

Bottom-up Inference

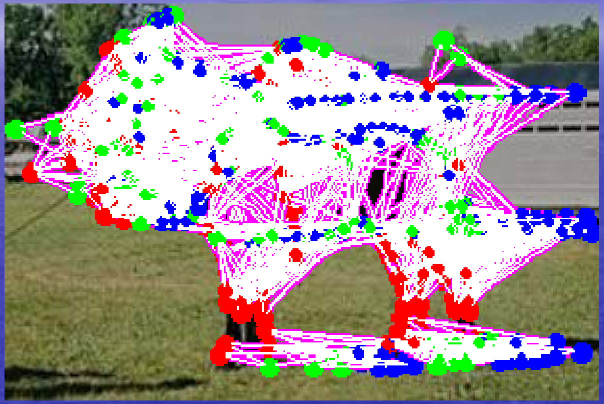
- Hierarchy



- Current Model



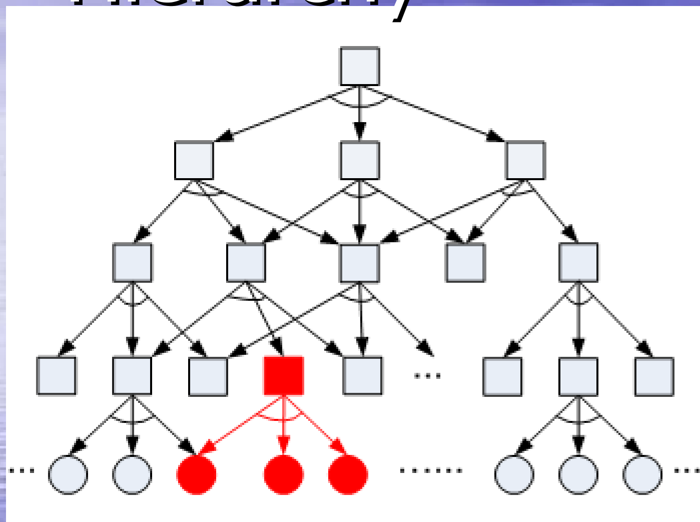
- Parsed Instances



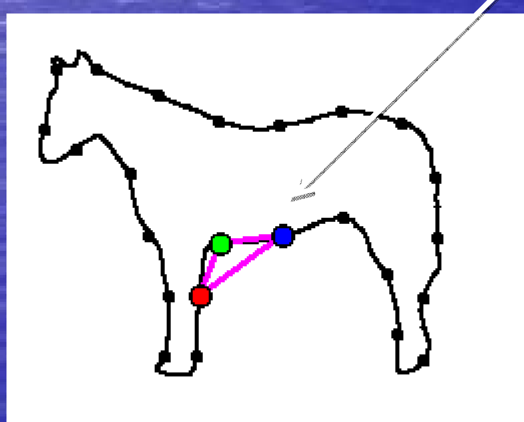
↑ **Step 1: Composition**

Bottom-up Inference

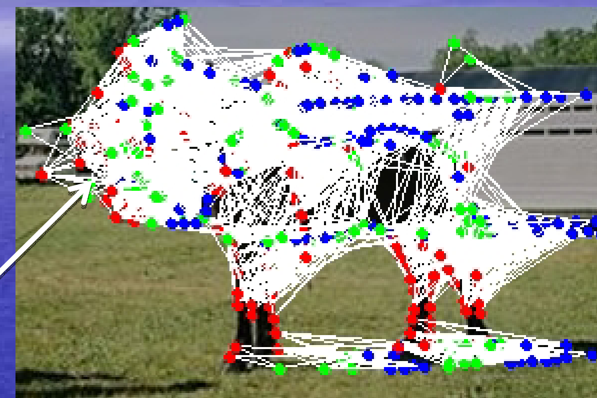
- Hierarchy



- Current Model



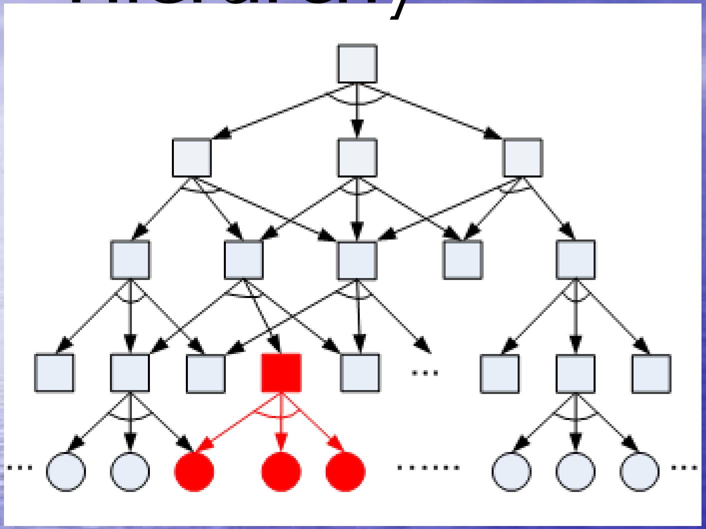
- Parsed Instances



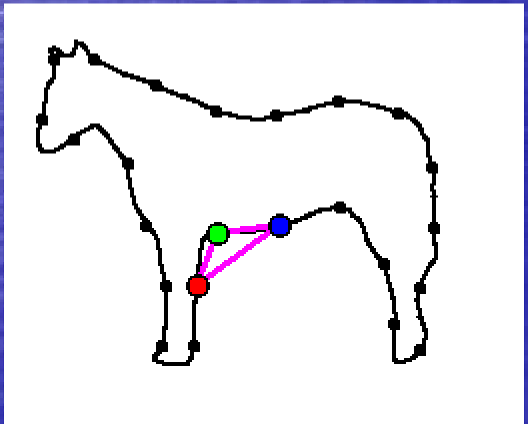
Step 2:
Pruning

Bottom-up Inference

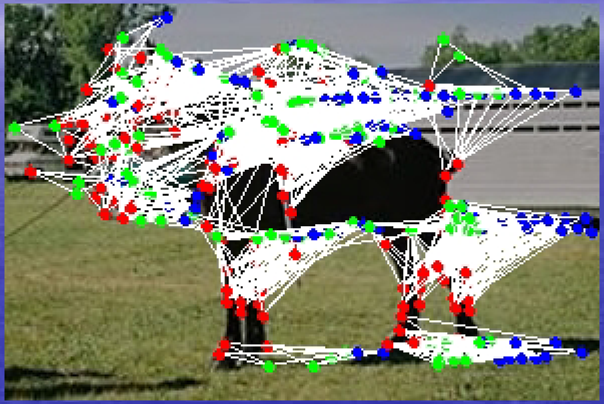
- Hierarchy



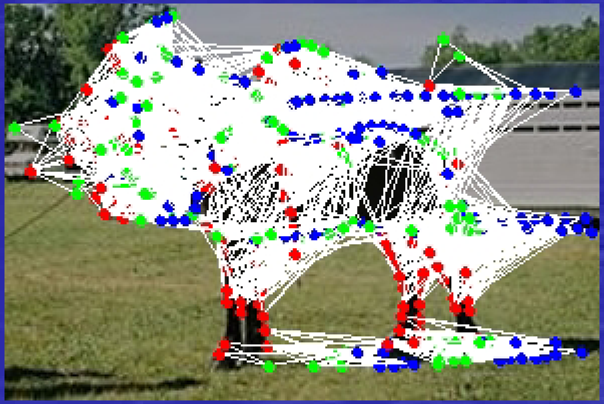
- Current Model

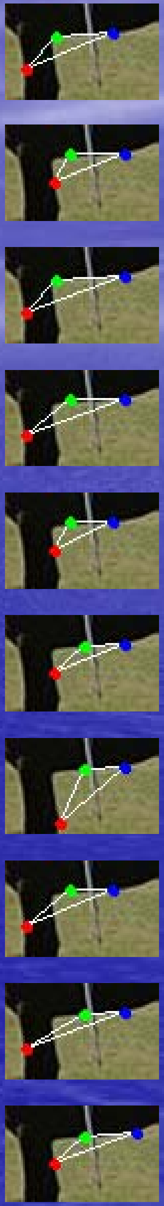


- Parsed Instances



Step 3: Surround Suppression

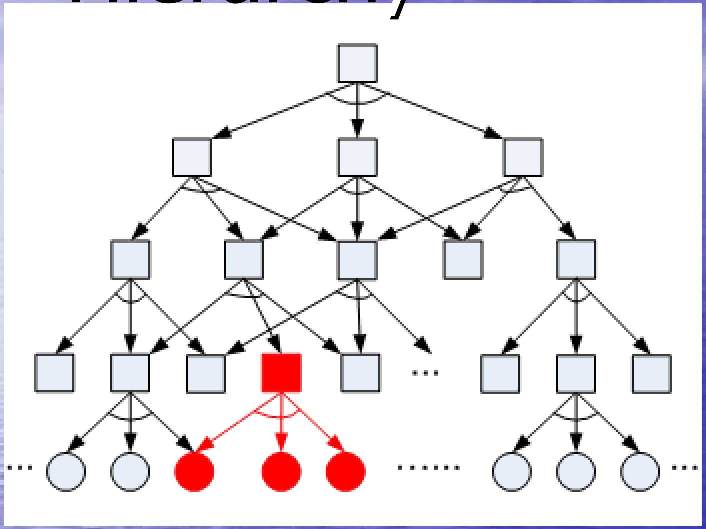




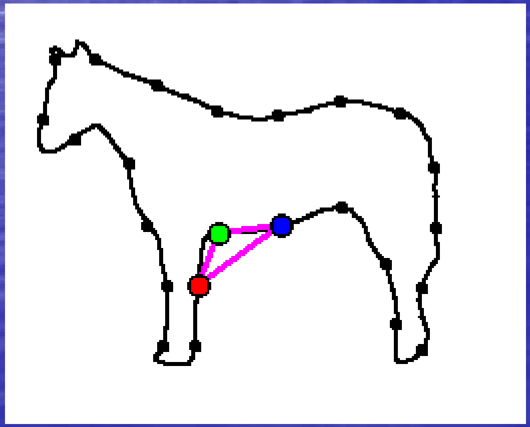
……

Bottom-up Inference

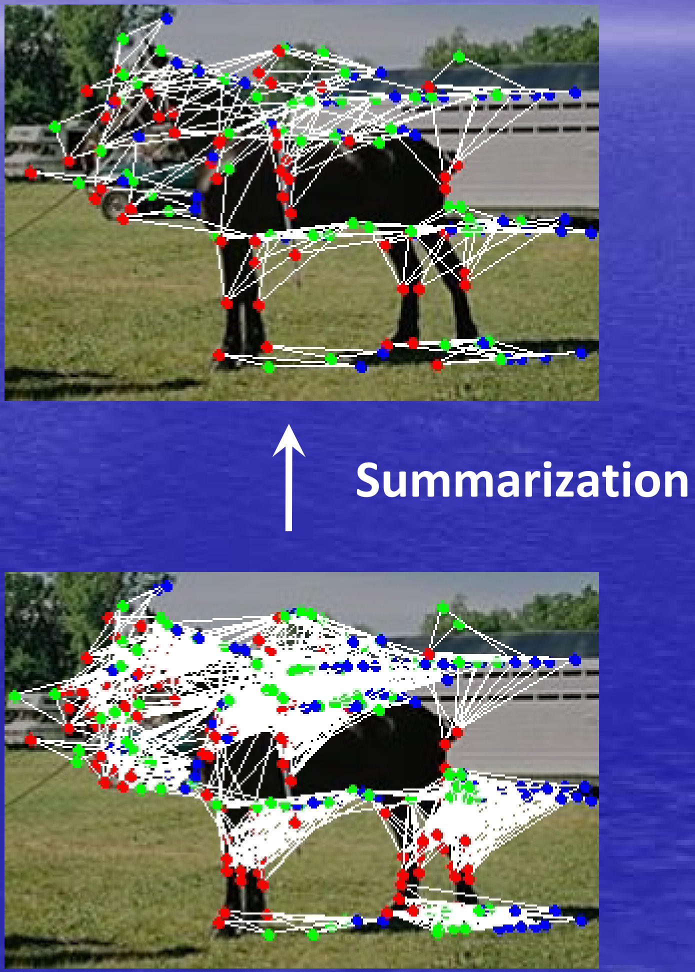
- Hierarchy



- Current Model

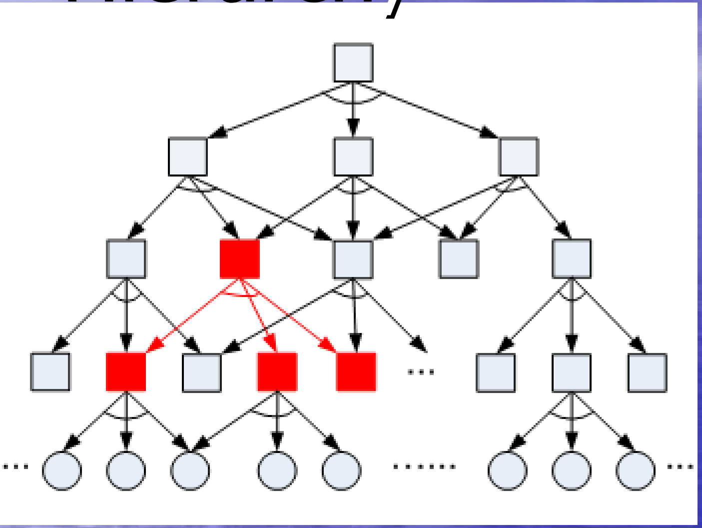


- Parsed Instances

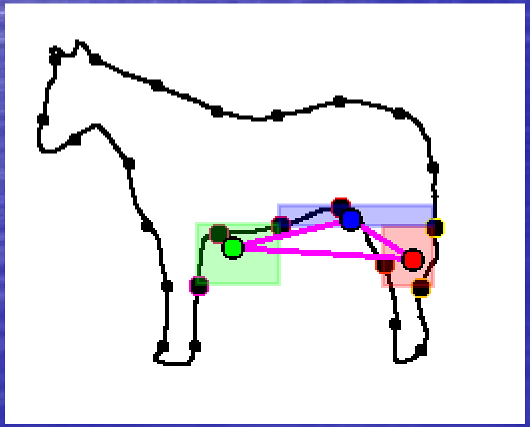


Bottom-up Inference

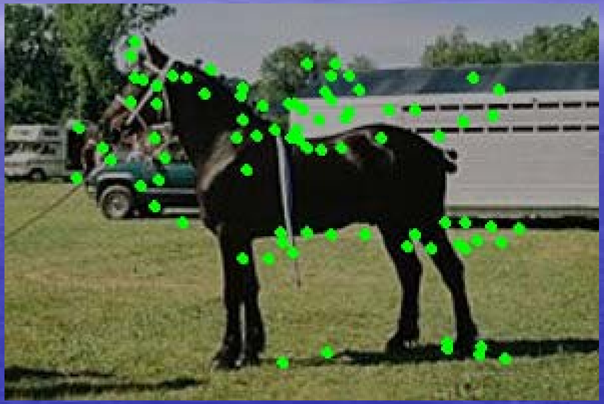
- Hierarchy

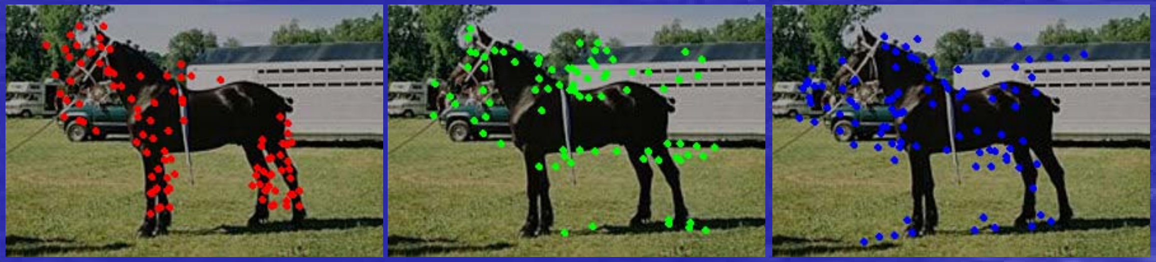


- Current Model



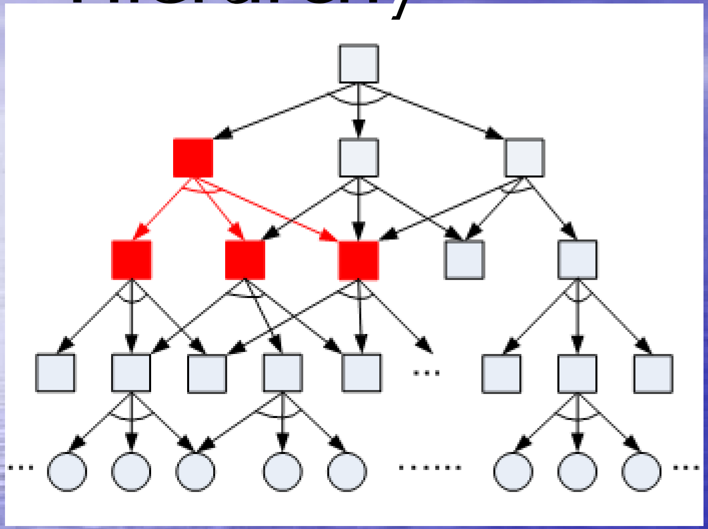
- Parsed Instances



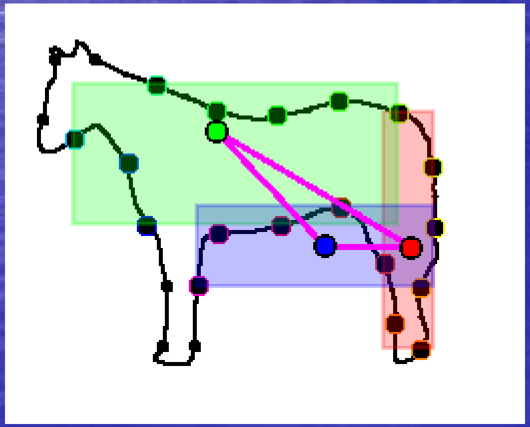


Bottom-up Inference

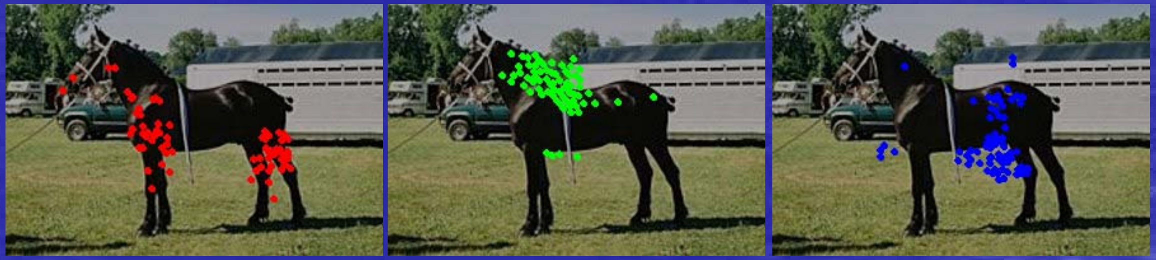
- Hierarchy



- Current Model

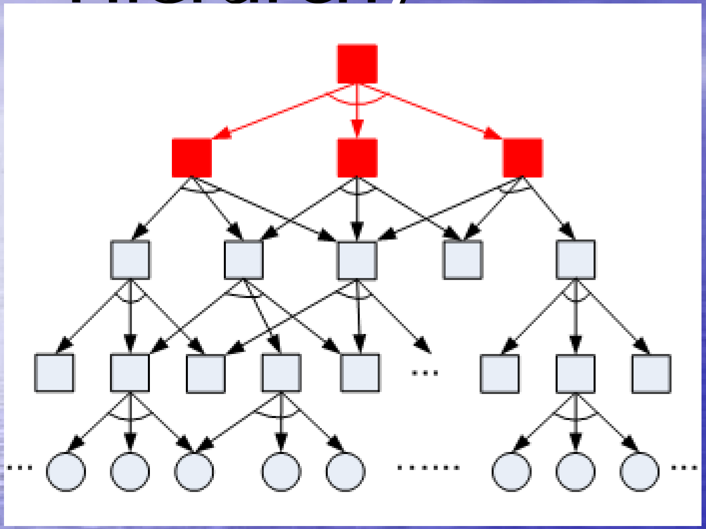


- Parsed Instances

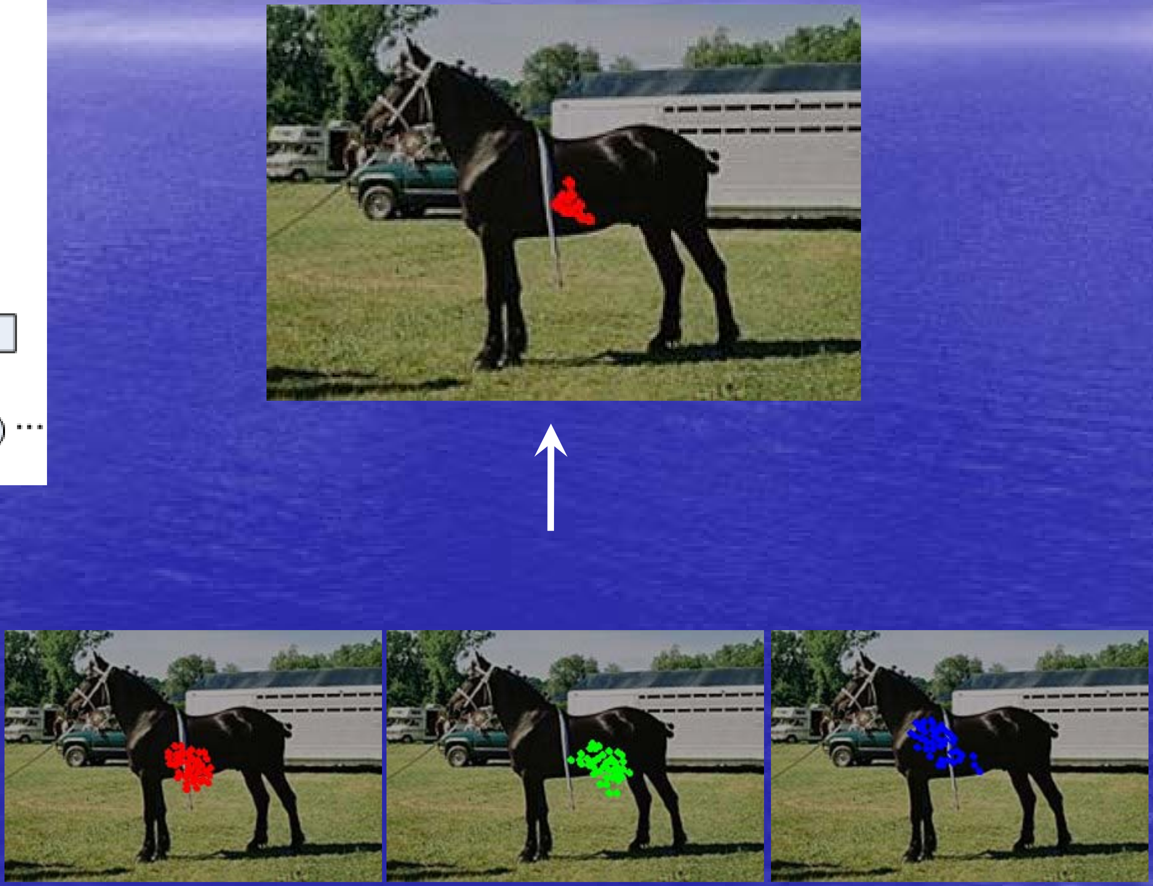


Bottom-up Inference

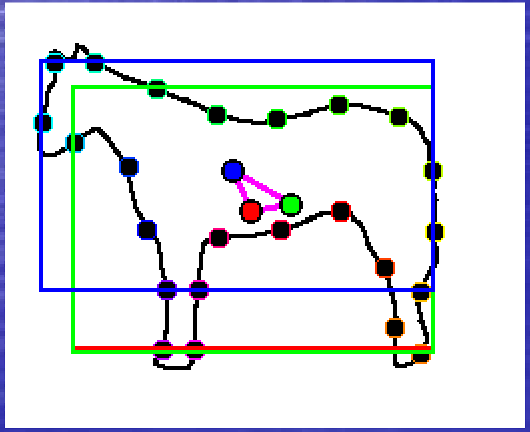
- Hierarchy



- Parsed Instances

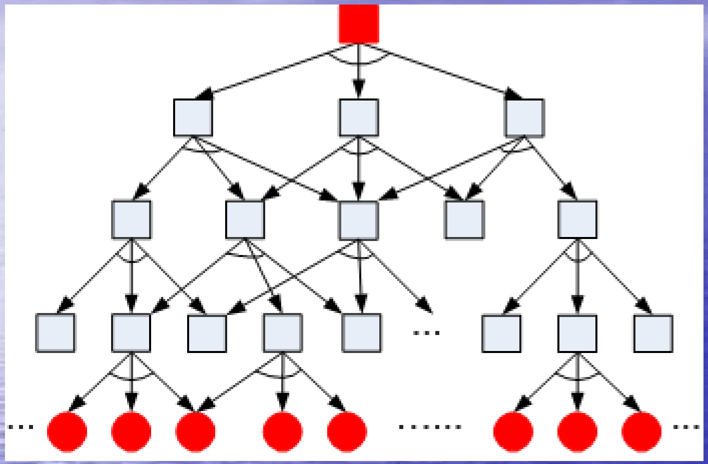


- Current Model



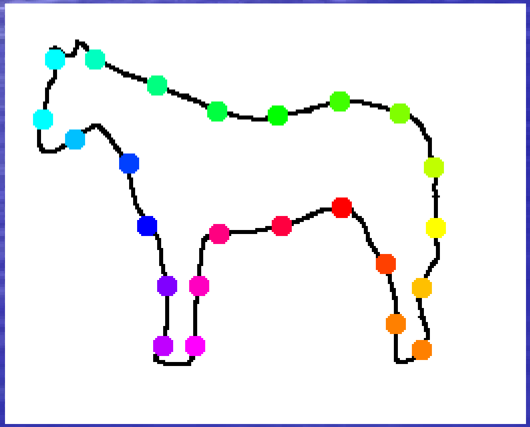
Bottom-up Inference

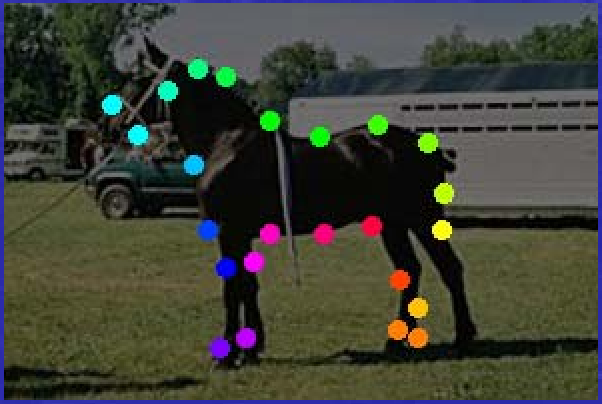
- Hierarchy

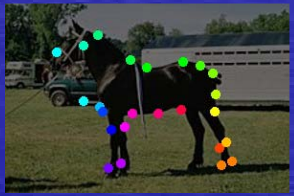


- Parsed Instances

- Current Model



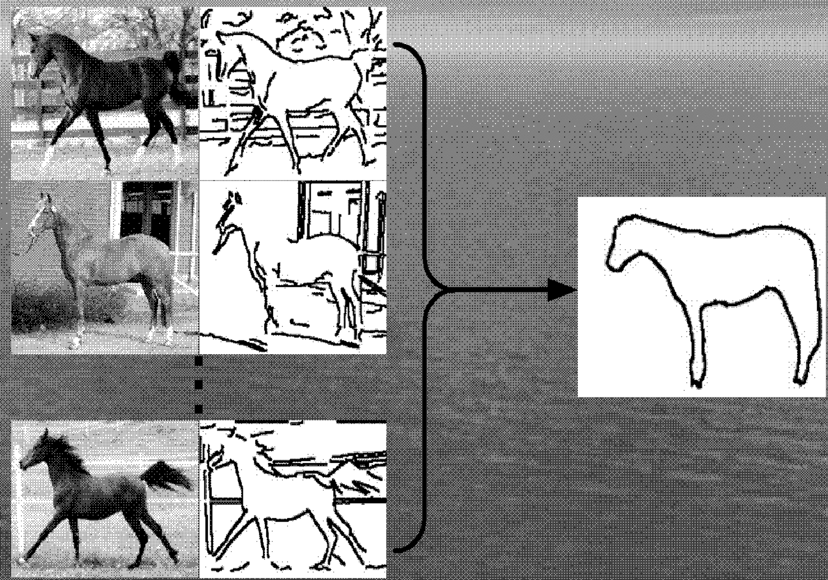


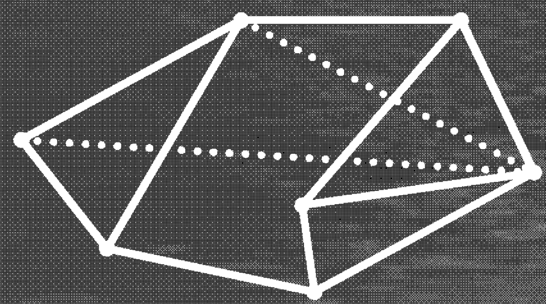


......

Unsupervised Structure Learning

- Unsupervised Structure Learning: induce a structure (vertexes and edges) and estimate its parameters
- Combinatorial Explosion Problem
  - M=150 object features
  - N =5000 total features
  - Big Ambiguity: Edgelets
  - Naïve brute force enumerations: O(M^N)
  - Greedy method:
    - Sparseness: M=6, N=100
    - Low ambiguity: more powerful features





Unsupervised Structure Learning

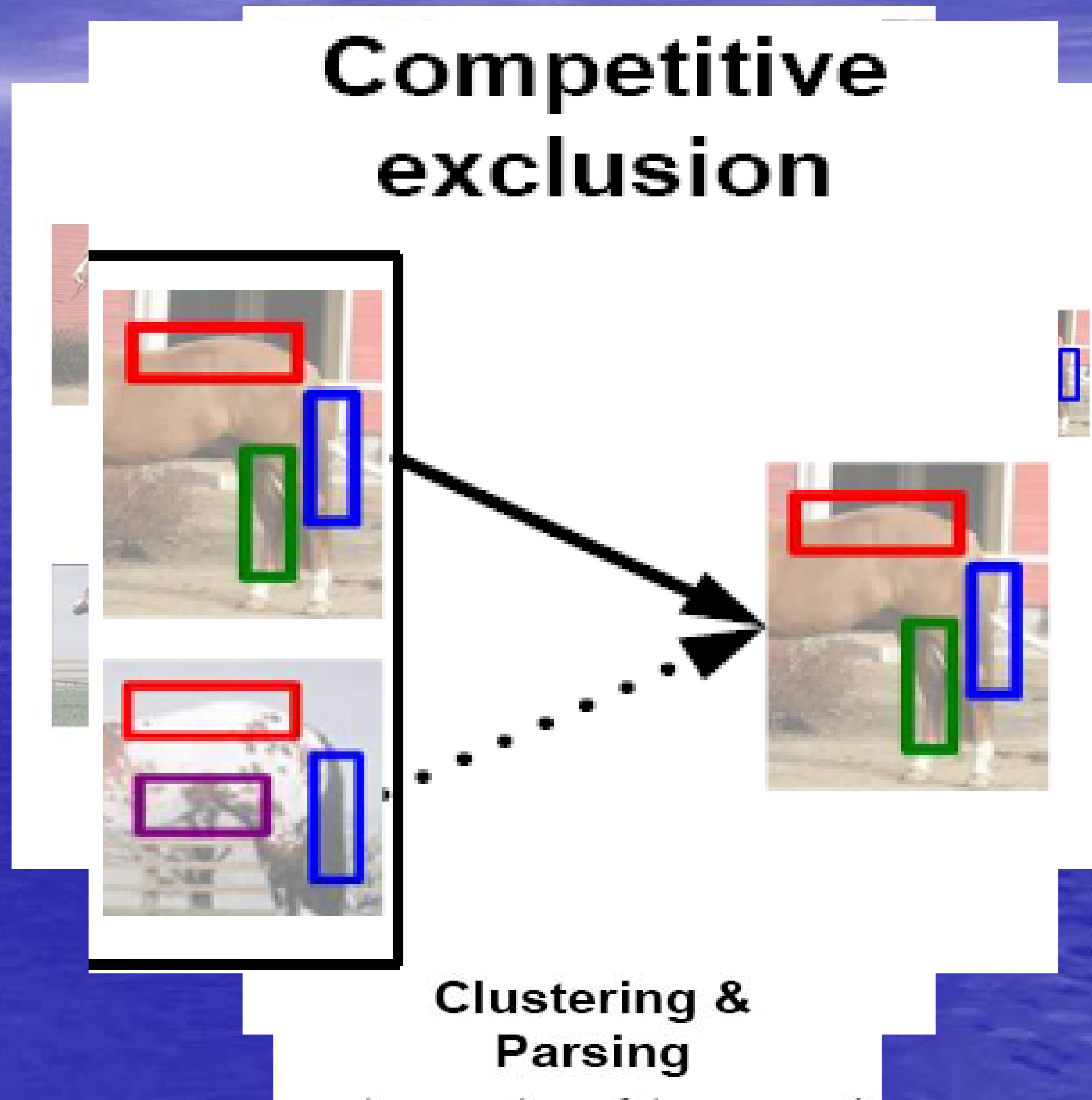
- Procedure: Bottom-Up and Top-Down
- Three principles:
  - Hierarchical Composition: combine elementary structures (danger combinatorial explosion)
  - Suspicious Coincidence
  - Competitive Exclusion
- Complexity: linear in the height of a hierarchy (empirically)

Bottom-Up Learning

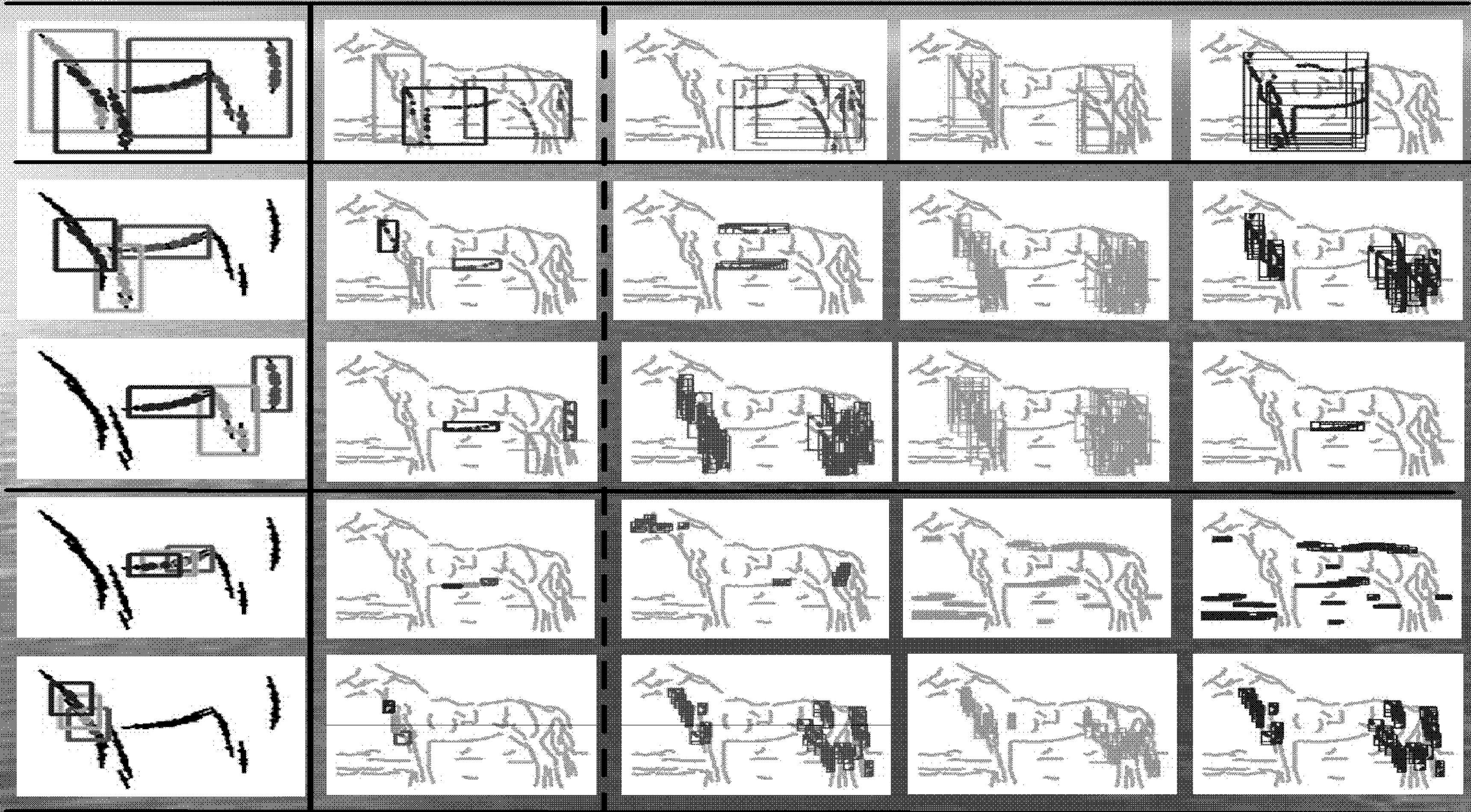
Repeat from low levels to high levels

1. Composition: combine instances from level L
2. Clustering: compose concepts at level L+1
3. Parsing: get responses of concepts
4. Pruning: prune out non-frequent concepts
5. Competitive Exclusion: prune out the similar concepts

Until no new compositions are formed (The number of layers is automatically decided by the algorithm)

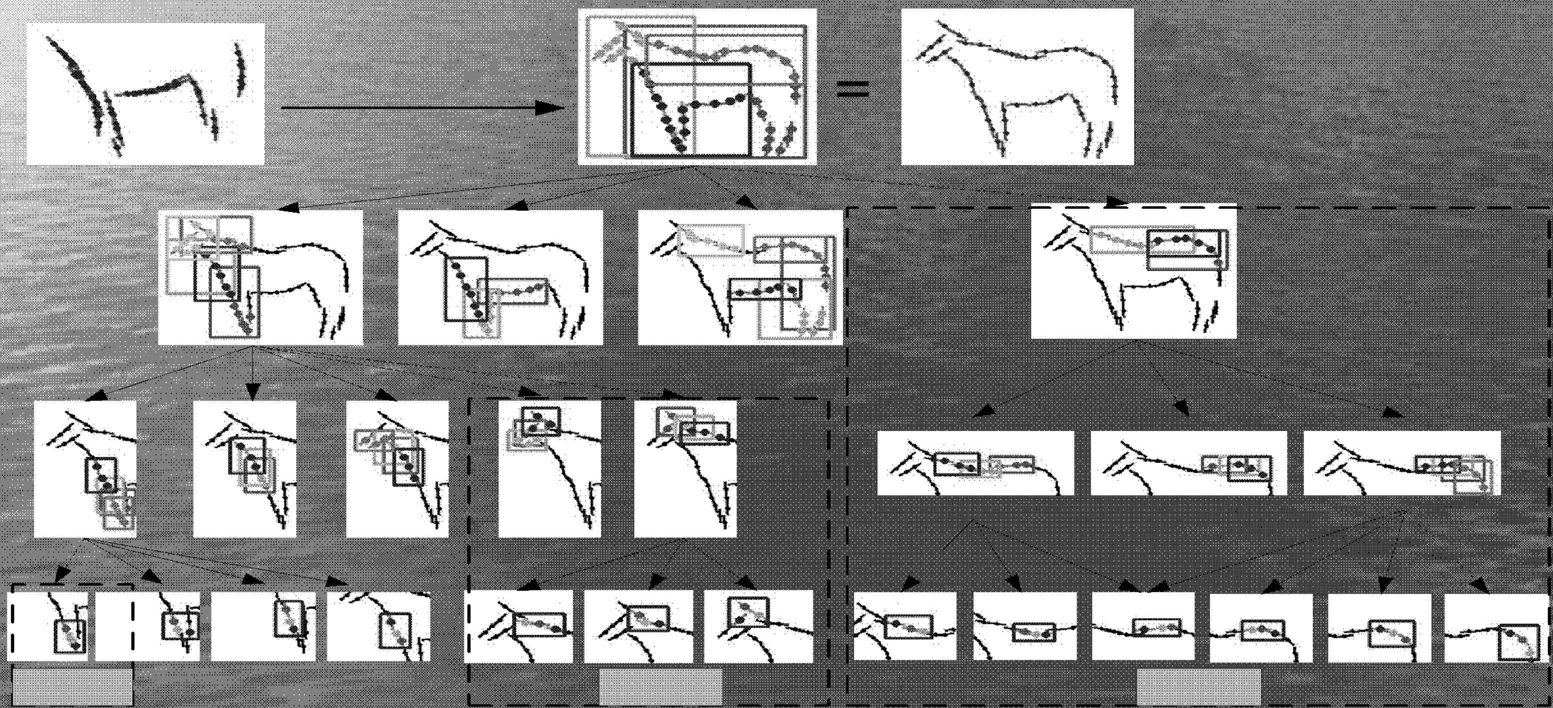


Concepts (structures) learnt by the bottom-up procedure



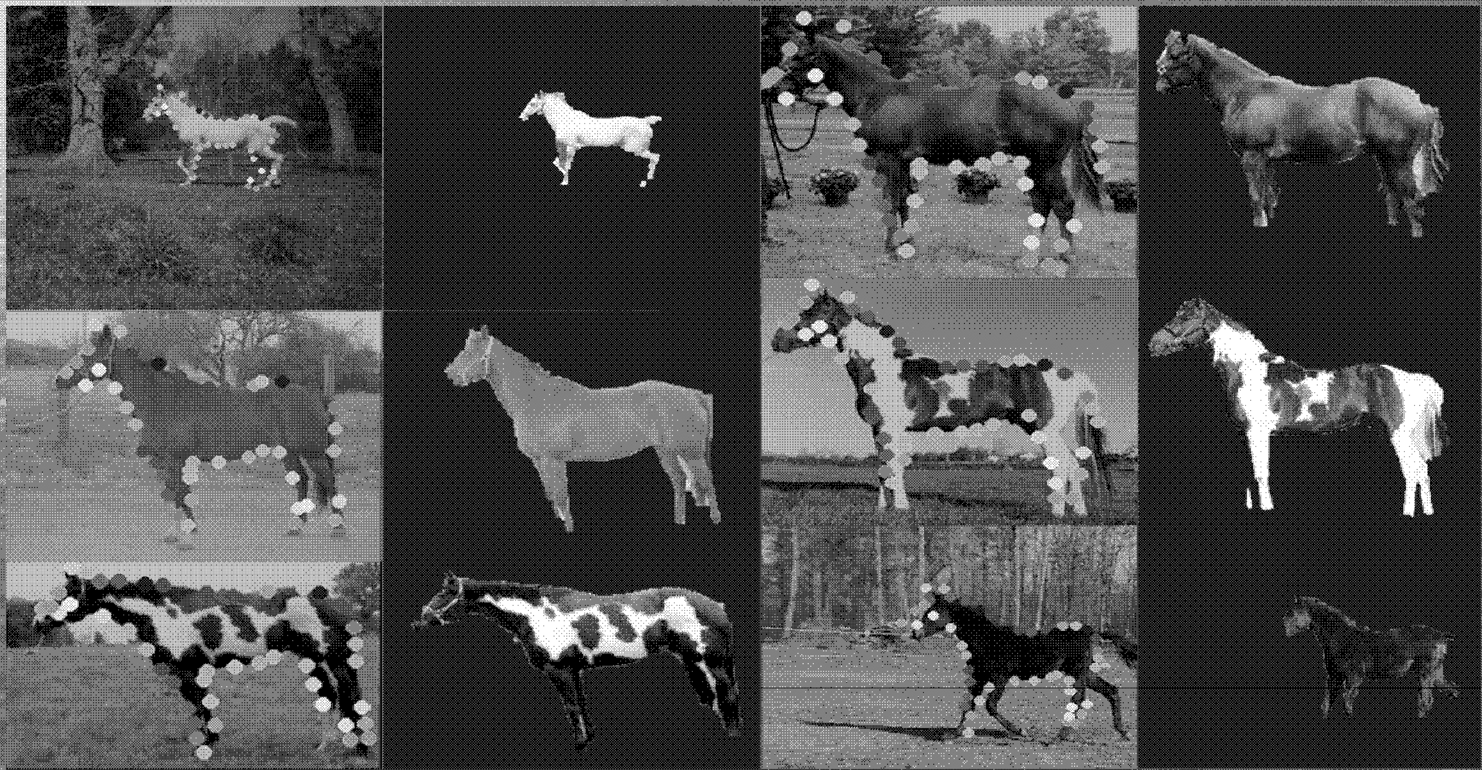
Top-Down Learning

- Fill in the missing parts caused by 1) competitive exclusion 2) pruning.
- Examine every node of the hierarchy.



Experiments: Deformable Object Segmentation and Parsing

- Weizmann Horse Dataset: 12 training images (no labeling) and 316 testing images (with ground truth)



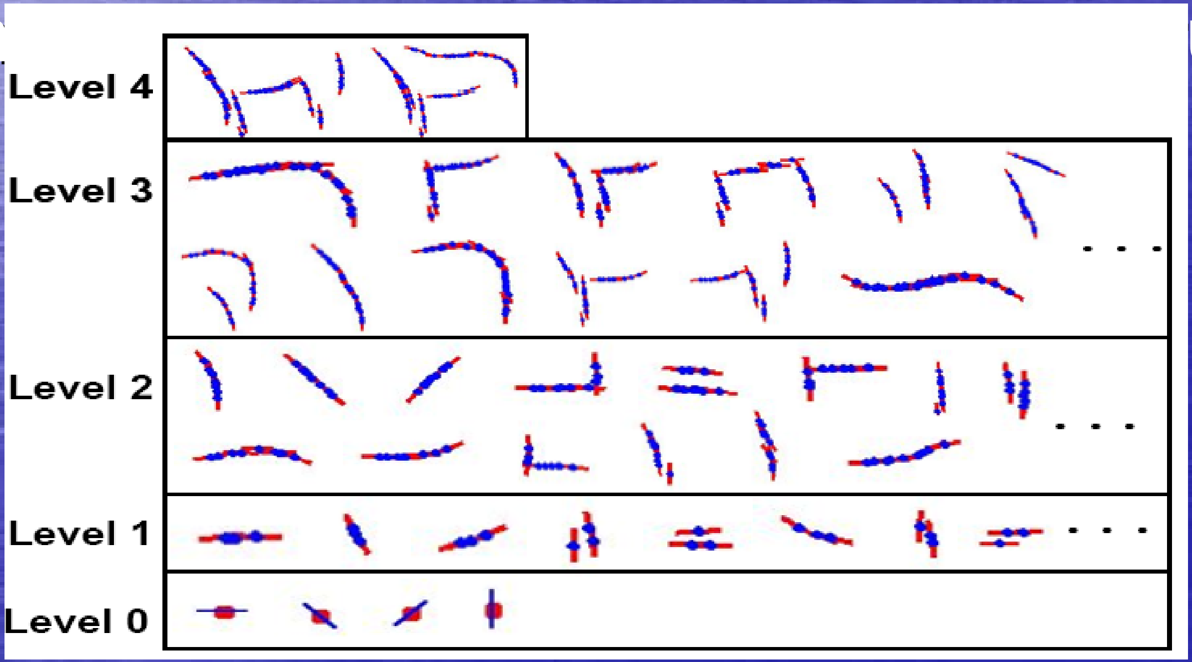
Comparisons

- Comparable to supervised learning methods

Method	Train	Test	Segmentation	Speed
Our method	12	316	93.3	16.9s
Ren [11]	172	172	91.0	–
Borenstein [21]	64	328	93.0	–
LOCUS [22]	20	200	93.1	–
OBJ CUT [23]	N/A	5	96.0	–
Levin [12]	N/A	N/A	95.0	–

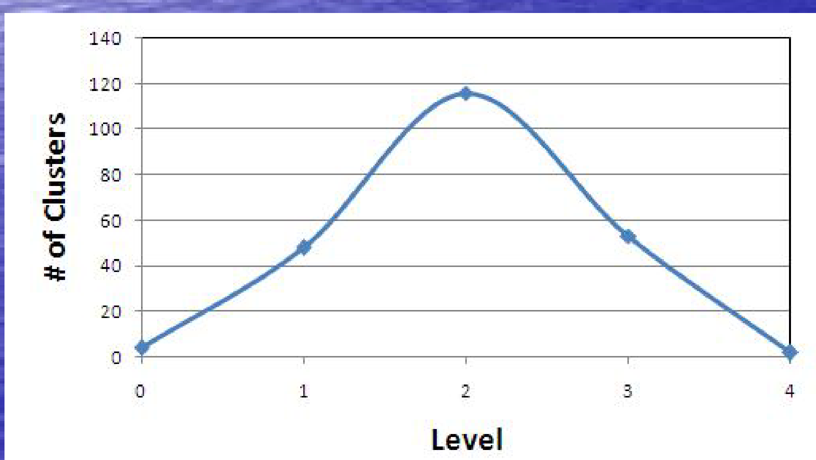
Analysis I: From Generic Feature to Object Structure

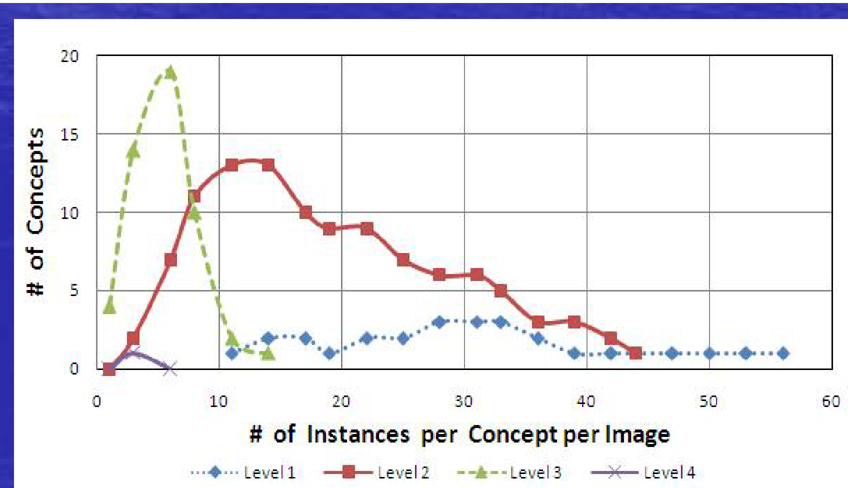
- Unified descriptors
- Unified learning: bridge the gap between the gen… …uctures



Analysis II: Multi-Level Computational Complexity

L	Composit.	Clusters	Prune	Com. Exe.	Time
0				4	1s
1	167431	14684	262	48	117s
2	2034851	741662	995	116	254s
3	2135467	1012777	305	53	99s
4	236955	72620	30	2	9s

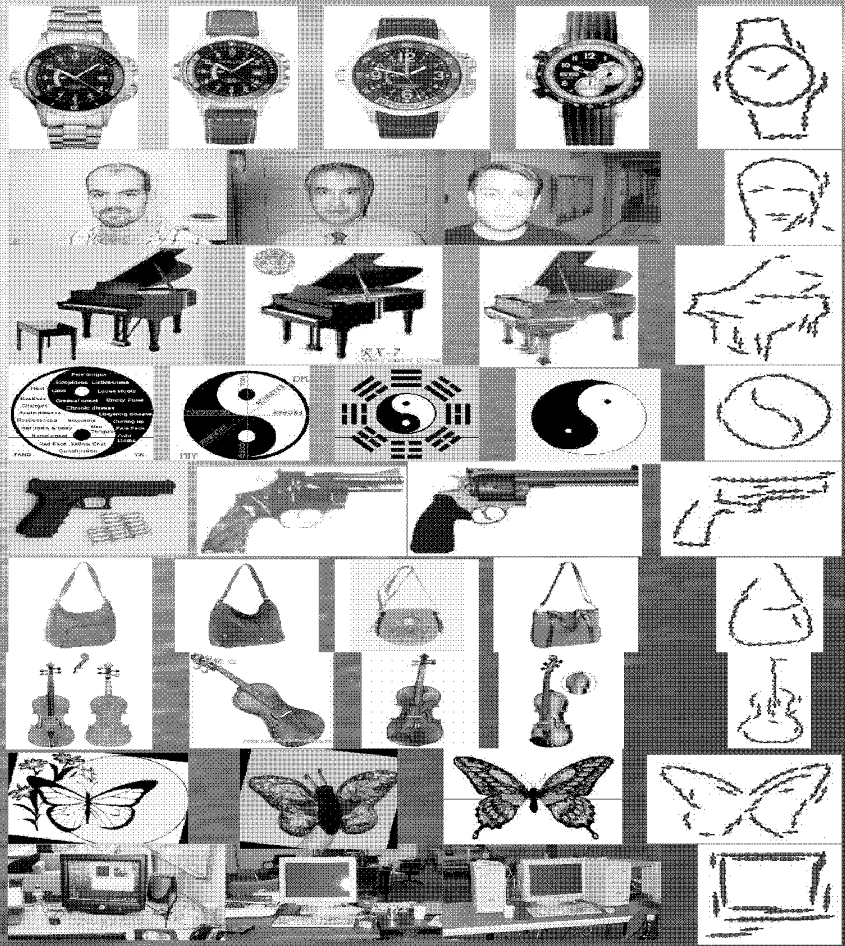


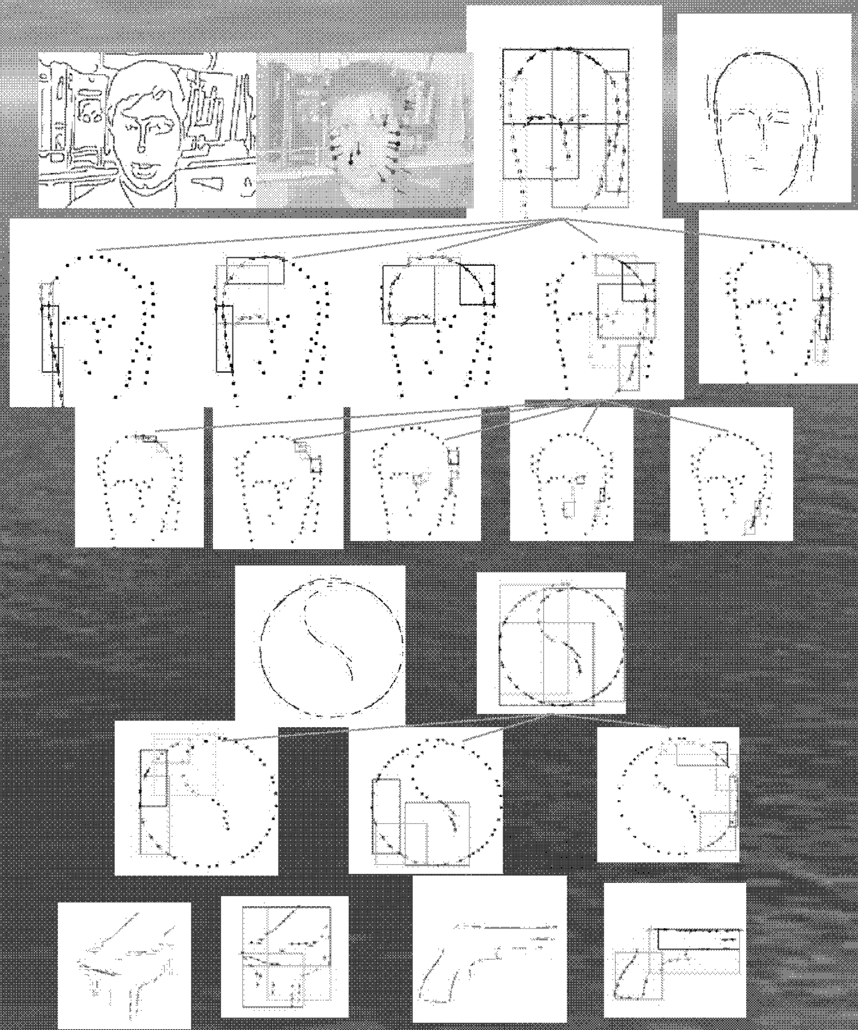


Feasibility of scaling up

- Short-term goal: 100 objects and 1000 images
- CPU and memory costs:
  - 10 images: 5 minutes , 320 Megabytes
  - 20 images: 10 minutes , 550 Megabytes
  - 50 images: 60 minutes, 1900 Megabytes
  - 1000 images: 2 days, 40 gigabytes (Prediction)

Analysis III: More Objects





Summary

- Hierarchical Composition Model
- Rapid Inference/Parsing
- Rapid Unsupervised Structure Learning