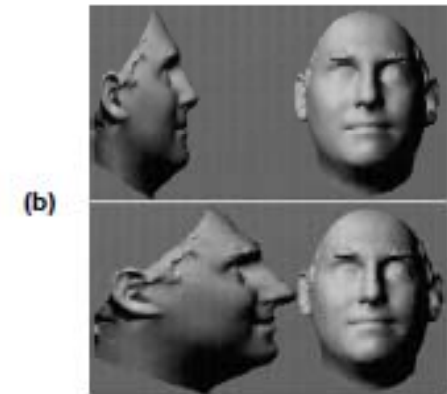
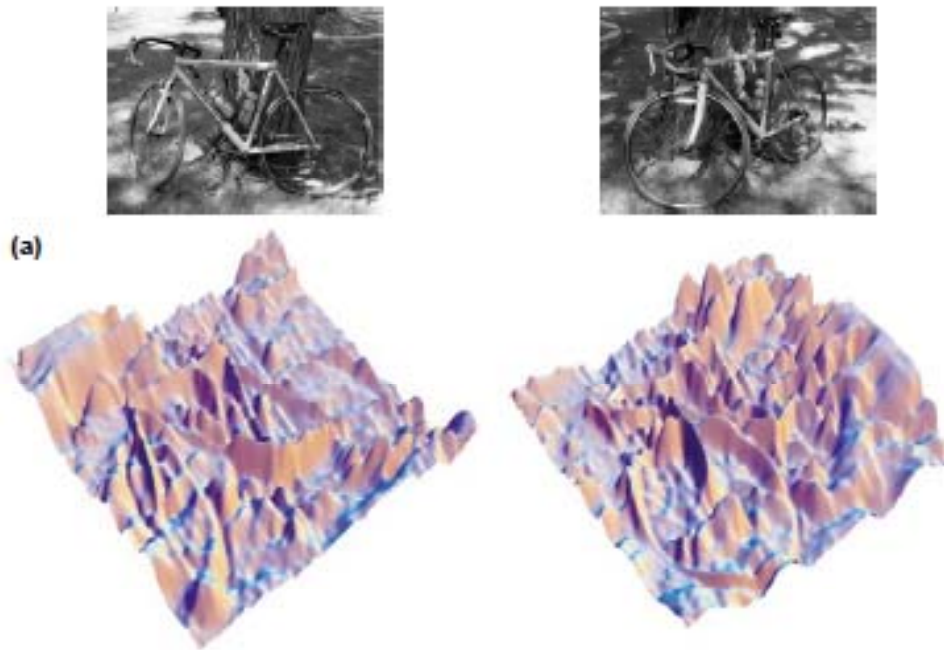


Vision: Overview

Alan Yuille

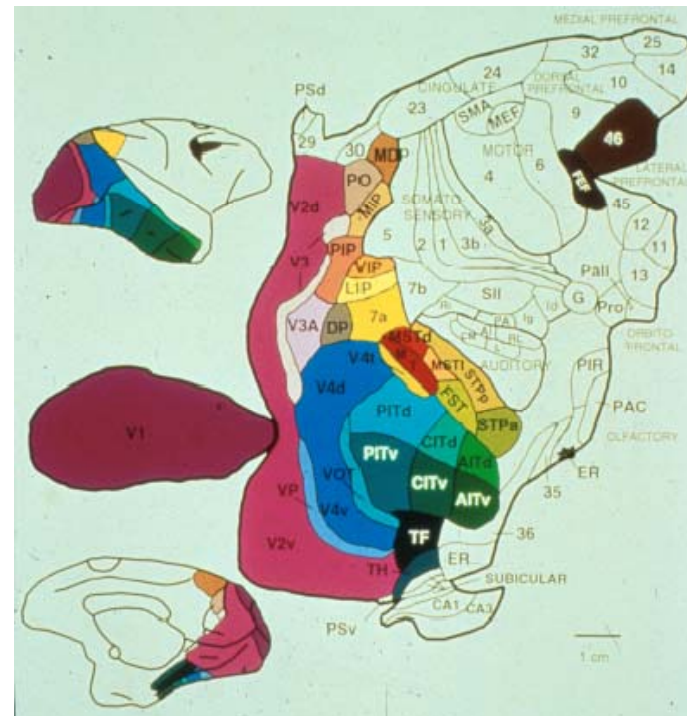
Why is Vision Hard?

- Complexity and Ambiguity of Images. Range of Vision Tasks.
- More 10×10 images -- $256^{100} = 6.7 \times 10^{240}$ -- than the total number of images seen by all humans throughout history 3×10^{21} .
- (50 billion people, live 20 billion seconds, 30 image per second)



Why does Vision seem easy?

- Because we devote roughly half our cortex to vision.
- Understanding vision means understanding half the cortex.



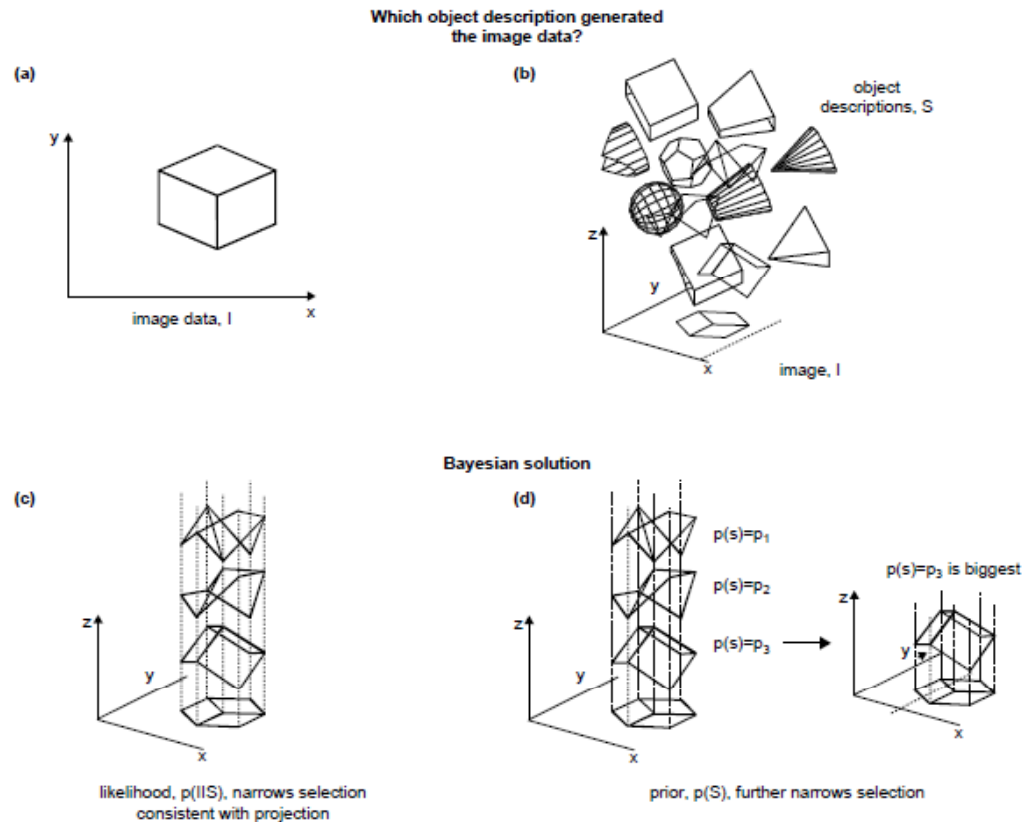
Bayes and Vision.



- History of Bayes and Vision dates to the early 1980's and before. (Ulf Grenander's pattern theory, 1960's).
- Vision as an inverse inference problem.
- Decode images by inverting image formation.
- As argued by Gibson and Marr, this requires knowledge about the world Natural constraints (Marr), Ecological constraints (Gibson).
- Bayesian formulations are natural. Constraints are priors and can be learnt from examples.

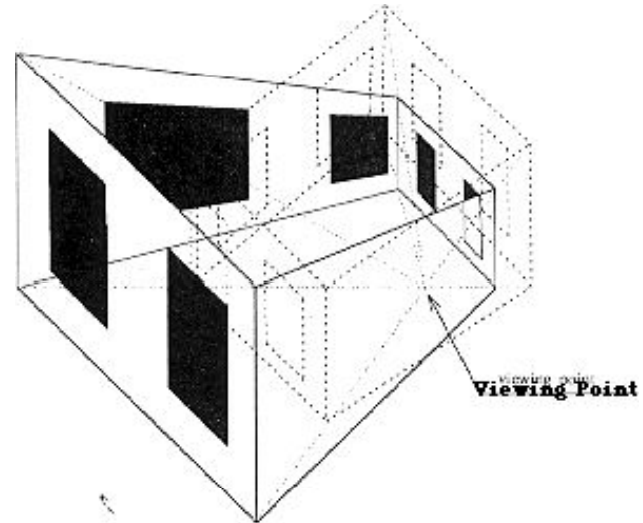
Bayes for Vision

- Courtesy of Pavan Sinha (MIT)
- The likelihood is not enough.



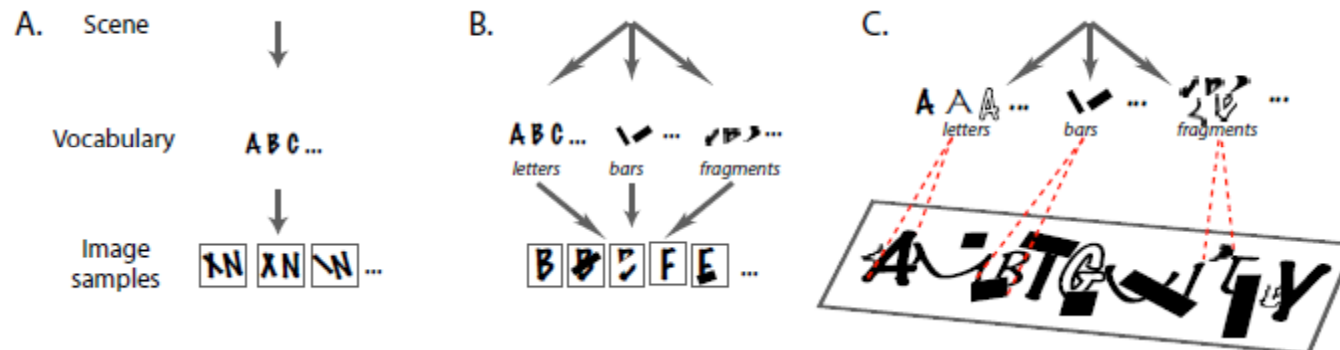
When priors are violated

- Who is bigger?



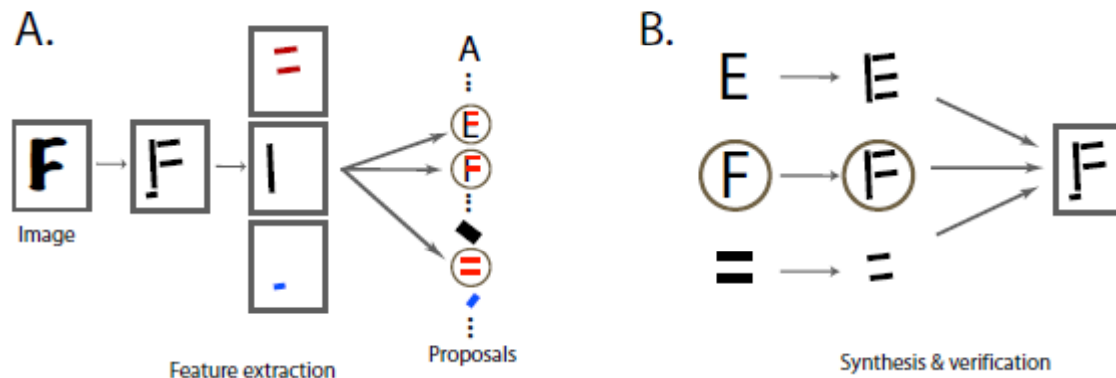
Models for Generating Images:

- Grammars (Grenander, Fu, Mjolsness, Biederman).
- Simple to Complex Grammars: Easy to hard Inference



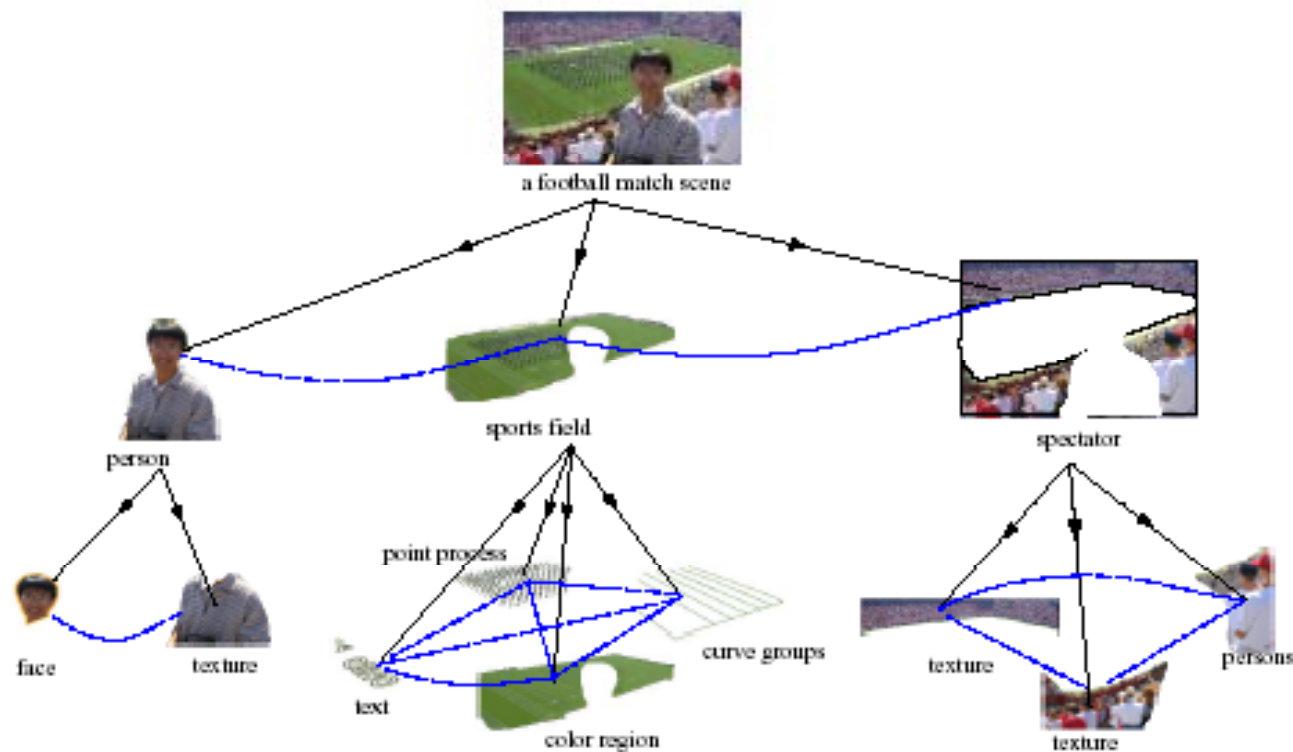
Analysis by Synthesis

- Analyze an image by inverting image formation.
- Proposals and Verification



Can we do this for Real Images?

- Image Parsing:
- Learn probabilistic models of the visual patterns that can appear in images.
- Interpret/understand an image by decomposing it into its constituent parts.



Vision Goals and Tasks

- Vision is often formalized as low, middle, and high-level.
- This seems to map onto different parts of the visual cortex (V1, V2,..., IT). (Poggio's Talk).
- High level vision relates very naturally to other aspects of cognition – reasoning, language.

Some Vision Goals (SC Zhu et al)

- Understanding objects, scenes, and events.
Reasoning about functions and roles of objects, goals and intentions of agents, predicting the outcomes of events.

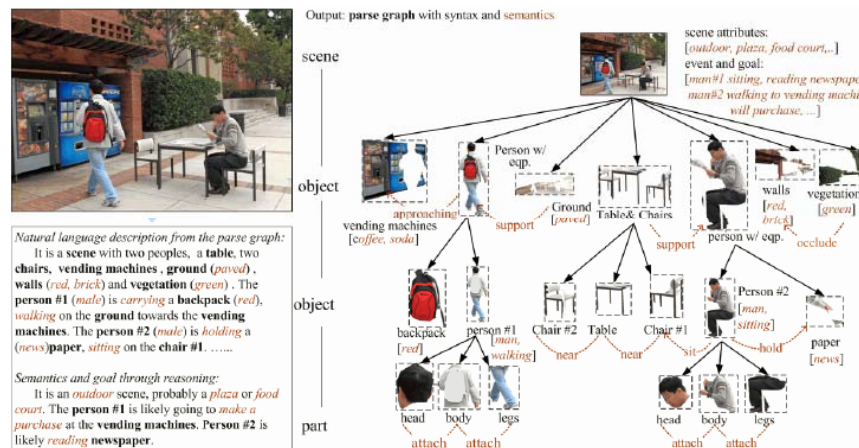


Figure 1. Example of image understanding. Analysis of the image (top-left) produces a parse graph (right) representing hierarchically objects, contextual relations, and semantic associations (in italic orange font) for attributes, functions, roles, and intents. The parse graph maybe converted to a description in natural language (bottom-left).

Converting Parse Graphs to Language

- Illustration: Perona and Fei-Fei Li.


Image shown to subjects	40ms	80ms	107ms	500ms
	“Possibly outdoor scene, maybe a farm. I could not tell for sure.”	“There seem to be two people in the center of the scene.”	“ People playing rugby. Two persons in close contact, wrestling, on grass. Another man more distant. Goal in sight.”	“Some kind of game or fight. Two groups of two men. One in the foreground was getting a fist in the face. Outdoors, because I see grass and maybe lines on the grass? That is why I think of a game, rough game though, more like rugby than football because they weren't in pads and helmets...”

Figure 2. Human subjects reporting on what he/she saw in an image shown for different presentation durations (PD=27, 40, 67, 80, 107, 500ms). From Fei-Fei and Perona [26].

Reasoning about Objects in 3D Space

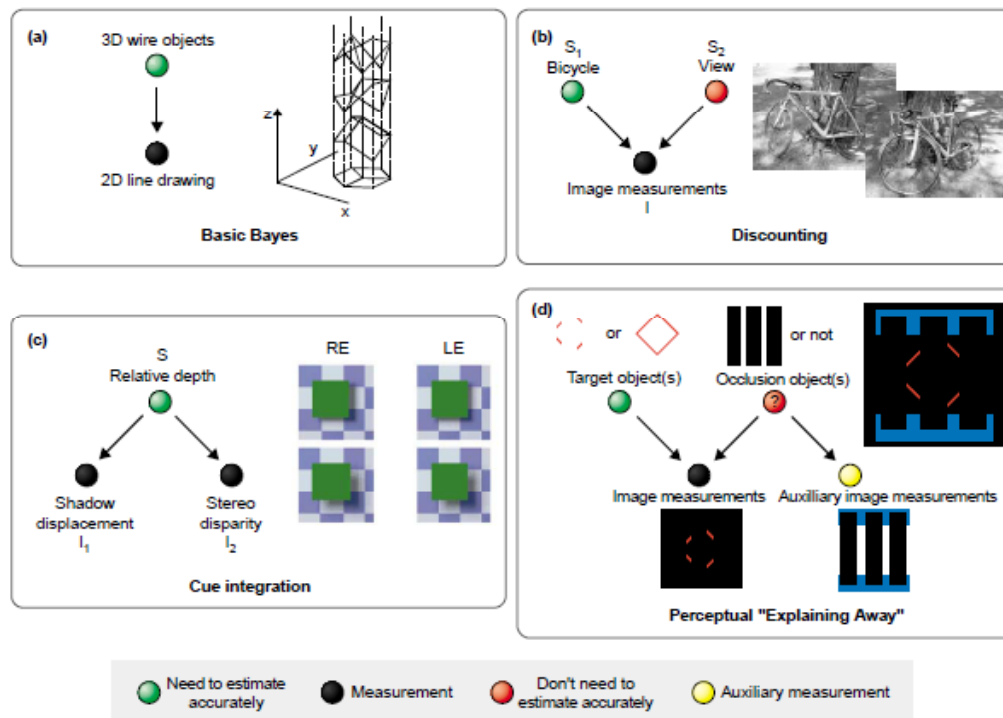
- Understanding the 3D scene structure enables reasoning.



Figure 9. Placing objects in a consistent geometric frame, such as children playing soccer, allows reasoning about objects in 3D space. Results from Koller's group ICCV09 [37]

Graphical Models

- Graphical Models give a nice way to formulate vision problems.
- Different types of Interactions.



Cue Combination

- How to couple different visual cues?
- Uncertainty – statistics relations between different factors that cause the image. What factors are independent? Which are not?
- Studies show (e.g., Larry Maloney) that humans can combine visual, and other perceptual cues, optimally and make optimal decisions.

How optimal are humans?

- Vision researchers can design Ideal Observer Models.
- These can be used to benchmark human performance compared to an ideal observer who know how the stimuli are generated.
- Humans typically perform poorly compared to the benchmark.
- But for certain tasks – like motion estimation – it appears that humans use an ecological prior, based on the statistics of natural images, instead the ‘prior’ used by the experimenter to construct the stimuli.
- Conjecture – humans are optimal for ecological stimuli.

Optimal Vrs. Slow and Smooth

- Optimal for Experiment – vrs. Slow-and-smooth (e.g., HongJing Lu).
- Humans perform orders of magnitude worse than an ideal model which knows the probabilistic model that generated the stimuli.
- But humans perform similarly to a model that assumes a prior of slow-and-smooth motion, similar to the prior measured from the statistics of natural image sequences.

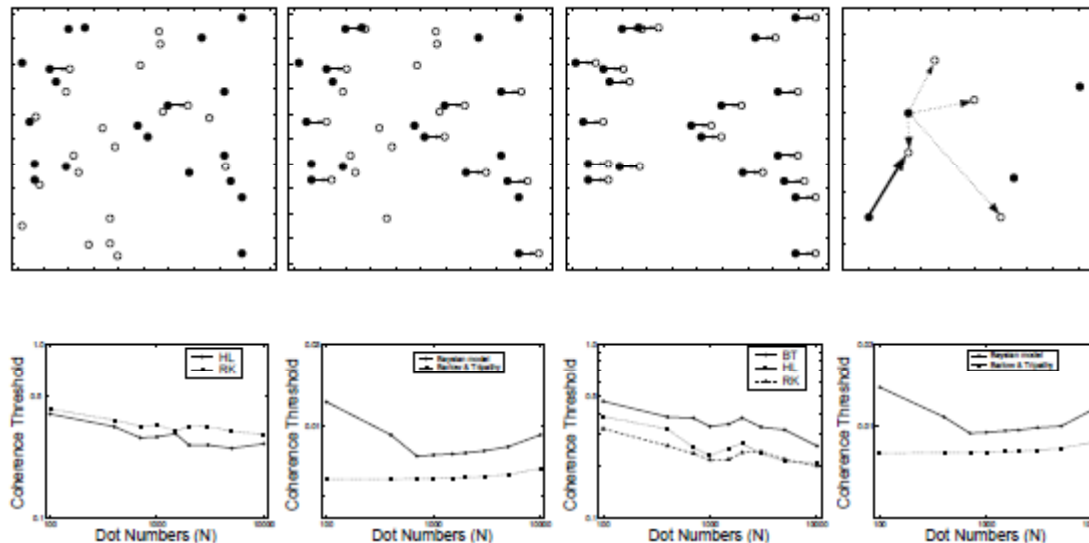


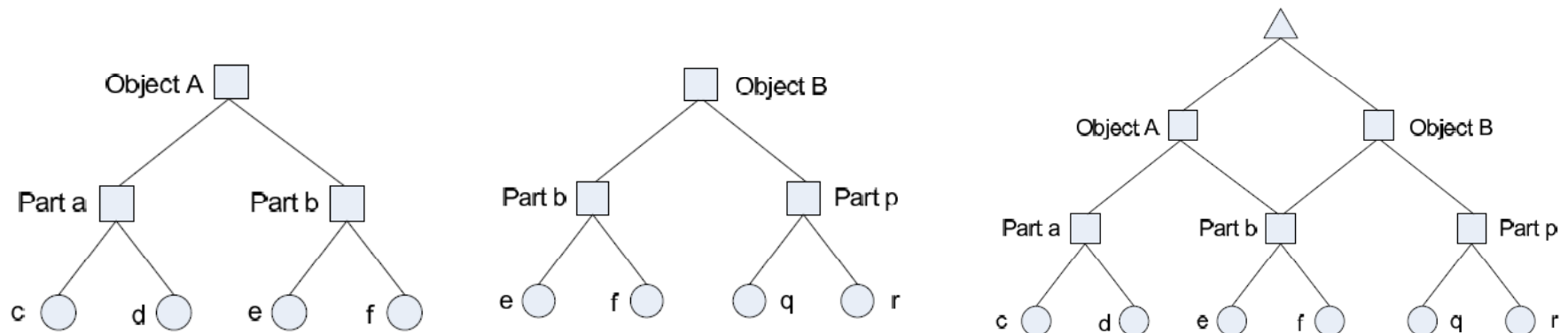
Figure 3: The left two panels show detection thresholds – human subjects (far left) and BIO and BT thresholds (left). The right two panels show discrimination thresholds – human subjects (right) and BIO and BT (far right).

Stochastic Grammars and Compositional Models

- Objects are composed of parts. These are composed of subparts, and so on.
- Semantic parts – e.g., arms and legs of humans.
- Structured graphical models, probability defined over these models, inference algorithms.
- Learning the models with, or without, known graph structure.
- Talks by Geman, Bienenstock, Yuille, Zhu, Poggio.

Key Idea: Compositionality

- Objects and Images are constructed by compositions of parts – ANDs and ORs.
- The probability models for are built by combining elementary models by composition.
- Efficient Inference and Learning.



Why compositionality?

- (1). Ability to transfer between contexts and generalize or extrapolate (e.g. , from Cow to Yak).**
- (2). Ability to reason about the system, intervene, do diagnostics.**
- (3). Allows the system to answer many different questions based on the same underlying knowledge structure.**
- (4). Scale up to multiple objects by part-sharing.**

“An embodiment of faith that the world is knowable, that one can tease things apart, comprehend them, and mentally recompose them at will.”

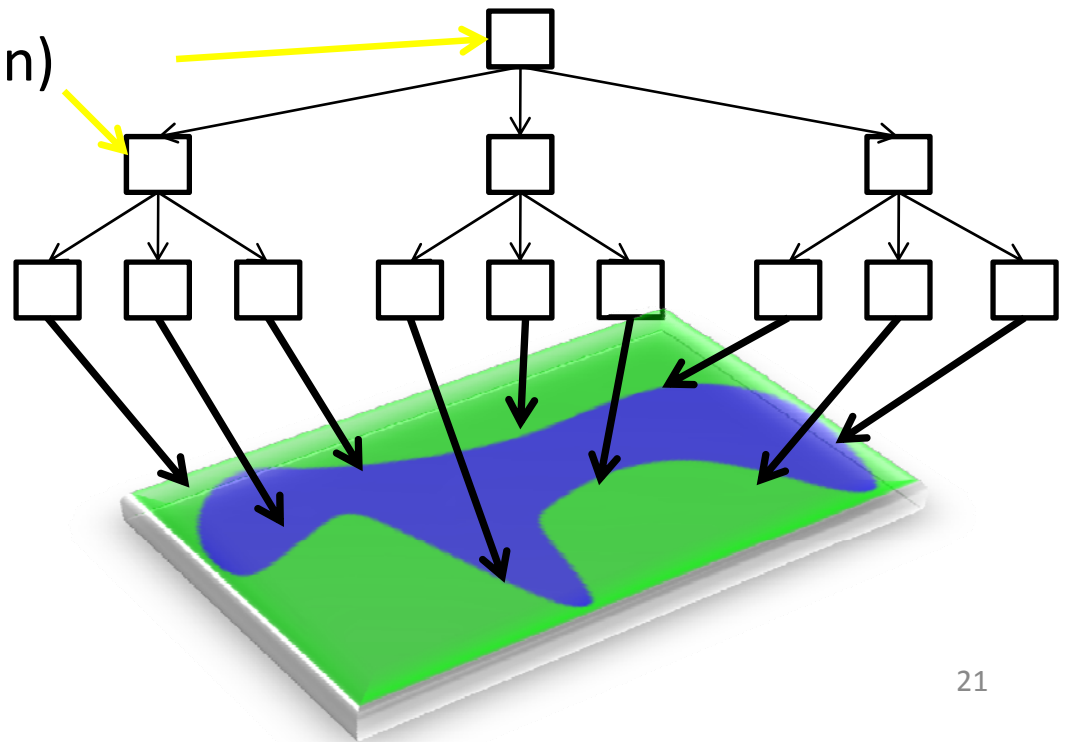
“The world is compositional or God exists”.

Horse Model (ANDs only).

Nodes of the Graph represents parts of the object.

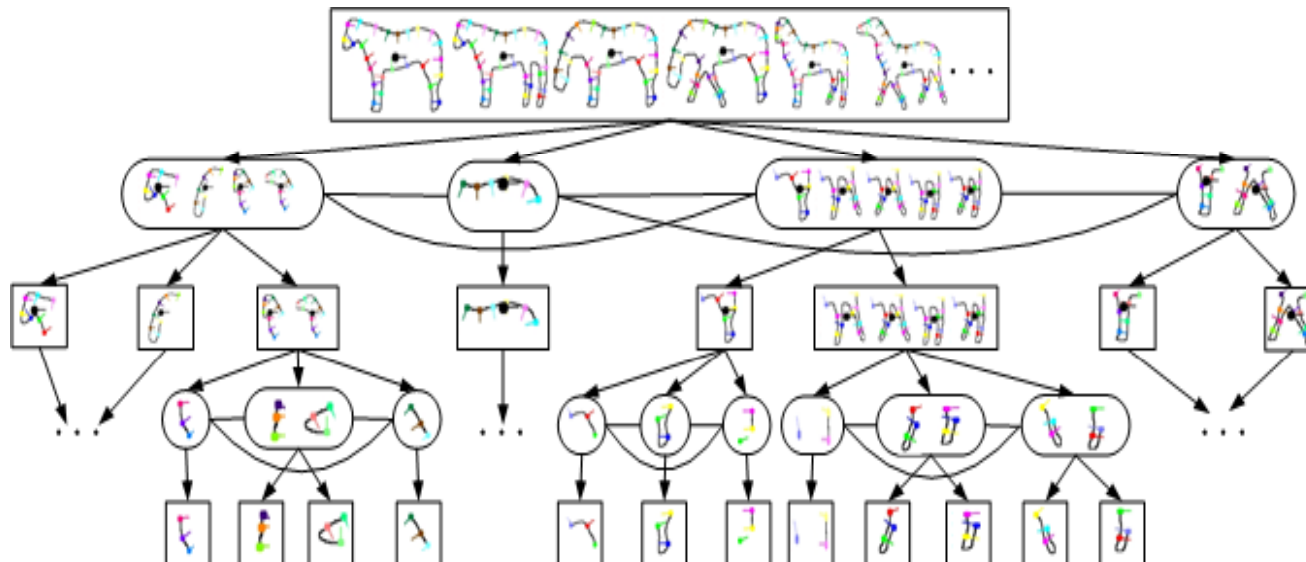
Parts can move and deform.

y : (position, scale, orientation)



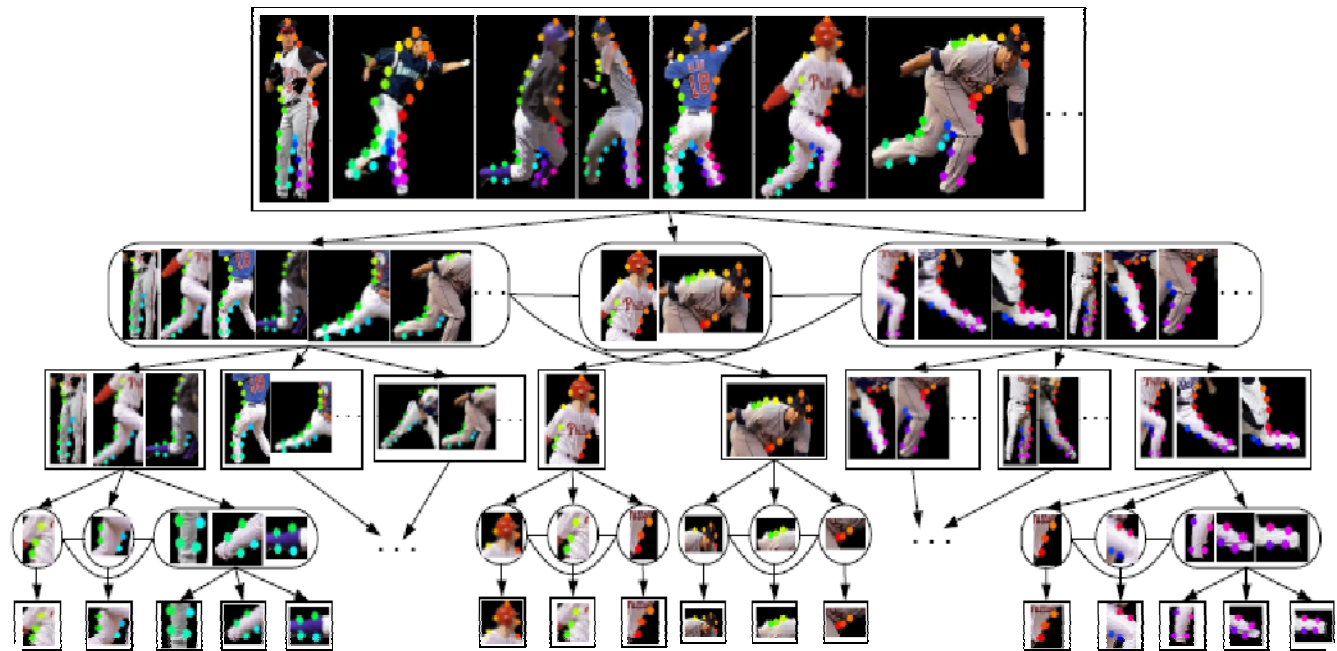
AND/OR Graphs for Horses

- Introduce OR nodes and switch variables.
- Settings of switch variables alters graph topology – *allows different parts for different viewpoints/poses:*
- Mixtures of models – with shared parts.



AND/OR Graphs for Baseball

- Enables RCMs to deal with objects with multiple poses and viewpoints (~100).
- Inference and Learning by bottom-up and top-down processing:



Results on Baseball Players:

- Performed well on benchmarked datasets.
- Zhu, Chen, Lin, Lin, Yuille CVPR 2008, 2010.



Unsupervised Structure Learning

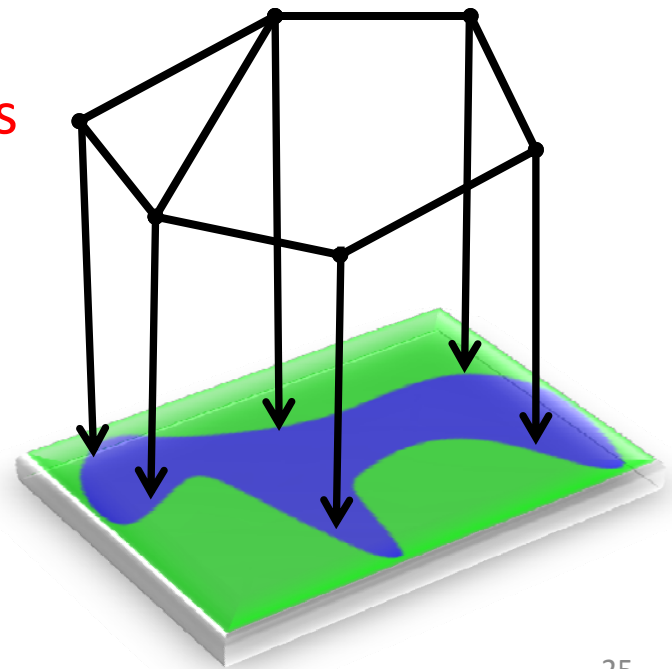
- Task: given 10 training images, no labeling, no alignment, highly ambiguous features.
 - Estimate Graph structure (nodes and edges)
 - Estimate the parameters.



Correspondence is
unknown

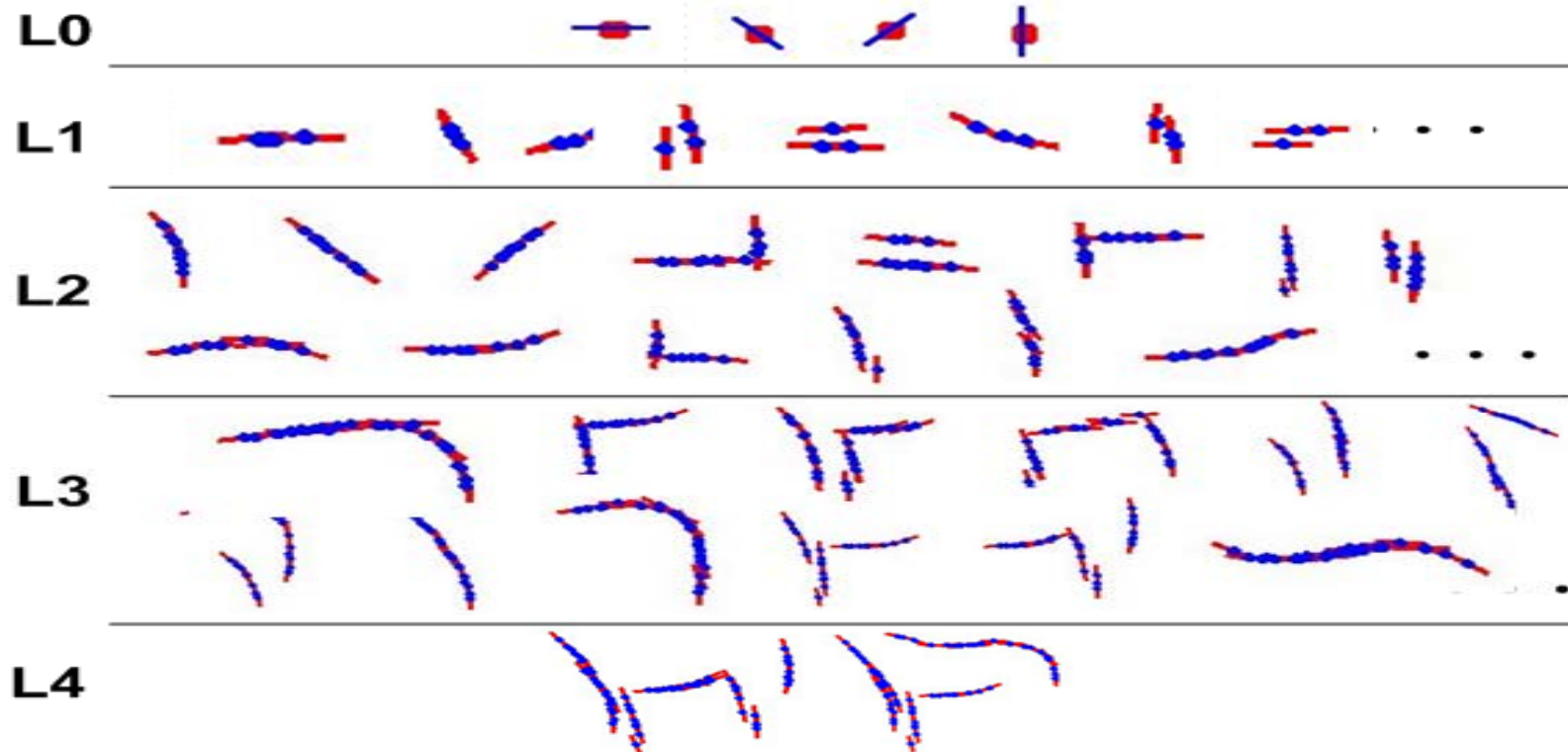


Combinatorial
Explosion problem

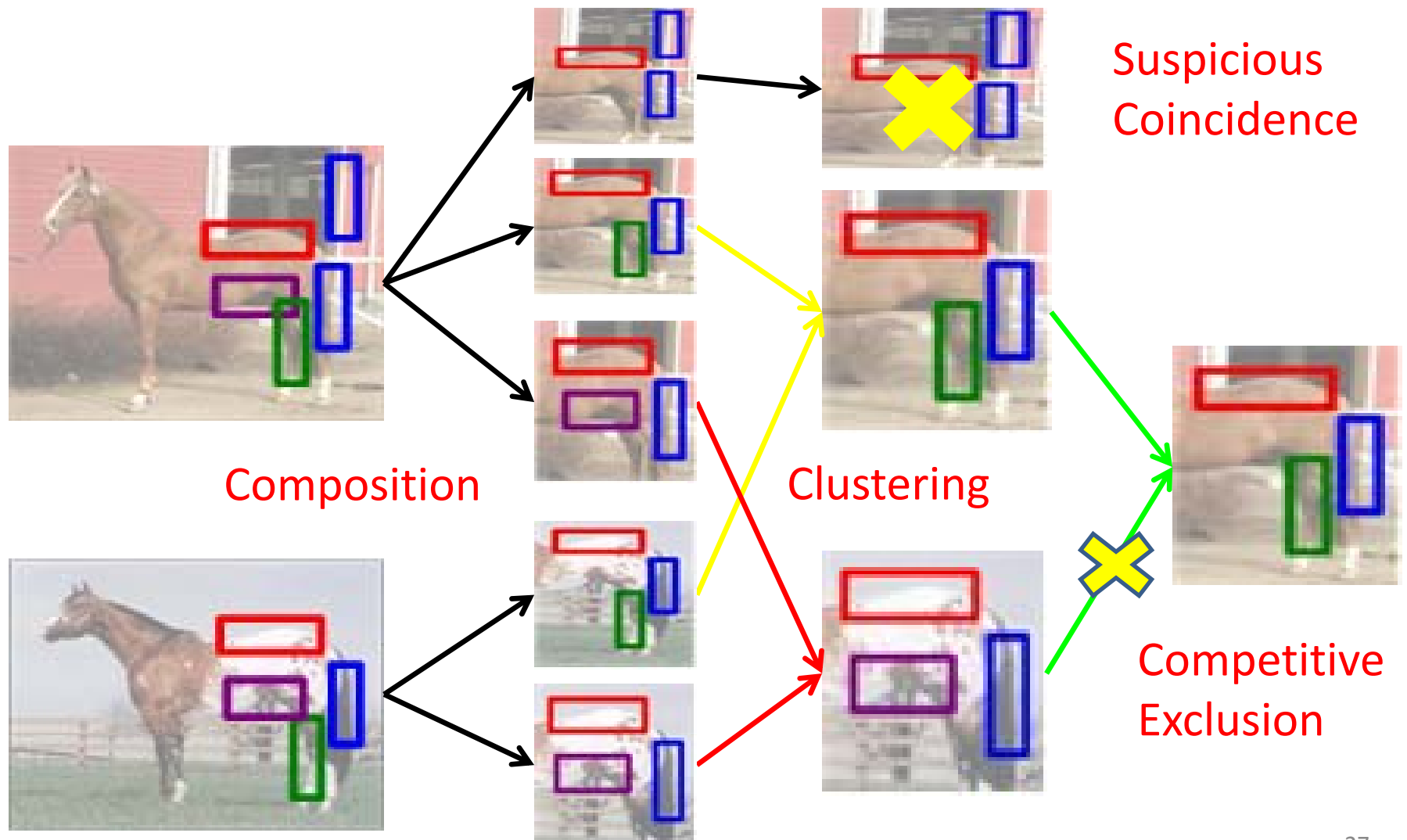


The Dictionary: From Generic Parts to Object Structures

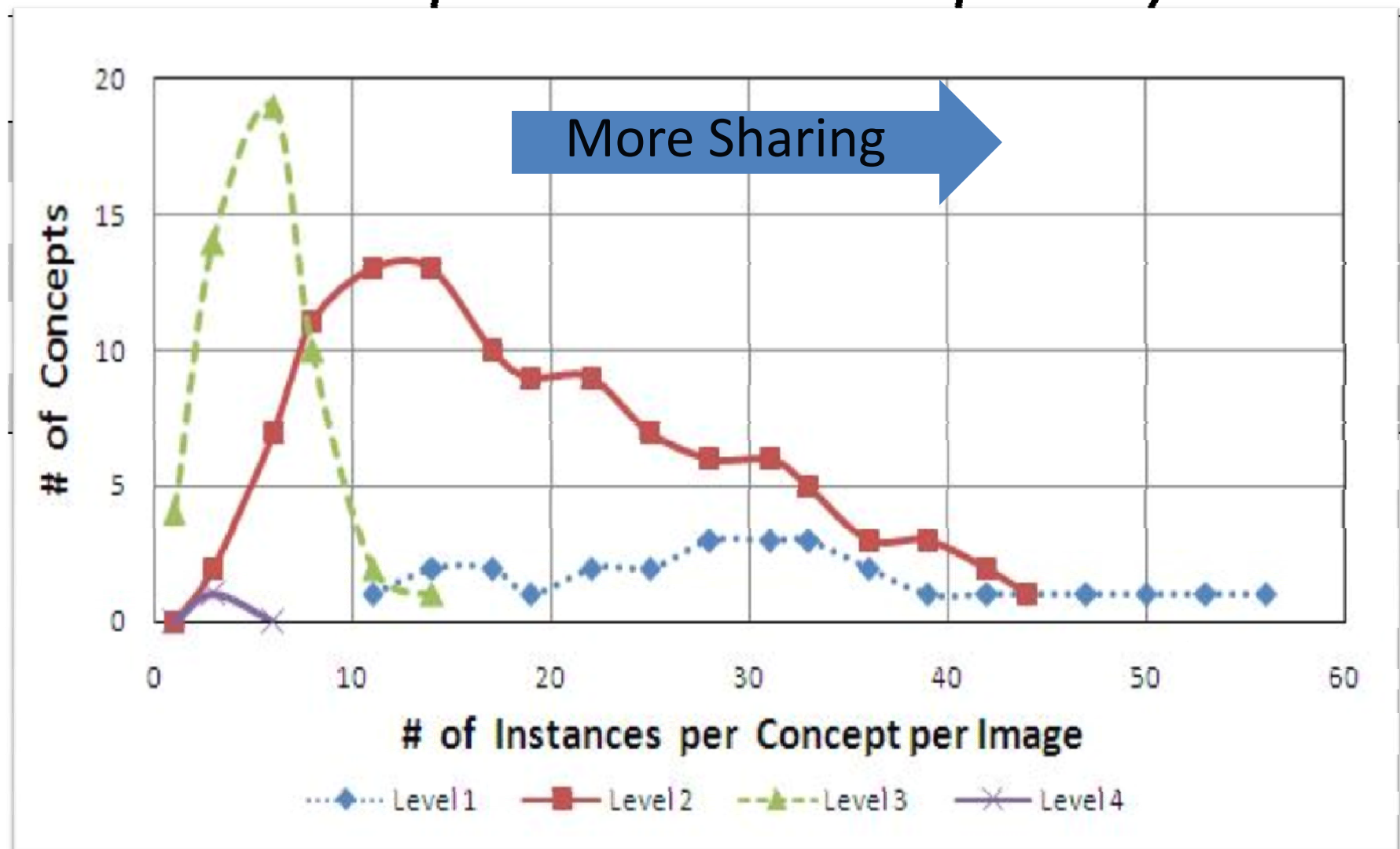
- Unified representation (RCMs) and learning
- Bridge the gap between the generic features and specific object structures



Bottom-up Learning

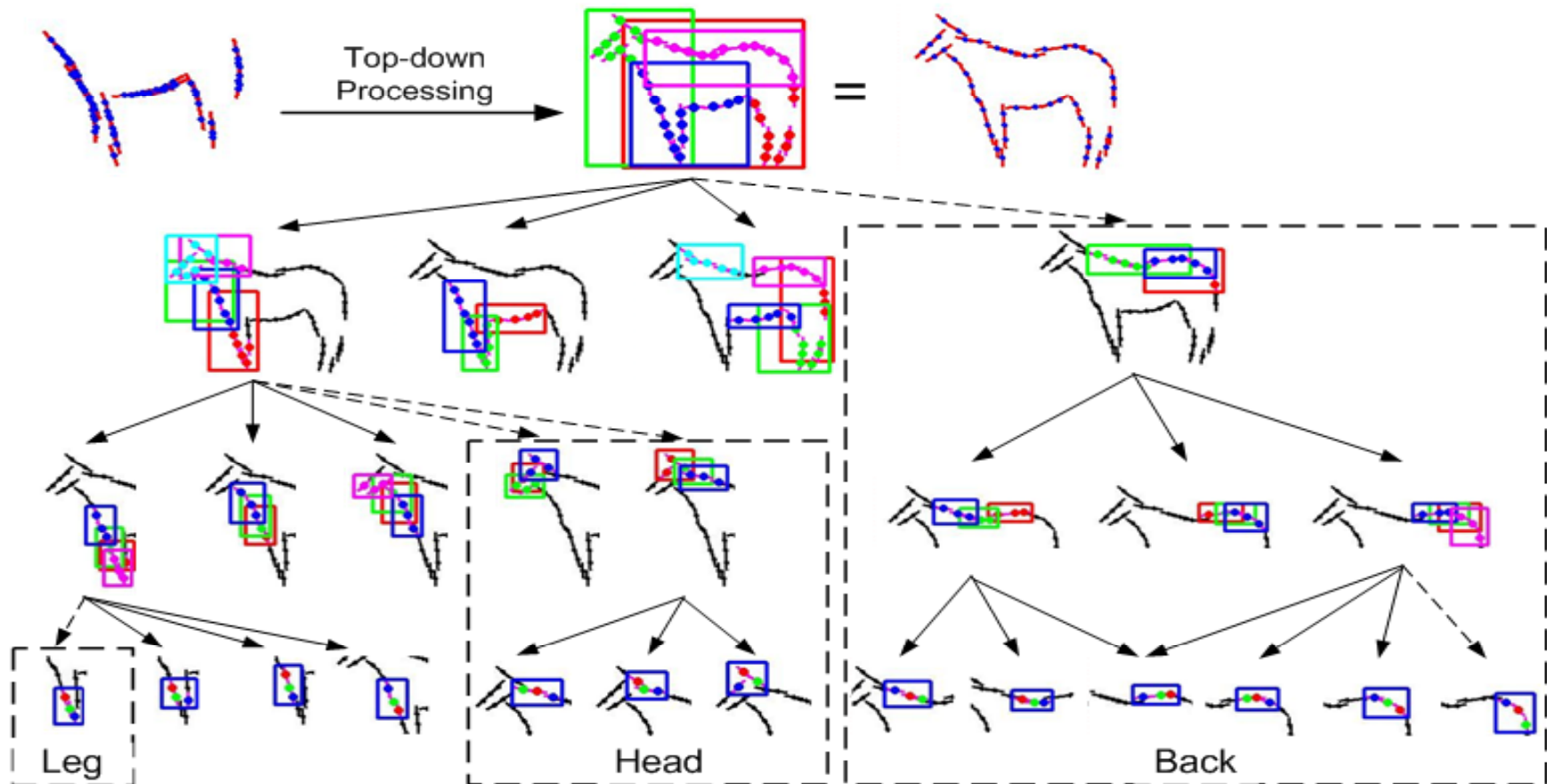


Dictionary Size, Part Sharing and Computational Complexity



Top-down refinement

- Fill in missing parts
- Examine every node from top to bottom



Part Sharing for multiple objects

Strategy: share parts between different objects and viewpoints.

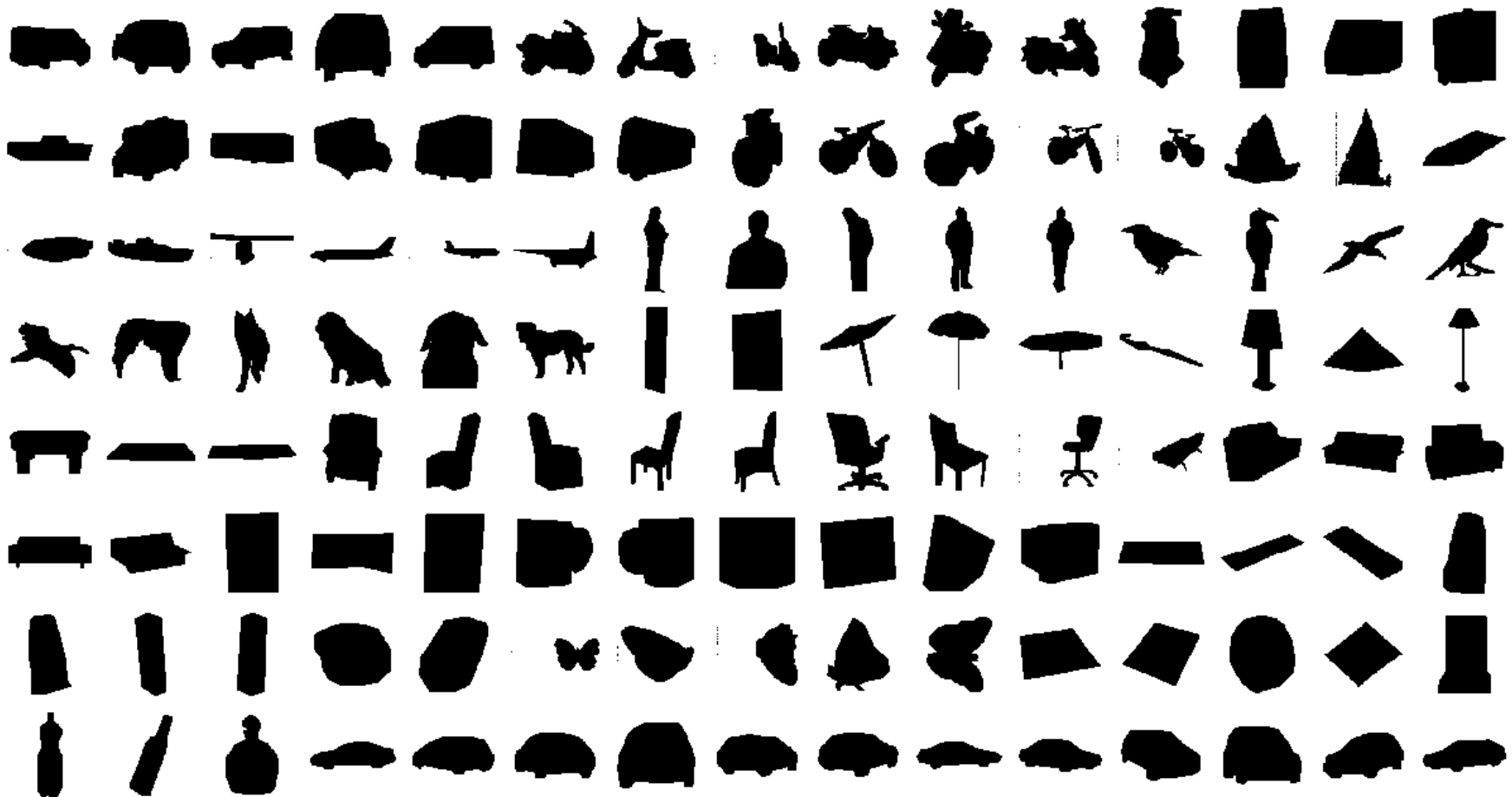


Learning Shared Parts

- Unsupervised learning algorithm to learn parts shared between different objects.
- Zhu, Chen, Freeman, Torralba, Yuille 2010.
- Structure Induction – learning the graph structures and learning the parameters.
- Supplemented by supervised learning of masks.

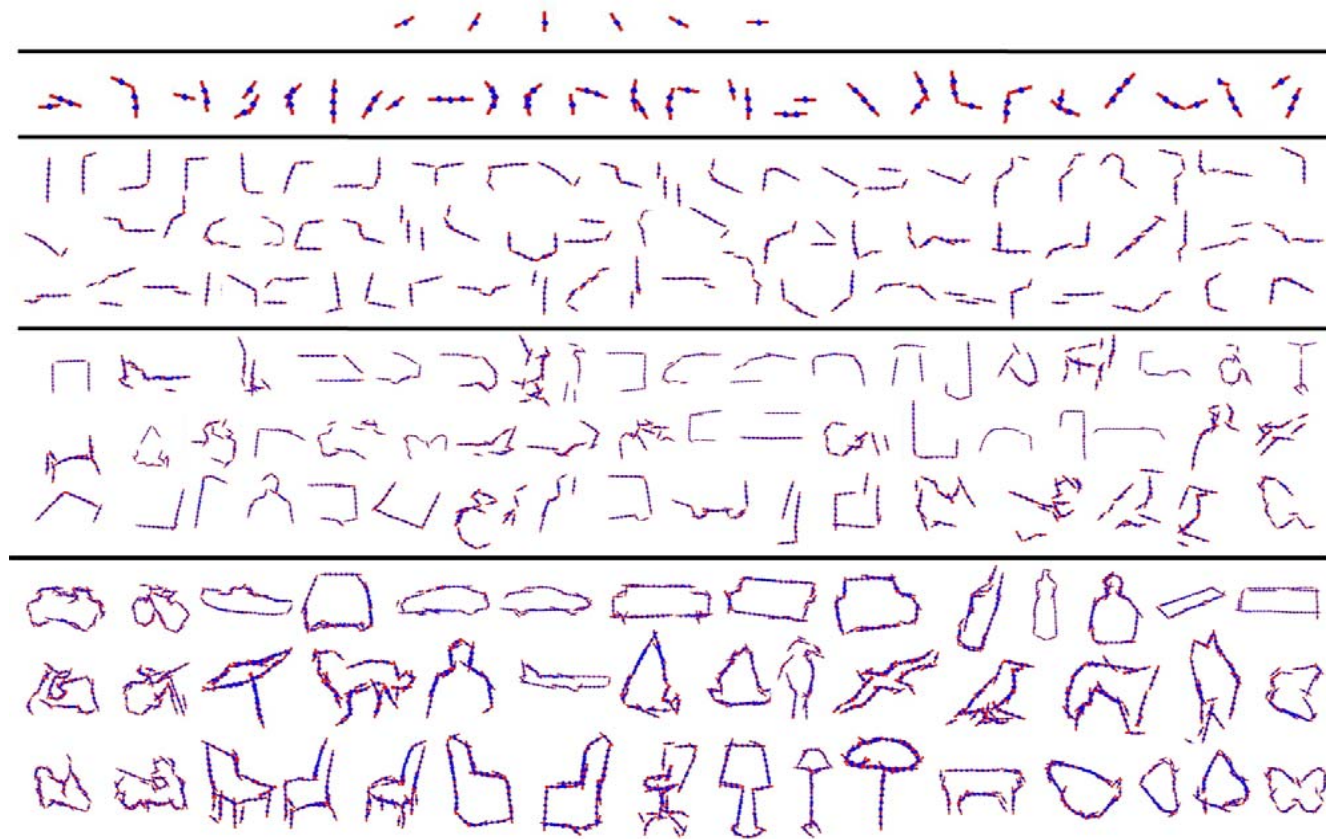
Many Objects/Viewpoints

- 120 templates: 5 viewpoints & 26 classes

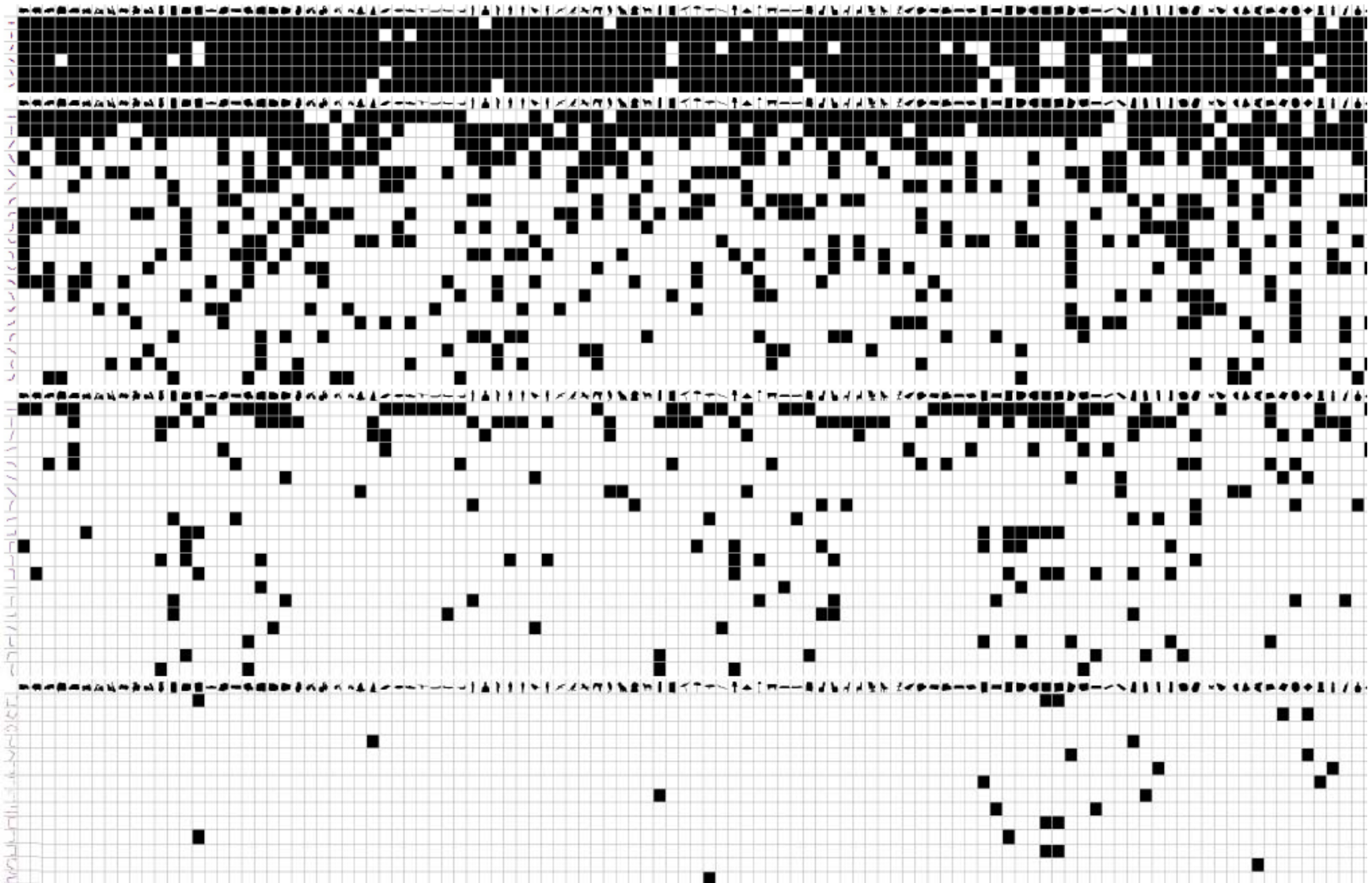


Learn Hierarchical Dictionary.

- Low-level to Mid-level to High-level.
- Learn by suspicious coincidences.

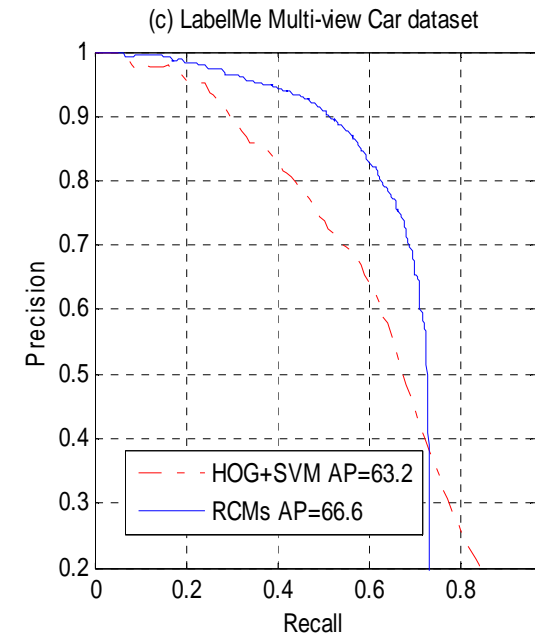
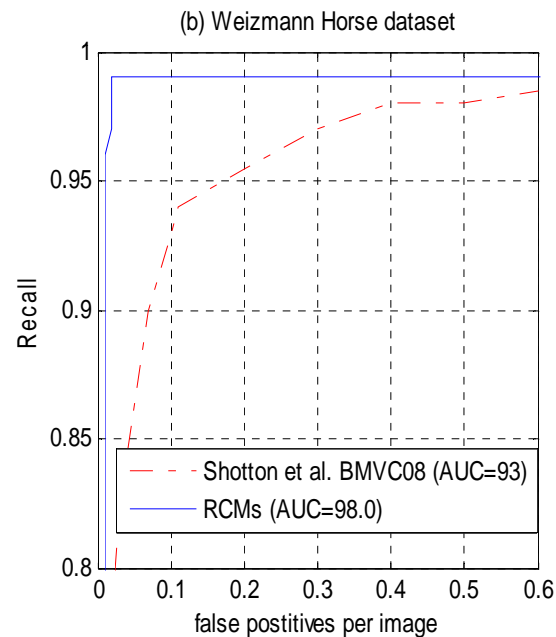
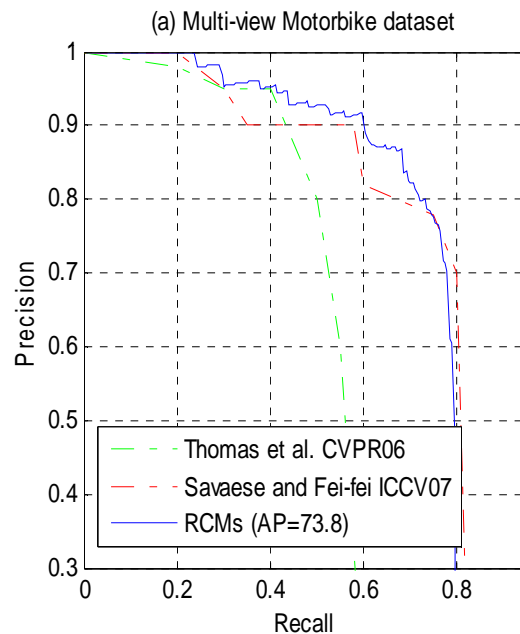


Part Sharing decreases with Levels



Multi-View Single Class Performance

- Comparable to State of the Art.



Conclusions

- Principles: Recursive Composition
 - Composition -> complexity decomposition
 - Recursion -> Universal rules (self-similarity)
 - Recursion and Composition -> sparseness
- A unified approach – object detection, recognition, parsing, matching, image labeling.
- Statistical Models, Machine Learning, and Efficient Inference algorithms.
- Extensible Models – easy to enhance.
- Scaling up: shared parts, compositionality.
- Trade-offs: sophistication of representation vrs. Features.
- *The Devil is in the Details.*