Unsupervised Learning of Compositional Models for Visual Objects

A.L. Yuille
Dept. Statistics, UCLA
Dept. Brain and Cognitive Engineering,
Korea University.

Compositional Models of Objects

- *Compositional Models represent objects in terms of object parts and their spatial relations.*
- These parts are represented recursively in terms of subparts (with spatial relations), and sub-subparts,…
- *Detecting an object also estimates the positions of its parts and subparts automatically.*
- Composition allows explicit part-sharing, yielding big gains in computational efficiency (**Saturday talk**).
- This talk describes *unsupervised learning algorithms which learn representation of objects*.

Compositional Models: Examples

- Examples: Models of Baseball Players and Horses.
- *Executive Summary*: High-level nodes encode coarse descriptions of object. E.g., centroid position
- Details (e.g., leg positions) are specified by lower-level nodes.

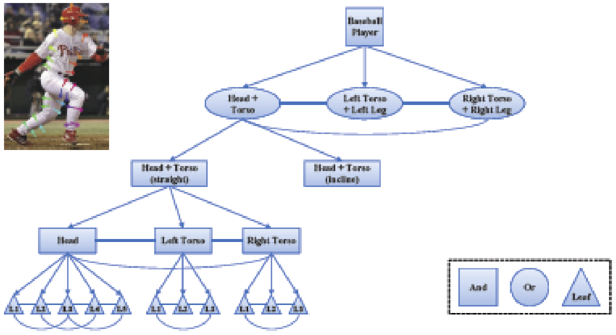
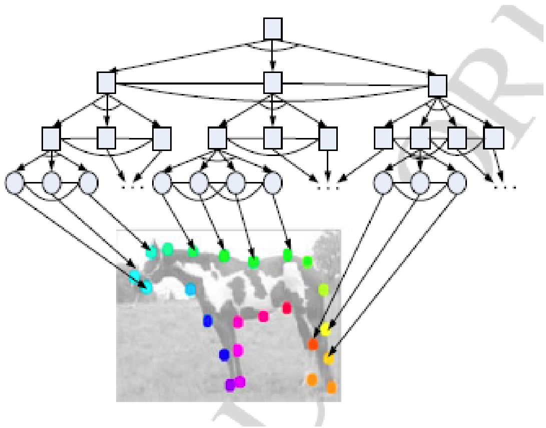
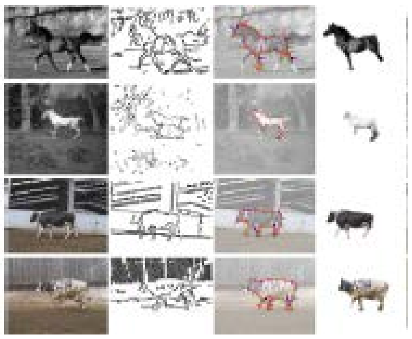


Figure 4. The AND/OR Graph Model (Zhu, Chen, Lin, & Yuille, 2010). The Baseball player is an AND of the head and torso, and left and right legs, but the head is an OR of straight head and torso or an inclined head and torso (top left).





Prior and Related Work

- Prior work on compositional and grammatical models of vision: typically hand-specifies the graphical and grammatical structures of the models. Although the parameters are learnt.
- S. Geman, S. Todorovic, SC Zhu, D.B. Mumford, L. Zhu, A.L. Yuille, P. Felzenzswalb, C. Williams.
- It is desirable to learn the structure of these models automatically.

Advantages of Explicit Representations

Compositionality

Construct models by composing smaller elements.

This enables:

(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.

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

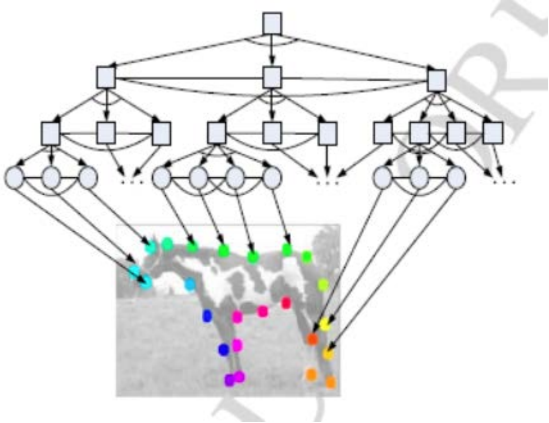
"The world is compositional or God exists". S. Geman.

Mathematics of Compositional Models

Part 2: Compositionality

- Build models for elementary components.
- What is a Compositional Model?
- A probability distribution defined over a graph specified by parent-child relations:

$$\prod_{\nu} \mathrm{P}(\vec{\mathrm{x}}_{ch(\nu)}|\mathrm{x}_{\nu})\mathrm{P}_{\mathrm{p}}(\mathrm{x}_{\mathrm{root}})$$



Key Property: Modularity

- The probability distribution of an object is composed from parts composed of subparts.
- *This enables you to make new distributions – by extracting one part of the object and replacing it by a different parts. Or by altering the parameters of the parts (e.g., making a leg thicker).*
- These changes can be done in a modular manner.
- *More generally, construct a distribution by building it from elementary parent-child components.*
- Modularity enables us to learn the distributions, one parent-child clique at a time.

Parent-Child components: basic building blocks

- Parent-Child determinism:

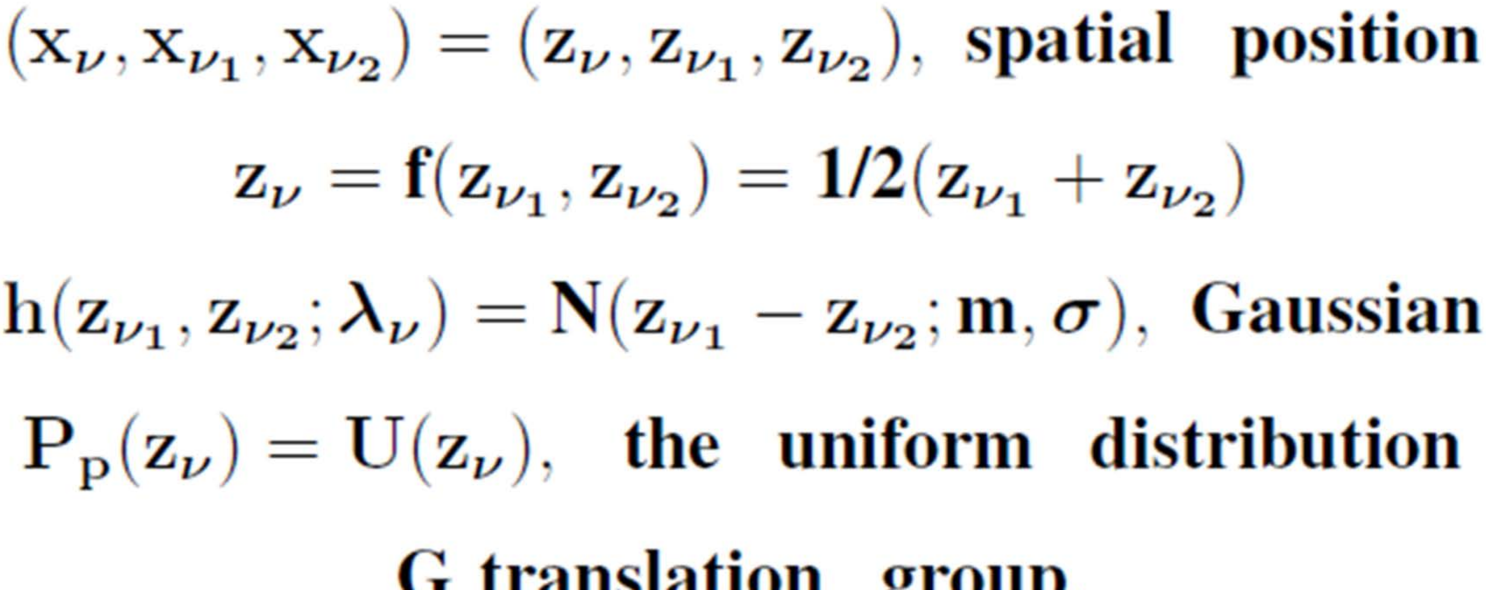
$$\mathbf{P}(\vec{\mathrm{x}}_{ch(\nu)}|\mathrm{x}_{\nu}, \boldsymbol{\lambda}_{\nu}) = \boldsymbol{\delta}(\mathrm{x}_{\nu} - \mathbf{f}(\vec{\mathrm{x}}_{ch(\nu)}))\mathbf{h}(\vec{\mathrm{x}}_{\nu}; \boldsymbol{\lambda}_{\nu})$$

- f() is a determinist function
- Executive summary – e.g., parent node encodes average position.
- h() spatial relations between child nodes. Specified by parameter lambda.
- Prior propagation: If object has uniform prior position, then subpart has uniform prior.

$$\mathbf{P}_{\mathrm{p}}(\mathrm{x}_{\nu_i}) = \sum_{\vec{\mathrm{x}}_{ch(\nu)}/\mathrm{x}_{\nu_i}} \boldsymbol{\delta}(\mathrm{x}_{\nu} - \mathbf{f}(\vec{\mathrm{x}}_{ch(\nu)}))\mathbf{h}(\vec{\mathrm{x}}_{ch(\nu)}; \boldsymbol{\lambda}_{\nu})\mathbf{P}_{\mathrm{p}}(\mathrm{x}_{\nu})$$

Parent-Child Example:

- Executive summary: parent node take mean position of child nodes.
- Spatial relations between parts are specified by Gaussian distribution on relative positions.



$$(\mathrm{x}_{\nu}, \mathrm{x}_{\nu_1}, \mathrm{x}_{\nu_2}) = (\mathrm{z}_{\nu}, \mathrm{z}_{\nu_1}, \mathrm{z}_{\nu_2}), \;\; \textbf{spatial} \;\; \textbf{position}$$

$$\mathrm{z}_{\nu} = \mathbf{f}(\mathrm{z}_{\nu_1}, \mathrm{z}_{\nu_2}) = \mathbf{1/2}(\mathrm{z}_{\nu_1} + \mathrm{z}_{\nu_2})$$

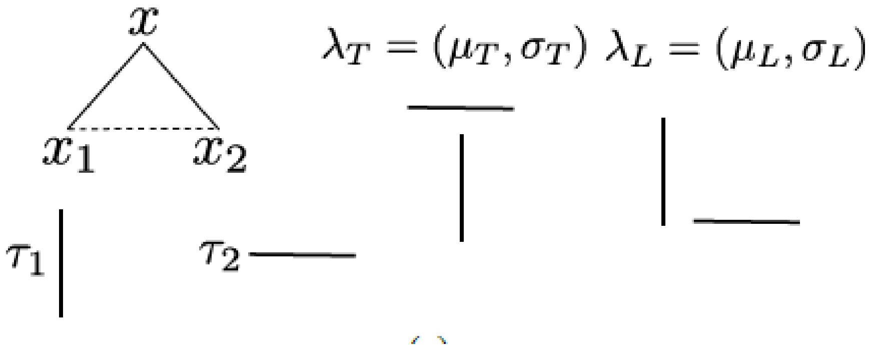
$$\mathrm{h}(\mathrm{z}_{\nu_1}, \mathrm{z}_{\nu_2}; \lambda_{\nu}) = \mathbf{N}(\mathrm{z}_{\nu_1} - \mathrm{z}_{\nu_2}; \mathbf{m}, \boldsymbol{\sigma}), \;\; \textbf{Gaussian}$$

$$\mathrm{P}_{\mathrm{p}}(\mathrm{z}_{\nu}) = \mathrm{U}(\mathrm{z}_{\nu}), \;\; \textbf{the} \;\; \textbf{uniform} \;\; \textbf{distribution}$$

$$\textbf{G} \;\; \textbf{translation} \;\; \textbf{group}$$

Compositional Models for T and L

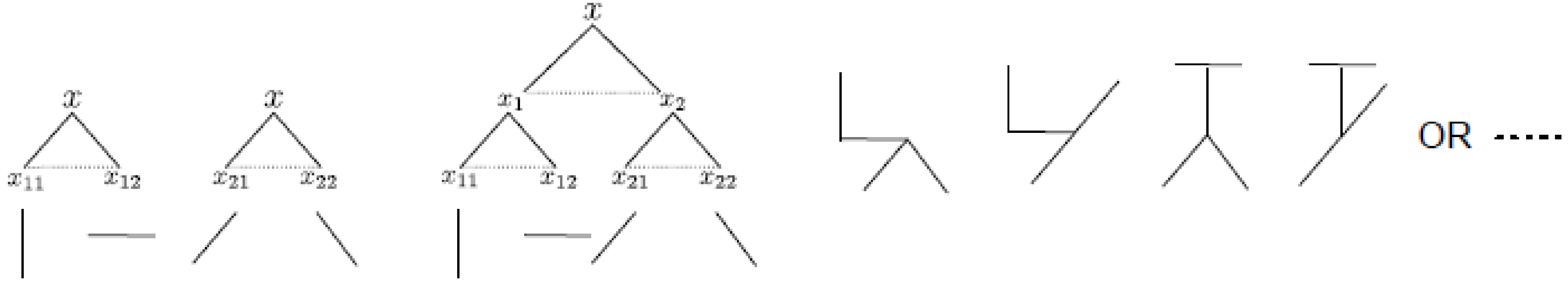
- How to make a T or an L?
- *Dictionary* of Level-0 models:
- E.g., horizontal or vertical bars.
- Level-1 model – T or L – is a composition of two Level-0 models plus spatial relations..
- Child nodes: horizontal or vertical bars.



$\lambda_T = (\mu_T, \sigma_T)$ $\lambda_L = (\mu_L, \sigma_L)$

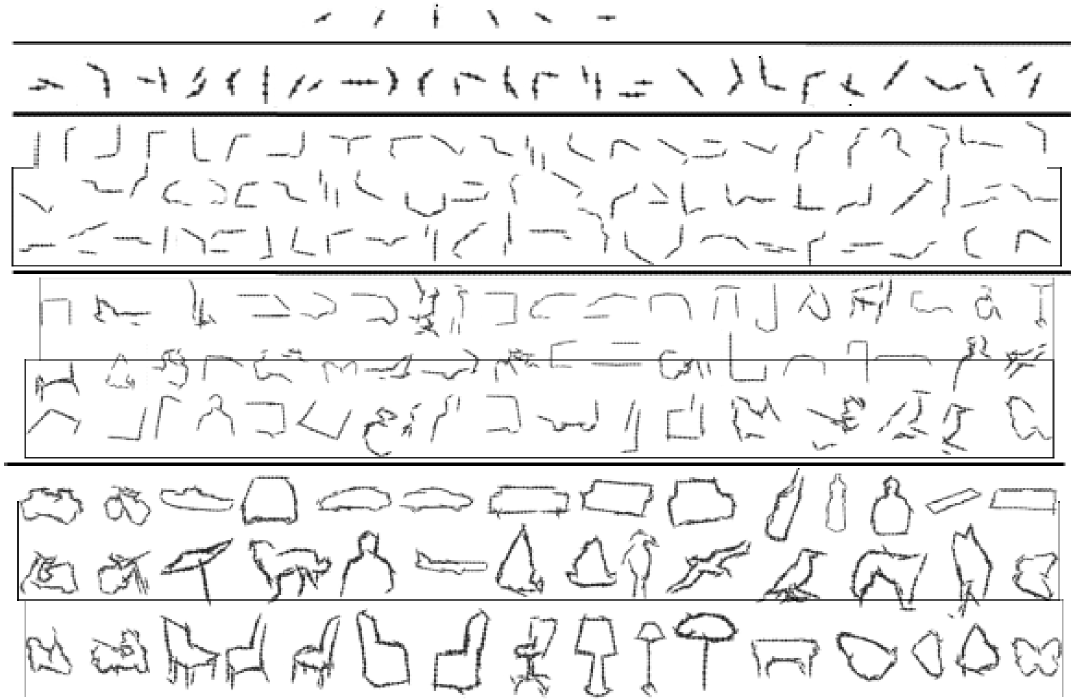
Compositional Learning: dictionaries.

- Start with a dictionary of Level-0 models.
- Learn a Dictionary of Level-1 models by combining models from the Level-0 dictionary.
- Repeat to build Level-2 dictionaries and high-level dictionaries.



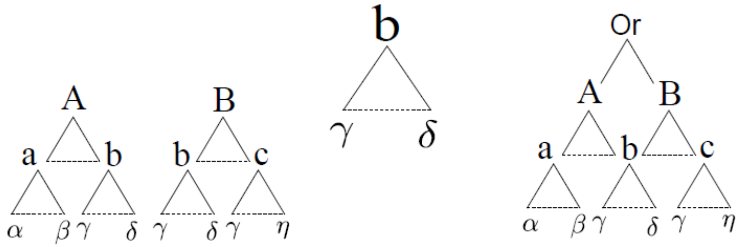
Examples of dictionaries for 120 objects.

- The mean shapes of elements of dictionaries at: Level-0, Level-1, Level-2 Level-3, Level-4.
- Note: the dictionaries are probability distribution, but we only show their mean shapes.



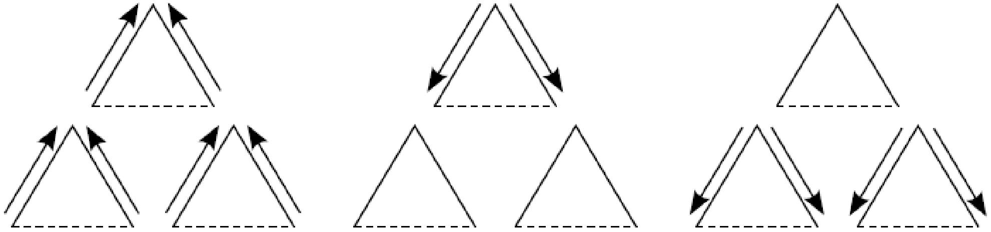
Multiple Objects:

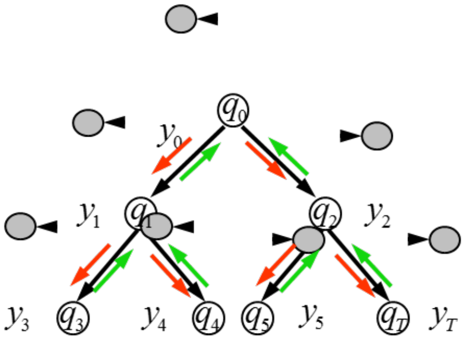
- Multiple objects can be represented in terms of these hierarchical dictionaries.
- This enables part-sharing between objects – dictionary elements used in several objects.
- *Part-Sharing enables efficient learning, representation and inference.* (**Saturday talk**).



Inference on Compositional Models.

- We perform inference using Dynamic Programming (message passing).
- Bottom-Up propagates local hypotheses to obtain consistent top-level interpretations.
- Top-down disambiguates local hypotheses.





- Discussed in detail in **Saturday**. Inference can be parallelized.

Unsupervised Learning

- Automatically learns a hierarchical set of dictionaries.
- Method: clustering, efficient encoding.
- Theory: parallel search through set of possible generative models of the data.

- ***Number of levels is determined automatically by the algorithm.***

How to Learn Compositional Models?

- Cocktail party problem – object in cluttered background.

- Hard Learning Problems: (unsupervised)
- Do not know the graph structure of the model (e.g., no. of levels)
- (ii) Do not know the assignment of leaf nodes of the model to the data.
- (iii) Do not know the model parameters (lambdas).

Strategy: Exploit Modularity

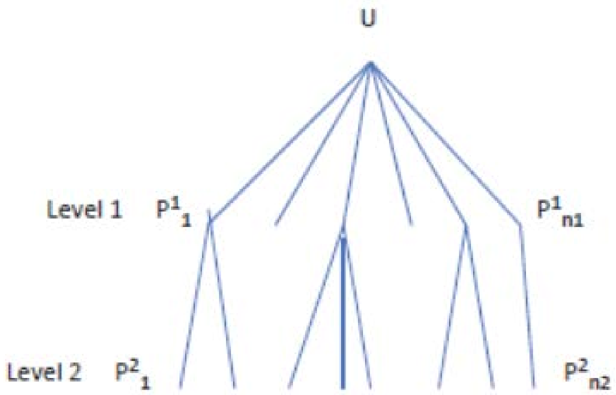
Start by learning the lowest levels of the dictionaries – i.e. the smallest parts.

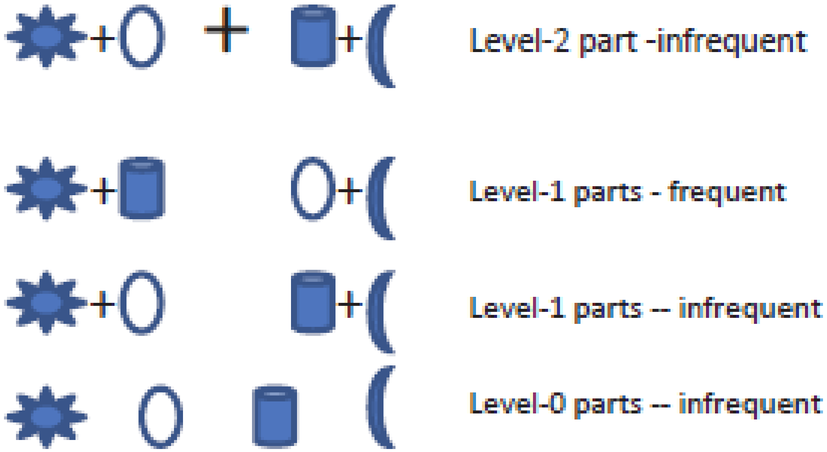
Learn these dictionary elements separately. Allow for overlap – we can enforce consistency later.

Each dictionary element gives an encoding of the data which is better than the encoding by the root model (uniform distribution).

Proceed level by level. Build new models by composition from models at lower levels. Impose consistency of assignments during composition.

Parallel Search in Model Space



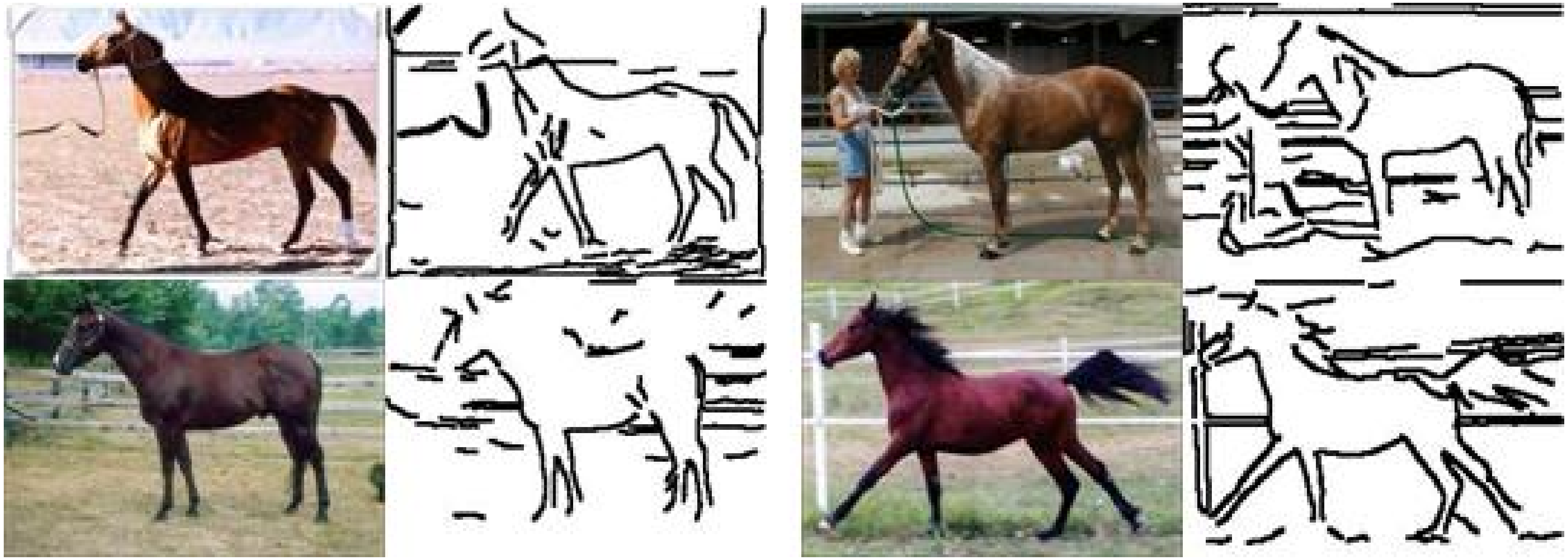


Low-level models may perform poorly by themselves, but may combine well to form good high-level compositions.

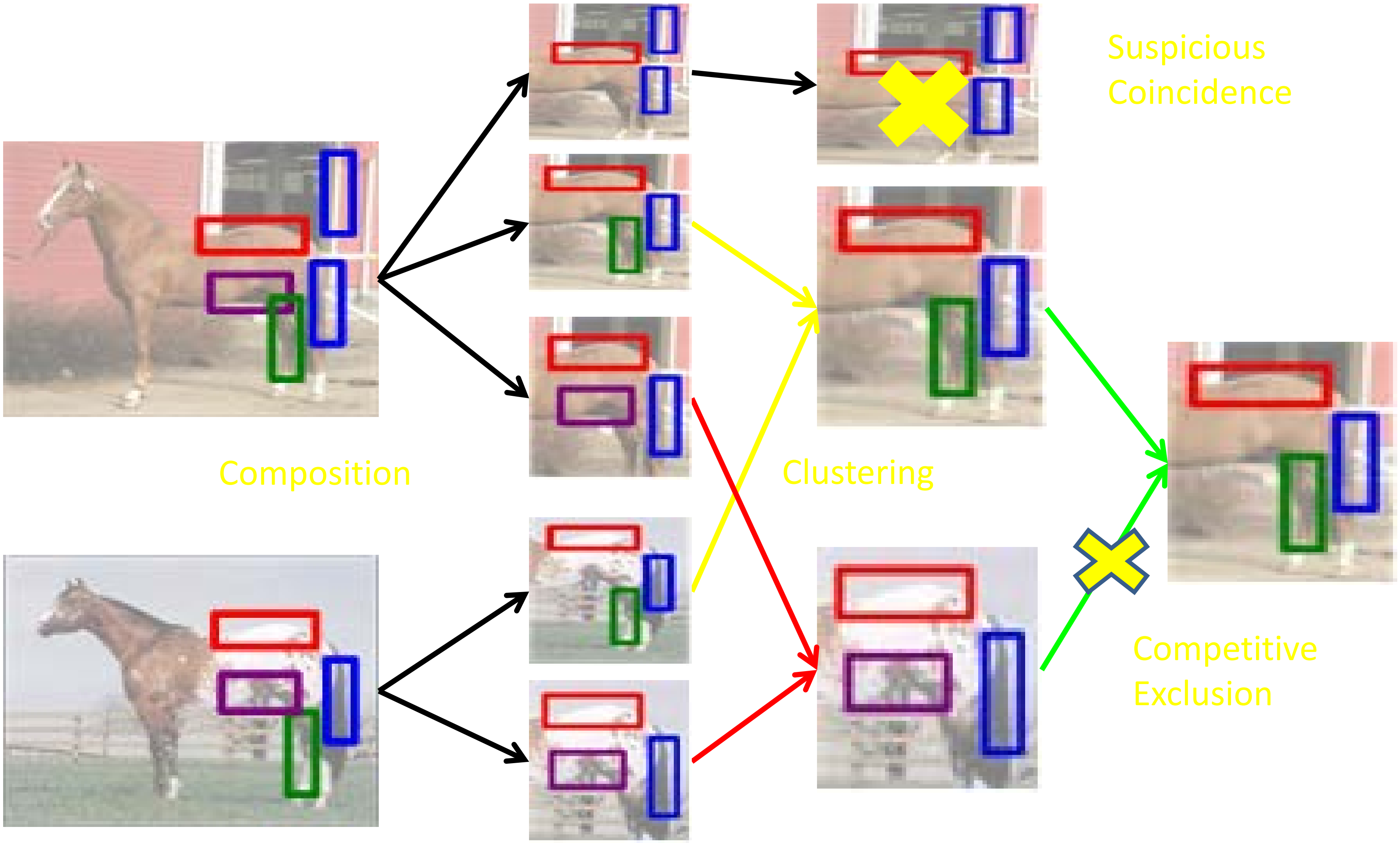
Do not reject weak models too soon.

Horse Dataset: L.Zhu et al. 2008.

- Input Images: Horse Dataset. 10 images used for training. 300 for testing.

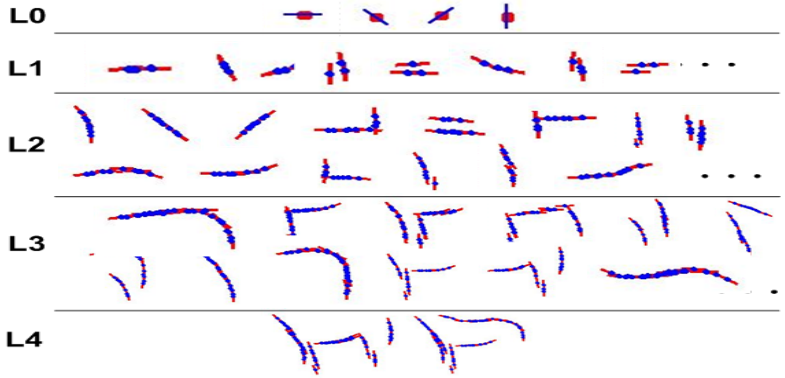


Compositional Learning



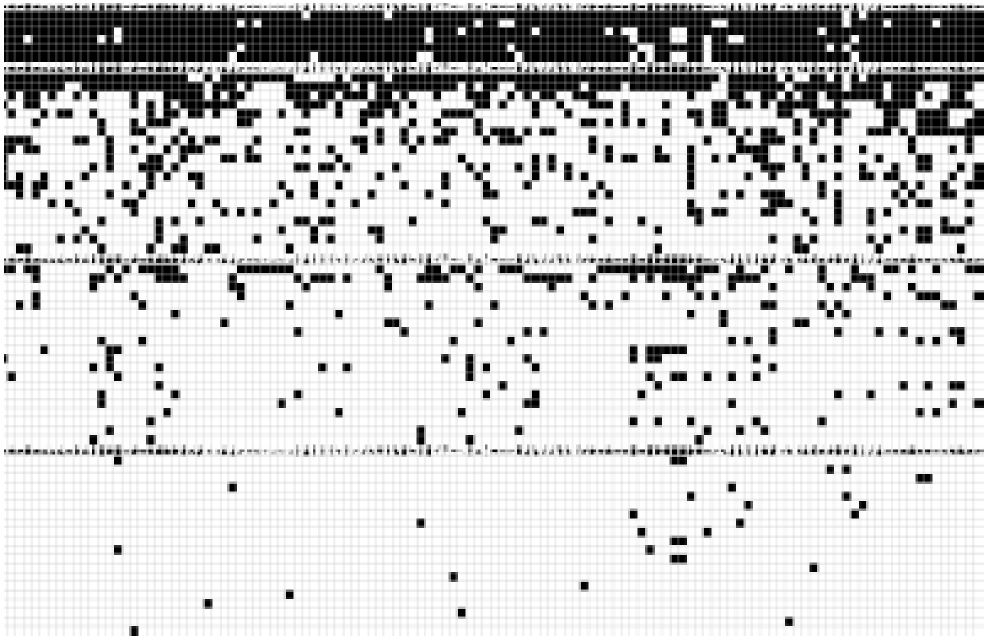
Generic Parts to Object Structures

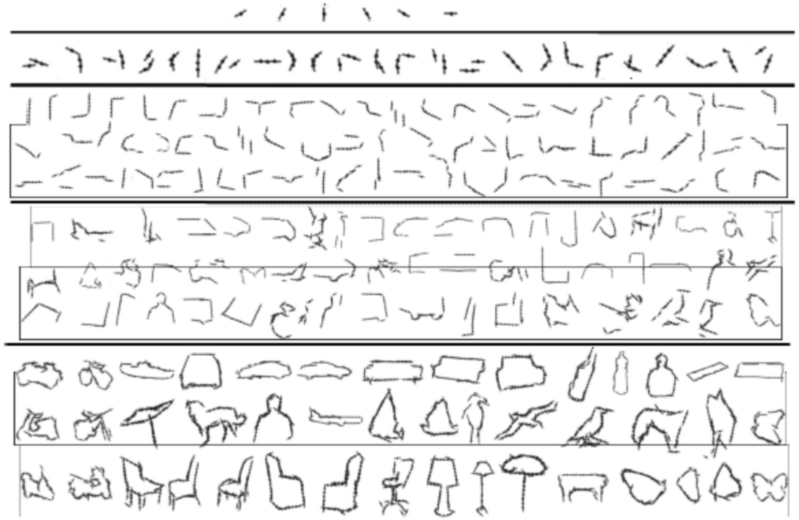
- As we go up the hierarchy, the dictionaries mimic the features used in low-level, mid-level, and high-level vision.
- *E.g., Mid-level gives ‘gestalt rules’*
- High-level is specific to the object. Low-, and mid-level are more generic.



Dictionaries and part sharing.

- Sharing of parts between 120 objects (horizontal)
- Vertical – Level-1, :Level-2,..
- Part sharing is very frequent at low levels. But less sharing at higher levels.

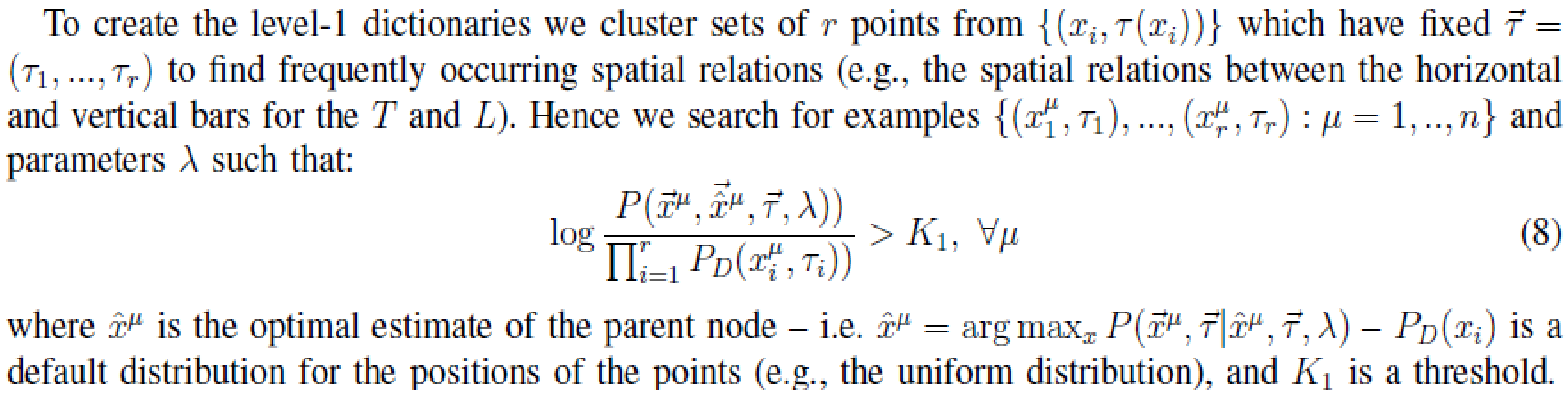




Brief Mathematical Descriptions

- The input to computational learning are a set of images. We assume a set M_0 of level-0 dictionary models, which are pre-specified – e.g., edge detectors.
- For each image, we obtain a set of points with their corresponding types: $(x_i;\ tau(x_i))$.
- The type – tau – indicates the element of the level-0 dictionary (e.g., horizontal or vertical bar).

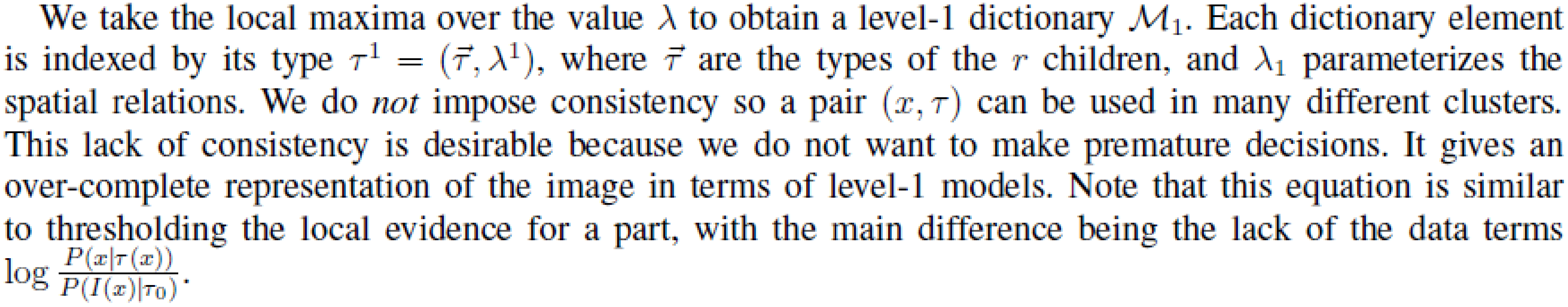
Brief Mathematics: Better encoding.



To create the level-1 dictionaries we cluster sets of r points from $\{(x_i, \tau(x_i))\}$ which have fixed $\vec{\tau} = (\tau_1, ..., \tau_r)$ to find frequently occurring spatial relations (e.g., the spatial relations between the horizontal and vertical bars for the T and L). Hence we search for examples $\{(x_1^\mu, \tau_1), ..., (x_r^\mu, \tau_r) : \mu = 1, .., n\}$ and parameters λ such that:

$$\log \frac{P(\vec{x}^\mu, \vec{\hat{x}}^\mu, \vec{\tau}, \lambda))}{\prod_{i=1}^{r} P_D(x_i^\mu, \tau_i))} > K_1, \ \forall \mu \qquad (8)$$

where $\hat{x}^\mu$ is the optimal estimate of the parent node – i.e. $\hat{x}^\mu = \arg\max_x P(\vec{x}^\mu, \vec{\tau} | \hat{x}^\mu, \vec{\tau}, \lambda)$ – $P_D(x_i)$ is a default distribution for the positions of the points (e.g., the uniform distribution), and K_1 is a threshold.



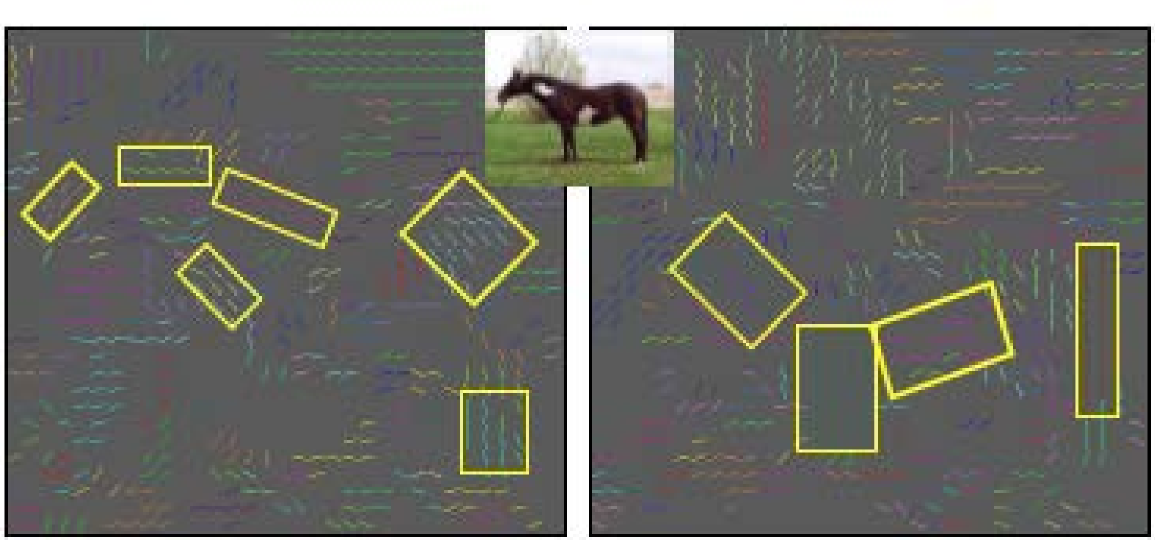
We take the local maxima over the value λ to obtain a level-1 dictionary $\mathcal{M}_1$. Each dictionary element is indexed by its type $\tau^1 = (\vec{\tau}, \lambda^1)$, where $\vec{\tau}$ are the types of the r children, and λ_1 parameterizes the spatial relations. We do *not* impose consistency so a pair (x, τ) can be used in many different clusters. This lack of consistency is desirable because we do not want to make premature decisions. It gives an over-complete representation of the image in terms of level-1 models. Note that this equation is similar to thresholding the local evidence for a part, with the main difference being the lack of the data terms $\log \frac{P(x|\tau(x))}{P(I(x)|\tau_0)}$.

What Inputs to Use?

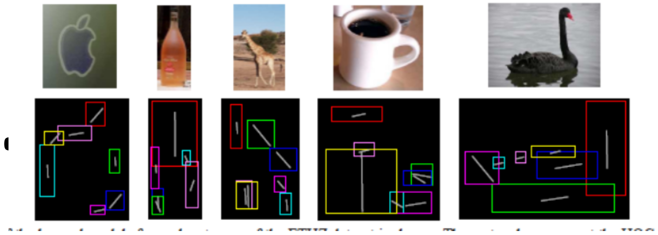
- The work described above uses edges as inputs. But alternative features can be used.
- For example, we can use HOG-Bundles (Mottaghi and Yuille 2011). These are built from HOG features by local spatial grouping.
- Note: edges have disadvantages because there are many of them and have similar properties. HOG-bundles are fewer and easier to differentiate.

HOG-Bundles

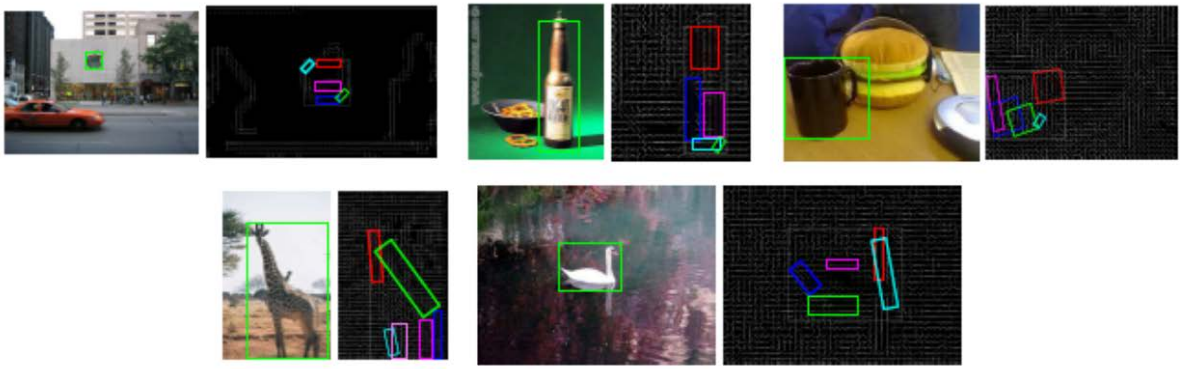
- Start with HoG-bundle representation of images.
- Hog-bundles: HOGs detect edges – HOG-bundles group by proximity and collinearity.
- HOG bundles often correspond (roughly) to parts of object.



ETHZ dataset.

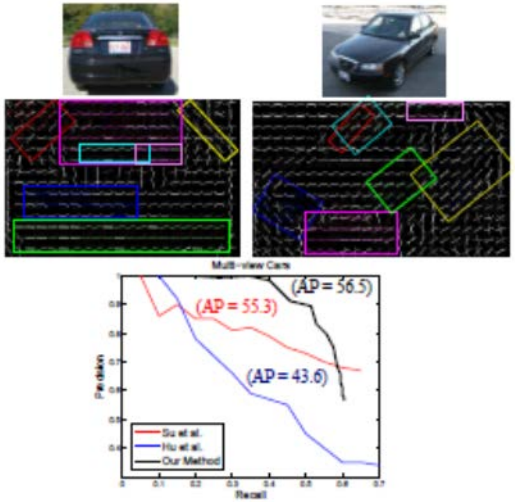


- Learnt models for each category of the ETHZ dataset. Rectangles represent the HOG bundles.
- No. of parts and relative position/orientation is learnt automatically.



Multiview Car Dataset

- Learns models for different viewpoints (automatically). Test on Car dataset (Su, Sun, Fei Fei, Savarase 2009).
- Performance was best – expect for methods with explicit 3D car models.



Any Relation to Neurons?

- There are some interesting relations to work by L. Valiant in Circuits of the Mind.
- Valiant studies how sets of “neuroids” could automatically store memories of conjunctions,.
- His more recent work considers memorizing conjunctions of conjunctions – analogous to higher level compositions.
- His interest was in Random Access Memory models. But the same ideas could be used for compositional models.

Summary

- Compositional Models represent objects explicitly in terms of parts, subparts, and spatial relations.
- This explicitness enables diagnostics and transfer.
- Unsupervised learning – learns dictionaries bottom-up exploiting modularity.
- Part-sharing – makes learning efficient.
- Efficiency of Inference and representation and parallel implementation (**Saturday talk**).
- But will they work on Pascal or ImageNet?..

References:

- S. Geman et al.. Composition Systems. Quarterly of Applied Mathematics, 60. 2002.
- D.M. Mumford and A. Desolneux. Pattern Theory. 2010
- L. Zhu et al. Unsupervised Structure Learning. ECCV. 2008.
- L. Zhu et al. Part and Appearance Sharing. CVPR 2010.
- R. Mottaghi and A.L. Yuille. A compositional approach to learning part-based models for single and multi-view object detection. ICCV. 2011.