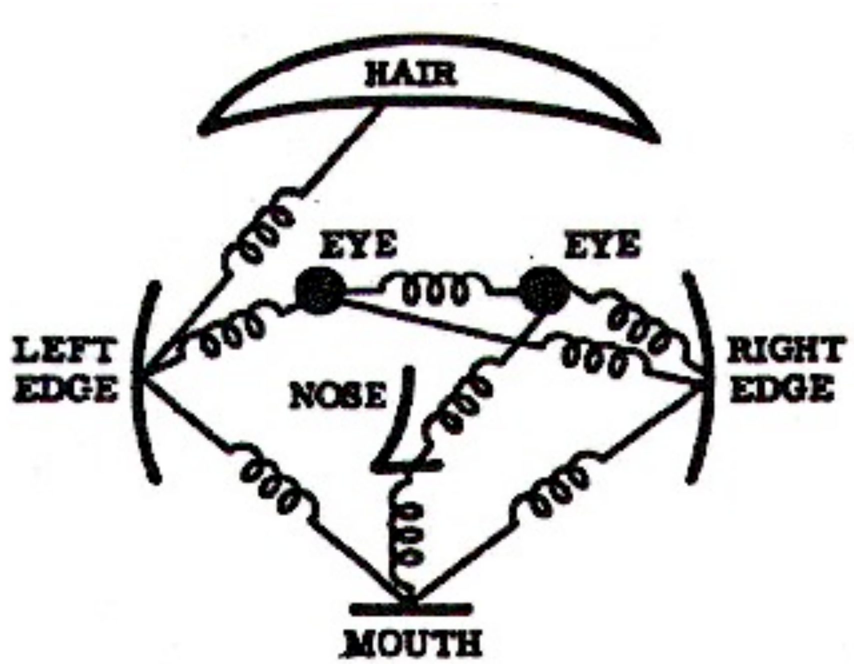
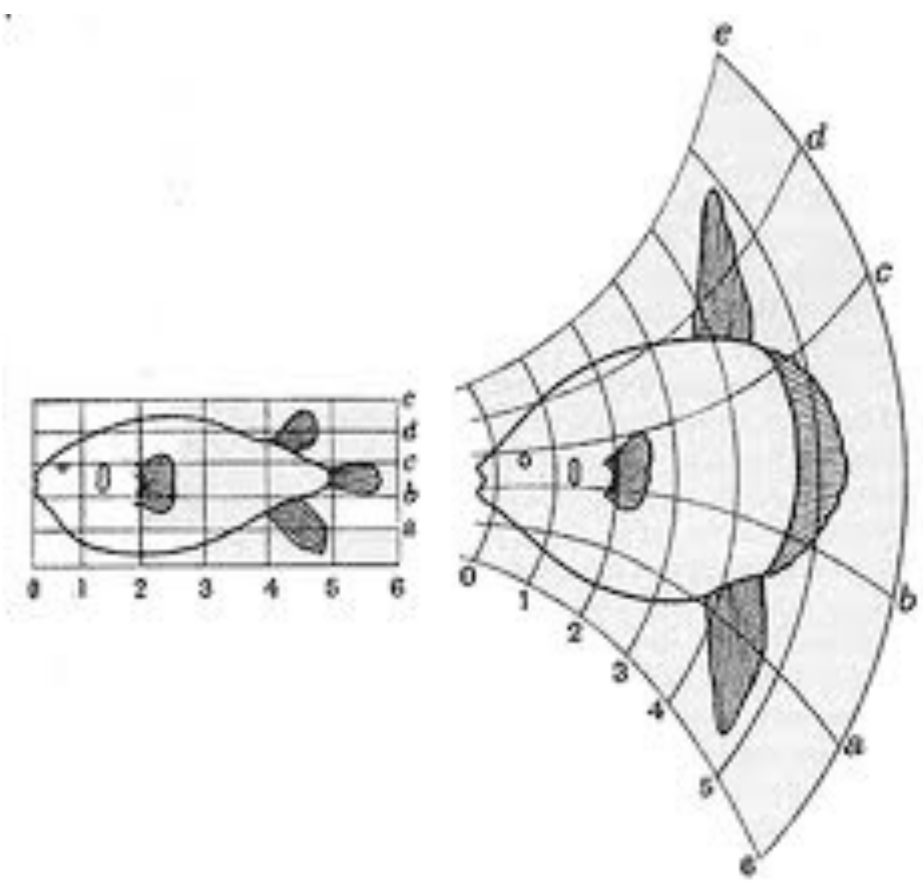
Compositional models

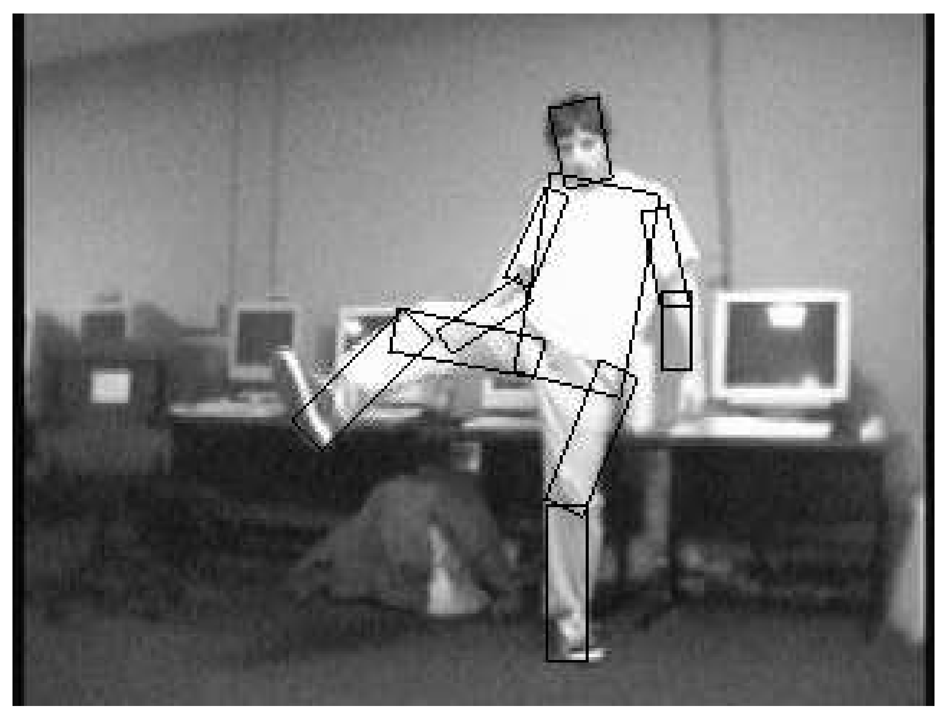
Pedro Felzenszwalb
Brown University

Deformable models

- Can take us a long way...
- But not all the way

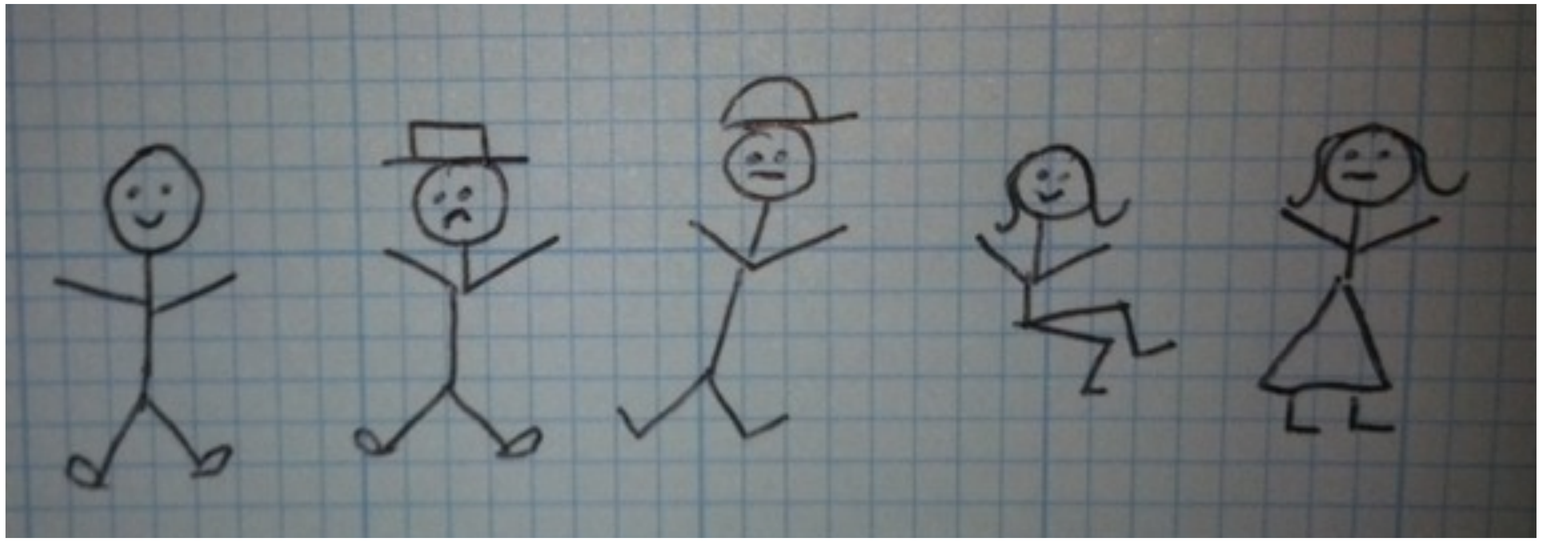






Structure variation

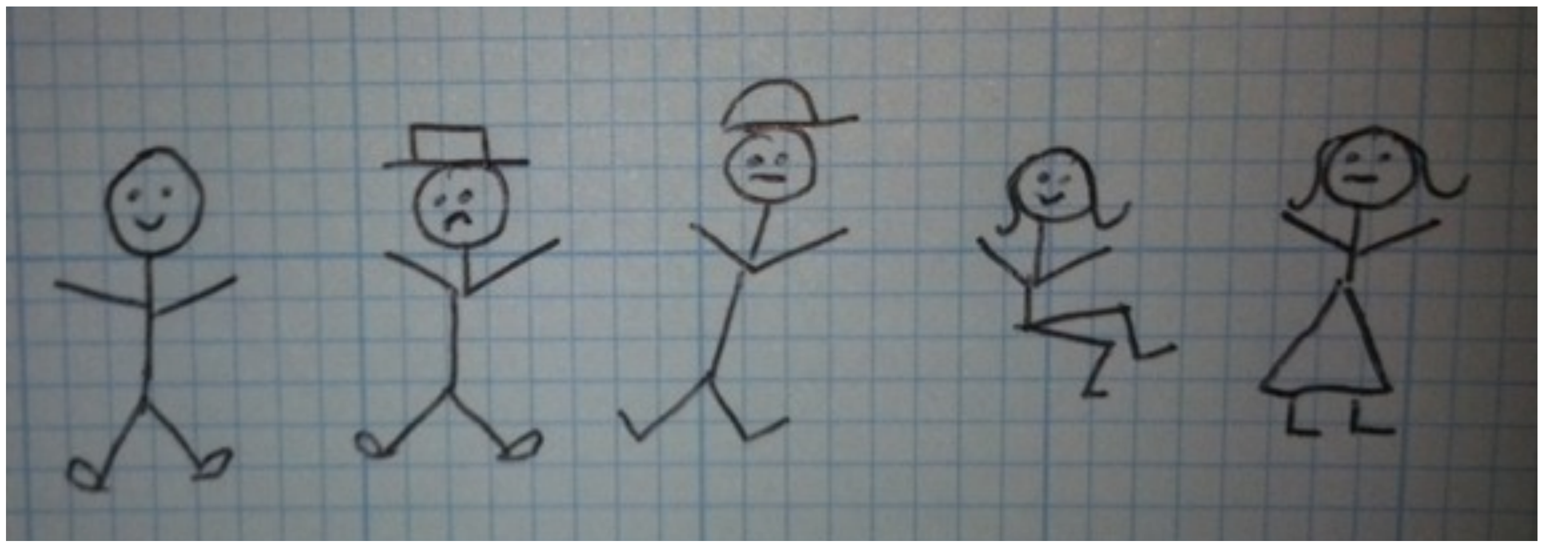
- Object in rich categories have variable structure



- These are NOT deformations
- There is always something you never saw before
- Mixture of deformable models? too many combined choices
- Bag of words? not enough structure
- Non-parametric? doesn't generalize

Structure variation

- Object in rich categories have variable structure



- These are NOT deformations
- There is always something you never saw before
- Mi ... ces
- Ba

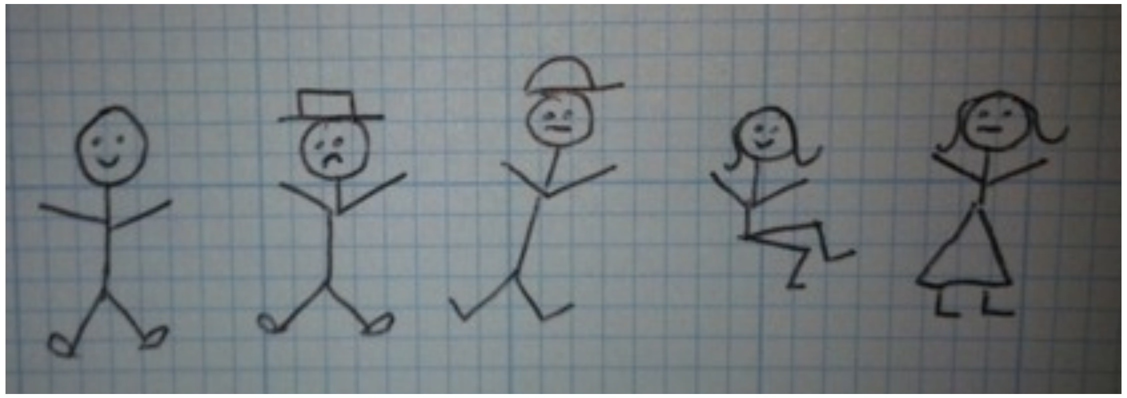
Compositional model

- Non-parametric? doesn't generalize

Object detection grammars

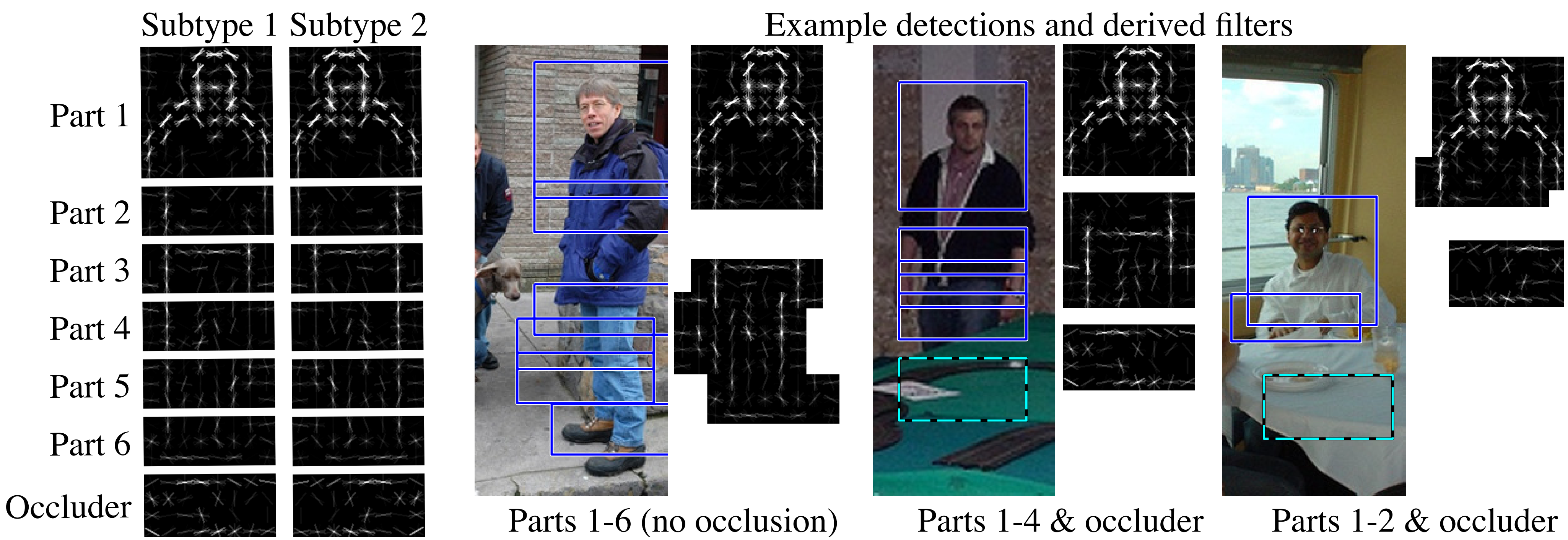
- Pictorial structure model with variable structure

- Stochastic context-free grammar
  - Generates tree-structured model
  - Springs connect symbols along derivation tree
  - Appearance model associated with each terminal



- person -> face, trunk, arms, lower-part
- face -> hat, eyes, nose, mouth
- face -> eyes, nose, mouth
- hat -> baseball-cap
- hat -> sombrero
- lower-part -> shoe, shoe, legs
- lower-part -> bare-foot, bare-foot, legs
- legs -> pants
- legs -> skirt

Person detection grammar



- Instantiation includes a variable number of parts
  - 1,...,k and occluder if k < 6
- Parts can translate relative to each other
- Parts have subtypes
- Parts have deformable sub-parts (not shown)
- Beats all other methods on PASCAL 2010 (49.5 AP)

Building the model

- Type in any non-recursive grammar

$$
\begin{aligned}
Q(\omega) &\xrightarrow{s_k} \{\ Y_1(\omega \oplus \delta_1), \ldots, Y_k(\omega \oplus \delta_k), O(\omega \oplus \delta_{k+1})\ \} \\
Q(\omega) &\xrightarrow{s_6} \{\ Y_1(\omega \oplus \delta_1), \ldots, Y_6(\omega \oplus \delta_6)\ \}
\end{aligned}
$$

$$
Y_p(\omega) \xrightarrow{0} \{\ Y_{p,t}(\omega)\ \}
$$

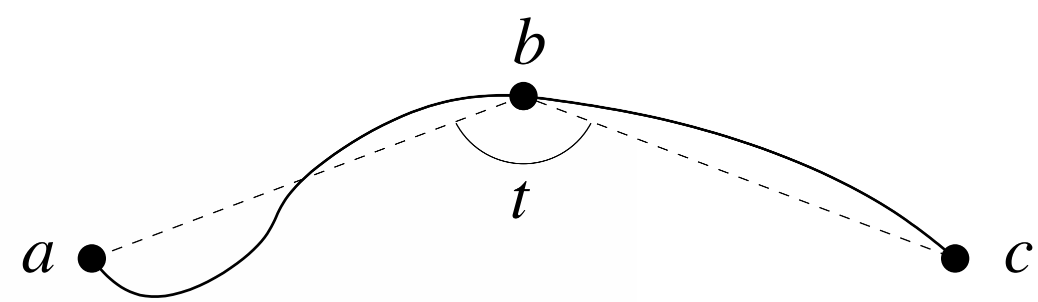
$$
O(\omega) \xrightarrow{0} \{\ O_t(\omega)\ \} \qquad O_t(\omega) \xrightarrow{\alpha_t \cdot \phi(\delta)} \{\ A_t(\omega \oplus \delta)\ \}
$$

$$
\begin{aligned}
Y_{p,t}(\omega) &\xrightarrow{\alpha_{p,t} \cdot \phi(\delta)} \{\ Z_{p,t}(\omega \oplus \delta)\ \} \\
Z_{p,t}(\omega) &\xrightarrow{0} \{A_{p,t}(\omega), W_{p,t,r,1}(\omega \oplus \delta_{p,t,r,1}), \ldots, W_{p,t,r,N_p}(\omega \oplus \delta_{p,t,r,N_p})\} \\
W_{p,t,r,u}(\omega) &\xrightarrow{\alpha_{p,t,r,u} \cdot \phi(\delta)} \{A_{p,t,r,u}(\omega \oplus \delta)\}
\end{aligned}
$$

- Train parameters from bounding box annotations
  - Production costs
  - Deformation models
  - HOG filters for terminals

Salient contours

- Curve(a,b) + Curve(b,c) --> Curve(a,c)



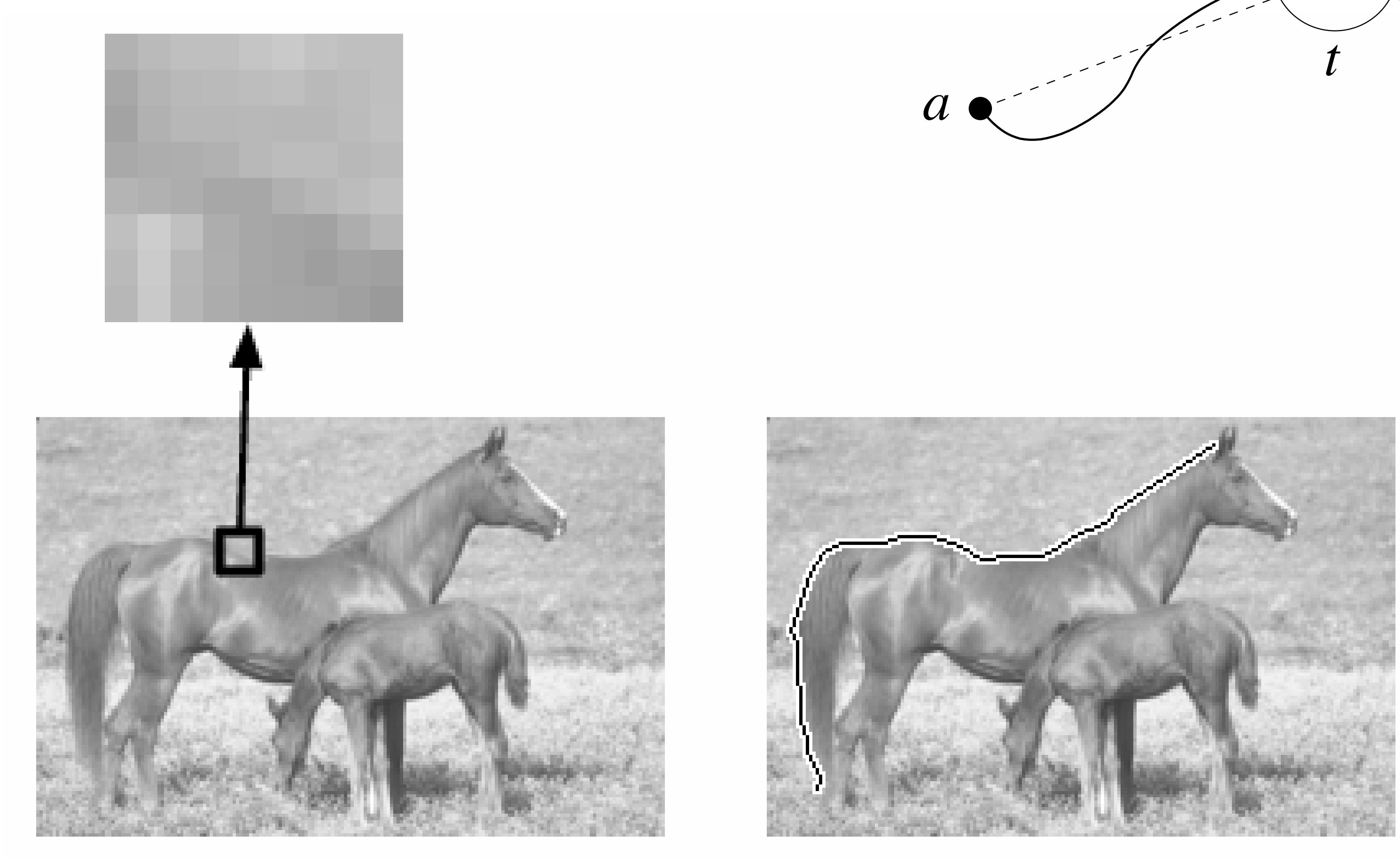
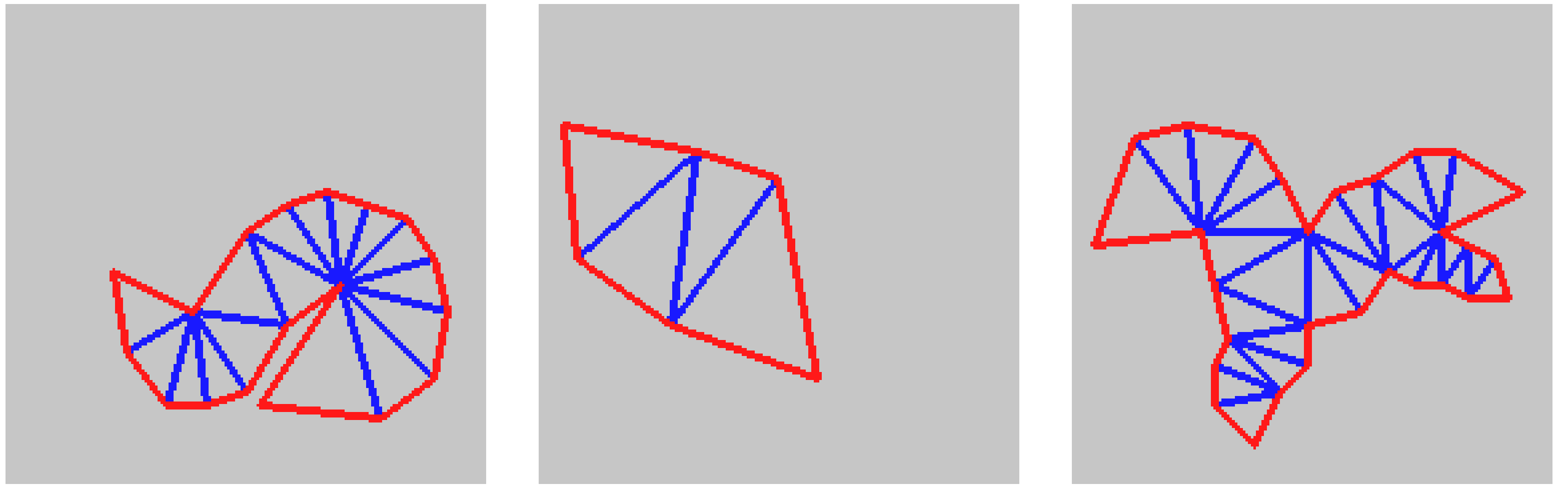


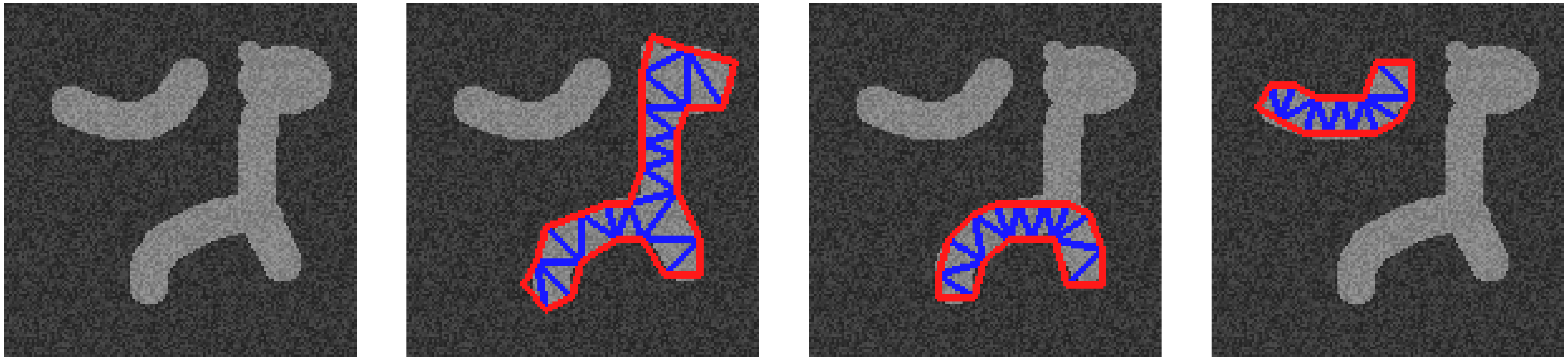
Figure 20: An example where the most salient curve goes over locations with essentially no local evidence for a the curve at those locations.

Shapes / Regions

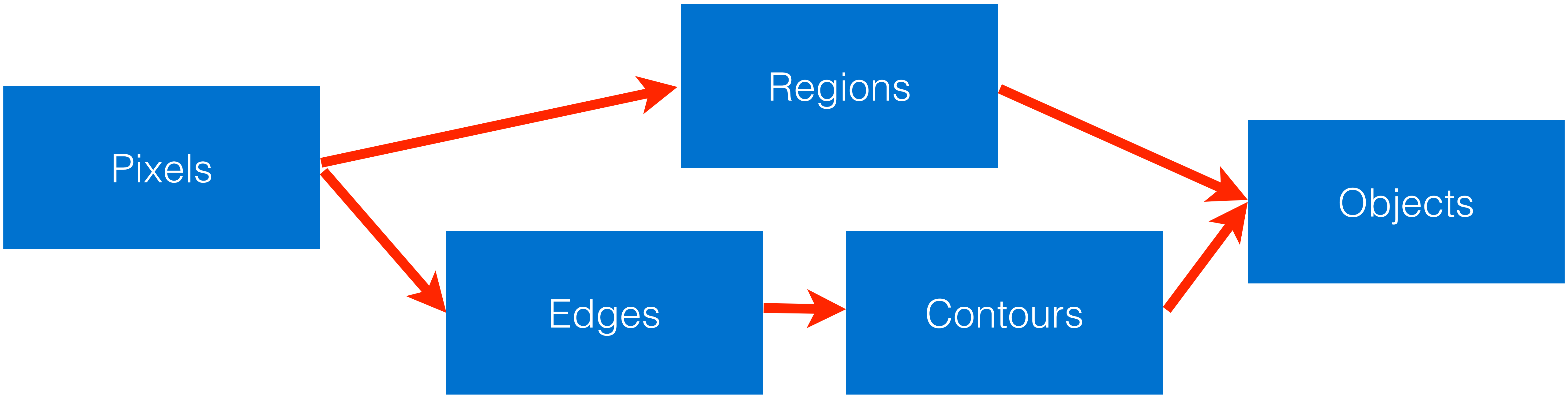
Samples from stochastic context-free shape grammar



"Matching" to images
(samples from posterior)



Processing pipeline

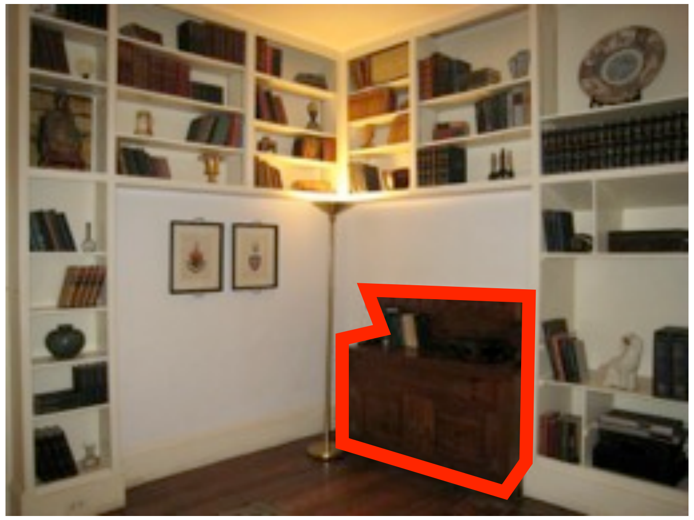


- Vision system have multiple processing stages
- Compositional model: each stage builds structures by grouping structures from previous stages
  - Single parsing problem
  - Avoids intermediate decisions
    (high-level information influences low-level interpretations)

Computation

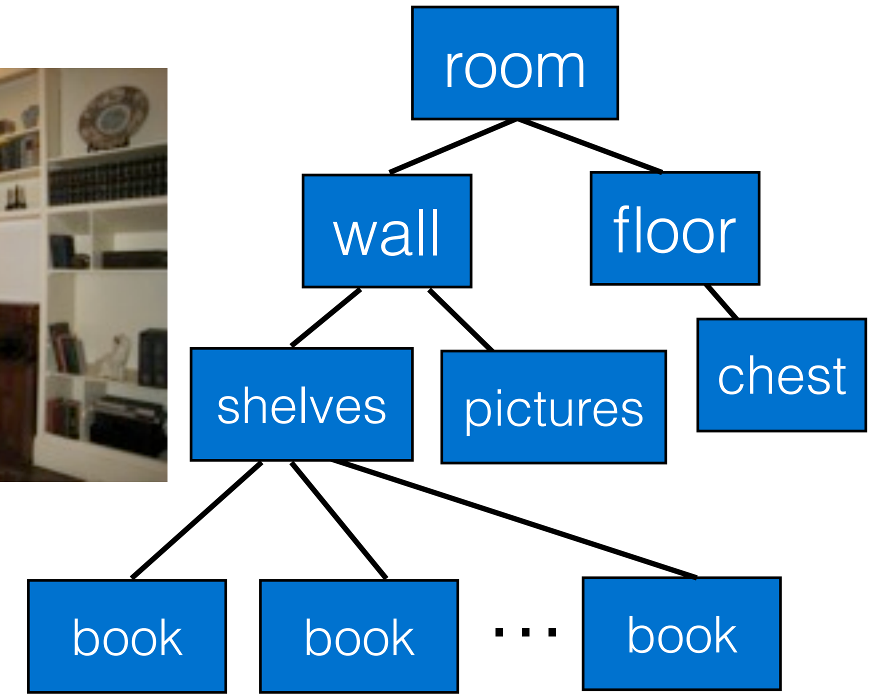
- Context-free or Context-sensitive?
- Even context-free models lead to hard parsing problem
  - Too many constituents!
  - String of length n have $O(n^2)$ substrings
  - Images with n pixels have $O(2^n)$ regions

G E T I K D S W O W Z Q E

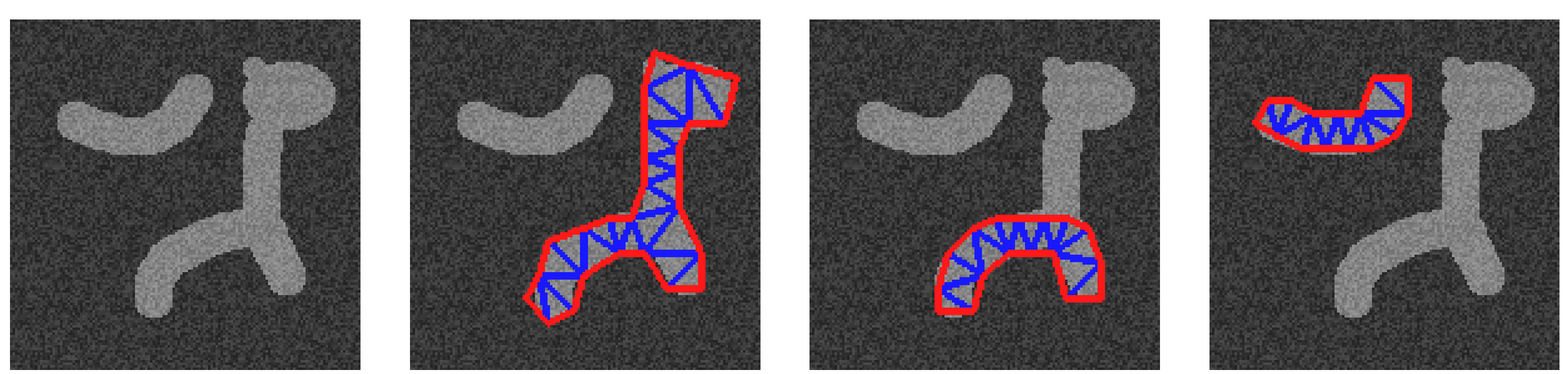


Alternative parsing problems

1. Whole image parsing
   - Explains every pixel exactly once
   - Hard



2. Find light derivations within an image
   - Expansion of start symbol into terminals
   - Explains part of the image
   - May explain the same pixel more then once
   - Efficient

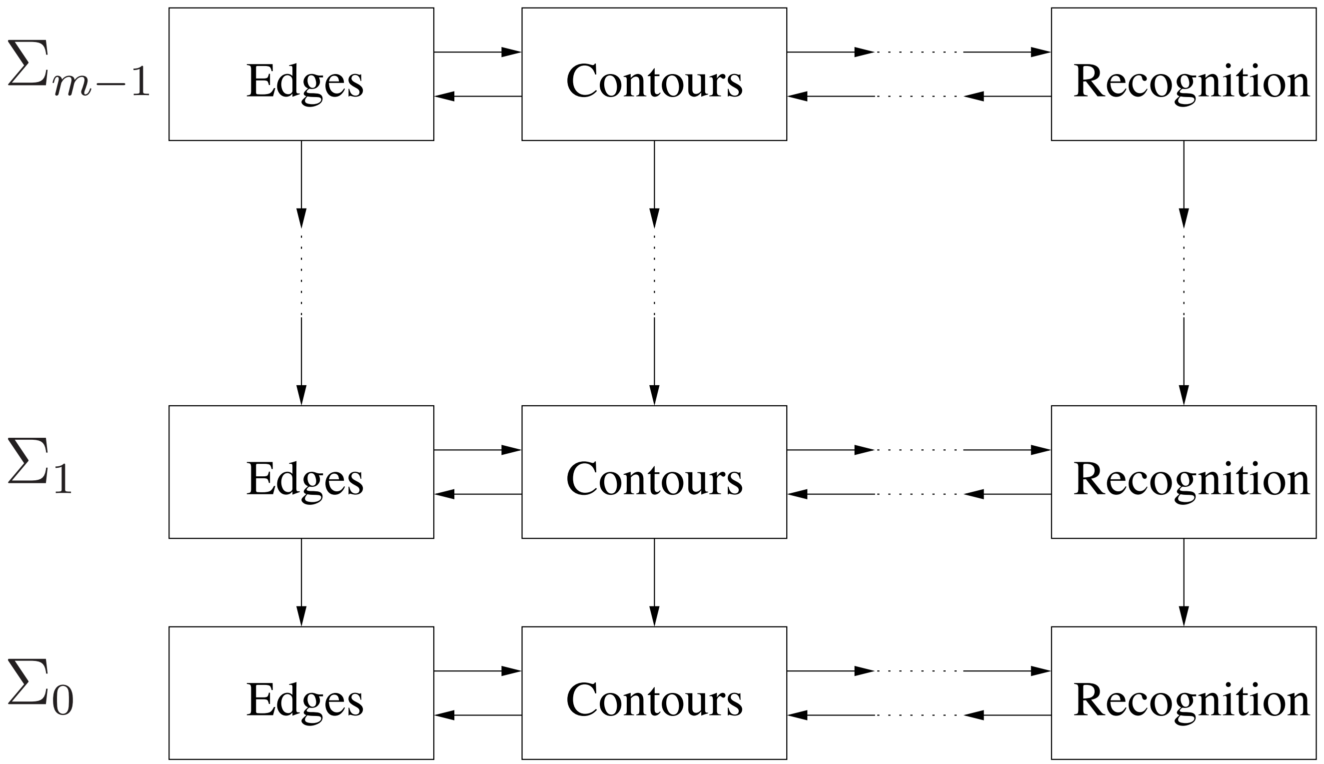


Computation

- Bottom-up
  - Repeated grouping structures (KLD / A*LD)
- Top-down
  - Repeated refining with backtracking (AO*)
- Bottom-up + Top-down
  - Bottom-up computation guided by top-down influence
  - Coarse derivations provide heuristic guidance for finding finer structures (HA*LD)

Coarse-to-fine

- Model abstraction f : S_i --> S_{i+1}
  - lower resolution
  - coarsen labels
    horse --> animal --> piecewise smooth object
- Coarse computation guides finer computation



Challenges

- Whole image parsing (with context-free grammars)
  - Restrict possible constituents
  - LP relaxation
  - DDMCMC
- Learn object grammars from weakly labeled data
  - PASCAL VOC
- Build a complete processing pipeline unifying segmentation and recognition