Object Category Detection: Parts-based Models

Computer Vision

CS 543 / ECE 549

University of Illinois

Derek Hoiem

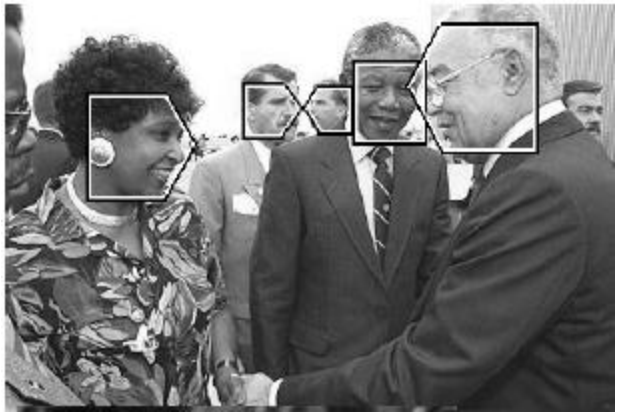
Administrative stuff

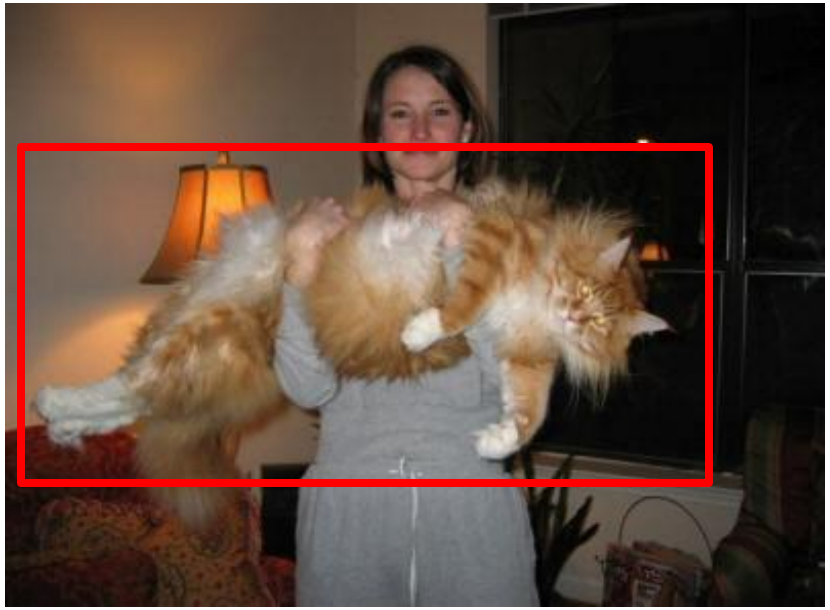
- HW 5 due Tues
- I'm out of town Sun to Tues night
  - Unresponsive on Monday
  - Amin Sadeghi will teach Tues
- Final project posters presented May 4 (1:30-4:30pm)
  - Final paper: due May 8

Goal: Detect all instances of objects

Cars

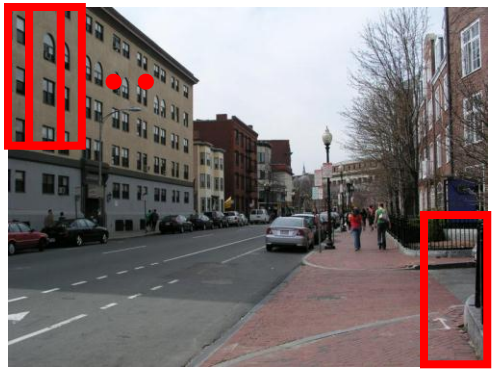
Faces





Cats

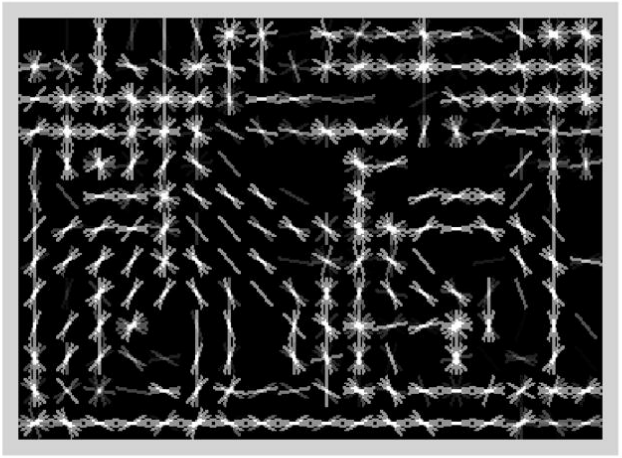
Last class: sliding window detection



Object model: last class

- Statistical Template in Bounding Box
  - Object is some (x,y,w,h) in image
  - Features defined wrt bounding box coordinates

Image



Template Visualization

Images from Felzenszwalb

Last class: statistical template

- Object model = log linear model of parts at fixed positions

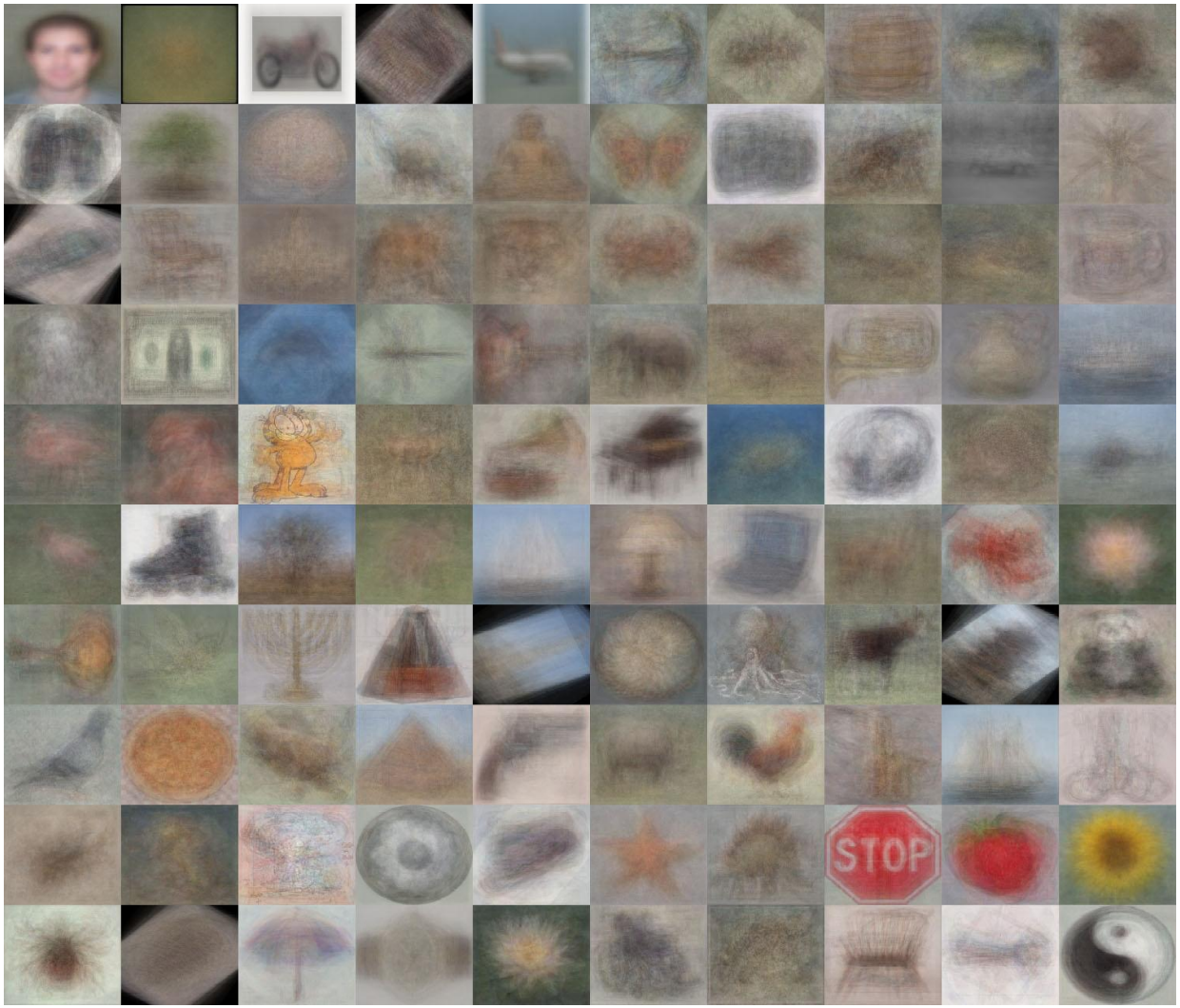
+3 +2 -2 -1 -2.5 = -0.5 ? > 7.5

Non-object

+4 +1 +0.5 +3 +0.5 = 10.5 ? > 7.5

Object

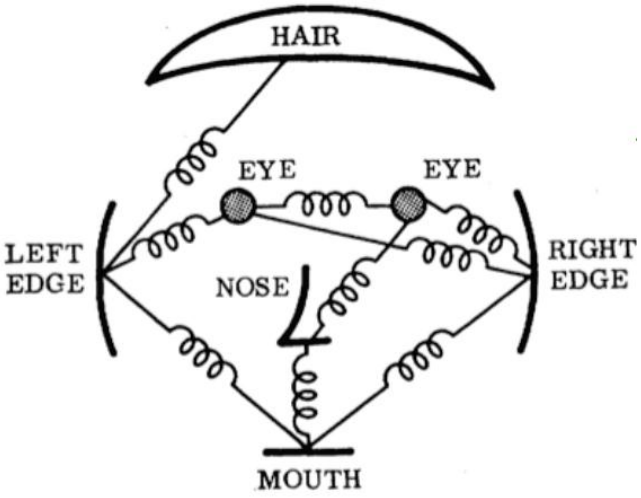
When do statistical templates make sense?

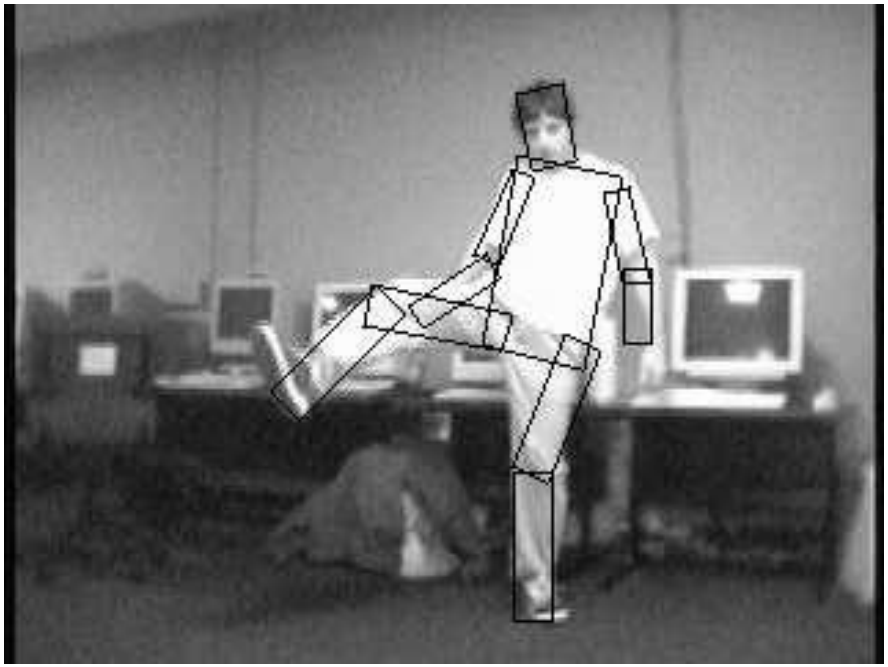


Caltech 101 Average Object Images

Object models: this class

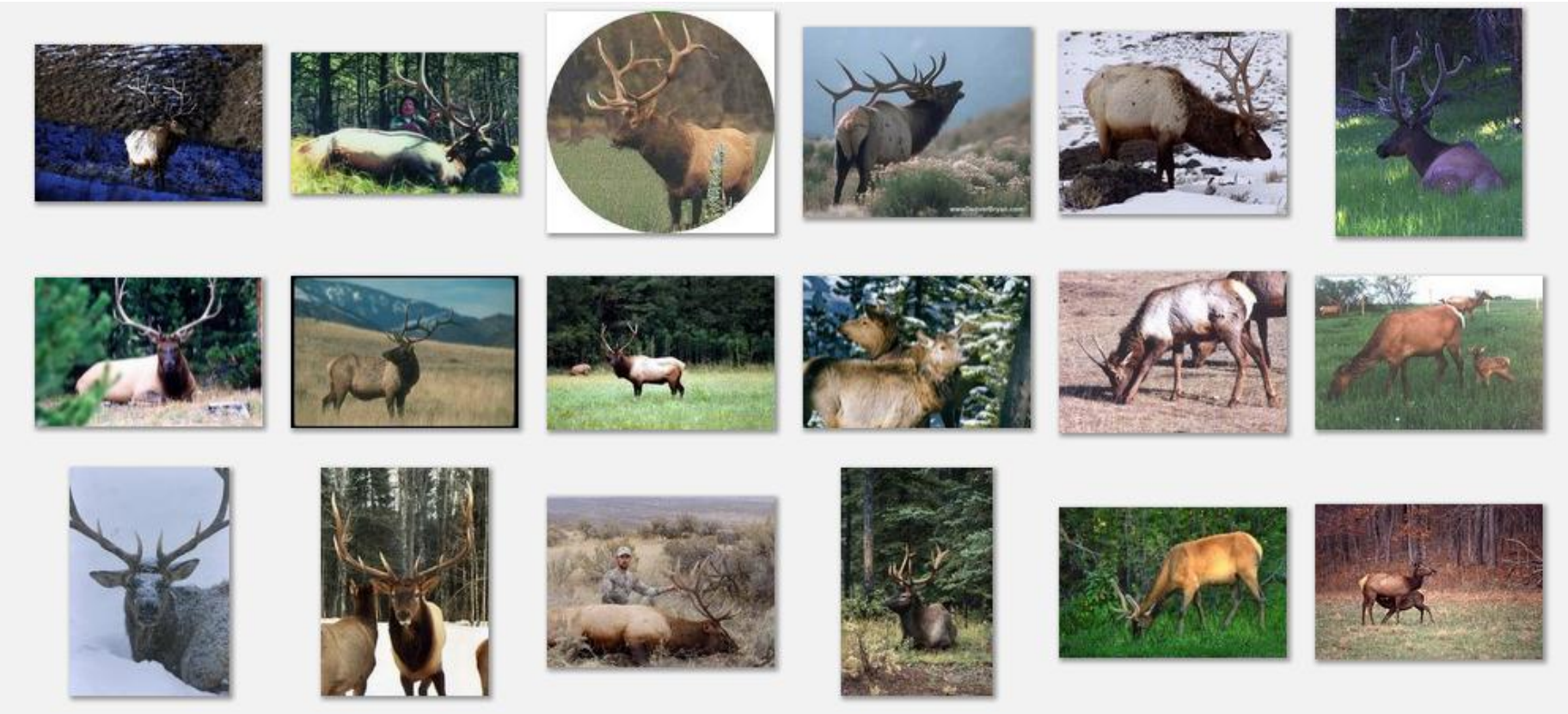
- Articulated parts model
  - Object is configuration of parts
  - Each part is detectable





Images from Felzenszwalb

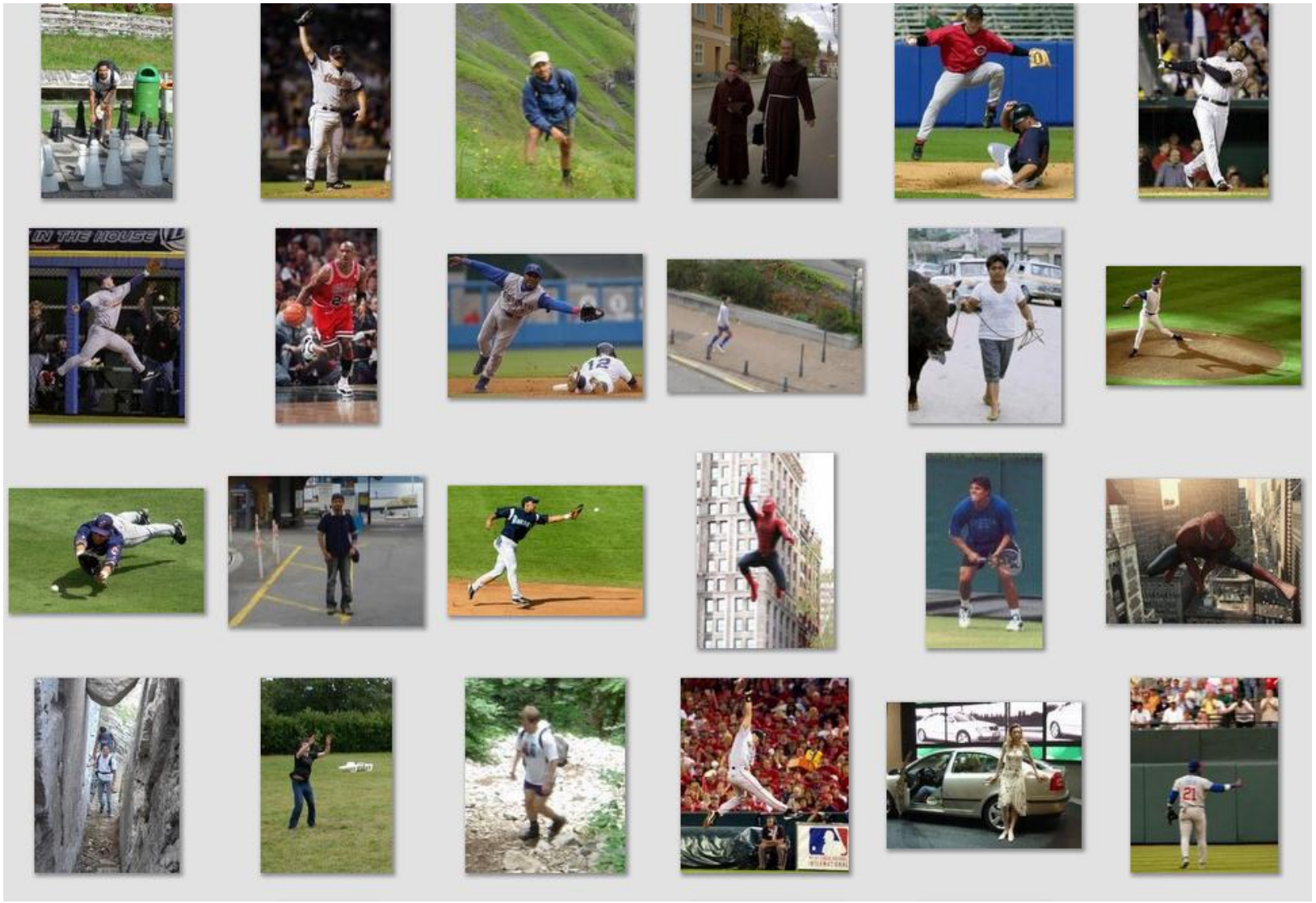
Deformable objects



Images from Caltech-256

Slide Credit: Duan Tran

Deformable objects



Images from D. Ramanan's dataset

Slide Credit: Duan Tran

Compositional objects

Parts-based Models

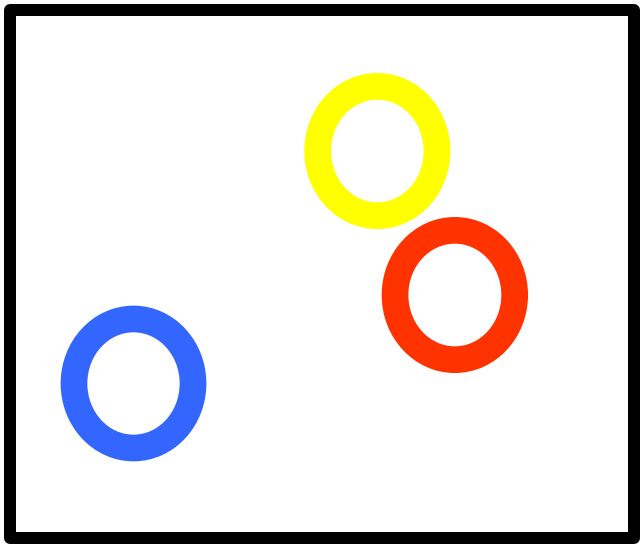
Define object by collection of parts modeled by

1. Appearance
2. Spatial configuration

Slide credit: Rob Fergus

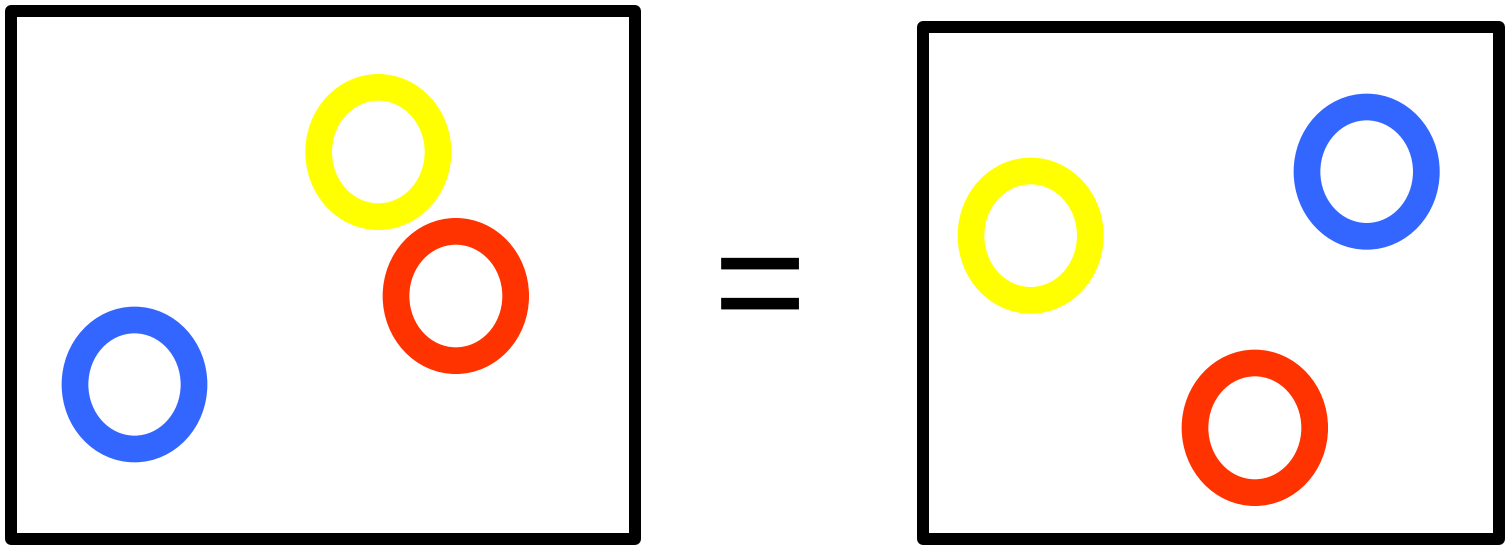
How to model spatial relations?

- One extreme: fixed template



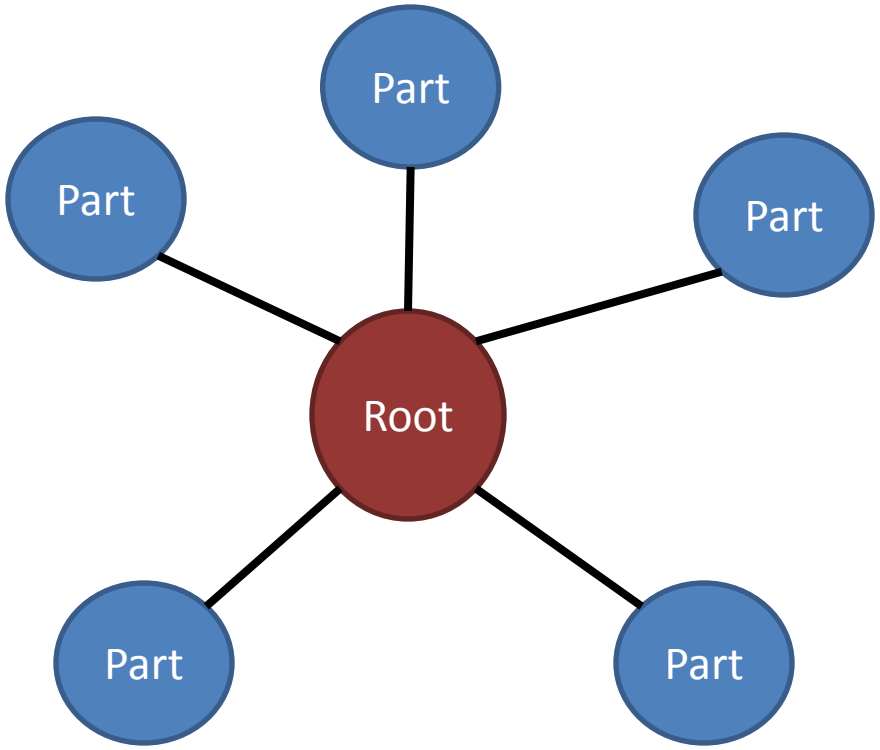
How to model spatial relations?

- Another extreme: bag of words



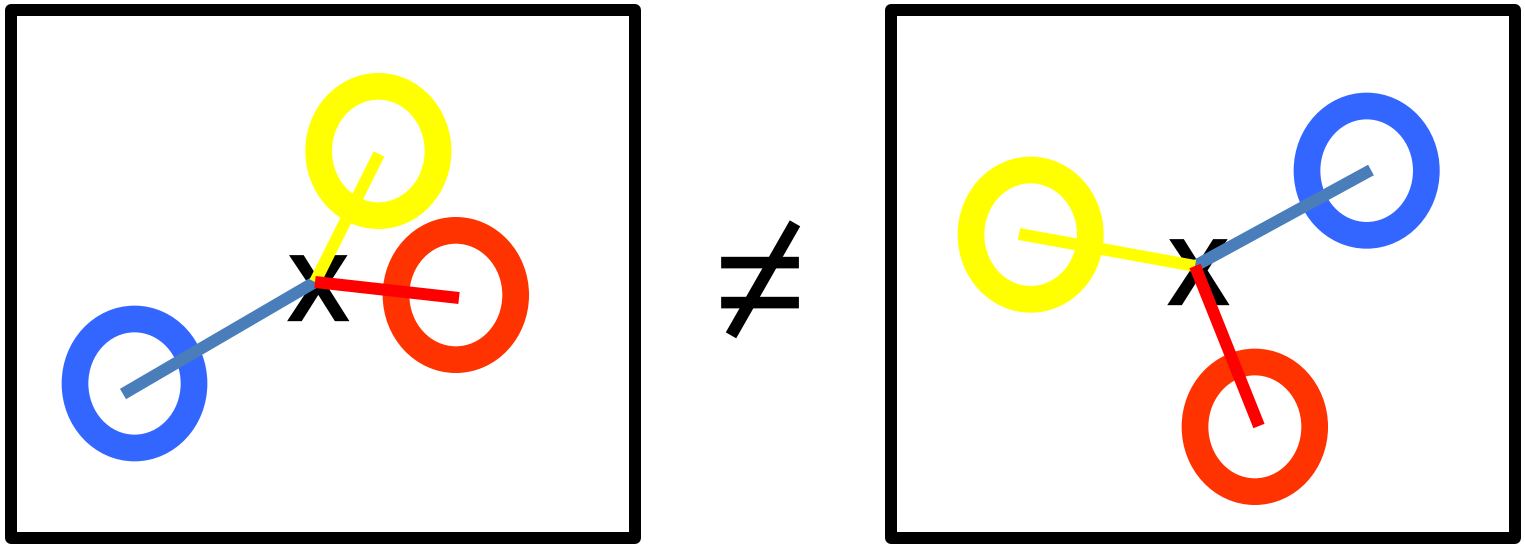
How to model spatial relations?

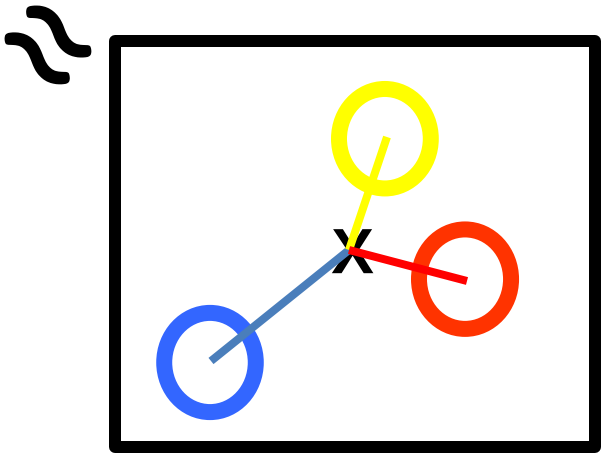
- Star-shaped model

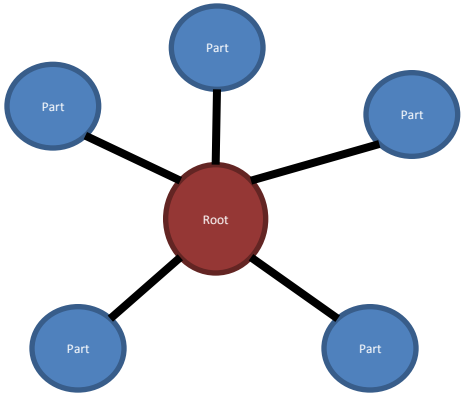


How to model spatial relations?

- Star-shaped model

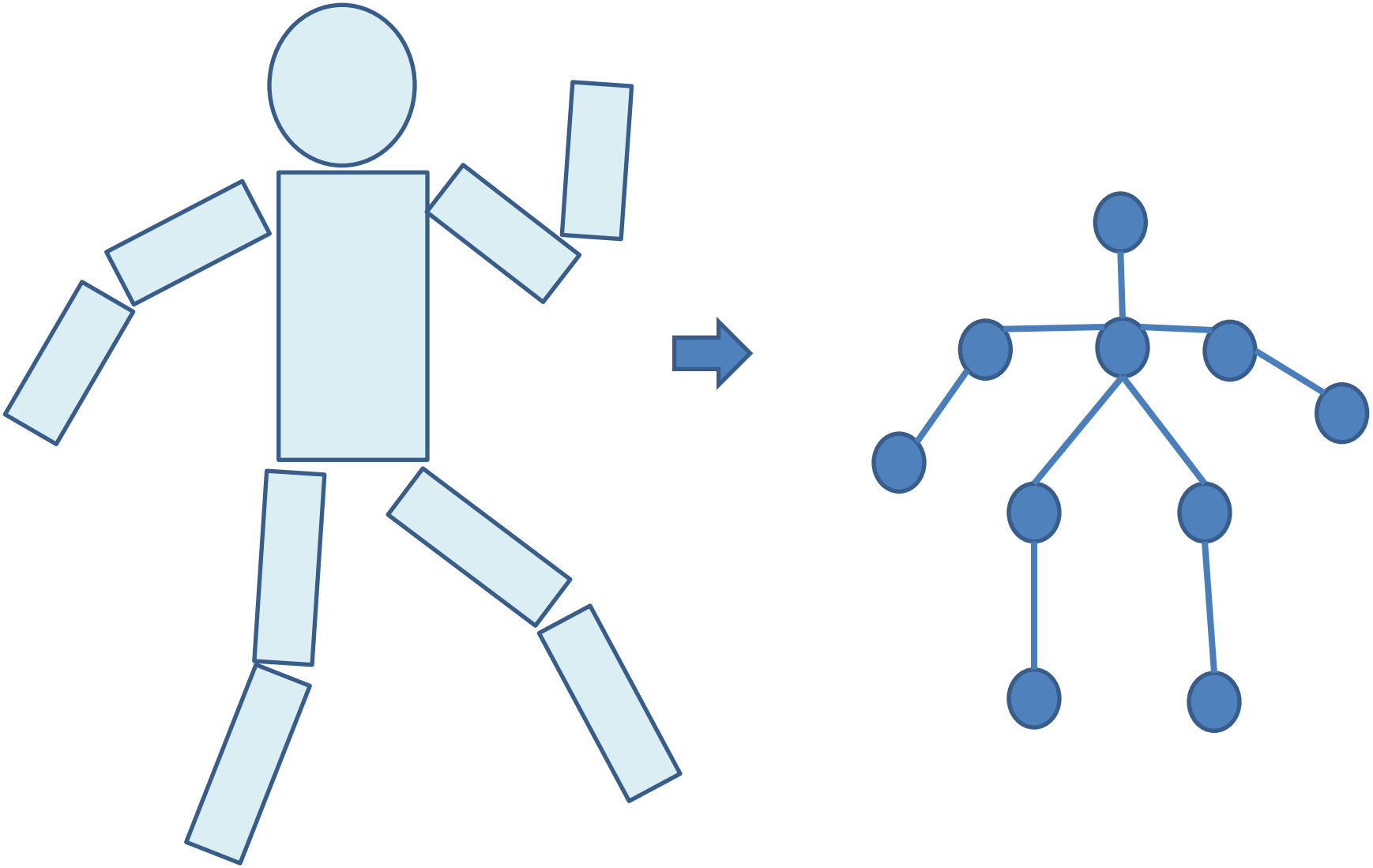






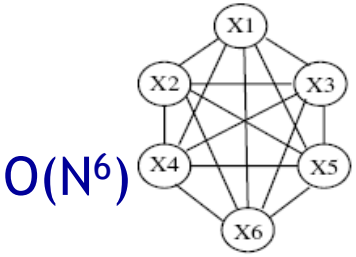
How to model spatial relations?

- Tree-shaped model



How to model spatial relations?

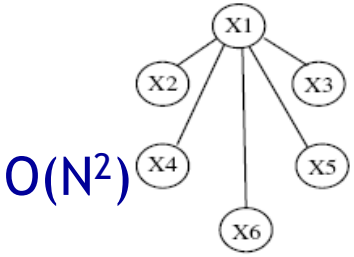
- Many others…



O(N⁶)

a) Constellation

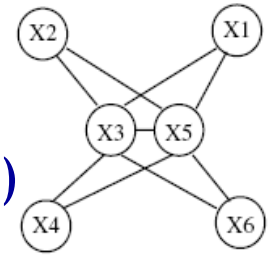
Fergus et al. ’03
Fei-Fei et al. ‘03



O(N²)

b) Star shape

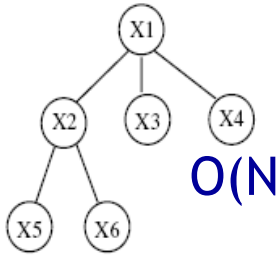
Leibe et al. ’04, ‘08
Crandall et al. ‘05
Fergus et al. ’05



O(N³)

c) k-fan ($k = 2$)

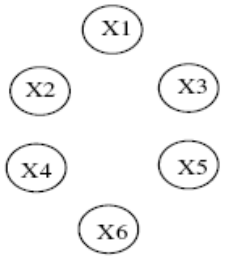
Crandall et al. ‘05



O(N²)

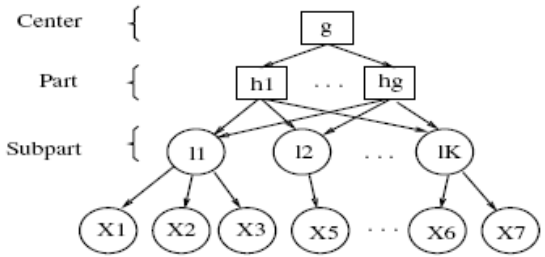
d) Tree

Felzenszwalb & Huttenlocher ‘05



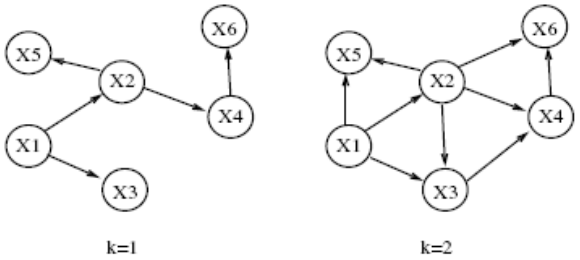
e) Bag of features

Csurka ’04
Vasconcelos ‘00



f) Hierarchy

Bouchard & Triggs ‘05



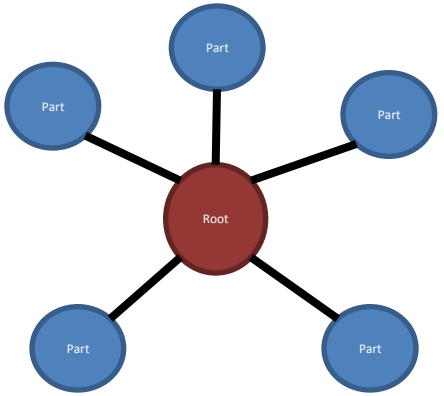
g) Sparse flexible model

Carneiro & Lowe ‘06

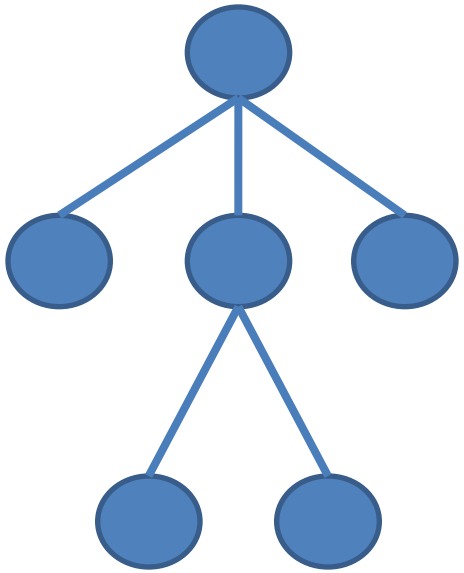
from [Carneiro & Lowe, ECCV’06]

Today's class

1. Star-shaped model
   - Example: ISM
     - Leibe et al. 2004, 2008



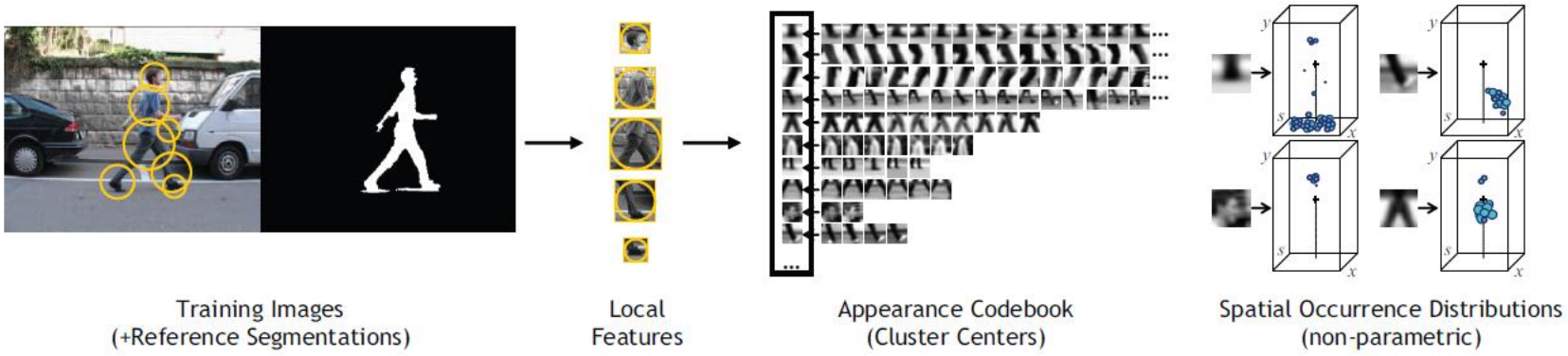
2. Tree-shaped model
   - Example: Pictorial structures
     - Felzenszwalb Huttenlocher 2005



ISM: Implicit Shape Model

Training overview

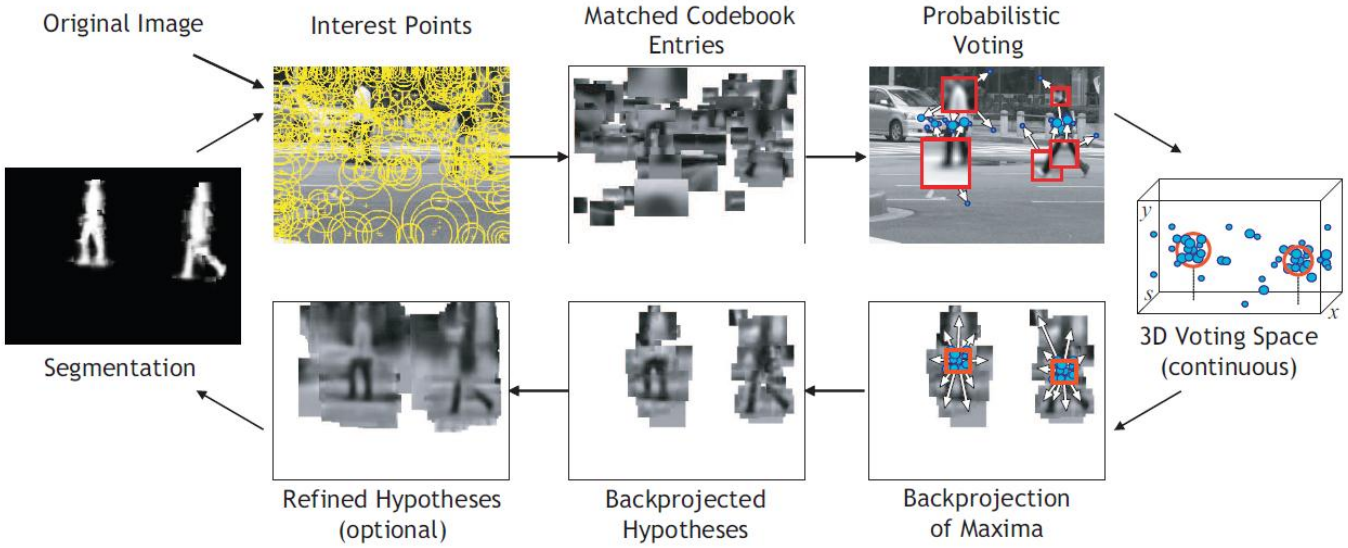
- Start with bounding boxes and (ideally) segmentations of objects
- Extract local features (e.g., patches or SIFT) at interest points on objects
- Cluster features to create codebook
- Record relative bounding box and segmentation for each codeword



ISM: Implicit Shape Model

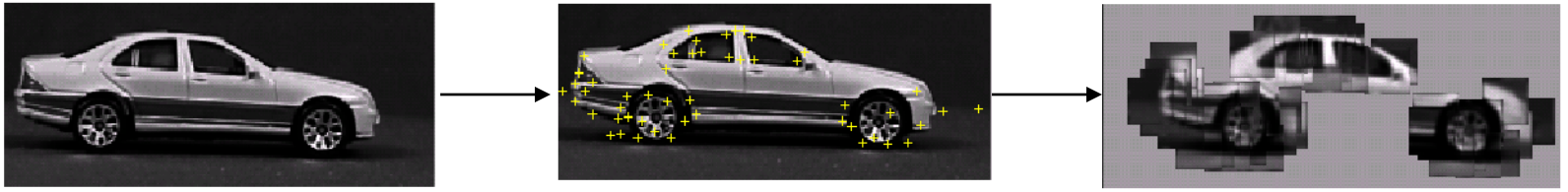
Testing overview

- Extract interest points in test image
- Softly match to codebook entries
- Each matched codeword votes for object bounding box
- Compute modes of votes using mean-shift
- Check which codewords voted for modes
- Refine

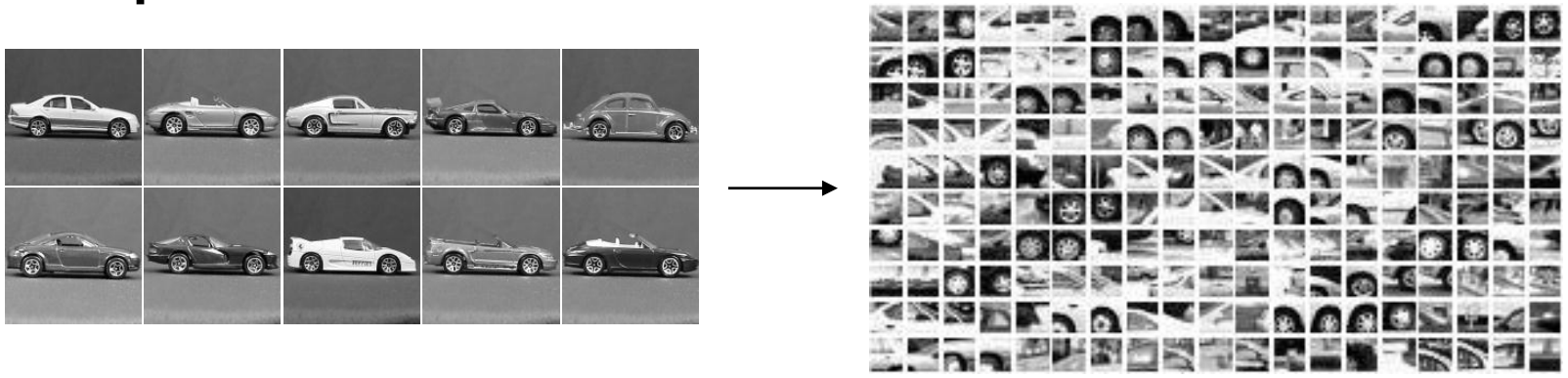


Codebook Representation

- Extraction of local object features
  - Interest Points (e.g. Harris detector)
  - Sparse representation of the object appearance



- Collect features from whole training set
- Example:



K. Grauman, B. Leibe

Agglomerative Clustering

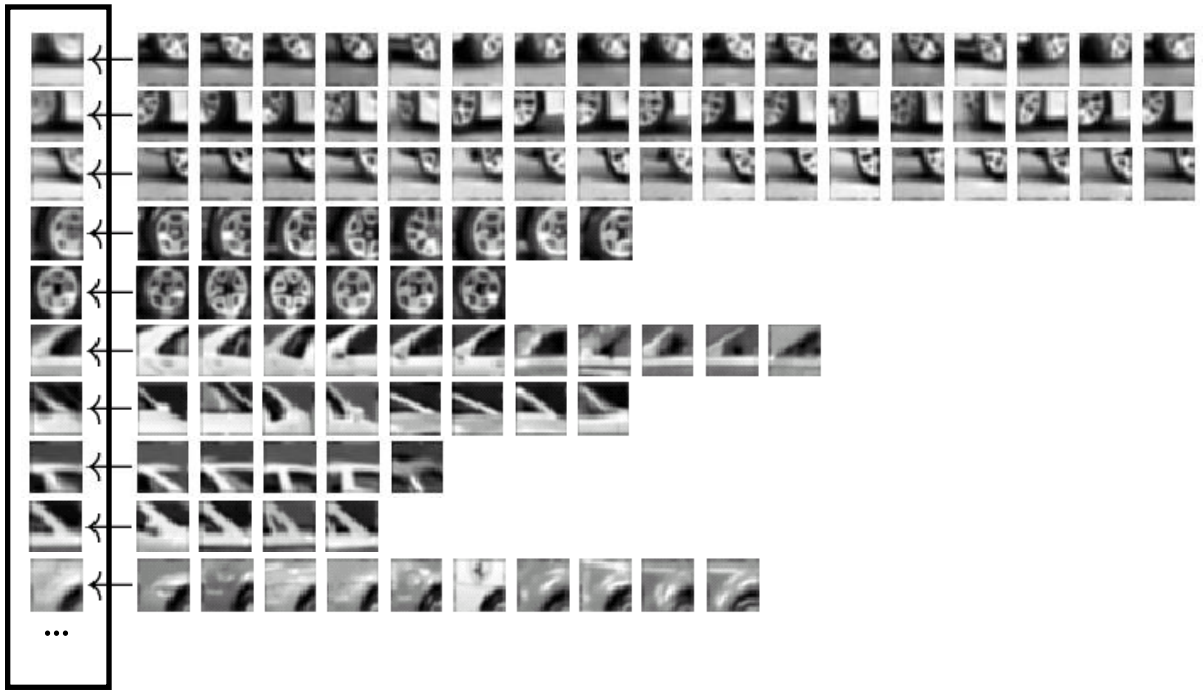
- **Algorithm (Average-Link)**
  1. **Start with each patch as a cluster of its own**
  2. **Repeatedly merge the two most similar clusters X and Y, where the similarity between two clusters is defined as the average similarity between their members**

$$\mathrm{sim}(X,Y) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\mathrm{sim}(x^{(i)}, y^{(j)})$$

  3. **Until** $\mathrm{sim}(X,Y) < \theta$

- **Commonly used similarity measures**
  - **Normalized correlation**
  - **Euclidean distances**

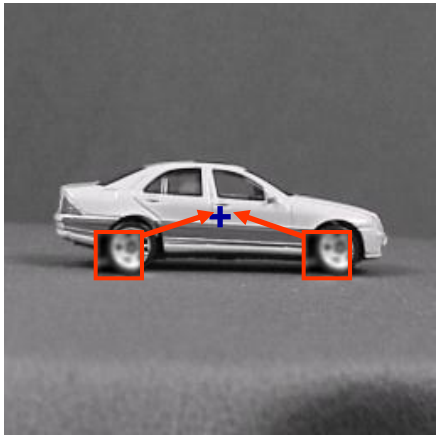
Appearance Codebook



- **Clustering Results**
  - Visual similarity preserved
  - Wheel parts, window corners, fenders, ...
  - Store cluster centers as *Appearance Codebook*

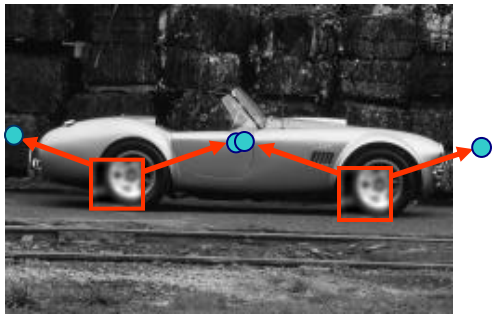
Voting with Local Features

- For every feature, store possible "occurrences"



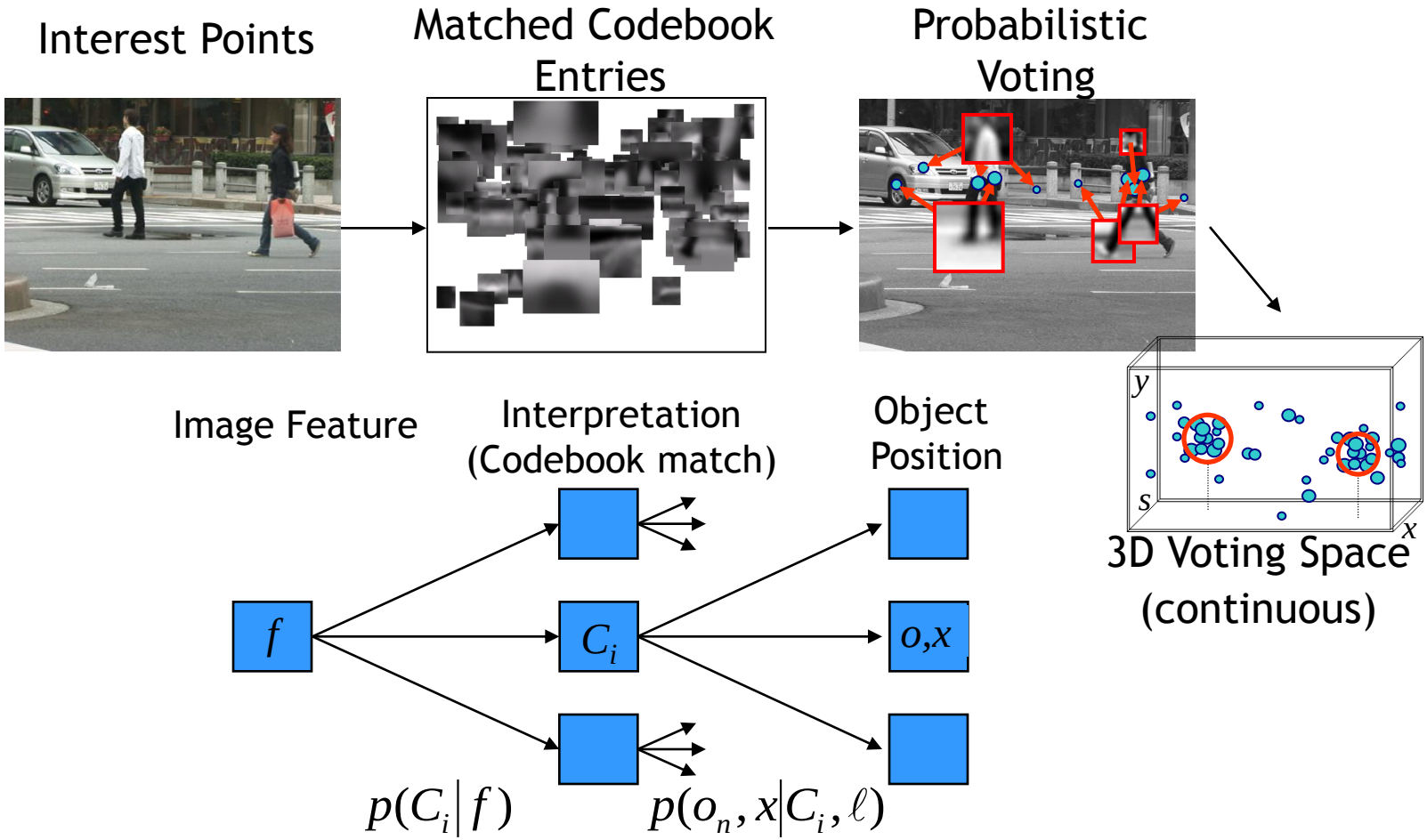
Record relative size and scale of object

- For new image, let the matched features vote for possible object positions



$$x_{vote} = x_{img} - x_{occ}(s_{img}/s_{occ})$$
$$y_{vote} = y_{img} - y_{occ}(s_{img}/s_{occ})$$
$$s_{vote} = (s_{img}/s_{occ}).$$

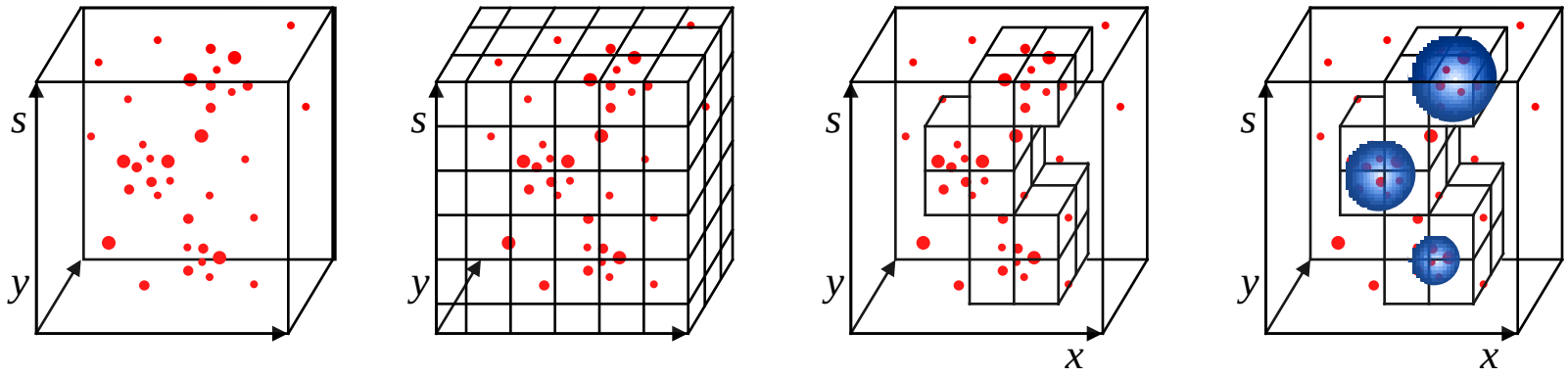
Implicit Shape Model - Recognition



$$p(o_n, x|f, \ell) = \sum_i p(C_i|f)\, p(o_n, x|C_i, \ell)$$

[Leibe04, Leibe08]

Scale Voting: Efficient Computation

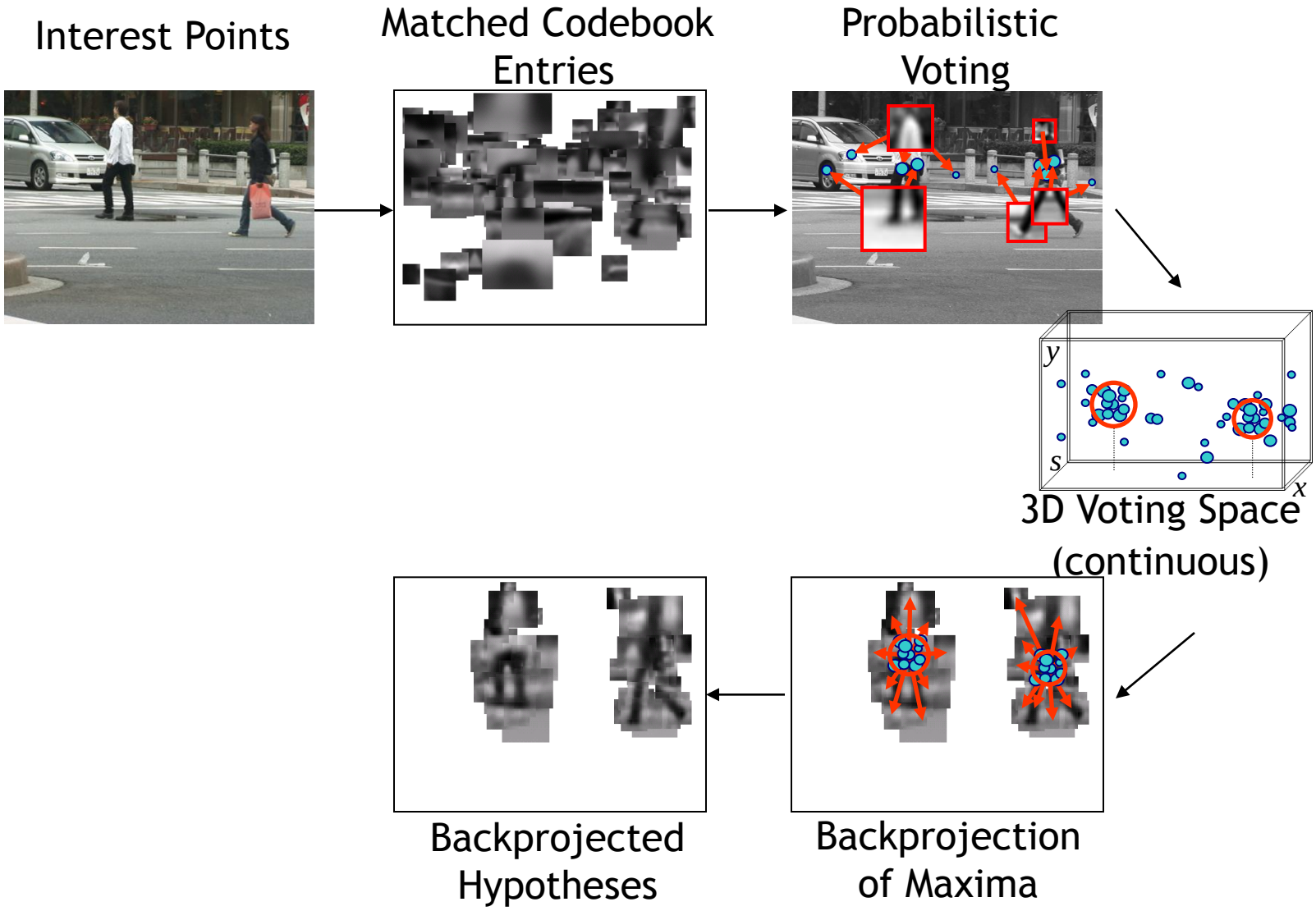


Scale votes | Binned accum. array | Candidate maxima | Refinement

- **Mean-Shift formulation for refinement**
  - **Scale-adaptive *balloon density estimator***

$$\hat{p}(o_n, x) = \frac{1}{V_b} \sum_k \sum_j p(o_n, x_j | f_k, \ell_k) K\left(\frac{x - x_j}{b}\right)$$

Implicit Shape Model - Recognition

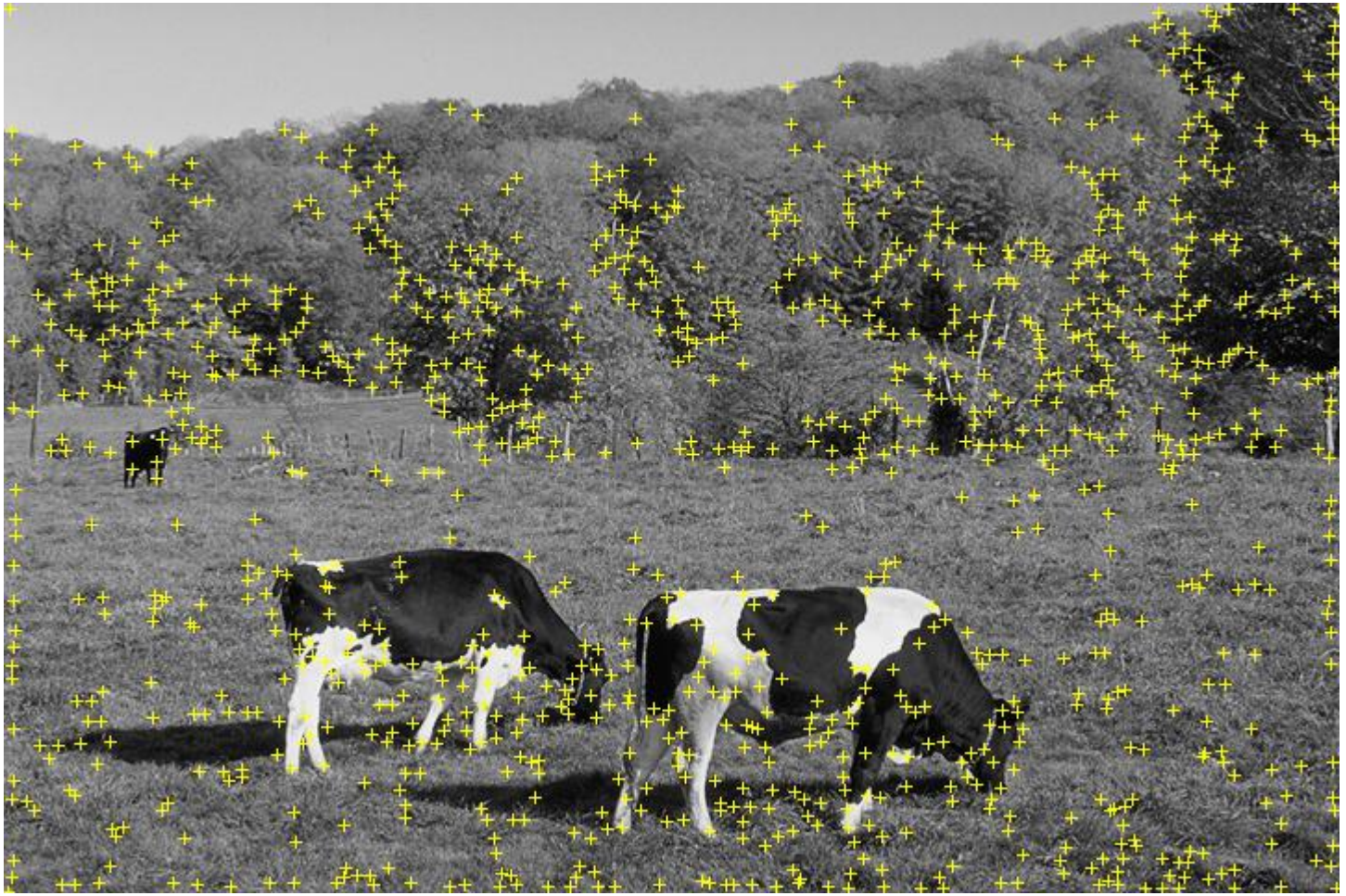


[Leibe04, Leibe08]

Example: Results on Cows

Original image

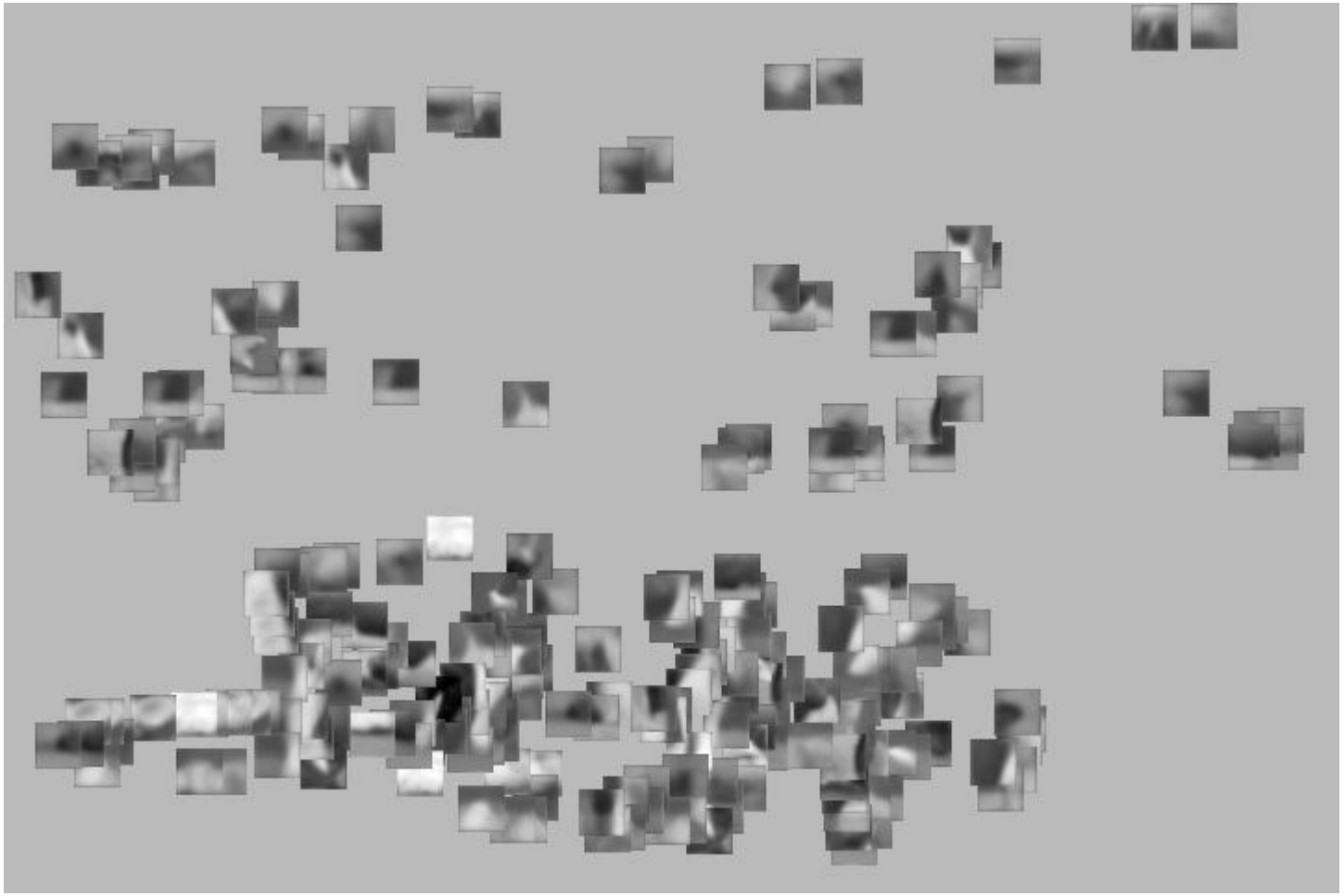
Example: Results on Cows



Interest points

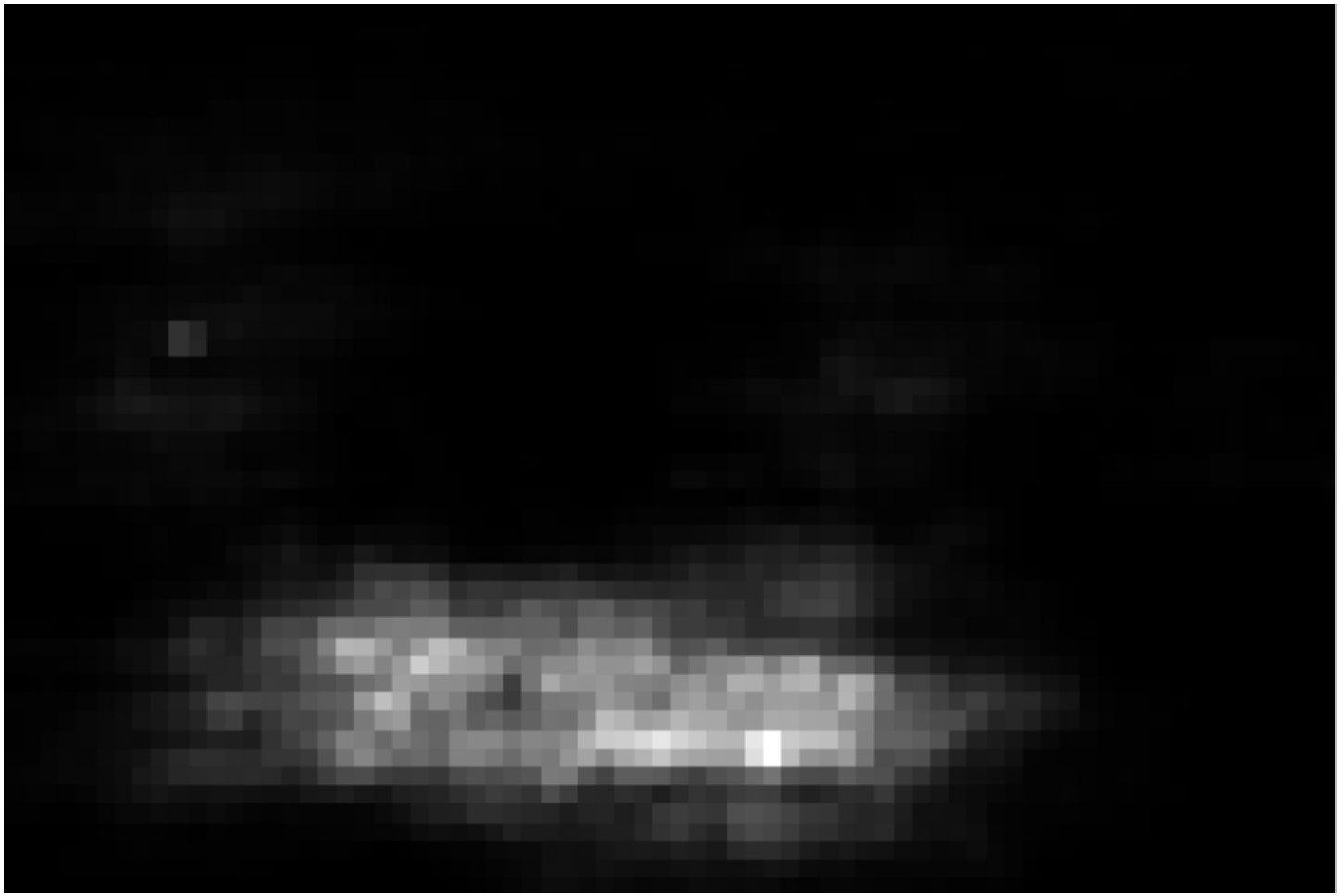
K. Grauman, B. Leibe

Example: Results on Cows



Matched patches

Example: Results on Cows



Prob. Votes

Example: Results on Cows

1st hypothesis

Example: Results on Cows

2nd hypothesis

Example: Results on Cows

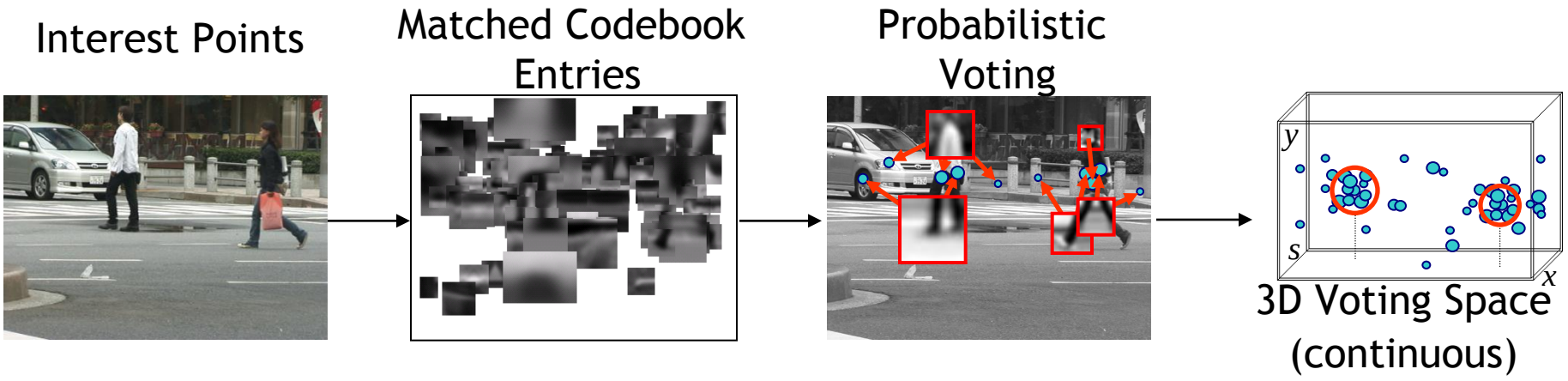
3rd hypothesis

ISM: Detection Results

- Qualitative Performance
  - Robust to clutter, occlusion, noise, low contrast

K. Grauman, B. Leibe

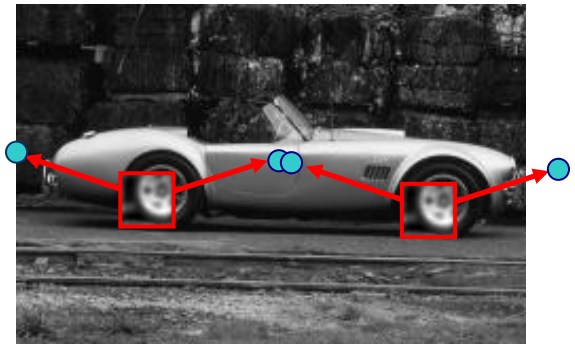
Beyond bounding boxes

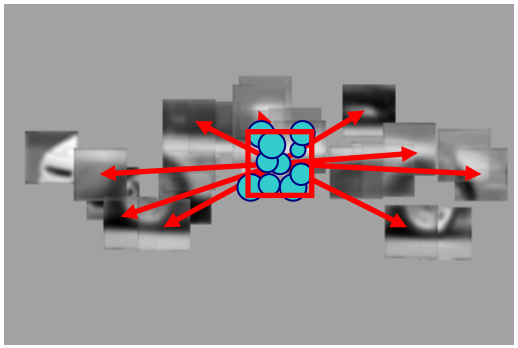


Backprojected codewords can vote:

- Pixel segmentation
- Part layout
- Pose
- Depth values

Segmentation: Probabilistic Formulation





- **Influence of patch on object hypothesis (vote weight)**

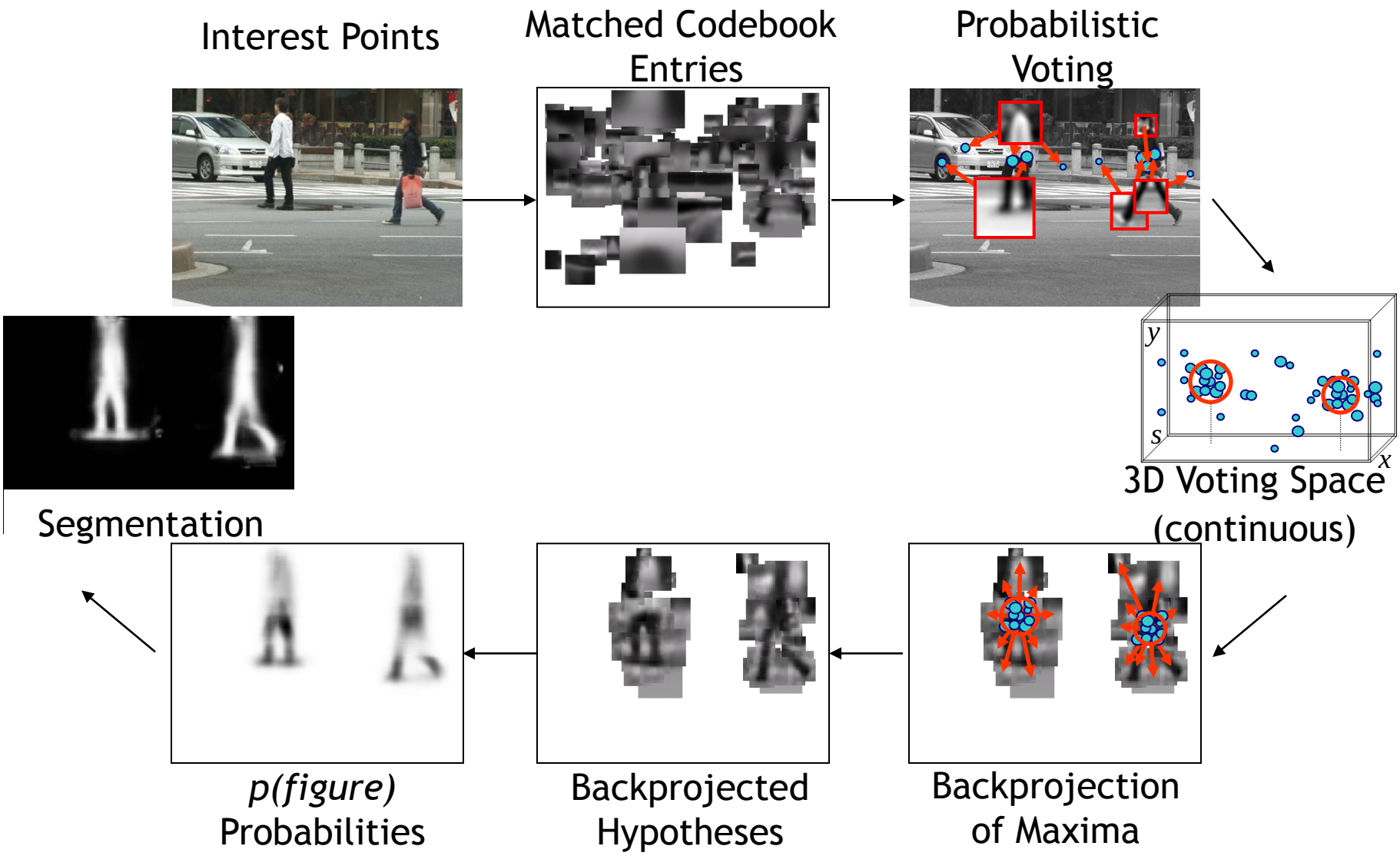
$$p(f,\ell|o_n,x)=\frac{\sum_i p(o_n,x|C_i)p(C_i|f)p(f,\ell)}{p(o_n,x)}$$

- **Backprojection to features f and pixels $\mathbf{p}$:**

$$p(\mathbf{p}=figure|o_n,x)=\sum_{\mathbf{p}\in(f,\ell)} \underbrace{p(\mathbf{p}=figure|f,\ell,o_n,x)}_{\text{Segmentation information}} \underbrace{p(f,\ell|o_n,x)}_{\text{Influence on object hypothesis}}$$

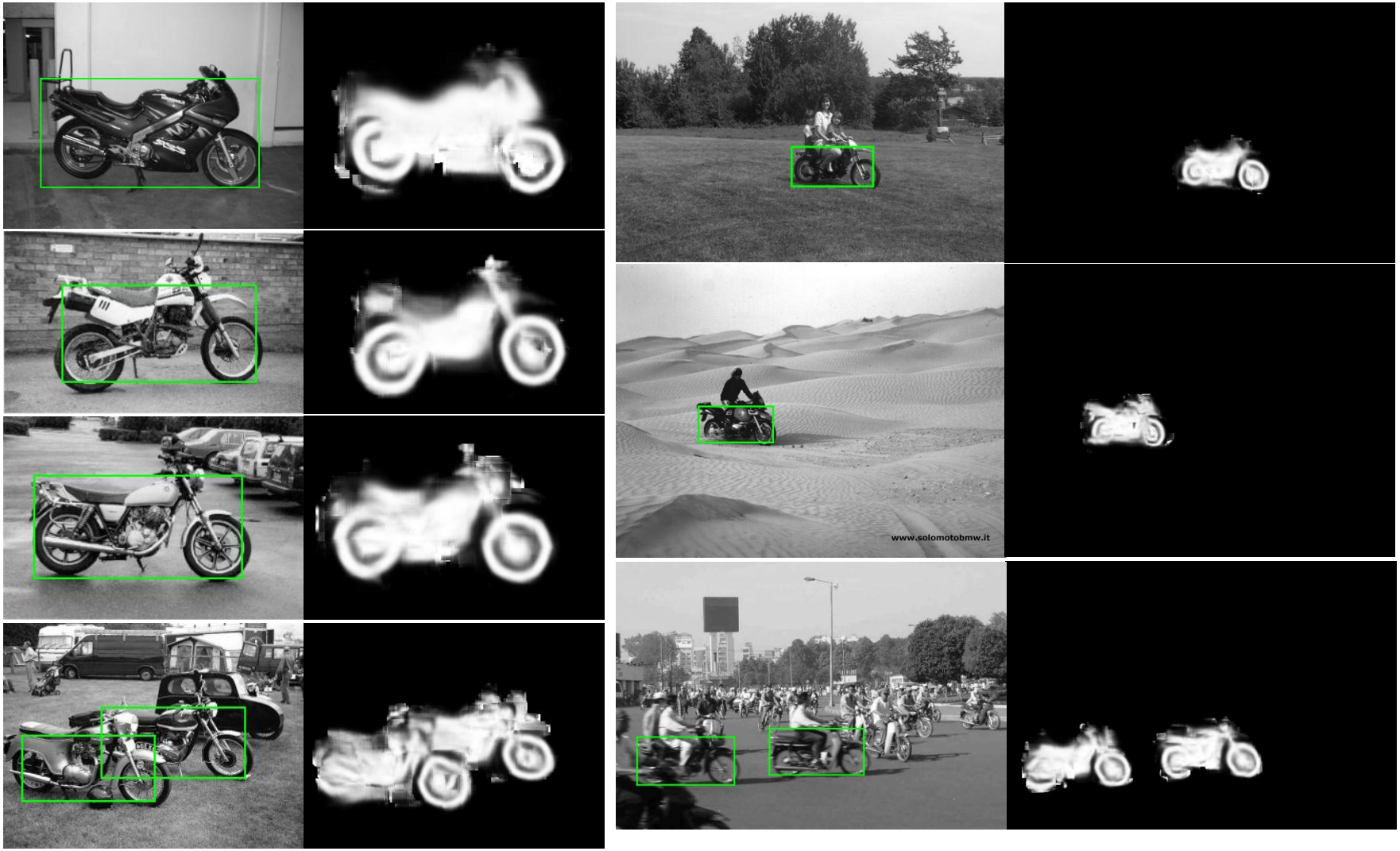
[Leibe04, Leibe08]

ISM – Top-Down Segmentation



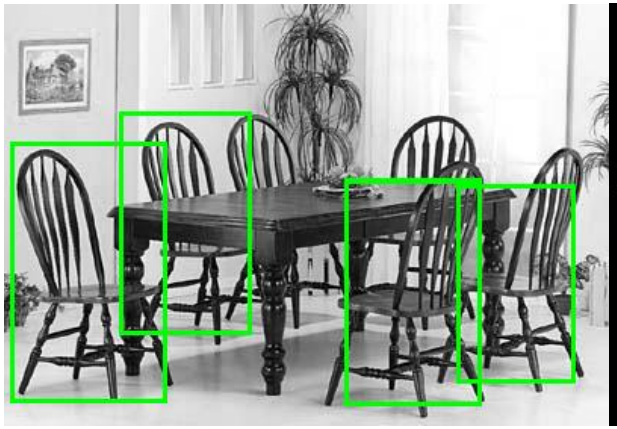
[Leibe04, Leibe08]

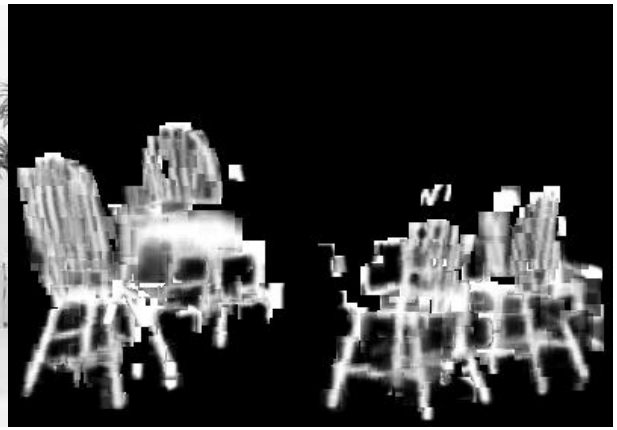
Example Results: Motorbikes



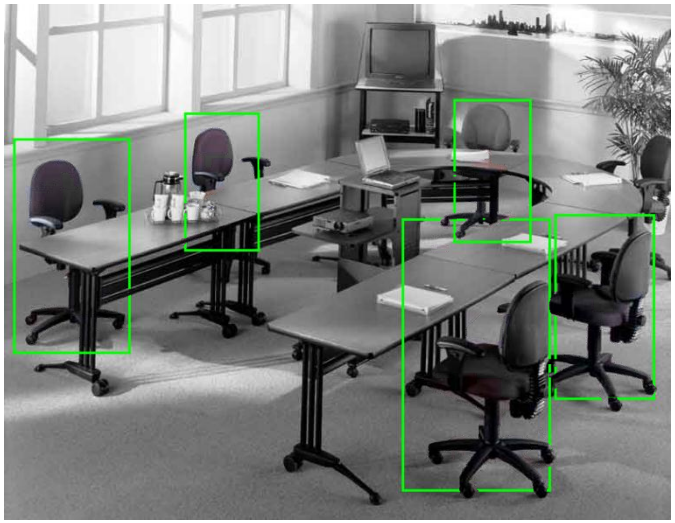
K. Grauman, B. Leibe

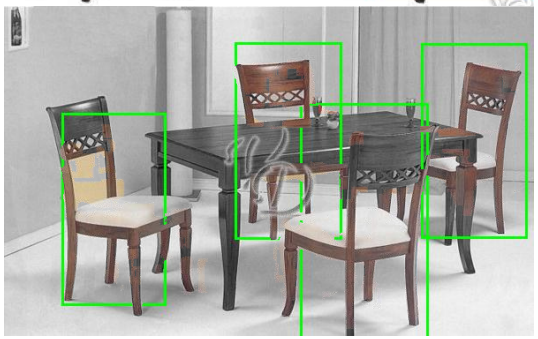
Example Results: Chairs

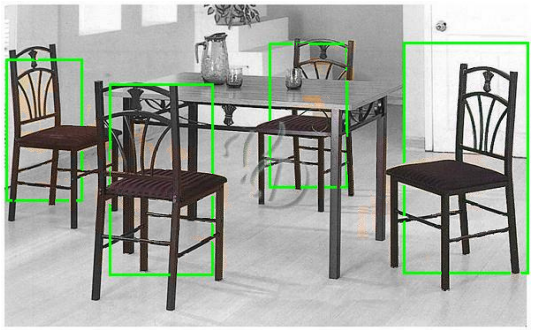




Dining room chairs

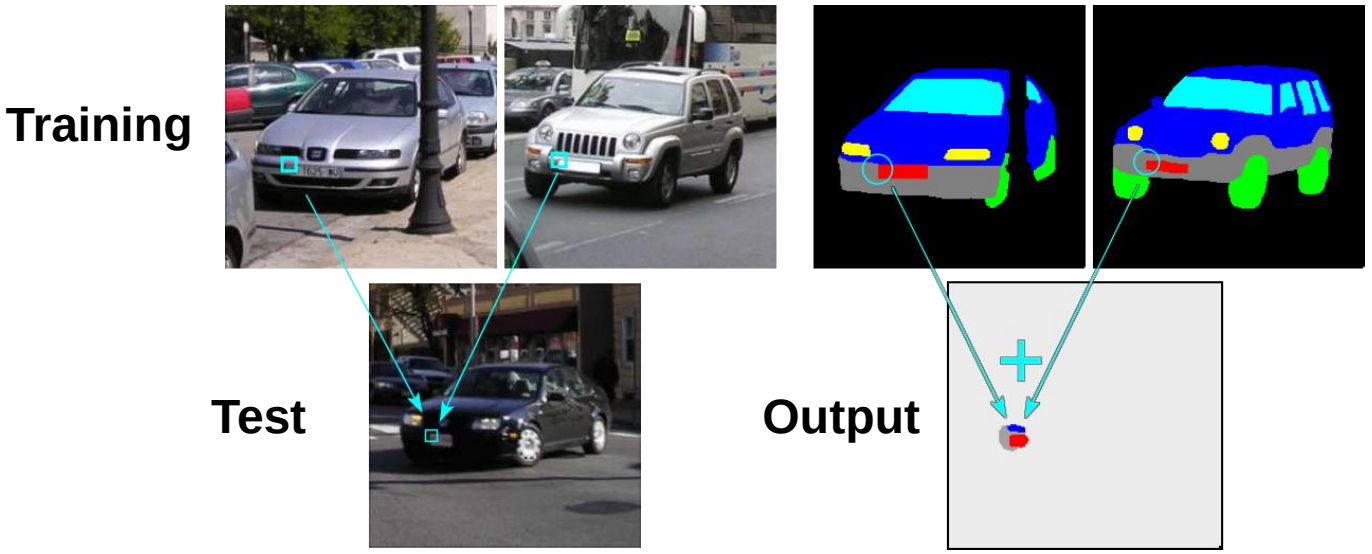






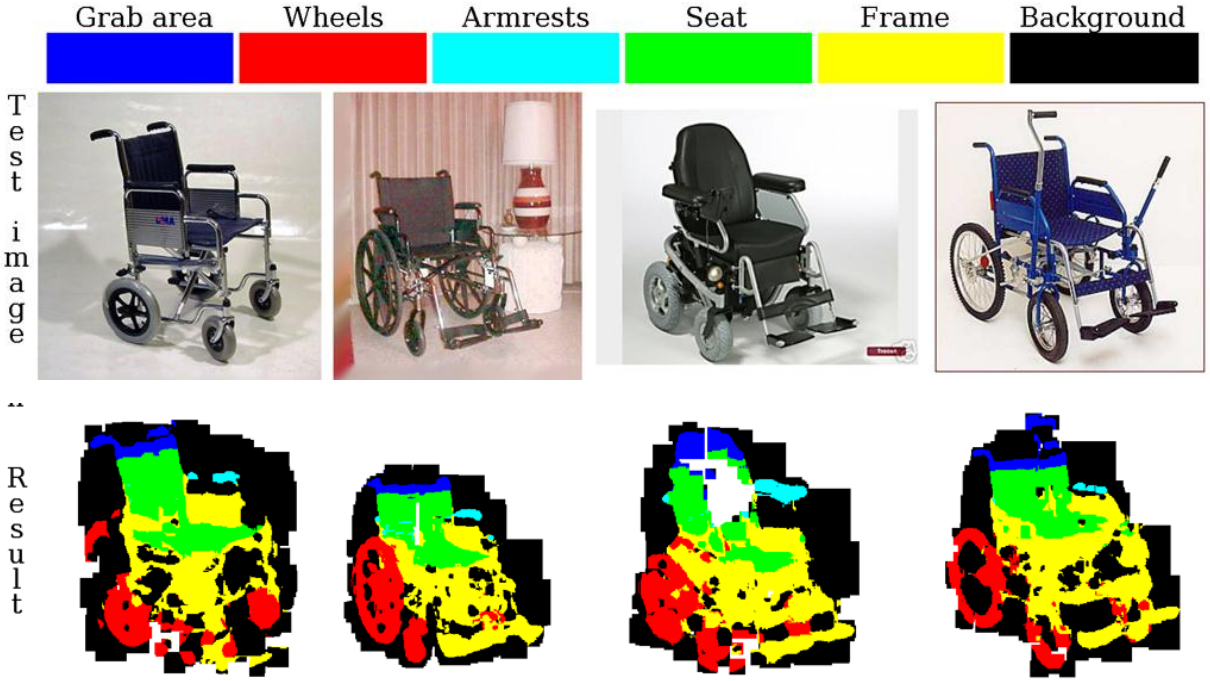
Office chairs

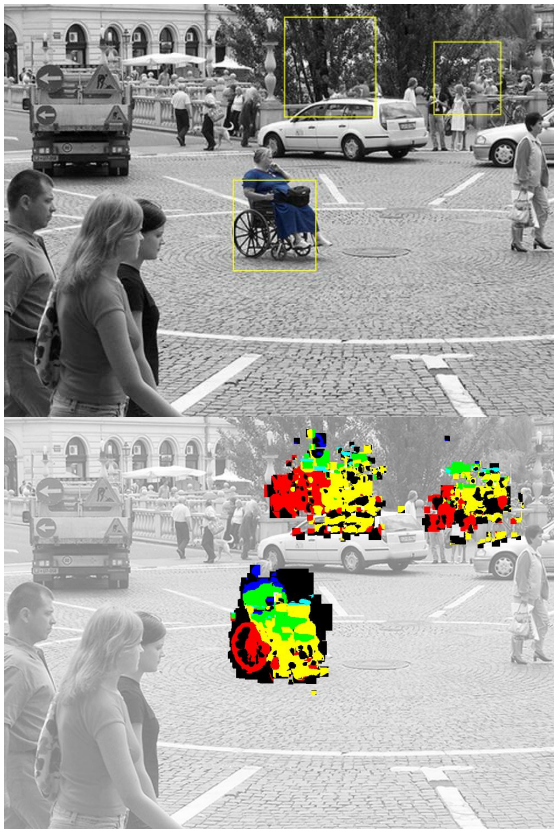
Inferring Other Information: Part Labels

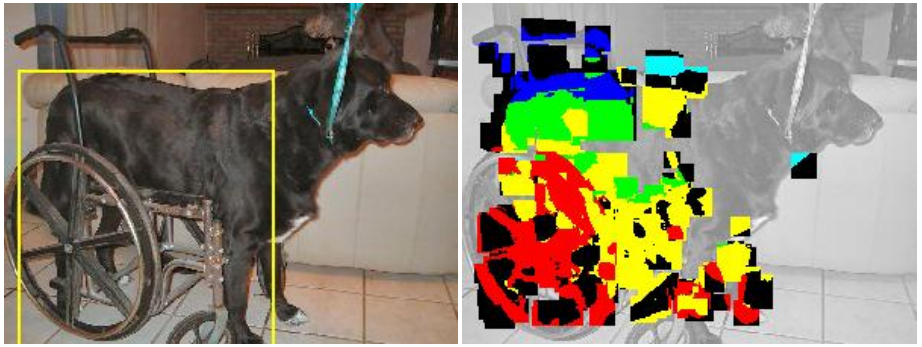


[Thomas07]

Inferring Other Information: Part Labels (2)

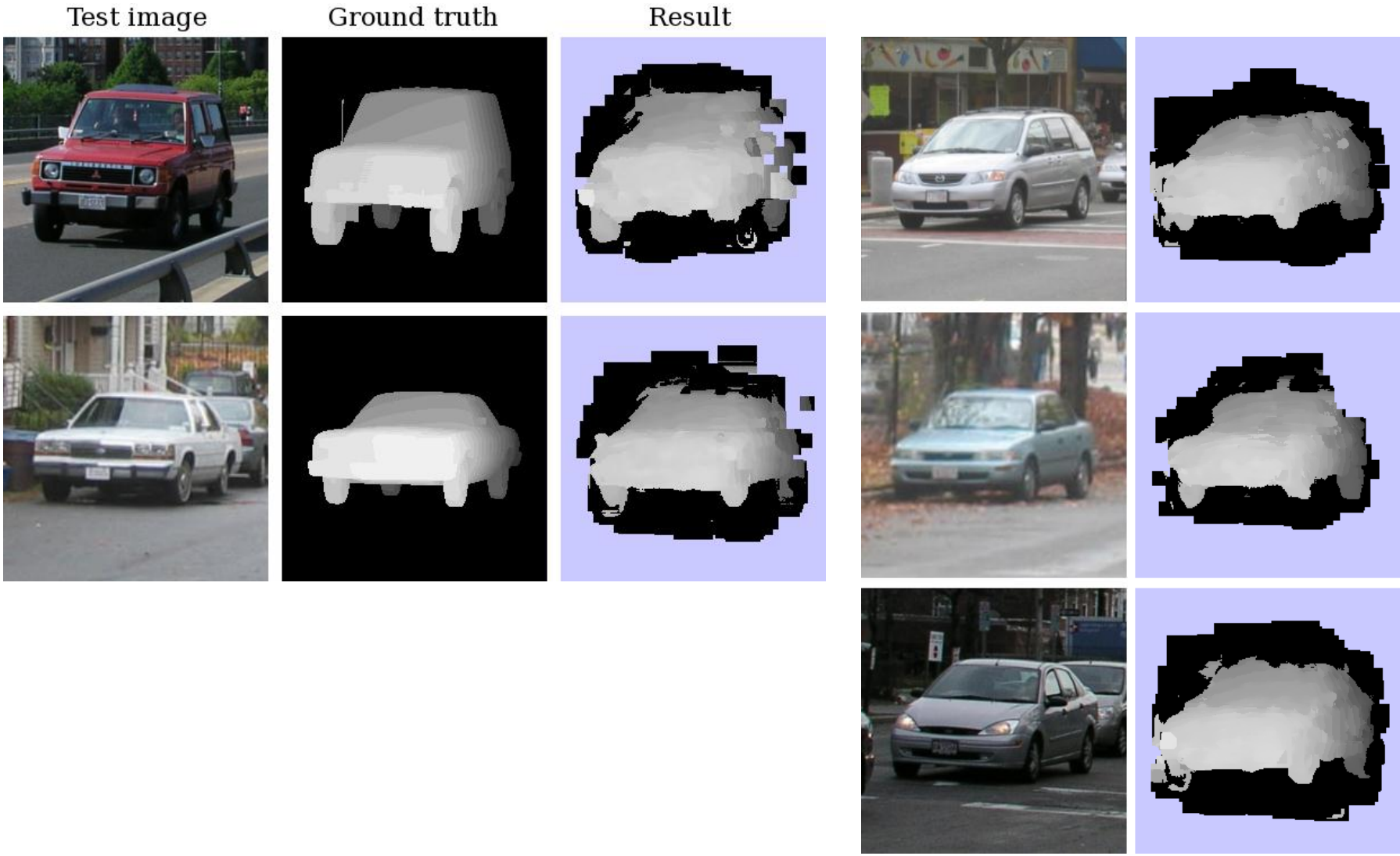






[Thomas07]

Inferring Other Information: Depth Maps

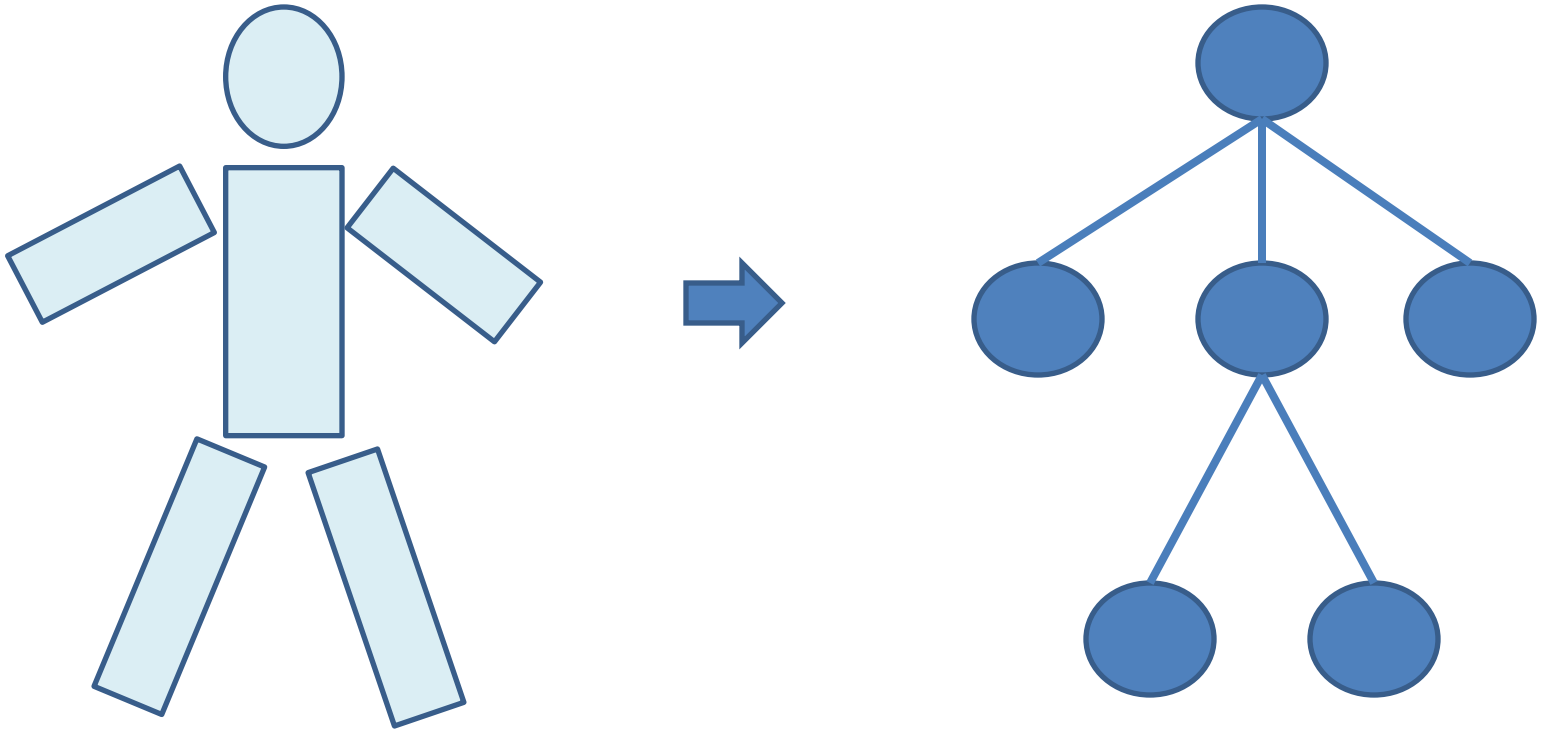


[Thomas07]

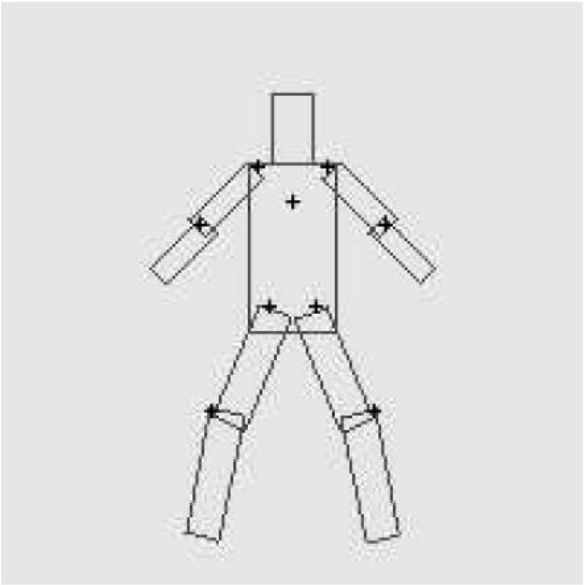
2 minute break

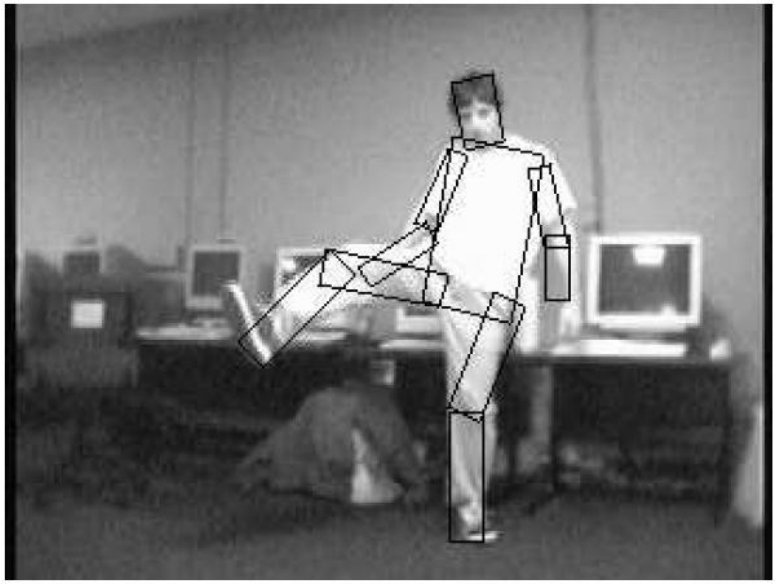
- Comparing a sliding window/part detector
  - What are the advantages of the part detector?
  - What are the advantages of the sliding window detector?

Tree-shaped model



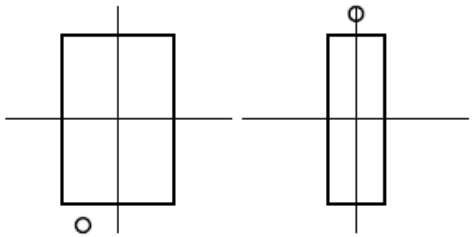
Pictorial Structures Model



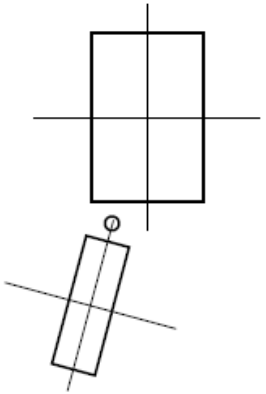


Part = oriented rectangle

Spatial model = relative size/orientation



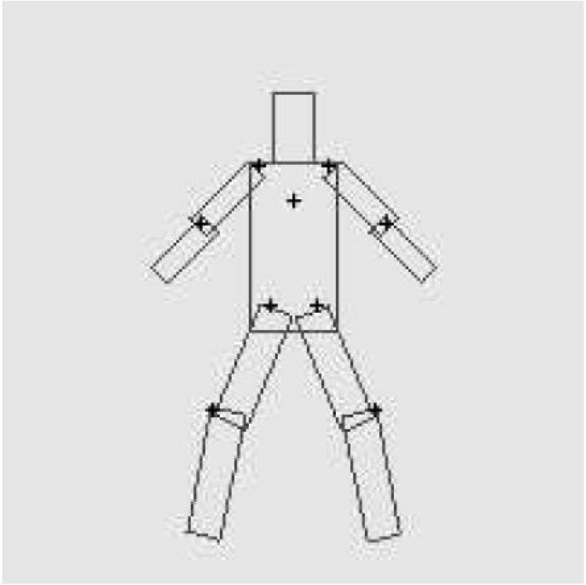
a



b

Felzenszwalb and Huttenlocher 2005

Pictorial Structures Model



$$P(L|I,\theta) \propto \left(\prod_{i=1}^{n} p(I|l_i, u_i) \prod_{(v_i,v_j)\in E} p(l_i, l_j|c_{ij}) \right)$$

Appearance likelihood

Geometry likelihood

Modeling the Appearance

- Any appearance model could be used
  - HOG Templates, etc.
  - Here: rectangles fit to background subtracted binary map

- Can train appearance models independently (easy, not as good) or jointly (more complicated but better)

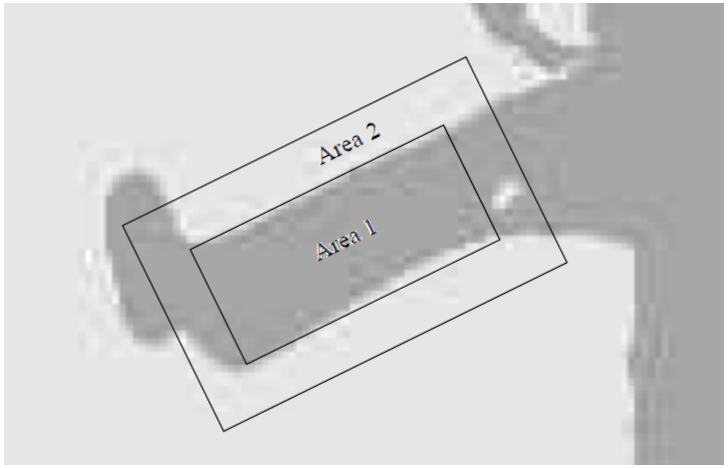
$$P(L|I,\theta) \propto \left(\prod_{i=1}^{n} p(I|l_i, u_i) \prod_{(v_i,v_j)\in E} p(l_i, l_j|c_{ij}) \right)$$

Appearance likelihood

Geometry likelihood

Part representation

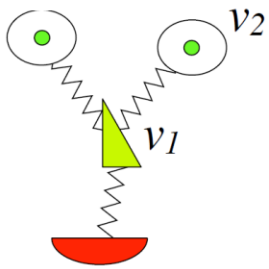
- Background subtraction



Pictorial structures model

Optimization is tricky but can be efficient

$$L^* = \arg\min_L \left(\sum_{i=1}^{n} m_i(l_i) + \sum_{(v_i,v_j)\in E} d_{ij}(l_i, l_j) \right)$$



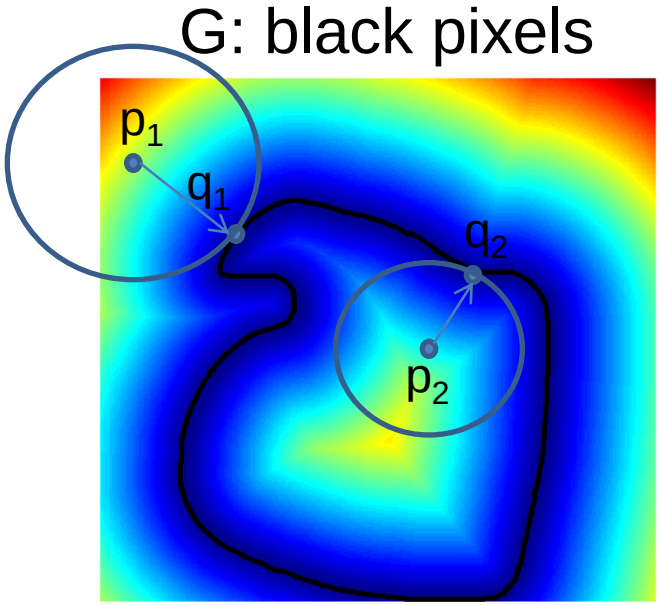
- For each l_1, find best l_2:

$$\text{Best}_2(l_1) = \min_{l_2} \left[m_2(l_2) + d_{12}(l_1, l_2) \right]$$

- Remove v_2, and repeat with smaller tree, until only a single part
- For k parts, n locations per part, this has complexity of $O(kn^2)$, but can be solved in ~$O(nk)$ using generalized distance transform

Distance Transform

- For each pixel p, how far away is the nearest pixel q of set S
  - $f(p) = \min_{q \in G} \; d(p, q)$
  - G is often the set of edge pixels



Distance Transform - Applications

- Set distances – e.g. Hausdorff Distance
- Image processing – e.g. Blurring
- Robotics – Motion Planning
- Alignment
  - Edge images
  - Motion tracks
  - Audio warping
- *Deformable Part Models*

Generalized Distance Transform

- Original form: $f(p) = \min_{q \in G} \; d(p,q)$
- General form: $f(p) = \min_{q \in [1,N]} \; m(q) + d(p,q)$

- For many deformation costs, $O(N^2) \rightarrow O(N)$

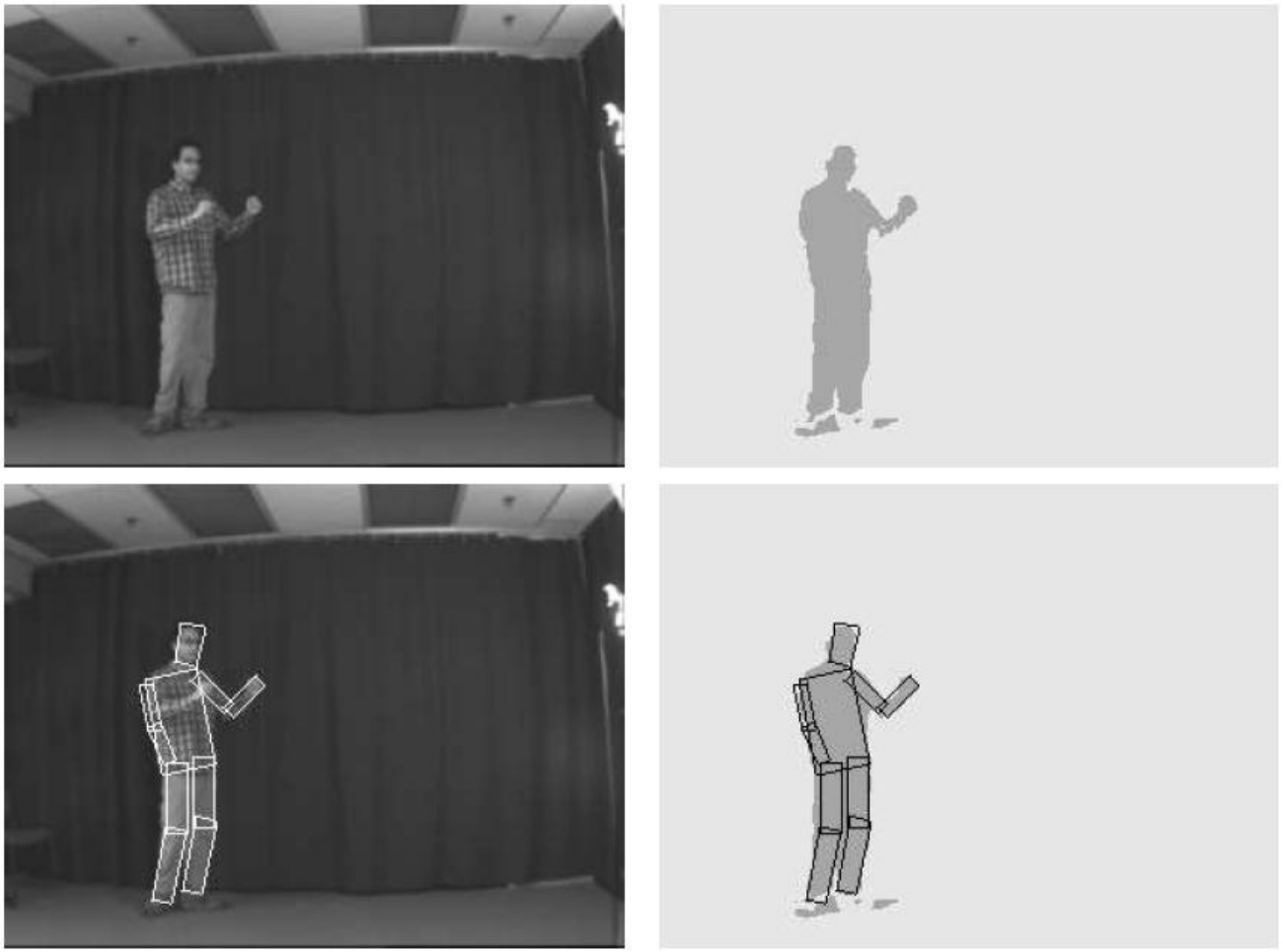
Quadratic $d(p,q) = \alpha(p-q)^2 + \beta(p-q)$

Abs Diff $d(p,q) = \alpha|p-q|$

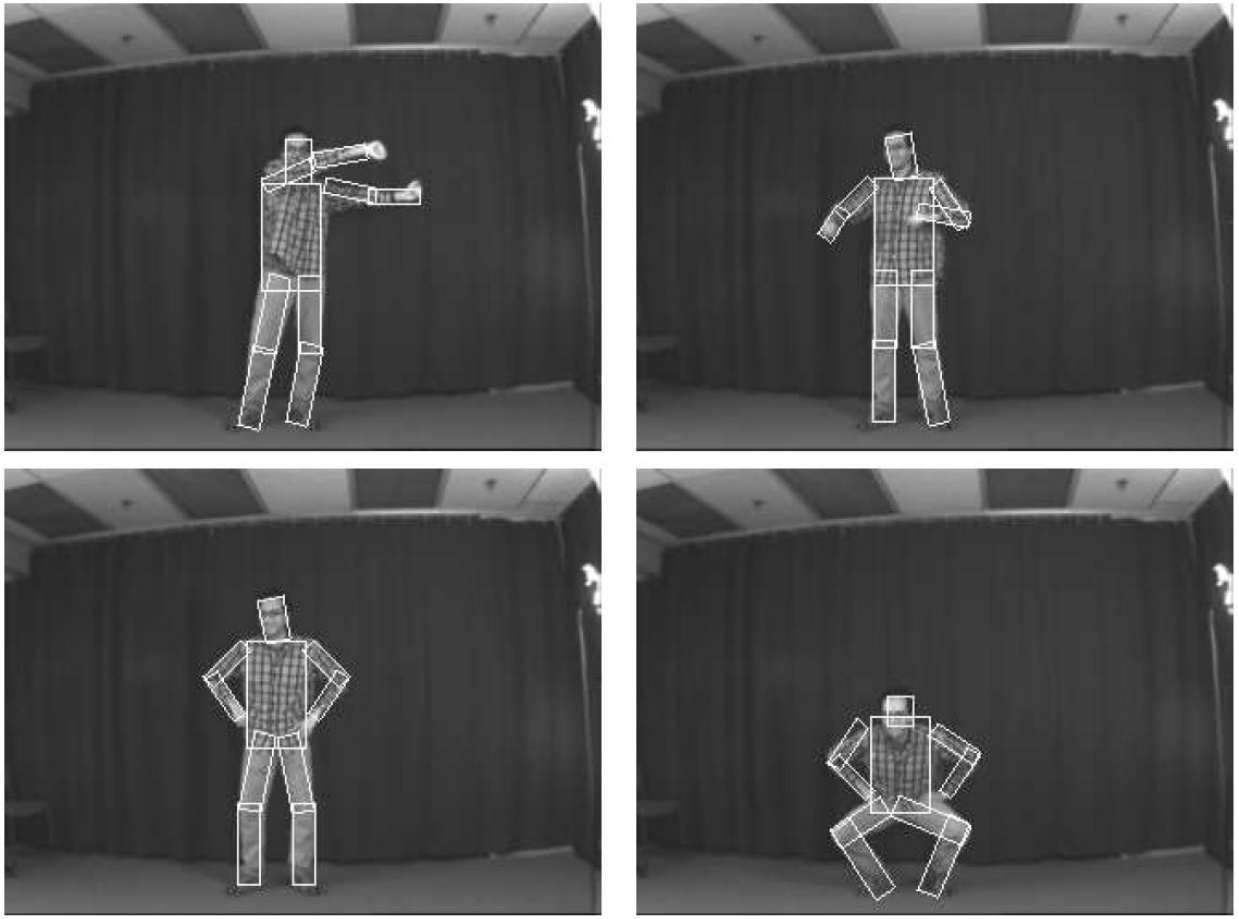
Min Composition $d(p,q) = \min(d_1(p,q), d_2(p,q))$

Bounded $d_\tau(p,q) = \begin{cases} d(p,q) & : |p-q| < \tau \\ \infty & : |p-q| \geq \tau \end{cases}$

Results for person matching



Results for person matching



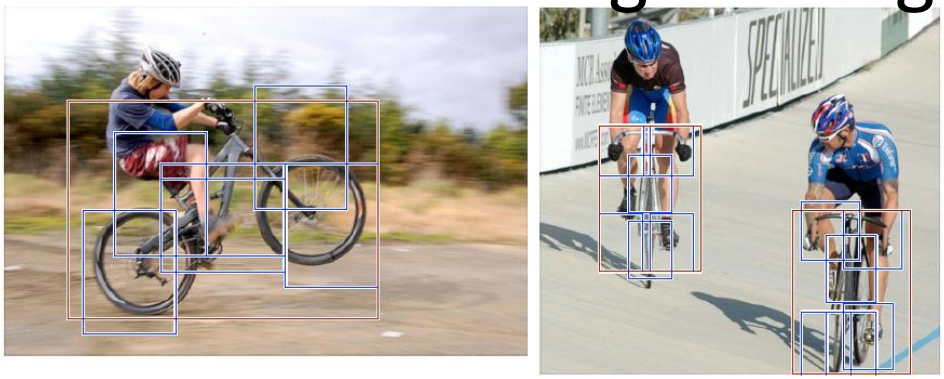
Enhanced pictorial structures

BMVC 2009

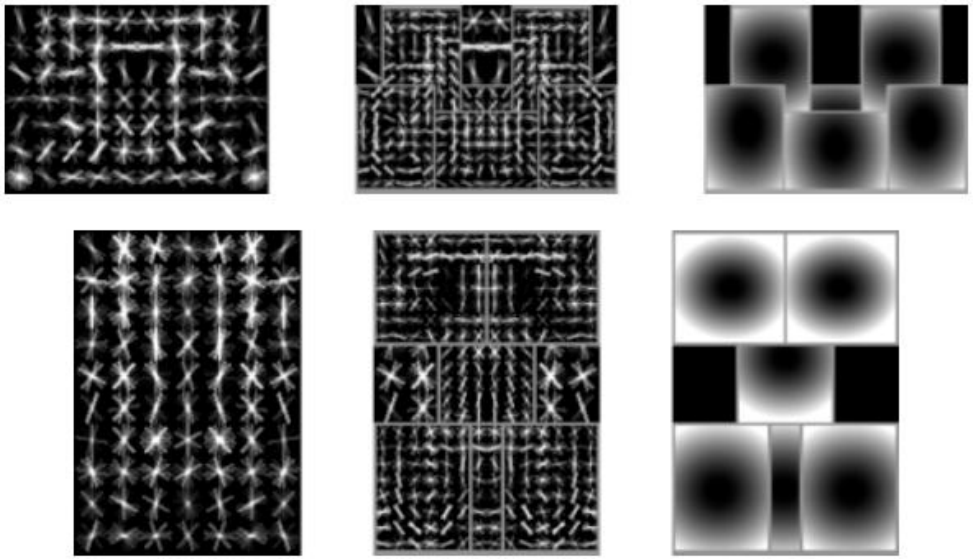
Deformable Latent Parts Model

Useful parts discovered during training

Detections



Template Visualization

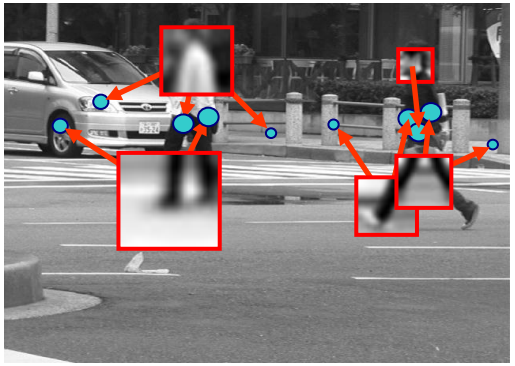


root filters coarse resolution | part filters finer resolution | deformation models

Felzenszwalb et al. 2008

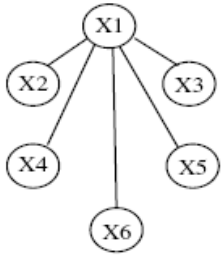
Things to remember

- Rather than searching for whole object, can locate “parts” that vote for object
  - Better encoding of spatial variation

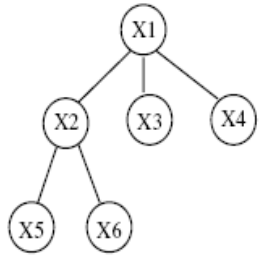


- These parts can vote for other things too

- Models can be broken down into part appearance and spatial configuration
  - Wide variety of models
- Efficient optimization can be tricky but usually possible



b) Star shape



d) Tree

Next classes

- Tues: Object tracking with Kalman Filters
  - Presented by Amin Sadeghi
  - HW 5 is due
- Thurs: Action Recognition
  - Presented by me