Structured Models in Computer Vision

Dan Huttenlocher

June 2010

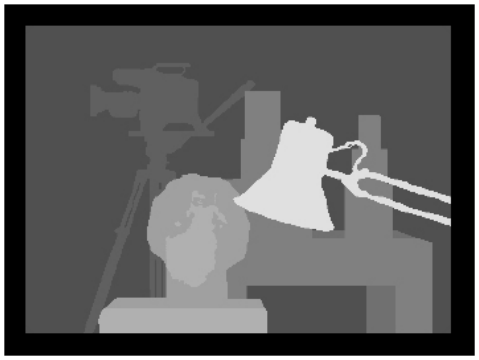
Cornell University
Faculty of Computing and Information Science

Structured Models

- Problems where output variables are mutually dependent or constrained
  - E.g., spatial or temporal relations
- Such dependencies often as important as input-output relations
- Historically studied in generative setting
  - HMM and MRF models
  - Often driven by specific problems
    - E.g., speech and low-level vision
  - Recently more general framework and discriminative methods

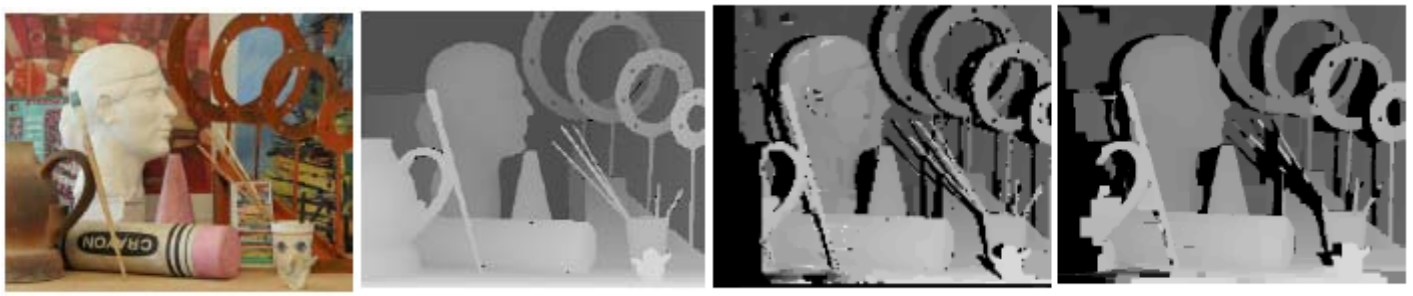
Structured Models in Vision

- Long history with MRF's dating to 1980's
  - Stereo, segmentation, sensor fusion
  - Output variables pixel labels, e.g., disparities
  - Fixed spatial dependency structure
  - Primarily prediction/inference and not learning
    - Hand-tuned energy functions for given problem



Structured Models in Low-Level Vision

- Hand-tuned models a limitation?
  - Few parameters, hard to get ground truth
- Yet structured learning does seem to help



  - Ground truth, max-likelihood CRF [SP07], max-margin [LH08]
  - Latter results compare favorably to best hand-tuned methods
  - Generalize well across datasets!

Structured Model for Stereo [LH08a]

- Data term: sampling-insensitive dissimilarity [BT98]
- Spatial term: linear function of disparity of neighboring pixels and local image gradient
- Sparse long-range edges of length 3^j, $j<k$
  - Max cliques size 2
  - Horizontal and vertical cliques
- Learn parameters using structured SVM
  - BP for finding (approx) most violated constraint
  - Loss function: number of bad unoccluded pixels

Learned Stereo Model Results

- Performs better than learned model [SP07], comparable to hand tuned [SS02][S+05]
- Generalizes reasonably well across datasets

Model \ Scene	**average**	Teddy	Cones
- Grid ($K = 1$), l_{std} loss	14.71	11.34	4.68
- Grid, l_{occl}	15.56	10.92	4.27
- Long-range ($K = 3$), l_{std}	**13.64**	8.89	3.94
- Long-range, l_{occl}	14.06	8.15	**3.77**
- [15] w/ 2 gradient bins	18†	11.3	10.7
- [15] w/ 6 gradient bins	20	14.5	16.8
- [16] w/ GC (non-learning)	–	16.5	7.70
- [18] (non-learning)	–	**6.47**	4.79

Train on Middlebury-2006	
- Long-range, l_{std}	15.73
- Long-range, l_{occl}	**14.96**

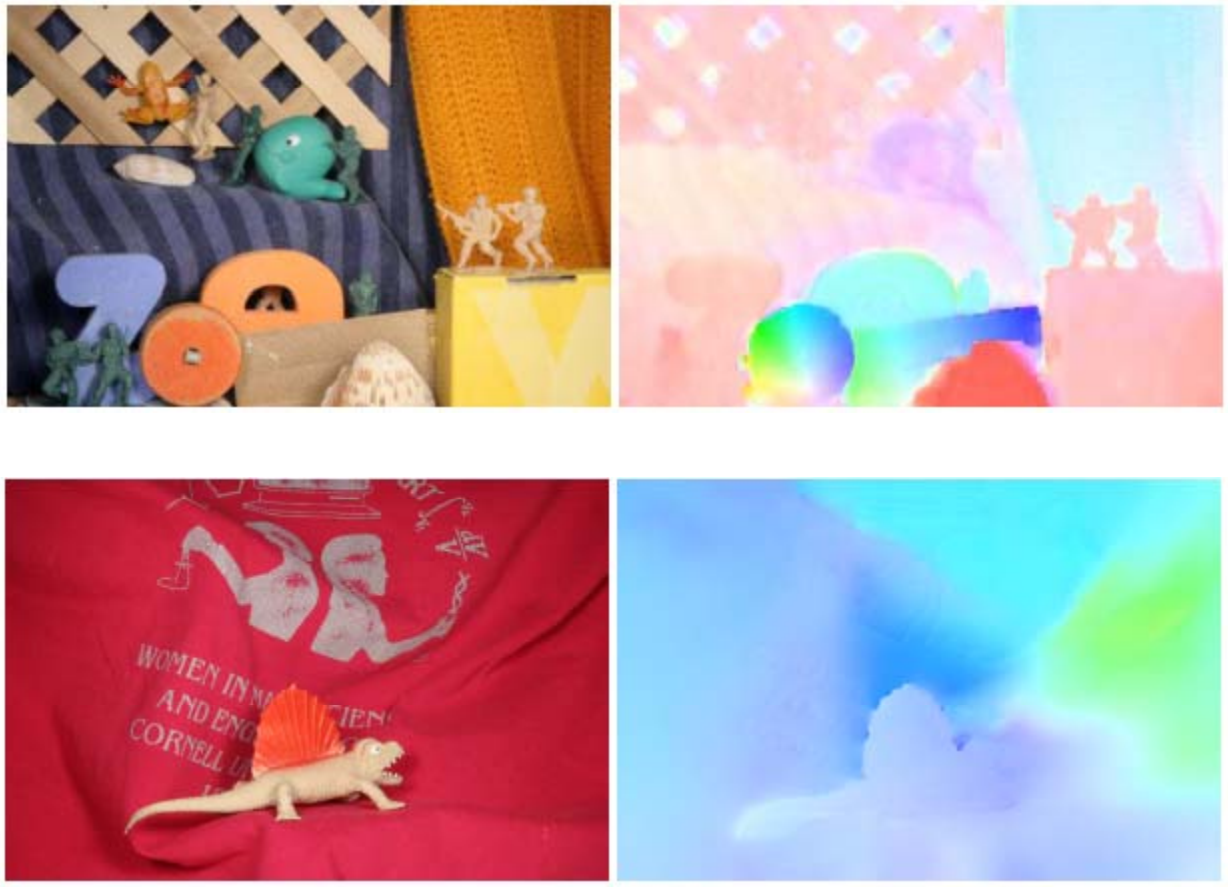
Learning for Optical Flow [LH08b]

- Continuous state MRF
  - Minimize training loss using SPSA (simultaneous perturbation stochastic approximation)
    - Measure loss using average endpoint error
    - Gradient-free method similar to finite difference (FDSA) but perturbing all model parameters
  - Achieves state of art performance with good generalization across images
    - Again compared to hand-tuned methods

Method\Sequence	Army	Mequon	Schefflera	Wooden	Grove	Urban	Yosemite	Teddy	**Average**
Our model	**6.84**	**8.47**	**12.5**	**8.40**	3.88	**6.32**	2.56	7.29	**7.03**
	0.18	**0.57**	**0.84**	**0.52**	**1.12**	1.75	0.13	**1.32**	**0.804**
Bruhn *et al.* [9]	10.1	9.84	16.9	14.1	3.93	6.77	**1.76**	**6.29**	8.71
	0.28	0.69	1.12	1.07	1.24	1.56	**0.10**	1.38	0.930

AAE
AEPE

Optical Flow Examples

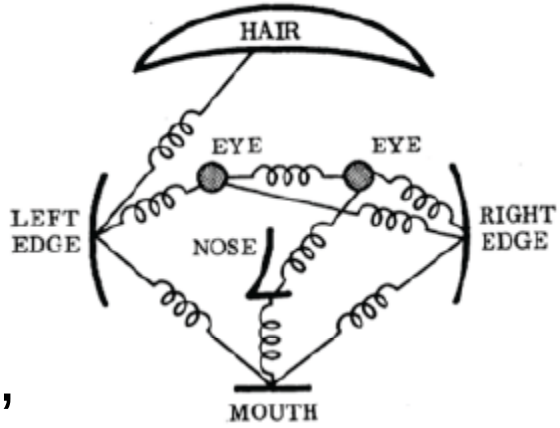


Object Category Recognition

- Three widely studied tasks
  - Image classification
    - Presence of absence of object category in image
  - Object category detection
    - Identifying instances and their locations
    - Possibly subparts including articulated parts such as human body pose
  - Object category segmentation
    - Identifying boundaries of instances – “mask”
- Structured models primarily in detection

Structured Models in Recognition

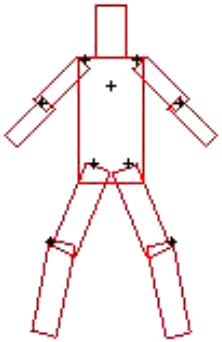
- Combination of local part appearance with spatial dependencies
- Energy minimization formulation of prediction problem – what parts where
- Long history
  - Dating at least to Fischler's Pictorial Structures in 1970's
  - Revisited in machine learning context in late 1990's by Fergus, Perona & Zisserman, Felzenszwalb & Huttenlocher, Forsyth & Ramanan

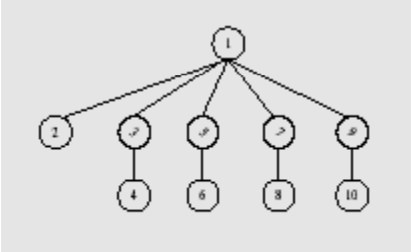


Success with Structured Models

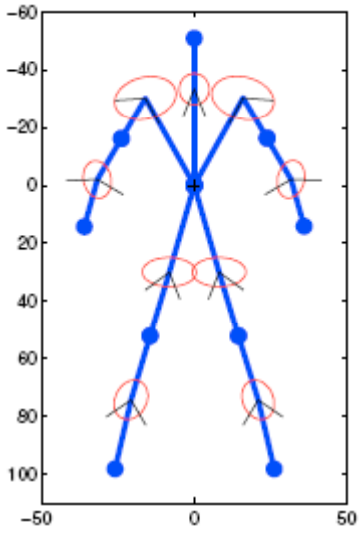
- Human body pose estimation particularly well suited to structured formulation
  - Body part appearances and kinematic dependencies among parts
  - Tree-structured constraints lead to natural dynamic programming formulation
  - Generalized distance transforms provide important additional efficiency [FH00, FH05]
  - Substantial improvements in learning of parts and spatial dependencies in past decade [RSB02], [R06], [FMZ08]
    - Some due to special case task assumptions

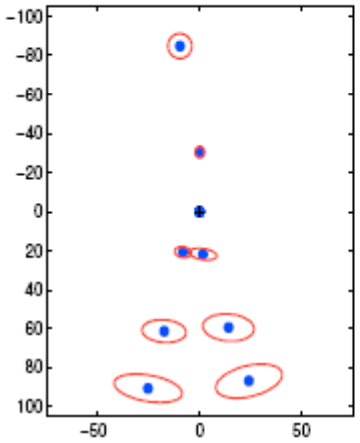
Human Body Pose Models

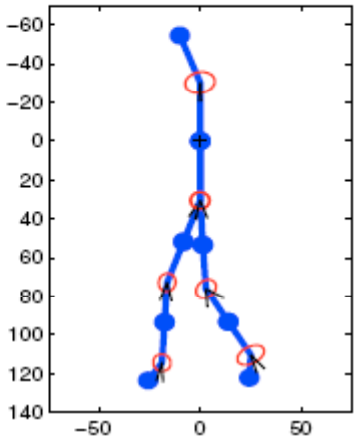




[FH00]





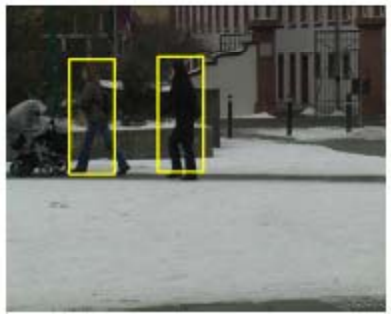
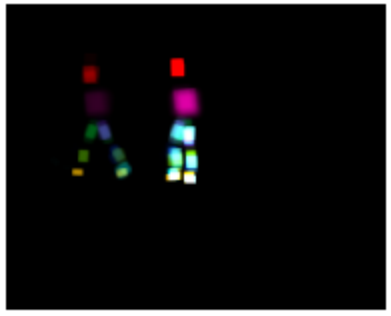


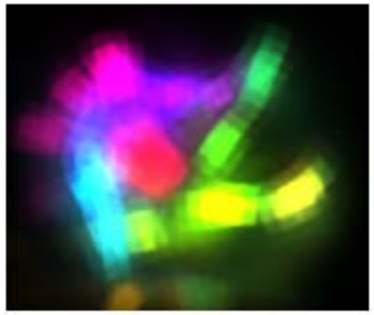
[ARS09]

Star vs. tree model uncertainty

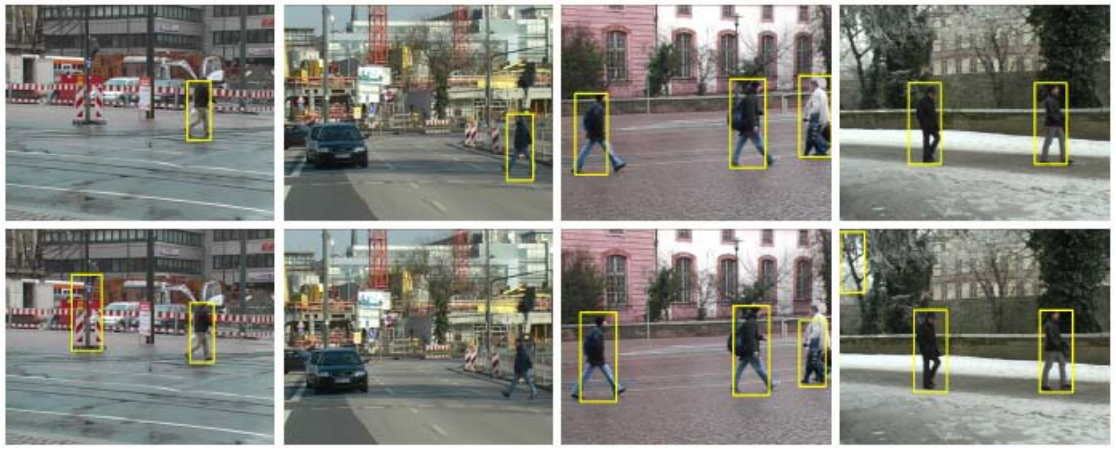
Human Body Detection and Pose Estimation

- Generic pictorial structures yielding state-of-art performance [ARS09]
  - Shape context part descriptors
    - Discriminatively trained AdaBoost classifiers
  - Normalized margin interpreted as likelihood in generative model
  - Part posteriors estimated using BP (exact)
- Detection and pose estimation



[ARS09] Results (vs [R06])

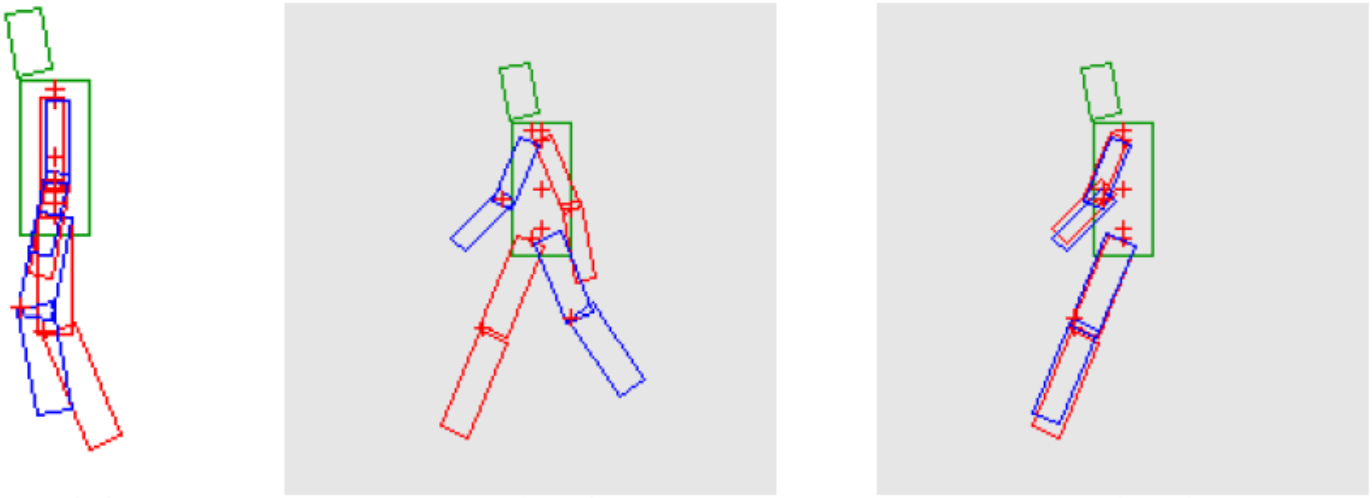


Code and data on web

8 vs. 0 6 vs. 5 3 vs. 4 7 vs. 3 9 vs. 6

Limitations of Kinematic Trees

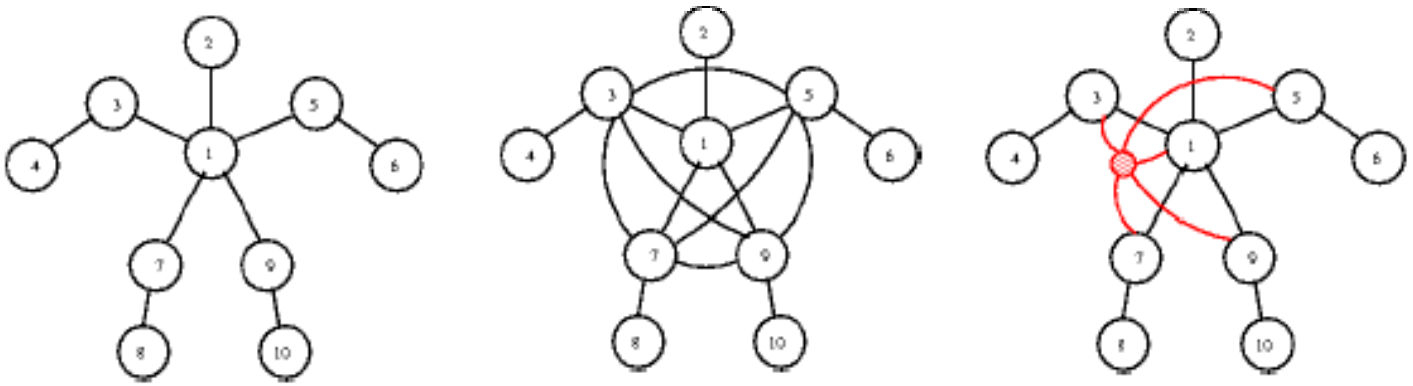
- Only represent relationships between connected parts (note still good proposals)
- Coordination between limbs not encoded
  - Critical for balance and many activities



Equally good under tree model

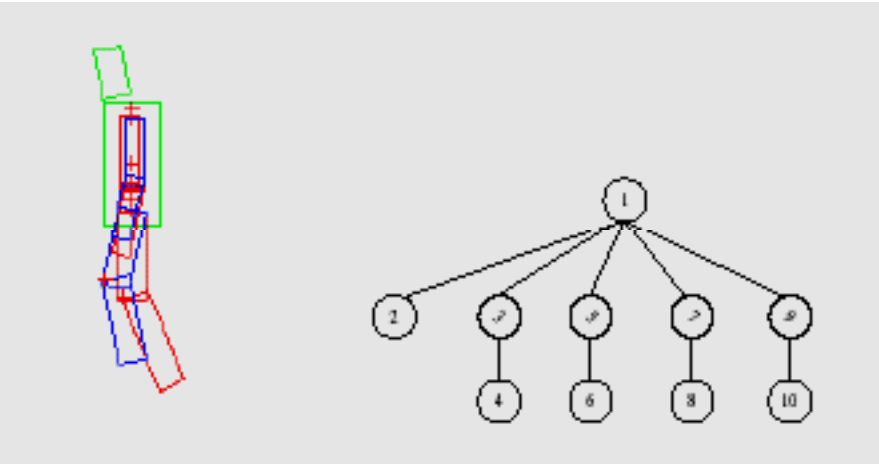
Non-Tree Models

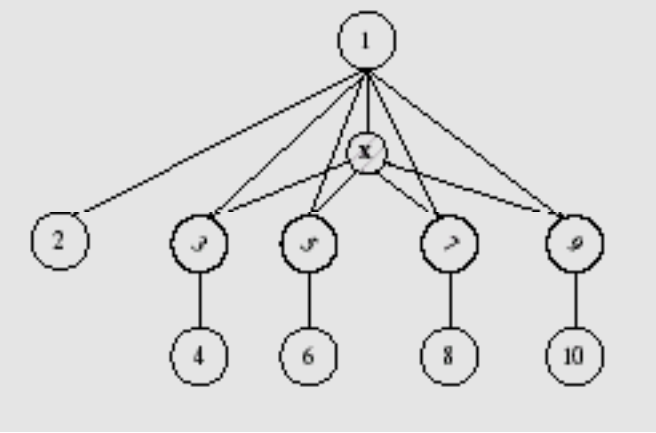
- Larger cliques to capture more dependency
  - Can quickly become computationally intractable
  - Exponential in largest clique size, parameter space for each node large
  - Alternative of introducing latent variables



A Latent Gait Variable for Humans

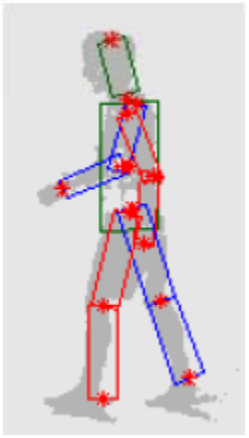
- Additional variable corresponding to common factor of limb coordination [LH05]
  - Consistency between limb positions, not captured by kinematic (skeletal) model
    - Rather than directly connecting limbs which creates large clique





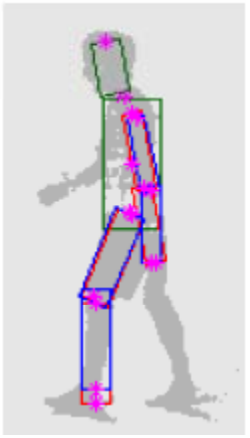
Example Using Brown MOCAP Data

- MAP estimate of best pose, single frame
  - Loopy models, but with small cliques

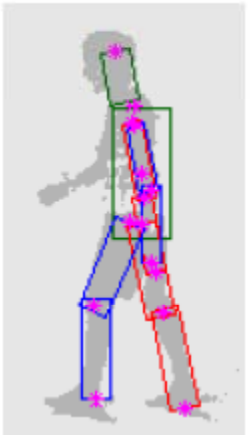


Ground Truth

Latent Variable Model



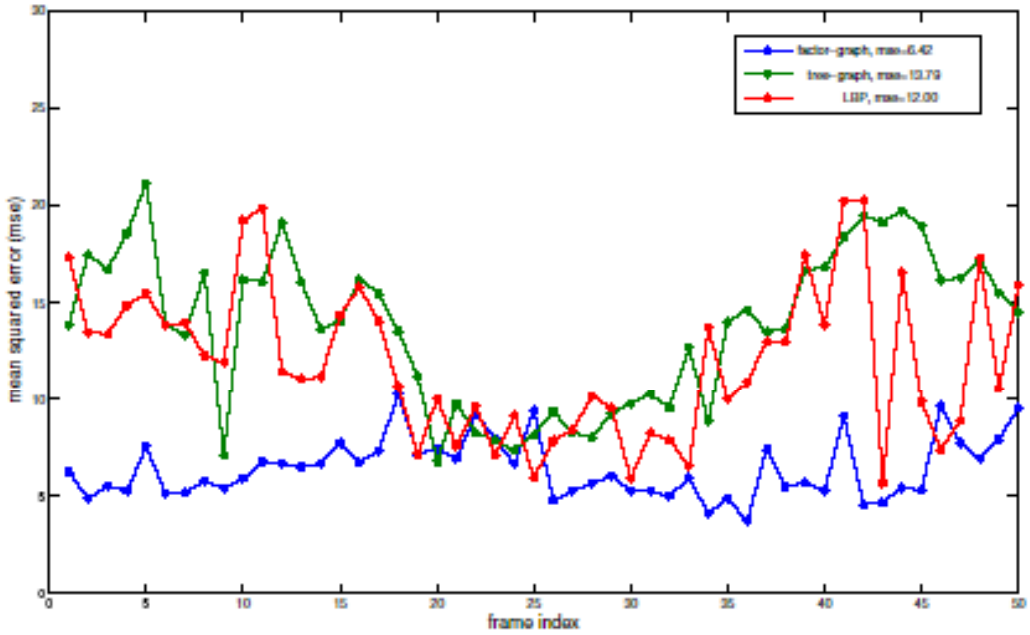
Tree Model



Larger Clique Using LBP (Pairwise)

Latent Gait Variable Helps

- Comparison using ground truth (MOCAP)
  - Latent gait variable model, tree structured model, model with large clique (loopy graph)
  - Better even than model with “more constraint”

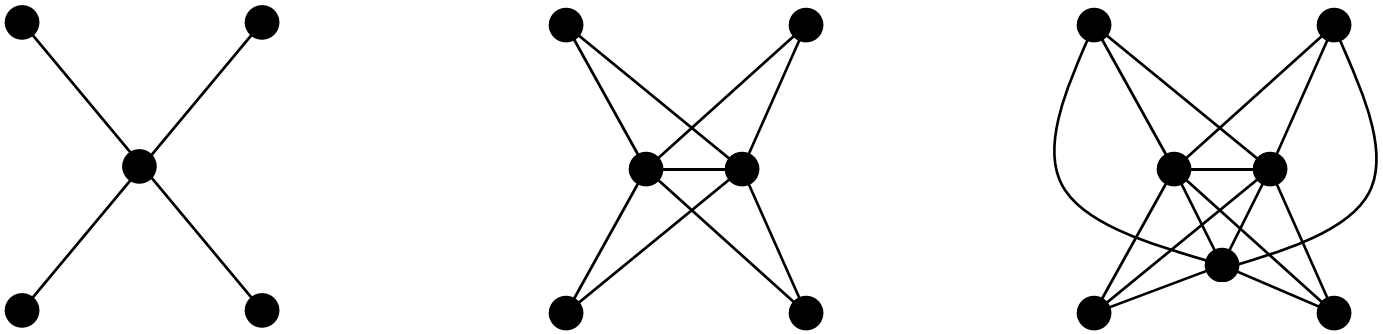


Object Category Recognition

- Most approaches to classification have not used structured models
  - Bag models, features or words (VQ features)
  - Scene-level descriptors such as gist
- More recently, use of weak structure in spatial pyramid matching [LSP06]
  - Considerable success over bag models for classification
  - Fixed structural model, prediction but not structure learning (analogous to MRF stereo)

Structured Models for Category Recognition

- K-fan, set of k reference nodes [CFH05]
  - Triangulated (decomposable)
  - Maximal clique for each non-reference node of size k+1
  - Complete graph, n-1 fan
- Weak spatial structure of 1-fan, star model, seems to be sweet spot (today)

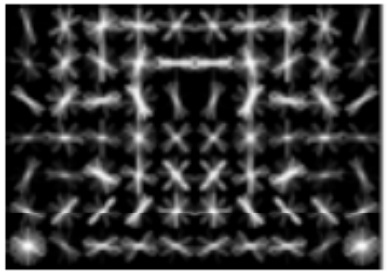
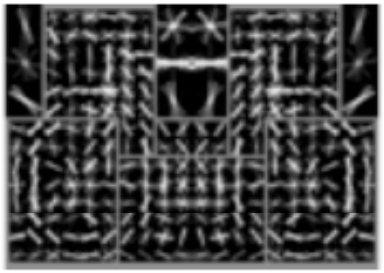
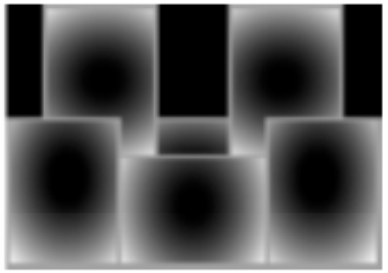
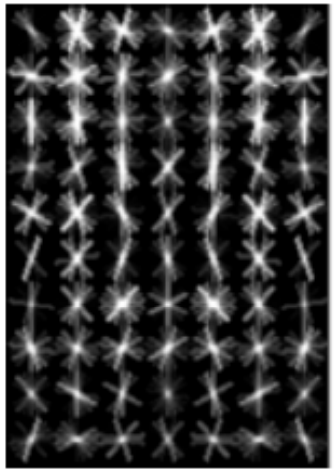
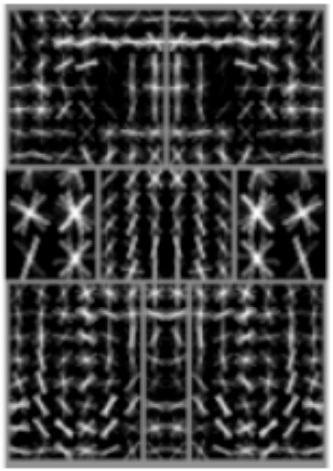
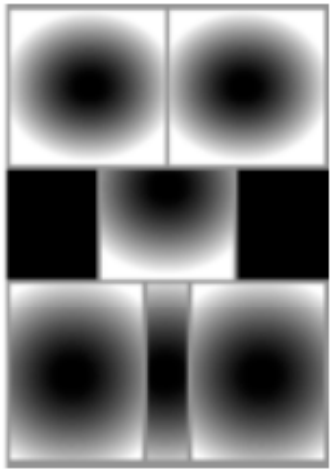


Structured Models for Recognition

- Improvements in structured learning and prediction driving state-of-art performance
  - Felzenszwalb et al, Pascal VOC 07-09
- HoG part models
  - Dense appearance
- Star-graph spatial model
  - Provides reference frame
- Discriminatively trained models
  - Latent SVM, weak labeling for training
- Mixture model for each category

Form of Model [FMR08][FGMR09]

- Two component bicycle model with 6 parts

Coarse Root Fine Parts Spatial Constraint

Score of Hypothesis

- Root w/n parts

Score of F at position p is $F \cdot \phi(p, H)$

$$\text{score}(p_0, \ldots, p_n) = \underbrace{\sum_{i=0}^{n} F_i \cdot \phi(H, p_i)}_{\text{“data term”}} - \underbrace{\sum_{i=1}^{n} d_i \cdot (dx_i^2, dy_i^2)}_{\text{“spatial prior”}}$$

filters

displacements

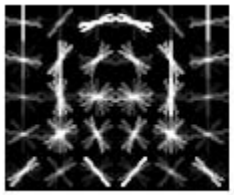
deformation parameters

$$\text{score}(z) = \beta \cdot \Psi(H, z)$$

concatenation filters and deformation parameters

concatenation of HOG features and part displacement features

Processing of Part Response

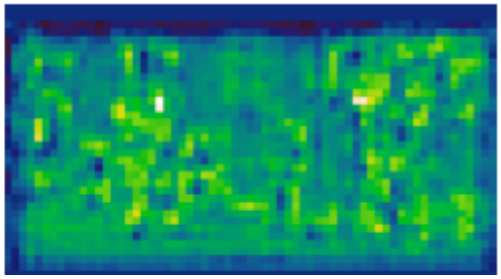


head filter

input image

Response of filter in l-th pyramid level

$$R_l(x, y) = F \cdot \phi(H, (x, y, l))$$

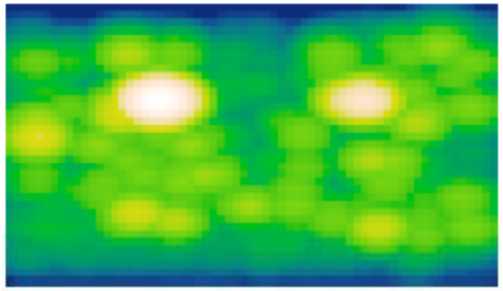


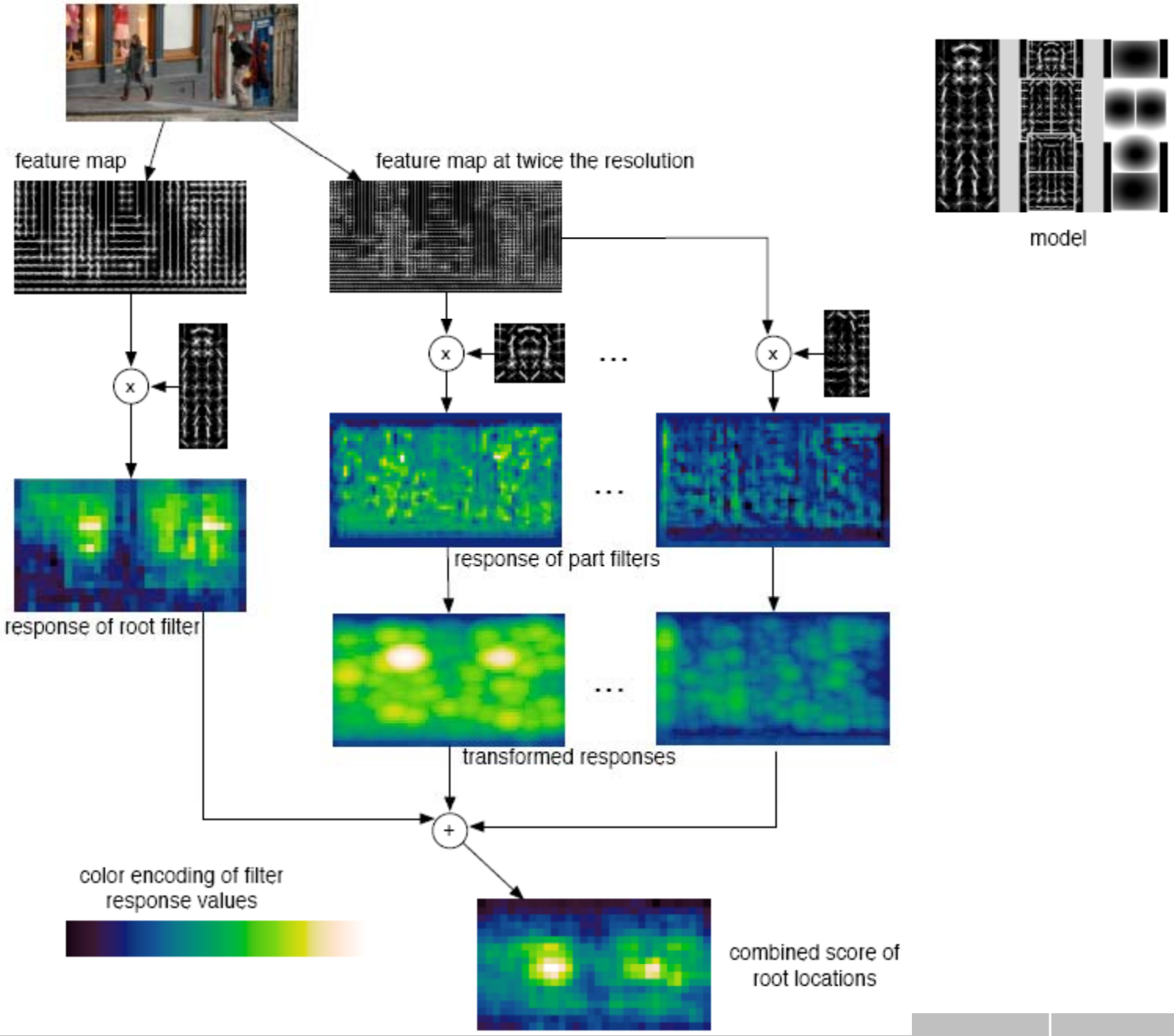
cross-correlation

Transformed response

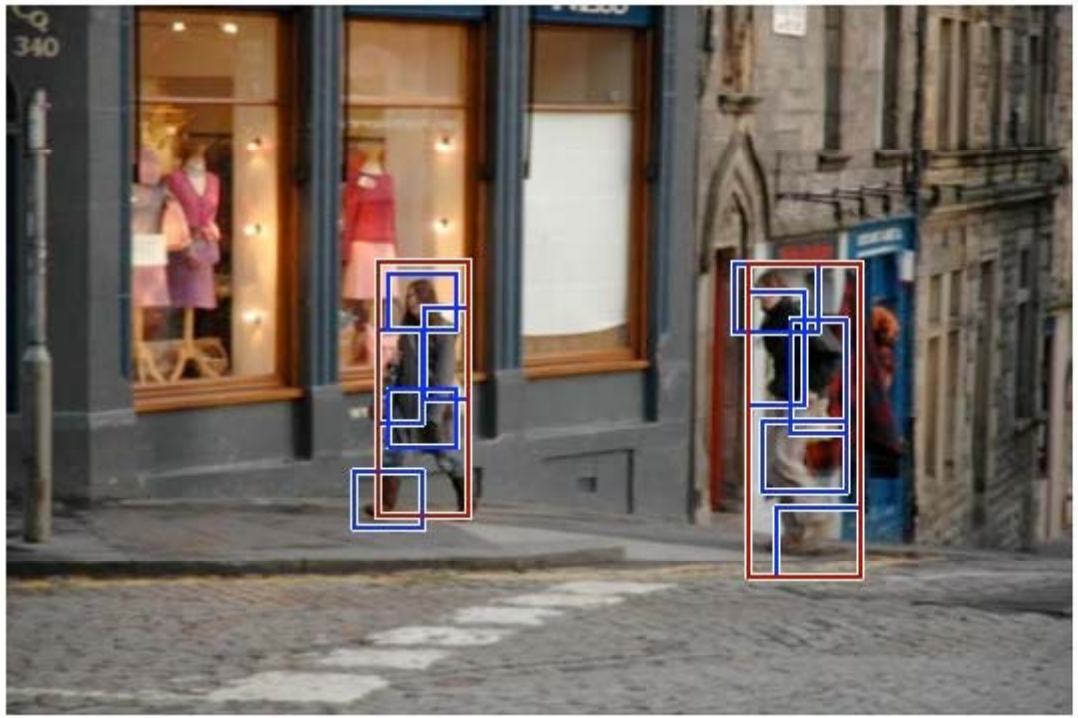
$$D_l(x, y) = \max_{dx,dy} \left(R_l(x + dx, y + dy) - d_i \cdot (dx^2, dy^2)\right)$$

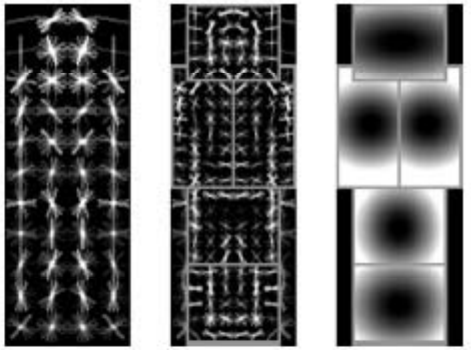
max-convolution, computed in linear time
(spreading, local max, etc)




feature map
feature map at twice the resolution
model
response of root filter
response of part filters
transformed responses
color encoding of filter
response values
combined score of
root locations

Example Results [FGMR09]

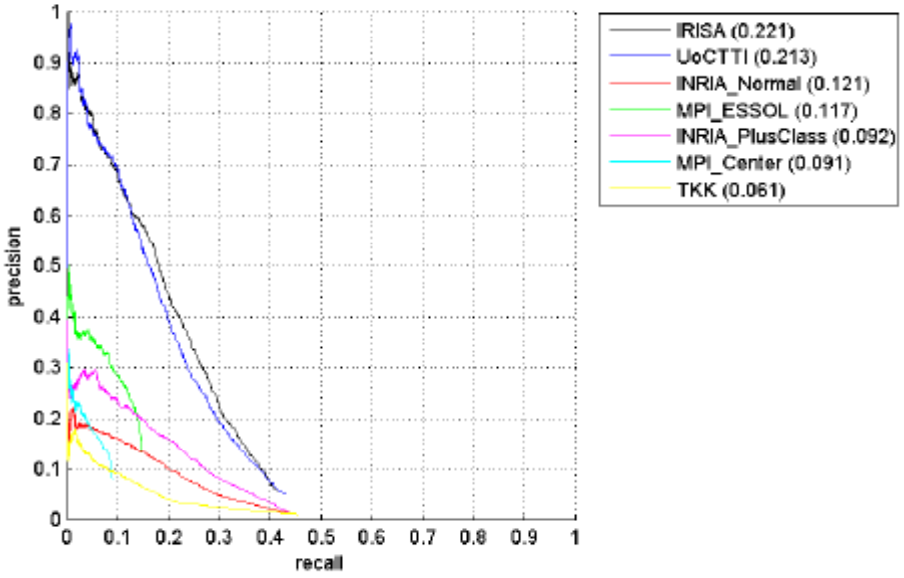


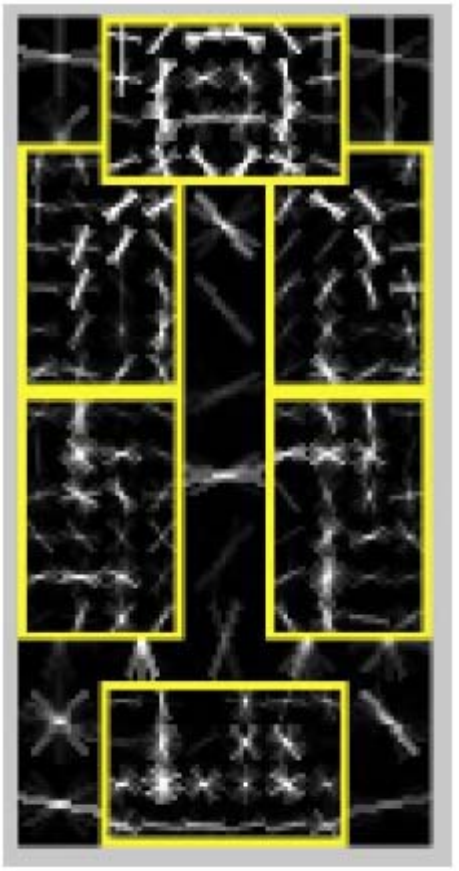


- After non-maximum suppression
- Fast: approx 1 sec to search all scales

PASCAL VOC 2007 Person Detection

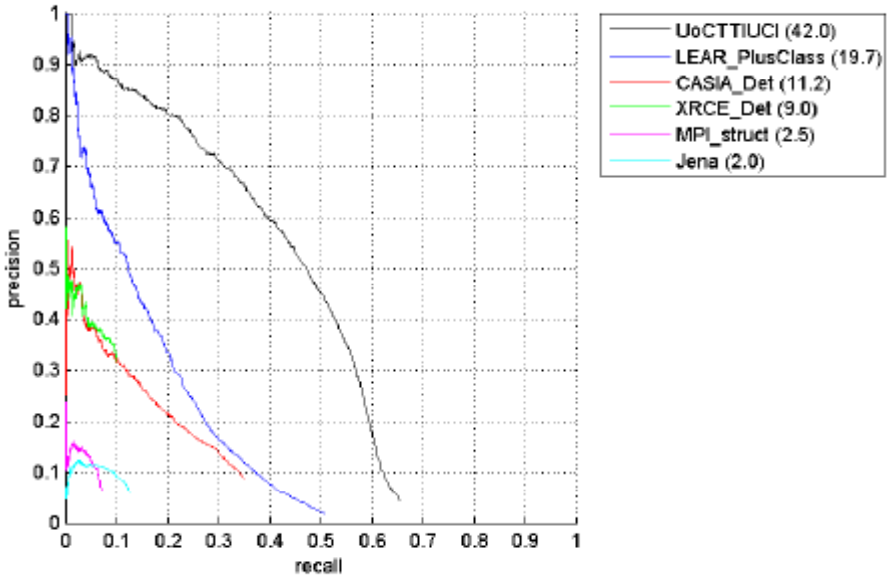
- Pictorial structure model
  - 45% precision at 20% recall

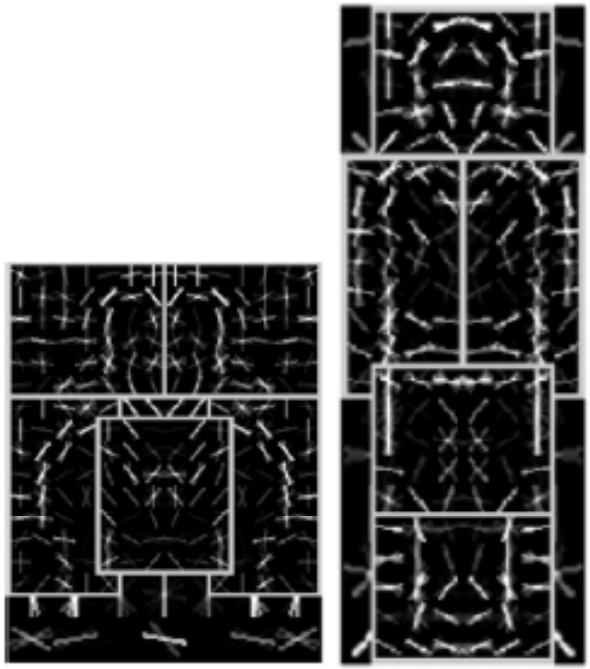




PASCAL VOC 2008 Person Detection

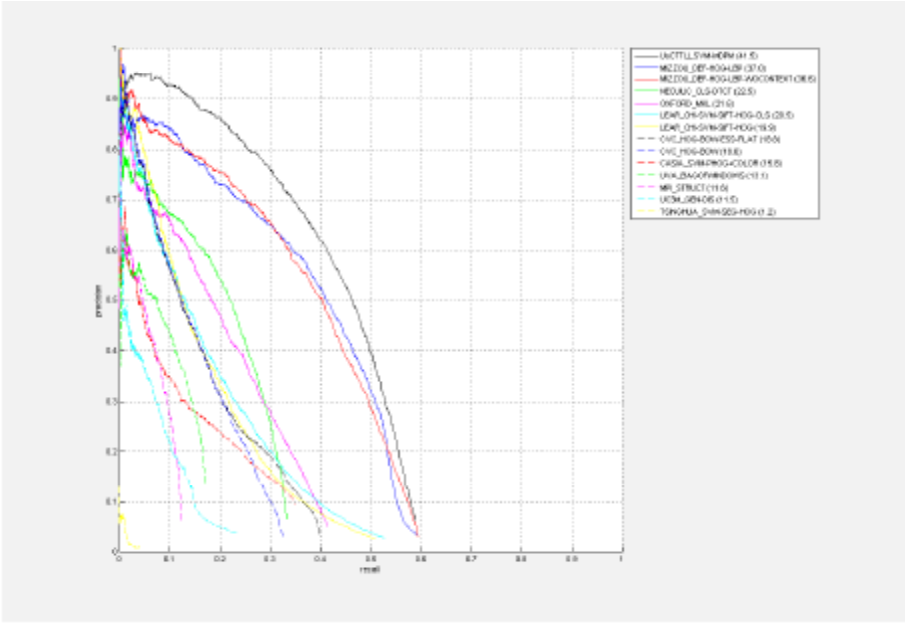
- Disjunction of two pictorial structures
  - 80% precision at 20% recall

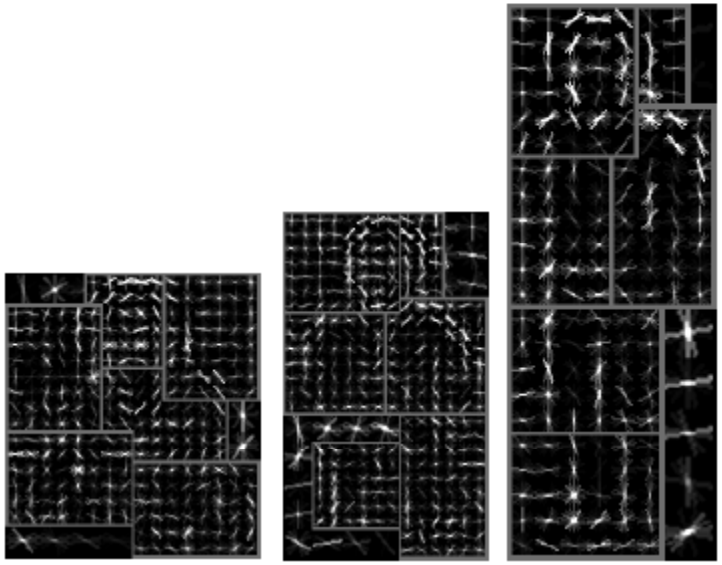




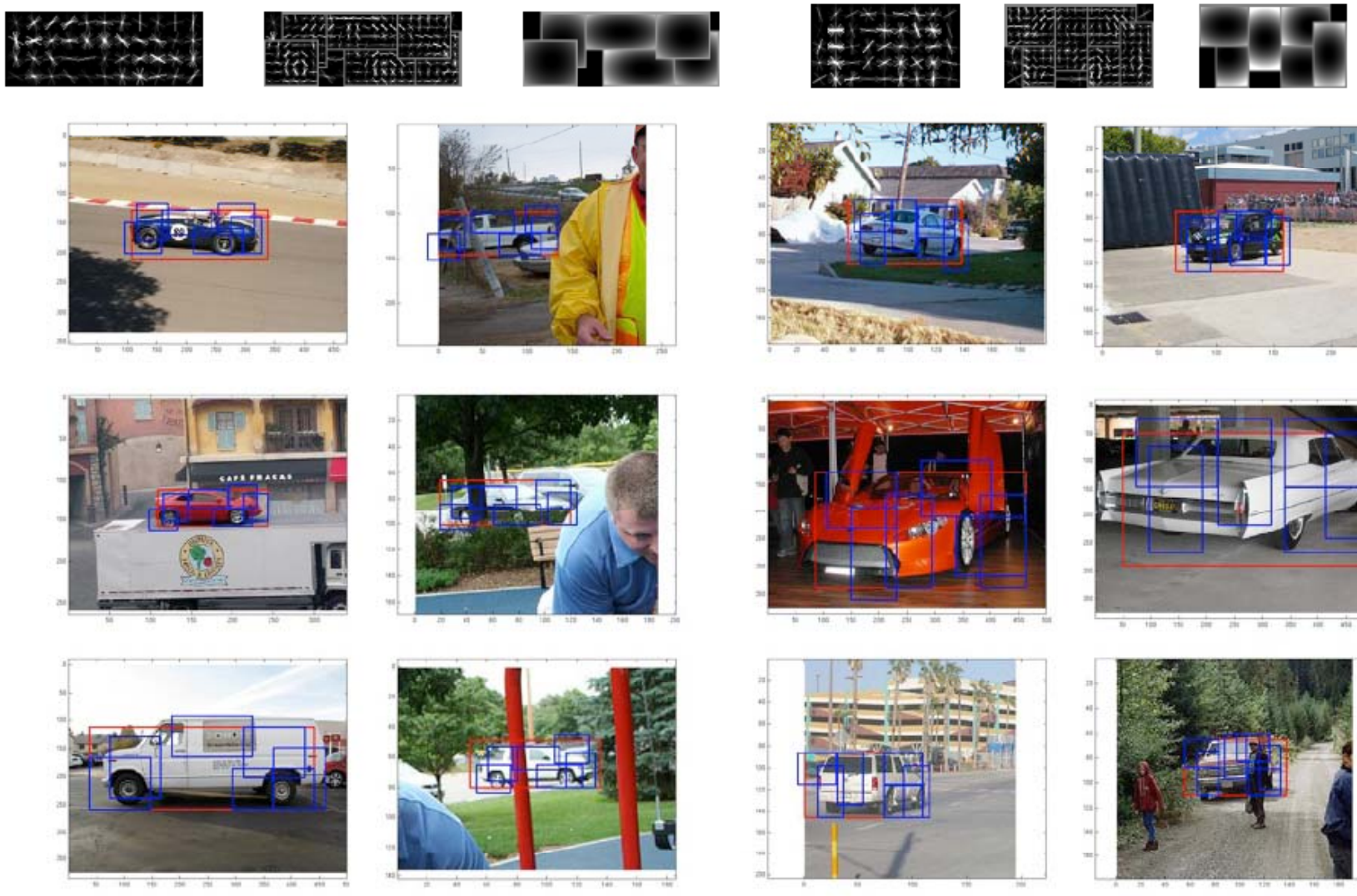
PASCAL VOC 2009 Person Detection

- Disjunction of three pictorial structures
  - 85% precision at 20% recall





Example Car Detections [FGMR09]



Segmentation by Detection

- Felzenszwalb et al also applied their detection method to segmentation
- Binary mask associated to each part of each class model to generate segmentation
  - Masks trained on segmentations
- Yields 3rd or 4th ranked segmentation results out of 21 entries in 2009 Pascal challenge (“comp5”)

What's Working for Recognition

- Algorithmic techniques
  - Energy minimization/optimization framework offers plenty of opportunity for efficient algorithms
    - Often dynamic programming which computes exact same answer only faster
    - Sometimes approximations, but often well studied elsewhere and thus well understood
- Weak labeling of parts
  - Latent svm and other methods for learning structural constraints without explicit training

What's Working (2)

- Power of star-graph models
  - Highly efficient and quite simple to implement
  - Provides "reference frame" for parts, but would seem to be fairly weak constraint
- Use and development of large margin discriminative learning techniques
  - Often coupled with probabilistic interpretation in generative inference for prediction
- Dense part descriptors such as HoG or finely sampled shape contexts

What's Not (Yet?) Working

- Contextual information in recognition
  - Current models not helping much compared to baseline performance without context
- Object categories with less regular spatial structure
  - E.g., cat, dog, bird
- Rare and partial instances
  - Explicit occlusion modeling?
- Large numbers of categories

Directions and Opportunities

- Continued improvement of optimization methods and learning techniques
  - Performance not yet asymptoting, e.g., 25% reduction in error
  - Speeds continuing to improve substantially
- Grammars (probabilistic) and more general representational schemes
- Combining scene-level and object-level modeling to benefit of both

Summary

- Structured model learning and inference widely useful – high "vision specific" content
- Learning (not only prediction) in low-level problems on pixel grid
  - Stereo, flow, denoising, segmentation
- Object category detection and segmentation
  - Natural means of combining appearance and spatial information
  - Rapid progress on algorithms, learning
  - Still lots to do, e.g., combining object and scene levels