Learning and Recognizing Visual Object Categories Without First Detecting Features

Daniel Huttenlocher
2007

Joint work with D. Crandall and P. Felzenszwalb

Cornell University
Faculty of Computing and Information Science

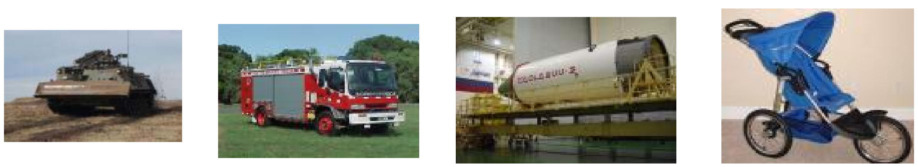
Object Category Recognition

- Generic classes rather than specific objects
  - Visual – e.g., bike

distinguished parts

  - Functional – e.g., chair

  - Abstract – e.g., vehicle

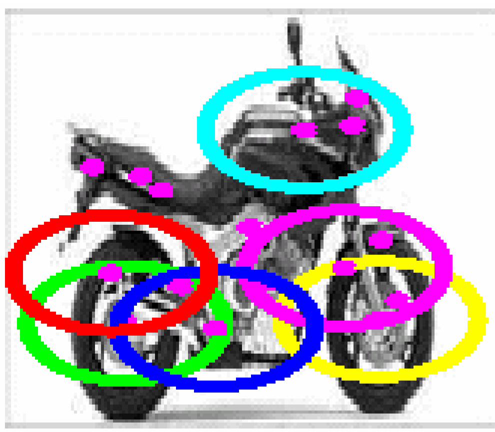


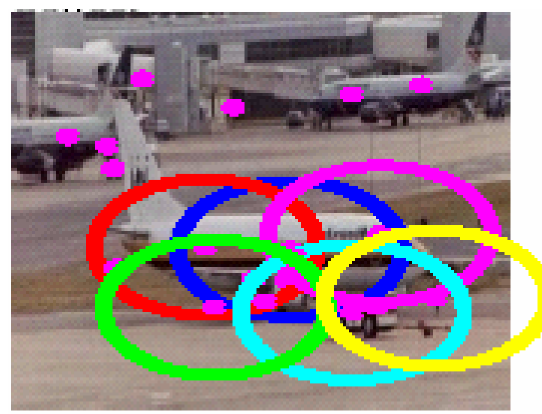
Recognition Tasks

- Classification and localization
  - Classification: presence or absence of an object
    - Image retrieval applications
  - Localization: where objects, and potentially subparts, are in an image
    - Applications that involve interacting with world
- Appearance and geometry
  - Appearance: local patterns of intensity or color
  - Geometry: global spatial configuration, e.g., arrangement of parts

Using Appearance and Geometry

- Most methods rely on feature detection
  - Find sparse affine-invariant feature or interest points such as corners
  - Have spatial model of how feature locations vary within category

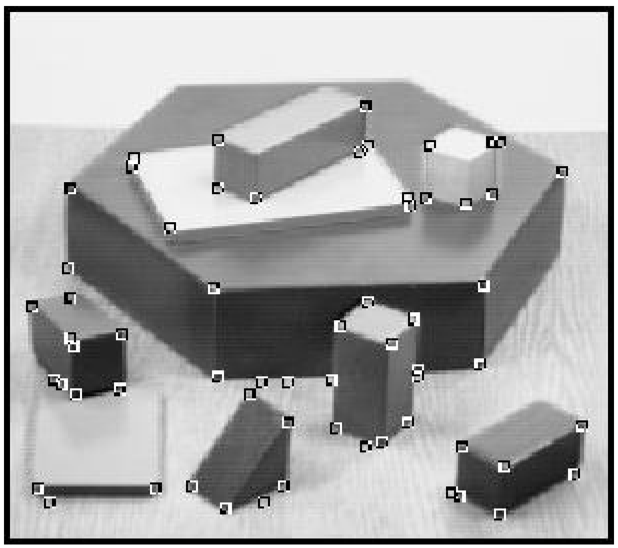




[FPZ03]

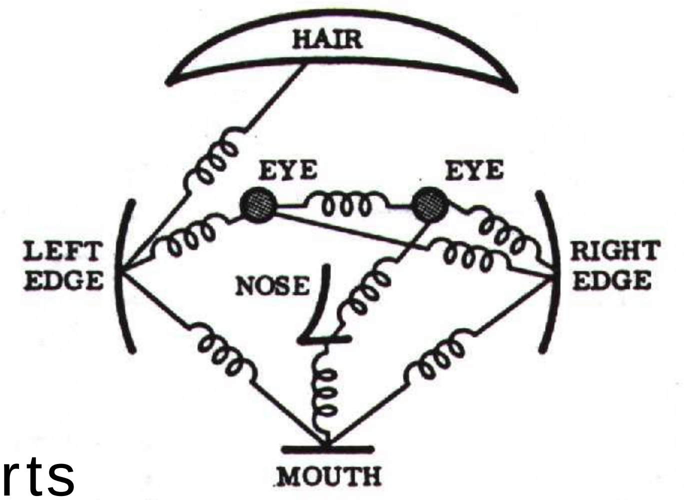
Problems With Feature Detection

- Local decisions about presence or absence of features are difficult and error prone
  - E.g., often hard to determine whether a corner is present without more context



Spatial Models Without Feature Detection

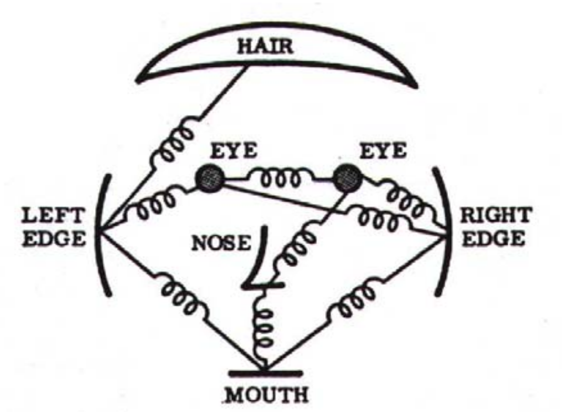
- Pictorial structures [FE73]
  - Parts arranged in deformable configuration
    - Match cost function for each part at each location
    - Deformation cost function for each connected pair of parts



- Intuitively natural notion of parts connected by springs
  - “Wiggle until fits”, <u>no individual feature detection</u>
  - Abandoned due to computational difficulty

Formal Definition of Model

- Undirected graphical model – MRF
  - Graph M=(V,E)
  - Parts V=(v_1, ..., v_n)
  - Spatial relations E={e_{ij}}
    - Gaussian on relative locations for pair of parts i,j
- Spatial prior $P_M(L)$
  - L=(ℓ_1, ..., ℓ_n) and each ℓ_i discrete configuration space
    - E.g., translation, rotation, scale



7 nodes
9 edges
(out of 21)

Object Detection

- Given image I and model M
  - Prior $P_M(L)$ distribution of spatial configurations
  - Likelihood $P_M(I|L)$ of image given configuration
- Evidence over all configurations L

$$\sum_L P_M(I|L)P_M(L) \propto \sum_L P_M(L|I)$$

- Or quality of best configuration (MAP est.)

$$\max_L P_M(I|L)P_M(L) \propto \max_L P_M(L|I)$$

  - Also localizes parts, maximizer L^*
  - Energy minimization, negative log

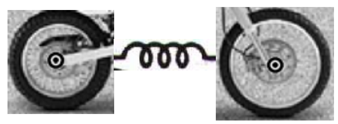
Pictorial Structures Version 2

- Efficient algorithms for certain types of pictorial structure models
  - Tree- or fan-like underlying graph structures and likelihood that factors [FH00,FH05,CFH05]
    - Dynamic programming techniques
- Issue of learning models [CH06]
  - Using weak supervision, where training data specifies presence of object but not location
- Better performance than approaches that rely on detected features [CFH05,FPZ05]

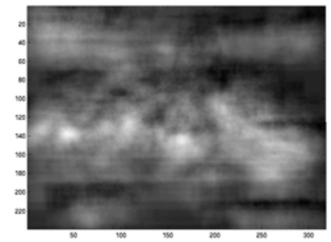
Single Overall Estimation Problem

- Likelihood of image given each part at each location
  - E.g., edge probability templates, translation

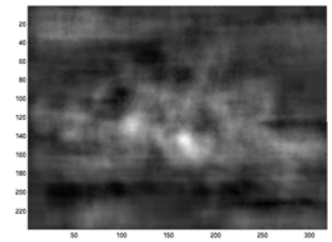
I



v_1 v_2

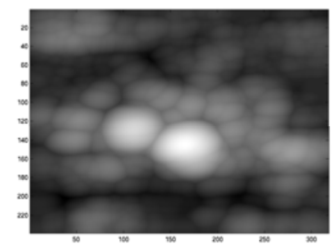


$P_M(I|\ell_1)$



$P_M(I|\ell_2)$

- How well fits spatial model
  - No error-prone feature detection
  - Tractability depends on graph



$\max_{\ell_1} P_M(I|\ell_2) P_M(\ell_1,\ell_2)$

Fast Methods

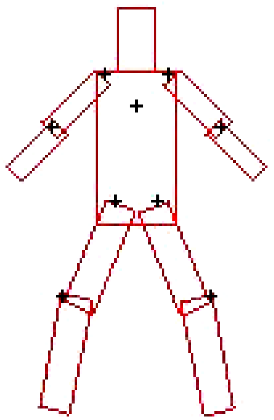
- Spatial term based on relative location of pairs, allows convolution-like operations

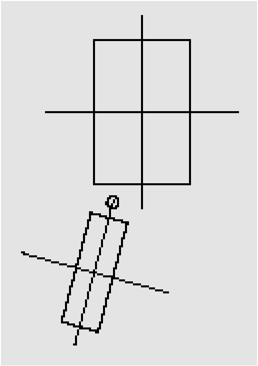
$$P_M(\ell_i,\ell_j) \propto \rho(\ell_i-\ell_j)$$

- Acyclic spatial models with n parts, m locs
  - Best match (MAP estimate) [FH00, FH05]
    - Linear time methods for min convolution yield O(mn) time, generalized distance transforms
  - All configurations (marginals) [FH05]
    - Using FFT O(mnlogmn) time
      - For Gaussian, binomial filters O(mn) time
    - Fast sampling of good candidate matches

Tree Structured Models

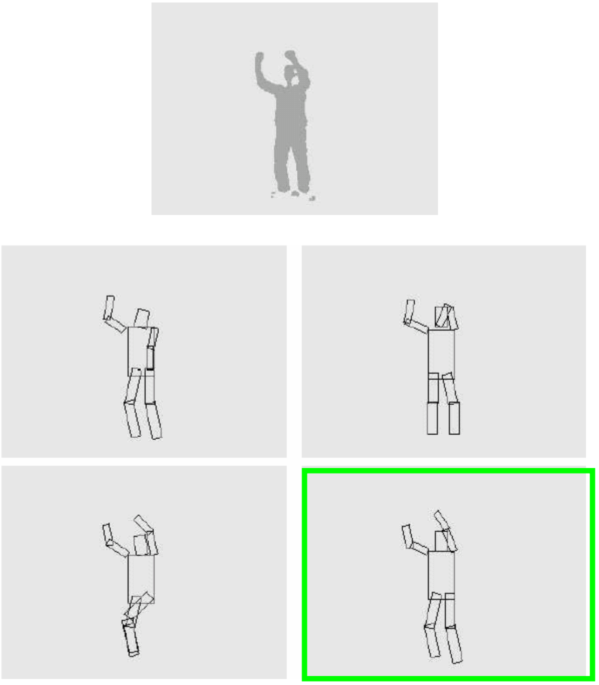
- Kinematic structure of animate objects
  - Skeleton forms tree
  - Parts as nodes, joints as edges
- 2D image of joint
  - Spatial configuration for pair of parts
  - Relative orientation, position and scale (foreshortening) – 4D





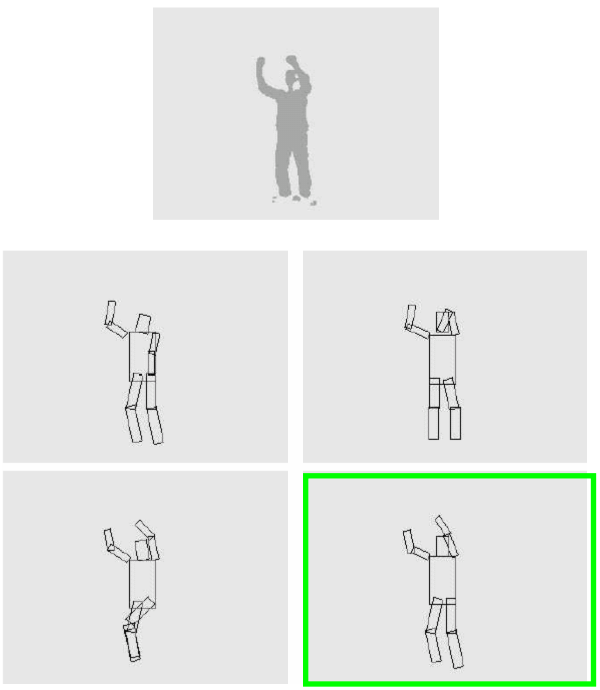
Sampling

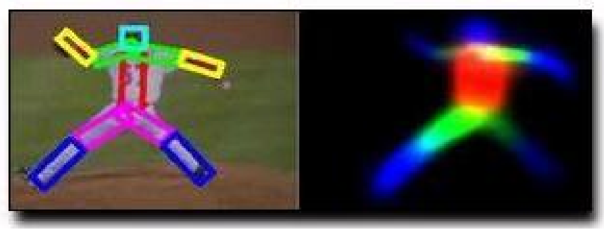
- Compute (factored) posterior distribution
  - Sampling for diversity not approximation
- Efficiently generate sample configurations
  - Sample recursively from a “root part”
- Approximation to POP distribution [AT07]
  - Likelihood that does not over count evidence for overlapping parts



Sampling For Human Body Pose

- Compute (factored) posterior distribution
- Efficiently generate sample configurations
  - Sample recursively from a “root part”





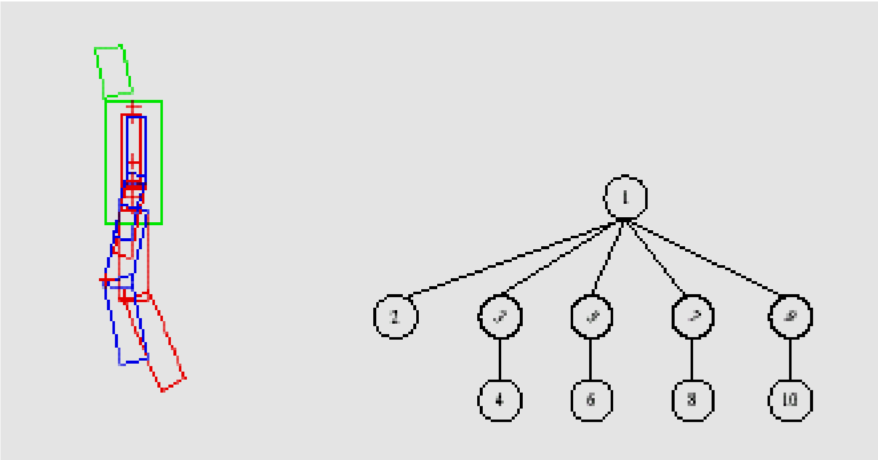
Used by 2D human pose detection techniques, e.g. [RFZ05]

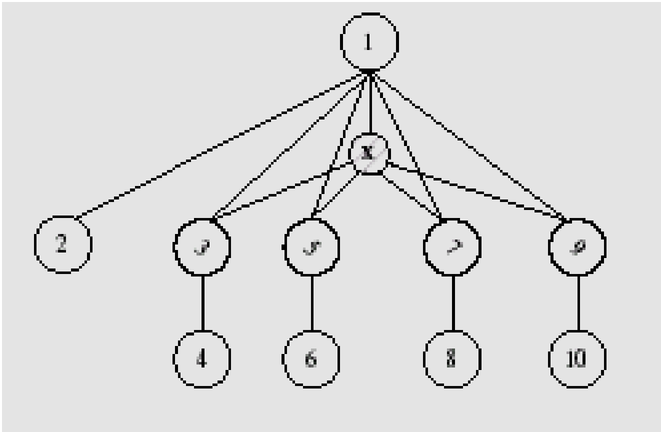
Spatial Structure in Model

- Going beyond trees while preserving computational tractability
- Adding latent variable(s) to models [LH05]
  - Correspond to overall model parameters rather than parts
  - Need to ensure no large cliques in resulting graph as computation increases exponentially
- K-fans [CFH05]
  - Generalization of star graph to root set of size k rather than single root node
  - Depth one and low tree width

A Latent Gait Variable for Humans

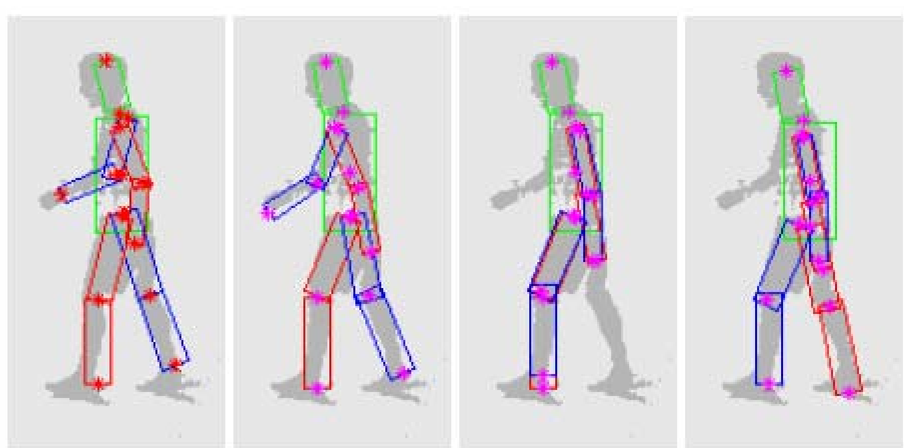
- Introduce additional variable corresponding to common factor [LH05]
  - Capture consistency between limb positions, not captured by kinematic (skeletal) model
    - Rather than directly connecting limbs which creates large clique



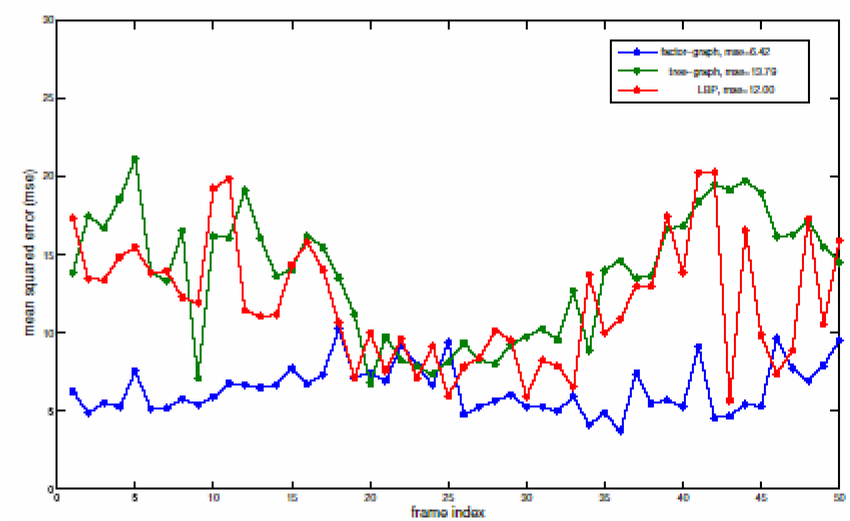


Latent Gait Variable Helps

- 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"



Ground Truth | Latent Var. | Tree | Large Clique



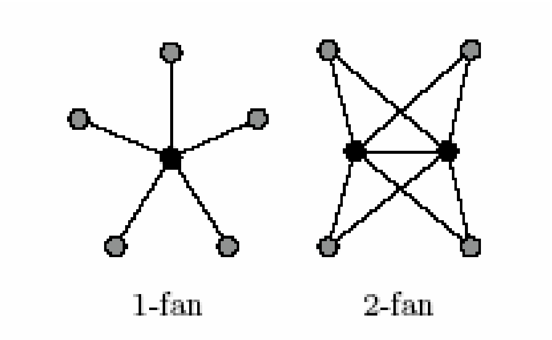
K-fan Models

- Prior factors according to graph of spatial constraints between parts

$$P_M(L) = \prod_C \Psi_C(L_C)$$

  – Product over maximal cliques of triangulated graph, L_C locations of corresponding parts

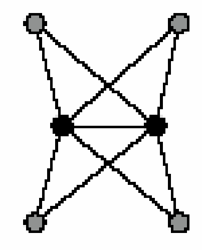
- K-fan generalizes star graph structure
  – Cliques of size k+1 for k central nodes
  – Exact discrete inference in $O(nm^k)$ time for n parts and m locations per part, using fast convolution methods



Spatial Prior for k-Fan

- Let $R \subseteq V$ be set of reference parts, "center"

$$P_M(L) = P_M(L_R) \prod_{v_i \in R'} P_M(\ell_i | L_R)$$



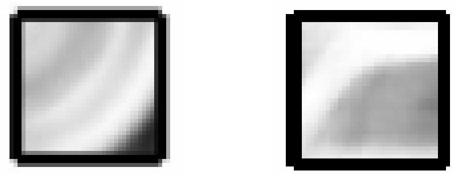
  - Where L_R vector of locations for R
    $L_R=(\ell_1, \ldots, \ell_k)$ for $R=(v_1, \ldots, v_k)$
- Makes explicit that part locations are independent conditioned on reference set
  - Product over non-reference parts, R'
- Geometric interpretation in terms of parts defining "reference frame"

Edge-Based Part Models

- Assume likelihood factors
  - Foreground product over parts
  - Background product over pixels

$$P_M(I|L) = \prod_i g_i(I,l_i) \prod_p b_p(I)$$

- Foreground model simple edge template
  - Probability of an edge at each pixel
  - Use vector of probabilities for four possible orientations
  - Slight dilation to account for discretization



Single Estimation Approach

- Single estimation more accurate (and faster) than sparse feature detection
  - Optimization for star or 2-fan [CFH05,FPZ05] vs. feature detection for joint Gaussian [FPZ03]
  - 6 parts under translation, Caltech-4 dataset
  - Single class, equal ROC error

	Airplane	Motorbike	Faces	Cars
Feat. Det. [FPZ03]	90.2%	92.5%	96.4%	90.3%
Est.-Star [FPZ05]	93.6%	97.3%	90.3%	87.7%
Est.-Fan [CFH05]	93.3%	97.0%	98.2%	92.2%

Learning Models

- [FPZ05] uses feature detection to learn models under weakly supervised regime
  - Know only which training images contain instances of the class, no location information
- [CFH05] does not use feature detection but requires extensive supervision
  - Know locations of all the parts in all the positive training images
- [CH06] weak supervision without relying on feature detection

Weakly Supervised Learning

- Consider large number of initial patch models to generate possible parts
  - Ranked by likelihood of data given part
- Generate all pairwise models formed by two initial patches
- Consider all sets of reference parts for fixed k
- Greedily add parts based on pairwise models to produce initial models
  - One per reference set

Learning Spatial Model

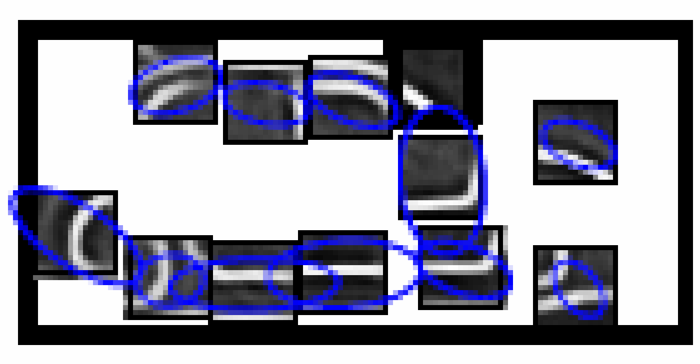
- Estimate pairwise spatial models for all pairs of patches – maximum likelihood
- Consider all k-tuples as root sets
- Use pairwise models to approximate true spatial model
  - Exact for 2-cliques (1-fan, star graph)
- Use EM to update model
  - Iteratively improve both appearance and spatial models

A More Accurate Form of Model

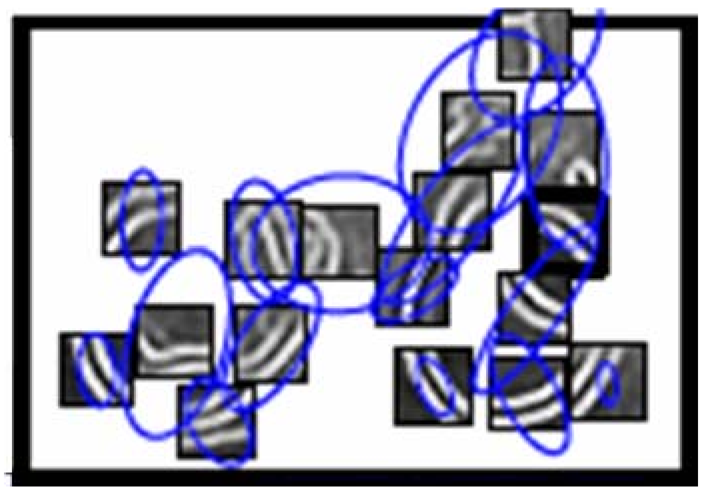
- Independent part appearance can over count evidence when parts overlap
  – Address by changing form of image likelihood
- POP – patchwork of parts [AT07]
  – More accurate model that accounts for overlapping parts
  – Average probabilities of patches that overlap
    - Distribution does not factor, can't compute efficiently
    - Can sample efficiently from factored distribution and then maximize POP criterion

Example Learned Models

- Star graph (one fan)
  - 24x24 patches
  - Reference part in bold box
  - Blue ellipse 2σ level set of Gaussian



Side View of Car



Side View of Bicycle

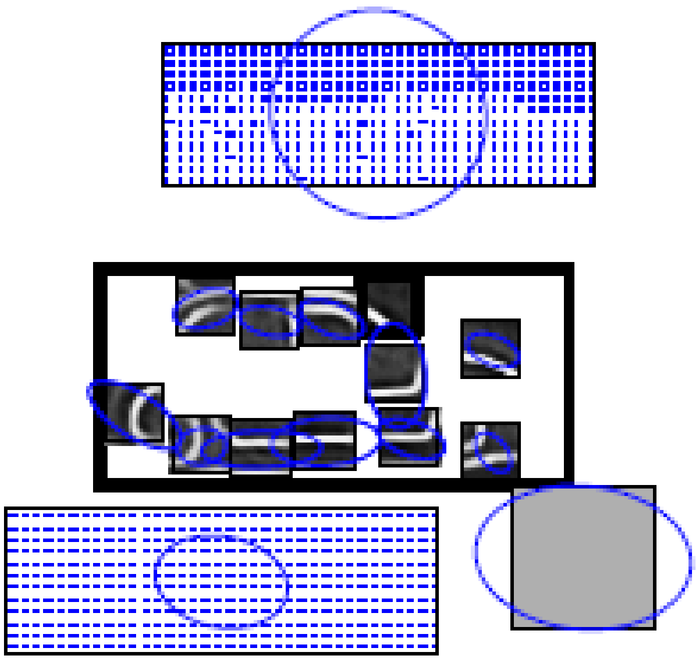
Adding Local Context to Models

- Spatial relations not only among parts of object but also object and background
  - E.g., vehicles on roads, often in front of buildings
  - Less predictable relative locations than object parts within a category
- Use coarser appearance models
  - Less predictable appearance of “scene parts”
- Augment spatial model using two-level hierarchy

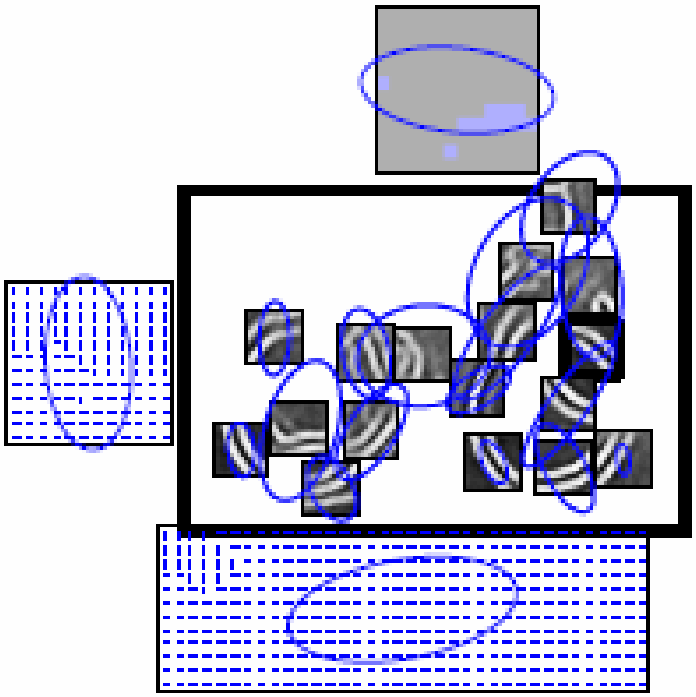
Composite Model

- Learn 1-fan (star graph) object model as before
- Learn 1-fan context model with bounding box as root and parts external to object
  - Lower resolution image
  - Various patch sizes
  - Edge, color and surface orientation descriptions
- Gaussian relating high resolution model root part to low resolution bounding box

Example Learned Models



Side View of Car



Side View of Bicycle

Recognition Results

- Four categories from PASCAL 06 VOC
  - Manmade objects: bicycle, bus, car, motorbike
  - Localization (detection) task
    - Search over translation and scale
    - Standard success measure used in VOC, overlap of detected object with ground truth > 50%
    - Report mean average precision
- Training with weak supervision
  - Use object bounding box
    - For scene model
    - To separate multiple instances in images

Comparison of Results

- Composite model with scene information substantially increases accuracy
- Better in terms of mean average precision than entries in VOC challenge
  - One method rather than several different methods

Object class	Obj. model only	Scene + obj. model	Best VOC result
Bicycle	0.421	0.498	0.440
Bus	0.172	0.185	0.169
Car	0.429	0.458	0.444
Motorbike	0.342	0.388	0.390

Example Results

Summary

- Detection and localization without doing feature detection
  - For common object class datasets, faster and more accurate than spatial models using feature detection
- Role of spatial structure
  - Latent structural variable such as human "gait" can substantially improve localization
- Role of local context
  - Including scene parts in model can substantially improve localization

More Details

P. Felzenszwalb and D. Huttenlocher, "Pictorial Structures for Object Recognition", Intl. J. of Computer Vision, v. 61, pp. 55-79, 2005.

D. Crandall and D. Huttenlocher, "Composite models of objects and scenes for category recognition", Proceedings of CVPR, 2007.