Graphical Models for Computer Vision

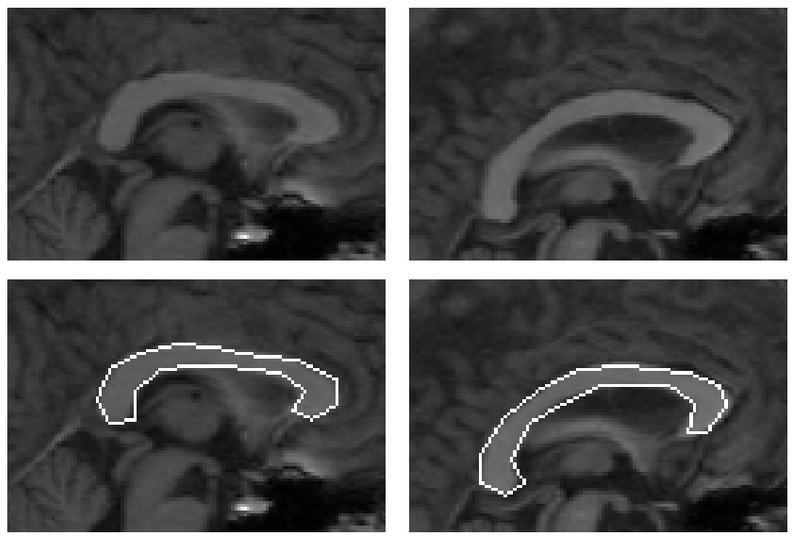
Pedro F Felzenszwalb
Brown University

Joint work with
Dan Huttenlocher, Joshua Schwartz,
Ross Girshick, David McAllester, Deva Ramanan,
Allie Shapiro, John Oberlin

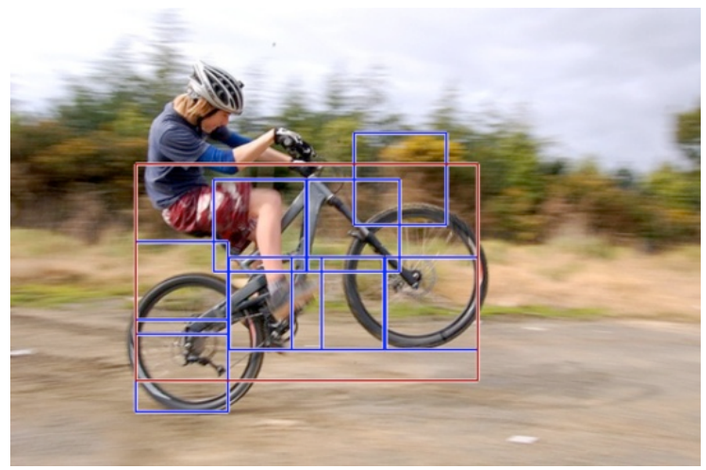
Vision Problems

Low-level vision ⟷ High-level vision

Restoration



Segmentation



Recognition

Depth estimation

Bayesian Framework

- Bayesian approach
  - We observe an image Y
  - Hidden variables X --- depth map, object labels, etc.
  - Vision involves statistical inference --- $P(X|Y)$
- Challenges
  - Building good models for X and Y
  - Thousands of random variables and large state spaces

Image Restoration

Object Detection

Multi-scale Models

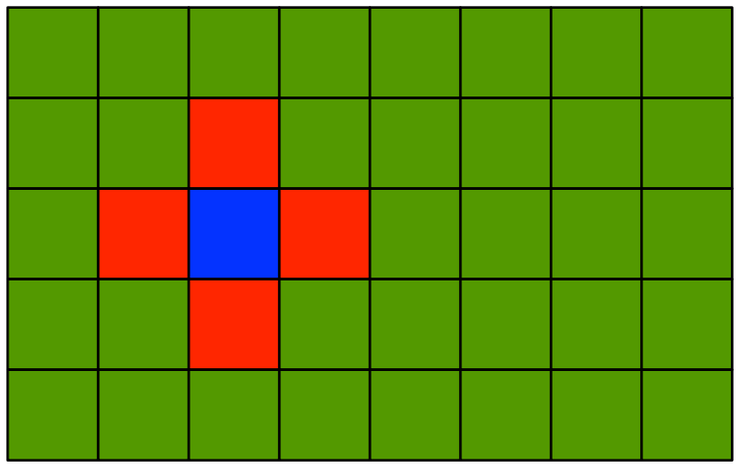
Image Restoration

- Random variables
  - X : clean picture
  - Y : observed image

X

Y

- $P(X)$: Markov random field
  - Nearby pixels tend to be similar
  - Markov blanket = 4 neighbors



- $P(Y|X)$: iid noise at each pixel
  - $Y_i = X_i + e_i$

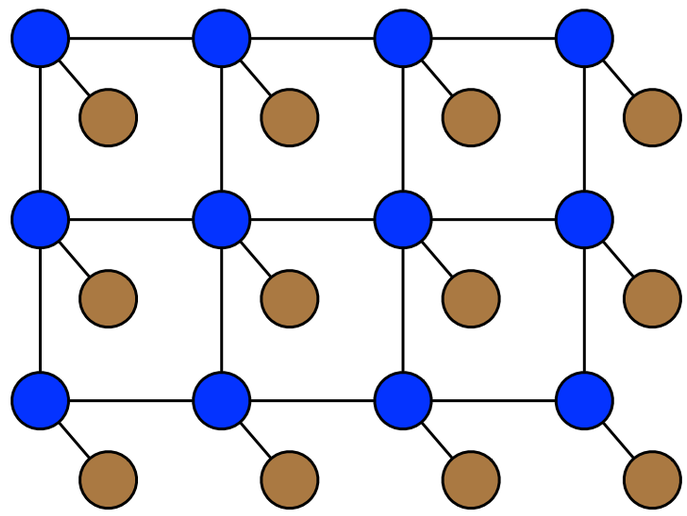
MAP estimation

- Minimize -log $P(X|Y)$

$$E(X) = \sum_i D(X_i, Y_i) + \sum_{ij} V(X_i, X_j)$$

  - D enforces consistency with the data (-log $P(Y|X)$)
  - V enforces smoothness (-log $P(X)$)

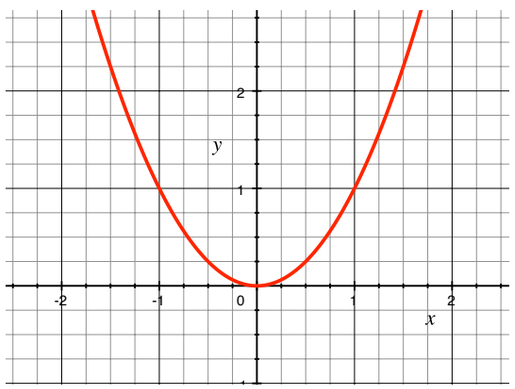
- Computational burden
  - huge number of variables
  - large state spaces
  - high treewidth

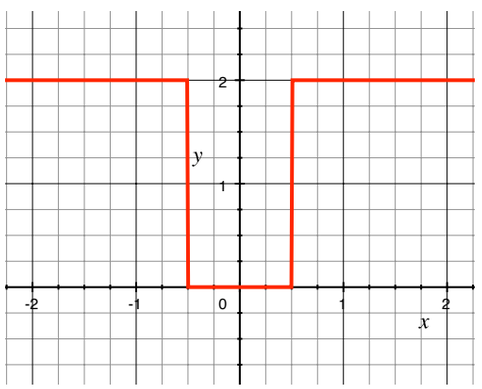


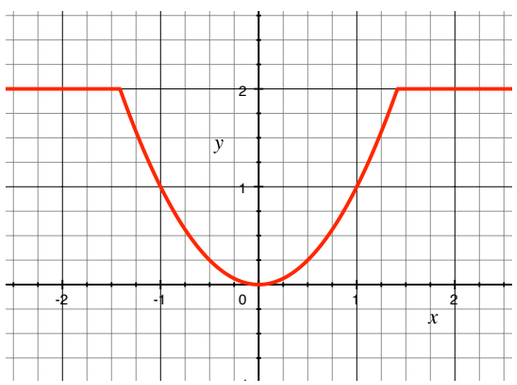
Discontinuity Costs

Y

MAP with different choices for $V(a,b) = W(a\text{-}b)$







Quadratic

X is smooth

Potts

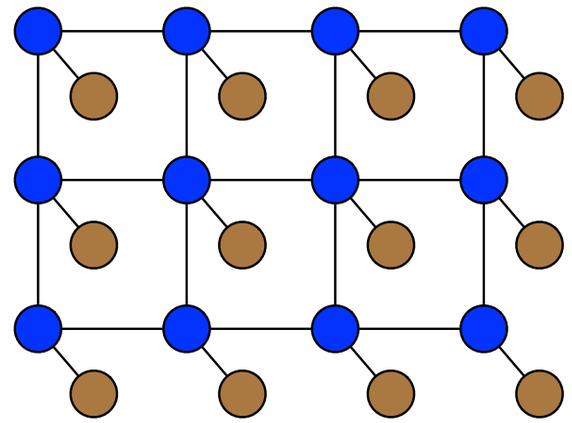
X is piecewise constant

Truncated quadratic

X is piecewise smooth

Computation

- MCMC (simulated annealing)
  - very general
  - slow...
- Graph-cuts
  - huge impact
  - works extremely well on restricted models
- Loopy belief propagation
  - very general
  - can be very fast

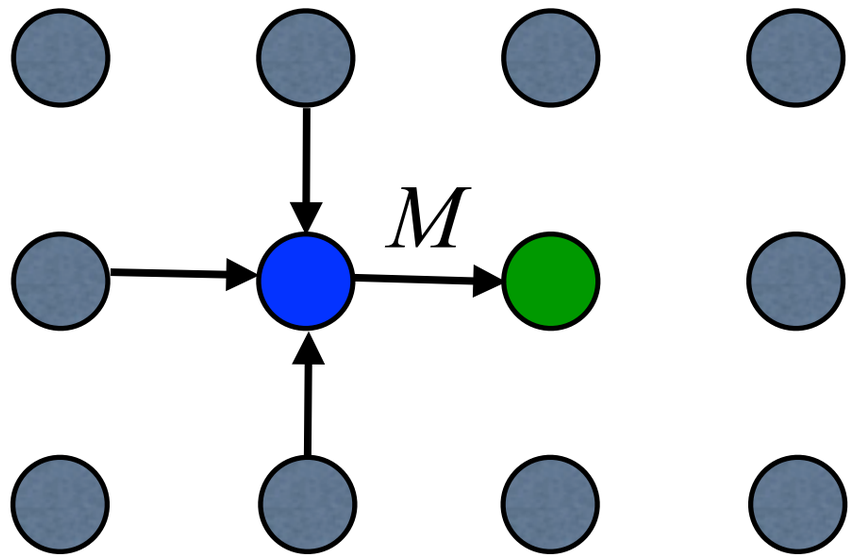


Runtime of Loopy BP

- Runtime depends on
  - Time for computing a message
  - Number of message updates for convergence

- Can exploit special problem structure to address both

Message Computation

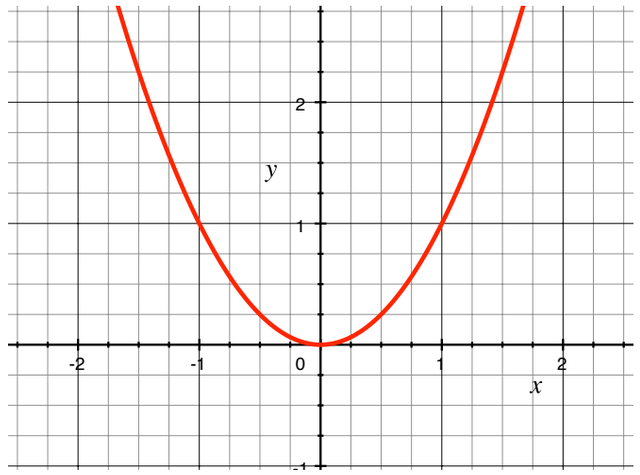


- $M(b) = \min_a(V(a,\mathrm{b}) + M_1(a) + M_2(a) + M_3(a) + D(a))$
- $M(b) = \min_a(V(a,\mathrm{b}) + H(a))$
  - k possible values for each pixel
  - $O(k^2)$ time by "brute-force"

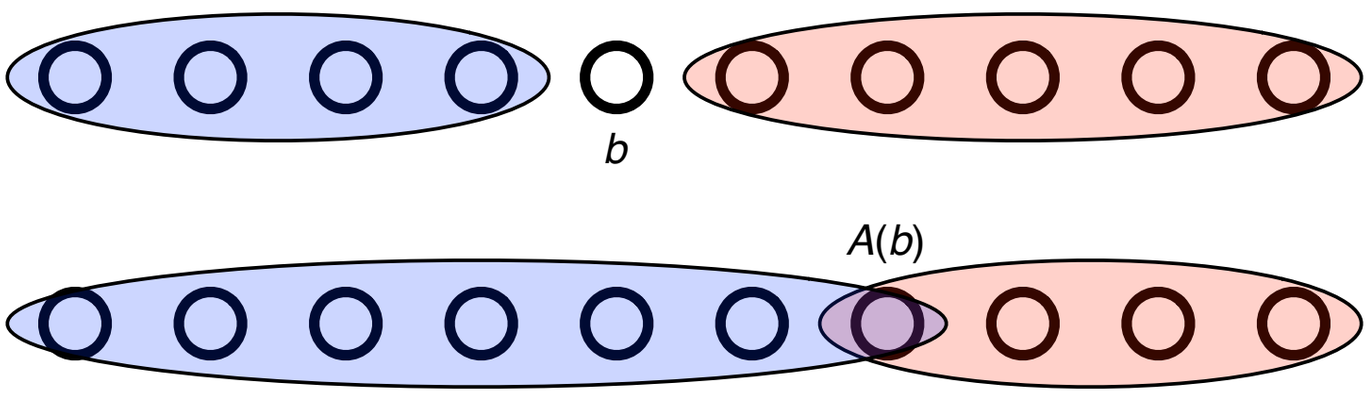
Fast Message Computation

- $M(b) = \min_a(V(a,\mathrm{b}) + H(a))$
  - States are integers and $V(a,b) = W(b\text{-}a)$
  - $M(b) = \min_a(W(b\text{-}a) + H(a))$
- Convolution of H and W in the (min,+) semi-ring
  - No known general fast algorithm like FFT
  - Best general algorithm $O(k^2/\log(k))$
  - Fast methods for restricted *W (we can pick W)*

Fast Min-Convolution



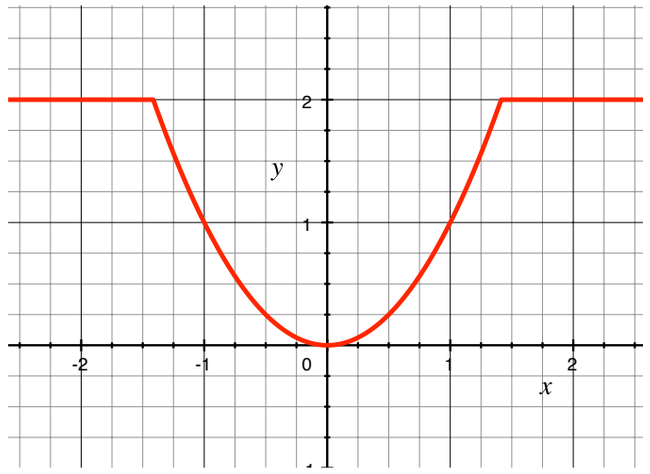
- $\hat{M}(b) = \text{argmin}_a(W(b\text{-}a) + H(a))$
- Assume $W(x)$ is convex
  - If $b' \geq b$ then $\hat{M}(b') \geq \hat{M}(b)$ --- "no crossings"
  - O(k log k) divide and conquer method



  - A little more work to get O(k) method

Fast Min-Convolution

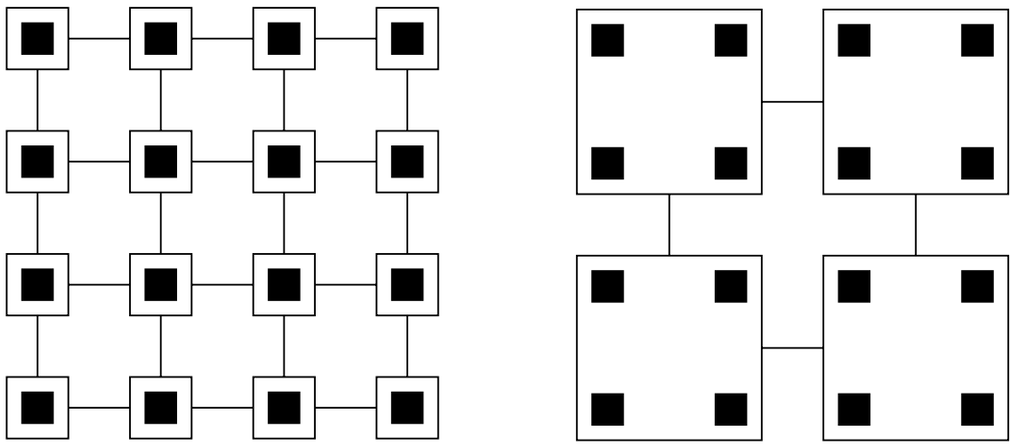
- If $W(x) = \min(E(x), F(x))$
  - $M_W(b) = \min(M_E(b), M_F(b))$



- For truncated quadratic W
  - E is quadratic
  - F is constant
  - Both convex - two O(k) computations plus O(k) to combine

Multi-Grid

- Number of updates for convergence is large
  - Information needs to propagate across the whole image
- Define a hierarchy of problems
  - Use messages from one level to initialize the one below
  - Good initialization leads to fast convergence



Hierarchical Algorithm

- Number of levels = log of image size
- LBP converges after ~10 iterations at each level (500x500 image)

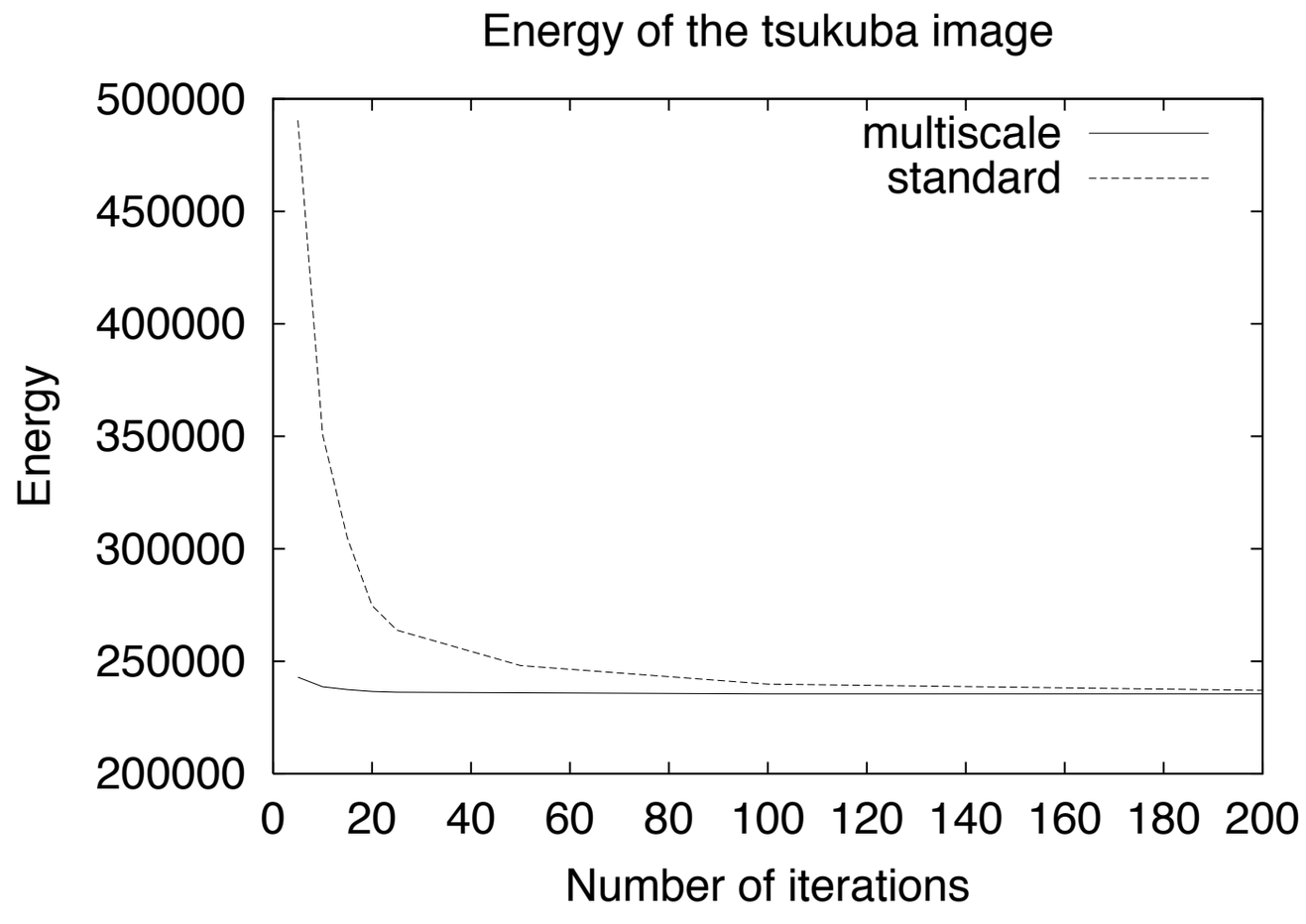
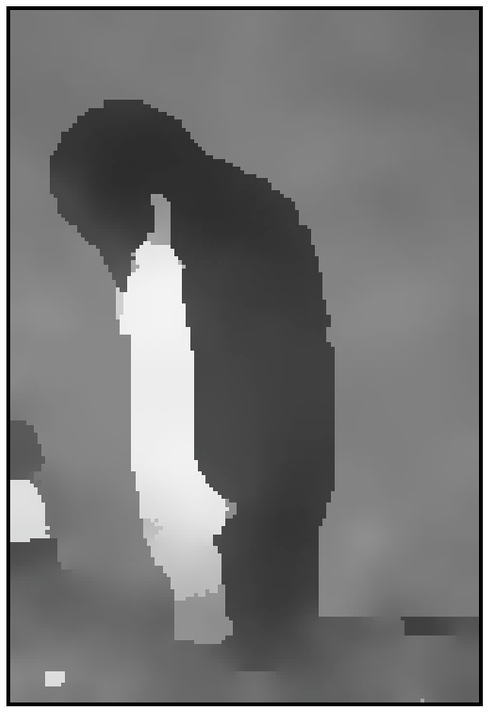


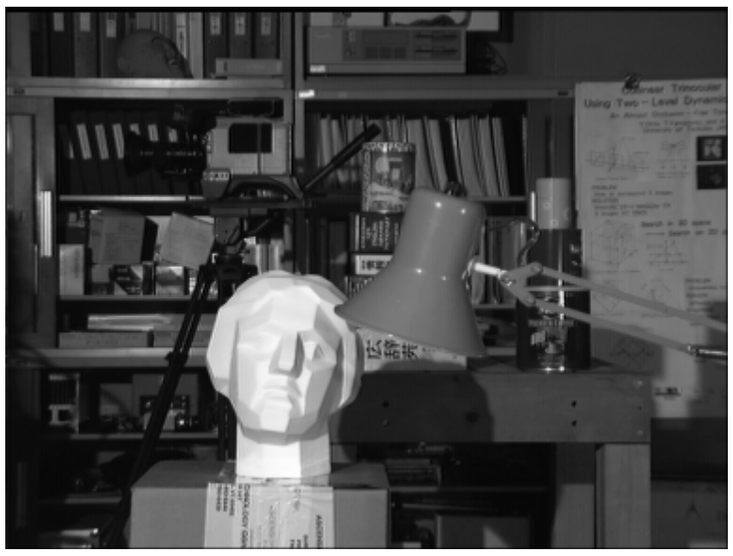
Image Restoration

- Truncated quadratic discontinuity costs
- Quadratic data terms, no data for masked pixels
- 256 states per pixel, propagating over large areas

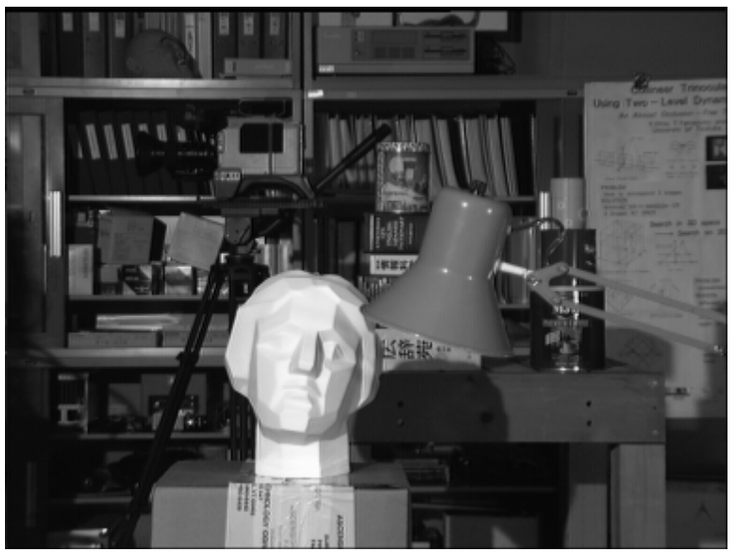


Stereo Depth Estimation

- State-of-the-art accuracy at frame-rate
- Simple, elegant model



left camera



right camera

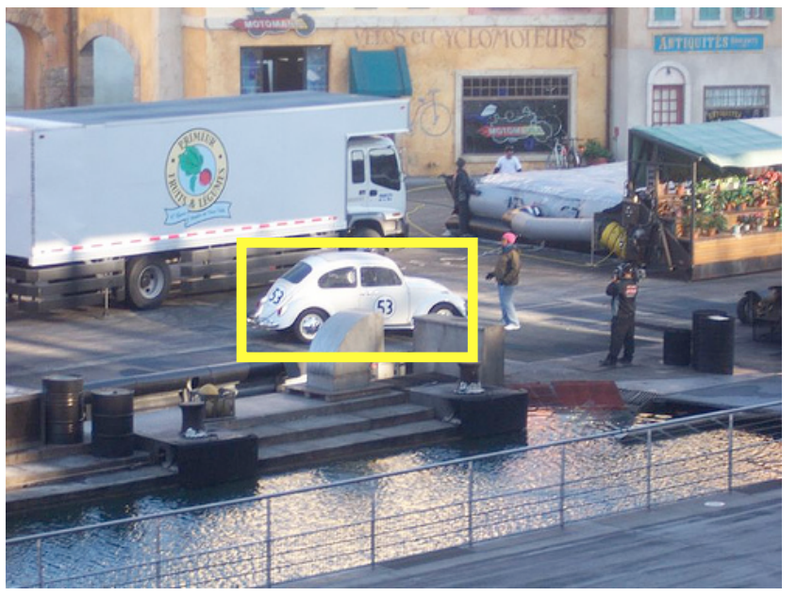
disparities

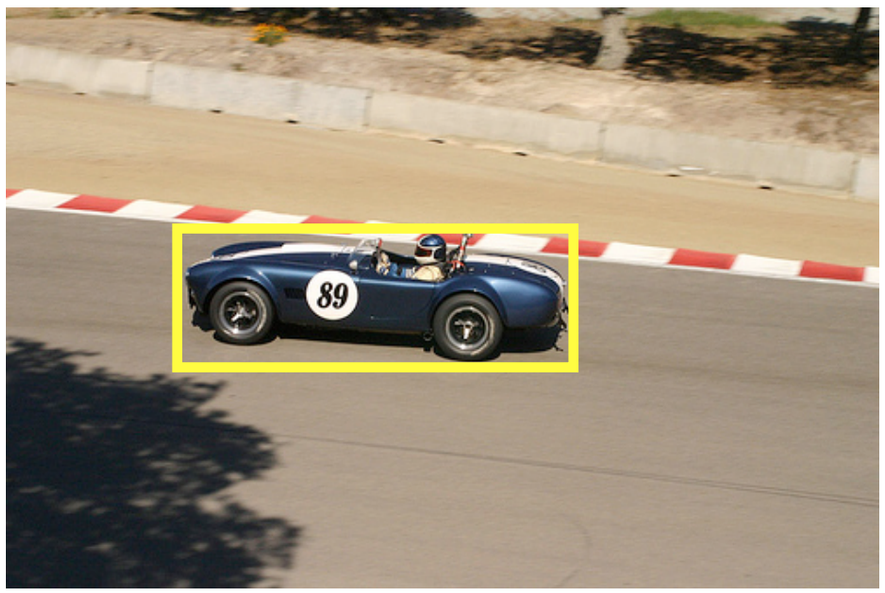
Image Restoration

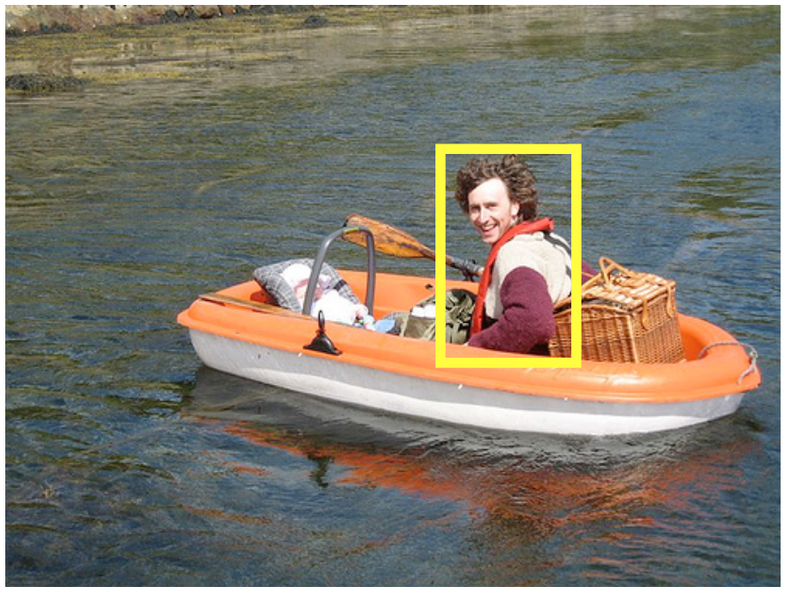
Object Detection

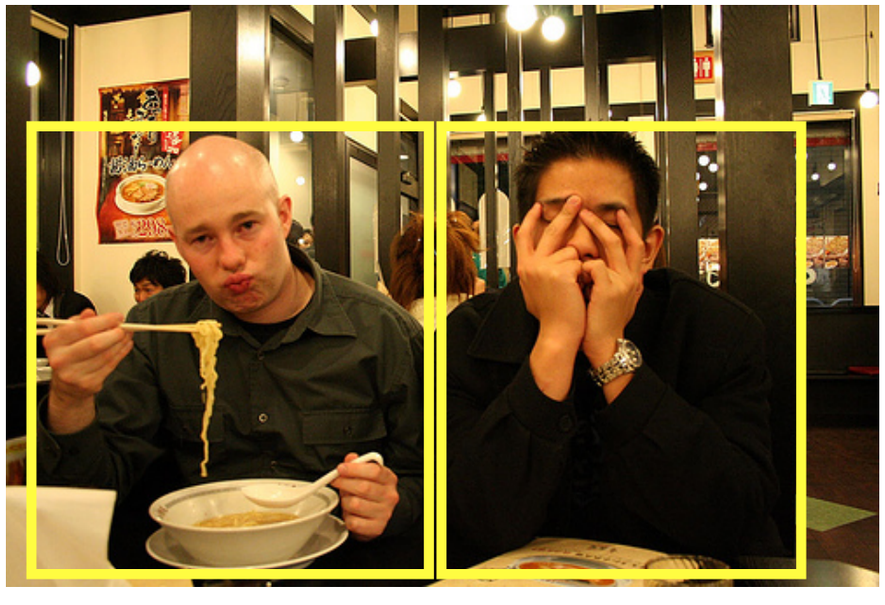
Multi-scale Models

Object Detection



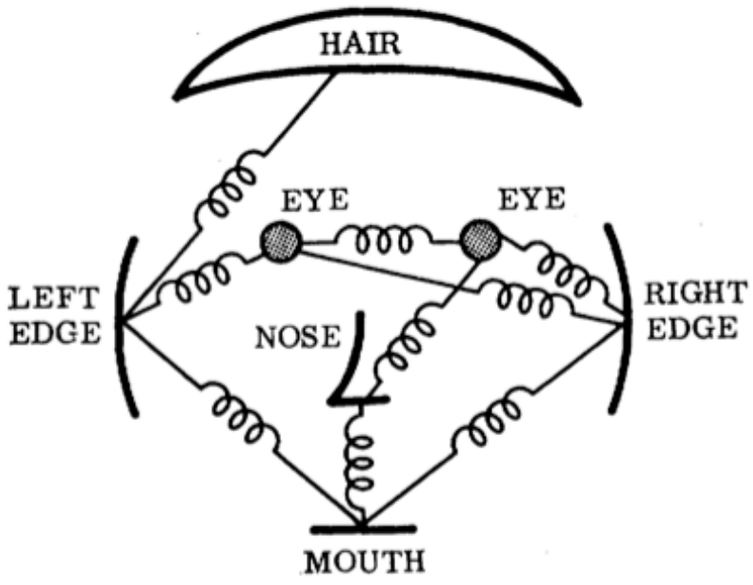






[PASCAL VOC dataset]

Part-Based Models

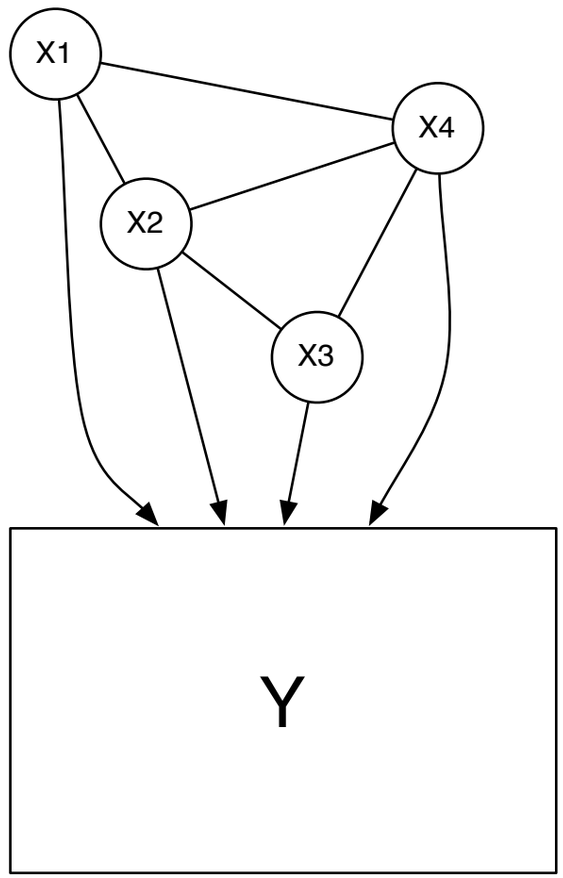


[Fischler, Elschlager 73]

- Each part has an appearance model
- Flexible geometric arrangement
  - Easier to model appearance of part than whole object
  - Factorization leads to better generalization

Graphical Model

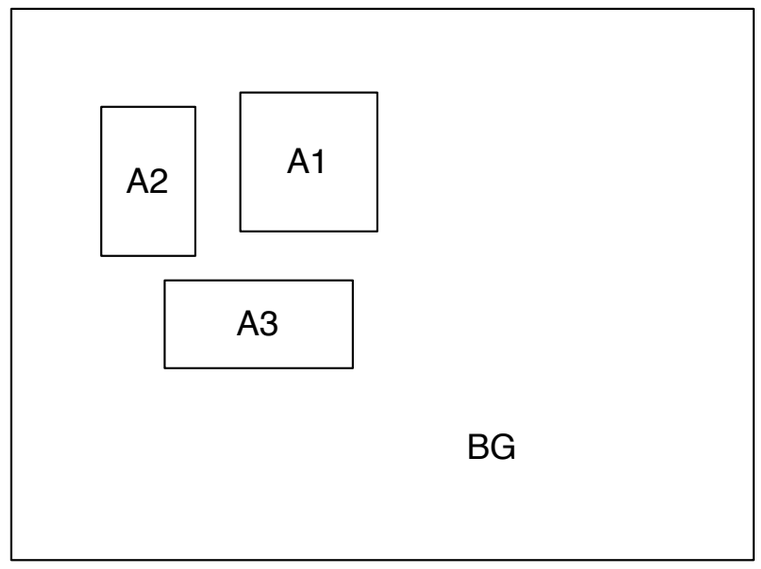
- Object with n parts
- Random variables
  - $X = (X_1 \ldots X_n)$: object configuration
  - X_i : location/pose of one part
  - Y : image



- $P(X)$: Markov random field
  - captures which geometric configurations are likely
- $P(Y|X)$: part appearance models + background model

Data Model

- We would like $P(Y|X)$ to factor
- Assume
  - 1) pixels (features) in background are iid
  - 2) parts don't overlap



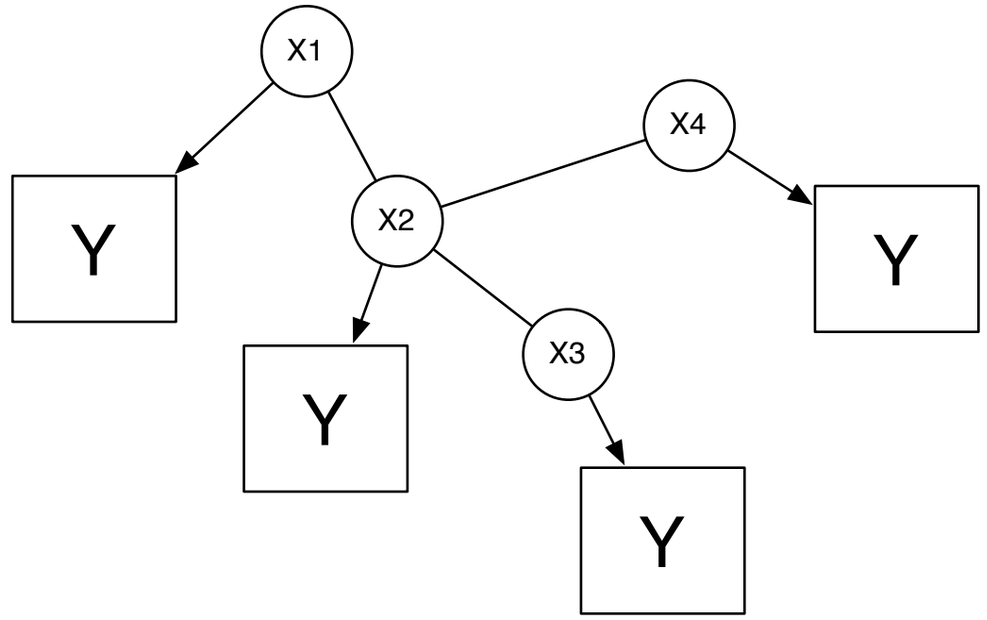
$$P(Y|X) = \left(\prod_i P_i(A_i)\right) P_{bg}(BG)$$

$$P(Y|X) = \left(\prod_i P_i(A_i)/P_{bg}(A_i)\right) P_{bg}(Y)$$

$$P(Y|X) \propto D_i(X_i)$$

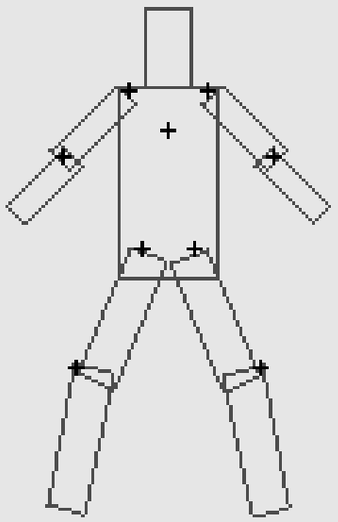
$\mathrm{BG} = Y \setminus \{A_1, .., A_3\}$

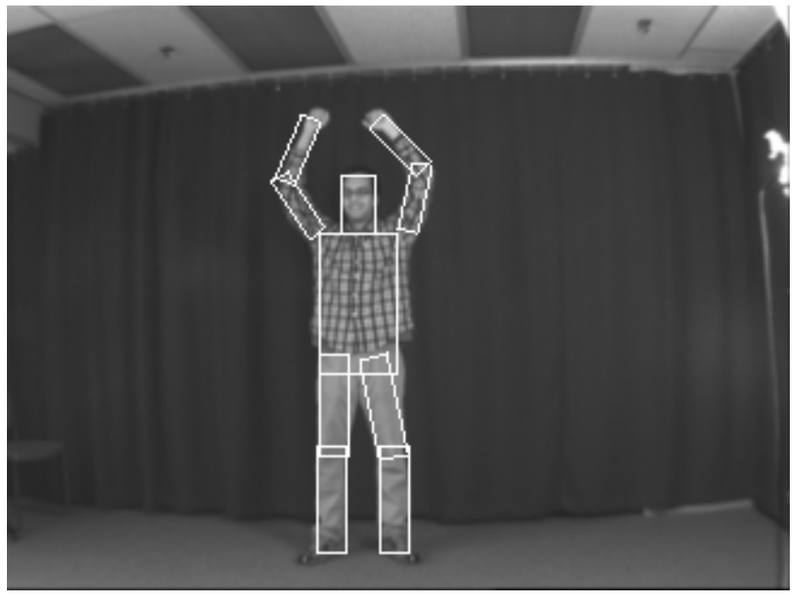
Inference

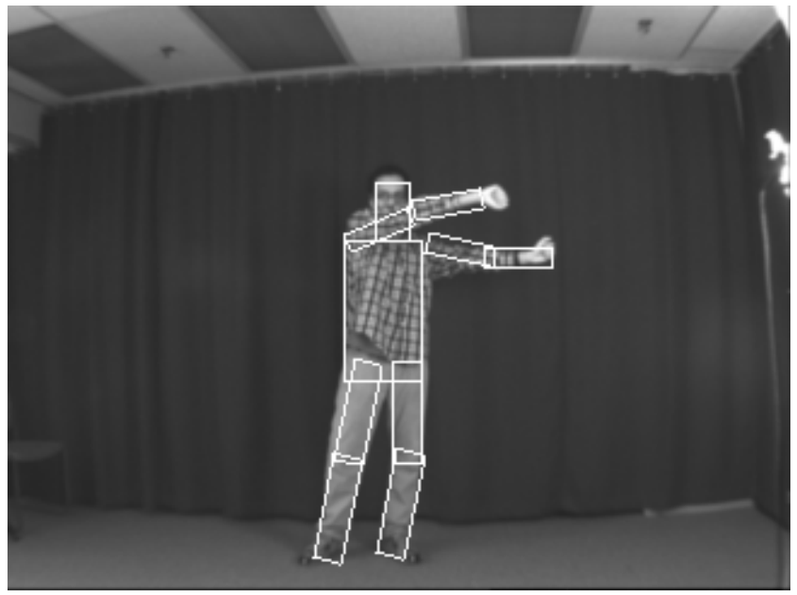


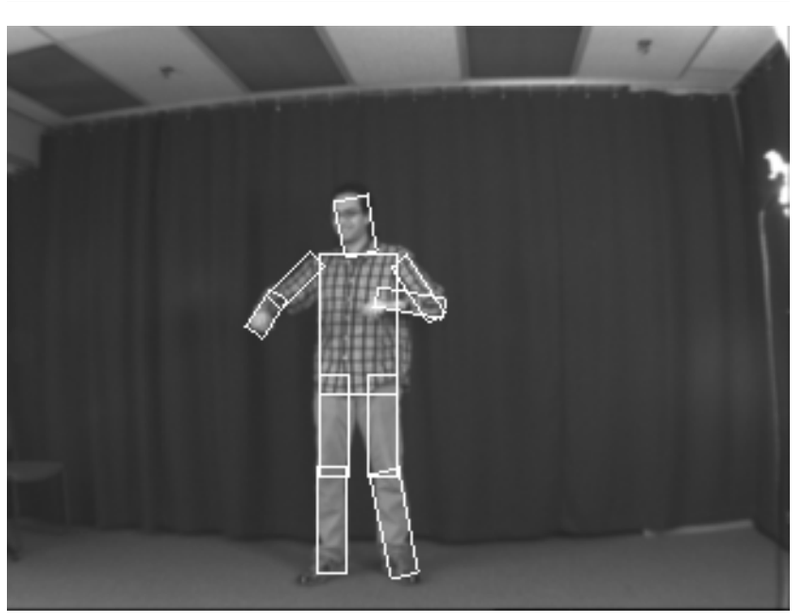
- Fully factored data model
- Further assume $P(X)$
  - tree-structured
  - pairwise relative positions
- Fast MAP estimation using min-convolutions
  - O(nk) time , n = number of parts, k = state space
  - As fast as detecting each part separately

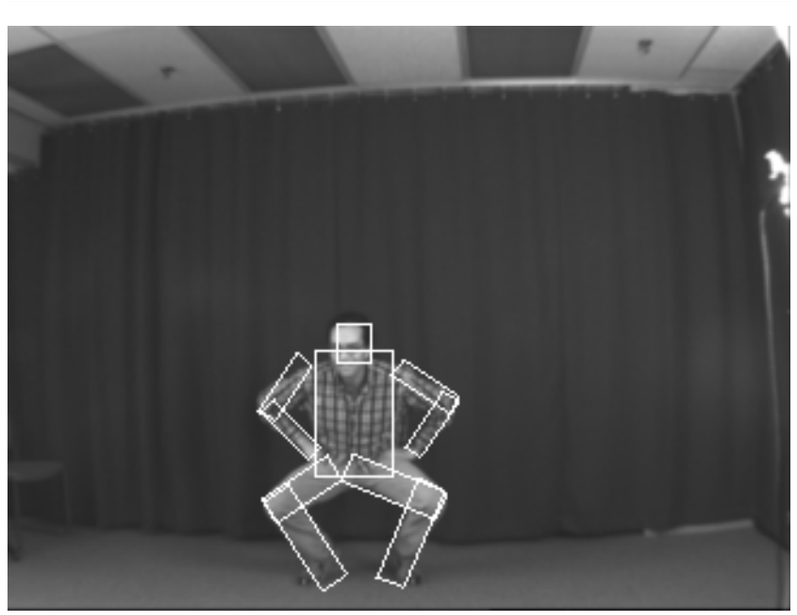
Human Pose Estimation



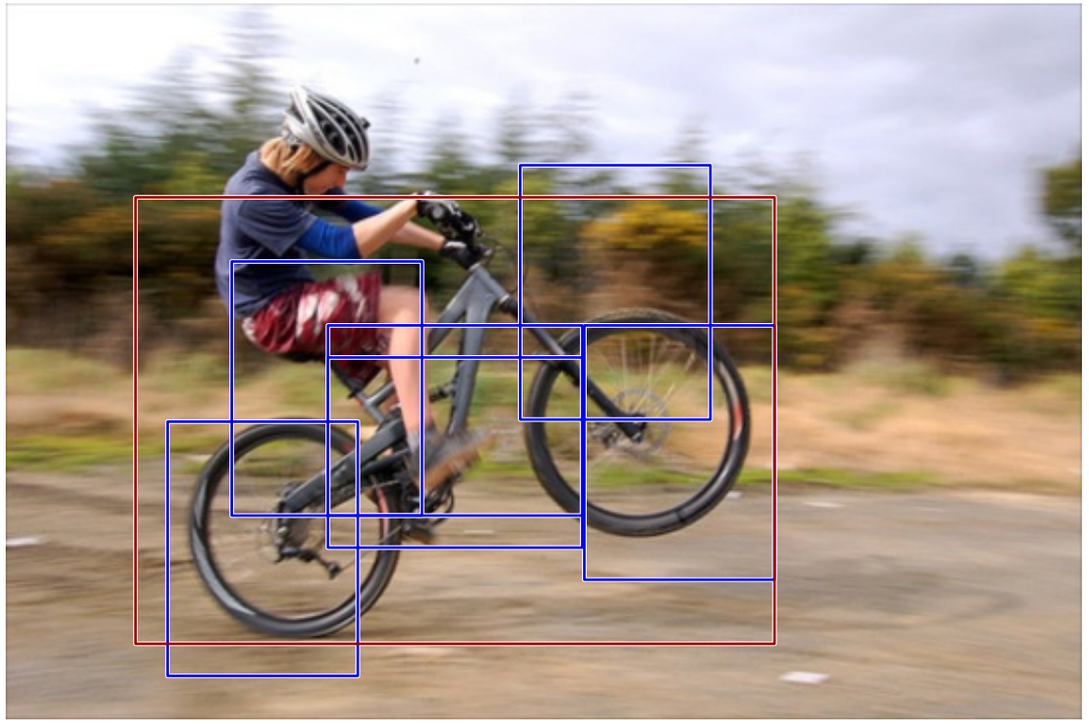


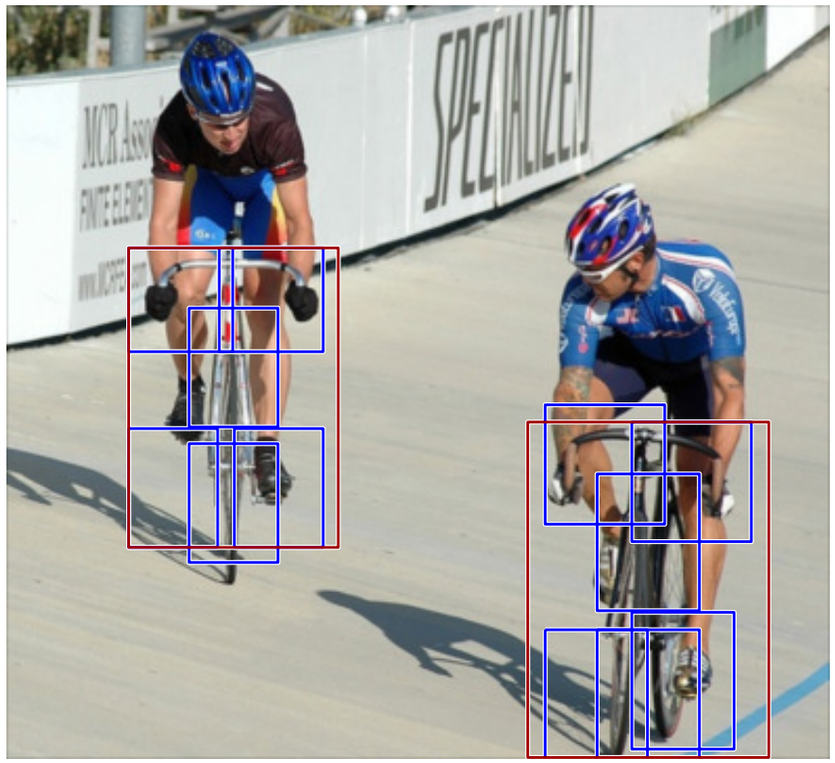






Object Category Detection



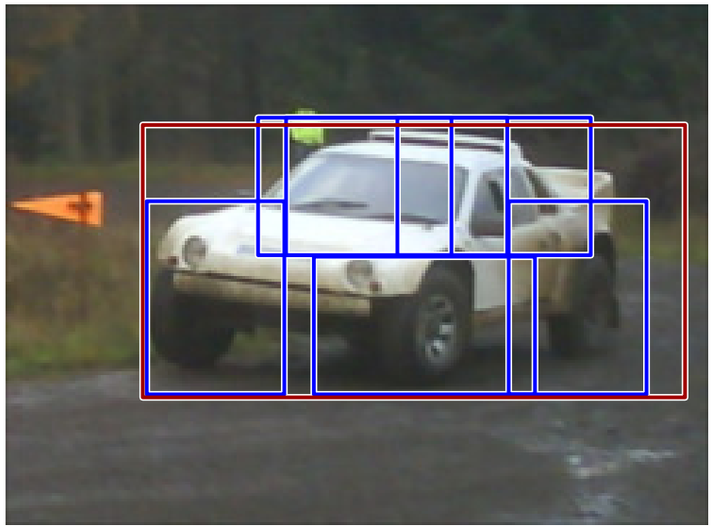


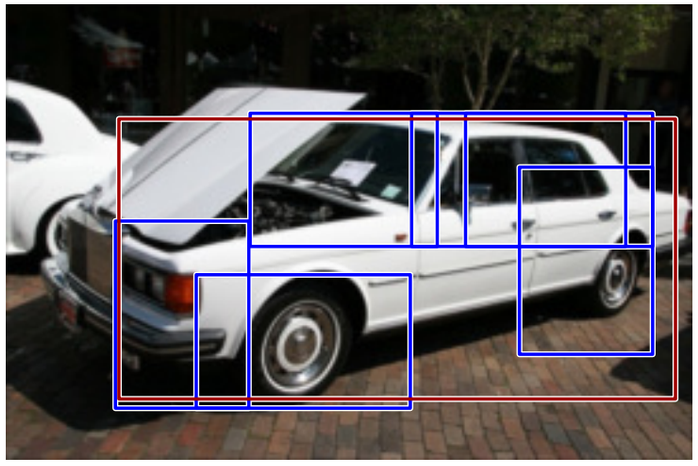
- Mixture of part-based models for each category
- Photometric invariant features (HOG)
- Discriminative learning (Latent SVM)
- Leading approach on PASCAL VOC benchmark

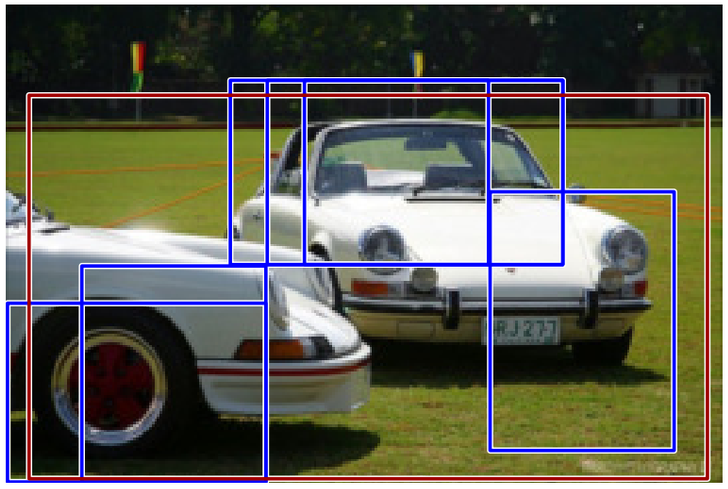
Car

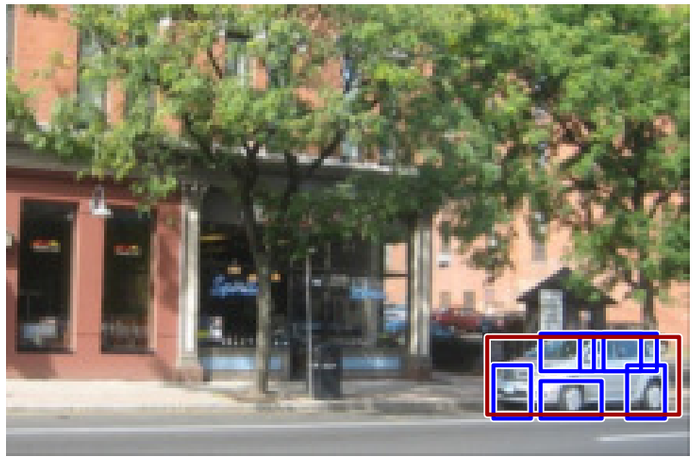
high scoring true positives

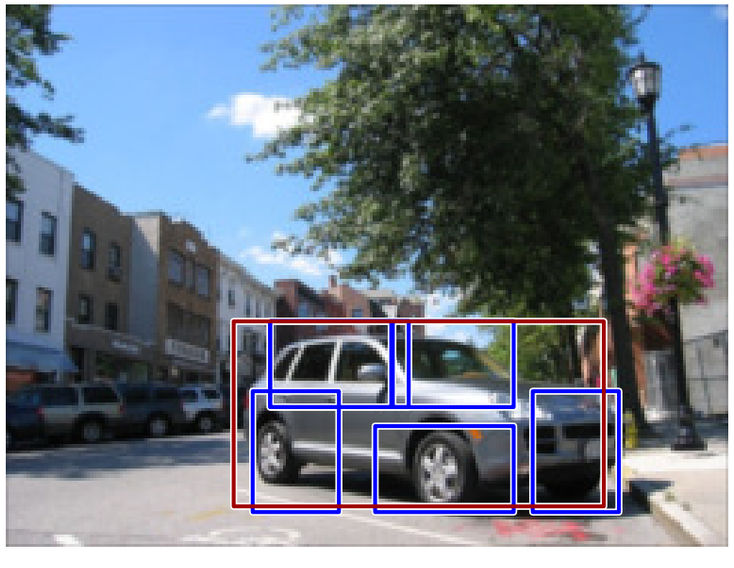
high scoring false positives

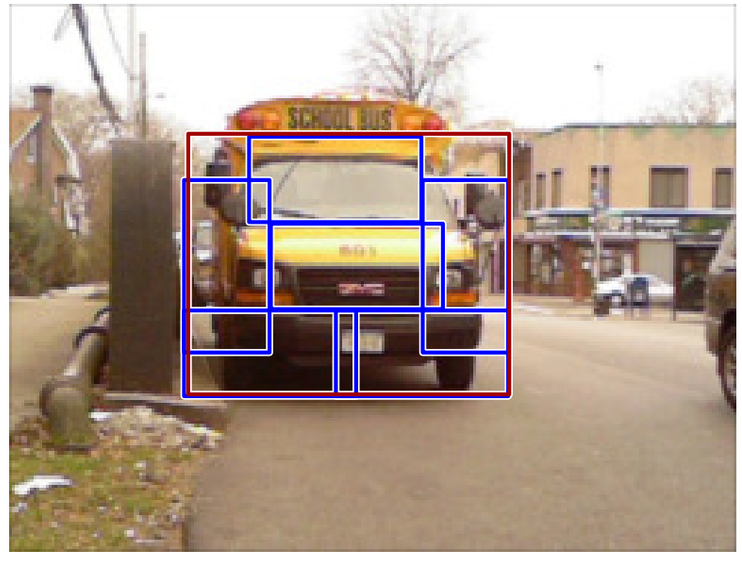










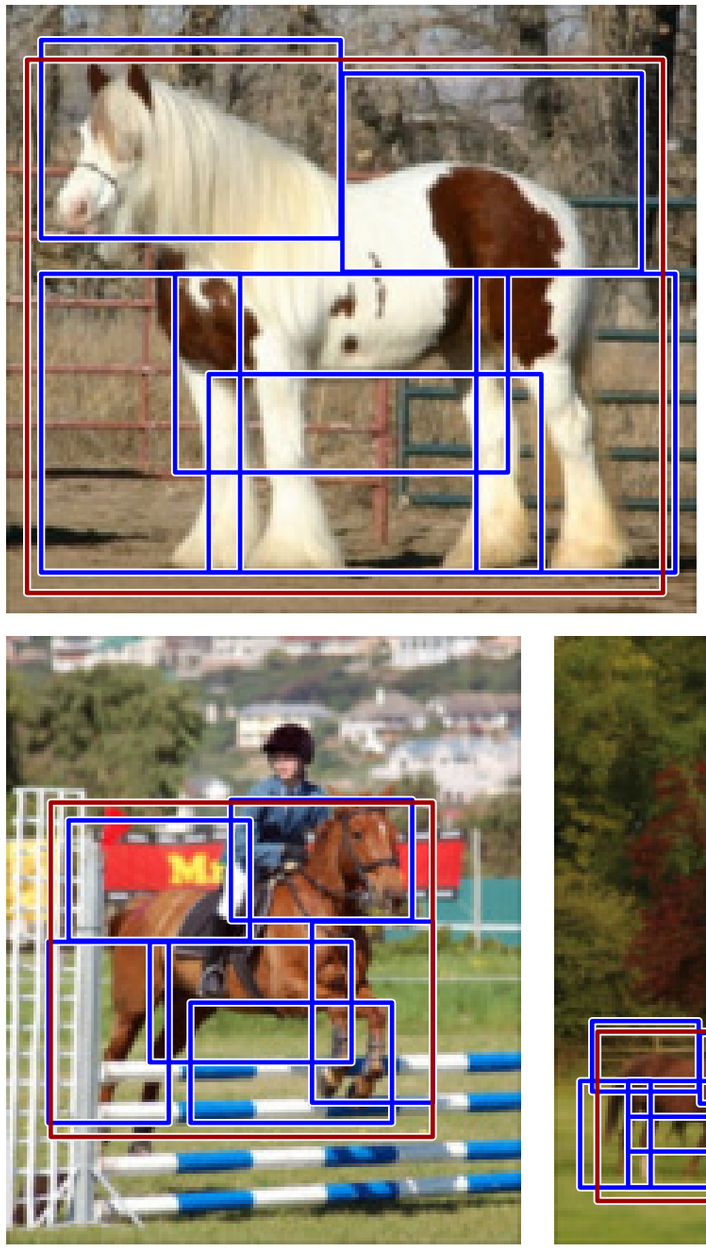


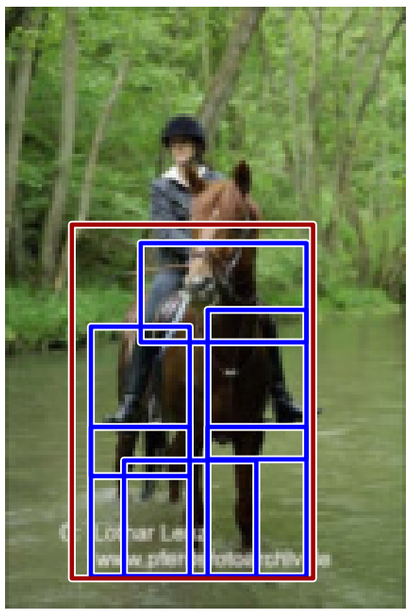
[PASCAL VOC dataset]

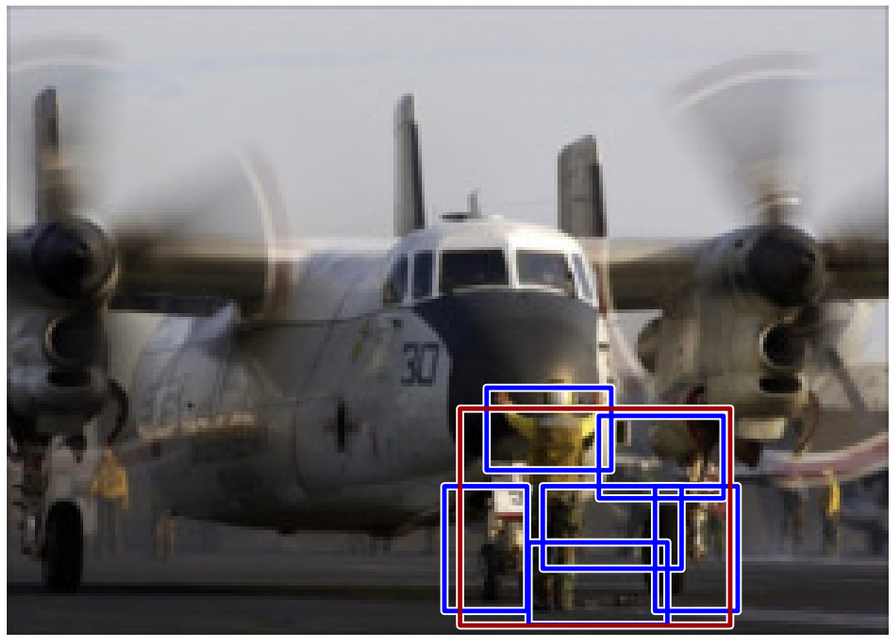
Horse

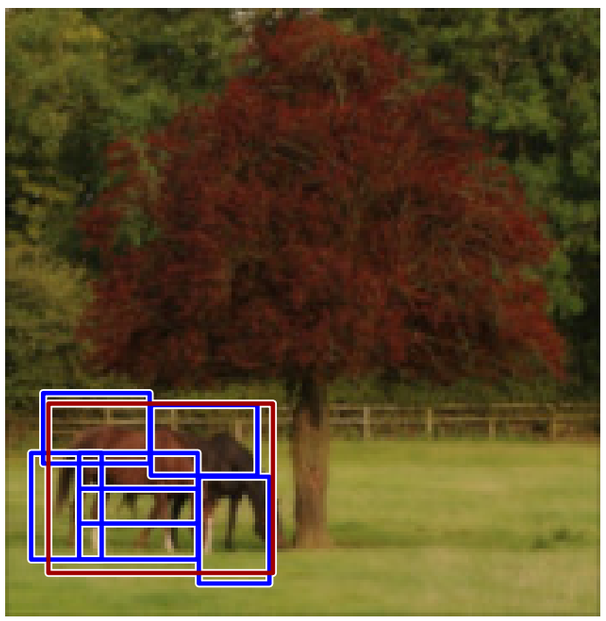
high scoring true positives

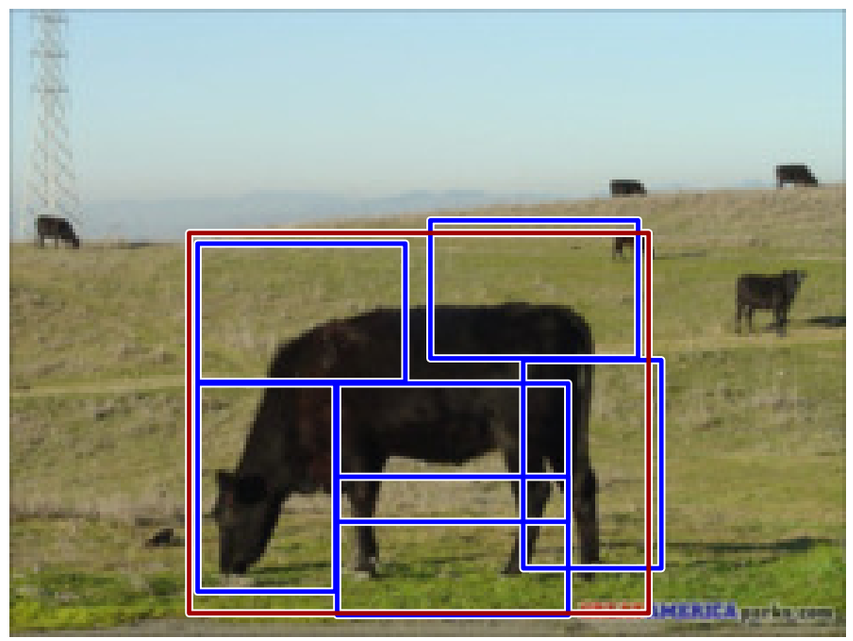
high scoring false positives











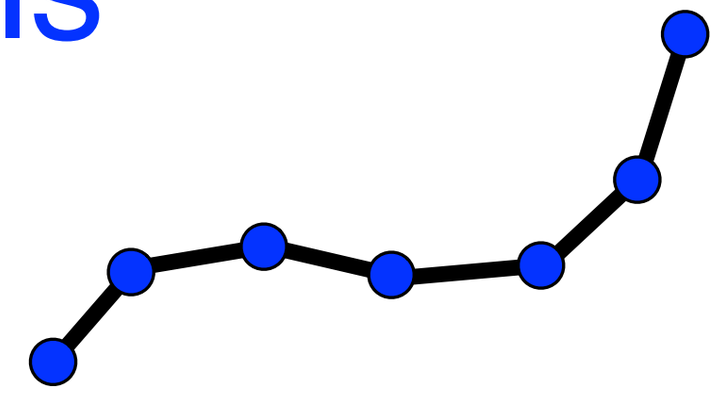
[PASCAL VOC dataset]

Image Restoration

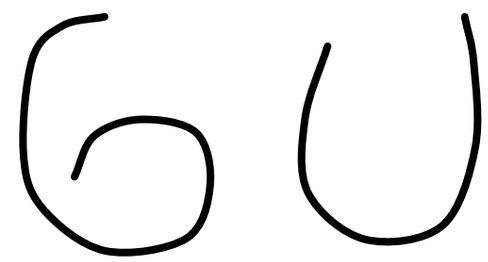
Object Detection

Multi-scale Models

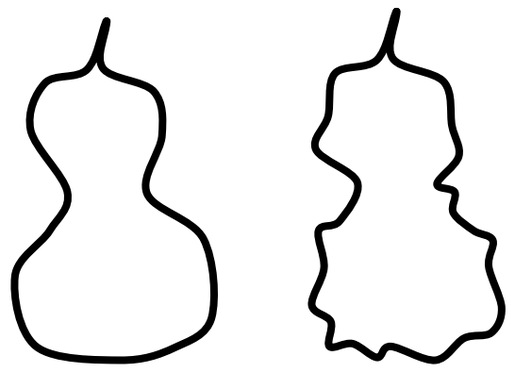
Curve Models



- Model for curve
  - Sequence of control points
  - Markov model $P(X)$ captures local shape
  - Drift: hard to control accumulation of local variation



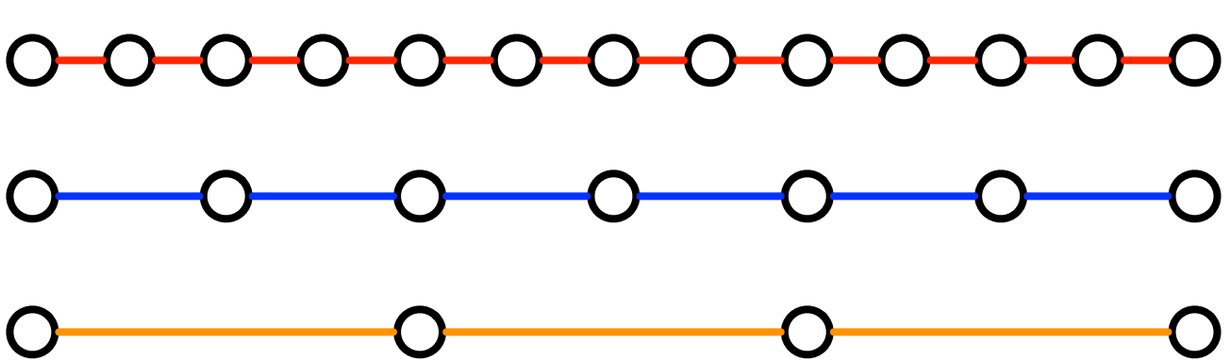
Locally these look very similar

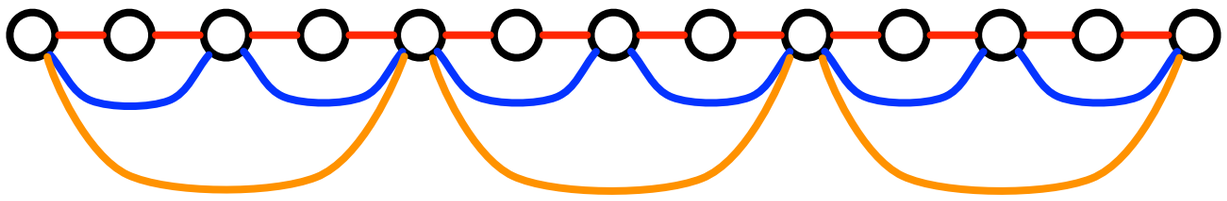


Locally these look very different

Multi-Scale Sequence Model

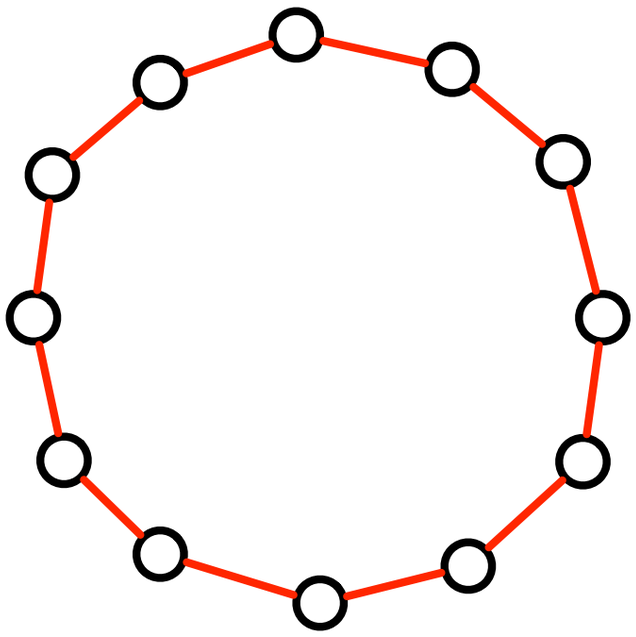
- Capture local properties at multiple resolutions
  - Subsample A to get B
  - local property of B = non-local property of A



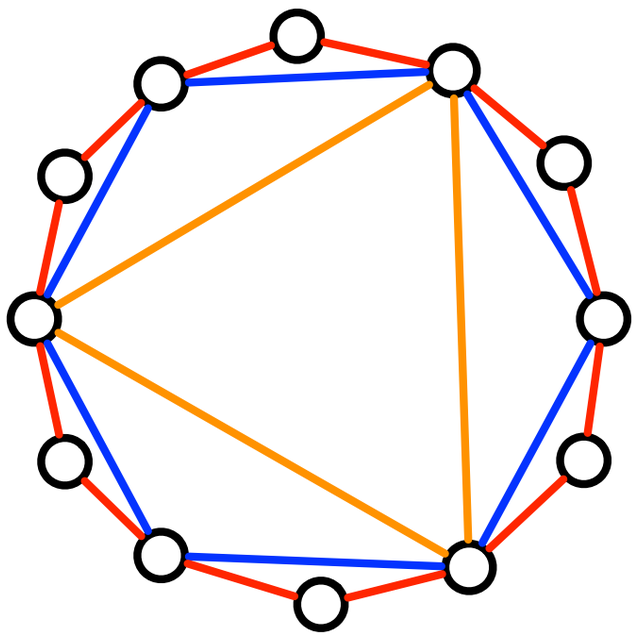


full model
tree-width = 2

Models for Closed Curves



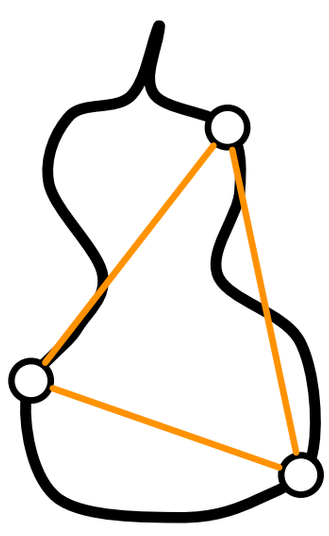
1st order Markov model

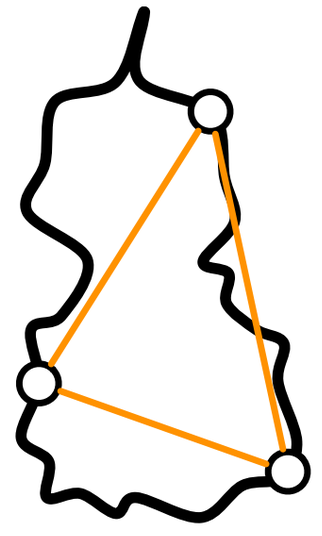


Multi-scale model

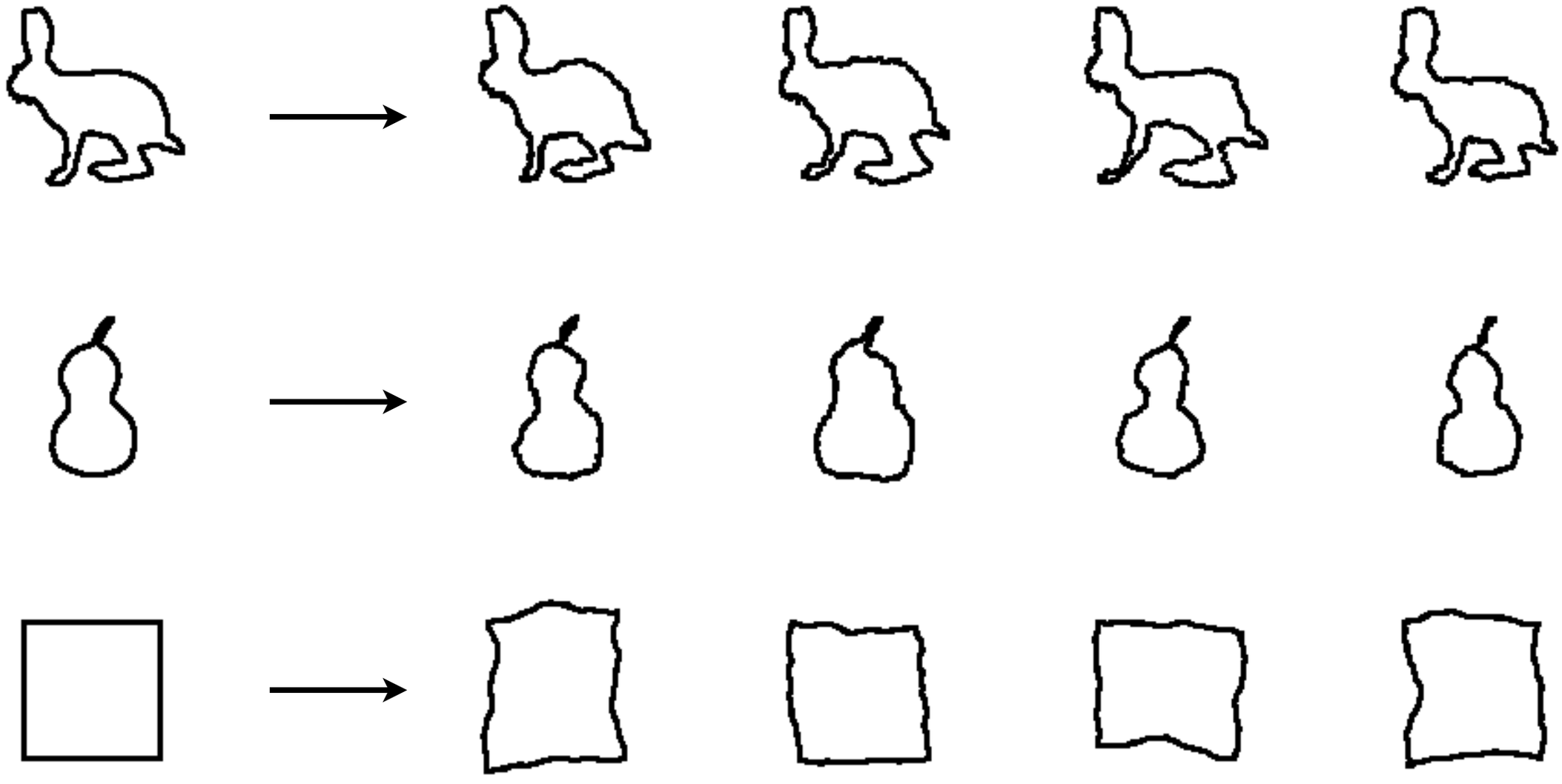
Both graphs have tree-width 2

Multi-scale model captures global shape properties





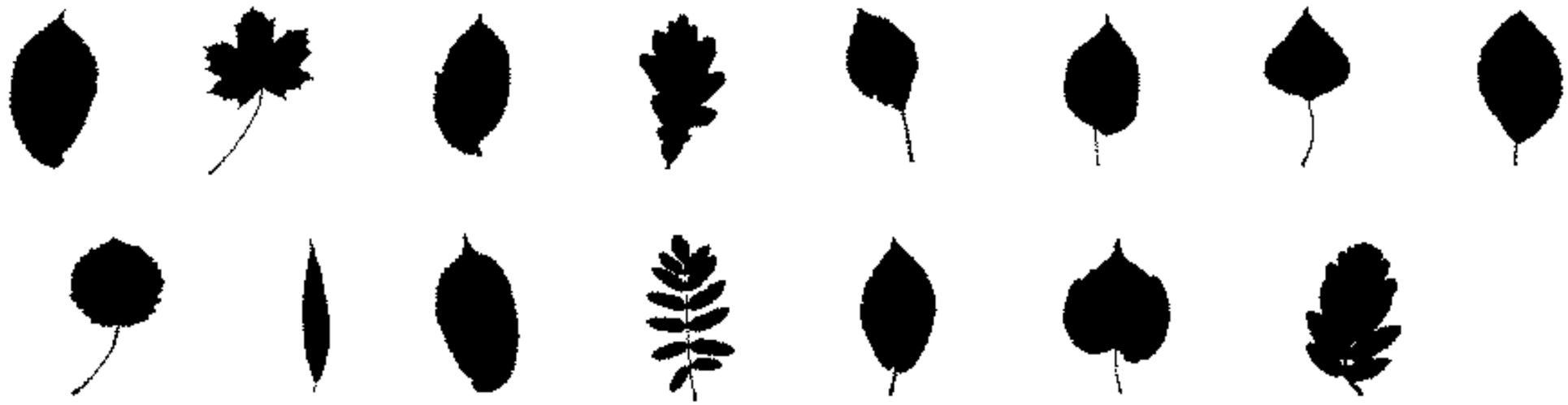
Samples from P(X)



Multi-scale model captures global shape properties

Shape Recognition

Swedish leaf dataset



Nearest neighbor classification	
Multi-scale model	96.28
Inner distance	94.13
Shape context	88.12

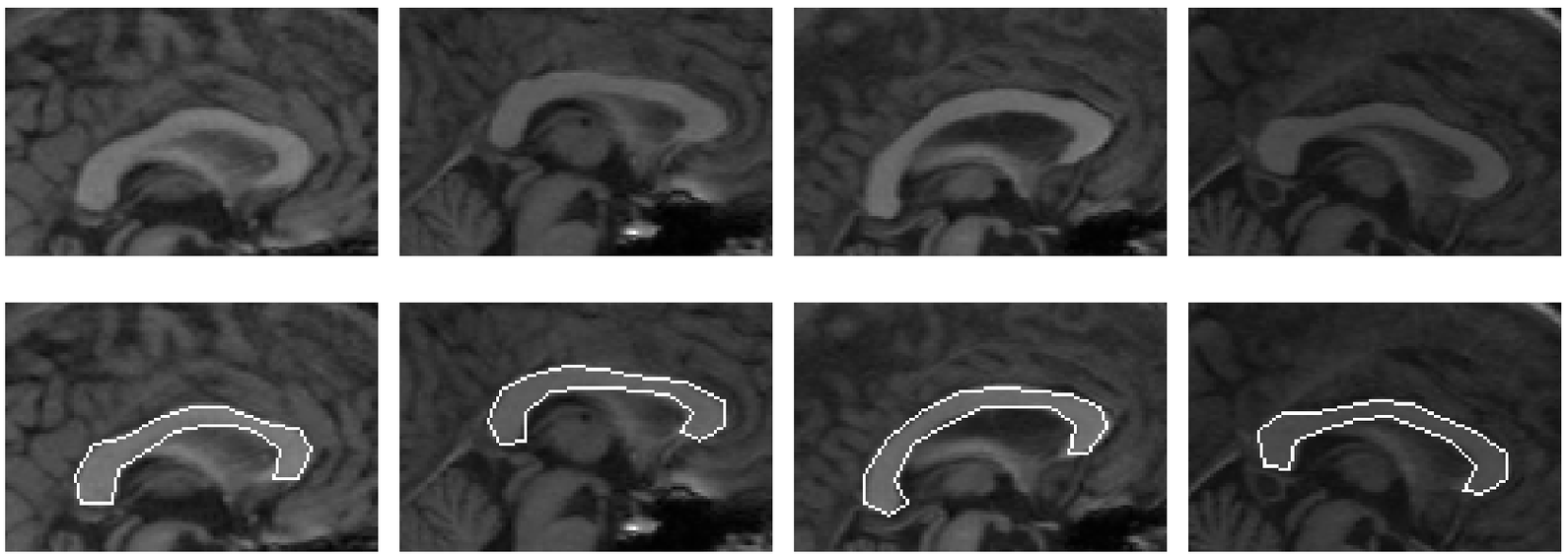
15 species

75 examples per species

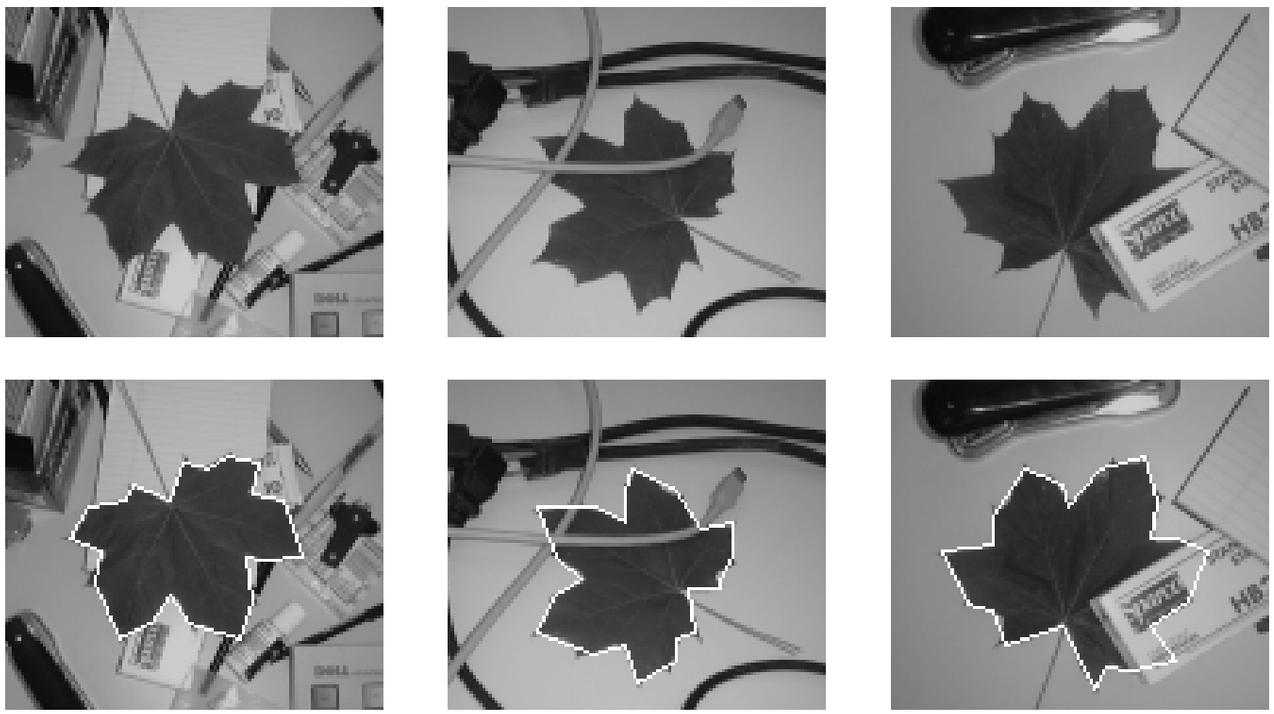
(25 training, 50 test)

Shape Detection

Model



Model



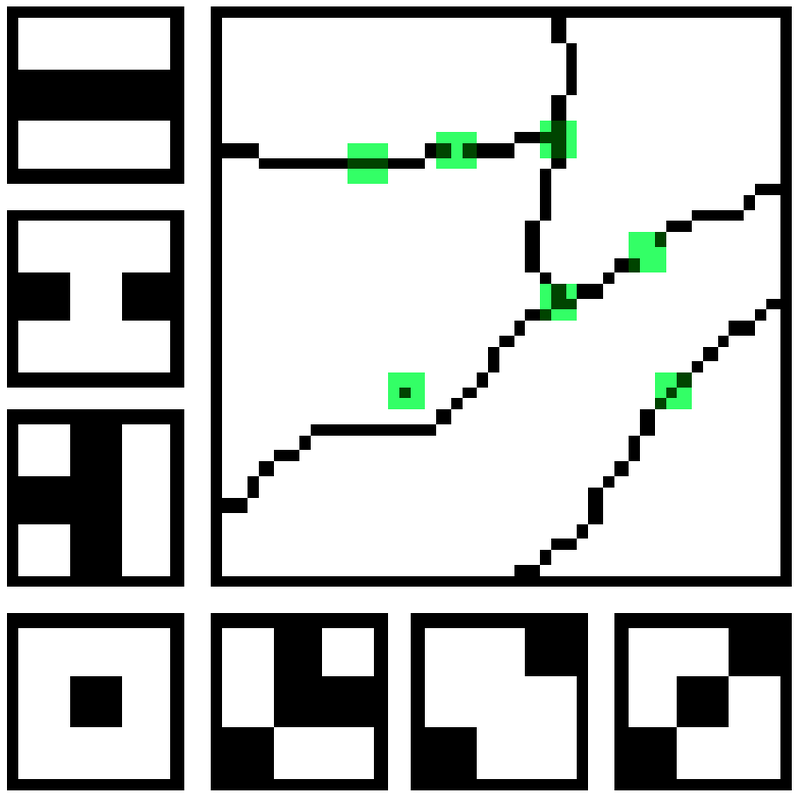
Boundary Detection

[BSDS]

- Lots of regularities
  - continuity, smoothness, closure, parallel lines, symmetry
- Can we build a “low/mid-level” model for $P(X)$?

Local Patterns

- Look at each 3x3 patch C
  - 512 possible patterns



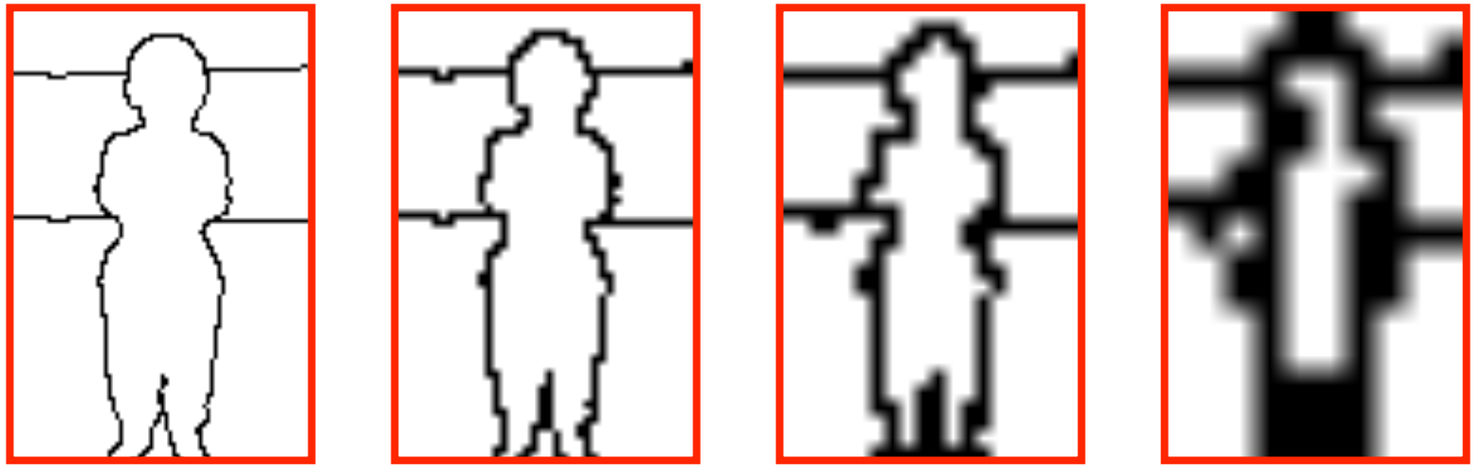
- Energy model

$$E(X) = \sum_C V(X_C)$$

  - Parameterized by 512 costs
  - Symmetries reduce to ~100

- Capture continuity, frequency junctions, etc.

Coarse Local Patterns

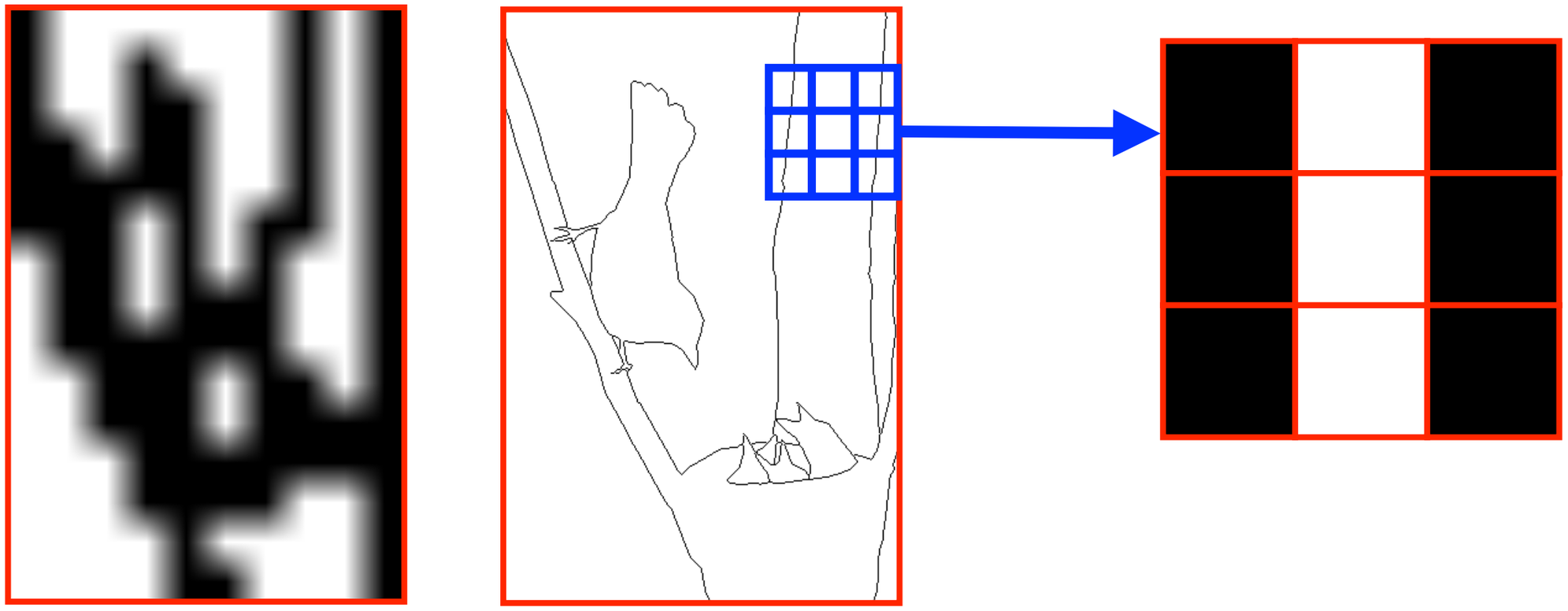


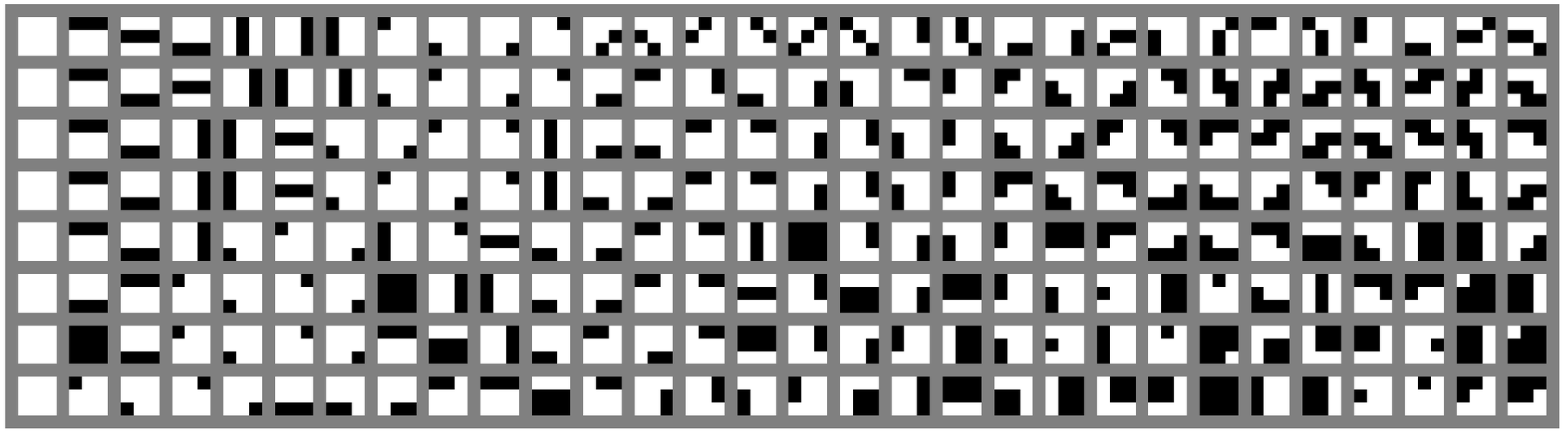
- Coarsenings
  - $X^1 \ldots X^K$
  - X^{i+1} is a function of X^i
- Look at 3x3 blocks at all resolutions

$$E(X) = \sum_i \sum_C V^i(X_C^i)$$

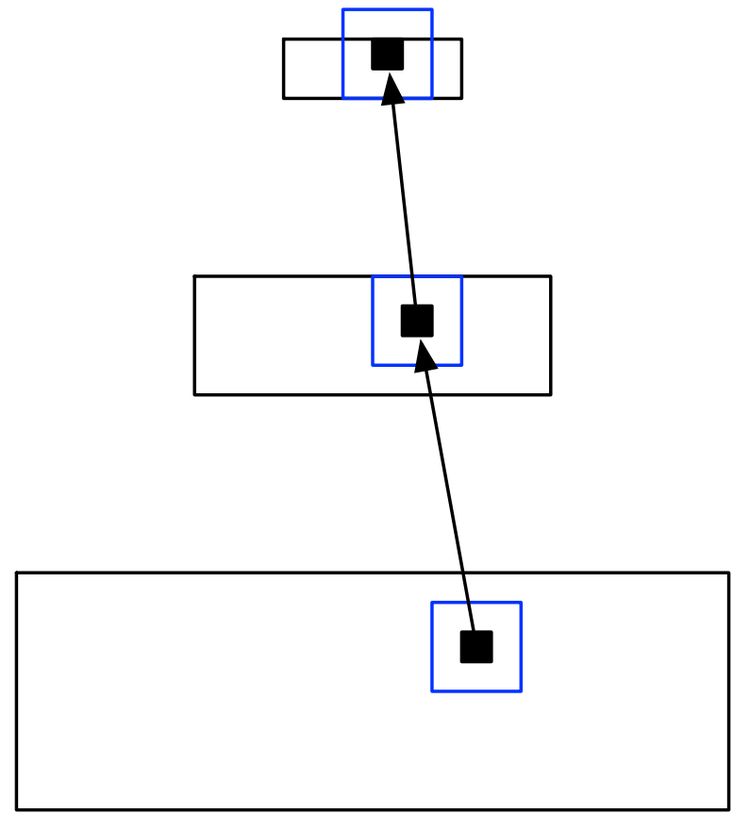
  - $V^i \neq V^j$

Coarse Patterns





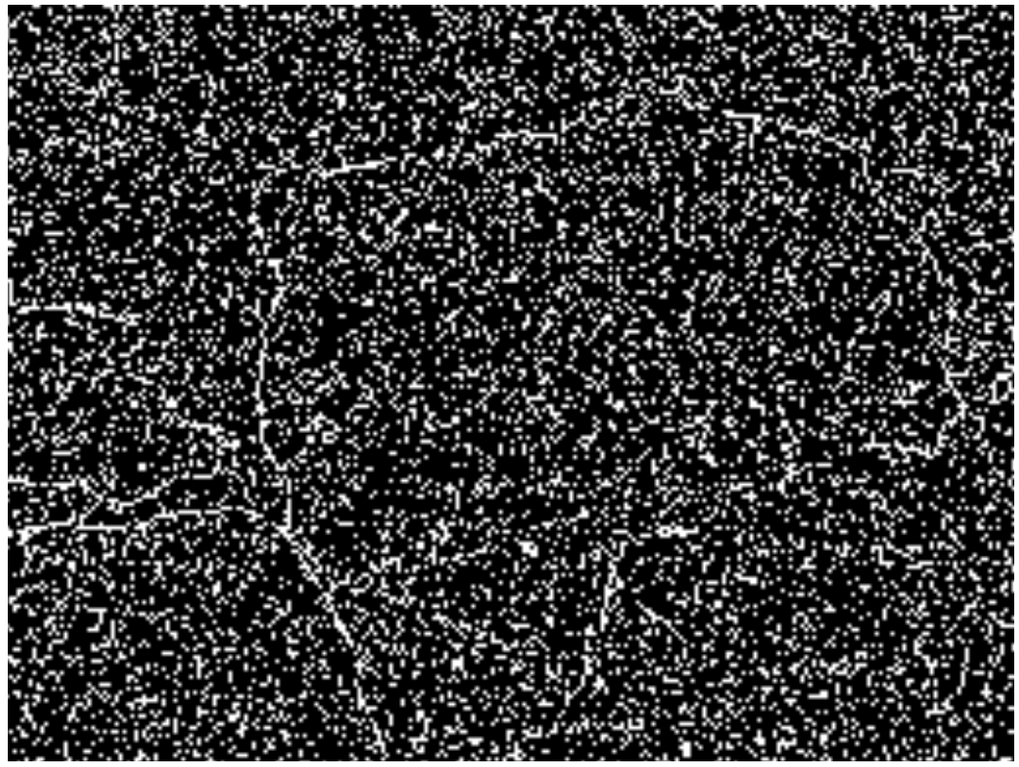
MCMC Inference



- Repeatedly update pixels
- $P(X)$ is not Markov
  - 3x3 block in X^K might depend on whole picture
- Efficient MCMC via multi-scale representation
  - Energy difference is local over X^1 ... X^K

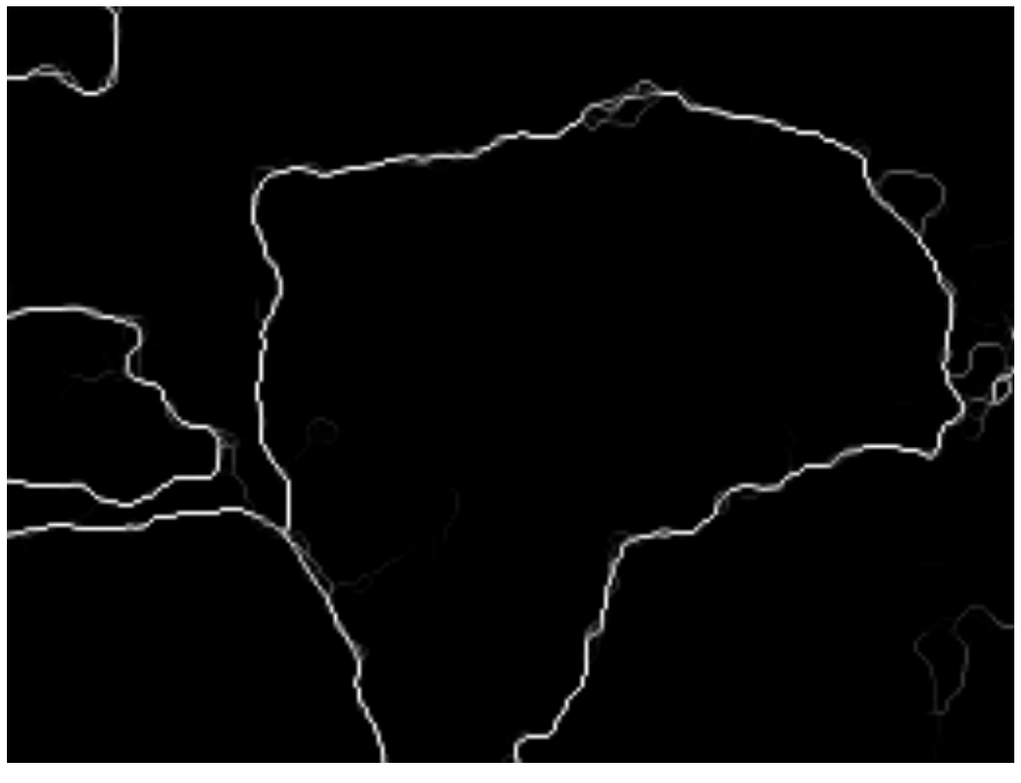
Restoring noisy images

iid noise
20% flipped



Y

$P(X_i=1|Y)$



X

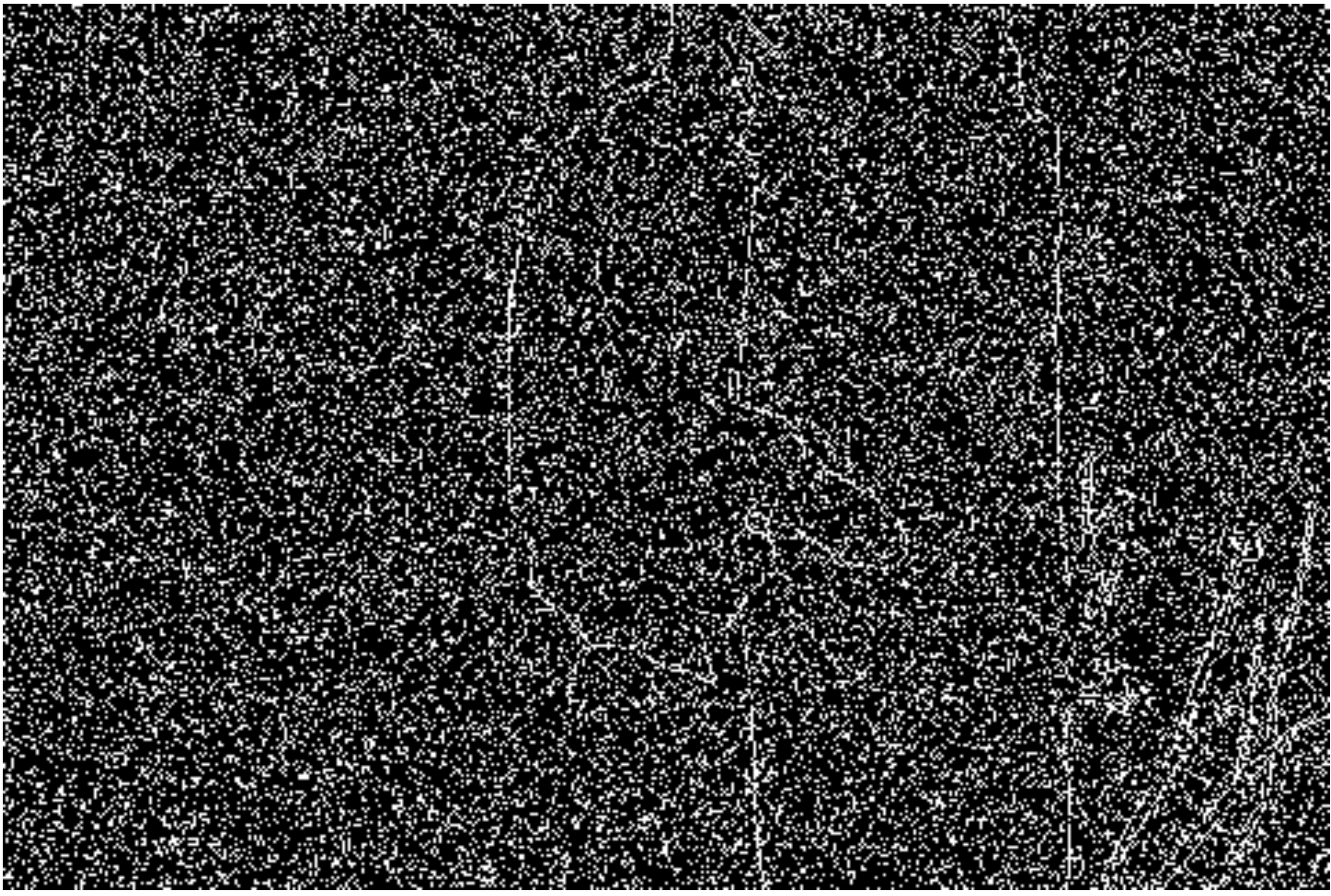
Restoring noisy images

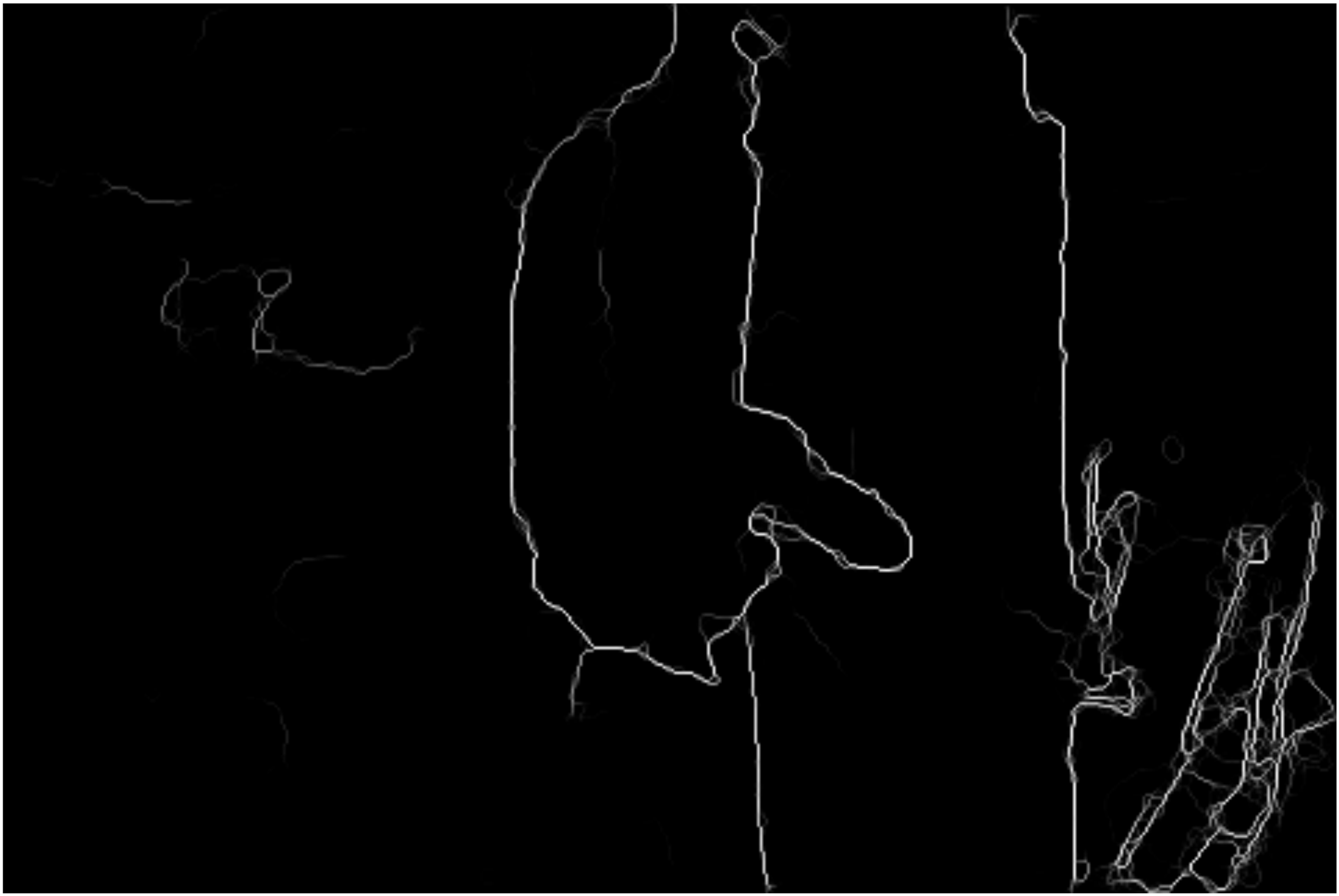
iid noise
20% flipped

Y

$P(X_i=1|Y)$

X





Summary

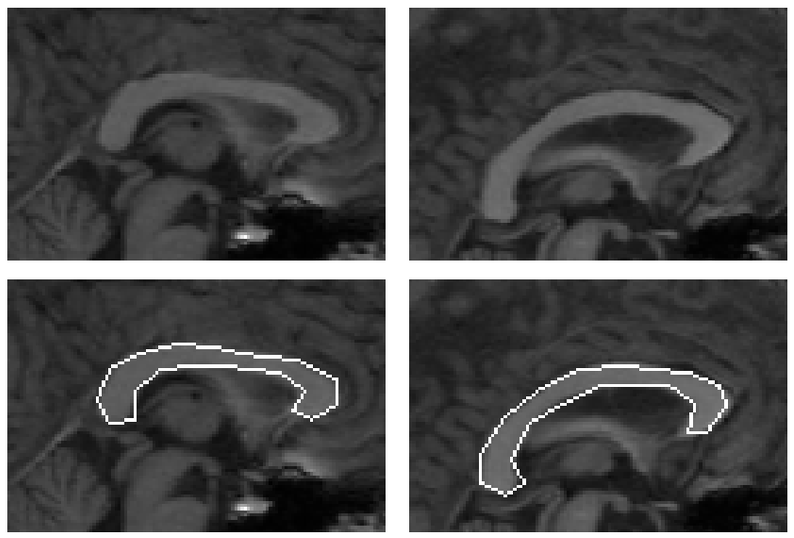
- Graphical models permeate computer vision
  - Image restoration
  - Depth estimation
  - Segmentation / Edge detection
  - Object Recognition

- A lot of work to do in object recognition/detection
  - Better data models
  - Structure variation
- Need better priors for low/mid-level vision

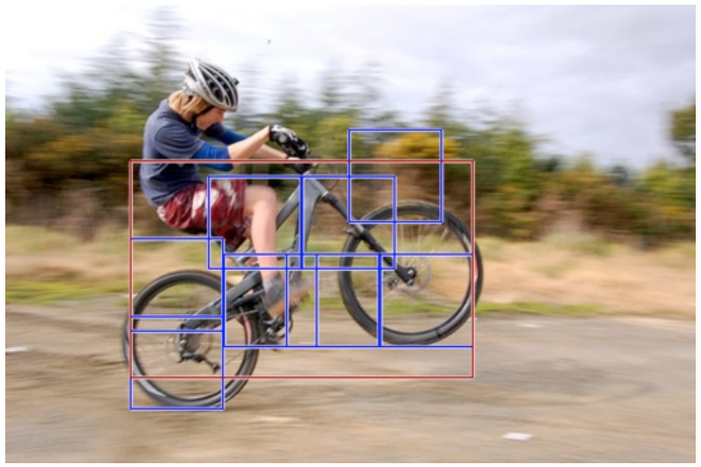
Thank you

Low-level vision ⟷ High-level vision

Restoration



Segmentation



Recognition

Depth estimation