Algorithms for Markov Random Fields in Computer Vision

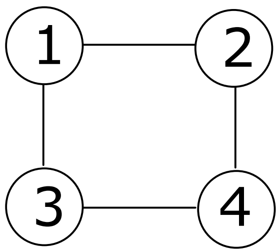
Dan Huttenlocher

November, 2003

(Joint work with Pedro Felzenszwalb)

Random Field

- Broadly applicable stochastic model
  - Collection of n sites $\mathcal{S}$
  - Hidden variable x_i at each site i
  - Label set $\mathcal{L}$
    - Each site takes on label $\ell \in \mathcal{L}$
  - Neighborhood system $\mathcal{N}$
    - $\mathcal{N}_i$ neighbors of site i
    - Explicit dependencies between neighbors



- Graphical model with undirected edges
  - Graph $\mathcal{G}=(\mathcal{S},\mathcal{N})$
  - $\mathcal{N}_i$ set of nodes with edges incident on i

Markov Random Field (MRF)

- Random field with Markov property

$$P(x_i \mid x_{\mathcal{S}\backslash i}) = P(x_i \mid x_{\mathcal{N}i})$$

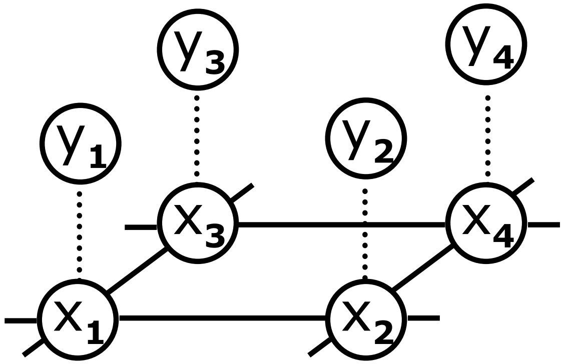
  - Where $\mathcal{S}\backslash i$ denotes set $\mathcal{S}$ excluding element i
- Standard simplification (abuse) of notation
  - Probability of r.v. x_i taking on value v, $P(x_i=v)$ abbreviated as $P(x_i)$
- Conditional probabilities depend only on neighborhood
  - Probability of x_i taking on some value same given all other nodes as given just neighbors

MRF's for Low Level Vision

- Grid graph
  - Sites are pixels; up, down, left, right neighbors
  - Neighborhood enforces spatial coherence
  - Observed value y_i at each site (pixel)
- Applies to many pixel-oriented problems
  - Naturally expressed as posterior probability of labels given observations, P(x|y)
    - Stereopsis, labels are depths (disparities)
    - Optical flow, labels are motion vectors
    - Restoration, labels are intensities (colors)
  - [Geman & Geman, 1984]

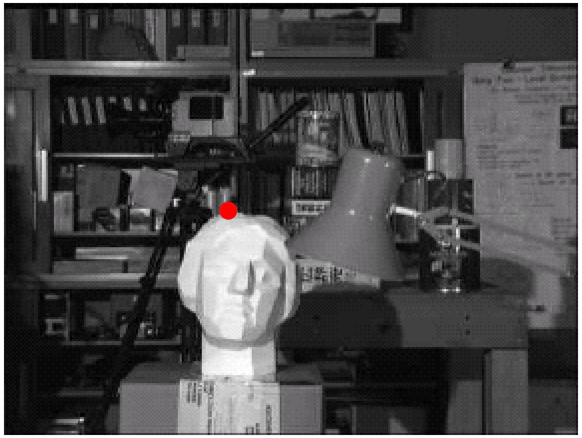
MRF Stereo

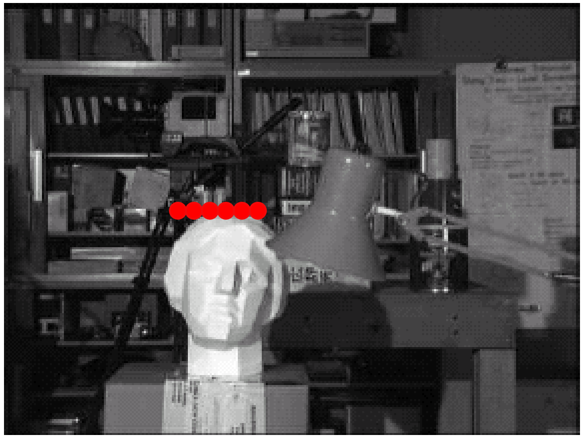
- Given two images, estimate depth at each pixel



Depths

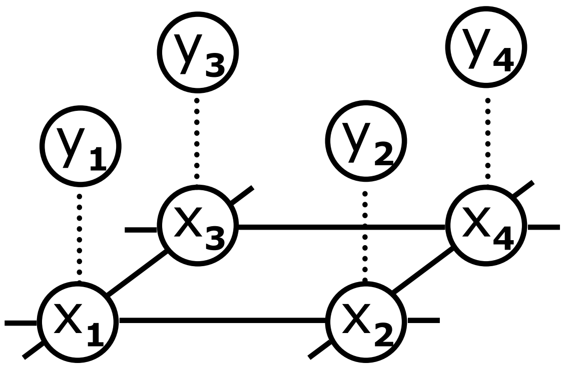
(Tens of labels)





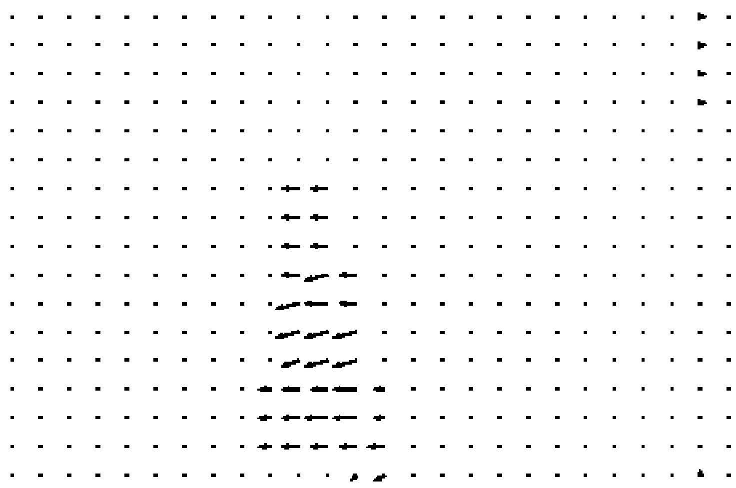
MRF Motion

- Given two images, estimate motion vector at each pixel



First image

Second image

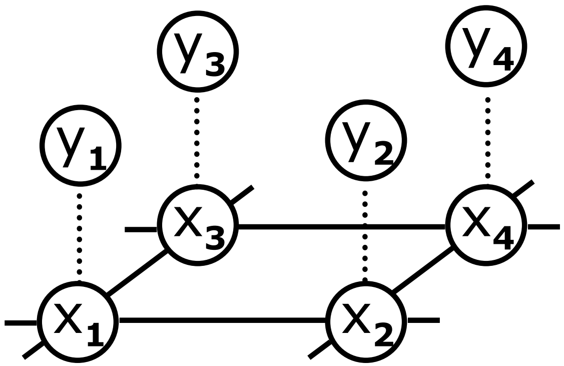


Flow vectors

(Hundreds of labels)

MRF Image Restoration

- Given image corrupted by noise, estimate original image
  - Intensity/color for each pixel

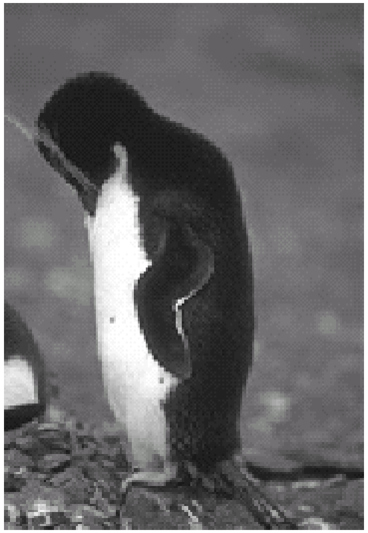


Intensities

(Hundreds of labels)

Corrupted

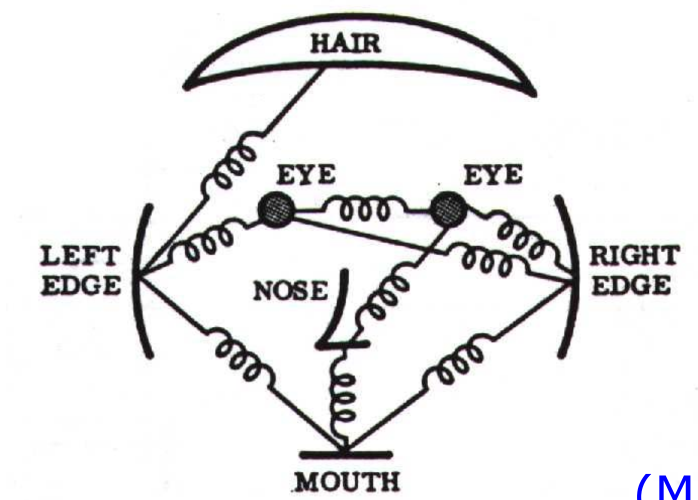
Restoration

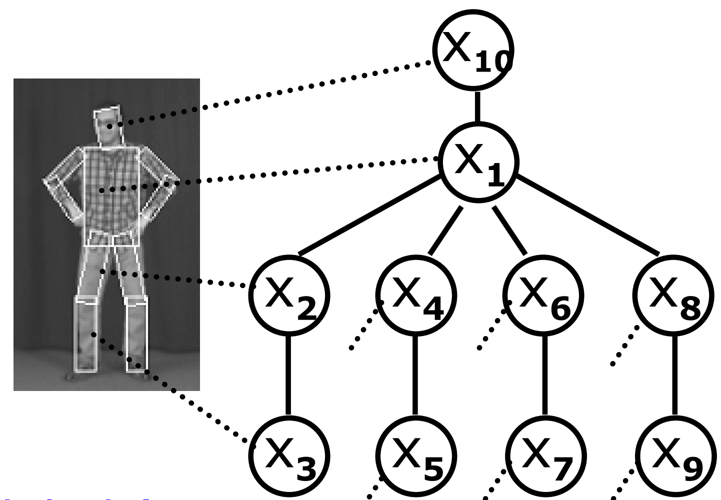


Original

MRF's for High Level Vision

- Given image, estimate location of object
  - Pictorial structure model
    - Parts represented as local image patches
    - Spring-like connections between pairs of parts
  - Non-MRF formulation [Fischler&Elschlager, 1973]





(Millions of labels)

Locations

Using MRF’s

- Given MRF model and observed values, infer most likely values of hidden variables
- Learn MRF parameters from examples
- Note analogous problems for hidden Markov models (HMM’s)
  - Chains are equivalent to HMM’s
  - Generalization to sites and neighborhoods rather than temporal (ordered) dependency
- Both inference (estimation) and learning problems are hard for general MRF’s

MRF Inference (Estimation)

- Find labelings that have high probability given observations (posterior)

  $$P(x|y)=P(x_1,x_2, \ldots, x_n \mid y_1,y_2, \ldots y_n)$$

- Standard Bayesian estimation problem

  $$P(x|y) \propto P(y|x)P(x)$$

  - Likelihood P(y|x) of observations given labels
    - Reasonable to assume independence, factor

      $$P(y|x) = \prod_{\mathbf{i}\in\mathcal{S}} P(y_{\mathbf{i}}|x_{\mathbf{i}})$$

  - Prior P(x) of labelings
    - MRF conditional probability $P(x_i|x_{\mathcal{S}\backslash i}) = P(x_i|x_{\mathcal{N}\mathbf{i}})$ not directly useful for factoring this joint distr.

Factoring the Prior

- MRF equivalent to Gibbs random field (GRF)
  - Hammersley-Clifford theorem (1971)
- In GRF prior is factored over cliques $\mathcal{C}$ of underlying graph $\mathcal{G}=(\mathcal{S},\mathcal{N})$

$$P(x) \propto \exp(-\textstyle\sum_{c\in\mathcal{C}} V_c(x_c))$$

  - Clique potential V_c function of labels for clique
- Cliques=edges for chains, trees, four-connected grids (cliques size 2)

$$P(x) \propto \exp(-\textstyle\sum_{(i,j)\in\mathcal{N}} V_{ij}(x_i,x_j))$$

  - Often also written $P(x) \propto \prod_{(i,j)\in\mathcal{N}} \Psi_{ij}(x_i,x_j)$

Tractable Inference Problem

- Posterior distribution factors

  $$P(x|y) \propto \prod_{i \in \mathcal{S}} P(y_i|x_i) \prod_{(i,j) \in \mathcal{N}} \Psi_{ij}(x_i,x_j)$$

- Maximize posterior
  - MAP estimate, $\text{argmax}_{\mathbf{x}} P(x|y)$
  - Sample high probability values of x
- Common to express as corresponding <u>energy minimization</u> problem
  - Costs (negative log probabilities)

  $$\sum_{i \in \mathcal{S}} D_i(x_i,y_i) + \sum_{(i,j) \in \mathcal{N}} V_{ij}(x_i,x_j)$$

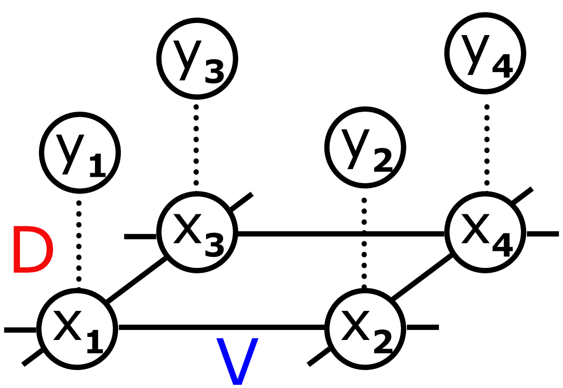
Match to Data — Local Consistency

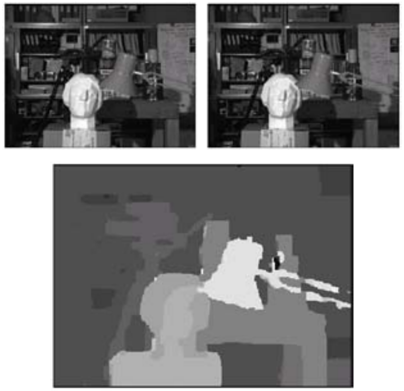
Back to Vision Problems

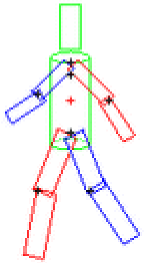
- Intuitive local meanings of energy function

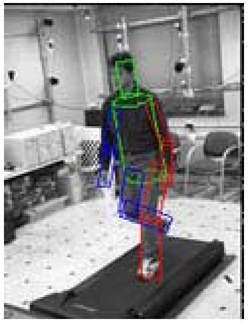
$$\sum_{i \in \mathcal{S}} D_i(x_i, y_i) + \sum_{(i,j) \in \mathcal{N}} V_{ij}(x_i, x_j)$$

  - For both low-level and high-level problems
    - Spatial coherence for stereo, motion, restoration
    - Spring-like connections for multi-part objects
- Global: equivalent to maximizing P(x|y)









Remaining Computational Issues

- Exponential number of labelings

  $O(k^n)$ where $|\mathcal{L}|=k$

- Efficient algorithms if no loops in the graph (i.e., chain or tree)
  - Viterbi algorithm $O(k^2 n)$
  - NP hard in most cases for grid graph
    - E.g., some two-label problems poly-time (min cut)
- For practical purposes a dead end
  - Low level vision: heuristic search methods like annealing slow and unreliable
  - High level vision: quadratic in millions of labels

Recent Algorithmic Advances

- Approximations for grid graph
  - Characterization of local minima
    - Graph cuts [Boykov, Veksler & Zabih, 1999]
    - Loopy belief propagation [Weiss&Freeman, 1999]
  - Best stereo algorithms now almost all use either GC or LBP
- O(nk) algorithm for tree – many labels
  - For pictorial structures where clique potential is a weighted quadratic distance, $s\|x_i-x_j\|^2$
    - Based on generalization of distance transforms [Felzenszwalb&Huttenlocher, 2000]

Still Limited Applicability

- Large label sets often impractical
  - Grid graphs
    - Optical flow (motion estimation)
    - Image restoration
  - Chains (HMM's)
    - Inference on time series data
- Graph cuts and belief prop slow compared to local methods
  - Several minutes for stereo pair compared to second or less for methods not based on MRF's
    - Choice of speed versus accuracy

New Results Address These Limits

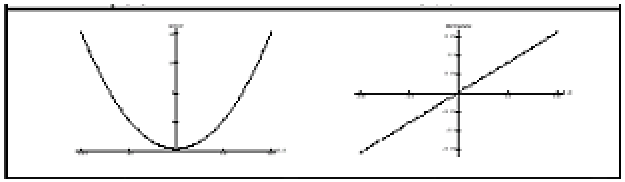
- Running time linear in number of labels for commonly used clique potentials V_{ij}
  - For Viterbi and BP algorithms
  - Efficient computation of min-transform
    - Potentially applicable to other combinatorial optimization problems
- Hierarchical method for LBP on grid graph
  - LBP is an iterative messaging passing method
    - Number of iterations generally proportional to diameter of graph
    - Hierarchy enables constant number of iterations

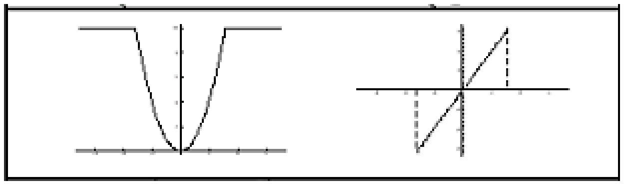
Form of Clique Potentials

- $V_{ij}(x_i,x_j)$ commonly based on measure of difference between labels x_i,x_j
  - Linear: $\sigma|x_i-x_j|$
  - Quadratic: $\sigma(x_i-x_j)^2$
  - Potts: 0 when $x_i=x_j$, τ otherwise
  - Truncated linear: $\min(\tau, \sigma|x_i-x_j|)$
  - Truncated quadratic: $\min(\tau, \sigma(x_i-x_j)^2)$

Spring-like (Linear, Quadratic)

Spatially Coherent (Potts, Truncated linear, Truncated quadratic)





Truncation allows for discontinuities (non-coherence)

Dependence on Number of Labels

- Viterbi and min-sum BP both involve min-transform of some f for each site i

$$h(x_i)=\min_{x_j} (V_{ij}(x_i,x_j) + f(x_j))$$

- Cost of label x_i at node i, $h(x_i)$
  - Depends on cost computed at neighbor j plus discontinuity cost (clique potential)
    - Seek best x_j for each x_i – miminization
- Explicit computation by considering pairs x_i,x_j leads to $O(k^2)$ term in running time

Potts Min-Transform

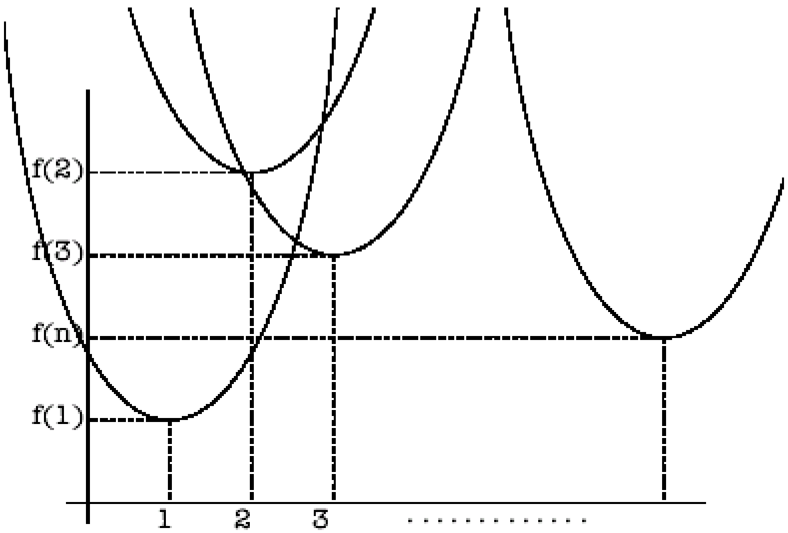
- The min-transform can be computed in O(k) time for the Potts model
  - Penalty of τ when labels disagree 0 when agree
- Straightforward re-arrangement of terms

$$h(x_i)=\min(\min_{x_j} f(x_j)+\tau, f(x_i))$$

  - Because $V_{ij}(x_i,x_j)$ is 0 when $x_i=x_j$, τ otherwise
  - No need to explicitly consider pairs only two cases
    - Same labels, value of h is same as f (penalty 0)
    - Different labels, value of h is best f plus penalty

Quadratic Min-Transform

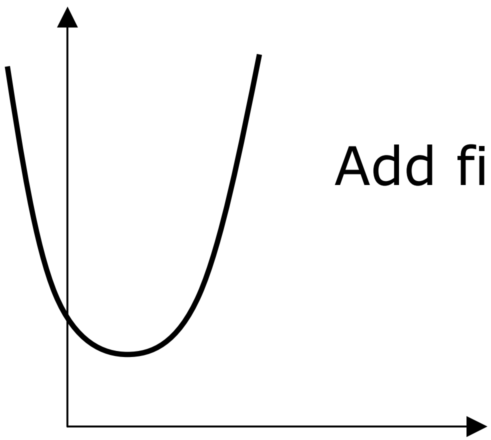
- Compute $h(x_i)=\min_{x_j}(\sigma(x_i-x_j)^2+f(x_j))$
  - Geometric view: in one dimension, lower envelope of arrangement of k quadratics
    - Each rooted at $(x_j, f(x_j))$



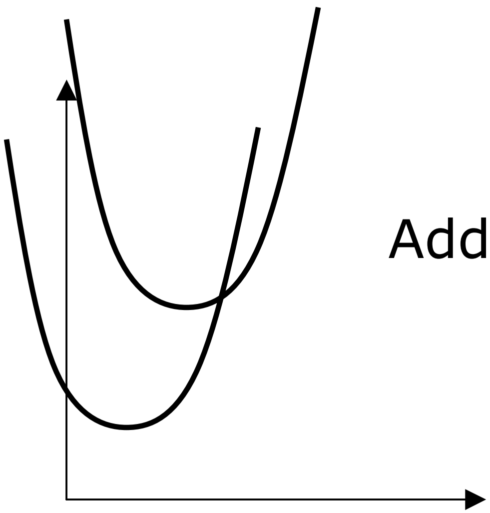
Algorithm for Lower Envelope

- Quadratics ordered $x_1 < x_2 < \ldots < x_k$
- At step j consider adding j-th quadratic to LE of first j-1 quadratics
  - Maintain two ordered lists
    - Quadratics currently visible on LE
    - Intersections currently visible on LE
  - Compute intersection of j-th quadratic and rightmost quadratic visible on LE
    - If right of rightmost visible intersection add quadratic and intersection to lists
    - If not, this quadratic hides at least rightmost quadratic, remove it and try again

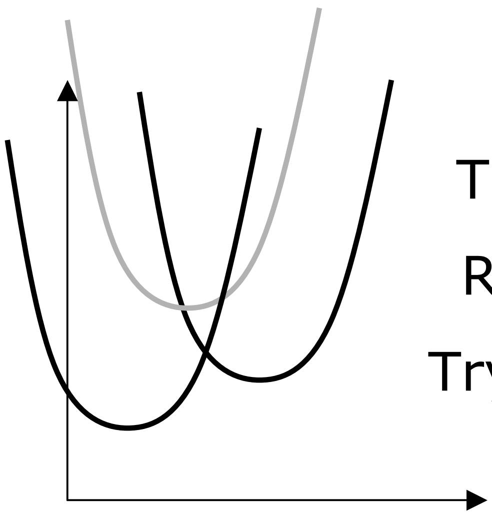
LE Algorithm



Add first



Add second



Try adding third

Remove second

Try again and add

…

Running Time of LE Algorithm

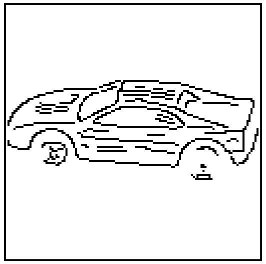
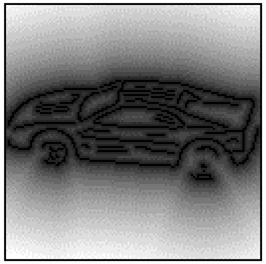
- Considers adding each of k quadratics just once
  - Intersection and comparison constant time
  - Adding to lists constant time
  - Removing from lists constant time
    - But then need to try again
- Simple amortized analysis
  - Total number of removals O(k)
    - Each quadratic once removed never considered for removal again
- Thus overall running time O(k)

Linear Time Min-Transform Method

- Calculating $\min_{x_j}(\sigma(x_i-x_j)^2+f(x_j))$ from LE
  - Fill in vector of values based on visible quadratics and intersections
  - Exact calculation followed by rasterization
- Overall algorithm about 30 lines of c code
  - Very fast in practice
- Generalizes to higher dimensions
  - Consider two dimensions u,v
    - First pass to compute min u^2 (or min v^2) distance
    - Subsequent pass on result of first pass computes min u^2+v^2 distance

Other Applications of Min-Transform

- (Squared) Euclidean distance transform
  - Distance to nearest "on" pixel in binary image
  - Previous algorithms complex because think of operating on point sets rather than functions

- Combinatorial optimization problems
  - Minimizations involving sum of cost and distance

$$\min_{\mathbf{y}}(\sigma \| x-y \| + f(y))$$

Min-Transform for Viterbi

- For chain $x=(x_1, \ldots, x_n)$ the Viterbi algorithm computes

$$\min_{\mathbf{x}} \textstyle\sum_{\mathbf{i}} D(x_{\mathbf{i}},y_{\mathbf{i}}) + V(x_{\mathbf{i}},x_{\mathbf{i-1}})$$

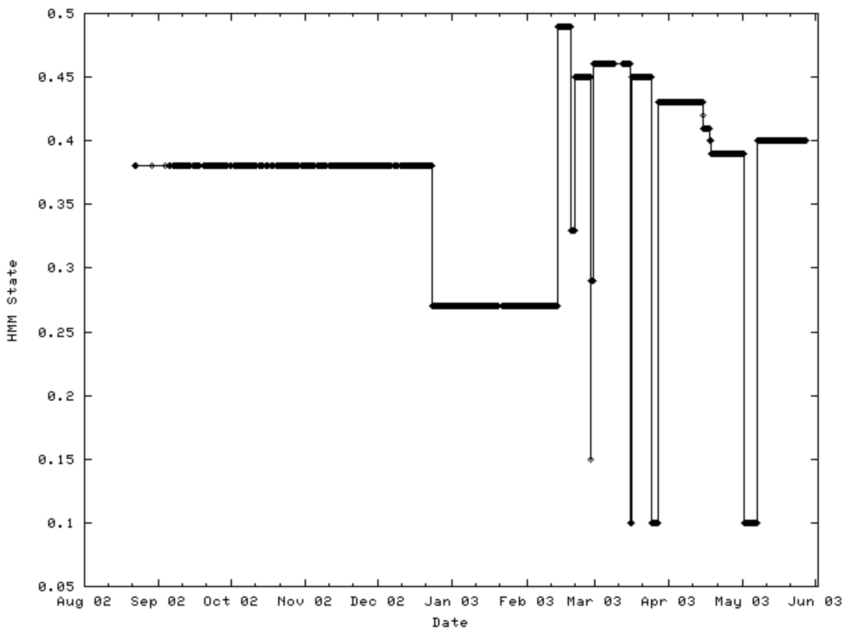
  - Using recurrence

$$s_{\mathbf{i}}(x_{\mathbf{i}}) = D(x_{\mathbf{i}},y_{\mathbf{i}}) + \min_{\mathbf{x_{i-1}}} (s_{\mathbf{i-1}}(x_{\mathbf{i-1}}) + V(x_{\mathbf{i}},x_{\mathbf{i-1}}))$$

- Use min-transform algorithm to compute second term of recurrence in O(k) time
  - For quadratic, Potts, truncated quadratic
  - Simpler method for linear, truncated linear
- O(nk) overall, n steps in recurrence

Coin Tossing Example

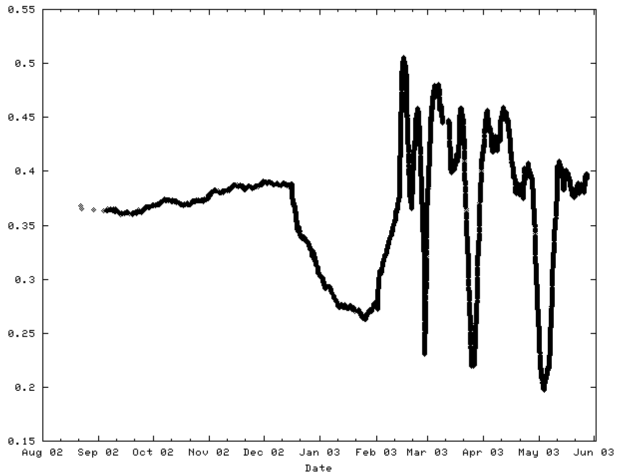
- Estimate bias of “changing coin” from sequence of observed {H,T} values
  - Labels correspond to possible bias values, e.g., .100, …, .900
  - Data costs
    $-\log P(H|x_i)$
    $-\log P(T|x_i)$
  - Clique potential truncated quadratic

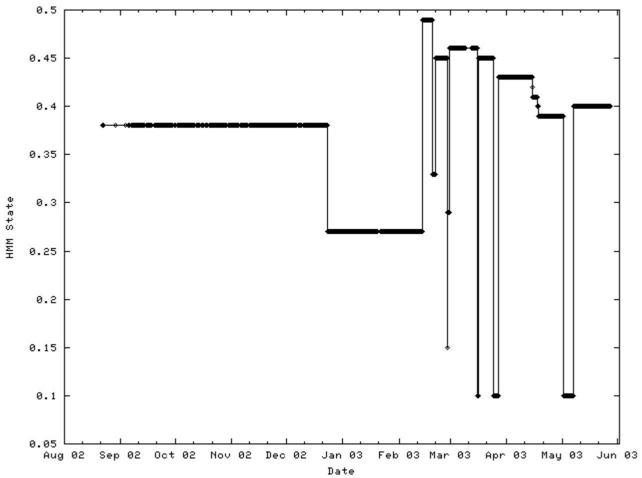


  - [Felzenszwalb, Huttenlocher & Kleinberg, 2003]

Power of Stochastic Model

- Infer instantaneous (discretized) probability from observed H,T sequence
- Detect changes in hidden value
- Contrast with linear approach such as weighted windowed average





Loopy Belief Propagation

- Iterative message passing from each site to neighbors
  - Several variants, consider min-sum which matches our energy minimization formulation
  - Message $m_{i,j,t}$ sent from site i to j at time t

$$m_{i,j,t}(x_j) = \min_{x_i} [V(x_i,x_j)+D(x_i,y_i) + \textstyle\sum_{k\in\mathcal{N}_i\setminus j} m_{k,i,t-1}(x_i)]$$

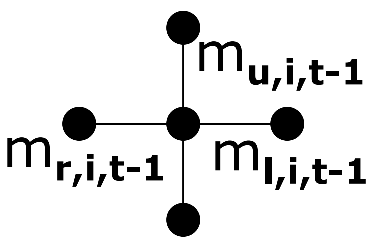
    - Based on neighbors of i other than j, at step t-1
  - After T iterations each node computes label minimizing (maximizing "belief")

$$b_i(x_i)=D(x_i,y_i)+\textstyle\sum_{k\in\mathcal{N}_i} m_{k,i,T}(x_i)$$

Schematic of LBP on Grid

- Each node computes four messages
  - Think of neighbors as up, down, left, right
- Example, message to send down from i

$$m_{i,d,t}(x_d) = \min_{x_i} [V(x_i,x_d)+D(x_i,y_i)+m_{r,i,t-1}(x_i) + m_{l,i,t-1}(x_i)+ m_{u,i,t-1}(x_i)]$$



Compute $m_{i,d,t}$

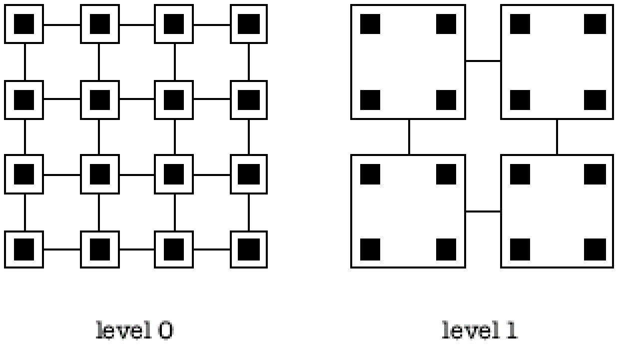
- Min-transform so $O(k)$ not $O(k^2)$

About LBP

- For grids works well in practice
  - Convergence properties not totally understood
- Number of iterations T proportional to diameter of grid
  - For most vision problems need to propagate information from distant parts of grid
- An improvement to LBP on grid
  - Initialize messages to values that reflect propagation from distant sites
    - Use a multi-scale method to do so
  - Only constant number of iterations required

Multi-Scale LBP on Grid

- Node corresponds to block of pixels that are all assigned a single label
  - $2^\ell x 2^\ell$ block at level ℓ of hierarchy
- Short paths in coarse level graphs
- Final messages at level ℓ initialize level ℓ-1
- Other multi-scale BP methods change problem definition
  - Use hierarchical graph

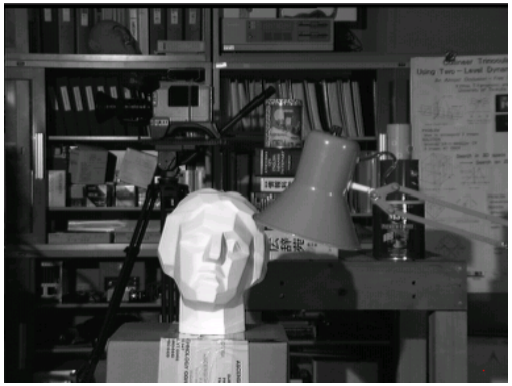


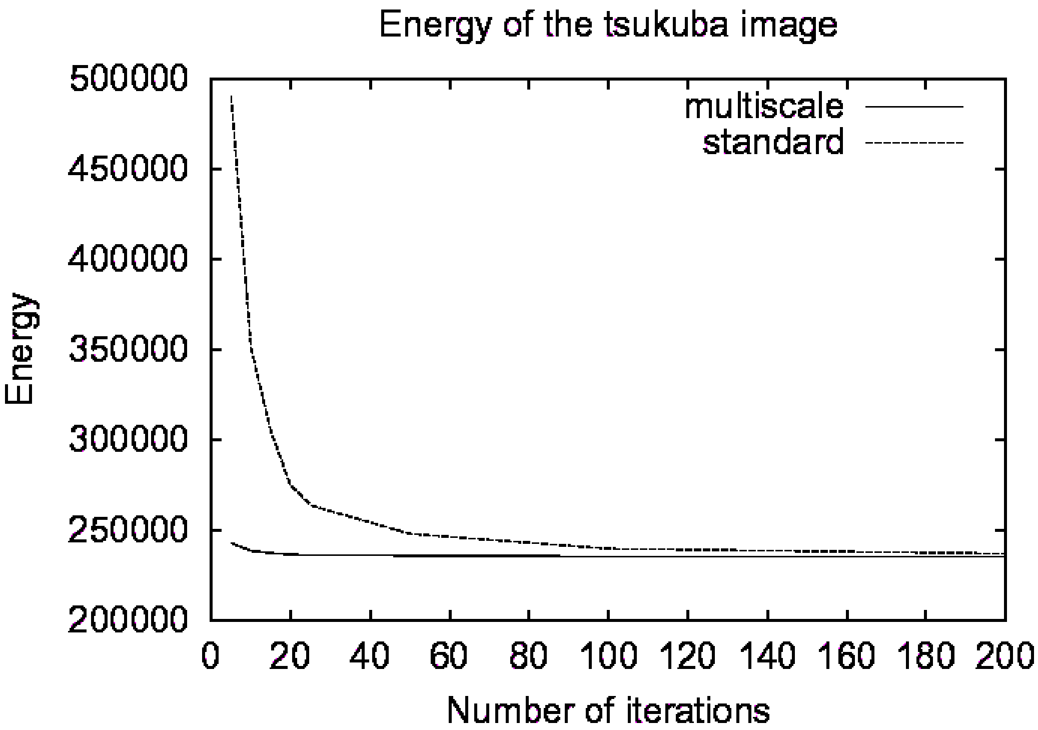
Multi-Scale Method

- Clique potential V, same at all levels
  - Based on pair of labels for two nodes
    - Each node assigned one label (for all pixels)
- Data cost D, sum of data costs for pixels
  - Corresponds to likelihood of observed data given single label for all pixels
  - Differs from other multi-scale methods
    - Not lower resolution image
      - E.g., Gaussian pyramid, smooth and sub-sample
    - Only lower resolution estimation problem
      - To speed message propagation

Hierarchical Method Converges Fast

- Example for stereo matching
  - Truncated linear clique potential





Fast MRF Methods

- Makes MRF’s practical for many problems
  - Vision, comparable to speed of local methods
    - Stereo matching, 1 sec per pair
    - Visual motion estimation, 4 secs per pair
    - Image restoration, 4 secs per image
    - Human body pose recovery, 30 secs per image

      (640x480 images, 2 gHz Pentium 4)
  - Time series, large label sets (state spaces)
- Compared with previous methods
  - GC and standard LBP take minutes for stereo, other vision problems not feasible
  - HMM’s only feasible for small state spaces

Still Plenty To Do

- Better understanding of why LBP and hierarchical method work well on grids
  - "Large moves" – many labels set together
  - Characterization of local minima found
- Related techniques for sum-product BP algorithm
  - Important for problems such as motion where sub-pixel interpolation desirable
- Problems where parameters of MRF not known (or learned) a priori
  - E.g., "multi modal" imagery