

Hidden Markov Trees for Statistical Signal/Image Processing




Xiaoning Qian

ECEN689–613 Probability Models

Texas A&M University

Part I

Papers

-  M. S. Crouse, R. D. Nowak, R. G. Baraniuk, *Wavelet-Based Statistical Signal Processing Using Hidden Markov Models*, TSP, **46**(4), 1998.
-  H. Choi, R. G. Baraniuk, *Multiscale Image Segmentation Using Wavelet-Domain Hidden Markov Models*, TIP, **10**(9), 2001.
-  J. Romberg, M. Wakin, H. Choi, R. G. Baraniuk, *A Geometric Hidden Markov Tree Wavelet Model*, dsp.rice.edu, 2003.

Part II

Wavelet Transform

What is a wavelet?

- [Wikipedia](#): A wavelet series representation of a square-integrable function is with respect to either a **complete, orthonormal** set of basis functions, or an **overcomplete set of Frame** of a vector space (also known as a Riesz basis), for the Hilbert space of square integrable functions.

What is a wavelet?

- The main idea of wavelets is from the idea of function representations. Wavelets are closely related to **multiscale/multiresolution analysis**:
 - Decompose functions into different scales/frequencies and study each component with a resolution that matches its scale.
- Wavelets are a class of a functions used to localize a given function in both space and scaling/frequency.
- For more information:
<http://www.amara.com/current/wavelet.html>

An example – Haar basis

Example

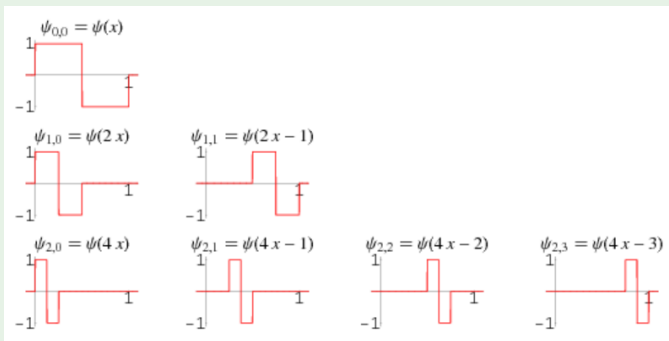
- Haar wavelet: the wavelet function (mother wavelet) $\psi(t)$; scaling function (father wavelet) $\phi(t)$:

$$\psi(t) = \begin{cases} 1 & 0 \leq t < 1/2 \\ -1 & 1/2 \leq t < 1 \\ 0 & \text{otherwise} \end{cases}; \quad \phi(t) = \begin{cases} 1 & 0 \leq t < 1 \\ 0 & \text{otherwise} \end{cases}.$$

- “Daughter” wavelets: $\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}} \psi(\frac{t-b}{a})$, a –scale; b –shift;
 $\psi_{J,K}(t) = \psi(2^J t - K)$.
- Multi-dimensional wavelet – tensor product of 1-dimensional wavelet

An example – Haar basis

Example



Why wavelet?

- Wavelets are localized in both space and frequency whereas the standard Fourier transform is only localized in frequency.
- Multiscale analysis
- Less computationally complex
- ...

Wavelet transform

- Continuous wavelet transform (CWT):

$$z(t) = \int_{\mathbf{R}} W^{\psi}\{z\}(a, b) \psi_{a,b}(t) db$$

$$W^{\psi}\{z\}(a, b) = \int_{\mathbf{R}} z(t) \psi_{a,b}^*(t) dt$$

$$\int_{\mathbf{R}} \psi_{a,b}(t) \psi_{c,d}^*(t) dt = \delta_{ac}(t) \delta_{bd}(t)$$

Wavelet transform

- Discrete wavelet transform (DWT):

$$z(t) = \sum_K u_K \phi_{J_0, K}(t) + \sum_{J=-\infty}^{J_0} \sum_K w_{J, K} \psi_{J, K}(t)$$

$$w_{J, K} = \int z(t) \psi_{J, K}^*(t) dt$$

$$\int \psi_{J', K'}(t) \psi_{J, K}^*(t) dt = \delta_{JJ'}(t) \delta_{KK'}(t)$$

Properties for wavelet transform

- Locality: Each wavelet is localized simultaneously in space and frequency.
- Multiresolution: Wavelets are compressed and dilated to analyze at a nested set of scales.
- Compression: The wavelet transforms of real-world signals tend to be sparse.

“Secondary” properties may be useful.

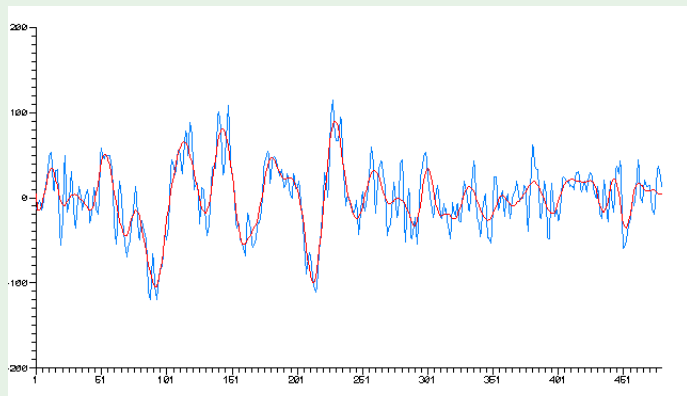
- Clustering: If a particular wavelet coefficient is large/small, adjacent coefficients are very likely to also be large/small.
- Persistence: Large/small values of wavelet coefficients tend to propagate across scales

Part III

Signal processing problems with wavelet applications

Denoising or signal detection

Example



Denoising or signal detection

- Note: The signal model is in the wavelet domain.
- Signal model:

$$w_i^k = y_i^k + n_i^k,$$

where w_i^k is the i th wavelet coefficient by transforming the k th sample. And the task of denoising or detection is to estimate y_i^k .

- Traditional assumption is that they follow independent Gaussian distribution. n_i is the white noise, adaptive thresholding is enough for denoising based on the “compression” property.

Image segmentation

Example

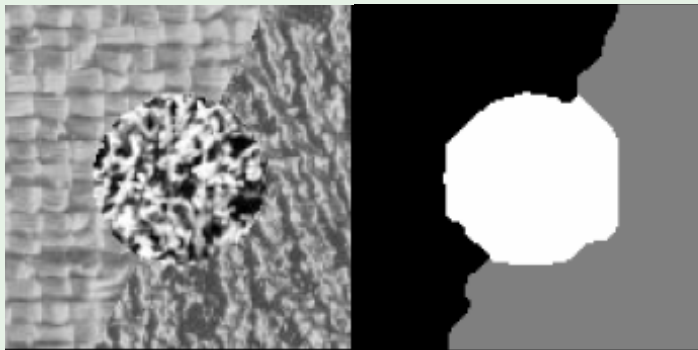


Image segmentation

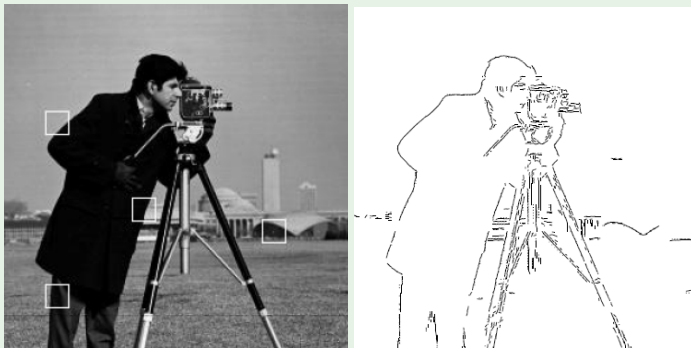
- Modeling the statistical dependency in images.
- Image model: $f(x_r|c_i)$, where c_i are the labels for different objects in an image, x_r are image regions with the same label; $c = \{c_i, \forall i\}$ can be considered as a random field while x_r is the observation.
- The model for c can be considered as prior knowledge.
- Maximum likelihood segmentation: $\max_c \prod_r f(x_r|c)$
- Maximum A Posteriori segmentation: $\max_c \prod_r f(x_r|c)f(c)$
- Note: The model can be either in the image domain or the wavelet domain;

Multiscale image segmentation

- Multiscale image segmentation: **window size**
- Note: the model in multiscale segmentation is again in the wavelet domain now; **the label random field is in the quadtree structure.**
- Different statistical properties for wavelet coefficients correspond to different image regions.
- Singularity structures (edges) have large wavelet coefficients (useful for heterogeneous regions).

Multiscale image segmentation

Example



Basic assumptions in these applications

- **Independent Gaussian** for wavelet coefficients
- Better assumptions?
- Secondary properties?

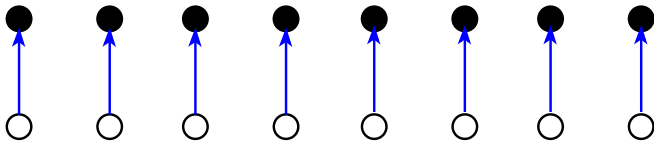
Part IV

Hidden Markov Trees

- General settings: c – random field (latent/hidden variables); x – observations
- Independent c : $f(c_i)$ and $f(x|c) = \prod f(x_i|c_i)$
- Markov random field (hidden Markov model):
 $f(c_i|x, c) = f(c_i|N_i)$ and $f(x|c) = \prod f(x_i|c_i)$
- Conditional random field: $f(c_i|x, c) = f(c_i|x, N_i)$

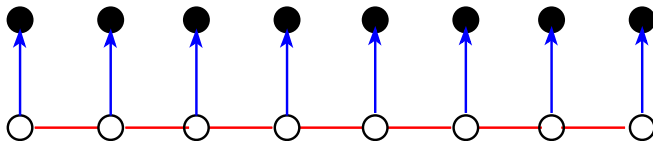
Independent c

- Simplest assumption: c 's are all independent: $f(c_i)$ and $f(x|c) = \prod f(x_i|c_i)$
- Classification algorithms



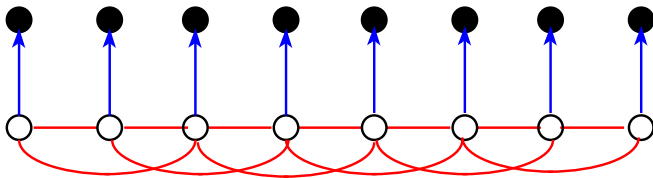
- c follows a Markov chain structure:

$$f(c_i|x, c) = f(c_i|c_{i-1}, c_{i+1})$$
- EM algorithms



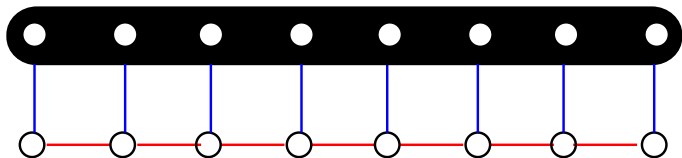
More general hidden Markov model

- c has a complex neighbor structure:
$$f(c_i | x, c) = f(c_i | c_{i-2}, c_{i-1}, c_{i+1}, c_{i+2})$$
- EM algorithms



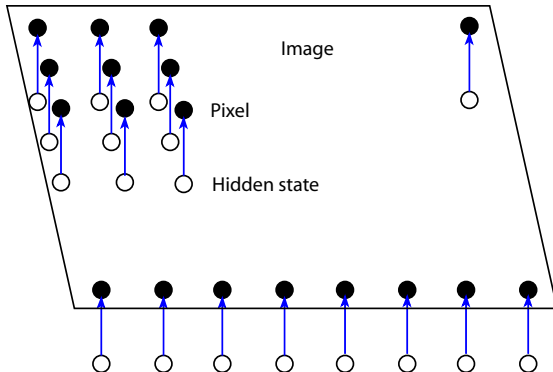
Conditional random field

- c has a Markov structure globally conditioned on x :
$$f(c_i|x, c) = f(c_i|x, N_i)$$
- We usually assume that the probability (or transition function and state function) has some special form.
- Belief propagation algorithms



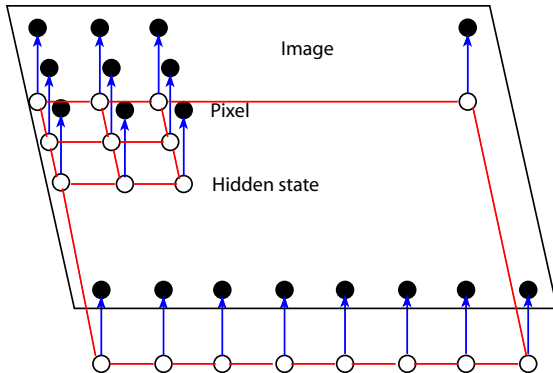
Graphical model in the image domain

- Independent model for homogeneous image regions
- Simple classifiers for pixel intensities



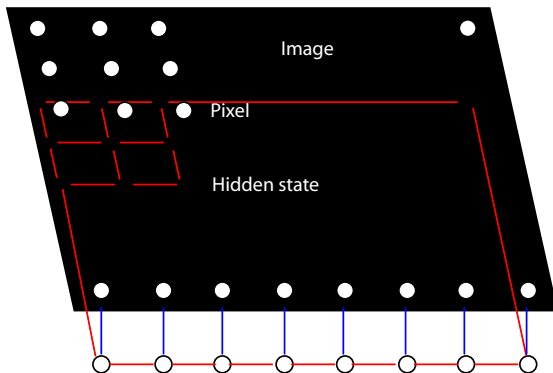
Graphical model in the image domain

- Markov random field for noisy images or texture images
- Adding prior on hidden states for “neighbors”



Graphical model in the image domain

- Conditional random field for more complicated appearance



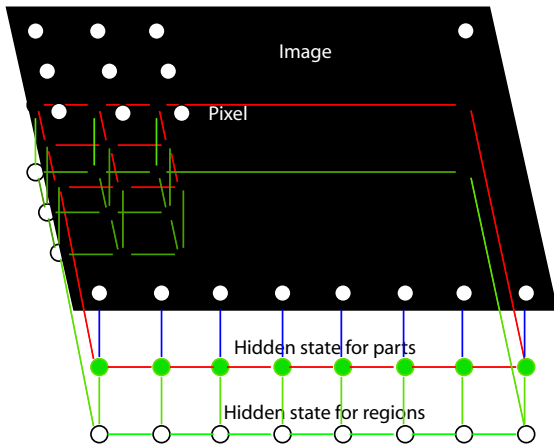
Graphical model in the image domain

- Hidden random field for image regions with different parts



Graphical model in the image domain

- Hidden random field for image regions with different parts



What is an appropriate model?

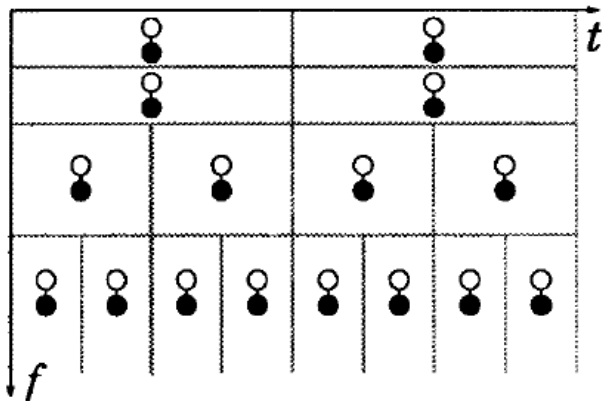
- Tradeoff between accuracy and complexity.
- Small sample size
- Overfitting ...

Hidden Markov trees for wavelet coefficients

- Residual dependency structure (nested structure) – “secondary” properties;
- A model that reflects these properties would be appropriate, flexible but not too complicated;
- Nested multiscale graph (tree to be specific) model:

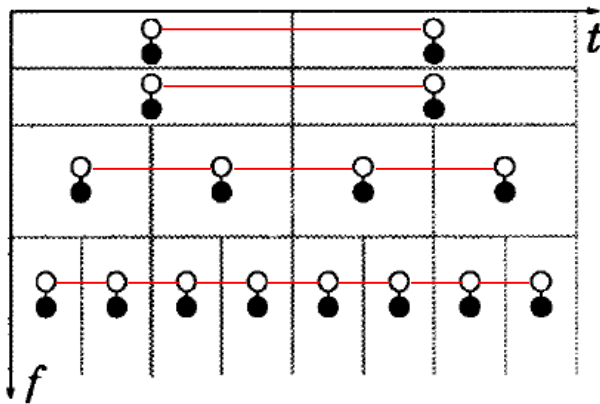
Independent mixture model for wavelet coefficients

- Mixture model provides appropriate approximation for non-Gaussian real-world signals.



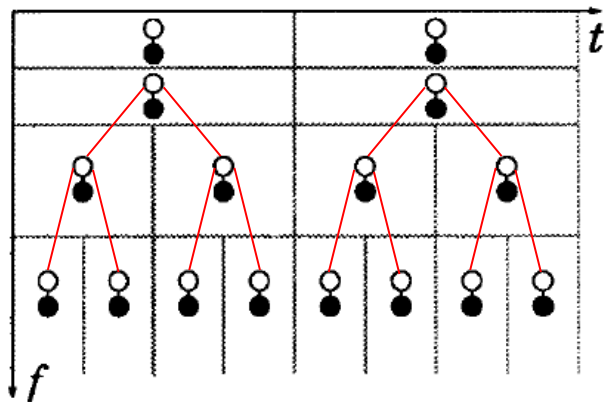
Hidden Markov chain model for wavelet coefficients

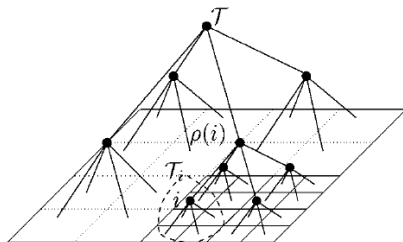
- Hidden Markov chain at the same scale:



Hidden Markov tree for wavelet coefficients

- Dependence across scales according to the “secondary properties” of wavelet coefficients:





Probabilities in hidden Markov trees

- For a single wavelet coefficient, as the real world signal is always non-Gaussian, we model it with a mixture model:

$$f(w) = \sum_m f(w|c = m)f(c = m)$$

- Independent mixture model; Hidden Markov chain model; **Hidden Markov tree** model
 - for the tree root c_0 : $f(c_0)$;
 - for the tree nodes other than root – transition probability: $f(c_i|c_{\rho(i)})$, where $\rho(i)$ denotes the parent node of i .

Parameters in HMT

- $f(c_0)$, $f(c_i|c_{\rho(i)})$;
- mixture means and variance: $\mu_{i,m}$, $\sigma_{i,m}^2$
- Notice the conditional independence properties of the model.

Problems of HMT

- As all of the graphical models, we need to solve
 - Training the model;
 - Computing the likelihood with the given observations;
 - Estimating the latent/hidden states.

Expectation–Maximization algorithm

- General settings for estimation: $\max_{\theta} f(x|\theta)$ (ML) or $\max_{\theta} f(\theta|x) \Leftrightarrow \max_{\theta} f(x|\theta)f(\theta)$ (MAP).
- EM algorithm provides a greedy and iterative way to solve the general estimation problem based on the hidden/latent variables c .
- $\log f(x|\theta) = \log f(x, c|\theta) - \log f(c|x, \theta)$. Since this is the iterative algorithm, we take the expectation with respect to c with the estimated parameters θ^{k-1} :

$$\begin{aligned} \int \log f(x|\theta) f(c|x, \theta^{k-1}) dc &= \int \log f(x, c|\theta) f(c|x, \theta^{k-1}) dc \\ &\quad - \int \log f(c|x, \theta) f(c|x, \theta^{k-1}) dc \end{aligned}$$

Expectation–Maximization algorithm

- Jensen's inequality:

$$\int \log f(c|x, \theta) f(c|x, \theta^{k-1}) dc \leq \int \log f(c|x, \theta^{k-1}) f(c|x, \theta^{k-1}) dc$$

- To guarantee the increase of likelihood $\log f(x|\theta)$, we only need to solve:

$$\theta^k = \arg \max_{\theta} \int \log f(x, c|\theta) f(c|x, \theta^{k-1}) dc$$

- Hence, **E-step** is for computing $f(c|x, \theta^{k-1})$; **M-step** is to solve the above optimization problem.

Training hidden Markov trees with EM

- In HMT, $\theta = \{f(c_0), f(c_i|c_{\rho(i)}), \mu_{i,m}, \sigma_{i,m}^2\}$, where i denotes each wavelet coefficient; m denotes each component in the mixture.
- Update the similar equation:

$$\theta^k = \arg \max_{\theta} \int \log f(w, c|\theta) f(c|w, \theta^{k-1}) dc$$

- We need several tricks to complete the EM algorithm here since we do not have an easy form for $f(w, c|\theta)$.

Training hidden Markov trees with EM

- The main task to estimate the marginal state distribution $f(c_i = m|w, \theta)$ and the parent-child joint distribution $f(c_i = m, c_{\rho(i)} = n|w, \theta)$.
- Based on the conditional independence we have for HMT, we can write: $f(c_i = m, w|\theta) = f(w_{T_i}|w_{\hat{T}_i}, c_i = m, \theta)f(c_i = m, w_{\hat{T}_i}|\theta) = f(w_{T_i}|c_i = m, \theta)f(c_i = m, w_{\hat{T}_i}|\theta) = \beta_i(m)\alpha_i(m)$; and similarly,
 $f(c_i = m, c_{\rho(i)} = n, w|\theta) = \beta_i(m)f(c_i|c_{\rho(i)})\alpha_{\rho(i)}(n)\beta_{\rho(i)\setminus i}(n)$.
- While $f(w|\theta) = \sum_m f(c_i = m, w|\theta) = \sum_m \alpha_i(m)\beta_i(m)$, we have these distributions expressed in terms of α, β .

Training hidden Markov trees with EM

- For the computation, we need to follow the downward algorithm from coarse to fine levels to estimate α 's and upward algorithm from fine to coarse levels to estimate β 's as described in the paper.
- M-step is simply the conditional means due to Gaussian assumption.
- Note the tricks to handle with K trees and tying.

Coming back to the denoising problem ...

- With the EM trained parameters, including $f(c_i = m|\mathbf{w}, \theta)$, σ'_{c_i} 's, and σ_n 's, the estimation for the signal is simple as solving the conditional mean estimates:

$$\mathbf{E}(y_i|\mathbf{w}, \theta) = \sum_m f(c_i = m|\mathbf{w}, \theta) \frac{\sigma_{i,m}^2}{\sigma_{i,m}^2 + \sigma_n^2} w_i$$

Image segmentation

- 2D hidden Markov trees
- Similar setting as in 1D signal model
- Difference:
 - Subband independence:

$$f(w|\Theta) = f(w_{LH}|\Theta_{LH})f(w_{HL}|\Theta_{HL})f(w_{HH}|\Theta_{HH}) \text{ (scaling);}$$
 - Leads to different expansion of α 's and β 's;
 - Context-based interscale fusion: prior $f(c)$: context vector
 - Different EM

Image segmentation

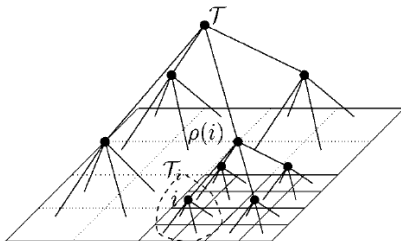
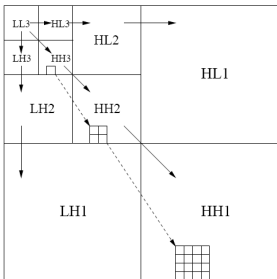
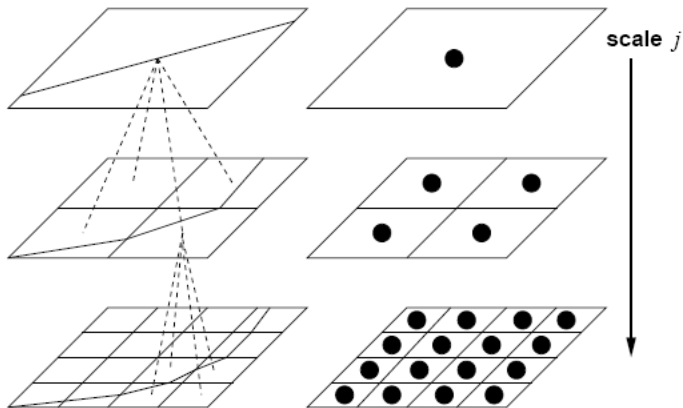


Image segmentation



Image segmentation



Extended hidden Markov trees

- Geometric hidden Markov trees:
 - Modeling contours explicitly;
 - Hidden state space: $c_i = \{d_m, \theta_m\}$
 - New conditional distribution of wavelet coefficients:
 $f(w_i|c_i) \propto \exp(-\text{dist}(w_i, e_m)^2 / (2\sigma_g^2))$, where e_m is the response for edges with fixed distance d_m and angle θ_m (filter banks)
 - New transition probability: $f(n|m) \propto \exp(-HD(l_m, l_n))$, where $HD(l_m, l_n)$ is the Hausdorff distance between lines determined by distance and angle restricted to a square in the plane.

