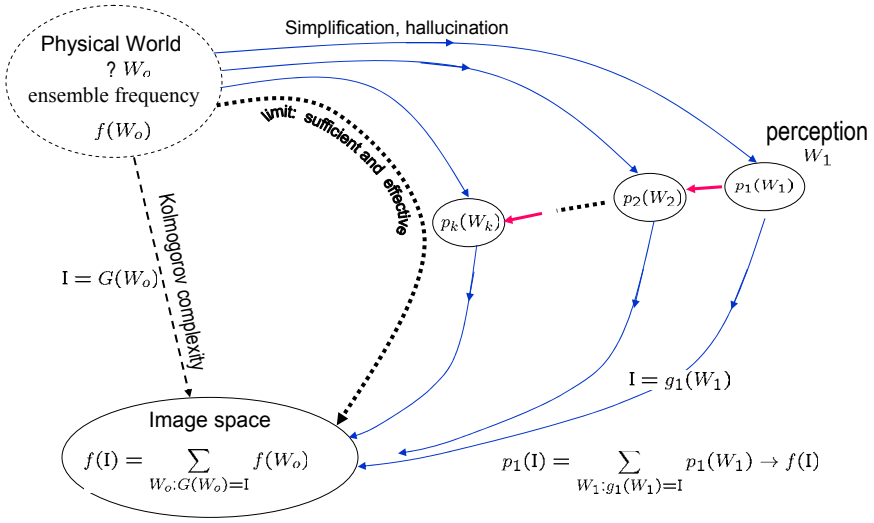
Seeing as Statistical Inference

Song-Chun Zhu

Center for Image and Vision Science
University of California, Los Angeles

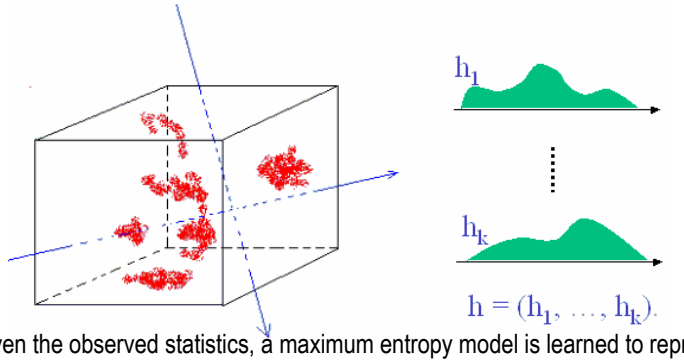
An overview of joint work with many students and colleagues

Pursuit of image models



Family I: Descriptive Modeling

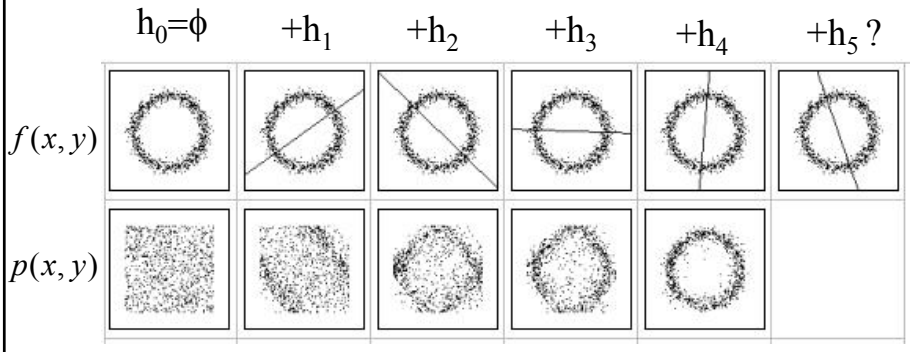
The extracted statistics are marginal distributions (histograms).



1. Given the observed statistics, a maximum entropy model is learned to reproduce the observed statistics.
2. The most informative statistics (features) is selected by minimizing the entropy of the max. ent. model

Leads to Markov random Field and Gibbs models

Toy Example: Estimating 1D manifold embedded in 2D



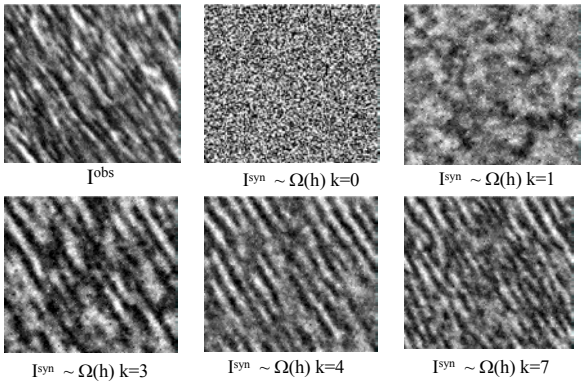
By Ce Liu, 2001.

Augmenting the model by minimax entropy

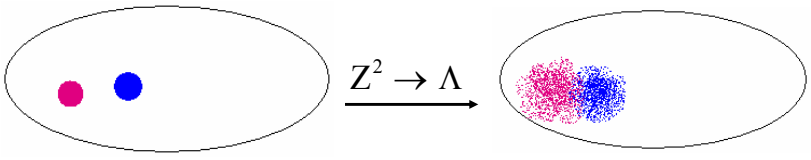
(Zhu,Wu, Mumford 95-97)

$$\text{a texture} = \Omega(h_C) = \{I : h(F_i * I) = h_{C,i}, i = 1, 2, ..., K, \Lambda \to Z^2\}$$

$h_c=(h_1, ..., h_K)$ are histograms of Gabor filters, i.e. marginal distributions of $f(I)$



Prob. Model derived from deterministic ensemble



texture ensembles :

$$\Omega(h_c)=\{I: h(I)=h_c\}$$

texture models :

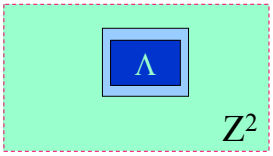
$$p(I_\Lambda \mid I_{\partial\Lambda}; \beta)$$

Markov random fields and FRAME models on finite lattice (Zhu, Wu, Mumford, 1997):

$$p(I_\Lambda \mid I_{\partial\Lambda}\ ; \beta) = \frac{1}{z(\beta)} \exp\{-\sum_{j=1}^{k} \beta_j h_j(I_\Lambda \mid I_{\partial\Lambda})\}$$

Prob. Model derived from deterministic ensemble

Wu and Zhu '99



Theorem

For a very large image from the Julesz ensemble

$$I \sim \Omega(h_c) = \{ I : h(I) = h_c \}$$

any local patch of the image I_Λ given its neighborhood follows a conditional distribution specified by a FRAME model $p(I_\Lambda \mid I_{\partial\Lambda} : \beta)$

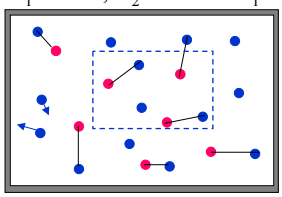
Theorem

As the image lattice goes to infinity, $f(I; h_c)$ is the limit of the FRAME model $p(I_\Lambda \mid I_{\partial\Lambda} : \beta)$, in the absence of phase transition.

A texture $\longleftrightarrow$ h_c $\longleftrightarrow$ β

Correspondence to ensembles in Stat. Physics

$N_1 = 10^{23}$, $N_2 = 10^{18}$ $N_1 + N_2 = 10^{23}$



Micro-canonical Ensemble = $\Omega(N, E, V) = \{ s : h(S) = (N, V, E) \}$

A large system with fix number of elements N, volume V, and energy E.

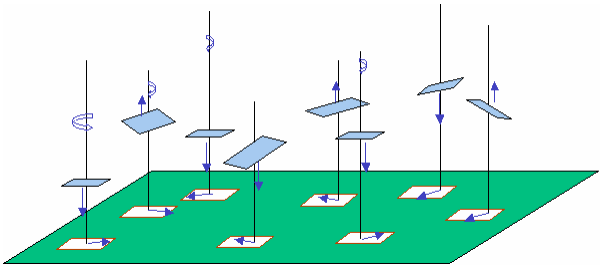
What are the basic elements in the ensemble of visual patterns?

The minimax entropy principle does not tell us about it !

Family II: layered generative modeling

Seeking Fundamental Image Elements (Isolated)

Sparse coding, Olshausen and Felds, 95
Transformed Component Analysis, Frey and Jojic 00
Textons, Leung and Malik 99, Guo, Zhu and Wu, 01,02, (Dated back to Julesz 70s)
Image primitives, Guo, Zhu and Wu, ICCV, 03 (Dated back to Marr)



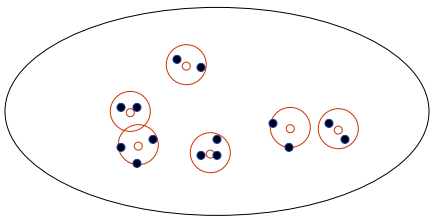
Minimizing the Kolmogorov entropy

The learning problem is to pursue the *best dictionary* so as to minimize the the Kolmogorov entropy.

Suppose we have a set of signals (in contrast to assuming a underlying density *f*)

$$\Omega = \{\mathrm{I}_1, \mathrm{I}_2, \cdots\}$$

which lie on a low-dimensional manifold with unknown dimension *H*. Suppose the ensemble is covered by at least $N(\varepsilon)$-balls with radius ε.



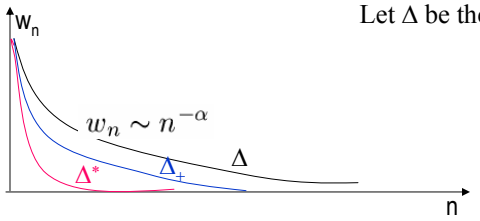
$$N(\epsilon) = (1/\epsilon)^H$$

$$H = \log_{1/\epsilon}^{N(\epsilon)} \rightarrow \log |\Omega|$$

Augmenting generative models

Therefore, the goal is to pursue the optimal bests so that the decreasing rate is fast, which corresponds to minimizing the Kolmogorov entropy.

Let Δ be the dictionary.



When the dictionary is orthogonal, there is a close relation between the Kolmogorov entropy H (or the dimension of signal) and the optimal decreasing rate of the coefficients.

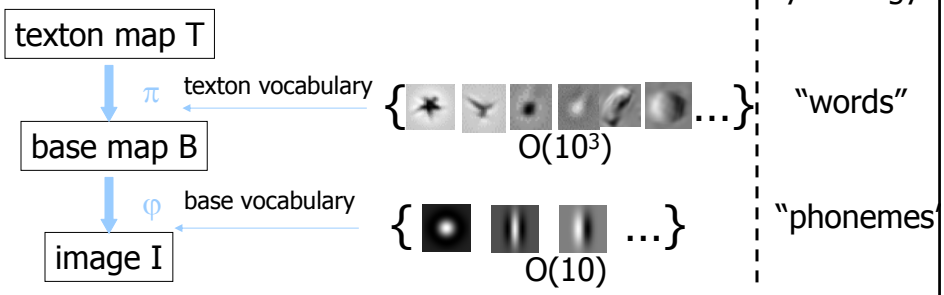
$$\alpha^* = \frac{1}{H} - \frac{1}{2}$$ (Donoho, 98)

A Three Level Generative image model

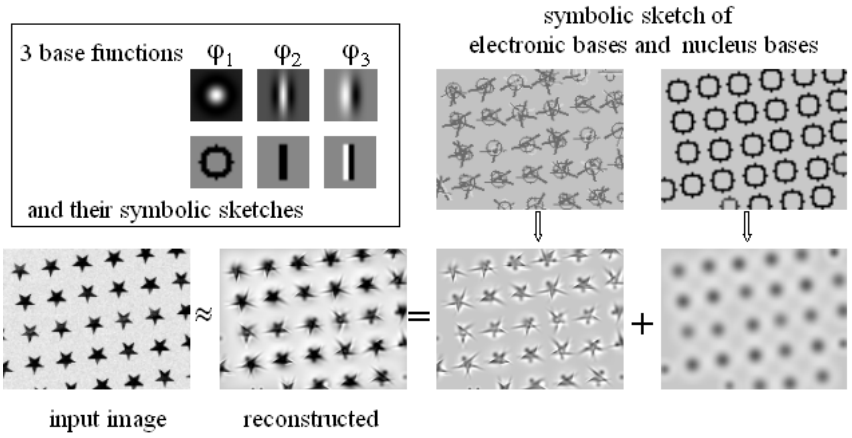
(Zhu et al 02)

Textons are defined as a vocabulary associated with a generative image model.

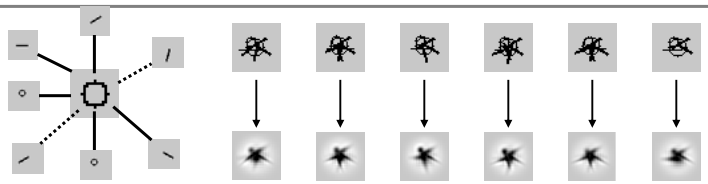
A two level image model:



Learning textons

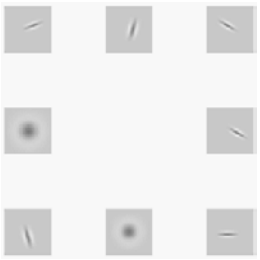


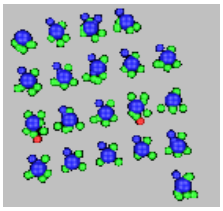
Texton with geometric variations



texton template

a sample of texton instances

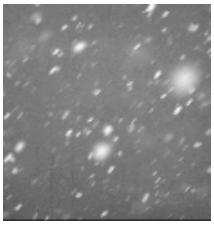


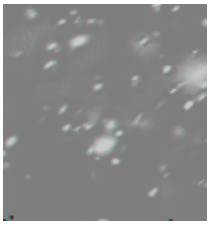


"atomic" model

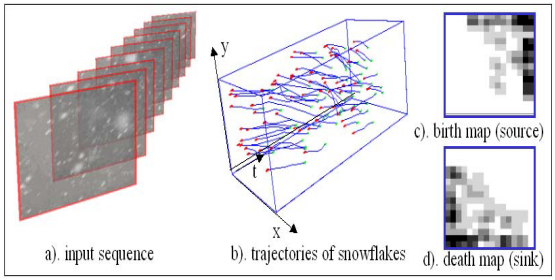
"Motons" --- Moving textons

Observed Sequence



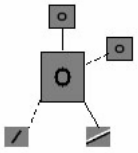


Synthesized Sequence

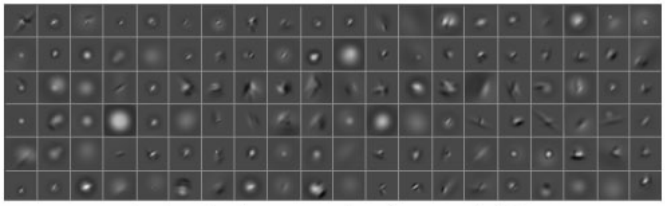


"Motons" --- Moving textons

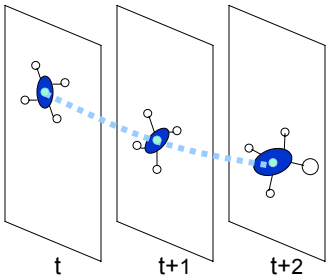
(Wang and Zhu 02,03)



a texton template π



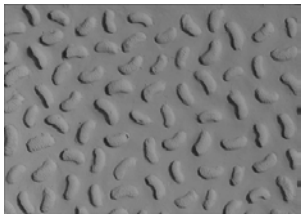
many texton instances randomly sampled from π



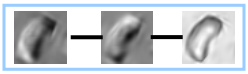
Lightons: textons with lighting variations

(Zhu et al 02)

photometric stereo images

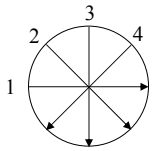


A lighton is a triplet

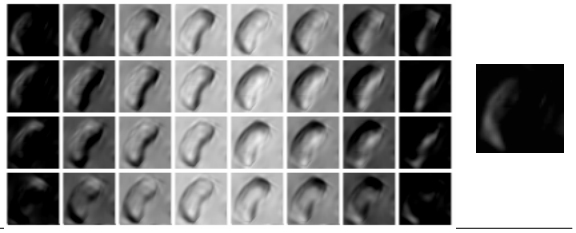


$B = \alpha_1 b_1 + \alpha_2 b_2 + \alpha_3 b_3$

Sampling the 3D elements under varying lighting directions



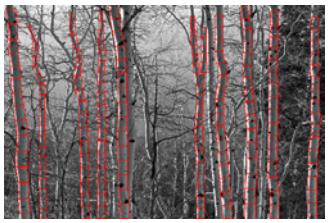
4 lighting directions

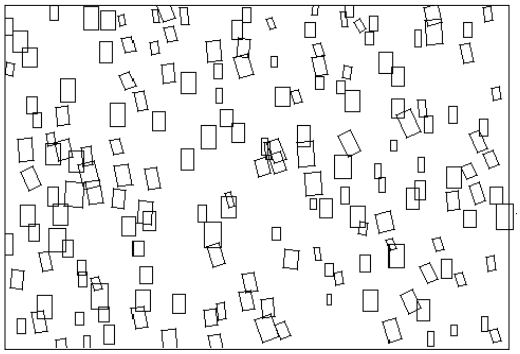


MCMC simulation of the "texton" process

Guo, Zhu, and Wu 01

The spatial relationship between textons are characterized by the MRF model (family I)





The textons form dynamic neighborhood (Mumford called mixed random fields)

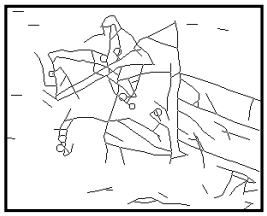
We realized that textons should not be modeled in isolation, and must pursue global structures

Primal Sketch Model

(Guo, Zhu and Wu, 2003)

org image

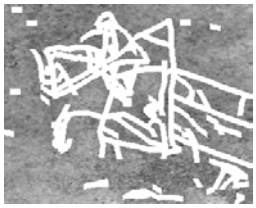
sketching pursuit process



sketches

syn image

=



synthesized textures

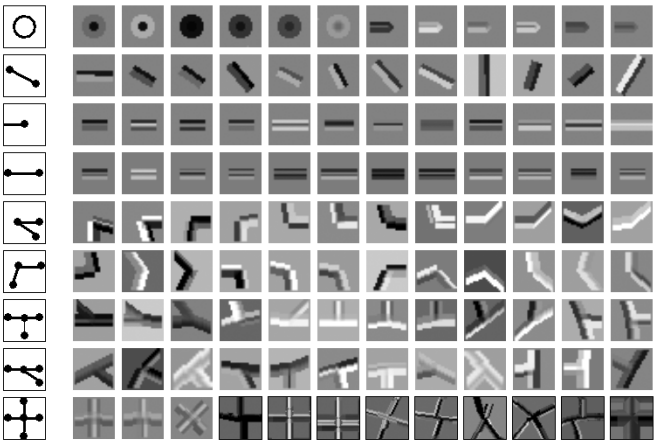
+

sketch image

Simple examples of the image primitive

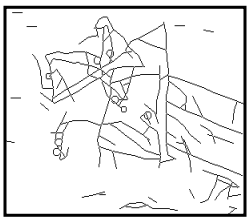
Learned texton dictionary

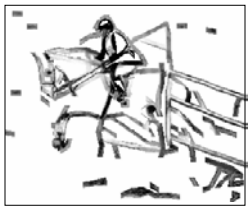
(Guo, Zhu and Wu, 2003)



(a) (b)

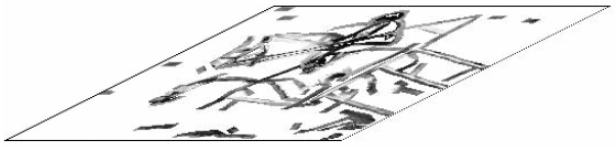
Primal Sketch: two-level model





Spatial MRF

dictionary Δ ⇩ generative

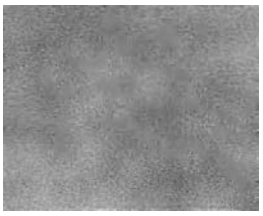


Texture MRF

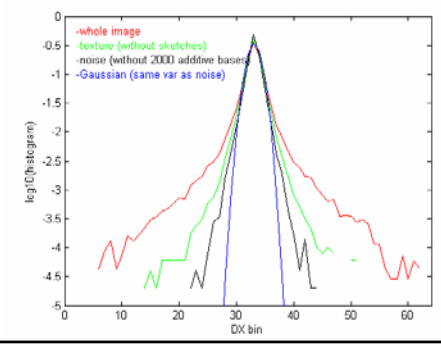
Ix Kurtosis is attributed to structures

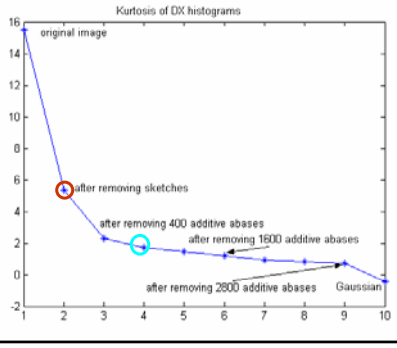
Input image

sketchable pixels removed

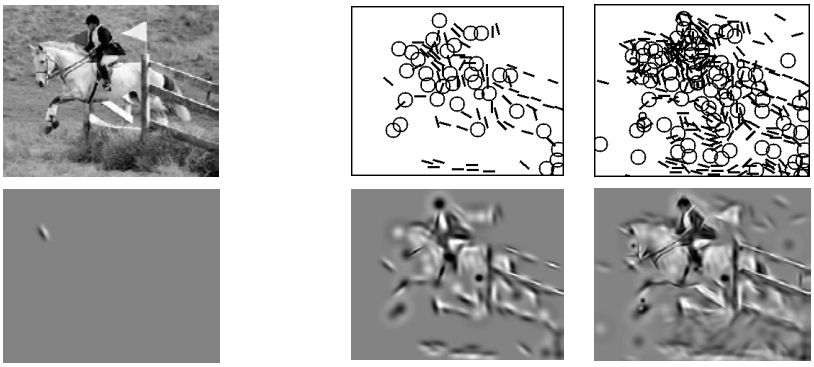


removing 800 additive bases



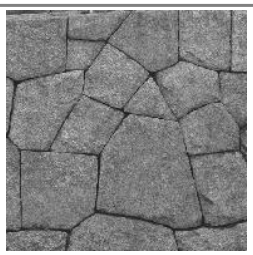


Comparison with linear additive bases

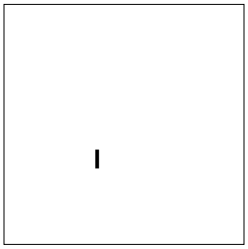


1. Missing the semantics structures
2. Not sparse enough!

More Example

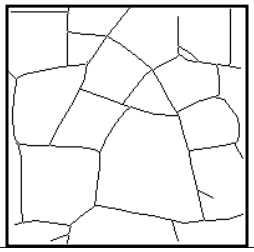


original image



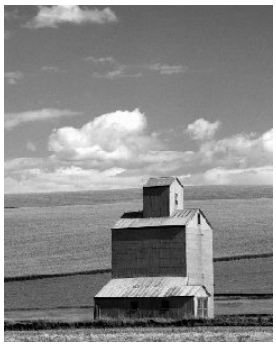
sketching pursuit process

synthesized image

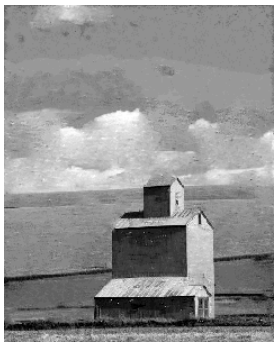


sketches

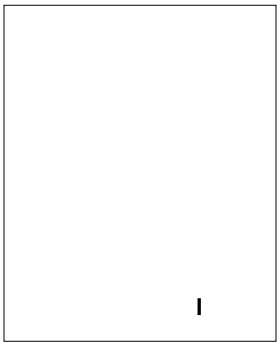
More example



original image

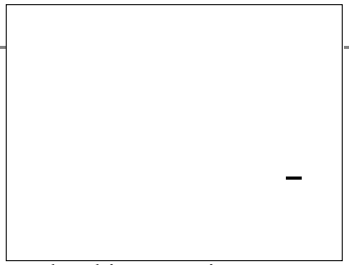


synthesized image



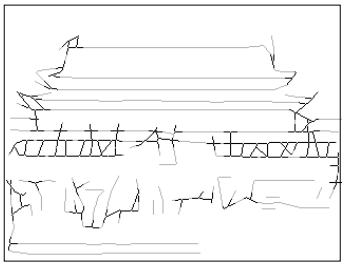
sketching pursuit process

original image



sketching pursuit process

synthesized image



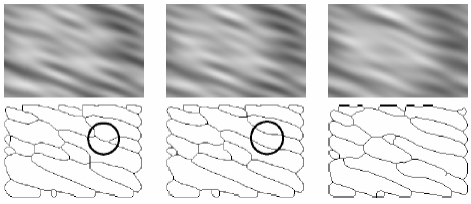
sketches

primal sketch over time: topological changes

(Wang and Zhu 2004)

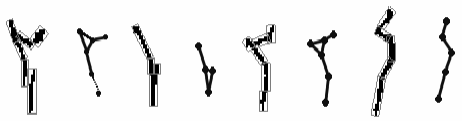
River sequence

Sketch sequence
shows topologic changes



Water sketch over time

Fire sketch over time



Summary: Generative and Descriptive manifolds

1. Textons are atomic structures in natural images.
   Each texton is specified by 4 types of intrinsic dimensions

   geometric, photometric, topological, and dynamic

$$\text{generative manifold}: \quad \Omega_D = \{I: \ I = g(W), \ W \in \Omega_d\}.$$

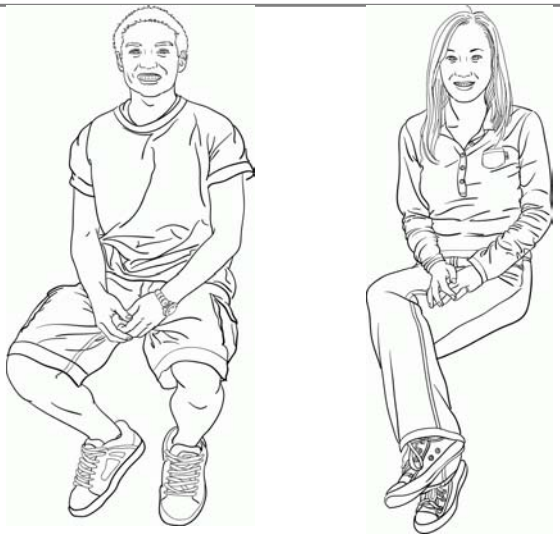
2. Textures can also be viewed as manifolds

$$\text{descriptive manifold}: \quad \Omega(h_c) = \{I: \ h(I) = h_c\}.$$

Theorems show: (1) works for low entropy regimes and (2) works for high entropy regimes

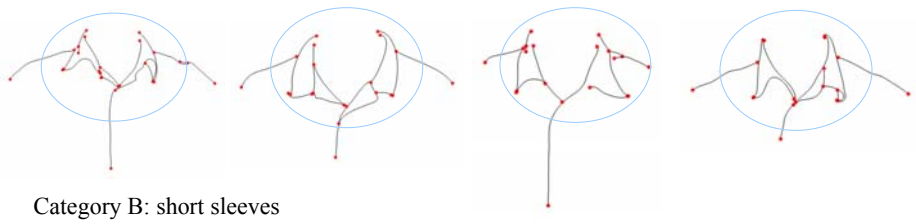
Climbing up the hierarchy of representation

Sketches of human figure

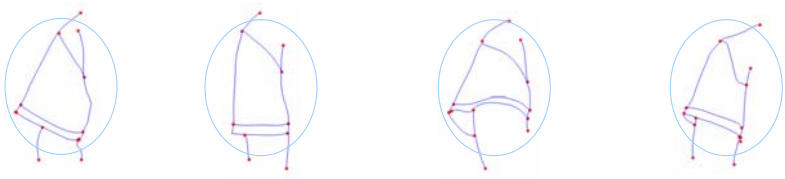


Supervised Learning of sub-graphs for parts

Category A: Collars

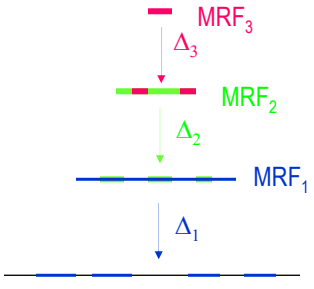


Category B: short sleeves

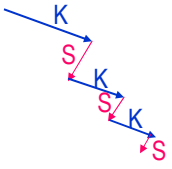


Principle: Integrating the two methods

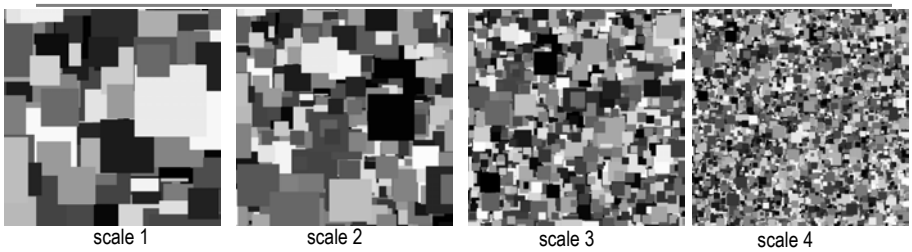
Now it is not hard to see that the modeling and learning process is to pursue hierarchic models by minimizing the Shannon entropy and Kolmogorov entropy in turns,



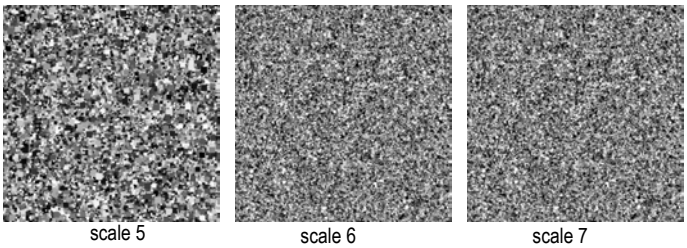
At each level, we run two steps
in reducing the entropy
K-step ---Transform the structures to higher layer
S-step --- Put a MRF/Gibbs model on the remainings



Challenges: model transitions in scale space



scale 1 scale 2 scale 3 scale 4



scale 5 scale 6 scale 7

We need a seamless transition between the MRF theory and the generative analysis

Nature image contains objects at a range of scales

This picture contains trees/leaves at four ranges of distance, over which our perception changes.

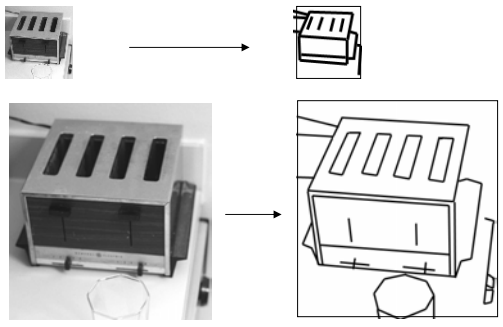
A: see individual leaves with sharp edge/boundary (occlusion model)
B: see leaves but blurry edge (additive model)
C: see a texture impression (MRF)
D: see constant area (iid Gaussian)

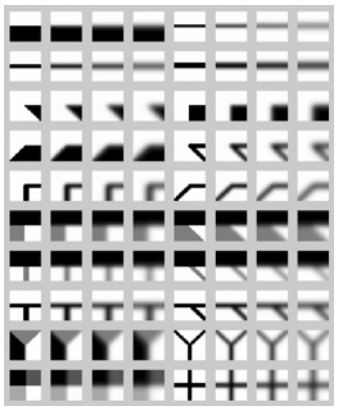
Three types of changes over scales

1. Catastrophic (texture to texton explosion),
2. graph grammatical spliting,
3. boundary sharpening

image I

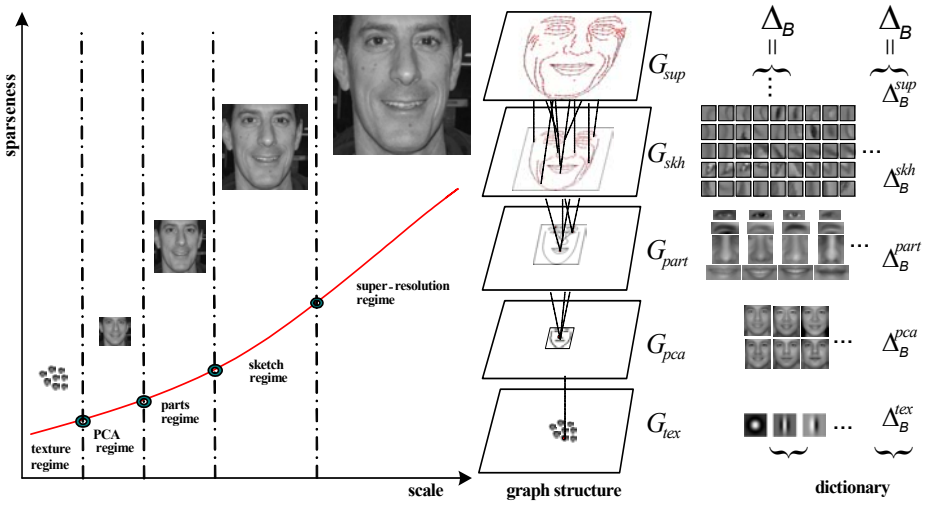
primal sketch G



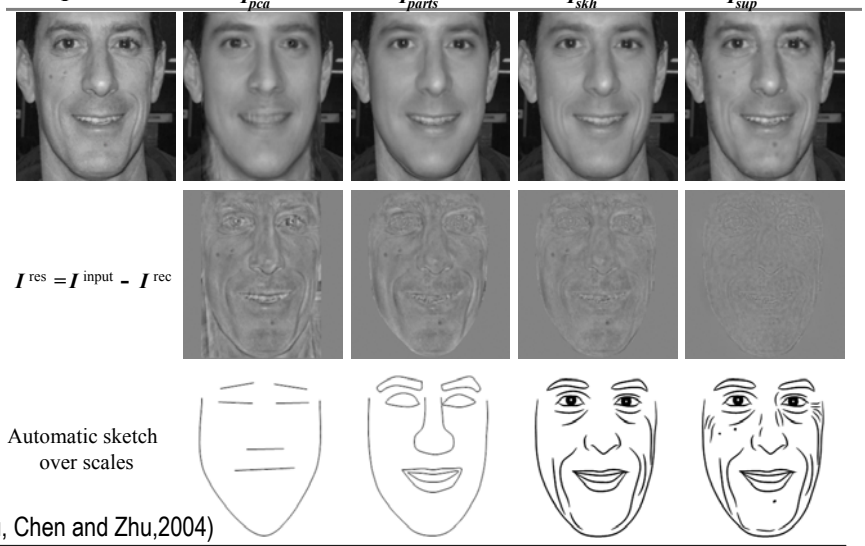


(Guo, Zhu and Wu, 2004)

"Scale – Regime" diagram (Xu, Chen and Zhu,2004)



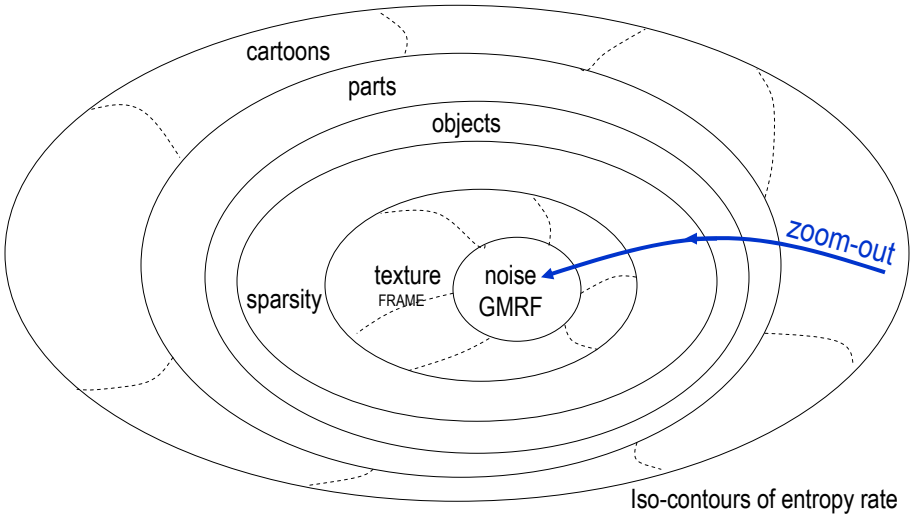
Example of reconstructed face by our model



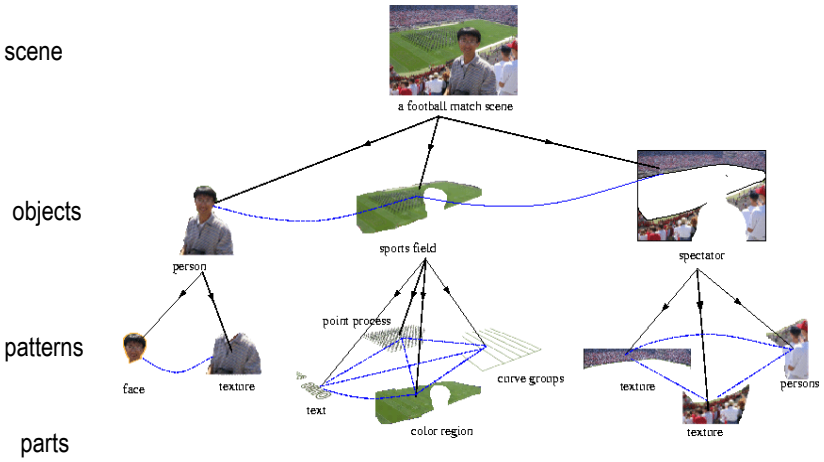
Automatic sketch over scales

(Xu, Chen and Zhu,2004)

Mapping the Image Universes at different entropy rate



Part 2: Generic Images parsing

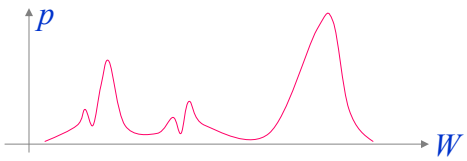


Example: parsing (Tu et al, 2000-2004)

Part 2: statistical Computing

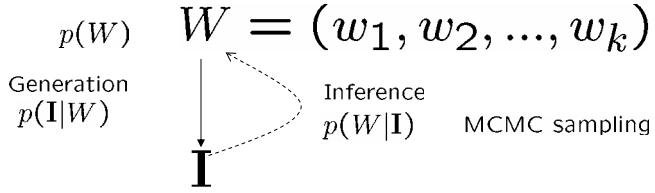
In statistics, we need to sample from the joint posterior probability.

$$(\mathrm{W}_1, \mathrm{W}_2, \ldots, \mathrm{W}_k) \sim p(\mathrm{W}|\mathrm{I}) \quad or \; p(\mathrm{I}|\mathrm{W})p(\mathrm{W})$$



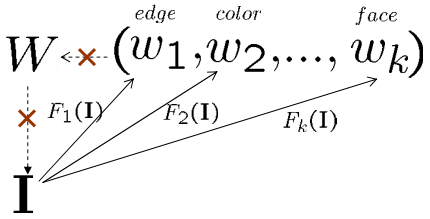
Generative vs. Discriminative Algorithms

Generative



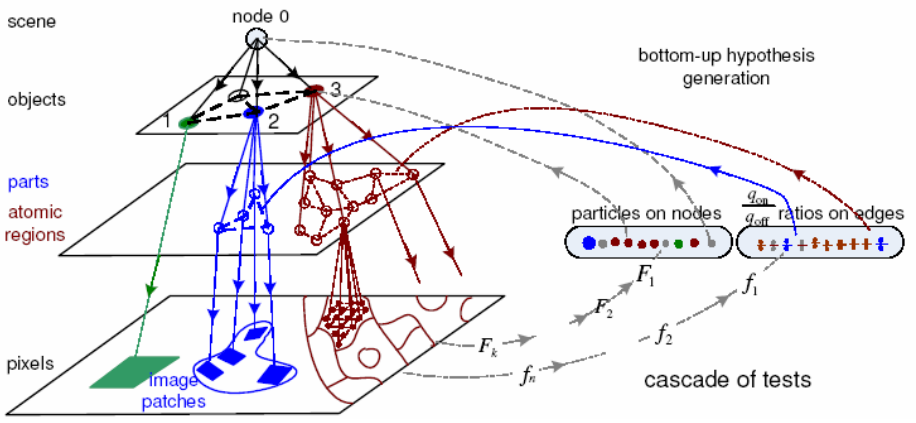
$$W^* = \arg\max \; p(W|\mathrm{I}) = \arg\max \; p(\mathbf{I}|W)p(W)$$

Discriminative

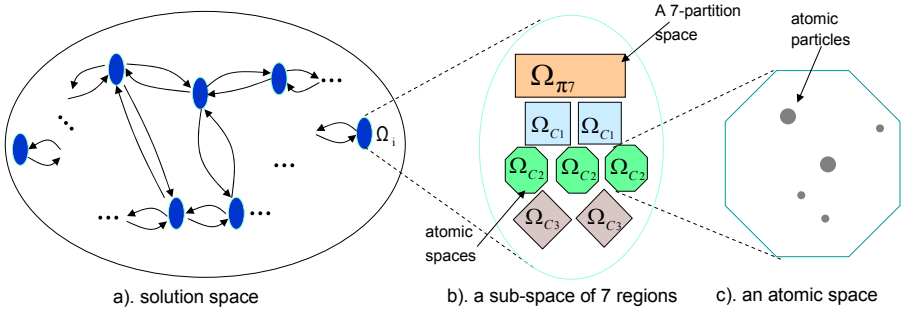


$$q(w_j|F_j(\mathbf{I})) \rightarrow p(w_j|\mathbf{I}), j = 1...k$$

Integrating generative and discriminative



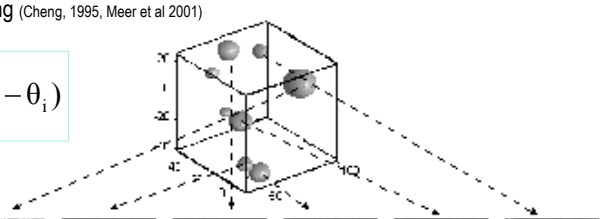
The Search/state Space



Example: Clustering in Color Space

Using Mean-shift clustering (Cheng, 1995, Meer et al 2001)

$$q(\theta|I) = \sum_{i=1}^{K} \omega_i \, g(\theta - \theta_i)$$

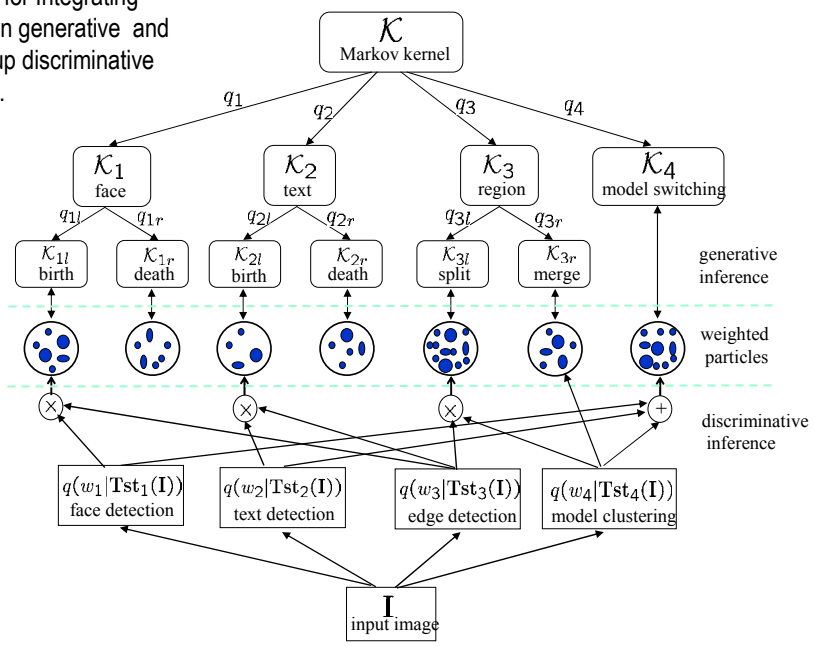


Input

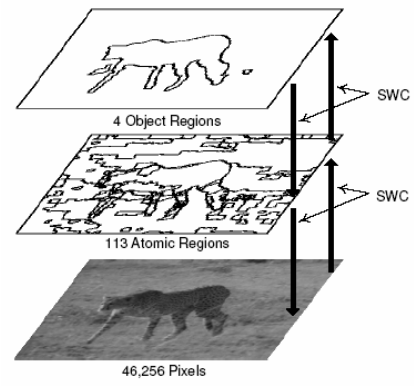
saliency maps 1 2 3 4 5 6

The brightness represents how likely a pixel belongs to a cluster.

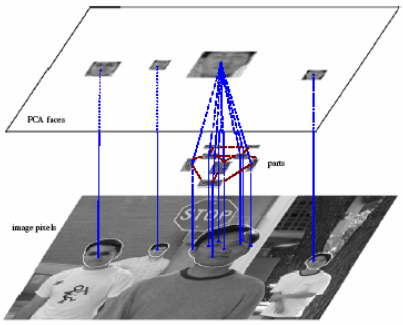
Diagram for Integrating Top-down generative and Bottom-up discriminative Methods.



Two Computing Mechanisms



(a) bottom-up graph construction



(b) Top-down graph construction

Alternating Bottom-up and Top-Down

Measuring the power of a discriminative Test

$$\delta(w|F_+) = KL(p(w|\mathbf{I})||q(w|Tst_t(\mathbf{I}))) - KL(p(w|\mathbf{I})||q(w|Tst_t(\mathbf{I}), F_+))$$
$$= MI(w||Tst_t(\mathbf{I}, F_+) - MI(w||Tst_t(\mathbf{I})) = KL(q(w|Tst_t(\mathbf{I}), F_+)||q(w|Tst_t(\mathbf{I})))$$

Measuring the power of sub-kernels

$$W_t \sim \mu_t(W) = \nu(W_0) \circ K_{a(1)} \circ K_{a(2)} \circ \cdots \circ K_{a(t)}$$

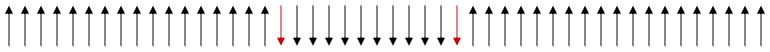
$$\delta_{a(t)} \stackrel{def}{=} KL(p(W|\mathbf{I})||\mu_t(W)) - KL(p(W|\mathbf{I})||\mu_{t+1}(W)) = KL(K_{a(t)}(W_t|W_{t+1})||p_{MC}(W_t|W_{t+1}))$$

Gibbs sampler with Ising / Potts models

$$p(I) = \frac{1}{Z}\exp\{-\beta \sum_{<s,t>} 1(I_s = I_t)\} = \frac{1}{Z}\prod_{<s,t>} \exp\{-\beta \cdot 1(I_s = I_t)\}, \quad <s,t> \in E_o$$

For example, in a 1D string of spins, suppose we use a Gibbs sampler to flip one spin at a time It has a p=½ probability for flipping the spin at the boundary. Flipping a string of length n will need on average

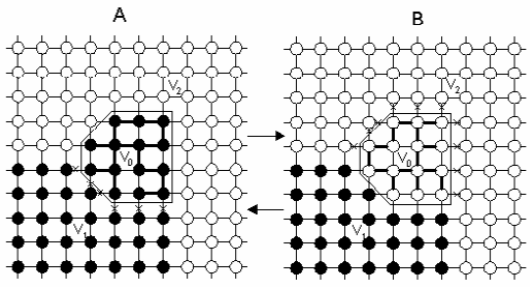
$t = 1/p^n = 2^n$ steps! This is exponential waiting time.



1/2 1/2

SW with Ising / Potts models

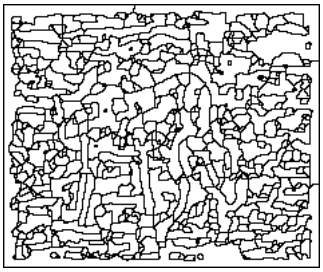
Swedsen-Wang (1987) is an extremely smart idea that flips a patch at a time. There are multiple interpretations. We explain it from the Metropolis-Hastings method.



Each edge in the lattice e=<s,t> is associated with a constant probability q.
If s and t have different labels at the current state, e is turned off.
If s and t have the same label, e is turned off with probability q.
Thus each object is broken into a number of connected components (subgraph).

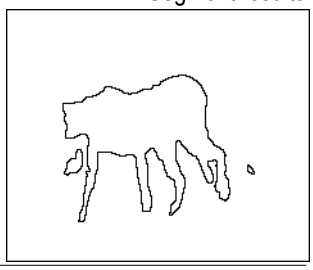
Convergence comparison: in sweep#

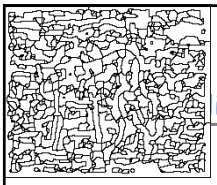
Input image



Atomic regions

Segment results

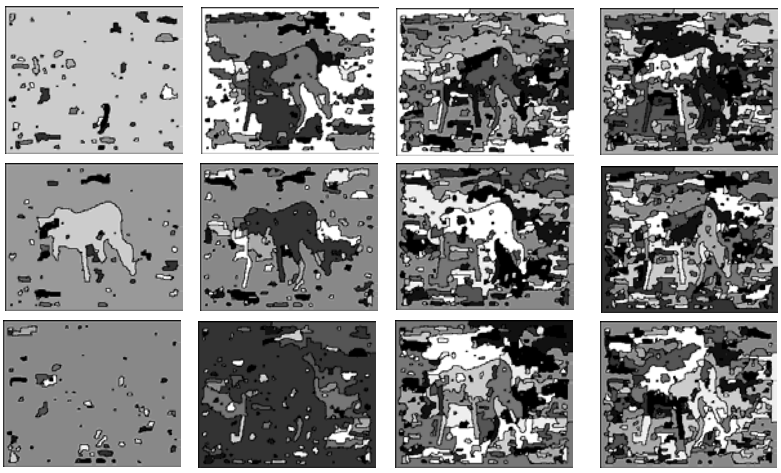




Graph partition/clustering:

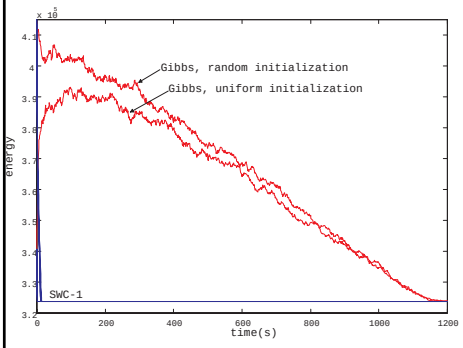
ing the discriminative model in the partition space

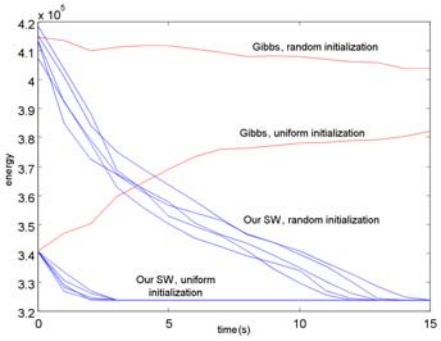
T=1 T=2 T=4 T=8



Convergence comparison: in cpu time

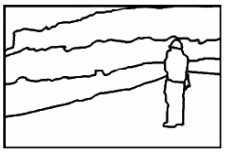
zoomed-in view

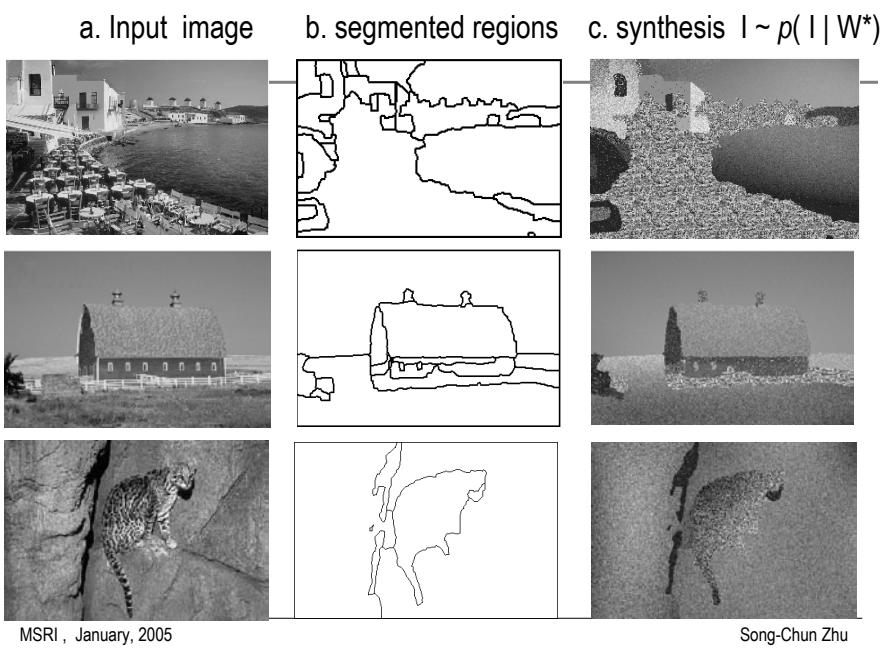




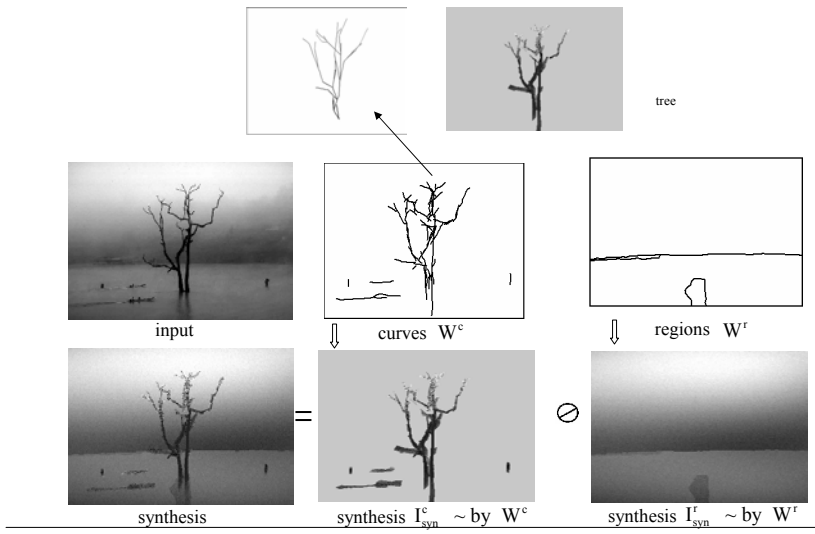
The Berkeley Benchmark Study

(David Martin et al, 2001)

test images	DDMCMC	manual segment	"error" measure
			0.1083
			0.3082
			0.5627



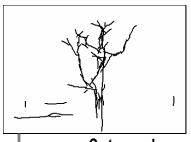
Parsing images into regions and curves



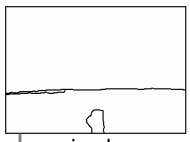
from image parsing to 3D

Example I: 3D reconstruction from a Single Image (Han and Zhu, 2003)

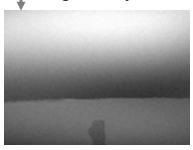
input I

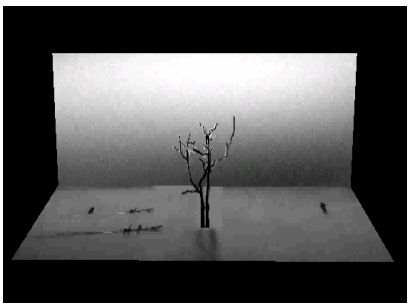


curve & tree layer



region layer



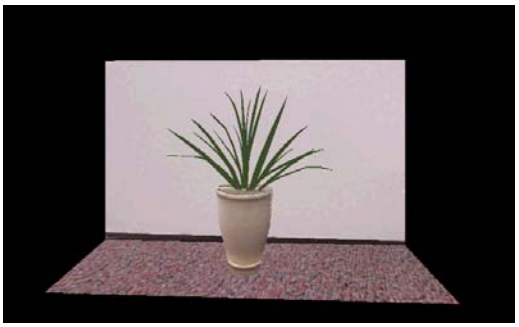


3D reconstruction and rendering

from image parsing to 3D

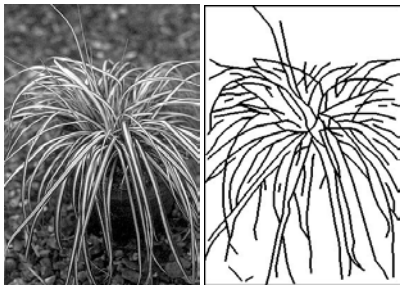
3D reconstruction (Han and Zhu, 2003)

input image



3D reconstruction from a single image

from image parsing to 3D



Input image sketch

Three new views

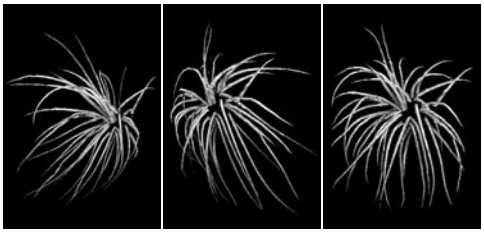
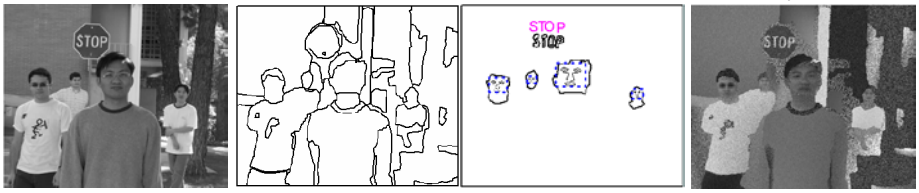


Image Parsing Results

Tu, Chen, Yuille, and Zhu, iccv2003

Input Regions Objects Synthesis



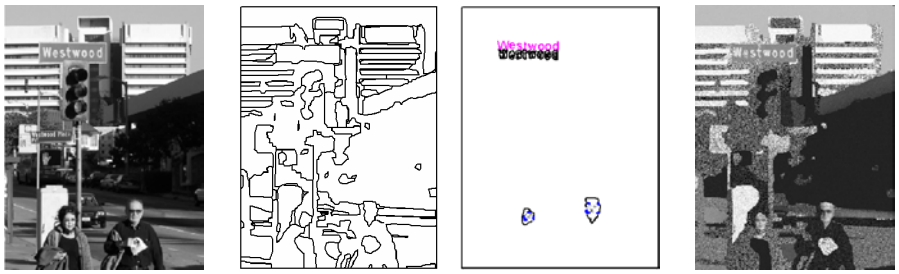
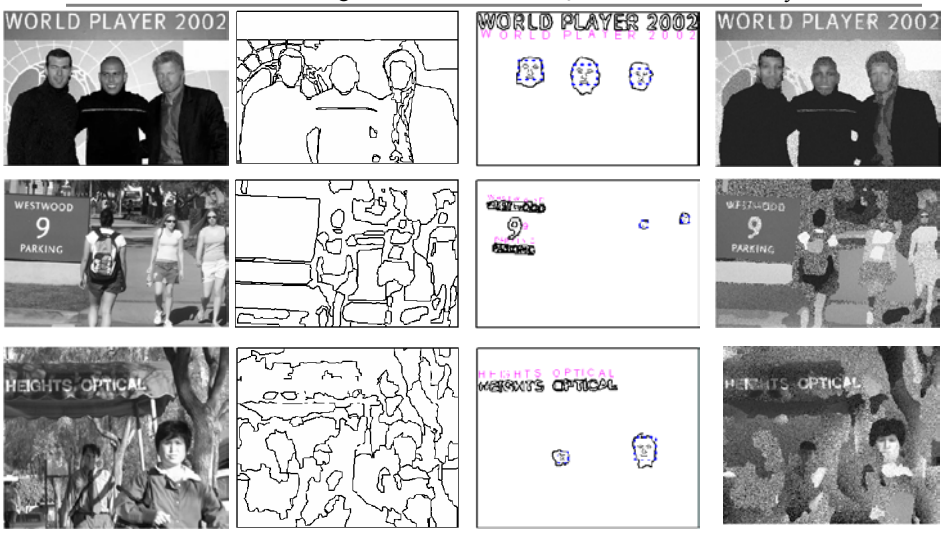


Image Parsing Results



Summary: We see three methods and we need to work on the cracks between them