Cluster Sampling and Data-Driven Markov Chain Monte Carlo

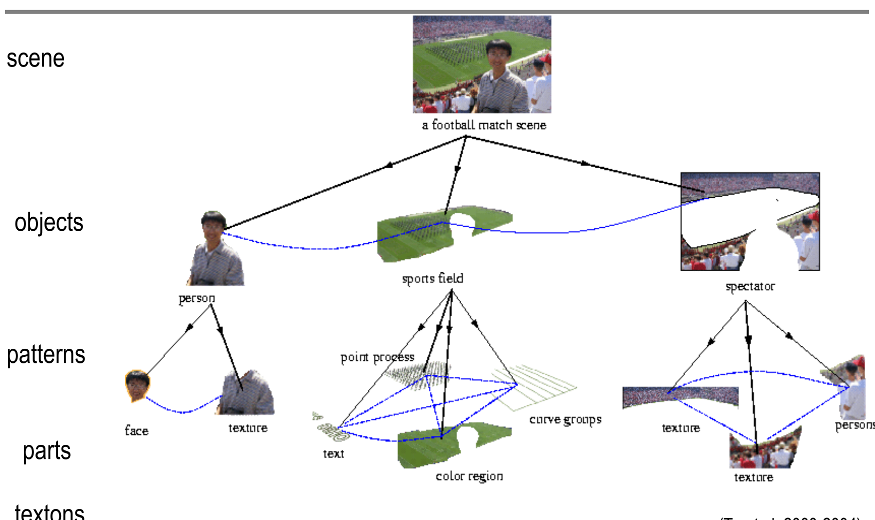
Song-Chun Zhu

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

Joint work with Adrian Barbu, Zhuowen Tu et al.

Introduction to computer vision:

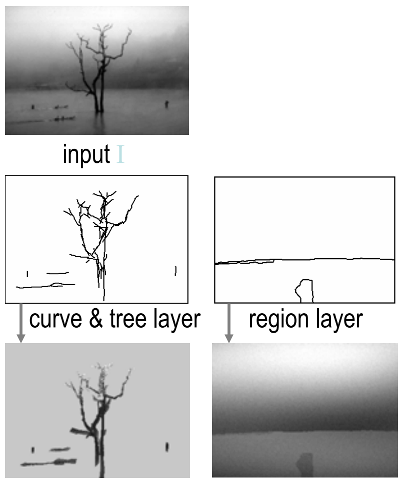
image parsing: decomposing images into their constituent visual patterns



(Tu et al, 2000-2004)

Introduction to computer vision:
3D scene construction

3D reconstruction from a Single Image



3D reconstruction and rendering

(Han and Zhu, 2003)

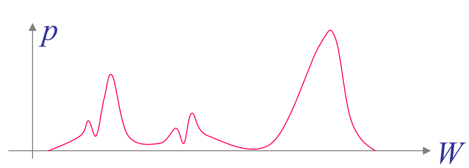
A Bayesian Formulation

Let I be an image and W the semantic representation of the world in I.

$$W^* = \arg\max_{W \in \Omega} \; p(W \mid I) = \arg\max_{W \in \Omega} \; p(I \mid W)\, p(W)$$

In statistics, we sample from a posterior probability to preserve ambiguities.

$$(W_1, W_2, \ldots, W_k) \sim p(W \mid I)$$

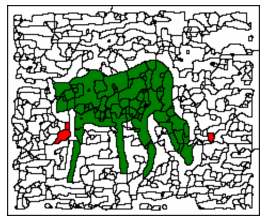


An example: image segmentation

Let π_n be the n-coloring of a lattice (image domain) Λ.

$$\pi_n = (R_1, ..., R_n), \quad \cup_{i=1}^{n} R_i = \Lambda, \; R_i \cap R_j = \emptyset \; i \neq j$$

input image



graph partition
(coloring/labeling)

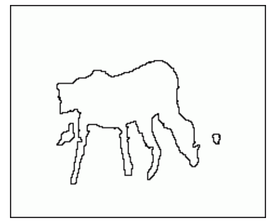


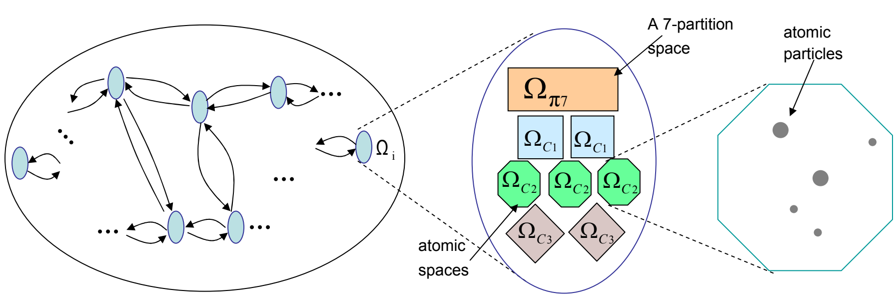
image segmentation result

The world representation is

$$W = (n, \pi_n, (\ell_i, \theta_i), i = 1, 2, ..., n)$$

(Barbu and Zhu, 2003)

The Search Space $\Omega \ni W$



a). solution space

b). a sub-space of 7 regions

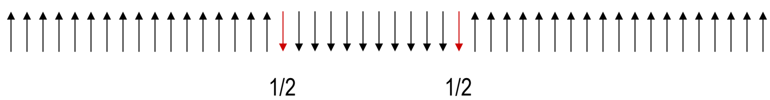
c). an atomic space

Any algorithm should be able to explore the whole space regardless its initialization. We design Markov chains that are "ergodic".

Graph (lattice) partitioning with Potts model being the prior

The Ising model (1920, two labels) and Potts model (1953, multiple labels) were used as a priori probabilities for segmentation (for fixed color n).

$$p(\pi_n) = p(C) = \frac{1}{Z}\exp\{\beta \sum_{<s,t>} 1(C_s = C_t)\},\ \beta > 0$$



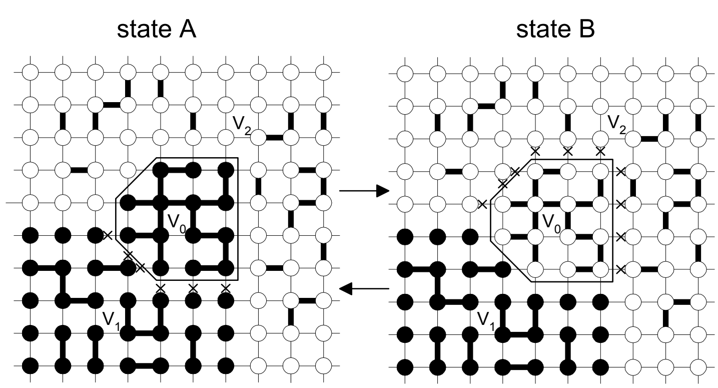
For single site Gibbs sampler (Geman and Geman 1984), the boundary spins are flipped with a p=½ probability. Flipping a string of length n will need on average

$t >= 1/p^n = 2^n$ steps!

This is exponential waiting time.

Swendsen-Wang for Ising / Potts models

Swedsen-Wang (1987) is a smart idea that flips a patch/cluster at a time.



Each edge in the lattice e=<s,t> is associated with probability $\rho=1-e^{-\beta}$.

Interpreting SW by data augmentation

One useful interpretation of SW is proposed by Edward and Sokal (1988) using the concept of data augmentation (Tanner and Wang 1987).

Augment the probability with auxiliary variables on the edges of the adjacency graph

$$U = \{u_{st} :< s, t > \in E\}$$

$$(C, U) \sim p_{ES}(C, U)$$

The augmented probability should have two nice properties,

1. The two conditional probabilities are easy to sample

$$U \sim p_{ES}(U|C) \qquad C \sim p_{ES}(C|U)$$

2. Its marginal probability on C is the target (Potts model in SW),

$$\sum_U p_{ES}(C, U) = p(C)$$

Interpreting SW by data augmentation

1. Flipping the edges by Bernoulli probability,

$$p_{ES}(U|C) = \prod_{<s,t>} p(\mu_{st}|C_s, C_t)$$

$$p(\mu_{st}|C_s, C_t) = \text{Bernoulli}(\rho \mathbf{1}(C_s = C_t))$$

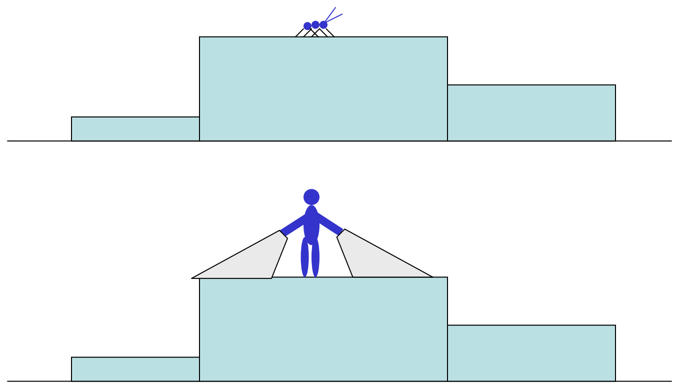
2. Flipping the color of a connected component (CCP) by uniform probability,

$$P_{ES}(C|U) = \text{unif}[\frac{\Omega_{\pi_n}}{CP(U))}]$$

CP(U) is a hard constraint that vertices in each connected component according to U has the same color. So we flip the ccp in the quotient space.

Intuition

Energy landscape



Conclusion: any two coloring states are connected in one step by SW if we flip the clusters all once.

Some theoretical results about SW

1. (Gore and Jerrum 97) constructed a "worst case"
   SW does not mix rapidly if G is a complete graph with n>2, and a certain β.

2. (Cooper and Frieze 99) had positive results
   If G is a tree, SW mixing time is O(|G|) for any b.
   If G has constant connectivity O(1), the SW has polynomial mixing time for $\rho <= \rho_0$.

3. (Huber 2002) proposed a method for exact sampling using bounding chain technique
   for small lattice with very low and very high temperature.

To engineers, the real limit of SW is that it is only valid for Ising/Potts models.
(A tiger contained in Potts' cage!)

Furthermore, it makes no use of the data (external fields) in forming clusters.

Our generalization

Barbu and Zhu (ICCV 03, CVPR04) extended SW in three aspects.

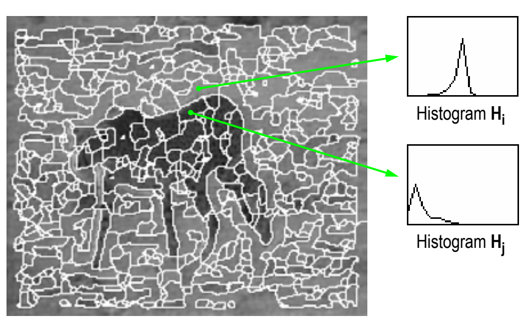
1. Generalize SW to arbitrary probabilities on graphs with variable color#.
   It can also be made into a generalized Gibbs sampler which flips a CCP at each step with simple weights on the conditional probabilities.

2. Using discriminative models (data-driven) for the edge probabilities
   The edge probability approaches the marginal posterior probability for how likely two sites s and t belong to the same color (object surface)

3. Hierarchical coloring in a multi-resolution pyramid representation.

Computing the edge weights by discriminative methods

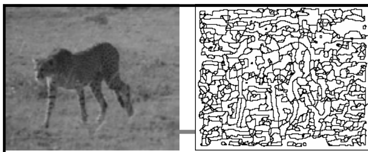
The edge probability is decided by local features



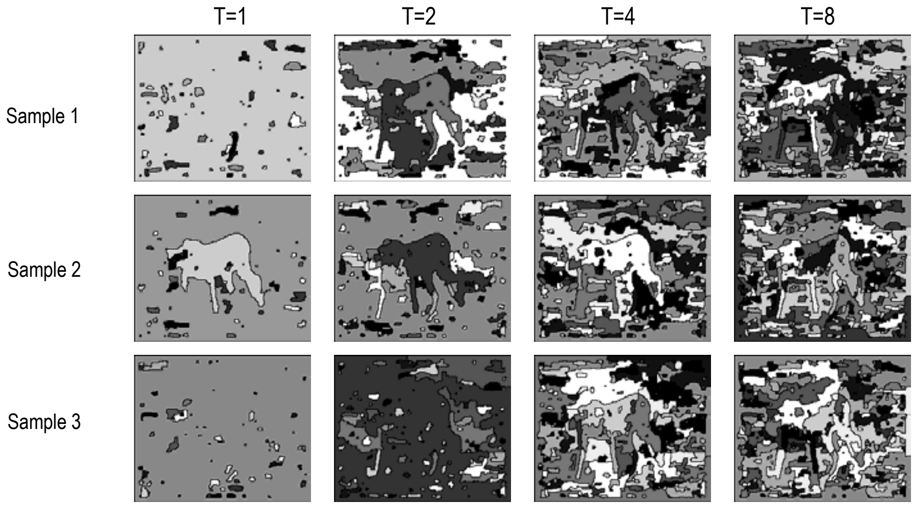
$$q_{st} = q(C_s = C_t | F_s, F_t) \rightarrow p(C_s = C_t | I)$$

$p(C_s = C_t | I)$ is a marginal probability of $p(W|I)$

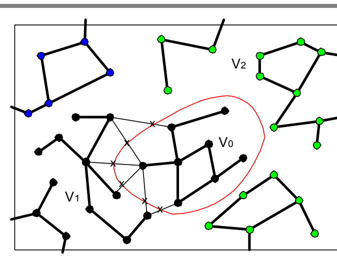
1. Konishi et al 01, Ren et al 04
2. Adaboost, Shapire 00



Clusters (connected components) by flipping the edge probabilities independently



Swendsen-Wang Cuts

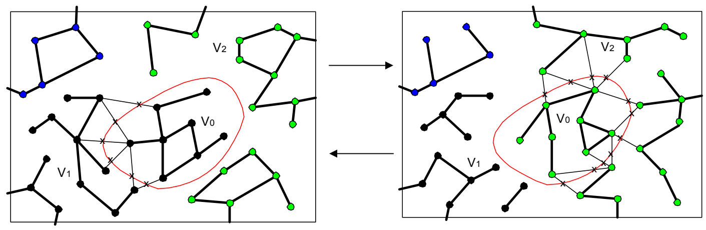


Definition: A *Swendsen-Wang cut* is the set of edges between a cluster (CCP) with other sites of the same color.

$$\mathrm{Cut}(V_0, V_1) = \{ < s, t > : s \in V_0, t \in V_1, C_s = C_t \}$$

Intuitively, this is the set of edges that must have been turned off for V_0 being a CCP.

Swendsen-Wang Cuts



State A

State B

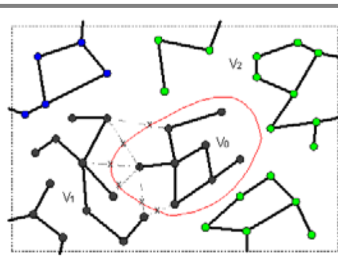
Theorem. The probability ratio for selecting CCP V_0 at states A and B is

$$\frac{q(B \to A)}{q(A \to B)} = \frac{\Pi_{e \in Cut(V_0, V_2)}(1 - q_e)}{\Pi_{e \in Cut(V_0, V_1)}(1 - q_e)}$$

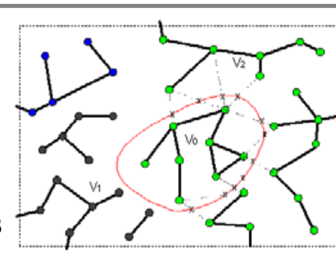
(Barbu and Zhu, 2003)

Same conclusion when multiple paths exist

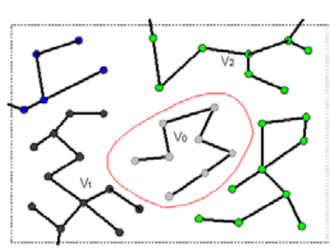
State A

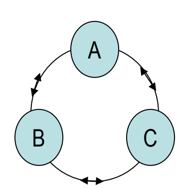


State B

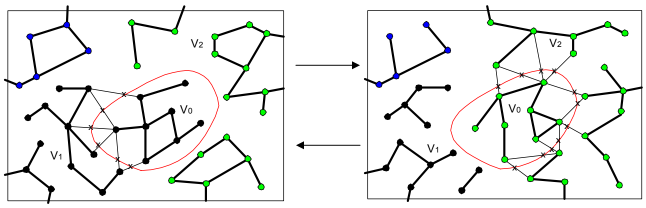


State C





Metropolis-Hasting Step



Theorem. The acceptance probability for flipping V_0 is

$$\alpha(A \to B) = \min(1, \frac{q(B \to A)}{q(A \to B)} \cdot \frac{p(B)}{p(A)})$$
$$= \min(1, \frac{\Pi_{e \in Cut(V_0, V_{l'})}(1-q_e)}{\Pi_{e \in Cut(V_0, V_l)}(1-q_e)} \cdot \frac{q(l \mid V_0, B)}{q(l' \mid V_0, A)} \cdot \frac{p(B)}{p(A)})$$

results in an ergodic and reversible Markov Chain.

Acceptance probability can be made always 1

$$\alpha(A \to B) = \min(1, \frac{\Pi_{e \in C(V_0, V_2)}(1-q_e)}{\Pi_{e \in C(V_0, V_1)}(1-q_e)} \cdot \frac{p(l_1 \mid V_0)}{p(l_2 \mid V_0)} \cdot \frac{p(B)}{p(A)}) = 1$$

If we select the label probability as

$$p(l_1 \mid V_0) = \Pi_{e \in C(V_0, V_1)}(1-q_e) \cdot p(V_1)$$
$$p(l_2 \mid V_0) = \Pi_{e \in C(V_0, V_2)}(1-q_e) \cdot p(V_2)$$

Zero rejection rate may not necessarily be an optimal design.

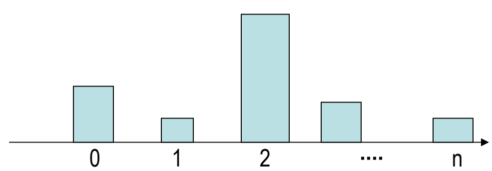
A generalized Gibbs sampler

We denote the probabilities on the SW-cuts $C(V_0, V_k)$ by weights

$$\varpi_k = \Pi_{e \in C(V_0, V_k)} (1 - q_e), \quad \mathrm{k} = 1, 2, \ldots, |\partial V_0|$$
$$\varpi_0 = 1 \quad \text{for a new label}$$

Flip the label of a CCP according to a condition probability weighted by the SW-weights

$$p(l_{\mathrm{k}} | V_0) = \varpi_k \cdot p(V_k), \quad \mathrm{k} = 0, 1, \ldots, \mathrm{n}$$



SW comes as a special case

Consider the reversible moves between states A and B by Metroplis-Hastings:

the proposal probability ratio is:

$$\frac{q(A \to B)}{q(B \to A)} = \frac{(1-q_o)^{|C(V_0,V_1)|}}{(1-q_o)^{|C(V_0,V_2)|}} = (1-q_o)^{|C(V_0,V_1)|-|C(V_0,V_2)|}$$

the probability ratio of the two states is:

$$\frac{p(A)}{p(B)} = \frac{\exp^{-\beta \cdot |C(V_0,V_2)|}}{\exp^{-\beta \cdot |C(V_0,V_1)|}} = \exp^{\beta \cdot (|C(V_0,V_1)|-|C(V_0,V_2)|)}$$

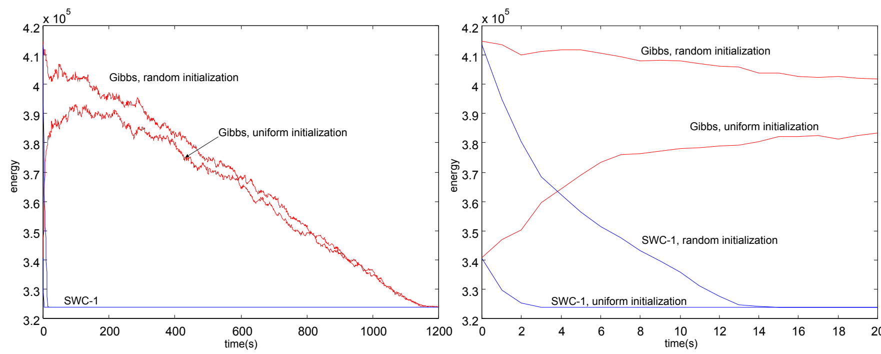
$$\alpha(A \to B) = \min(1, \frac{q(B \to A)}{q(A \to B)} \cdot \frac{p(B)}{p(A)}) = (\frac{e^{-\beta}}{1-q_o})^{|C(V_o,V_1)|-|C(V_o,V_2)|}$$

If we choose

$$q_o = 1 - e^{-\beta}$$

Then the acceptance probability is always 1.

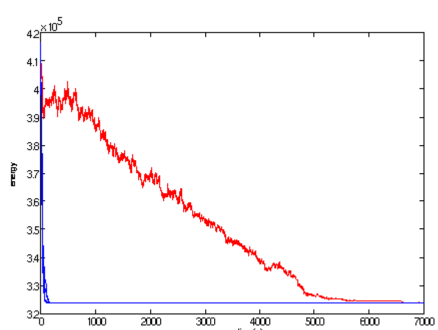
Comparison with the Gibbs sampler in CPU time



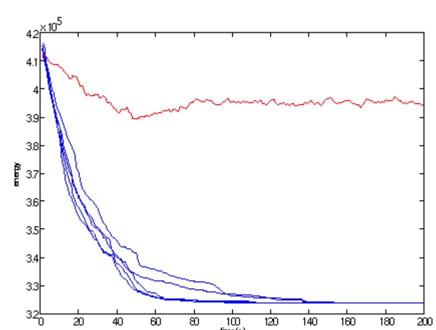
Convergence comparison of SWC-1 and the Gibbs sampler on the cheetah image, starting from a random state or from the state where all nodes have label 0. Right – zoom in view of the first 20 seconds.

Convergence comparison: in seconds

Another example

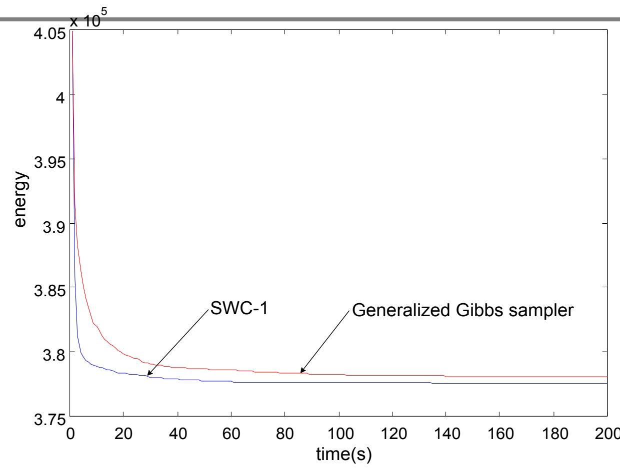


7000 seconds



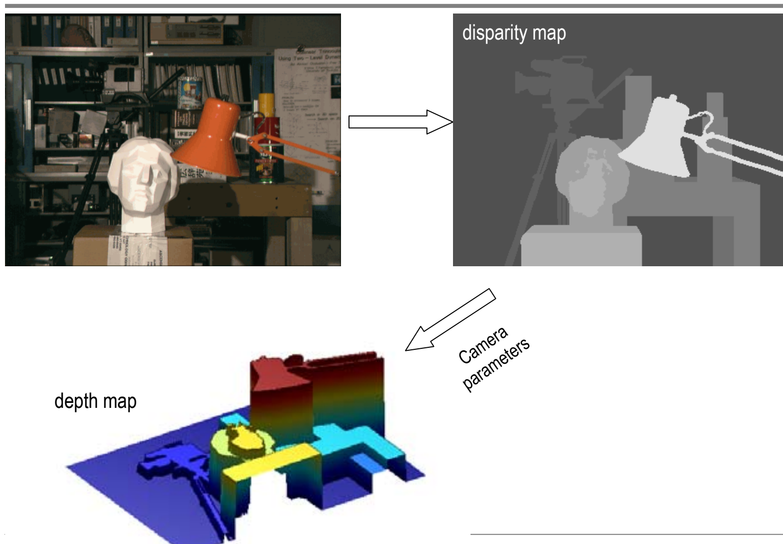
zoom-in view of the first 200 seconds

Comparison



starting from a random state.

Scene depth from stereo



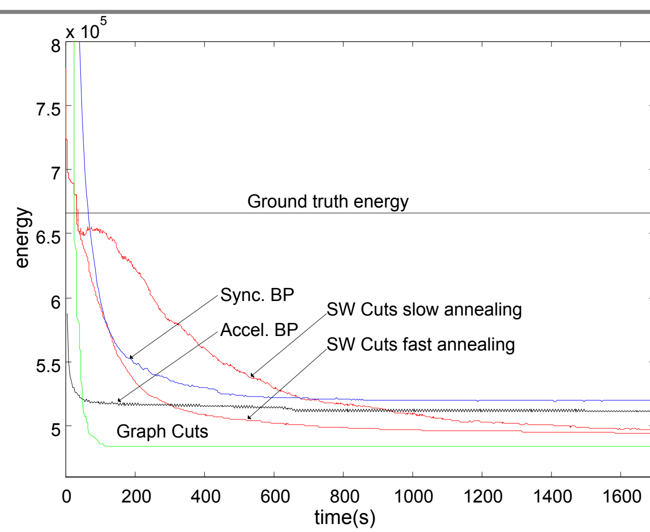
Examples on Stereo Reconstruction

Left image

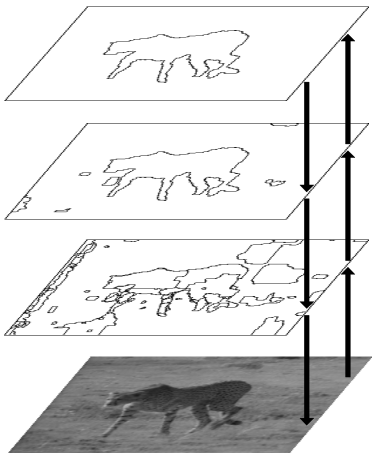
Ground truth

Segmentation result

Performance comparison with Graph Cuts and Belief propagation on a special (simplified) energy



Hierarchical partition and segmentation



X^2 – Level 2: Intensity regions are grouped into moving objects O_i with motion parameters θ_i

X^1 – Level 1: Atomic regions are grouped into intensity regions R_{ij} of coherent motion with intensity models H_{ij}

X^0 – Level 0: Pixels are grouped into atomic regions r_{ijk} of relatively constant motion and intensity

– motion parameters (u_{ijk}, v_{ijk})

– intensity histogram h_{ijk}

Motion segmentation examples

Input sequence

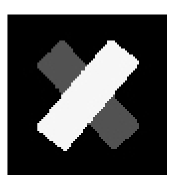
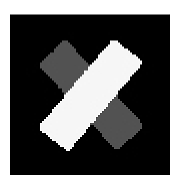


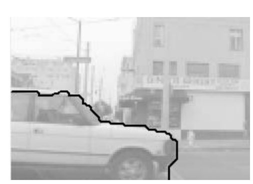
Image Segmentation



Motion Segmentation

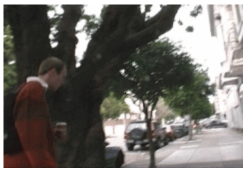
Input sequence

Image Segmentation



Motion Segmentation

Motion segmentation examples



Input sequence

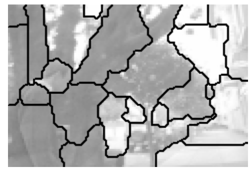
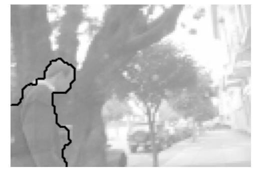


Image Segmentation



Motion Segmentation

Input sequence

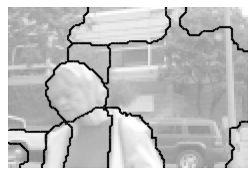
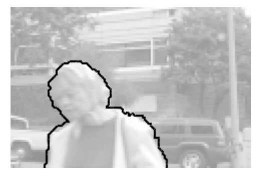


Image Segmentation



Motion Segmentation

Summary: Ideas to Improve MCMC Speed in Stat Literature

A main idea is to introduce auxiliary random variables:

$$x \sim \pi(x)$$

Augment x by variables:

T --- temperature (Simulated tempering, Narinari and Parisi, 92, Geyer and Thompson, 95)
s --- scale (Multi-grid sampling, Goodman and Sokal 88, Liu et al 94)
w --- weight (dynamic weighting, Liang and Wong 1996)
b --- bond (clustering, Swendsen-Wang, 87)
u --- energy level (slice sampling, Edwards and Sokal, 88 …)

The common problem is:
The Markov chain moves are designed a priori, without looking at the data.

Data-Driven Markov Chain Monte Carlo

Consider a reversible jump $W_A \leftrightarrow W_B$

$$\alpha(W_A \to W_B) = \min\left(1, \frac{p(W_B | \mathrm{I})\, G(W_B \to W_A)}{p(W_A | \mathrm{I})\, G(W_A \to W_B)}\right) \quad \text{or} \quad \min\left(1, \frac{p(W_B | \mathrm{I})\, q(W_A | W_B)}{p(W_A | \mathrm{I})\, q(W_B | W_A)}\right)$$

Without looking at the data, the pre-designed proposal probabilities are often uniform distributions, thus it is a blind (exhaustive) search !

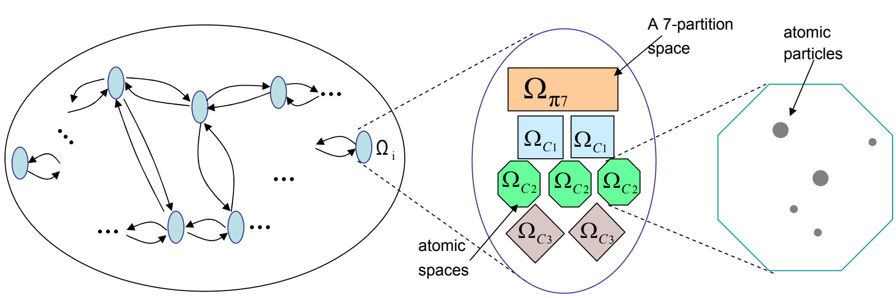
In DDMCMC,

$$\alpha(W_A \to W_B) = \min\left(1, \frac{p(W_B | \mathrm{I})\, q(W_A | W_B, \mathrm{I})}{p(W_A | \mathrm{I})\, q(W_B | W_A, \mathrm{I})}\right)$$

If $q(W_B | W_A, \mathrm{I}) \cong p(W_B | \mathrm{I}), \quad q(W_A | W_B, \mathrm{I}) \cong p(W_A | \mathrm{I})$

Then it may converges in a small number of steps !

Revisit the Search Space $\Omega \ni W$



a). solution space b). a sub-space of 7 regions c). an atomic space

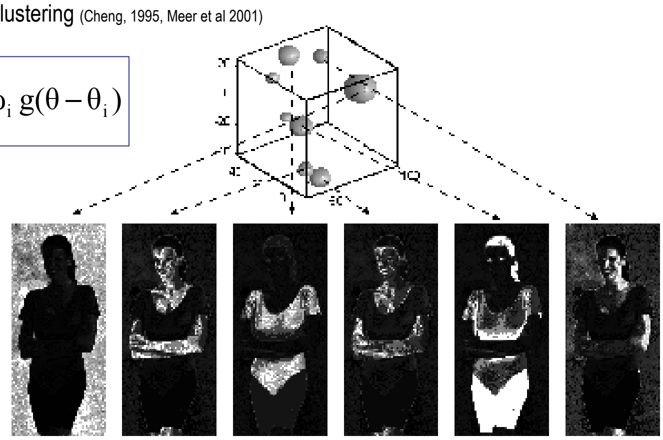
Any algorithm should be able to explore the whole space regardless its initialization. We design Markov chains that are "ergodic".

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.

Generative vs. Discriminative Algorithms

Generative

$p(W)$ $\quad W = (w_1, w_2, \ldots, w_k)$

Generation $p(\mathbf{I}|W)$

Inference $p(W|\mathbf{I})$ MCMC sampling

$\mathbf{I}$

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

Discriminative

$$W \leftarrow \times \ (w_1, w_2, \ldots, w_k)$$

edge, color, face

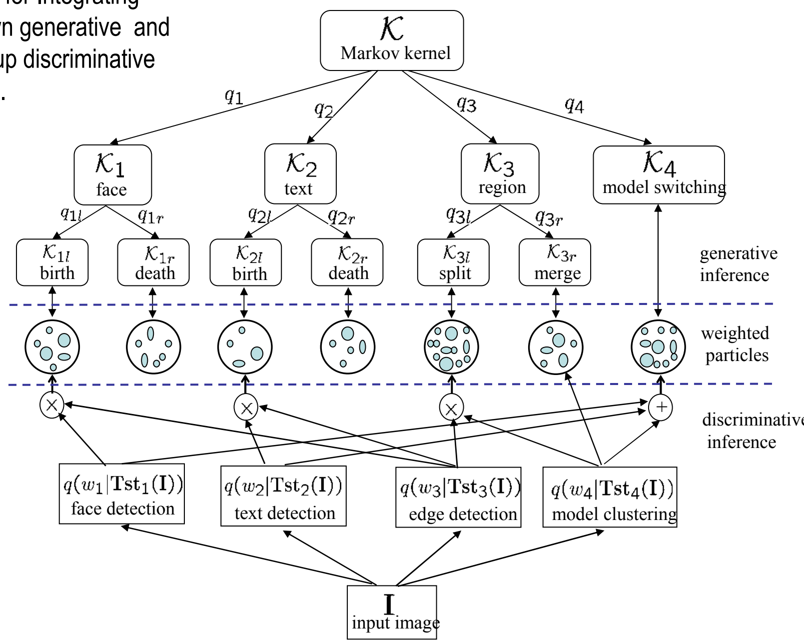
$F_1(\mathbf{I})$, $F_2(\mathbf{I})$, $F_k(\mathbf{I})$

$\mathbf{I}$

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

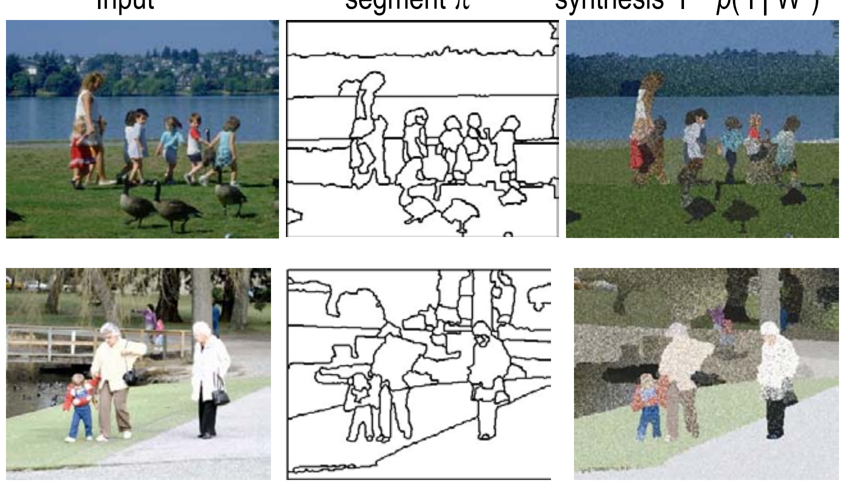
marginal posterior

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

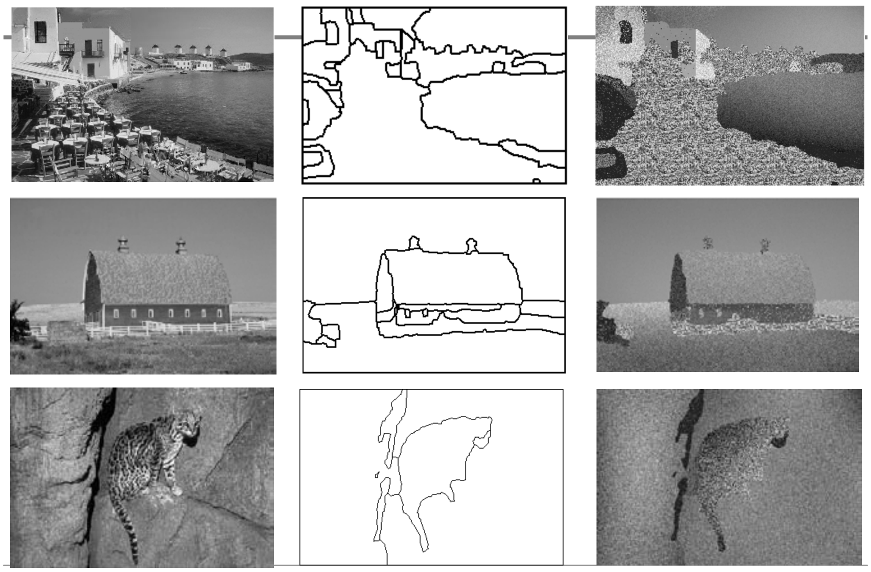


Experiments: Color Image Segmentation

Input   segment π^* synthesis I ~ p(I | W*)



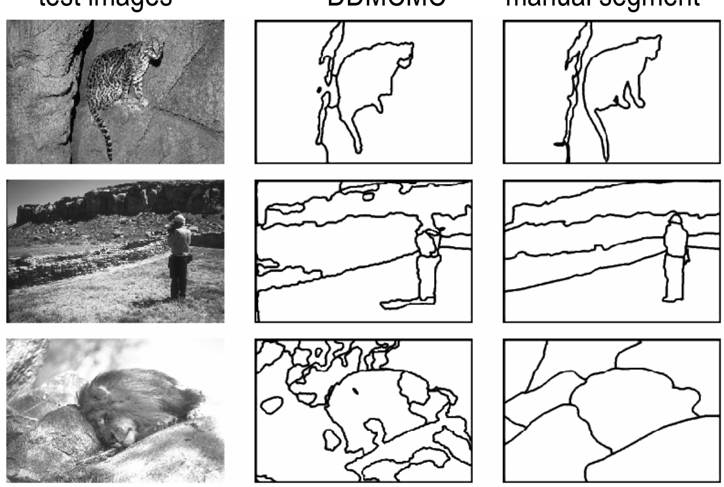
a. Input image   b. segmented regions   c. synthesis $I \sim p(I \mid W^*)$



The Berkeley Benchmark Study

(David Martin et al, 2001)

test images   DDMCMC   manual segment   "error" measure



"error" measure: 0.1083, 0.3082, 0.5627

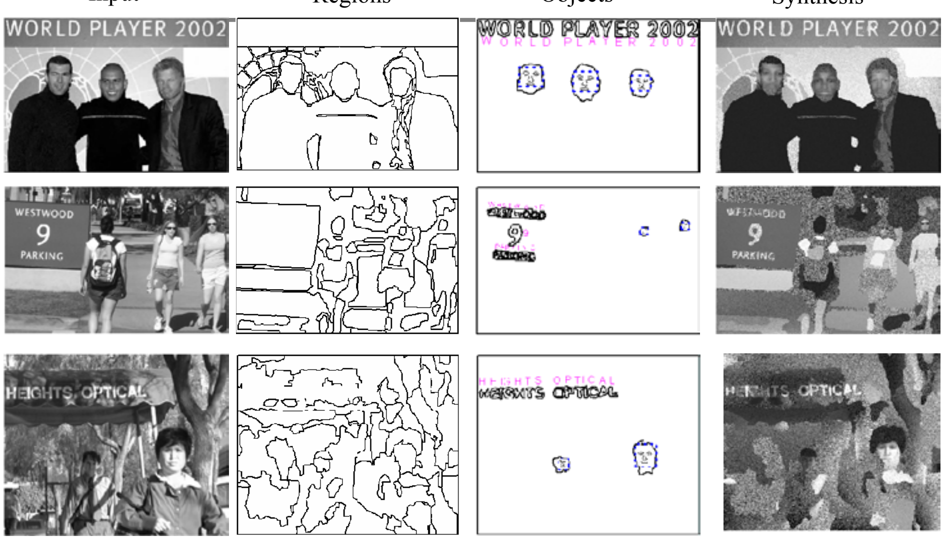
Image Parsing Results

Tu, Chen, Yuille, and Zhu, iccv2003

Input | Regions | Objects | Synthesis

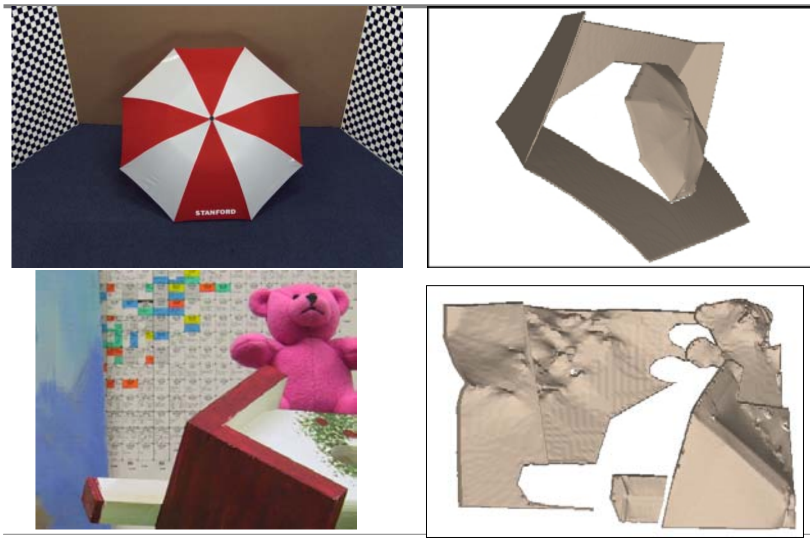
Image Parsing Results

Input | Regions | Objects | Synthesis

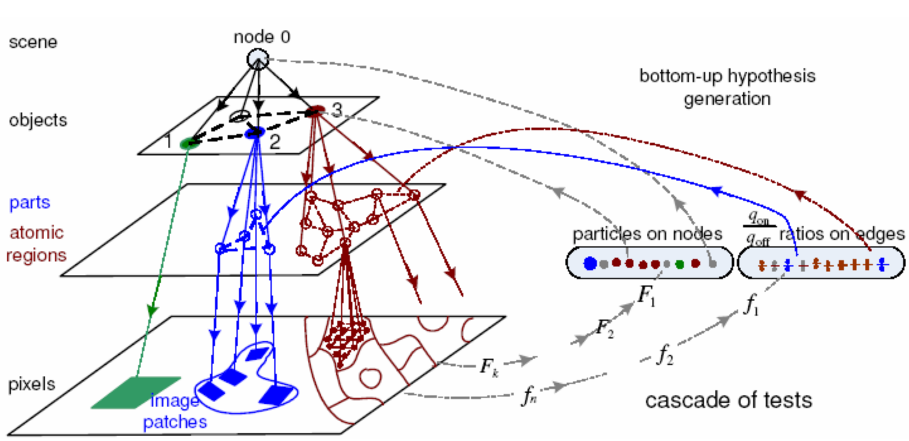


Statistics Dept. UC Berkeley, April, 2005, Song-Chun Zhu

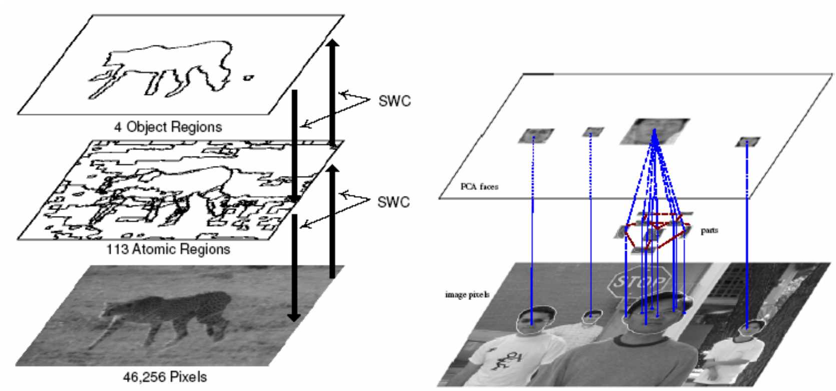
Examples on Stereo Reconstruction



Integrating generative and discriminative



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

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

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