Statistical Modeling of Visual Patterns: Part II

--- Bridging Wavelets with Markov Random Fields

Song-Chun Zhu

Departments of Statistics and Computer Science
University of California, Los Angeles

Note: Here the word "*wavelets*" is used in a very broad sense. What I really mean is an over-complete dictionary of image bases.
Since the previous lectures were focused on wavelets, this lecture plans to focus on the relationship to MRF models etc. So no discussion on wavelets specifically.

Outline of this lecture

1. A few slides to summarize Part I:
   --- from stat. physics to visual modeling.
2. Image Coding with over-complete dictionary:
   --- advantages and problems.
3. Textons --- the atoms of visual perception
4. Gestalt ensemble --- a Markov random field model of textons
5. Visual learning --- integrating the descriptive and generative models.

Natural images contain a wide variety of visual patterns

Visual Modeling: Knowledge Representation

1. What is a mathematical *definition* and *model* of a visual pattern ?

2. What are the *vocabulary* for these visual patterns ?
   By analogy to language, what are the phonemes, words, phrases, ...

3. Can these models and vocabulary be *learned* from natural images and video sequences?

Review: Definition of a Visual Pattern

Our concept of a pattern is an *abstraction* for an ensemble of *instances* which satisfy some statistical description:

For a homogeneous signal s on 2D lattice Λ, e.g. s = I,

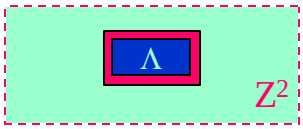
$$\text{a pattern} = \Omega(h_c) = \{ s:\ h(s) = h_c;\ \ \Lambda - Z^2 \},\quad f(s)=1/|\Omega(h_c)|.$$

h_c is a summary and s is an instance with details.

This equivalence class is called a Julesz ensemble (Zhu et al 1999)

Review: From definition to modeling

Equivalence of Julesz ensemble and FRAME models



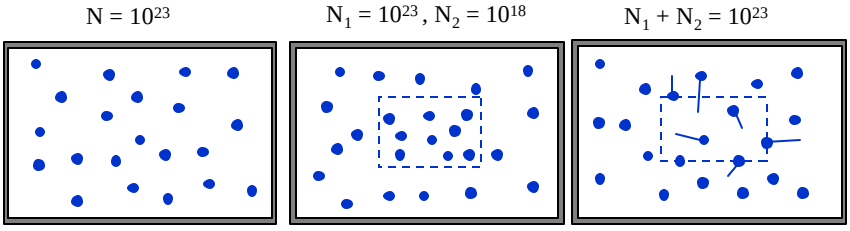
Theorem

For a very large image from the Julesz ensemble $I \sim f(I; h_c)$ any local patch of the image I_Λ given its neighborhood follows a conditional distribution specified by a FRAME model $p(I_\Lambda \mid I_{\partial\Lambda} : ß)$

$$\text{A visual pattern v} \longleftrightarrow h_c \longleftrightarrow \beta$$

Review: Ensembles in Statistical Mechanics

(See a stat. physics book by Chandler 1987)



Micro-canonical Ensemble Canonical Ensemble Grand-Canonical Ensemble

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

We assumed pixels, points in typical Gibbs models. This should be generalized.

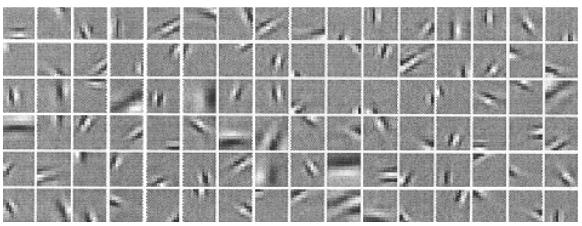
Very likely, the basic elements vary from different ensembles, and thus they need to be learned automatically from natural images --- for generic vision models.

Sparse coding (Olshausen and Fields, 95 nature).

Learning an over-complete image basis from natural images

An image I is assumed to be a linear addition of some image bases ψ_i, i=1,2, ……,n which are selected from an over-complete basis (dictionary).

$$I = \Sigma_i \alpha_i \psi_i + n, \quad \alpha_i \sim p(\alpha) \quad \text{iid}$$



It was said that these learned bases resemble cells in primate V1.

Advantages of Sparse Coding Model

Advantages --- from the perspective of modeling

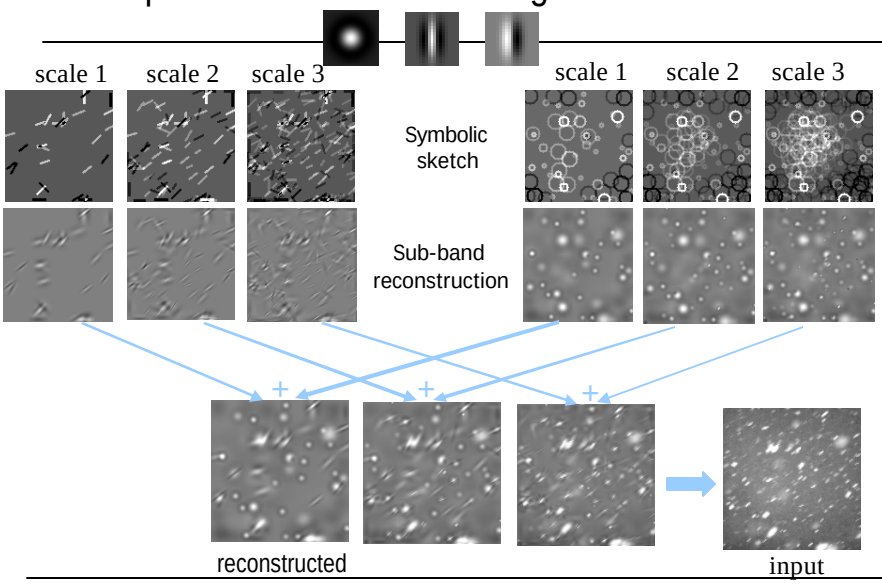
1. Dimension reduction

   --- The number of bases is often 100-times smaller than the number of pixels.

2. Variable decoupling

   --- the bases are less correlated than pixel intensities. In ideal case, one gets completely independent components, but it is impractical in natural image

Example of Coarse-to-fine image reconstruction



Remaining problems of Sparse coding

Remaining problems --- from the perspective of modeling

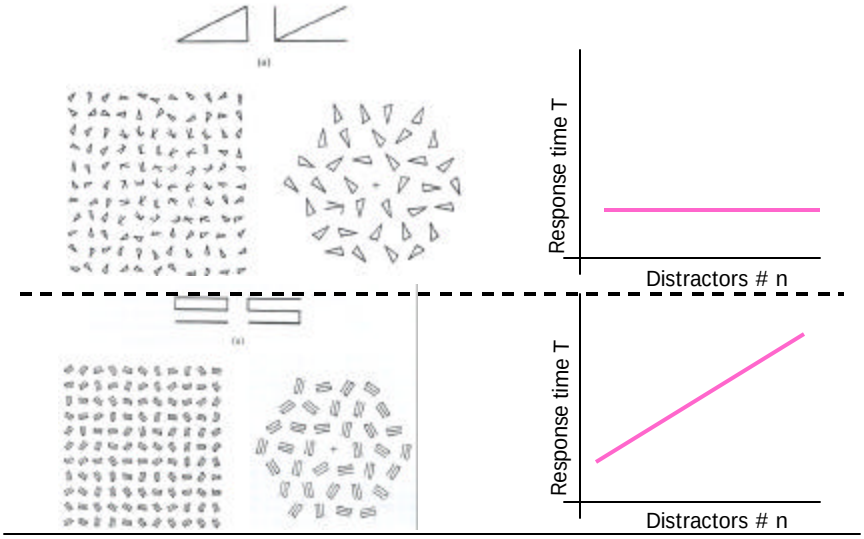
1. As bases are not independent,
   how do we model their spatial relationships and dependence?

2. What are the larger structures / image unit beyond base/wavelet representation?
   --- A similar question in neuroscience:
   what are the cell functions above V1?

   --- A similar question in psychophysics:
   what are the "textons"?

Textons: Fundamental elements in Visual perception

(Julesz, 1981 Textons: the fundamental elements of visual perception, Nature)



"Early Vision"

Julesz's heuristic (or axiom):

Textons are the fundamental elements in preattentive vision,

1. Elongated blobs
2. Terminators
3. Crossings

Texton plasticity:

Textons in human perception may change by training !

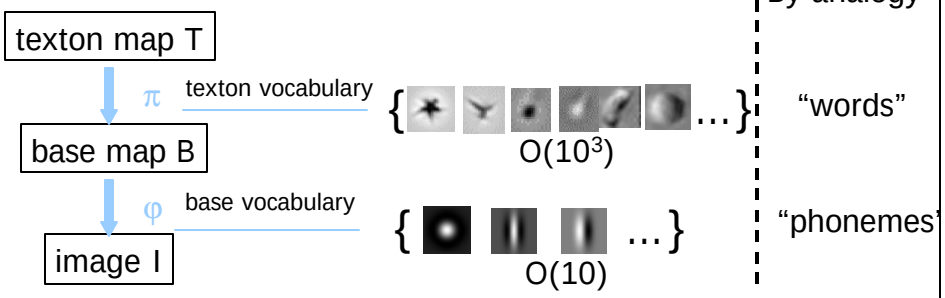
(Karni and Sagi, 1991)

What are the textons for the ensemble of natural images?

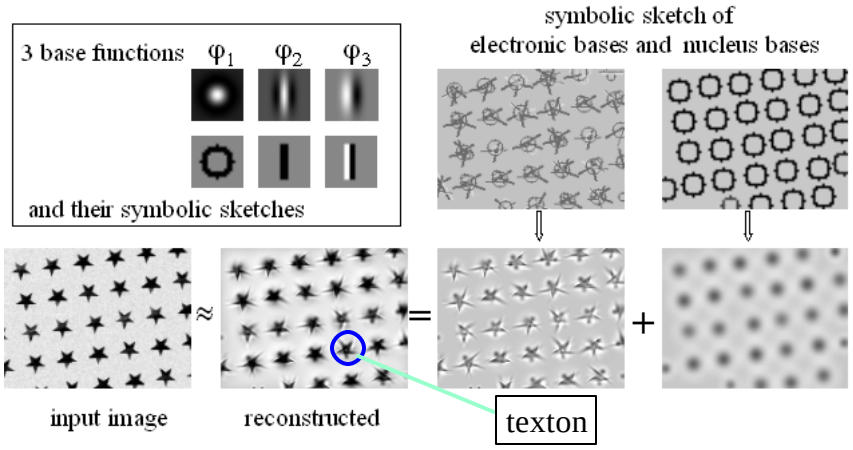
A Three Level Generative image model

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

A two level image model:



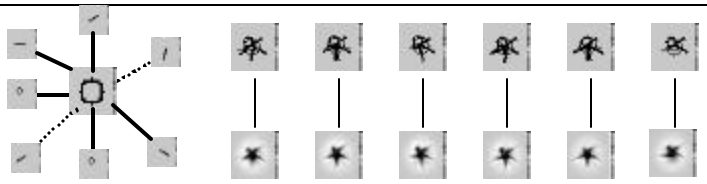
Example 1



With over-complete basis, the reconstruction often has 100 folds of dimension reduction.

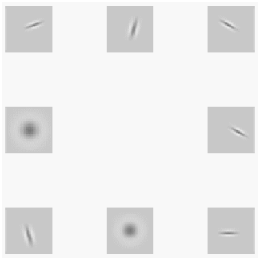
I: Geometric modeling

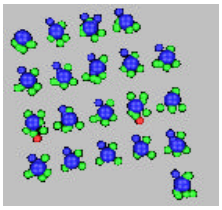
-- Texton with geometric variations



texton template

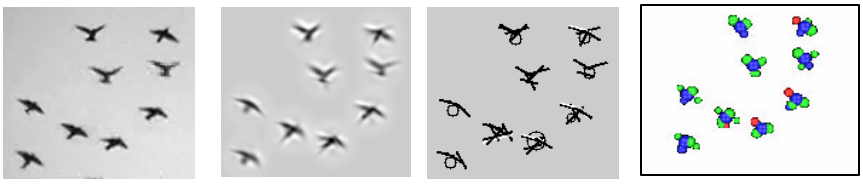
a sample of texton instances



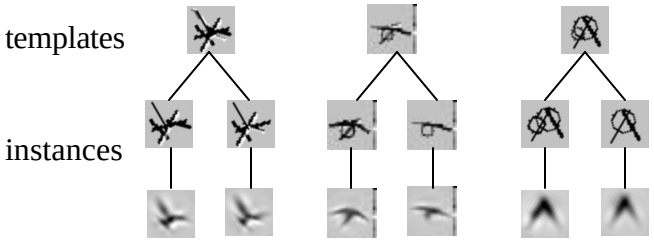


"atomic" model

Example 2



original image | reconstructed | symbolic sketch | atomic sketch

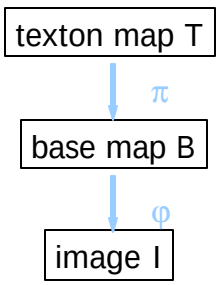


Learning with generative model

Image model:

$$p(I^{obs};?) = \iint p(I^{obs} \mid B; j)\ p(B \mid T; p)\ p(T; b)\, dB dT$$

$$? = (j, p, b)$$



θ is intrinsic to the ensemble of images

Learning by MLE:

$$?^* = \arg\min_{? \in O} KL(f \| p) = \arg\max_{? \in O} \log p(I^{obs};?)$$

Learning by Stochastic Gradient (like EM)

$$?^{*} = \arg\min_{?\in O} D(f \,\|\, p) = \arg\max_{?\in O} \log p(I^{obs};?) + e$$

$$\frac{\partial \log p(I^{obs};?)}{\partial ?} = \int \left(\frac{\partial \log p(I^{obs} \mid B; j)}{\partial j} + \frac{\partial \log p(B \mid T; p)}{\partial p} + \frac{\partial \log p(T; b)}{\partial ?}\right) \cdot p(B,T \mid I^{obs};?)\, dB\, dT$$

2. Regression, fitting.

3. Minimax entropy learning

1. Stochastic inference

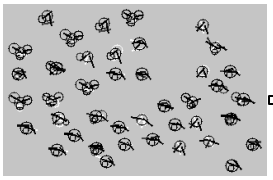
Algorithm by Markov chain Monte Carlo

The algorithm iterates two steps,

1. Given parameters $\theta=(\varphi, \pi, \beta)$,
   sample (B,T) from $p(B,T|I^{obs}; \theta)$, and compute the integral by Monte Carlo integration
   - Diffusion: diffuse bases, textons
   - Reversible jump: death/birth of bases, textons;
     switching of base types, texton types;
     assignment bases to textons
2. Given the sampled (B,T),
   update parameters $\theta=(\varphi,\pi, \beta)$ also be MCMC.

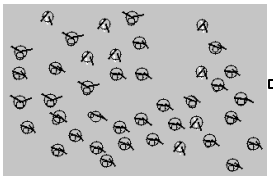
Example: bird texton

a). input image

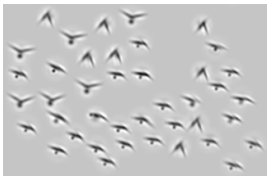


b). base map by bottom-up

c). reconstructed from b).

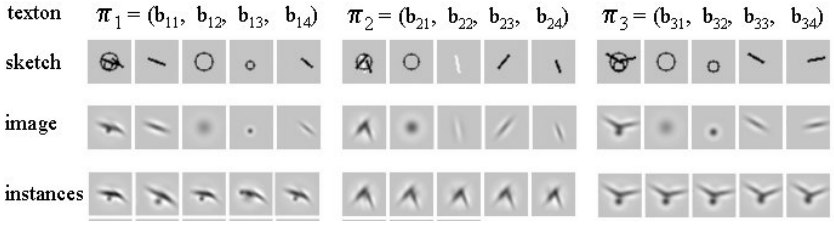


d). base map after learning

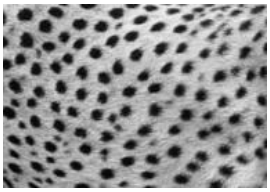


e). reconstructed from d).

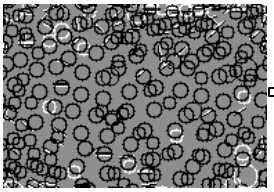
Bird Texton



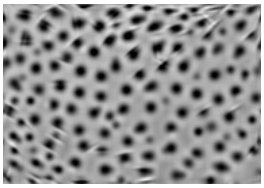
Example: cheetah dots



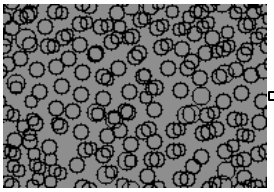
a). input image



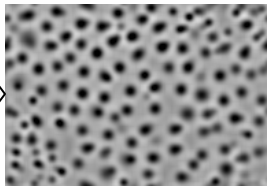
b). base map by bottom-up



c). reconstructed from b).



d). base map after learning



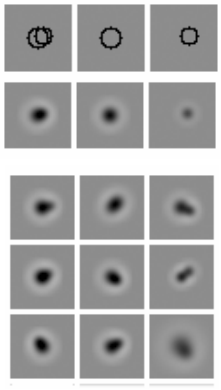
e). reconstructed from d).

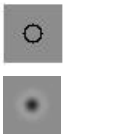
Texton – cheetah dots

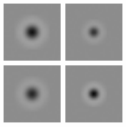
$\pi_1 = (b_{11}, b_{12})$ 9 π_1 instances π_2 4 π_2 instances

sketch

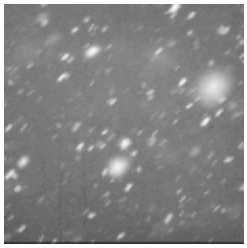
image







Textons in Motion



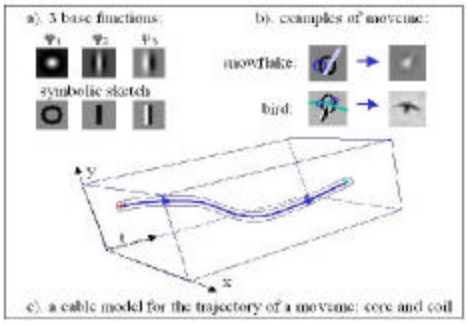


Fig. 1. A "cable model" for movemes.

$$C[t^b, t^e] = (\pi(t^b), \pi(t^b + 1), ..., \pi(t^e)).$$

Markov chain model

Then a Markov chain model is used for the texton trajectory (cable)

$$C[t^b, t^e] = (\pi(t^b), \pi(t^b + 1), ..., \pi(t^e)).$$

$$\pi(t) = A\pi(t-1) + B \cdot \pi(t-2) + C + DN(0, \sigma_0^2) \quad t \in [t^b + 2, t^e]$$
$$\pi(t^b + 1) = A'\pi(t^b) + C' + D\omega$$
$$(\pi(t^b), t^b) \sim P_B(\pi, \lambda), \quad (\pi(t^e), t^e - t^b) \sim P_D(\pi, \lambda).$$

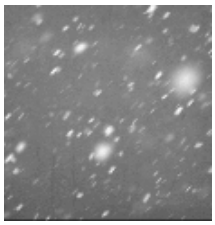
$$p(\pi(t)|\pi(t-1), \pi(t-2);\ A, B, C, D).$$

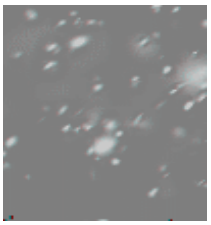
The total probability model includes ~300 parameters.

0. The deformable texton templates π *(~10-50 parameters)*
1. a conditional probability model *(~50 parameters)*
2. A birth probability for the sources of the elements P_B *(15 x 15 grid)*.
3. A death probability for the sinks of the elements P_D *(15 x 15 grid)*.

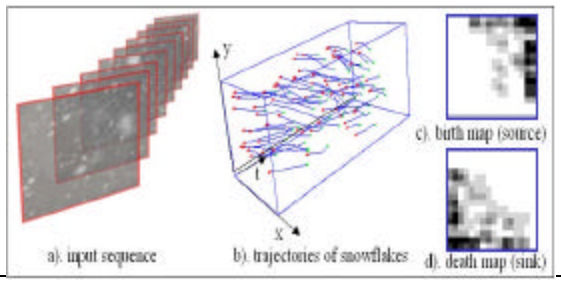
Moving textons : snow flakes

Observed Sequence



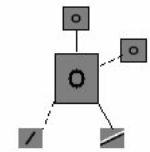


Synthesized Sequence

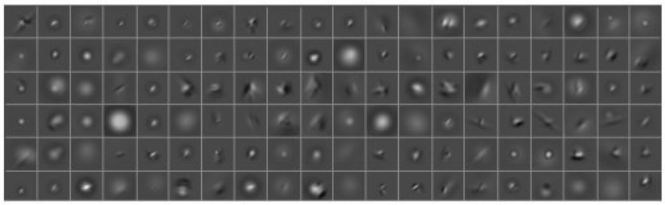


Moving textons: snow Flakes

For instance, a texton template π for the snowflake is shown below. Then 120 random snowflake instances are sampled randomly from π for a proof of variety.



a texton template π

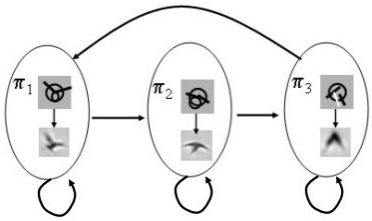


many texton instances randomly sampled from π

Moving textons: flying birds

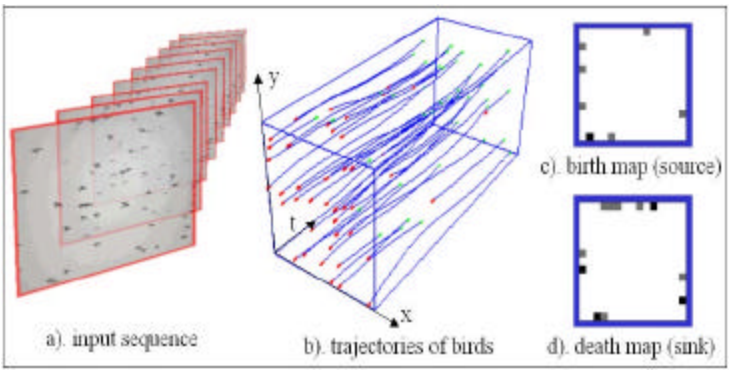
Observed Sequence

Synthesized Sequence



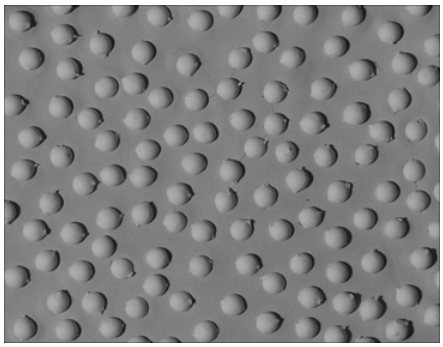
Experimental Results

- Flying Birds



Texton with lighting variations

Extend the generative model to 3D and lighting:
--- often accurate 3D depth is not recoverable or unnecessary

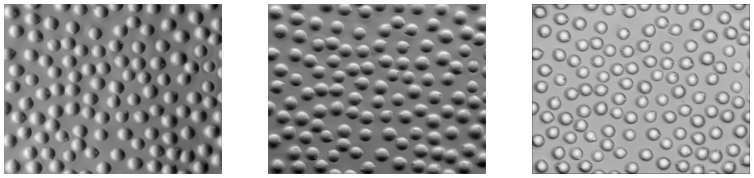


Some observed images at varying lighting conditions

Texton with lighting variations

By SVD, one can recover the three image bases b_1, b_2, b_3

$$I = s_1 b_1 + s_2 b_2 + s_3 b_3 \qquad (s_1, s_2, s_3) \text{ is the light.}$$

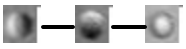


b_1 b_2 b_3

Each image base b is further decomposed into a linear sum of textons. Comparing with $I = \Sigma_i \alpha_i \psi_i$

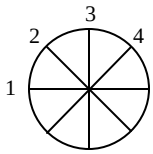
Texton with lighting variation

Each element is represented by a triplet of textons

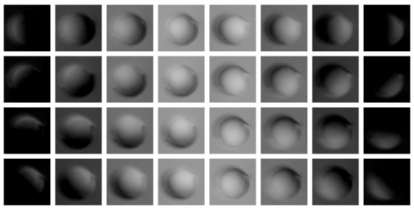
"lighton"? 

sketch 

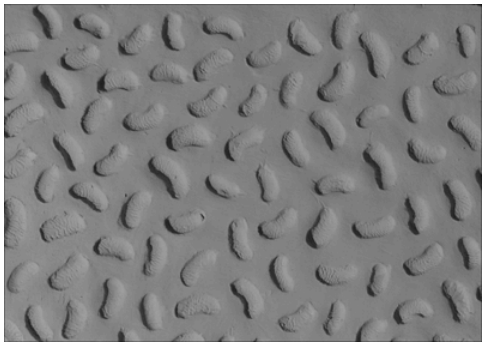
Sampling the 3D elements under varying lighting directions



4 lighting directions

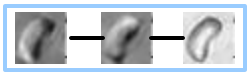


Example 2



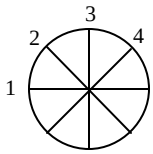
input images

Example 2

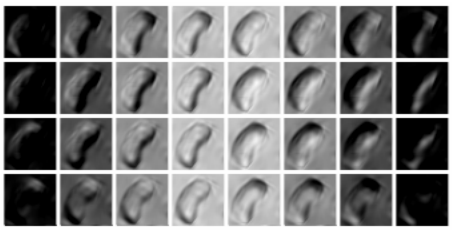
"Lighton" 

sketch 

Sampling the 3D elements under varying lighting directions



4 lighting directions

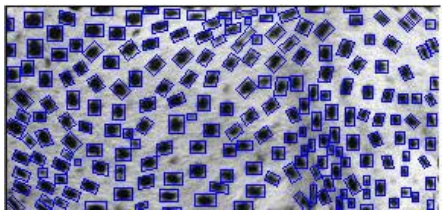


Summary: From Bases to Textons

Textons are atomic structures in natural images
Mathematically, textons form a vocabulary associated with a generative image model $p(I; \Theta)$, each texton is a triplet specified by 4 groups of parameters:

a texton = { $I=(s_1, s_2, s_3)' (b_1(\phi), b_2(\phi), b_3(\phi))$:
(x,y,σ,θ): similarity transform,
π: geometric deformations,
(A,B,C,D): dynamics,
lighting variations (s_1,s_2,s_3),
}

Modeling the Spatial Patterns of Textons



random graph process

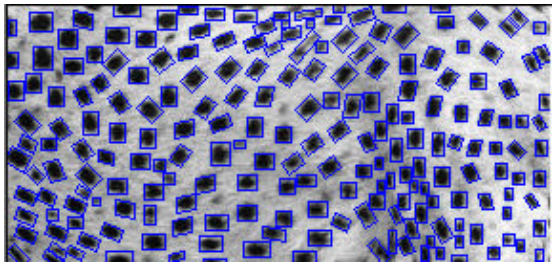
Treat a Texton Map as an Attributed Point Process

One layer of hidden variables: the texton map

? --- a "texton" (a min-template, a mother wavelet)

$$T = \{n, (x_i, y_i, ?_i, s_i, a_i), i = 1,2,..., n\}$$

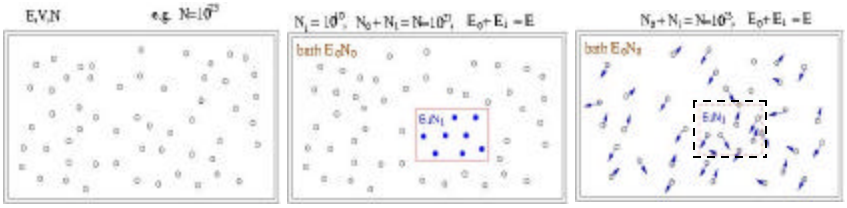
For texton #, translation, rotation, scale, contrast, …



Revisit Statistical Physics

Statistical physics studies macroscopic properties of systems with massive amount of elements

Three typical ensembles:



micro-canonical	canonical	grand-canonical
↓	↓	↓
Julesz	Gibbs	Gestalt

Markov Random Field Model for Texton Map

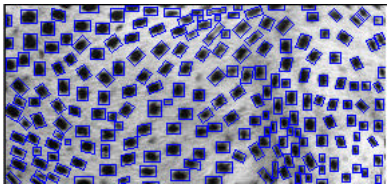
$$T = \{n, (x_i, y_i, ?_i, s_i, a_i), i = 1,2,..., n\}$$

The texton map is governed by a Gibbs distribution

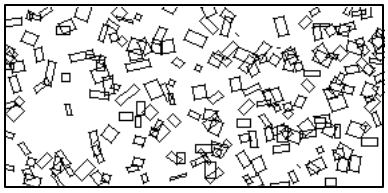
$$p(T; ß) = \frac{1}{Z} \exp\{-ß_o n - \sum_i <ß_i, h_i(T)>\}$$

Where β_o controls the density of the textons in unit area and h(T) captures some spatial statistics related to Gestalt psychology. The model can also be derived from maximum entropy principle. The parameters β are learned from MLE.

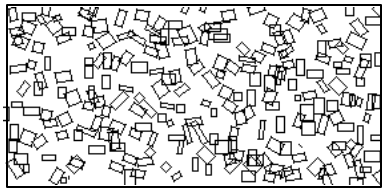
Experiment II: Modeling Texton Maps



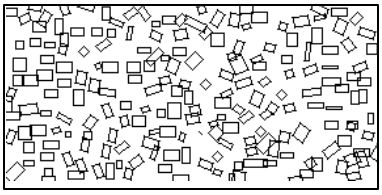
input



T=1



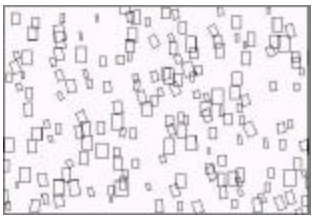
T=30



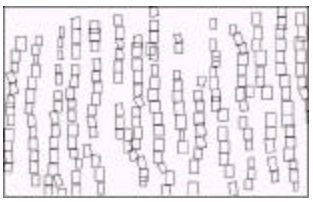
T=234

Experiment II: Modeling Texton Maps

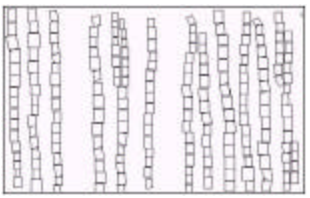
input



T=1

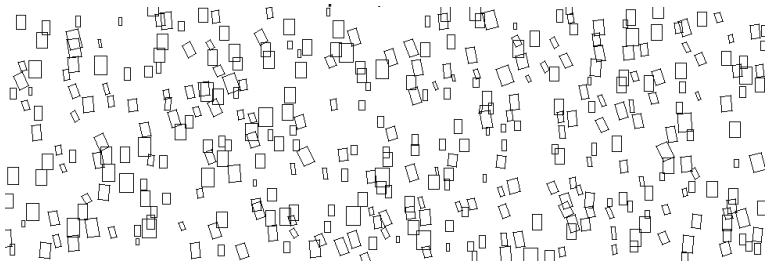


T=30

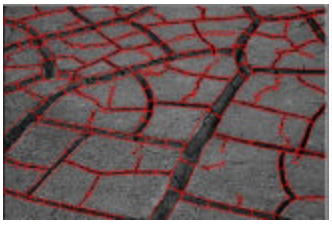


T=332

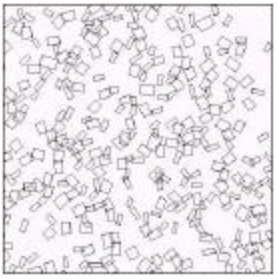
Modeling Texton Maps



Experiment II: Modeling Texton Maps



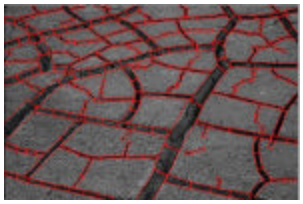
input



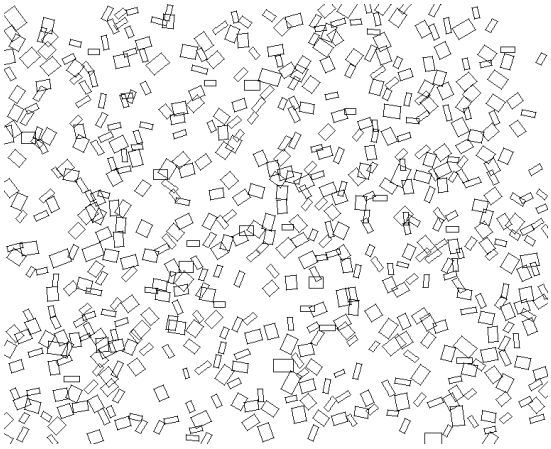
T=1

T=202

Modeling Texton Maps

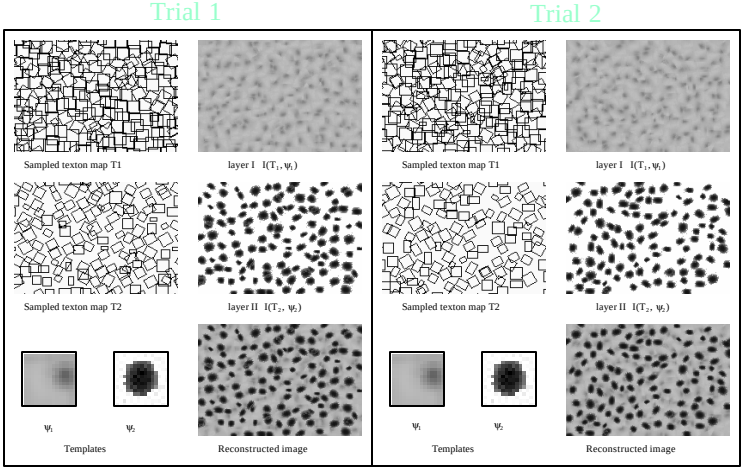


input



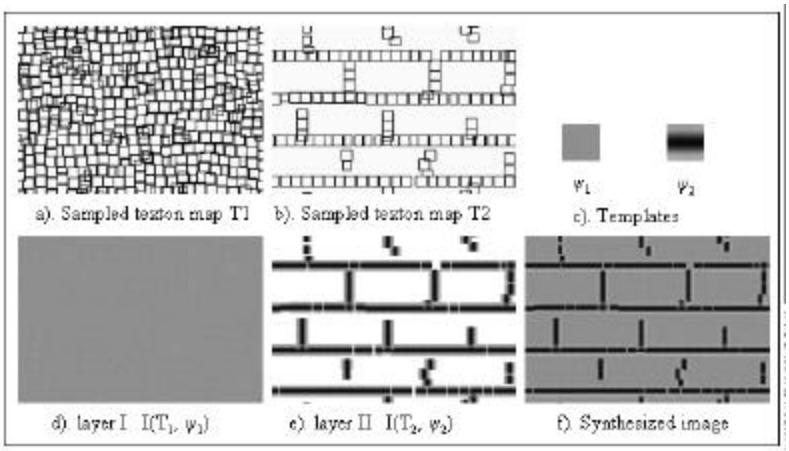
Experiment III: Integrated Model

Synthesizing cheetah skin patterns:



Experiment III: Integrated Model

Synthesizing brick patterns:



Integrating Two Visual Learning Paradigms

1. Descriptive methods:

   Markov random fields, minimax entropy learning, Julesz ensemble

   Characterize a pattern by statistical constraints on raw signals. Do not capture the semantic contents.

2. Generative methods:

   PCA, TCA (Frey and Jojic 99), Helmholtz machine (Hinton et al 95)
   wavelet / sparse coding (Mallat 93, Olshausen et al. 95, Donoho et al. 96-00)
   collage models (Lee and Mumford, 00)

   Characterize a pattern by generative bases: eigen-vectors, wavelets, Do not capture the spatial arrangement of the bases, and thus it can reconstruct but not synthesize realistic patterns.

Usually, a descriptive model has a trivial generative component and a generative model has a trivial descriptive components. The integration will lead to more advanced models and facilitate computation / inference.

Main References for This Lecture

Results presented in this lecture can be seen from the following papers.

1. S. C. Zhu, Y.N. Wu and D.B. Mumford, "Minimax Entropy Principle and Its Applications to Texture Modeling", *Neural Computation* Vol. 9, no 8, pp 1627-1660, Nov. 1997.

2. C.E Guo, S.C. Zhu, and Y. N. Wu, "Visual Learning by Integrating Descriptive and Generative Models", *ICCV01*. Long version to appear in *IJCV, 2003*.

3. S.C. Zhu, "Statistical Modeling and Conceptualization of Visual Patterns", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2003

4. S.C. Zhu, C. E. Guo, Y.N. Wu, and Y.Z. Wang, "What are Textons?", ECCV02.

5. Y.Z. Wang and S.C. Zhu, "A Generative Method for Textured Motion: Analysis and Synthesis", ECCV02.

Papers are available in Zhu's web page stat.ucla.edu/~sczhu