The image universe --- what are its structures?

Consider an image \( I \) with 256 x 256 pixels in 256 grey levels.

- The volume of image space \( |\Omega| = 2^{256 \times 256} = 10^{757,830} \)
- The volume of natural image ensemble \( |\Omega_p| \approx 2^{0.3 \times 256 \times 256} \approx 10^{5,718} \)
- The volume of images seen by humans \( \leq 10^{10} \times 10^{10} = 10^{20} \)

People believe that natural images reside in low dimensional manifolds.
This is only partially right.
1, Background on visual (appearance) manifolds

Image patches from a single object category are often found to form low dimensional manifolds.

e.g. ISOMAP, LLE:

But, people found that image patches of generic natural images do not follow this observation.

Looking at local, generic natural image statistics

Ruderman and Bialek 87, 94
Fields 87, 94
Zhu and Mumford 95-96
Chi and Geman 97-98
Huang and Mumford, 1999
Simoncelli etc 98-03

Here is an example of how real world data can be truly complex – non-Gaussian and highly kurtotic. This is an iso-density contour for a 3D histogram of log(range) images (2x2 patches minus their means) (Brown range image database, thesis of James Huang)
Patches in an object come from different subspaces


An example of low dimensional manifold:

A 3D element under varying lighting directions

Ref. S. Zhu/Xu/Guo/Wang, 2002-05 “What are textons?”
An example of low dimensional manifold:

By analogy: pictures of our universe

Star: low volume and high density --- like the explicit manifold for texton/primitive
Nebulous: high volume and low density --- like the implicit manifold for texture

Interchangeable concepts: entropy ~ dimension ~ log-volume

Summer School at Beijing, July, 2009
2. Pursuing Manifolds in the universe of image patches

\[ q = p_0 \rightarrow p_1 \rightarrow \cdots \rightarrow p_k \overset{\text{to}}{\sim} f \]

1. \( q = \text{unif()} \)
2. \( q = \delta() \)

Exchangeable concepts: a model \( p(I) \sim \) an image ensemble \( \Omega_p \sim \) a manifold \sim \) a cluster

Intuitive idea: a professor grading an exam

The full score (like dimension in our case) is 100. You have two ways:

For top students (high dimensional manifolds), you start from 100 and deduct points:

\[ 100 - 2 - 0 - 0 - 3 - 0 - 2 - 0 - 0 - 0 - 0 - 1 = 92 \]

For bottom students (low dimensional manifolds), you start from 0 and add points:

\[ 0 + 8 + 0 + 0 + 3 + 0 + 2 + 0 + 0 + 5 + 0 + 0 + 1 = 19 \]

In reality, suppose the exam is very long (just like the large image has >1M pixels), a student may have mixed performance, e.g. doing excellent in the 1st half and doing poorly in the 2nd half. Thus a most effective way is to use the two methods for different sections of the exam:

\[ (50 - 2 - 0 - 0 - 3 - 0) + (0 + 5 + 3 + 0 + 0 + 2) = 45 + 10 = 55 \]

In fact, most of the object categories are middle entropy manifolds and have mixed structures.
Manifold pursuit in the image universe

In a simple case: \( f \) is a Gaussian distribution

![Diagram showing eigen-values and examples]

Manifold pursuit by information projection

Given only positive examples from a class \( c \)

\[
\Omega_c^+ = \{ \mathbf{x}^{+k} \mid k = 1, 2, \ldots, M^+ \} \sim f(\mathbf{I})
\]

We pursue a series of models \( p \) to approach a underlying “true” probability \( f \)

\[
q = p_0 \rightarrow p_1 \rightarrow \cdots \rightarrow p_k \rightarrow f
\]

At each step, we augment the current model \( p \) to a new model \( p^* \)

\[
h^*_+ = \arg \max_p KL(f \mid p) - KL(f \mid p_+), \quad p^*_+ = \arg \max_p KL(p_+ \mid p)
\]

Subject to a projection constraint:

\[
E_{p^*_+}[h_+ (\mathbf{I})] = E_f[h_+ (\mathbf{I})] \approx \bar{h}_+
\]

\( h_+ (\mathbf{I}) \) is a feature statistics of image \( \mathbf{I} \)
Manifold pursuit by information projection

Solving the constrained optimization problem leads to the Euler-Lagrange equation:

\[
P^* = \arg \min_p \int p_q(t) \log \frac{p_q(t)}{p(t)} \, dt + \lambda_1 \int p_q(t) h_1(t) \, dt - h_0 + \lambda_2 \int p_q(t) \, dt - 1
\]

where

\[
p_k(l; \Theta_k) = \frac{1}{z_{k-1}} p_{k-1}(l; \Theta_{k-1}) e^{-\lambda_k h_k(l)} \exp \left\{ -\sum_{i=1}^{k} \lambda_i h_i(l) \right\}
\]

For \( q \) being a uniform distribution, we have

\[
q(l) = \frac{1}{z_0}
\]

Information projection

DellaPietra, DellaPietra, Lafferty, 97
Zhu, Wu, Mumford, 97

\[
p_0 = q \quad \Omega_1 = \{ p : E_p[h_1(l)] = E_f[h_1(l)] \}
\]

\[
p_1 \quad \Omega_2 = \{ p : E_p[h_2(l)] = E_f[h_2(l)] \}
\]

\[
KL(f \mid p) = KL(f \mid p^*) + KL(p^* \mid p)
\]

So the KL-divergence decreases monotonically.
A Maximin Learning Principle

A max-step: choosing a distinct feature and statistics

\[ h^*_+ = \arg \max_{h_+} KL(p_+ \mid p) \]

A min-step: given the selected feature constraint, computing the parameter

\[ \lambda^*_+ = \arg \min_{\lambda_+} KL(p_+ \mid p) \]

3. Two types of pure and atomic manifolds

**implicit** manifold

\[ \Omega = \{ I: h(I) = h_0 \} \]

\( h(I) \) is some image feature/statistics

**explicit** manifold

\[ \Omega = \{ I: I = g(w; \Delta) \} \]

g is a generation function,
w is intrinsic dimension
\( \Delta \) is a dictionary
Case 1: A texture pattern is an "implicit manifold"

A texture $= \Omega(h) = \{ 1: h_i(\Omega) = h_{c,i}, i = 1,2,\ldots,K \}$

$H_i$ are histograms of Gabor filters, i.e. marginal distributions of $f(I)$

(Zhu, Wu, Mumford 97, 99, 00)

Pursuing texture manifolds
More examples of the texture manifold (implicit)

Observed

MCMC sample

This is originally from statistical physics! Gibbs 1902

Statistical physics studies macroscopic properties of systems that consist of massive elements with microscopic interactions.

\[ N = 10^{33} \]

A state of the system is specified by the position of the \( N \) elements \( X^N \) and their momenta \( p^N \)

\[ S = (x^N, p^N) \]

But we only care about some global properties

Energy \( E \), Volume \( V \), Pressure, ...

Micro-canonical Ensemble

\[ \Omega(N, E, V) = \{ s : h(S) = (N, E, V) \} \]
**Equivalence of Julesz ensemble and FRAME / MRF models**

Theorem 1
For a very large image from the Julesz ensemble $I \sim f(I; h_z)$ any local patch of the image $I_\Lambda$ given its neighborhood follows a conditional distribution specified by a FRAME model $p(I_\Lambda | I_{\text{EA}} : \beta)$.

Theorem 2
As the image lattice goes to infinity, $f(I; h_z)$ is the limit of the FRAME model $p(I_\Lambda | I_{\text{EA}} : \beta)$, in the absence of phase transition.

$$p(I_\Lambda | I_{\text{EA}} : \beta) = \frac{1}{Z(\beta)} \exp\left\{ - \sum_{j=1}^J \beta_j I_\Lambda(h_j) \right\}$$

**Case 2: Learning active basis as deformable template**

A basis is an image space spanned by a number of vectors (e.g. Gabor/primitives)

$$B = (B_1, B_2, \ldots, B_k)$$

A car $= \Omega = \{I: I = \sum_i \gamma_i B_i \}$

A car template

(Gabor elements represented by bar)

An incoming car image:

With slight modification, this model can handle multi-views

Ref: Wu, Si. Gong, Zhu, ICCV 08 2008
Deformed to fit many car instances

A car template

Summer School at Beijing, July, 2009
Pursuing the active basis model (explicit manifold)

\( q(I) \): background distribution
(all natural images)

\( p(I) \): pursued model to approximate
the true distribution.

![Diagram](Image)

A running example

A car template consisting of
48 Gabor elements

Car instances
Experiment: learning and clustering

Learning active basis

EM clustering

Experiment: learning and detection

Y.N. Wu et al ICCV'07, IJCV'09
vs: Viola, Jones, 04
Template detection experiment

Matching $\Omega_p$ to $\Omega_f$: Push and Pull

(a) push and compress

(b) pull and expand
Summary: a second look at the space of image patches

4. Relations to the literature: psychophysics

(1) textures vs textons  (Julesz, 60-70s)

textons
Textons vs. textures

5, Frequency plot of the ex/implicit manifolds in natural images

implicit texture clusters (blue), explicit primitive clusters (pink).

Summer School at Beijing, July, 2009
Clustering in video

Summer School at Beijing, July, 2009

Examples in video

<table>
<thead>
<tr>
<th>explicit</th>
<th>implicit</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Explicit Examples" /></td>
<td><img src="image2.png" alt="Implicit Examples" /></td>
</tr>
</tbody>
</table>
Textons in motion

Observed Sequence

Synthesized Sequence

Ref. Y.Z. Wang, 2003

Summer School at Beijing, July, 2009

6. Primal sketch: integrating the two regimes

org image

sketching pursuit process

sketches

syn image

synthesized textures

sketch image

(Guo, Zhu, Wu, 2003-05)
manifolds of image primitives

Learned texton/primitive dictionary with some landmarks that transform and warp the patches

Primal Sketch is a two-level MRF model

Summer School at Beijing, July, 2009
Primal sketch example

input image

synthesized image

sketching pursuit process

sketches

Summer School at Beijing, July, 2009

Primal sketch example

original image

synthesized image

sketching pursuit process

Summer School at Beijing, July, 2009
7. Pursuing composite manifolds in the middle entropy regime

Learning Hybrid Image Templates

Learning object templates by manifold pursuit

The two types of models compete in learning the templates
Examples of the learned hybrid image templates

Mixing the implicit and explicit manifolds

Some examples of learn object templates

laptop  butterfly  pizza

cat head  tiger head  bear head
The epsilon-ball interpretation

Comparing with the HoG Representation

Dalal and Triggs, 2005; Felzenszwalb, Girshick, McAllester and Ramanan, 2007-09
8, Information scaling leads to manifold transitions!

Scaling (zoom-out) increases the image entropy (dimensions)

Wu, Zhu, Guo, 04,07

Information scaling leads to manifold transitions!

scale 1  scale 2  scale 3

scale 4  scale 6  scale 8

Ref:
Wu, Zhu, Guo, 2004-07, "From Information Scaling to Regimes of Statistical Models"
Coding efficiency and number of clusters over scales

A wide spectrum of categories from low to high entropy

<table>
<thead>
<tr>
<th>Edge</th>
<th>Bar</th>
<th>Two Parallel Lines</th>
<th>Cat</th>
<th>Dog</th>
<th>Lion</th>
<th>Tiger</th>
<th>Fur</th>
<th>Carpet</th>
<th>Grass</th>
<th>Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="edge.png" alt="Image" /></td>
<td><img src="bar.png" alt="Image" /></td>
<td><img src="two_parallel_lines.png" alt="Image" /></td>
<td><img src="cat.png" alt="Image" /></td>
<td><img src="dog.png" alt="Image" /></td>
<td><img src="lion.png" alt="Image" /></td>
<td><img src="tiger.png" alt="Image" /></td>
<td><img src="fur.png" alt="Image" /></td>
<td><img src="carpet.png" alt="Image" /></td>
<td><img src="grass.png" alt="Image" /></td>
<td><img src="noise.png" alt="Image" /></td>
</tr>
</tbody>
</table>

Entropy ~ Dimension ~ Log volume( manifold )
Transition of the manifolds through info. scaling

How are these manifolds related to each other?

perceptual scale space theory (Wang and Zhu 2005)

Summary on the representation

2 pure atomic image spaces

Texture

Primal Sketch

Texton

Scaling → Transition

where

2.1D Sketch 2.5D Sketch 3D Sketch

Graphlets → Parts → Objects → Scenes

what
Lecture 1.B

Stochastic Image Grammar in And-Or Graph
--- Modeling and Learning Object Categories

Song-Chun Zhu

University of California, Los Angeles, USA
Lotus Hill Research Institute, China


1, Representing Objects by Reconfigurable Graphs

~3,000 basic object categories.

Objects have large within-category variations in configurations
   Vehicles --- sedan, hunchback, van, truck, SUV, ...
   Clothes --- jacket, T-shirt, sweater, ....
   Furniture --- desk, chair, dresser, ...

Scenes have more flexible configurations
   a living room,
   an office,
   a street, ...

Summer School at Beijing, July, 2009
How do we define an object category?

Each object category is a set of “re-configurable” graphs that satisfy certain regulations in its structures and appearance.

This is actually a grammar in formal formulation.

It comes in many other names:
- Compositional models,
- Hierarchical models,
- Contextual models,
- ...

Formulation of Grammar by Chomsky 1957

A grammar is a 4-tuple: \( \mathcal{G} = (V_N, V_T, R, S) \)

<table>
<thead>
<tr>
<th>Type</th>
<th>Grammar</th>
<th>Production Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type - 0</td>
<td>Unrestricted</td>
<td>( \alpha \rightarrow \beta )</td>
</tr>
<tr>
<td>Type - 1</td>
<td>Context-sensitive</td>
<td>( \alpha A \beta \rightarrow \alpha \gamma \beta )</td>
</tr>
<tr>
<td>Type - 2</td>
<td>Context-free</td>
<td>( A \rightarrow \gamma )</td>
</tr>
<tr>
<td>Type - 3</td>
<td>Regular</td>
<td>( A \rightarrow \alpha ) ( A \rightarrow \alpha B )</td>
</tr>
</tbody>
</table>

The language of a grammar is the set of all valid sentences

\[ L(\mathcal{G}) = \{ \omega : S \xrightarrow{R^*} \omega, \omega \in V_T^* \} \]

\[ S \xrightarrow{\gamma_1, \gamma_2, \ldots, \gamma_n(\omega)} \omega \]
2. And-Or tree for Production rules

In a grammar, each non-terminal node has a number of alternative ways for expanding, and thus can be represented by an And-Or tree

\[ A ::= aB | a | aBc \]

A special property of image grammar is that any node can terminate or "ground" immediately.

Representing grammar by And-Or graph

A grammar production rule:

\[ A \rightarrow ab | cc \]

The language of a node A is the set of all valid configurations
The expressive power of and-or graphs

Consider a 2-layer And-Or tree with branching factor 3.

And-Or graph was used in the 1980s by Judea Pearl for heuristic search in AI. For example the 12-counterfeit coin problem.

Total: $1+3+9+27 = 30$ nodes with $81$ leaves

$(2 \cdot 3)^3 = 3^2 = 531,441$ configurations

Definition: And-Or graph, parse graphs, and configurations

Each category is conceptualized to a grammar whose language defines a set or *equivalence class* for all the valid configurations of the each category.
An example: the clock category

A parse graph of a bike
A relation is like a non-linear filter

Some examples

<table>
<thead>
<tr>
<th>Pos</th>
<th>Scale</th>
<th>Orientation</th>
<th>Contained</th>
<th>Hinged</th>
<th>Attached</th>
<th>Beating</th>
<th>Convertible</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
</tr>
<tr>
<td>Low Level Relationships</td>
<td>High Level Relationships</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A binary relation is set of links between selected nodes.

It is applied to selected sites and returns a value (scalar or binary). Suppose $A$ is a vector of attributes for all nodes

$A = (a_1, a_2, ..., a_n)$

$r_{ij} = f(a_i, a_j)$

3, Pursuing a probability model on the And-Or graph

Denote:

$G$ ---- a parse graph,

$U(G)$ ---- the set of Or-nodes in $G$,

$V(G)$ ---- the set of the And-nodes + leaf nodes in $G$

$R(G)$ ---- the set of relational links between nodes in $G$.

The probability model is defined as

$p(G; \Delta, R, \theta) = \frac{1}{Z} \exp \left\{ - \sum_{u \in U(G)} \lambda(u) - \sum_{v \in V(G)} \psi(v) - \sum_{r \in R(G)} \psi(r_{ij}) \right\}$

The first term alone stands for a SCFG. The second and third terms are Markov potentials.
Pursuing the relations by info. projection

Sampling Clock: we keep evolving $O(100)$ samples

Poway et al, 2006-07
**Learning and sampling a bike model**

Examples of sampling bicycles: Computer can dream

Dreaming is a process of learning

Poway et al, 2006-07
Examples from sampling cars
it is less satisfactory, as 3D perspectives are not accounted.

Iterative learning to match the statistics (histogram)

Results of the learning procedure.

(a) Histograms for four pairwise relationships at different iterations. The last iteration matches the observed histogram quite closely.

(b) The KL divergence between the current and target model as the relationship pursuit is performed.
Top-down prediction by sampling the missing part

<table>
<thead>
<tr>
<th>Stage 1</th>
<th>Stage 2</th>
<th>Stage 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>random prediction</td>
<td>correct scale</td>
<td>more relations</td>
</tr>
</tbody>
</table>

The blue parts are predicted by the learned models at various learning stages

Learning from a small training set & generalization by sampling

In our recognition experiment with 33 categories, each category has 50-60 samples, the sampling process improves the average recognition rate from 66% to 81%.
What is the smallest sample set for training?

Coverage results for 6 categories. We only need a small fraction of the training set to maximally cover the testing set.

4, A large scale human annotation project at Lotus Hill
An example: parse graph of a cat

An example: clip for surveillance video
LHI dataset: 1.8M images parsed, 1000s AoG nodes

5, Example: Representing An event by AoG
A parse graph for event instance

Video parsing by And-Or Graph
Examples of automated text generation

| Land_vehicle_359 approaches intersection_0 along road_0 at 57:27. It stops at 57:29. |
| Land_vehicle_360 approaches intersection_0 along road_3 at 57:31. |
| Land_vehicle_360 moves at an above-than-normal average speed of 26.5 mph in zone_4 (approach of road_3 to intersection_0) at 57:32. It enters intersection_0 at 57:32; it leaves intersection_0 at 57:34. There is a possible failure-to-yield violation between 57:27 to 57:36 by Land_vehicle_360. |
| Land_vehicle_359 enters intersection_0 at 57:35. It turns right at 57:39. It leaves intersection_0 at 57:36. It exits the scene at the top-left of the image at 57:38. |

Ref: Benjamin Yao et al “From image parsing to text generation”, 2009.
In collaboration with Mun Wai Lee at ObjectVideo Inc.

Summary on the representation

2 pure atomic image spaces: Texture, Texton

- Primal Sketch
- Scaling – Transition

2.1D Sketch → 2.5D Sketch → 3D Sketch

Graphlets → Parts → Objects → Scenes

where

what

Summer School at Beijing, July, 2009