Lecture 1.A

Pursuing Manifolds in the Universe of Images

--- Texture, Texton, Primal Sketch, and Object Template

Song-Chun Zhu

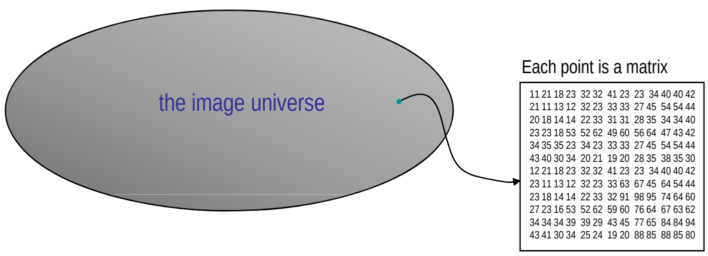
University of California, Los Angeles, USA
Lotus Hill Research Institute, China

Ref: S. C. Zhu, et al "Learning Explicit and Implicit Visual Manifolds by Information Projection", 2009.

The image universe --- what are its structures?

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

$$\text{The volume of image space } |\Omega_{\mathrm{I}}| = 2^{8\times256\times256} = 10^{157,830}$$
$$\text{The volume of natural image ensemble } |\Omega_{\mathrm{f}}| \cong 2^{0.3\times256\times256} \cong 10^{5,718}$$
$$\text{The volume of images seen by humans } \leq 10^{10}\times10^{10} = 10^{20}$$



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:
Saul and Roweis, 2000.

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

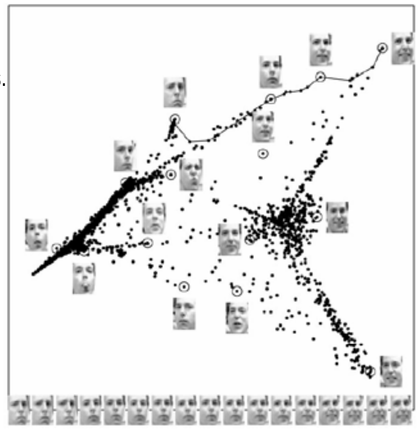
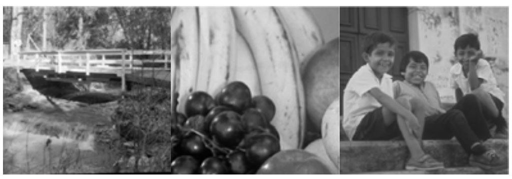
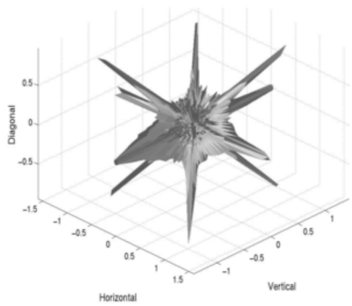


Fig. 3. Images of faces (*11*) mapped into the embedding space described by the first two coordinates of LLE. Representative faces are shown next to circled points in different parts of the space. The bottom images correspond to points along the top-right path (linked by solid line), illustrating one particular mode of variability in pose and expression.

Looking at local, generic natural image statistics

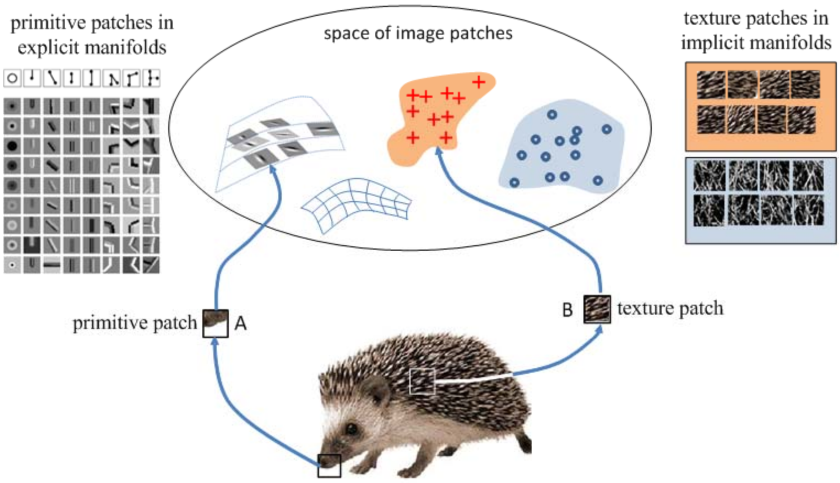
....
Ruderman and Bialek 87, 94
Fields 87, 94
Zhu and Mumford 95-96
Chi and Geman 97-98
Huang and Mumford, 1999
Simoncelli etc 98-03
.....





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

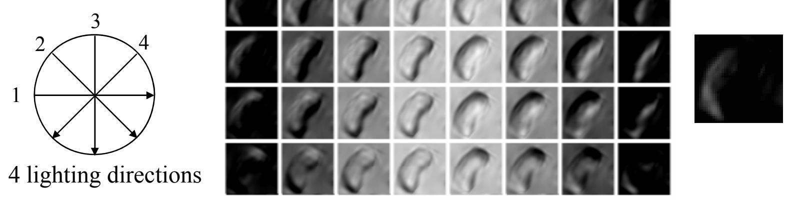


Ref: Z.Z. Si, H. Gong, Y.N. Wu, S.C. Zhu, "Learning Hybrid Image Templates", 2008-09.

An example of low dimensional manifold:

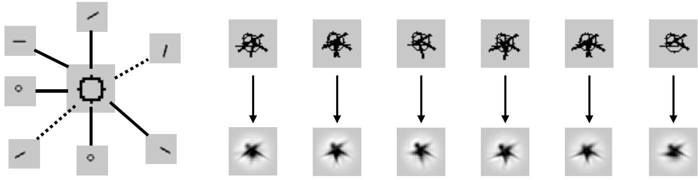
texton / primitive

A 3D element under varying lighting directions



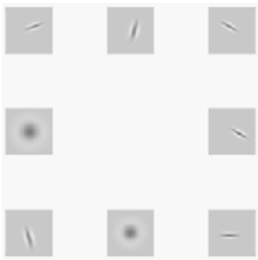
Ref. S. Zhu/Xu/Guo/Wang, 2002-05 "What are tetxons?"

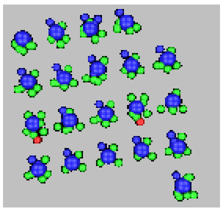
An example of low dimensional manifold:



texton template

a sample of texton instances





Guo/Zhu 2002-05

"atomic" model

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

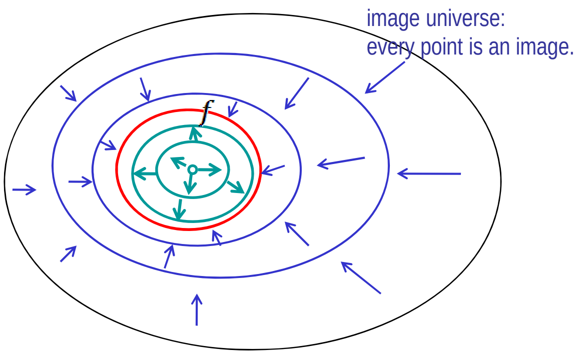
2, Pursuing Manifolds in the universe of image patches

f : target distribution; *p*: our model; *q*: initial model

$$q = p_0 \rightarrow p_1 \rightarrow \cdots \rightarrow p_k \quad to \quad f$$

1, $q = unif()$

2, $q = \delta()$



Exchangeable concepts: a model $p(I)$ ~ an image ensemble Ω_p ~ a manifold ~ 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 – 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

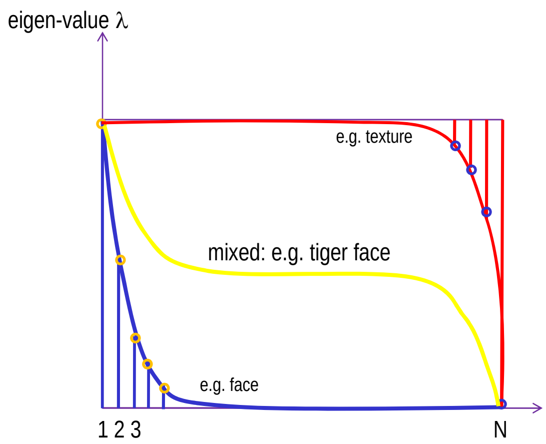
In reality, suppose the exam is very long (just like the large image has >1M pixels), a student may have mixed performance, e.g. doing excellent in the 1st half and doing poorly in the 2nd half. Thus a most effective way is to use the two methods for different sections of the exam.

(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



Manifold pursuit by information projection

Given only positive examples from a class c

$$\Omega_c^+ = \{\mathrm{I}_i^{obs};\quad i = 1,2,\dots,M^+\} \sim f(\mathrm{I})$$

We pursue a series of models *p* to approach a underlying "true" probability *f*

$$q = p_0 \to p_1 \to \cdots \to p_k \quad to \quad f$$

At each step, we augment the current model *p* to a new model p_+

$$h_+^* = \arg\max\; KL(f \mid p) - KL(f \mid p_+)$$
$$= \arg\max KL\;(p_+ \mid p)$$

Subject to a projection constraint:

$$E_{p_+}[\mathrm{h}_+(\mathrm{I})] = E_f[\mathrm{h}_+(\mathrm{I})] \cong \bar{h}_+$$

$\mathrm{h}_+(\mathrm{I})$ is a feature statistics of image I

Manifold pursuit by information projection

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

$$p_+^* = \arg\min \int p_+(I) \log \frac{p_+(I)}{p(I)}\, dI + \lambda_+ [\int p_+(I) h_+(I) dI - \bar{h}_+] + \lambda_o \int p_+(I) dI - 1]$$

$$p_k(I;\Theta_k) = \frac{1}{Z_{+,k}} p_{k-1}(I;\Theta_{k-1}) e^{-\lambda_k h_k(I)}$$

$$= \frac{1}{Z_k} q(I) \exp\{-\sum_{i=1}^{k} \lambda_i h_i(I)\}$$

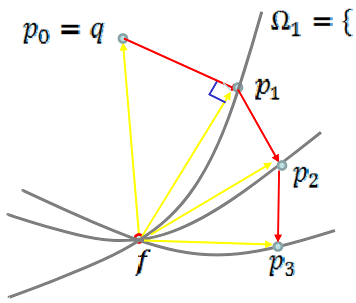
where

$Z_k = Z_{+,1} \cdot Z_{+,2} \cdots Z_{+,k}$ $\Theta_k = (\lambda_1, \lambda_2, \cdots \lambda_k)$

For q being a uniform distribution, we have $q(I) = \frac{1}{Z_o}$

Information projection

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



$\Omega_1 = \{p: E_p[h_1(I)] = E_f[h_1(I)]\}$

$\Omega_2 = \{p: E_p[h_2(I)] = E_f[h_2(I)]\}$

$$KL(f \mid p) = KL(f \mid p_+) + KL(p_+ \mid p)$$

So the KL-divergence decreases monotonically.

A Maximin Learning Principle

A **max-step**: choosing a distinct feature and statistics

$$h_+^* = \arg\max KL\,(p_+ \mid p)$$

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

$$\lambda_+^* = \arg\min KL(p_+ \mid p)$$

3, Two types of pure and atomic manifolds

implicit manifold

$$\Omega = \{\mathrm{I}:\ \mathrm{h(I)} = \mathrm{h_o}\}$$

h(I) is some image feature/statistics

explicit manifold

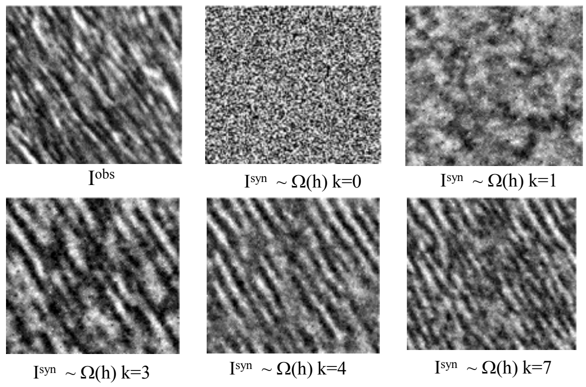
$$\Omega = \{\mathrm{I}:\ \mathrm{I} = \mathrm{g(w;\Delta)}\}$$

g is a generation function,
w is intrinsic dimension
Δ is a dictionary

Case 1: A texture pattern is an “implicit manifold”

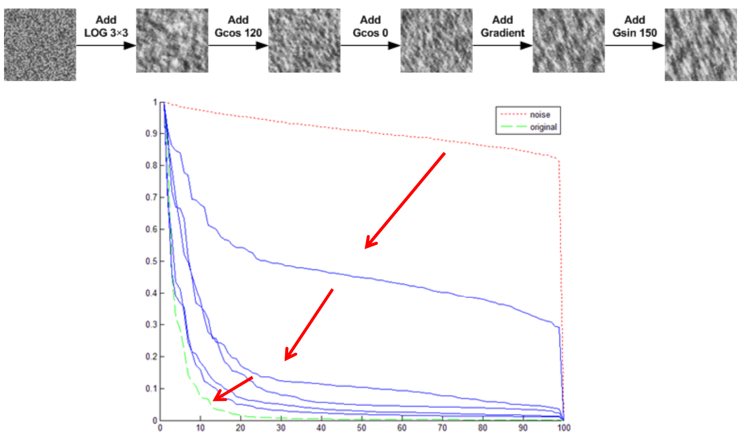
$$\text{a texture} = \Omega(h_c) = \{ I: \; h_i(I) = h_{c,i} \;\; , i = 1,2,\ldots,K \}$$

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



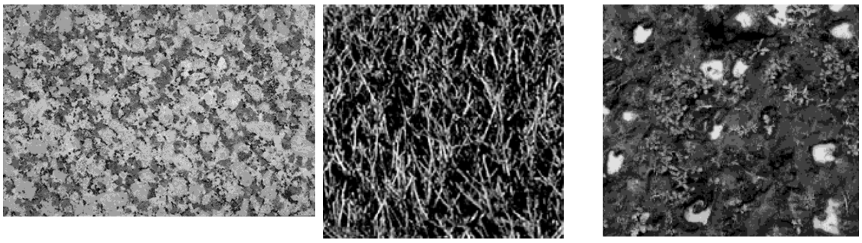
(Zhu,Wu, Mumford 97,99,00)

Pursuing texture manifolds

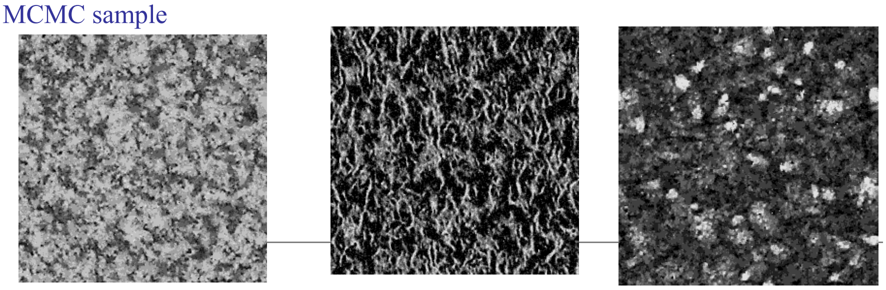


More examples of the texture manifold (implicit)

Observed



MCMC sample

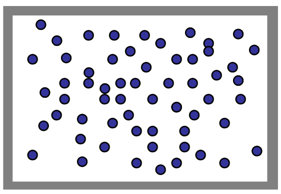


This is originally from statistical physics ! Gibbs 1902

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

e.g.: a tank of insulated gas or ferro-magnetic material

$N = 10^{23}$



Micro-canonical Ensemble

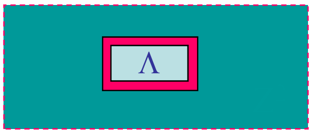
A state of the system is specified by the position of the N elements X^N and their momenta p^N

$$S = (x^N,\ p^N)$$

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



Zhu, Wu, Mumford, 1997
Wu and Zhu, 1999

Theorem 1

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

Theorem 2

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

$$\boxed{p(\mathrm{I}_\Lambda \mid \mathrm{I}_{\partial\Lambda} ; \beta) = \frac{1}{Z(\beta)} \exp\{-\sum_{j=1}^{k} \beta_j h_j(\mathrm{I}_\Lambda \mid \mathrm{I}_{\partial\Lambda})\}}$$

Case 2: Learning active basis as deformable template

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

$$\mathrm{B} = (\mathrm{B}_1, \mathrm{B}_2, \ldots, \mathrm{B}_k)$$

$$\text{A car} = \Omega = \{\mathrm{I}: \ \mathrm{I} = \sum_i \gamma_i \mathrm{B}_{i,\delta}\}$$

A car template

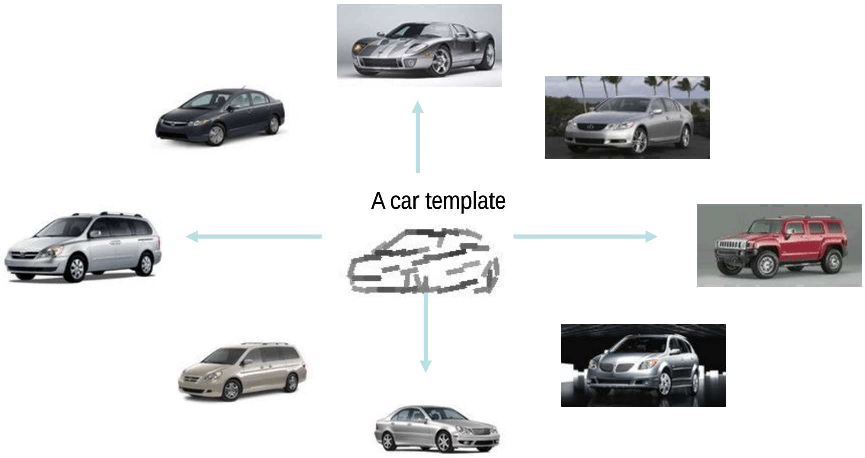
(Gabor elements represented by bar)

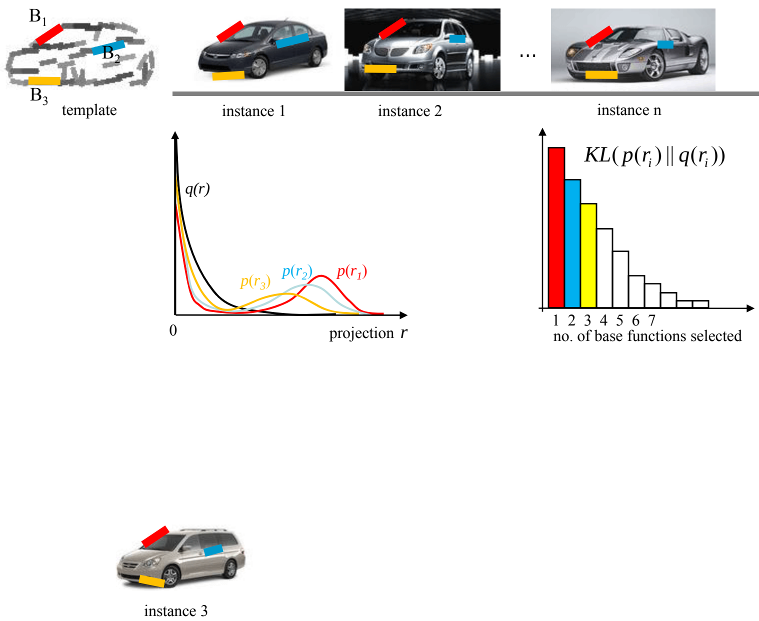
An incoming car image:

With slight modification, this model can handle multi-views

Ref: Wu, Si. Gong, Zhu, ICCV 08 2008

Deformed to fit many car instances

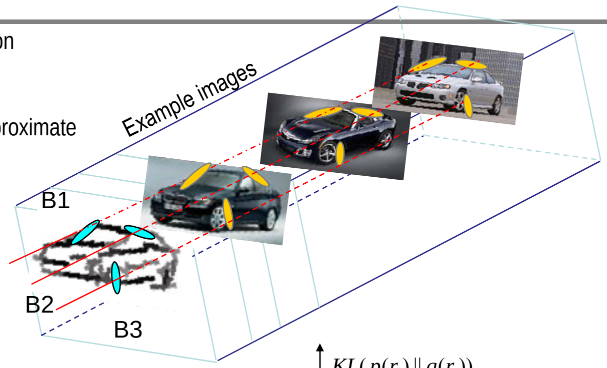


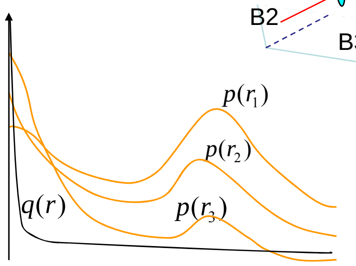


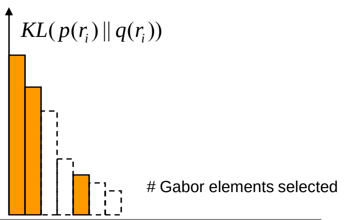
Pursuing the active basis model (explicit manifold)

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

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





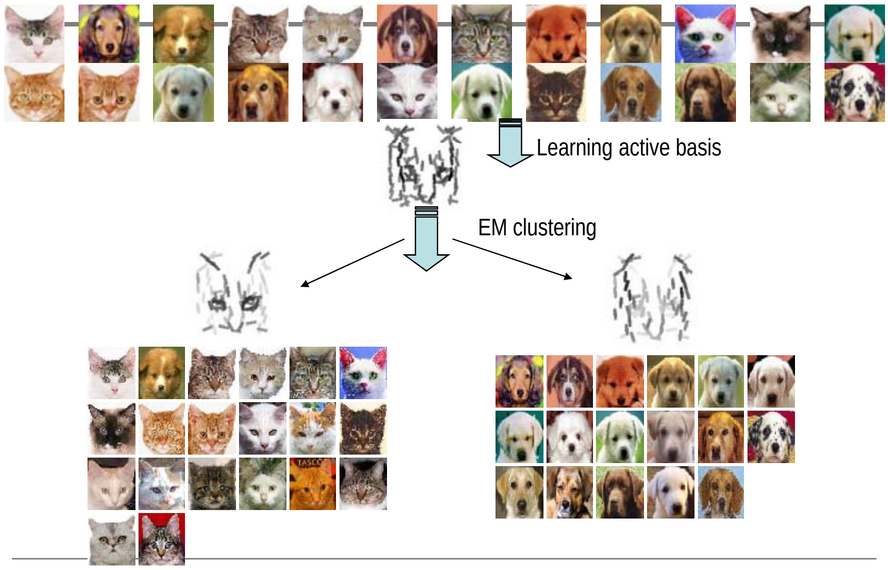


A running example

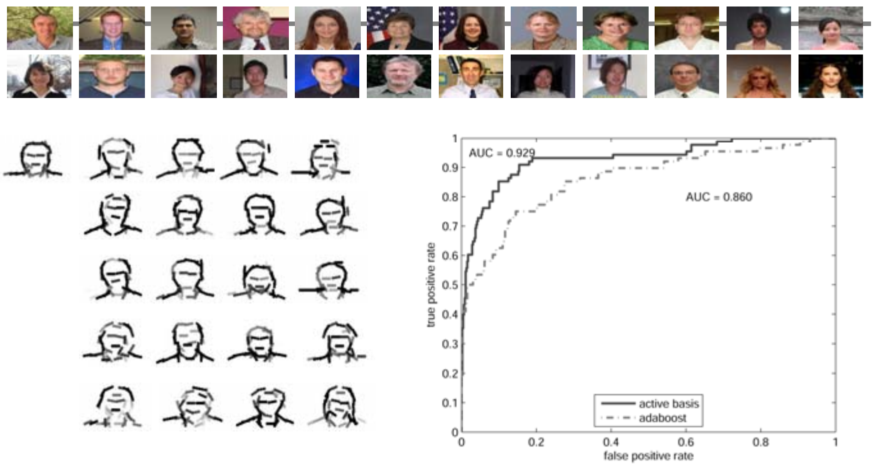
A car template consisting of
48 Gabor elements

Car instances

Experiment : learning and clustering



Experiment : learning and detection



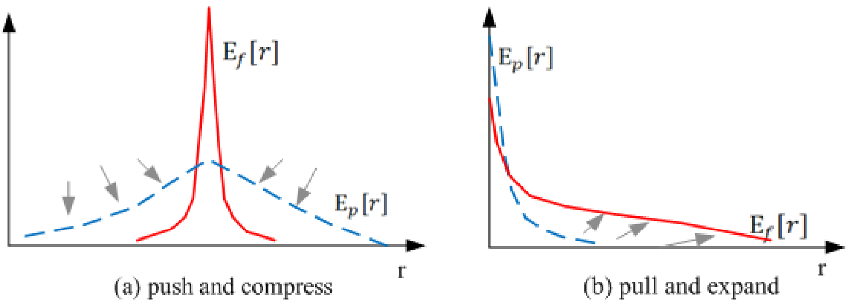
Y.N. Wu et al ICCV07, IJCV09

vs: Viola, Jones, 04

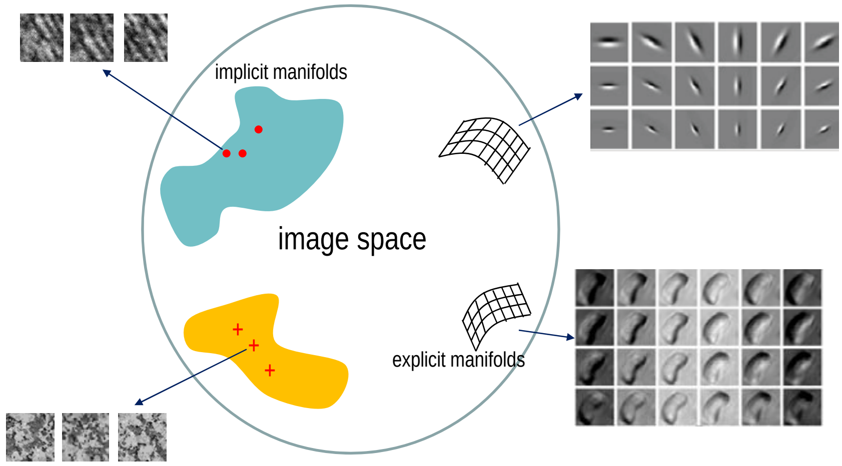
Template detection experiment

Y.N Wu et al. 2007

Matching Ω_p to Ω_f: Push and Pull



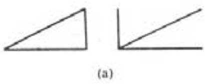
Summary: a second look at the space of image patches

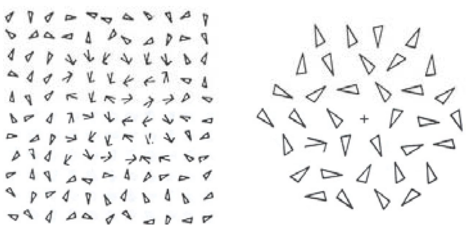


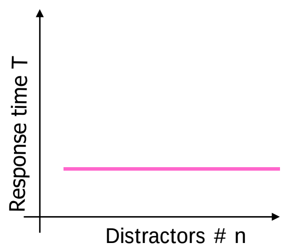
4, Relations to the literature: psychophysics

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

textons

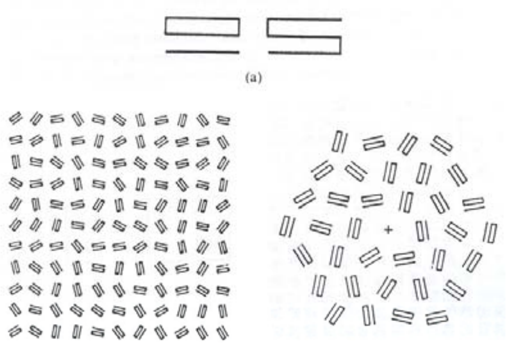


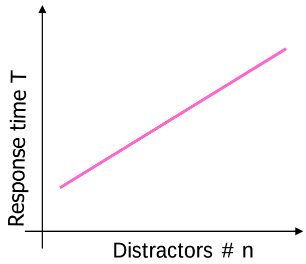




Textons vs. textures

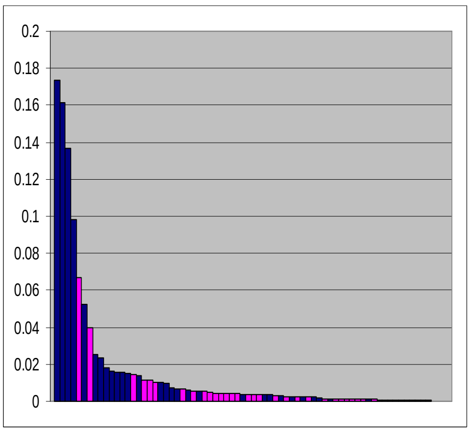
textures

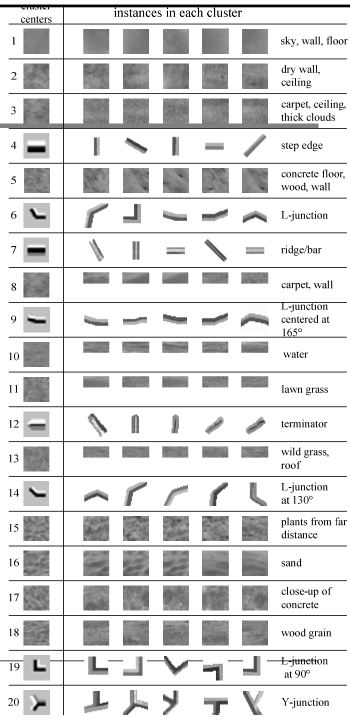




5, Frequency plot of the ex/implicit manifolds in natural images

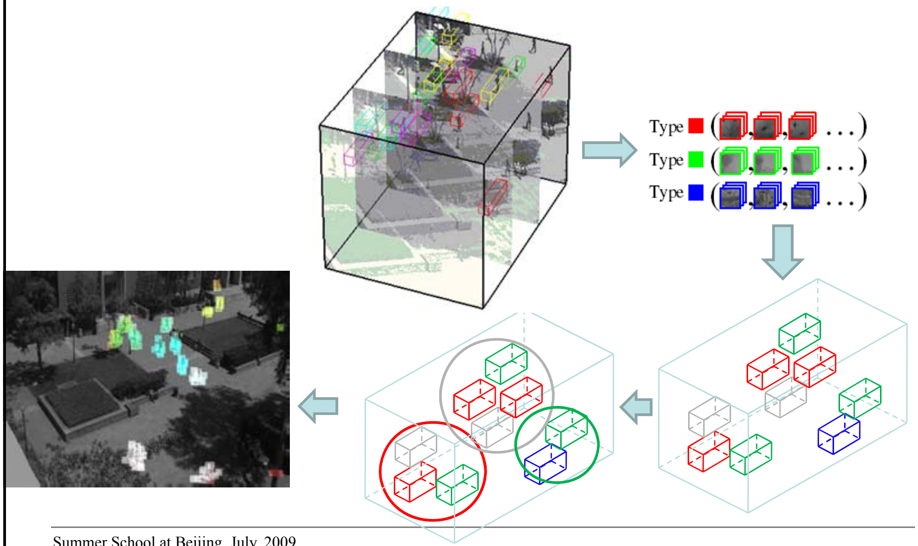
implicit texture clusters (blue),
explicit primitive clusters (pink).





Summer School at Beijing, July, 2009

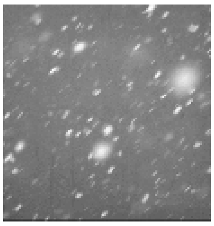
Clustering in video

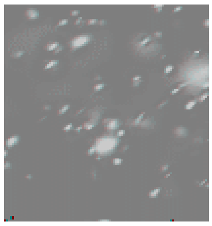


Examples in video

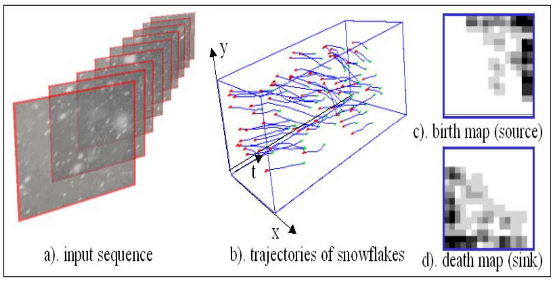
Textons in motion

Observed Sequence





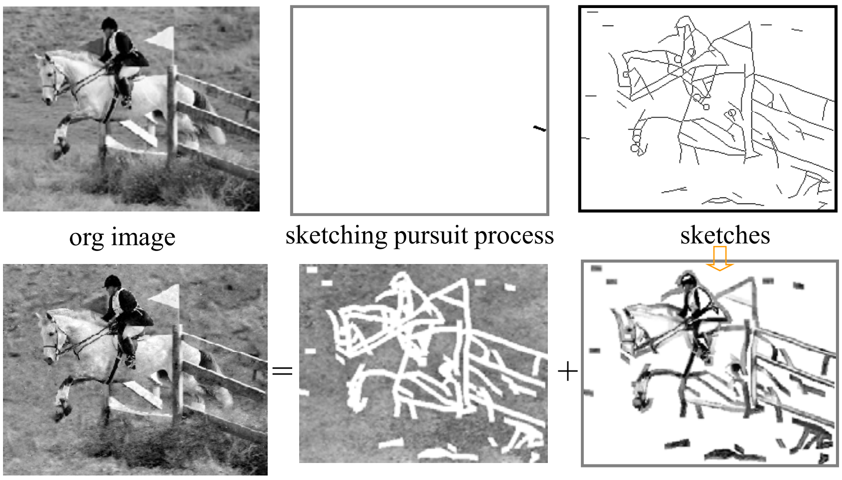
Synthesized Sequence



a). input sequence b). trajectories of snowflakes

Ref. Y.Z. Wang, 2003

6, Primal sketch: integrating the two regimes



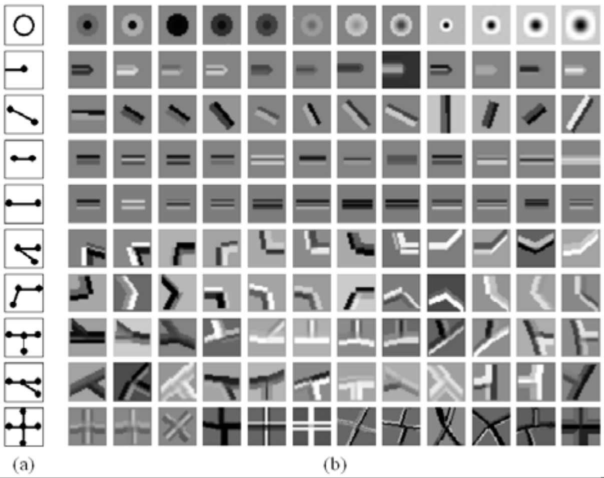
org image sketching pursuit process sketches

syn image = synthesized textures + sketch image

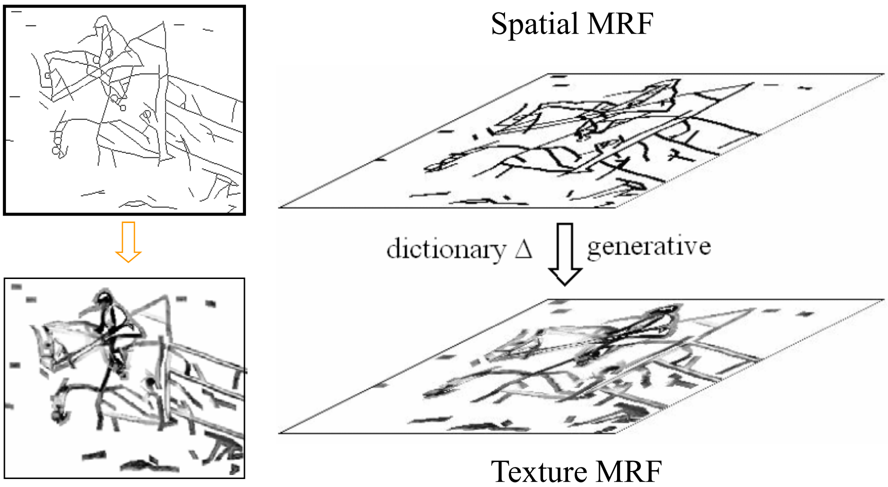
(Guo,Zhu,Wu, 2003-05)

manifolds of image primitives

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

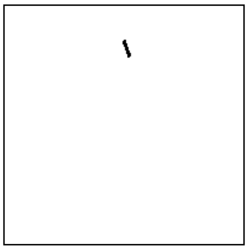


Primal Sketch is a two-level MRF model



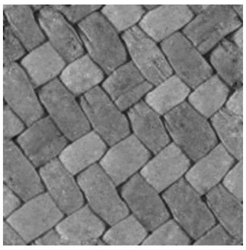
Primal sketch example

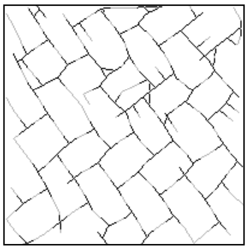
input image



sketching pursuit process

synthesized image

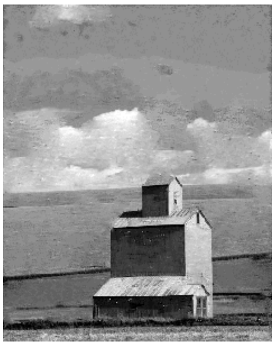




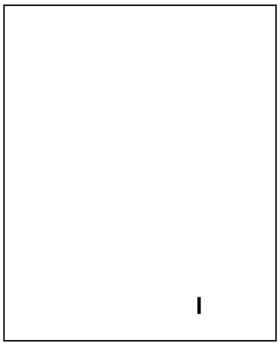
sketches

Primal sketch example

original image



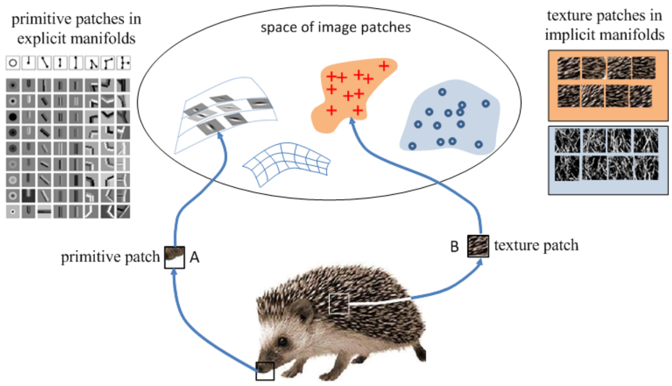
synthesized image



sketching pursuit process

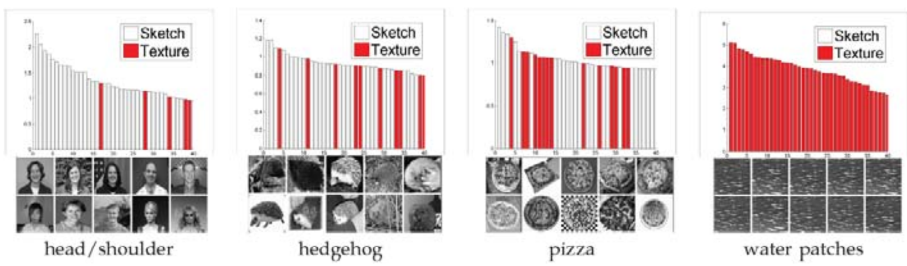
7, Pursuing composite manifolds in the middle entropy regime

Learning Hybrid Image Templates



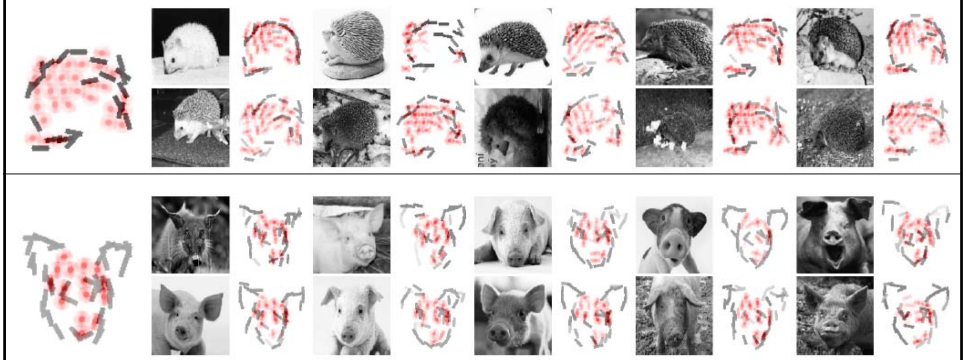
Learning object templates by manifold pursuit

The two types of models compete in learning the templates



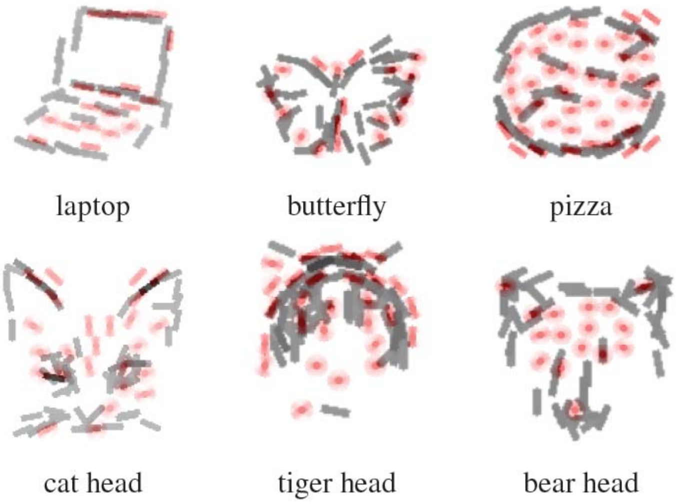
Examples of the learned hybrid image templates

Mixing the implicit and explicit manifolds

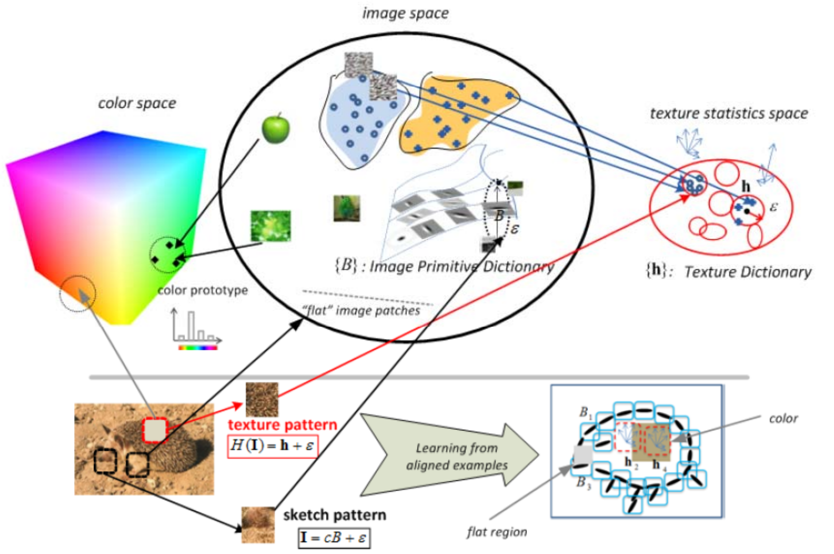


Z. Z. Si et al 2008-09

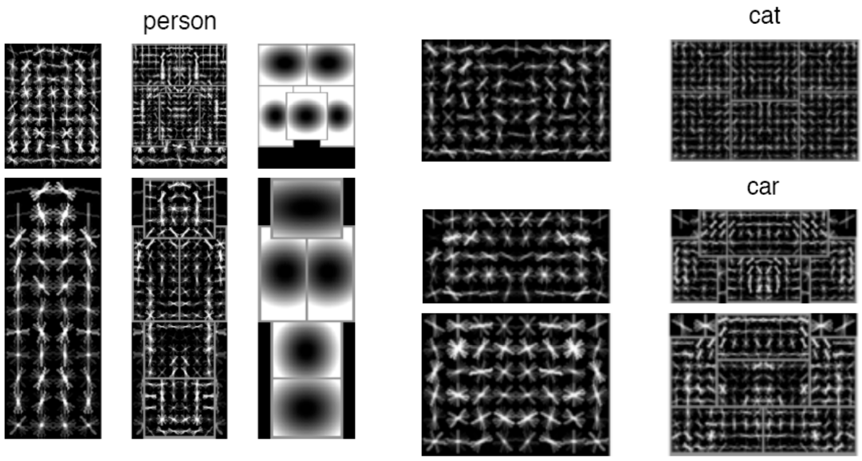
Some examples of learn object templates



The epsilon-ball interpretation



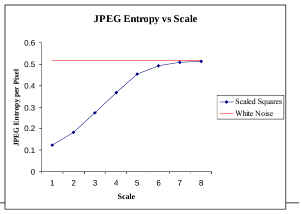
Comparing with the HoG Representation



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)

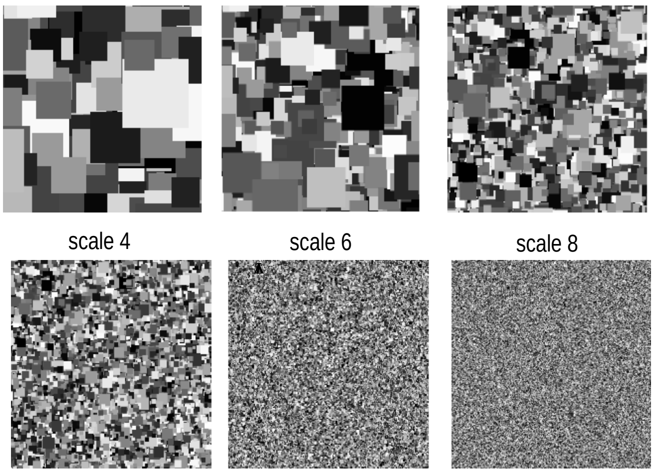


Wu, Zhu, Guo, 04,07

Information scaling leads to manifold transitions !

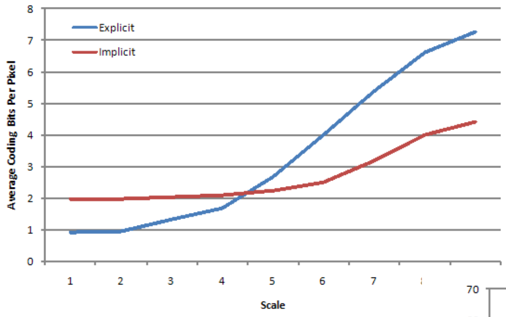
scale 1 scale 2 scale 3

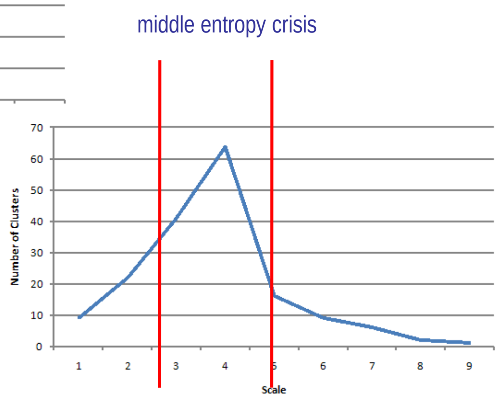
scale 4 scale 6 scale 8



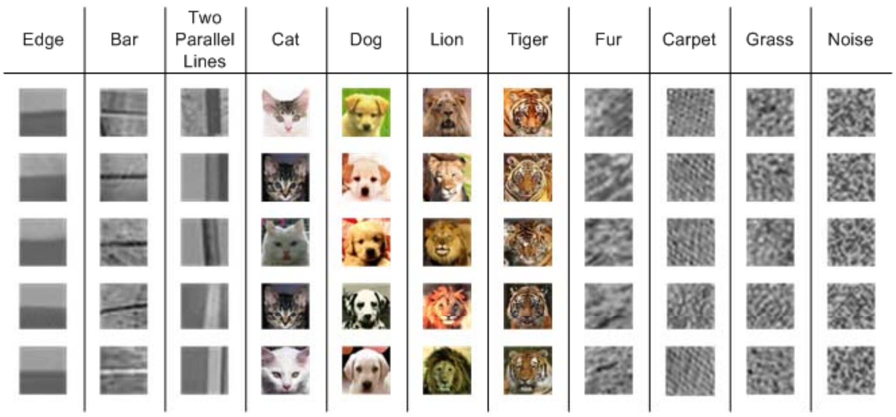
Ref:
Wu, Zhu, Guo, 2004-07,
"From Information Scaling to
Regimes of Statistical Models"

Coding efficiency and number of clusters over scales





A wide spectrum of categories from low to high entropy



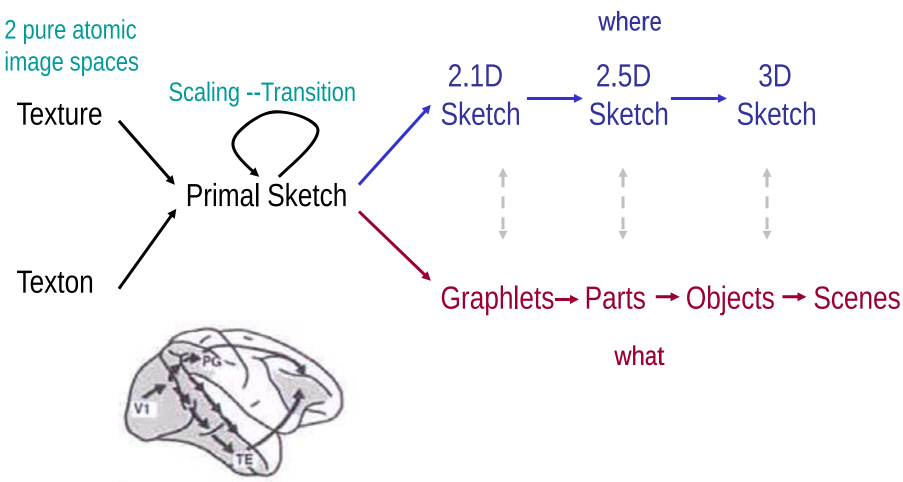
Entropy ~ Dimension ~ Log volume(manifold)

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



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

Ref: S.C. Zhu and D. Mumford, "A Stochastic Grammar of Images", *Foundations and Trends in Computer Graphics and Vision*, Vol.2, No.4, pp 259-362, 2006.

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, …

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: $\mathcal{G} = (V_N, V_T, \mathrm{R}, S)$

	Grammar	Production Rules
Type - 0	Unrestricted	$\boldsymbol{\alpha} \rightarrow \boldsymbol{\beta}$
Type - 1	Context-sensitive	$\boldsymbol{\alpha}\, \boldsymbol{A}\, \boldsymbol{\beta} \rightarrow \boldsymbol{\alpha}\, \boldsymbol{\gamma}\, \boldsymbol{\beta}$
Type - 2	Context-free	$\boldsymbol{A} \rightarrow \boldsymbol{\gamma}$
Type - 3	Regular	$\boldsymbol{A} \rightarrow \boldsymbol{a}$ $\boldsymbol{A} \rightarrow \boldsymbol{a}\, \boldsymbol{B}$

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

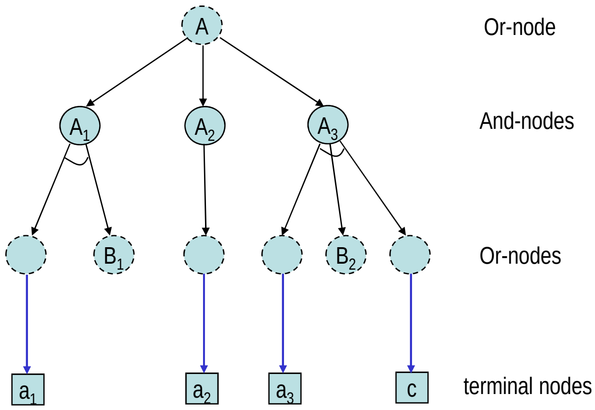
$$\mathbf{L}(\mathcal{G}) = \left\{\omega : \ S \overset{\mathrm{R}^*}{\Longrightarrow} \omega,\ \omega \in V_T^*\right\} \qquad S \overset{\gamma_1, \gamma_2, \cdots, \gamma_{n(\omega)}}{\Longrightarrow} \omega$$

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

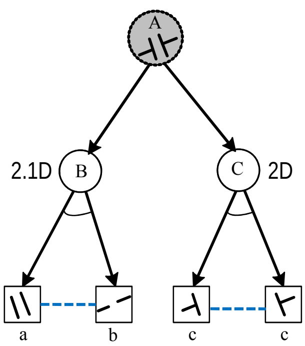
A special property of image grammar is that any node can terminate or “ground” immediately.

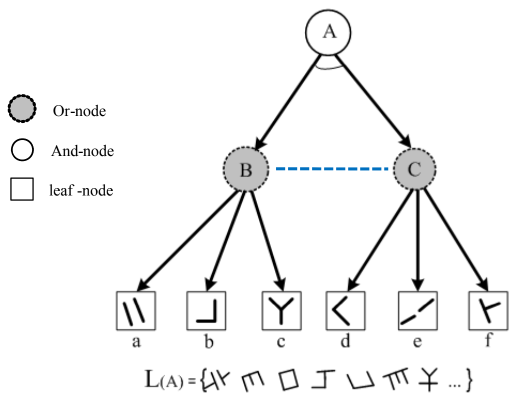


Representing grammar by And-Or graph

A grammar production rule:

$A \rightarrow ab \,|\, cc$

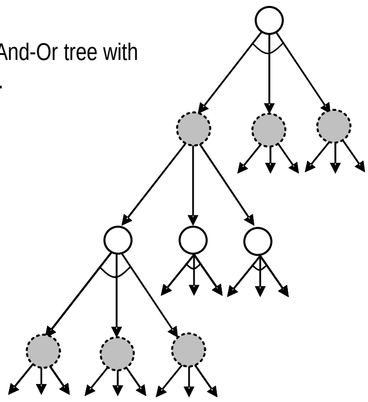




The language of a node A is the set of all valid configurations

The expressive power of and-or graphs

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

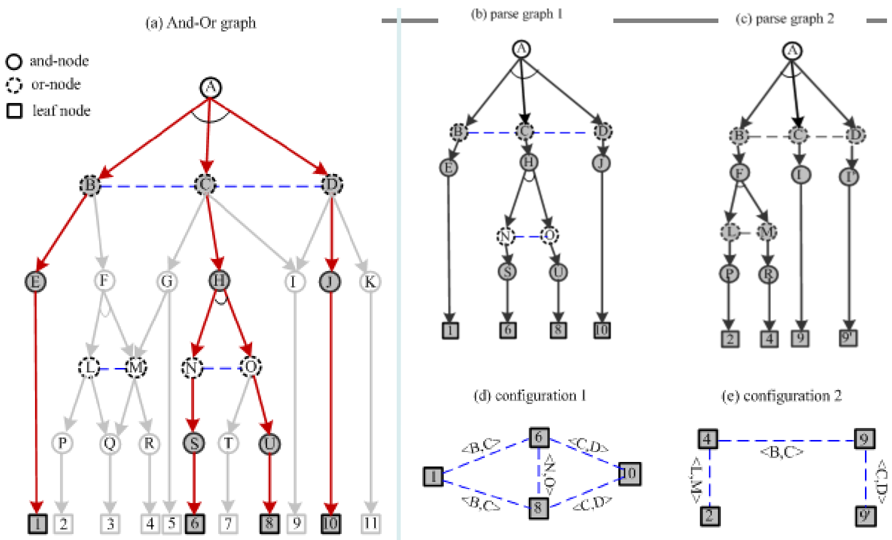


And-Or graph was used in the 1980s by Judea Pearl for heuristic search in AI.
For example the 12-counterfeit coin problem.

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

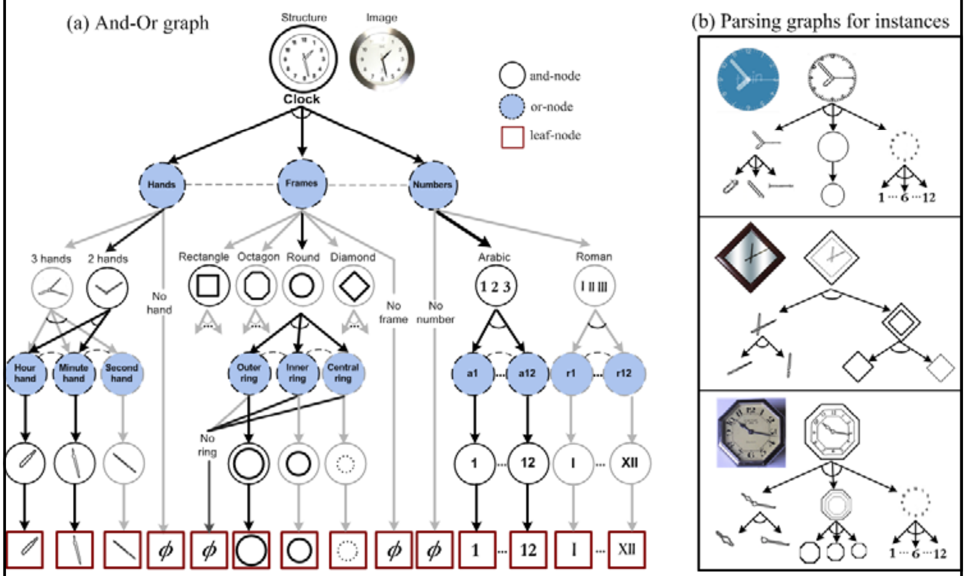
$(3x3^3)^3=3^{12}$ = 531,441configurations

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

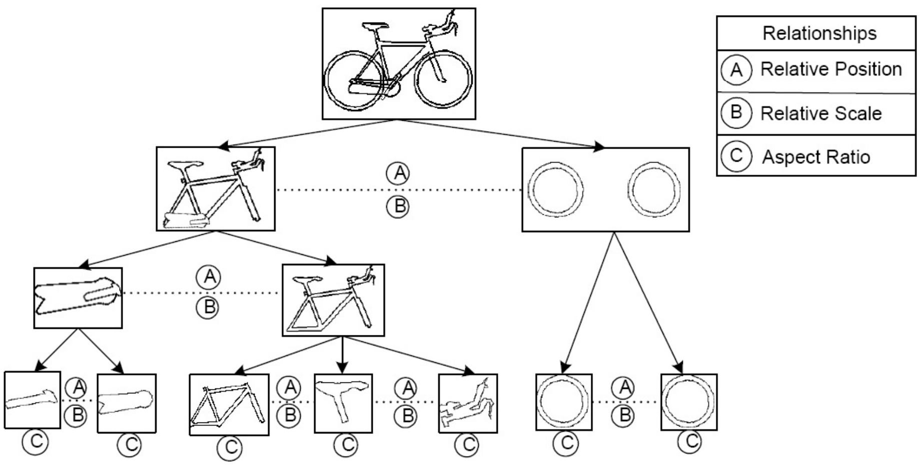


Each category is conceptualized to a grammar whose language defines a set or "equivalence class" for all the valid configurations of the each category.

An example: the clock category

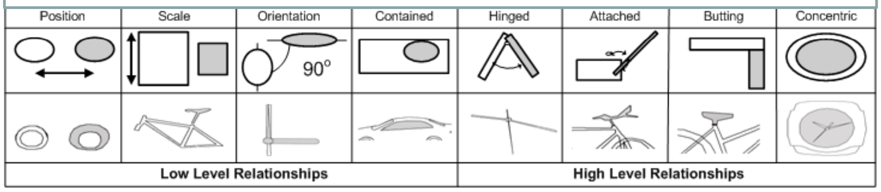


A parse graph of a bike



A relation is like a non-linear filter

Some examples



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

$$A = (a_1, a_1, \ldots, a_n)$$

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

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.

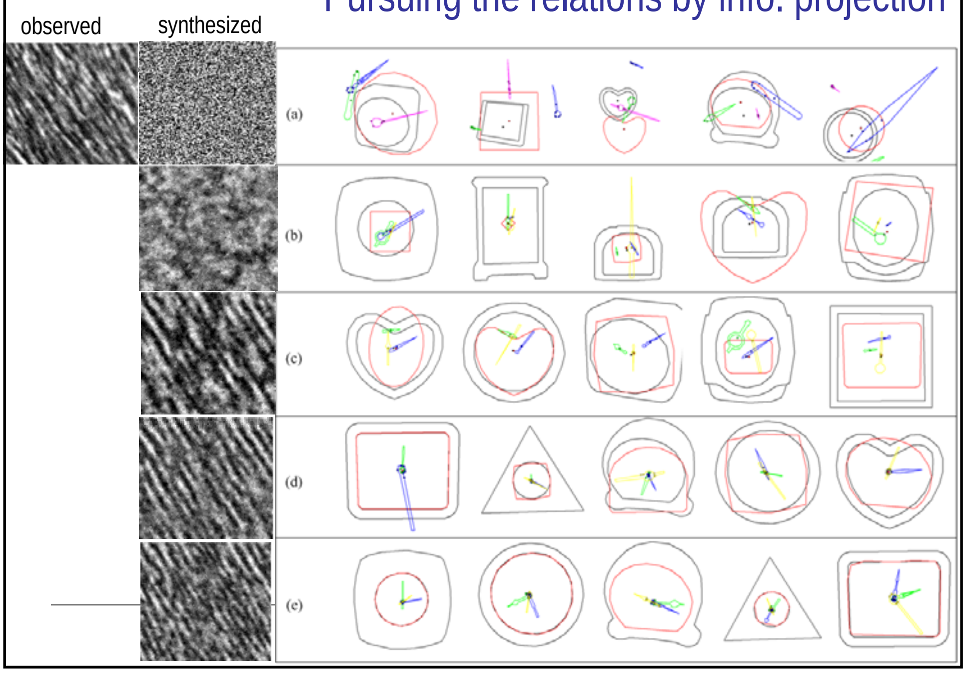
The probability model is defined as

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

The first term alone stands for a SCFG.
The second and third terms are Markov potentials.

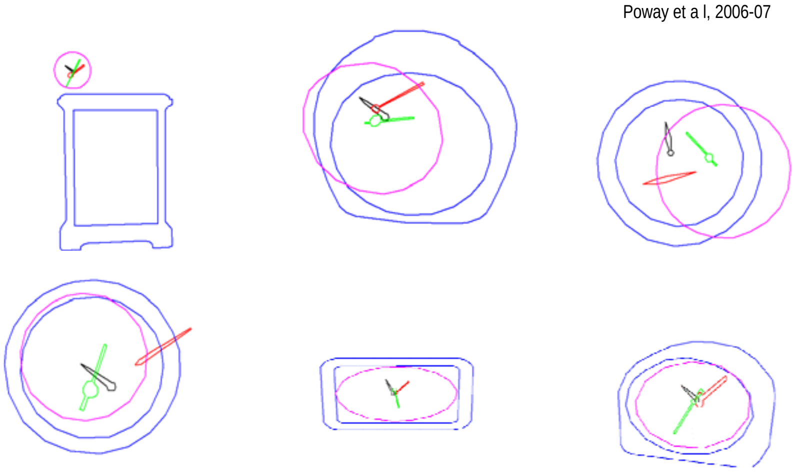
By analogy to texture modeling

Pursuing the relations by info. projection

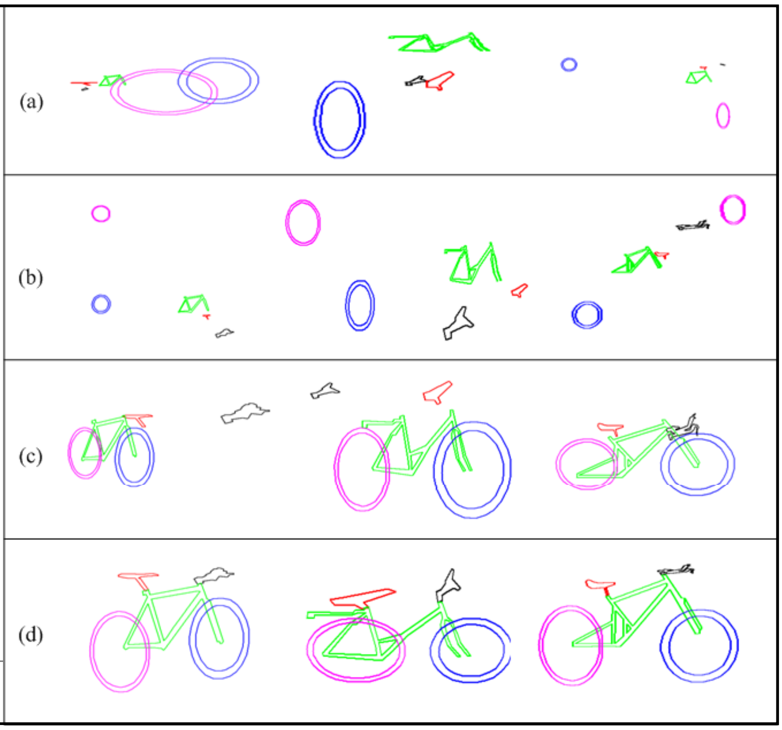


Sampling Clock: we keep evolving O(100) samples

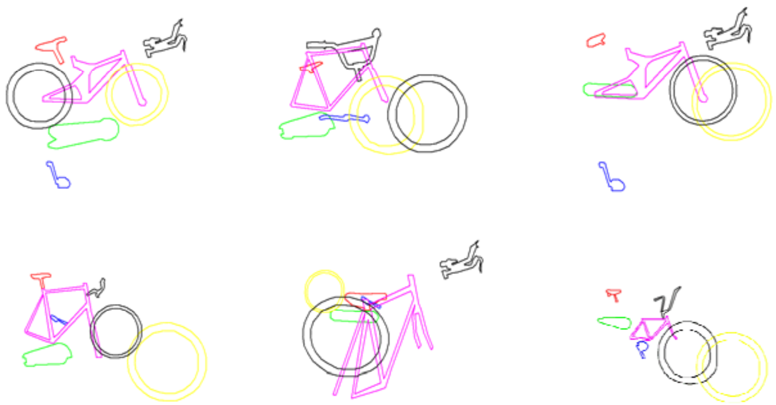
Poway et a l, 2006-07



Learning and sampling a bike model



Examples of sampling bicycles: Computer can dream

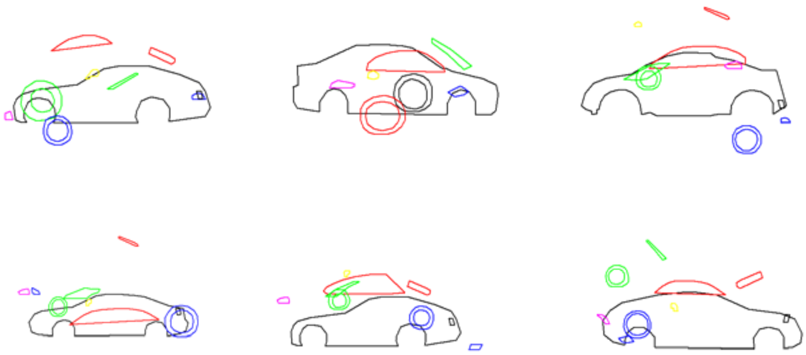


Dreaming is a process of learning

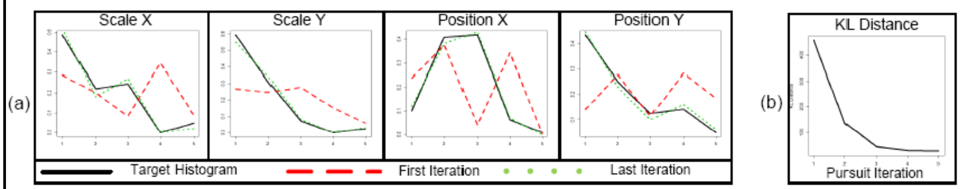
Poway et a l, 2006-07

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.

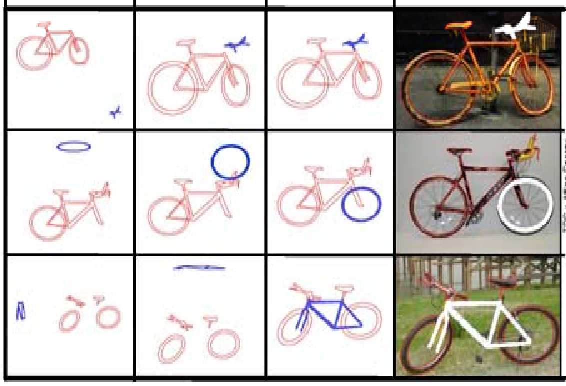
(a) Histograms for four pairwise relationships at different iterations. The last iteration matches the observed histogram quite closely.

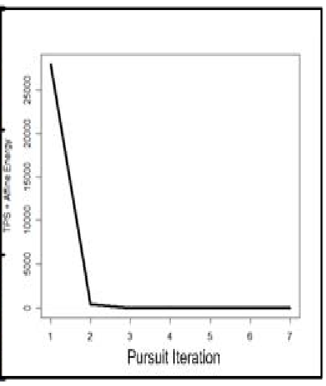
(b) The KL divergence between the current and target model as the relationship pursuit is performed.

Top-down prediction by sampling the missing part

Stage 1: random prediction | Stage 2: correct scale | Stage 3: more relations

Poway, et al 2006-07

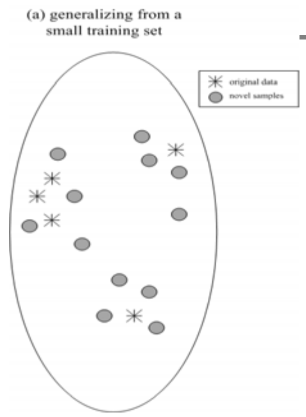


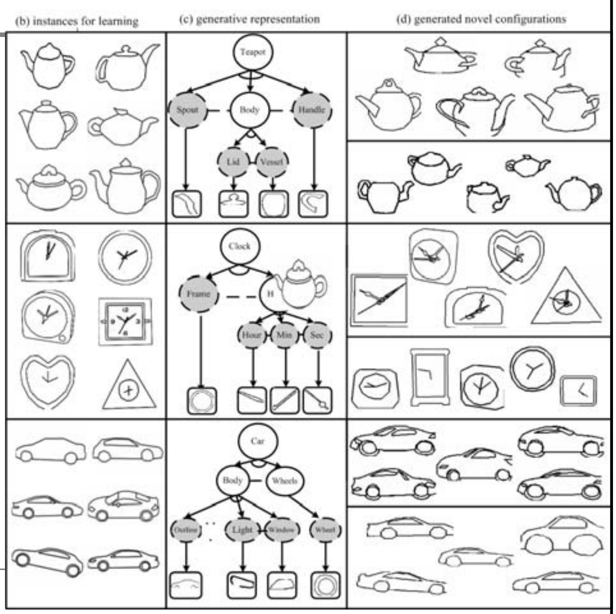


The blue parts are predicted by the learned models at various learning stages

Learning from a small training set & generalization by sampling

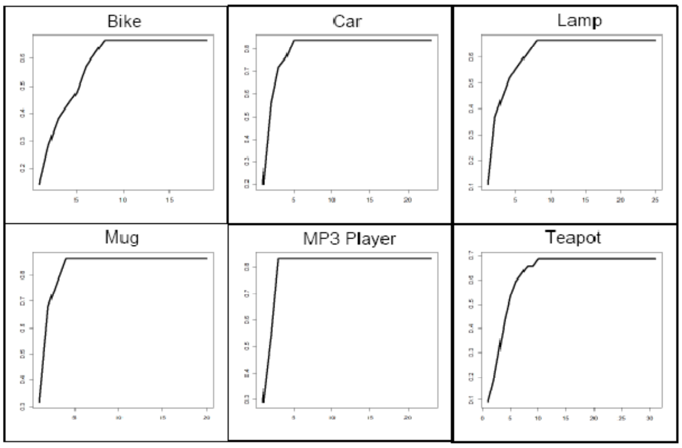
Poway, Yao, and Zhu, 2006-07





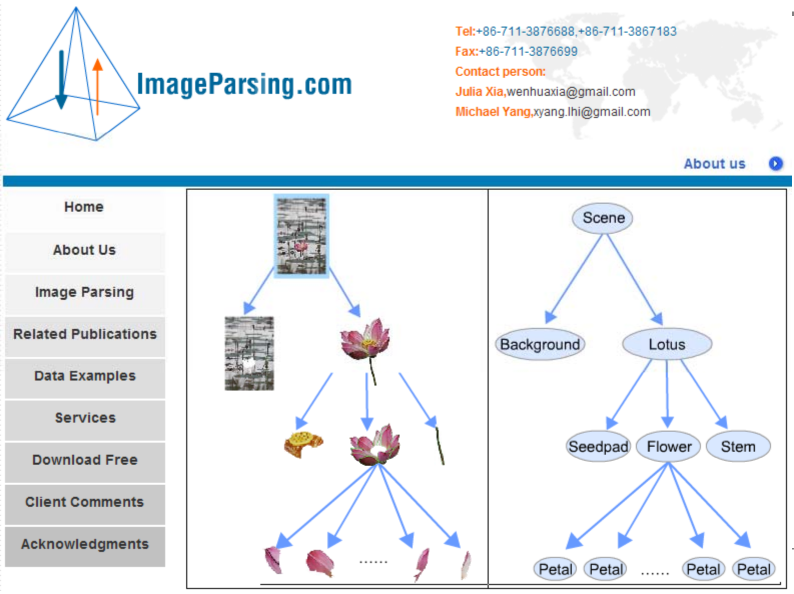
In our recognition experiment with 33 categories, each category has 50-60 samples. the sampling process improves the average recognition rate from 66% to 81%.

What is the smallest sample set for training?

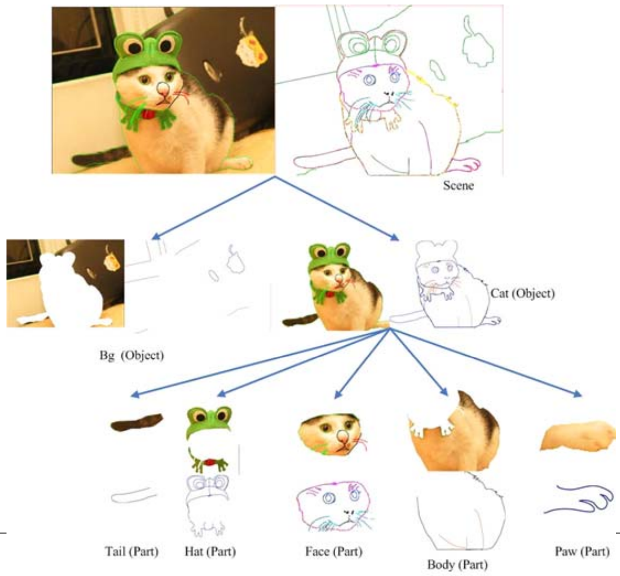


Coverage results for 6 categories. we only need a small fraction of the training set to maximally cover the testing set.

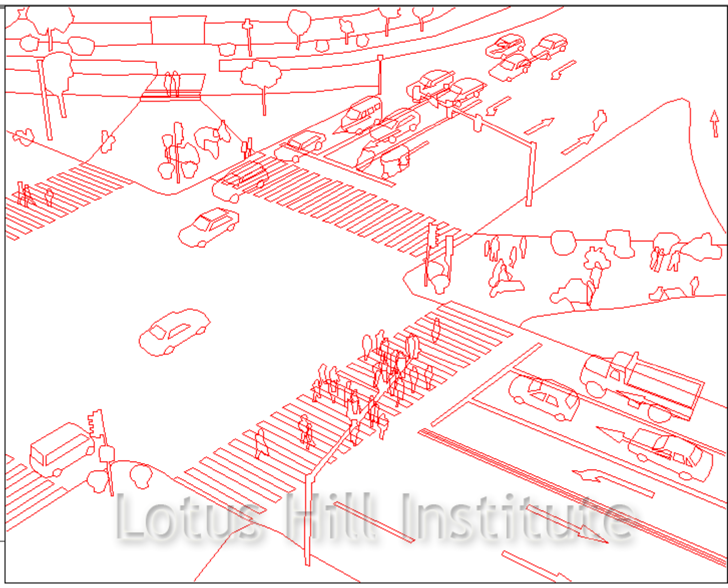
4, A large scale human annotation project at Lotus Hill



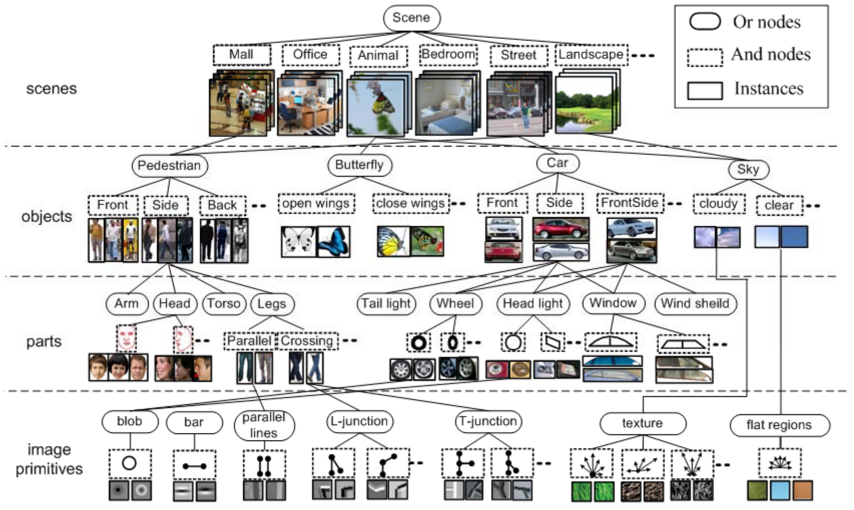
An example: parse graph of a cat



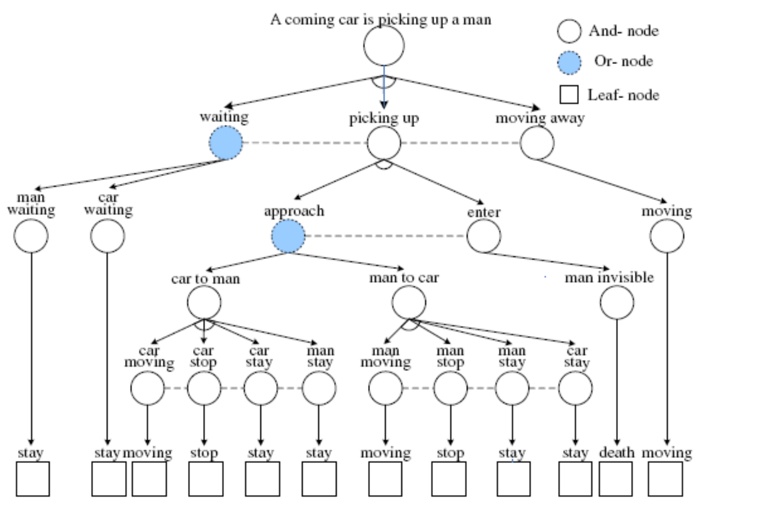
An example: clip for surveillance video



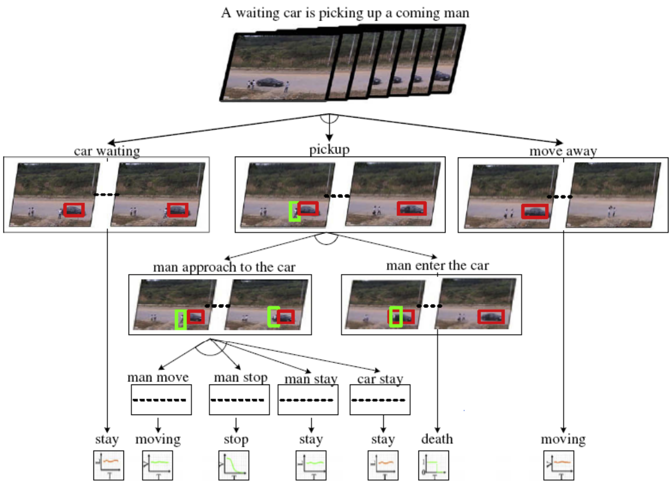
LHI dataset: 1.8M images parsed, 1000s AoG nodes



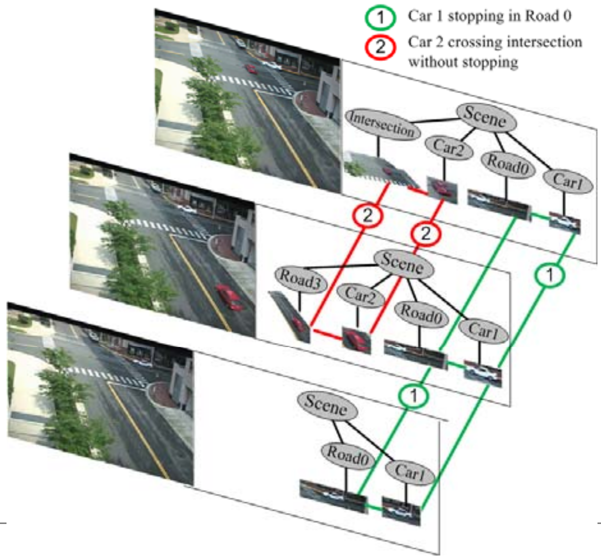
5, Example: Representing An event by AoG



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 57:18.

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

where

Scaling --Transition

2.1D Sketch → 2.5D Sketch → 3D Sketch

Texture → Primal Sketch

Texton → Primal Sketch

Graphlets→ Parts → Objects → Scenes

what