# From Primal Sketch to $2\frac{1}{2}$D Sketch

Song-Chun Zhu

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

Joint work with A. Barbu, C.E. Guo, F. Han, Y.Z. Wang, Y. N. Wu



## A classical model at low level vision

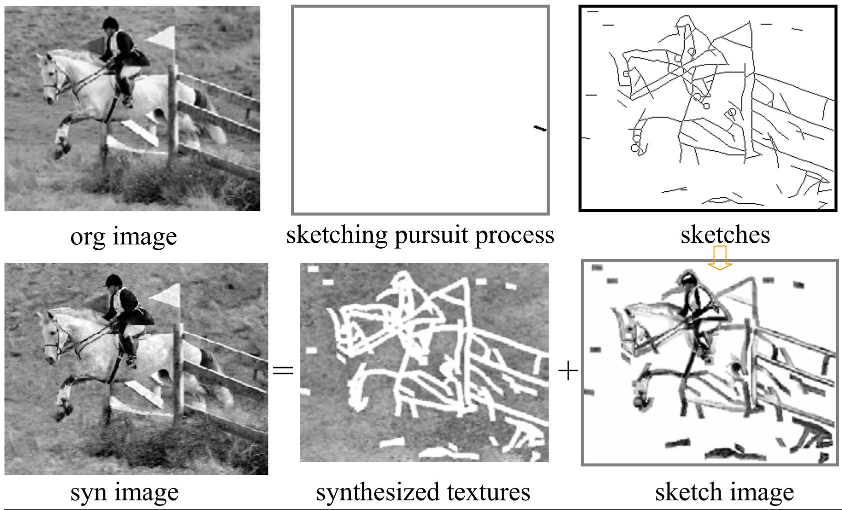
In the 1980s, a classical model came in many ways for low level vision and shape-from-X
Regularization theory, physically-based model, robust statistics, ….
**Line process** (Geman and Geman, 84) ,
**Weak membrane/thin-plate** (Zisserman and Blake, 85)
**Cartoon model** (Mumford-Shah, 89)

$$p(J,\Gamma) = \frac{1}{Z}\exp\{-\alpha\int\int_{\Lambda\backslash\Gamma}|\nabla J(x,y)|^2 dxdy - \beta||\Gamma||\}$$

1. Why is this potential function?
2. Why use this operator (filter)? How many are optimal?
3. Where is the “edge” from?

# Primal Sketch Model

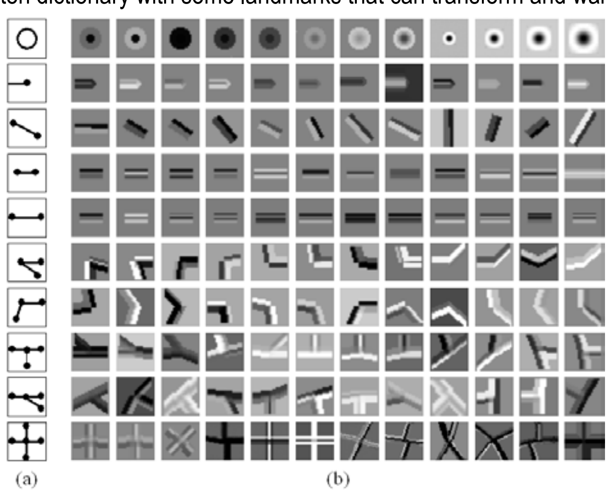
(Guo, Zhu and Wu, iccv03)



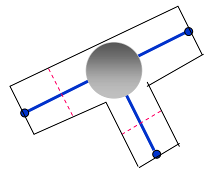



# Examples of the image primitive

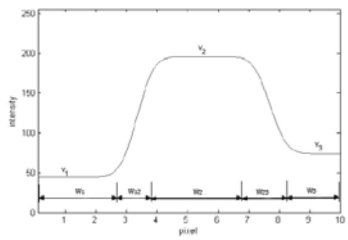
Learned texton dictionary with some landmarks that can transform and warp the patches



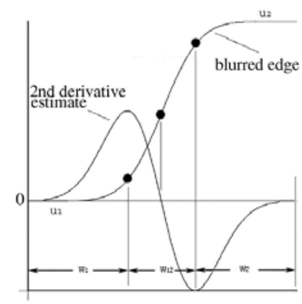
## Intensity profiles perpendicular to the axis



Bar




Step edge

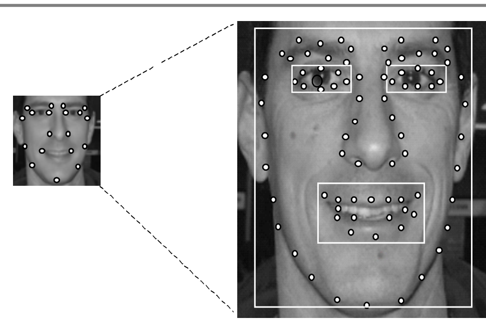



Similarly we model blobs, terminators, and blurred junctions.



## Image primitives are similar to the AAM model

Geometric: 2D warping
Photometric: variations

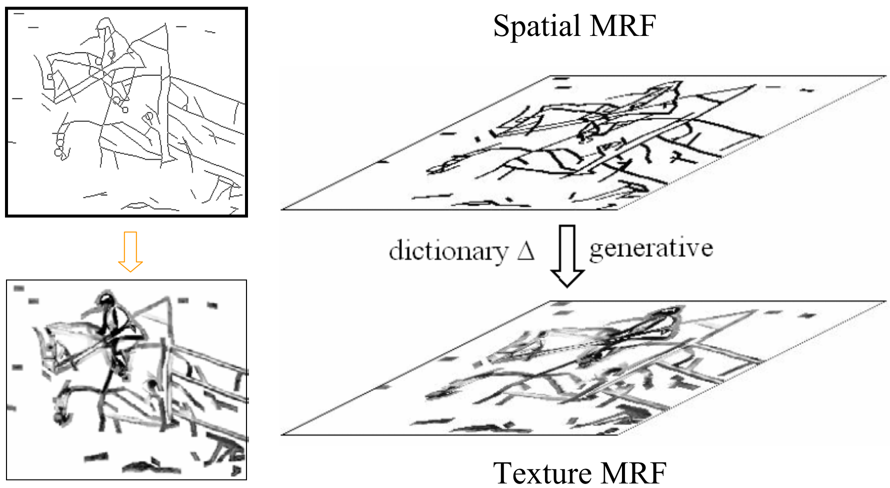


Extension:

1. Topological variability
2. Lighting modeling, e.g. folds for clothes
3. 3D geometry, e.g. different boundaries for stereo
4. dynamics, e.g. graphs in motion.

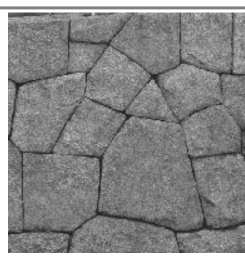
2 ½ D sketch will be much easier if we have visual knowledge coded.

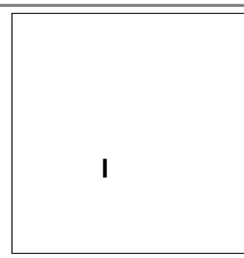
# Primal sketch: two-level representation

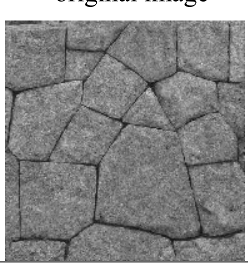


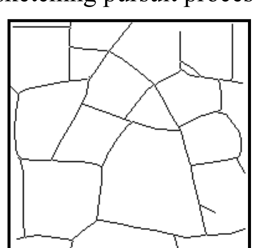



# More Example


original image

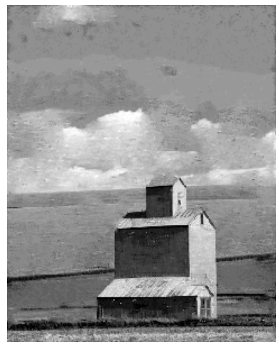

sketching pursuit process


synthesized image

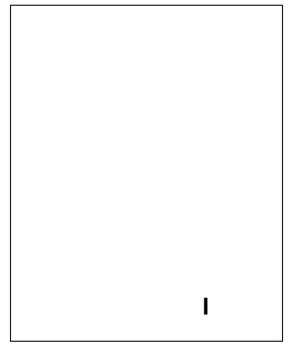

sketches

# More example

original image



synthesized image



sketching pursuit process



# The primal sketch model

1. The lattice is divided into two parts: sketchable and non-sketchable

$$\Lambda = \Lambda_{\mathrm{sk}} \cup \Lambda_{\mathrm{nsk}}$$

2. The sketchable part is divided into disjoint domains,

$$\Lambda_{\mathrm{sk}} = \cup_{k=1}^{K} \Lambda_{\mathrm{sk,k}}$$

Each domain is covered by a patch from a dictionary $\Delta_{\mathrm{sk}}$

$$\mathbf{I}_{\Lambda_{\mathrm{sk,k}}}(u,v) = \mathbf{B}_k(u,v) + n, \quad k = (\ell, x, y, \theta, \sigma, \alpha_{\mathrm{pho}}, \alpha_{\mathrm{wrp}})$$

Patches are aligned by landmarks (anchors) to form an attributed graph

$$S_{\mathrm{sk}} = (K,\ \{(\Lambda_{\mathrm{sk,k}}, \mathrm{B}_k) :\ k = 1, 2, ..., K\})$$

# The primal sketch model

3. The non-sketchable part is divided into homogeneous texture regions

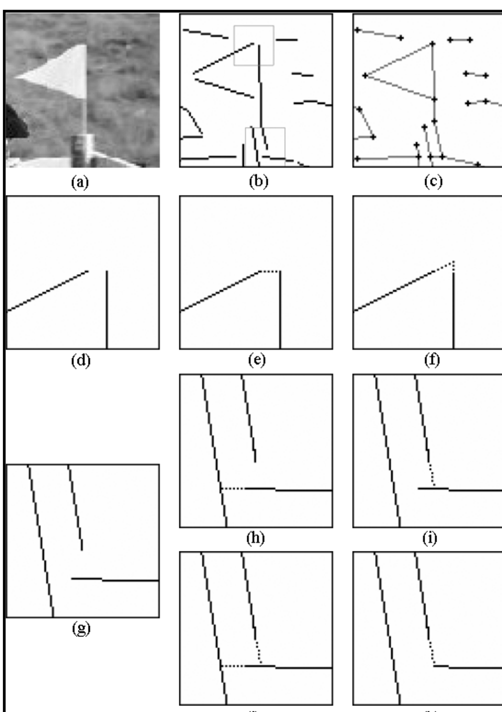
$$\Lambda_{\text{nsk}} = \cup_{i=1}^{n} \Lambda_{\text{nsk,i}}$$

Each region has a statistical summary $h_n$

$$S_{\text{nsk}} = (N,\ \{(\Lambda_{\text{nsk,i}}, h_i \leftrightarrow \beta_i) :\ n = 1, 2, ..., N\})$$

$$\begin{aligned} p(\mathrm{I}, S_{\text{sk}}, S_{\text{nsk}}; \Delta_{\text{sk}}, \Delta_{\text{nsk}}) \;&=\; \frac{1}{Z} \exp\{-E_{\text{sk}}(S_{\text{sk}}) - E_{\text{nsk}}(S_{\text{nsk}}) \\ &\quad - \sum_{k=1}^{K} \sum_{(x,y)\in\Lambda_{\text{nsk,k}}} (\mathrm{I}(u,v) - B_k(x,y))^2 \\ &\quad - \sum_{i=1}^{n} < \beta_i, h(\mathrm{I}_{\Lambda_{\text{nsk,i}}}) > \} \end{aligned}$$






# A zoom-in view

The algorithm works in two phases:

1. bottom-up sketching,
   like matching pursuit

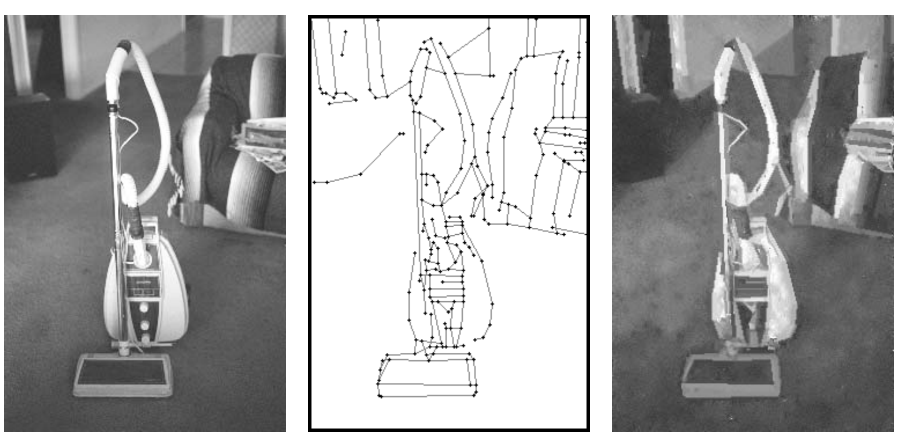
2. graph editing by operators
   to get the Gestalt field
   where the junctions are adjusted

# Reversible graph operators

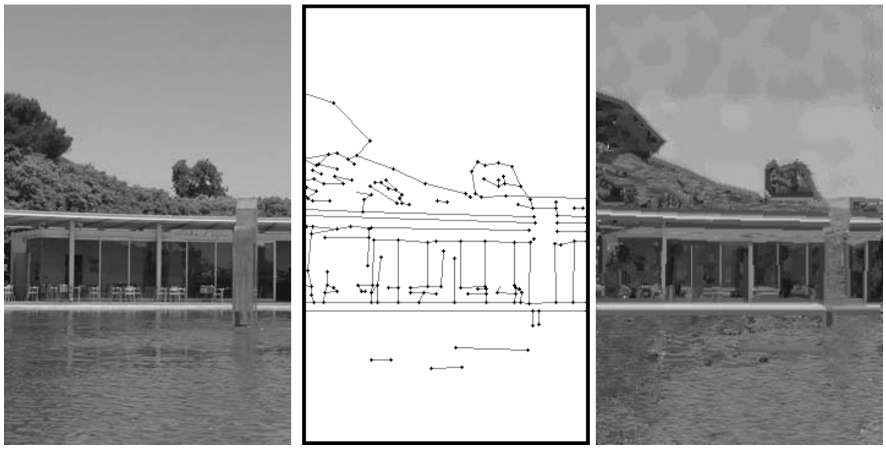
| operators | graph change | illustration |
|---|---|---|
| $O_1, O_1'$ | create / remove a stroke | Φ ⟺ |
| $O_2, O_2'$ | grow / shrink a stroke | ⟺ |
| $O_3, O_3'$ | connect / disconnect vertices | ⟺ |
| $O_4, O_4'$ | extend one stroke and cross / disconnect and combine | ⟺ |
| $O_5, O_5'$ | extend two strokes and cross / disconnect and combine | ⟺ |
| $O_6, O_6'$ | combine two connected strokes / break a stroke | ⟺ |
| $O_7, O_7'$ | combine two parallel strokes / split one into two parallel | ⟺ |
| $O_8, O_8'$ | merge two vertices / split a vertex | ⟺ |
| $O_9, O_9'$ | create / remove a blob | Φ ⟺ |
| $O_{10}, O_{10}'$ | switch between a stroke(s) and a blob | ⟺ |



# More examples



# More examples





# Manifold learning and entropy minimization

Let $\Omega_{nat}$ be the ensemble of natural images on large enough lattice. To measure the Volume/dimension of this manifold, we construct an ensemble $\Omega_\varepsilon$ which is an $\varepsilon$-cover of $\Omega_{nat}$ for a certain perceptual metric $\rho$.

$$\forall \mathrm{I} \in \Omega_{\text{nat}}, \ \exists \mathrm{J} \in \Omega_{\epsilon}, \ \text{ so that } \ \rho(\mathrm{I}, \mathrm{J}) \leq \epsilon.$$

The minimum $\varepsilon$-cover has size $\mathcal{N}(\Omega_{\text{nat}}, \rho, \epsilon)$
The $\varepsilon$-entropy of the natural image ensemble is

$$\mathcal{H}(\Omega_{\text{nat}}, \rho, \epsilon) = \log_2 \mathcal{N}(\Omega_{\text{nat}}, \rho, \epsilon)$$

In the literature, there are two ways for manifold learning using two perceptual metrics

1. generative models (Harmonic analysis)
2. descriptive models (Markov random fields)

## Explicit manifold learning

Generative models build the e-ensemble by explicit functions,

$$\Omega_{\text{gen}} = \{\mathbf{I}: \quad \mathbf{I} = g(W; \Delta_{\text{gen}}), \ W \in \Omega_W\}$$

W are the dimensions of the manifold $\Omega_{W:}$ geometric and photometric. The metric is the MSE,

$$\rho_{\text{gen}}(\mathbf{I}, \mathbf{J}) = \frac{1}{|\Lambda|} \sum_{x,y} (\mathbf{I}(x,y) - \mathbf{J}(x,y))^2$$

This ensemble has size $\mathcal{M}(\Omega_{\text{gen}}, \rho_{\text{gen}}, \epsilon)$

The ε-entropy of the ensemble is

$$\mathcal{H}(\Omega_{\text{gen}}, \rho_{\text{gen}}, \epsilon) = \log_2 \mathcal{M}(\Omega_{\text{gen}}, \rho_{\text{gen}}, \epsilon)$$

The objective is to find the optimal dictionary to minimize the discrepancy (KL-divergence),

$$\Delta^*_{\text{gen}} = \arg\min \{\mathcal{H}(\Omega_{\text{gen}}, \rho_{\text{gen}}, \epsilon) - \mathcal{H}(\Omega_{\text{nat}}, \rho_{\text{gen}}, \epsilon)\}$$



## Implicit manifold learning

Generative models build the e-ensemble by explicit functions,

$$\Omega_{\text{des}} = \{\mathbf{I}: \quad h(\mathbf{I}; \Delta_{\text{des}}) = h_o, \ h_o \in \Omega_h\}$$

h are the statistics/features extracted (projection of the image space).
The metric is on the projected statistics,

$$\rho_{\text{des}}(\mathbf{I}, \mathbf{J}) = ||h(\mathbf{I}) - h(\mathbf{J})||$$

This ensemble has size $\mathcal{M}(\Omega_{\text{des}}, \rho_{\text{des}}, \epsilon)$

The ε-entropy of the ensemble is

$$\mathcal{H}(\Omega_{\text{des}}, \rho_{\text{des}}, \epsilon) = \log_2 \mathcal{M}(\Omega_{\text{des}}, \rho_{\text{des}}, \epsilon)$$

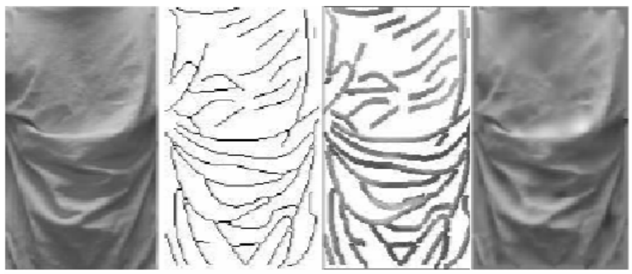
The objective is to find the optimal dictionary to minimize the discrepancy (KL-divergence),

$$\Delta^*_{\text{des}} = \arg\min \{\mathcal{H}(\Omega_{\text{des}}, \rho_{\text{des}}, \epsilon) - \mathcal{H}(\Omega_{\text{nat}}, \rho_{\text{des}}, \epsilon)\}$$

# Shape from shading with sketch

(Han and Zhu 05)

We take clothes as example.



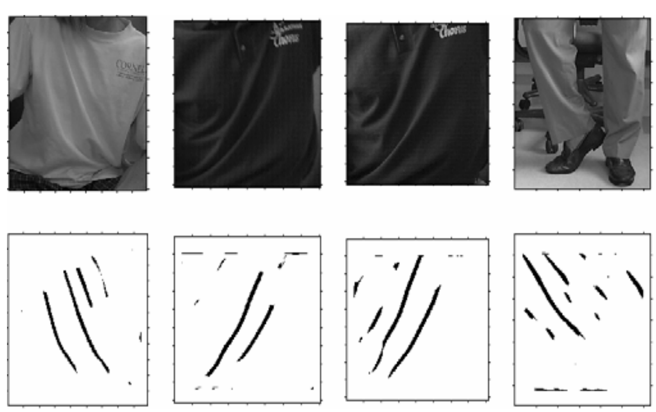
(a) input (b)folds graph $G$ (c)$\mathbf{I}_{\text{fd}}$ (d) Filling result



# Related work: fold detection by SVM

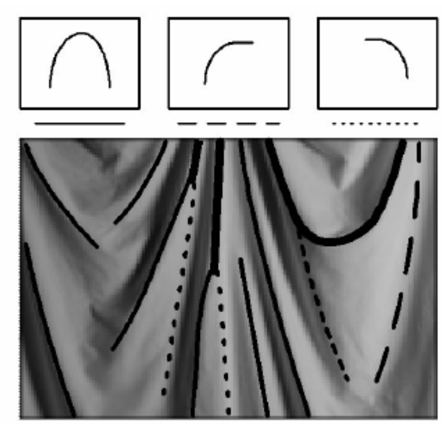
(Forsyth 97)

"Shading Primitives"



## Learning fold primitives

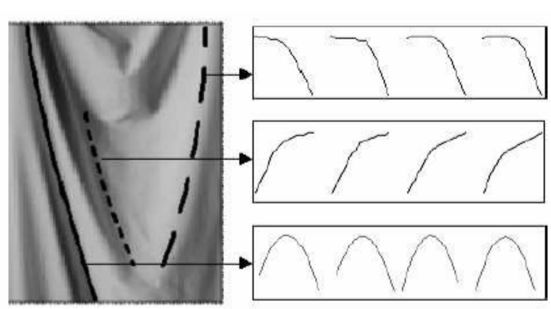
Three types of fold primitives





## Learning fold primitives

Model fold primitive profile by PCA



$$\mathbf{I}(x,y) = \eta n \cdot \vec{L} = R(p,q) = \eta \frac{-pl_1 - ql_2 + l_3}{\sqrt{p^2+q^2+1}}$$

# Learning fold primitives

1. We obtain the depth map of cloth surfaces by photometric stereo,
2. We draw the folds on the depth map manually
3. We learn the folds by surface fitting.

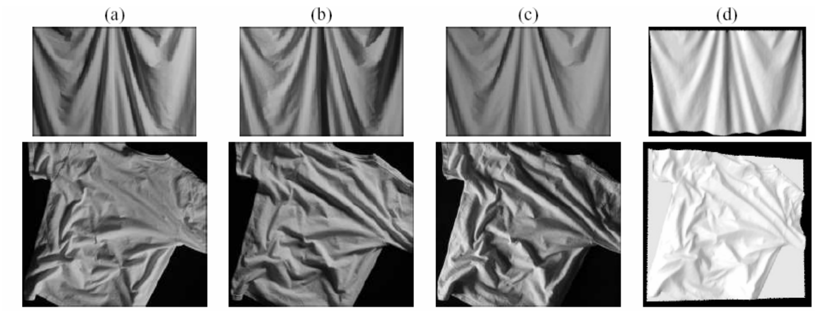
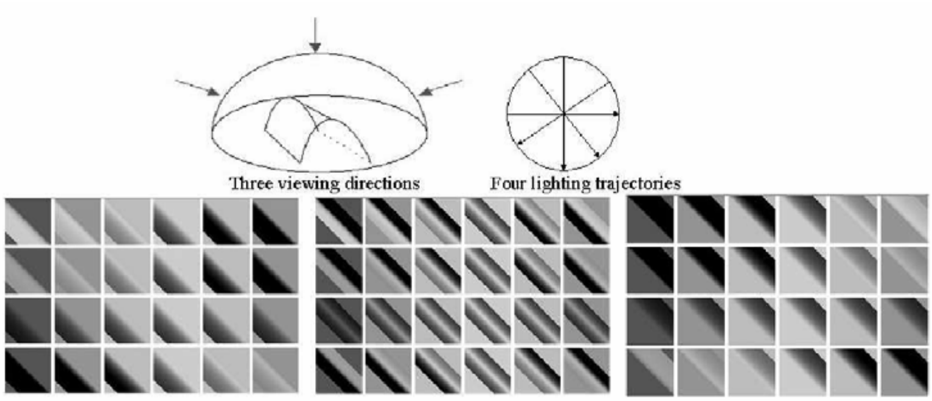



Figure 6: (a), (b), (c) are three images out of the sequence used to reconstruct the 3D cloth shape in (d).



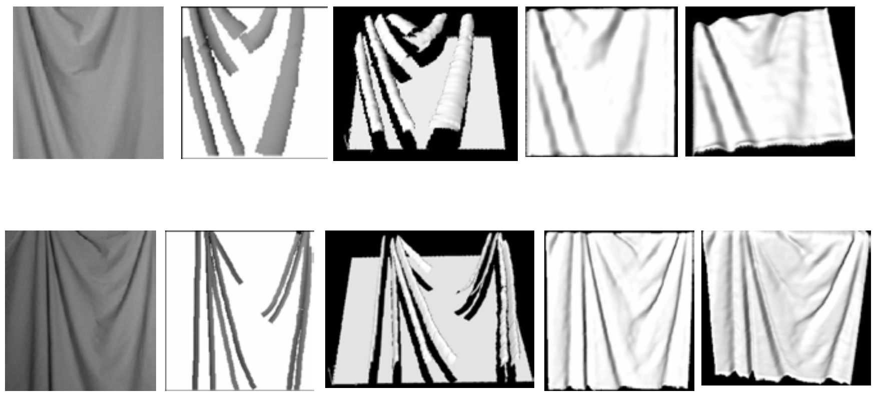
# Learning fold primitives



# Experimental results

(Han and Zhu 05)

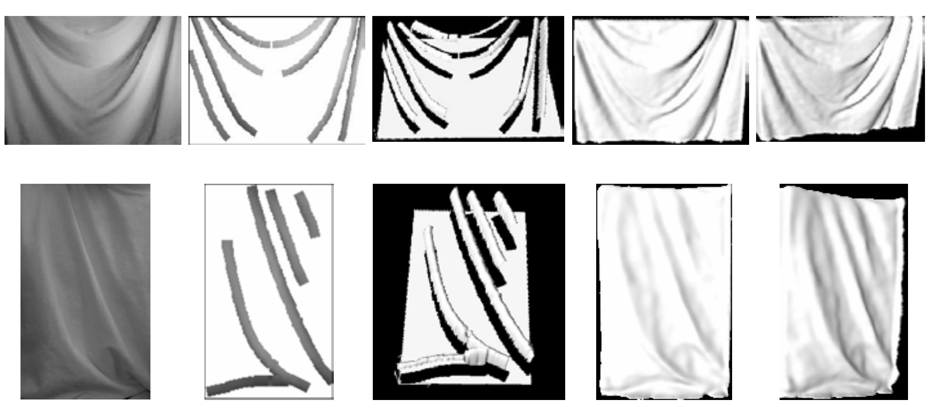
Input image    folds    surface of folds    full surface    novel view



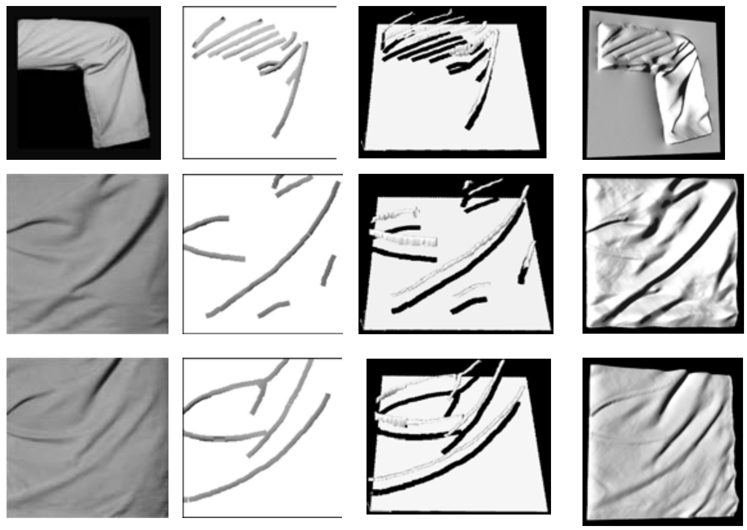


# Experimental results

Input image    folds    surface of folds    full surface    novel view

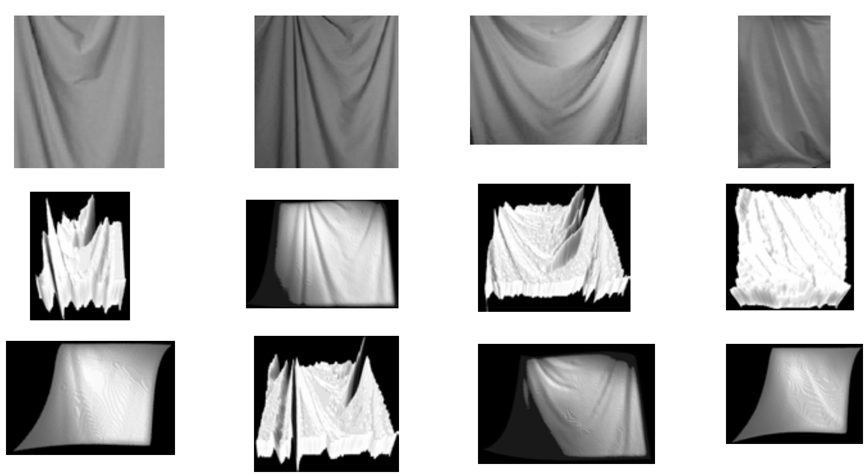


# Experimental results





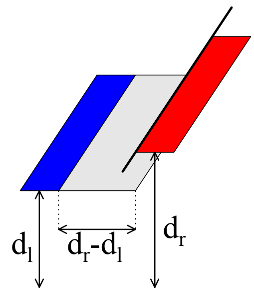
# Comparison: without folds



# Example on stereo vision (Barbu and Zhu 05)

Three types of edges:

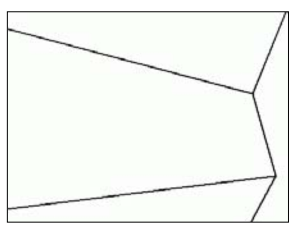
- Surface edges
- V edges where continuity is preserved but derivatives are different of the left and right of the edge
- Occlusion edges

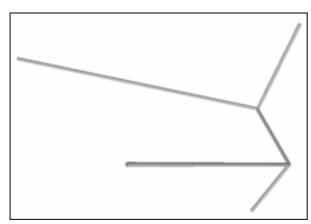


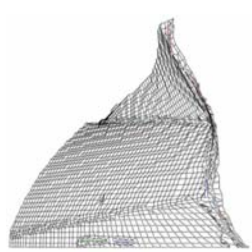


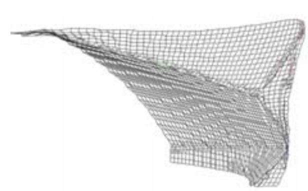

# Results on textureless surfaces

Original image

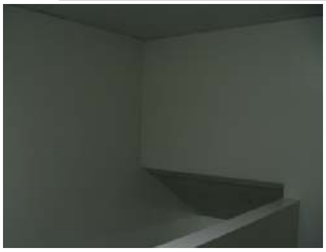




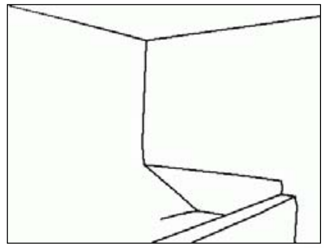


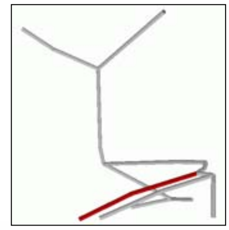


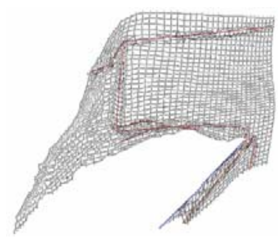
# textureless surfaces

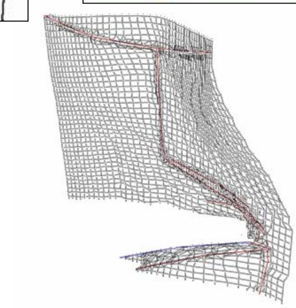


Original image





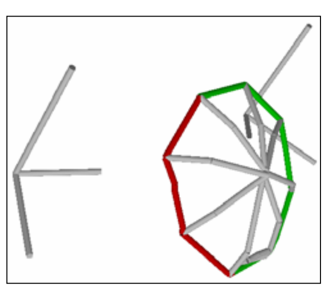


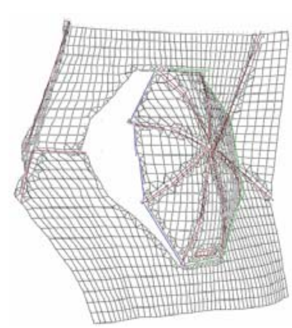




# Result on texture and textureless surfaces

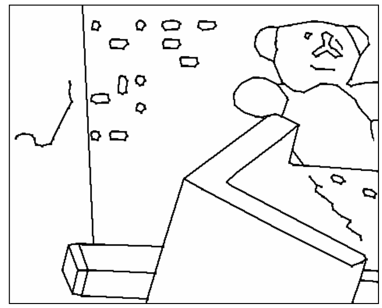
Original image
(Tomasi et al 04)





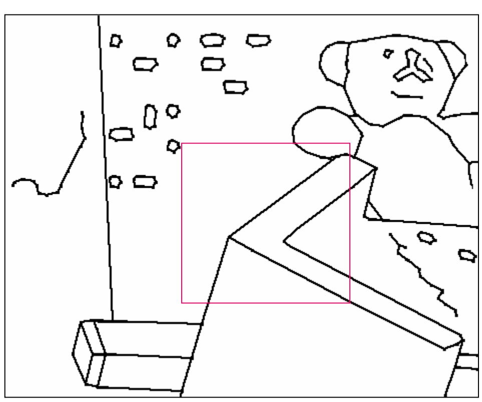
# A stereo image and its sketch

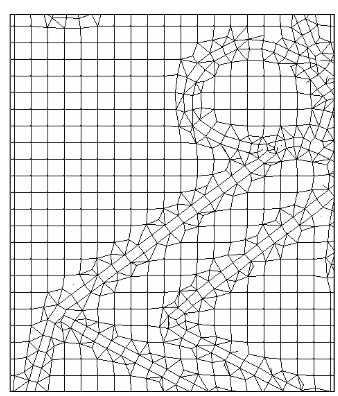
Original image
(Sziliski et al 02)





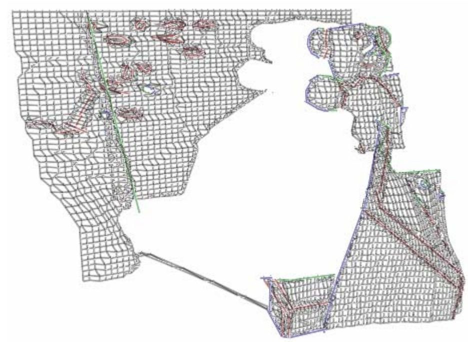
# Sketch and mesh





# More Results

Original image
(Sziliski et al 03)



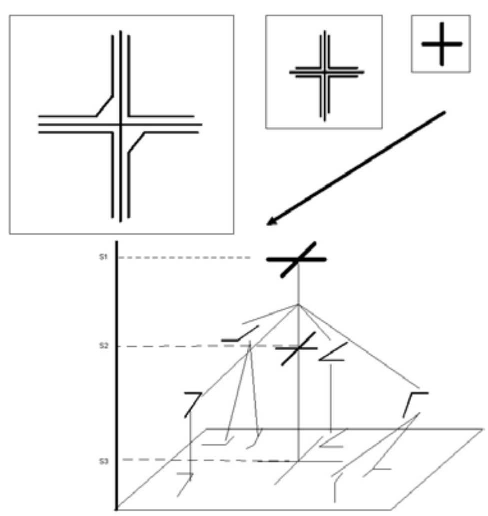


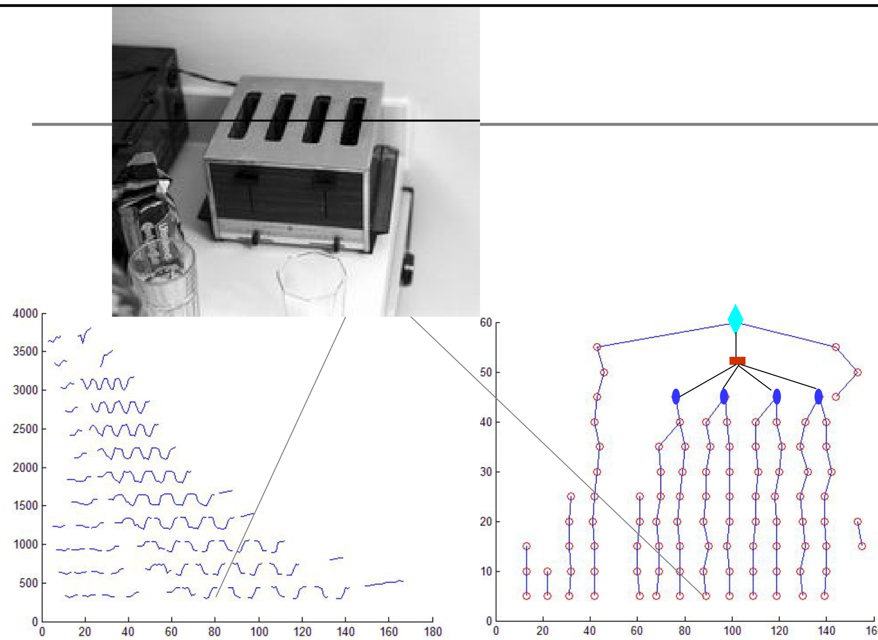
# Topologic changes over in scaleing

(Wang and Zhu 05)

The current scale-space theory is based on continuous Gaussian --Laplacian pyramids. While it is suitable for the retina and LGN, it is wrong for V1.

We need a new scale-space theory which is multi-layer of primal sketches

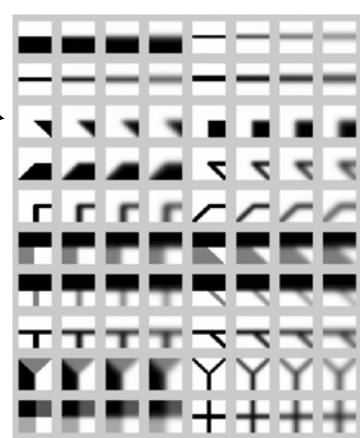






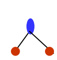

## What occurs in perception when up-scaling?

1. Image sharpening on boundaries



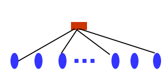
2. Mild jumps

   e.g. birth of a sketch, or split a bar to 2 edges
   ---- handled by graph grammar.



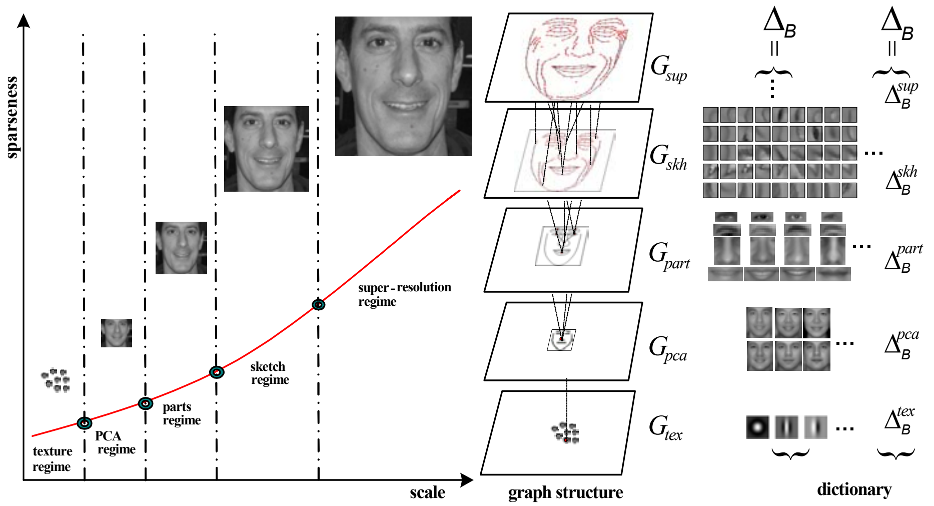
3. Catastrophic transition

   e.g. from texture to 100s primitives



# Scaling of Faces

(Xu, Chen, and Zhu, 05)



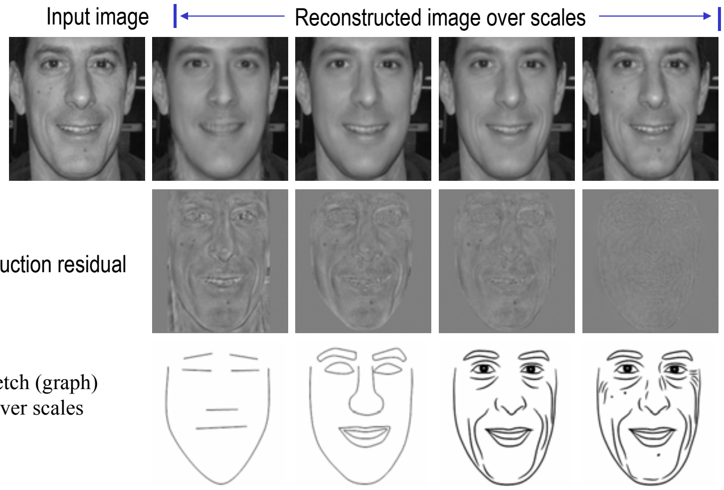



# Example of hierarchic graph of face

Input image | ← Reconstructed image over scales → |

Reconstruction residual

The sketch (graph) $G$ over scales

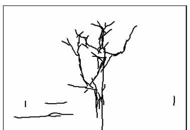


(Xu, Chen and Zhu, 2005)

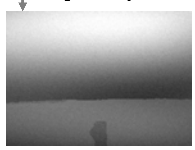
# from image parsing to 3D

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


input I


curve & tree layer


region layer




3D reconstruction and rendering



# from image parsing to 3D

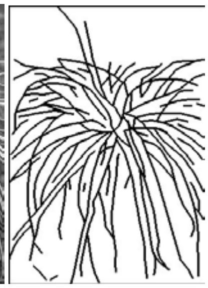
3D reconstruction (Han and Zhu, 2003)


input image


3D reconstruction from a single image

# from image parsing to 3D

Input image



sketch

Three new views