From Primal Sketch to 2½D 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)

\[
p(J, \Gamma) = \frac{1}{Z} \exp \left\{ -\alpha \int \int_{\Lambda \setminus \Gamma} |\nabla J(x, y)|^2 dx dy - \beta ||\Gamma|| \right\}
\]

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

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

Spatial MRF

Texture MRF

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

\[ \Lambda = \Lambda_{sk} \cup \Lambda_{n\text{sk}} \]

2. The sketchable part is divided into disjoint domains,

\[ \Lambda_{sk} = \bigcup_{k=1}^{K} \Lambda_{sk,k} \]

Each domain is covered by a patch from a dictionary \( \Delta_{sk} \)

\[ I_{\Lambda_{sk,k}}(u, v) = B_k(u, v) + n, \quad k = (\ell, x, y, \theta, \sigma, \alpha_{\text{phi}}, \alpha_{\text{wrp}}) \]

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

\[ S_{sk} = (K, \{ (\Lambda_{sk,k}, B_k) : k = 1, 2, \ldots, K \}) \]
The primal sketch model

3. The non-sketchable part is divided into homogeneous texture regions
\[ \Lambda_{nsk} = \bigcup_{i=1}^{n} \Lambda_{nsk}, \]

Each region has a statistical summary \( h_n \)
\[ S_{nsk} = (N, \{(\Lambda_{nsk,i}, h_i \leftrightarrow \beta_i): n = 1, 2, ..., N\}) \]

\[ p(I, S_{sk}, S_{nsk}, \Delta_{sk}, \Delta_{nsk}) = \frac{1}{Z} \exp\{-E_{sk}(S_{sk}) - E_{nsk}(S_{nsk}) \]
\[ - \sum_{k=1}^{K} \sum_{(x,y) \in \Lambda_{nsk,k}} (I(u,v) - B_{k}(x,y))^2 \]
\[ - \sum_{i=1}^{n} < \beta_i, h(I|\Lambda_{nsk,i}) > \]
# Reversible graph operators

<table>
<thead>
<tr>
<th>Operator</th>
<th>Graph Change</th>
<th>Illustration</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O_1$, $O_1'$</td>
<td>create / narrow a stroke</td>
<td><img src="image1" alt="Illustration" /></td>
</tr>
<tr>
<td>$O_2$, $O_2'$</td>
<td>grow / shrink a stroke</td>
<td><img src="image2" alt="Illustration" /></td>
</tr>
<tr>
<td>$O_3$, $O_3'$</td>
<td>connect / disconnect vertices</td>
<td><img src="image3" alt="Illustration" /></td>
</tr>
<tr>
<td>$O_4$, $O_4'$</td>
<td>extend one stroke and cross / disconnect and combine</td>
<td><img src="image4" alt="Illustration" /></td>
</tr>
<tr>
<td>$O_5$, $O_5'$</td>
<td>extend two strokes and cross / disconnect and combine</td>
<td><img src="image5" alt="Illustration" /></td>
</tr>
<tr>
<td>$O_6$, $O_6'$</td>
<td>combine two connected strokes / break a stroke</td>
<td><img src="image6" alt="Illustration" /></td>
</tr>
<tr>
<td>$O_7$, $O_7'$</td>
<td>combine two parallel strokes / split one into two parallel</td>
<td><img src="image7" alt="Illustration" /></td>
</tr>
<tr>
<td>$O_8$, $O_8'$</td>
<td>merge two vertices / split a vertex</td>
<td><img src="image8" alt="Illustration" /></td>
</tr>
<tr>
<td>$O_9$, $O_9'$</td>
<td>create / remove a blob</td>
<td><img src="image9" alt="Illustration" /></td>
</tr>
<tr>
<td>$O_{10}$, $O_{10}'$</td>
<td>switch between a stroke(s) and a blob</td>
<td><img src="image10" alt="Illustration" /></td>
</tr>
</tbody>
</table>

## More examples

![Vacuum Cleaner](image11)

![Diagram](image12)

![Image](image13)
More examples

Manifold learning and entropy minimization

Let $\Omega_{\text{nat}}$ 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 $\Omega_{\text{nat}}$ for a certain perceptual metric $\rho$.

$$\forall I \in \Omega_{\text{nat}}, \exists J \in \Omega_\varepsilon, \text{ so that } \rho(I, J) < \varepsilon.$$  

The minimum $\varepsilon$-cover has size $\mathcal{N}(\Omega_{\text{nat}}, \rho, \varepsilon)$

The $\varepsilon$-entropy of the natural image ensemble is

$$\mathcal{H}(\Omega_{\text{nat}}, \rho, \varepsilon) = \log \mathcal{N}(\Omega_{\text{nat}}, \rho, \varepsilon)$$

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,
\[ \Omega_{\text{gen}} = \{ I : \ I \ = \ g(W; \Delta_{\text{gen}}), \ W \in \Omega_W \} \]

\( W \) are the dimensions of the manifold \( \Omega_W \), geometric and photometric. The metric is the MSE,
\[ \rho_{\text{gen}}(I, J) = \frac{1}{|A|} \sum_{x,y} (I(x, y) - J(x, y))^2 \]

This ensemble has size \( \mathcal{M}(\Omega_{\text{gen}}, \rho_{\text{gen}}, \epsilon) \)
The \( \epsilon \)-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,
\[ \Omega_{\text{des}} = \{ I : \ h(I; \Delta_{\text{des}}) = h_o, \ h_o \in \Omega_h \} \]

\( h \) are the statistics/features extracted (projection of the image space). The metric is on the projected statistics,
\[ \rho_{\text{des}}(I, J) = ||h(I) - h(J)|| \]

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

We take clothes as example. (Han and Zhu 05)

(a) input  (b)folds graph $G$  (c)$f_{\text{est}}$  (d) Filling result

Related work: fold detection by SVM

"Shading Primitives"

(Forsyth 97)
Learning fold primitives

Three types of fold primitives

Model fold primitive profile by PCA

\[ I(x, y) = \eta \cdot \tilde{L} = R(p, q) = \frac{-pl_1 - ql_2 + t_3}{\sqrt{p^2 + q^2 + 1}} \]
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).
Experimental results

(Han and Zhu 05)

<table>
<thead>
<tr>
<th>Input image</th>
<th>folds</th>
<th>surface of folds</th>
<th>full surface</th>
<th>novel view</th>
</tr>
</thead>
</table>

MSRI workshop on low and middle level vision, Feb, 2005
Song-Chun Zhu
Experimental results

Comparison: without folds
Example on stereo vision

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
(Szoliski 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

(Xu, Chen and Zhu, 2005)
from image parsing to 3D

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

- Input image
- Curve & tree layer
- Region layer

3D reconstruction and rendering

3D reconstruction from a single image
from image parsing to 3D

Input image        sketch

Three new views

MSRI workshop on low and middle level vision, Feb, 2005
Song-Chun Zhu