Image Parsing

*Segmentation, Detection, and Recognition.*

Alan Yuille (UCLA).
Dept. Statistics and Psychology.
Thanks to D. Kersten (U. Minnesota)
Introduction: Mathematical Theories of Vision

- Want a Mathematical (Computational) Theory of Vision that:
  
  (i) *lets us to build* computer vision systems that *work in the real world.*
  
  (ii) *serves as an Ideal Observer model for evaluating biological vision.*
  
  (iii) motivates models of neural processing.
Introduction: Visual Realism.

- **Claim**: mathematical theories of vision need to model the visual environment.

- What are the ecological (Gibson) or natural constraints (Marr)?

- **Claim**: Designing a system that works with real images helps tell you what the real hard problems are.
Introduction: Image Parsing

- Task: take an input image and parse it into its constituent components.
- Components are objects (faces) and generic regions (shading, texture).
- Analogous to parsing a sentence “The cat sat on the mat” into nouns, verbs, etc. (precise connections later).
Introduction: Example

Input Image

Generic Regions

Letters/ Digits

Faces
Bayesian Inference: Expected Loss

- Parsing must estimate a representation $W^* \text{ (objects...)}$ from the image $I$.
- What is the best rule (algorithm) $d(.)$ to give solution $W^* = d(I)$?
- Pick rule $d(.)$ to minimize expected loss $R(d) = \sum P(W,I) \cdot L(W,d(I))$
- $L(W,d(I))$ is penalty for wrong answer.
- Depends on visual environment $P(W,I)$. 
Bayesian Inference: Generative Models.

- Best rule is select $W^*$ that maximizes $P(W,I)/P(I)$.

- Can express $P(W,I)/P(I)$ as (Bayes Rule):
  \[
  P(W,I)/P(I) = P(I|W) \frac{P(W)}{P(I)},
  \]

where:

(i) $P(I|W)$ is the probability of generating the image from $W$.
(ii) $P(W)$ is the prior on $W$. 

Bayesian Inference: Sinha’s Figure

Illustrates the use of:
$P(I|W)$ & $P(W)$
Bayesian Inference: Key Issues

(i) **Modeling**: How to model $P(I|W)$ and $P(W)$ for real images and scenes?

$P(I|W)$ is like computer graphics. But need mathematical models.

(ii) **Inference**: How to compute $W^*$?
Modeling: $P(I|W)$ & Generation.

- **Probabilistic Context Free Grammar (CFG).**
- Tree structure. Single node at top represents the entire image region.
- Probability of splitting a region into two.
- Probability of labeling a region – face, text, generic.
- Probability of generating intensity values in each regions.
- *(Prob. CFG’s used for speech & language).*
Modeling: Probabilistic CFG.

Full image Region

Probability of Split: Boundary of Split.

Region 1  Region 2

Region 3  Region 4

Probability of Region Label: Face, Text, Shading, etc.

Probability of Region Parameters given label.

Probability of Image of Region given label and Parameters.

- Generic Regions: (i) constant, (ii) clutter, (iii) texture, (iv) shading.

- Require models:
  \[ P(I(x,y) \mid \text{label, parameters}) \] (Tu & Zhu ’02).
  e.g. Gaussian for intensity in constant regions. parameters mean, variance.
  (Zhu & Yuille 1996)
Modeling: Synthesis from models

**Input:** region boundaries, region labels, region parameters.
Modeling: Synthesis of Objects

- Faces (front-on) and Text.
Modeling: $P(I|W)$ and $P(W)$.

Image decomposed into regions.

Probability of image is product of prob. of each region image.

\[
p(I|W) = \prod_{i=1}^{N} p(I_{R_i}|\theta_i, l_i)
\]

Region labels $l_i$,
Region parameters $\theta_i$

Also prior probabilities $P(W)$ for shapes of regions, parameters of face and text models.
Inference: Estimate $W^*$ from $I$

- Traditional models of vision are feedforward via intermediate level representations.
  \[ \text{Image} \rightarrow 2-1/2D \text{Sketch} \rightarrow \text{Objects}. \text{ (Marr)}. \]

- **Problem:** often very hard to construct these intermediate representations (on real images)

- **Claim:** intermediate level vision is ill-posed and ambiguous (hard to detect edges), but high level vision is well-posed (easy to detect faces).
Inference: Rapid Detection Faces/Text.

- There exist learning algorithms (e.g. Adaboost) that can be trained to detect faces and text in unconstrained images.

<table>
<thead>
<tr>
<th>Object</th>
<th>False Positive</th>
<th>False Negative</th>
<th>Images</th>
<th>Subwindows</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face</td>
<td>65</td>
<td>26</td>
<td>162</td>
<td>355,960,040</td>
</tr>
<tr>
<td>Face</td>
<td>918</td>
<td>14</td>
<td>162</td>
<td>355,960,040</td>
</tr>
<tr>
<td>Face</td>
<td>7542</td>
<td>1</td>
<td>162</td>
<td>355,960,040</td>
</tr>
<tr>
<td>Text</td>
<td>118</td>
<td>27</td>
<td>35</td>
<td>20,183,316</td>
</tr>
<tr>
<td>Text</td>
<td>1879</td>
<td>5</td>
<td>35</td>
<td>20,183,316</td>
</tr>
</tbody>
</table>

Table 1: Performance of AdaBoost at different thresholds.

- Error rate still too high: But much better than error rate for edges!
Inference: Feedforward/Feedback.

- Claim: low-level visual cues are ambiguous but fast. (Feedforward).
- High-level models are reliable but slow (Feedback).
- *High-level models needs to search over all parameters of models.*
- *Except – low-level cues for faces/text can be fast (AdaBoost).*
Inference: Generative Feedback

- Searching through high-level models can be done in a Bayesian spirit by "analysis through synthesis" Grenander/Mumford.

- Sample from the generative model $P(I|W)$ until you find the $W^*$ that best generates the image. Too slow!

- Mumford advocated this as a model for the brain – feedback connections.
Inference: DDMCMC

- Data Driven Markov Chain Monte Carlo (DDMCMC). Tu & Zhu.
- A fast way to do *Analysis by Synthesis*.
- *Feedforward*: low-level cues to propose high-level models (and model parameters).
- *Feedback*: high-level models generate the image and get validated.
- *Attraction*: Can prove that the DDMCMC will converge to best $W^*$. *But how fast?*
Inference: DDMCMC

- Search for $W^*$ by making moves in the solution space (split region, change label, etc. etc).

- Propose move with prob: $q(W \rightarrow W'|I)$

- Accept move with probability

\[
\alpha(W \rightarrow W') = \min(1, \frac{p(W'|I)}{p(W|I)} \cdot \frac{q(W' \rightarrow W|I)}{q(W \rightarrow W'|I)}).
\]

- The q’s are low-level cues (heuristics) which determine the speed of the algorithm but don’t affect the final answer.
Inference: propose/accept.

- “Man proposes, God disposes”.
  Sir Edwin H. Landseer R.A.
**Inference: DDMCMC & Segmentation.**

DDMCMC *using generic region models only is most effective way to segment images* (Tu/Zhu)

Evaluated on the Berkeley dataset. Ground truth from Berkeley students.

Errors often due to lack of knowledge of objects.
Inference: Image Parsing

- Use DDMCMC algorithm (feedforward and feedback).
- Generative models of generic regions and objects (faces, text).
- Proposals for faces and text from AdaBoost learning algorithm.
- Proposals for generic regions as for segmentation (edges, clustering, etc.)
Inference: Moves in Solution Space.
Feedforward/Feedback in Brain.

“High-level tells Low-Level to shut up”?  
Or “High-level tells Low-Level to stop gossiping”.

Kersten’ Lab.
Results: AdaBoost.

Boxes show faces & text detected by AdaBoost at fixed threshold.

Impossible to pick a threshold that gives no false positives/negatives on these two images.

Boxes show high probability proposals for faces & text.
The different region models can cooperate to explain the Image.

Generic “shaded region” processes detect the dark glasses, so the face model doesn’t need to “explain” that part of the data.

Advanced object models could allow for glasses.
Results: Scales, Cooperation.

Stop Sign.  Multiple scales.

Soccer Image.

Parking Image.  
**Glasses/Shaded.**  
9 detected as a generic region.  (cooperative).
Street: *Face model is used to reject fake AdaBoost candidates.*

Cooperativity – shadows on text explained as shaded regions.

Westwood: shaded region models needed to explain away glasses.
Summary: (I)

- **Image Parsing**: combines segmentation, detection, and recognition in a Bayesian framework.
- Feedforward proposals and feedback acceptance/rejection.
- *Non-traditional* – *no intermediate-level representation* (no data thrown away).
- Does this relate to the feedforward and feedback loops in the brain?
Summary II: Technical.


3. DDMCMC.

Summary III

- Are there limits to this approach?
- Can we add more objects, proposals, etc, and build a general purpose vision machine?
- *Need to study the visual environment and model it mathematically.*
- *Need to determine rapid search proposals (also environment driven).*
References.

- The DDMCMC algorithm appears in Tu, Zhu 2002.
- To obtain: go to visciences.ucla.edu/people & access Yuille and Zhu webpages.