Image Parsing

Segmentation, Detection, and Recognition.

Alan Yuille (UCLA).

Dept. Statistics and Psychology.

Work with A. Chen, Z.Tu and S.C. Zhu.

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

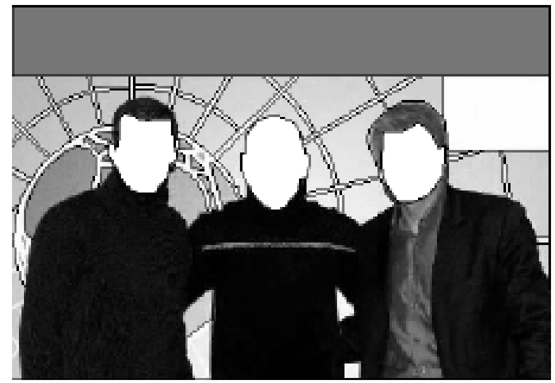
- Task: take an input image and *parse* it into its constituent components.
- Components are objects (faces) and generic regions (shading, texture).
- *Analogous to parsing a sentence "The cat sat on the mat" into nouns, verbs, etc.* (precise connections later).

Introduction: Example

Input Image

a. An example image

Generic Regions

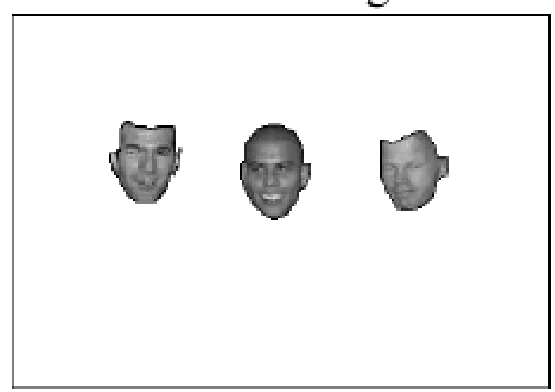


b. Generic regions

Letters/Digits

c. Text

Faces



d. Faces

Bayesian Inference: Expected Loss

- Parsing must estimate a representation W* (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) 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):

$$P(W,I)/P(I) = P(I|W)\,P(W)/P(I),$$

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)

Figure 2

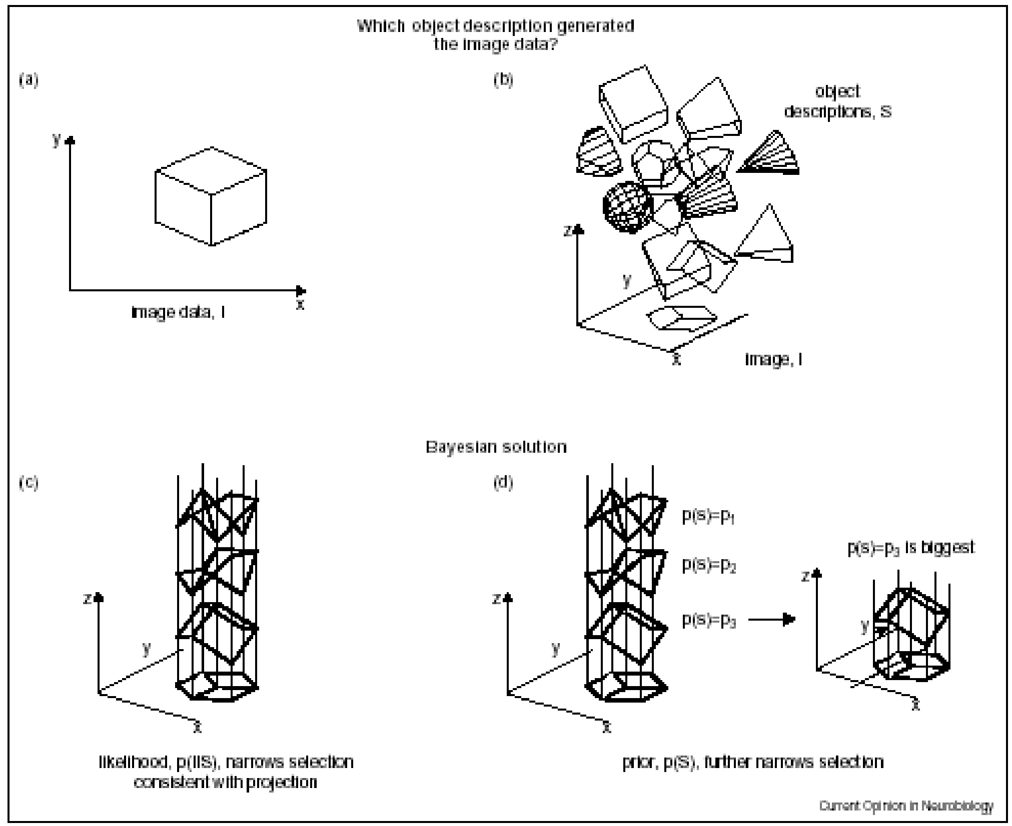


Illustration of Bayesian object perception **(a)** What 3D object caused the image of a cube? The likelihood (p(I|S)=p(image data|object descriptions)) constrains the possible set of objects to those consistent with the image data, but even this is an infinite set. **(b)** The prior knowledge probability (p(S)=p(object descriptions)) constrains the consistent set of 3D objects to those that are more probable in the world. **(c,d)** The probability over all instances is determined by the product of the likelihood and prior knowledge: p(S,I)=p(I|S)p(S)=p(object descriptions, image data). See the section on Basic Bayes for mathematical definitions. (Adapted from [64], © 1993 IEEE.)

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

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

Region 3 Region 4

Probability of Region Parameters given label.

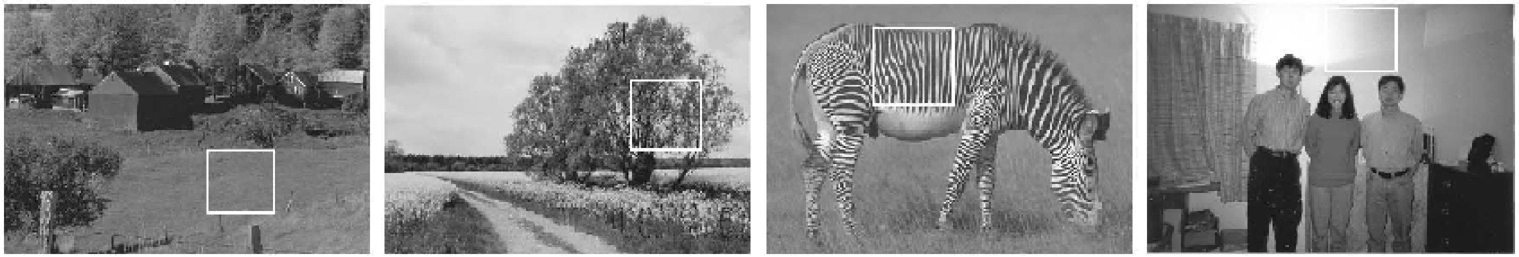
Region Label

Probability of Image of Region given label and Parameters.

Image of Region

Modeling: Prob. Images & Labels.

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



(a) (b) (c) (d)

- Require models:

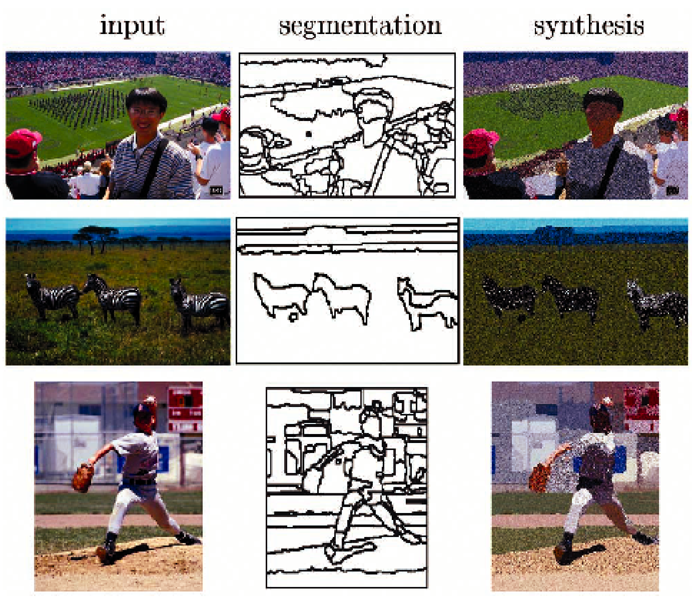
P(I(x,y) | label, parameters) (Tu & Zhu '02).

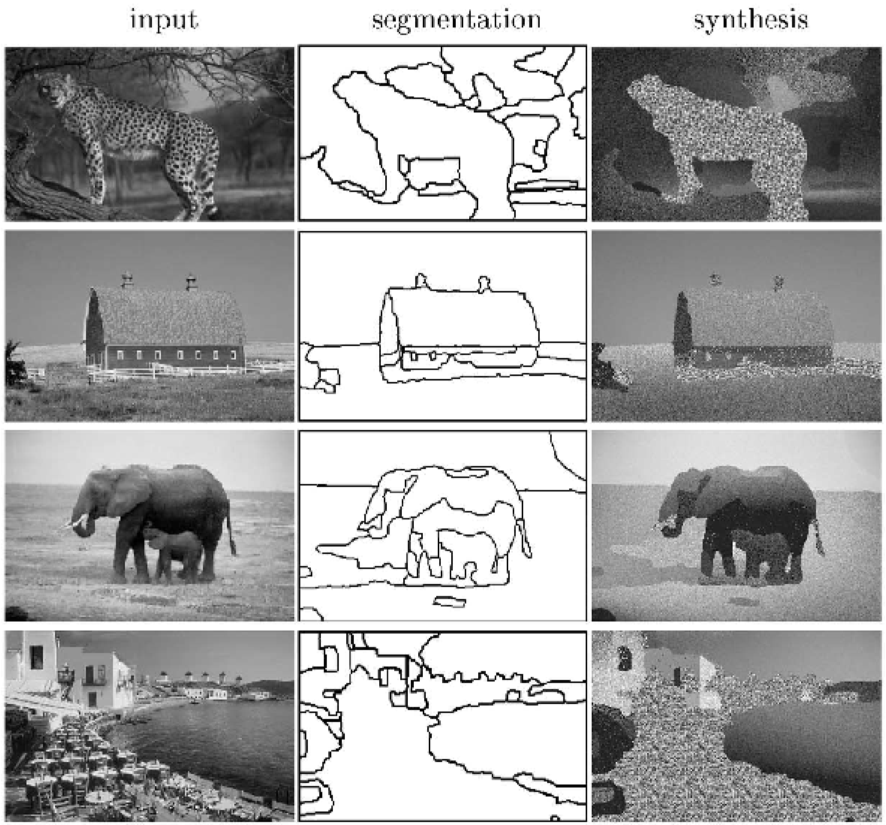
e.g. Gaussian for intensity in constant regions. parameters mean, variance.

(Zhu & Yuille 1996)

Modeling: Synthesis from models

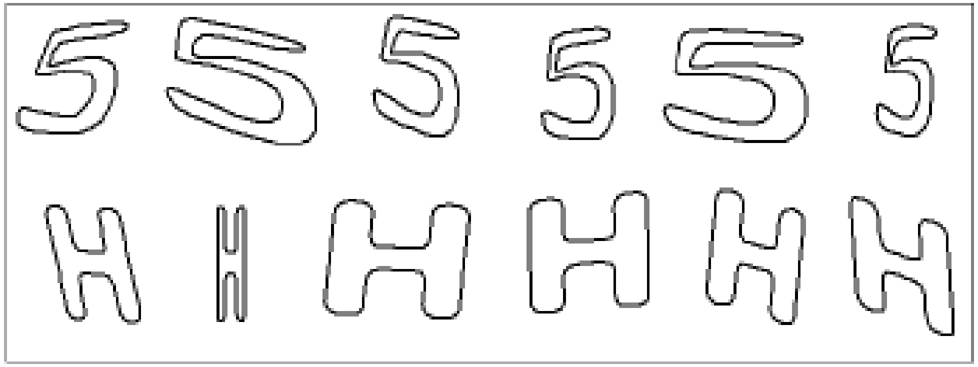
<u>Input: region boundaries,</u>
<u>Region labels,</u>
<u>Region parameters.</u>





Modeling: Synthesis of Objects

- Faces (front-on) and Text.



Modeling: P(I|W) and P(W).

Image decomposed Into regions.

$$\{R_i : i = 1, .., N\} \text{ s.t. } \cup_{i=1}^{N} R_i = R,$$

$$R_i \cap R_j = \emptyset \ \forall i \neq j \ \ \Gamma_i = \partial R_i$$

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

$$p(\mathbf{I}|W) = \prod_{i=1}^{N} p(\mathbf{I}_{R_i}|\theta_i, l_i)$$

Region labels l_i

Region parameters θ_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.

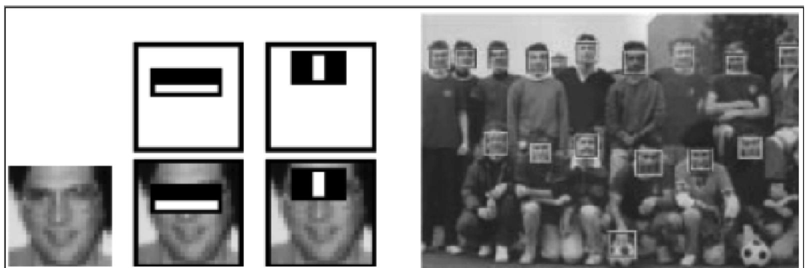
Image → 2-1/2D Sketch → Objects. (Marr).

- **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.



a. the first two face features

b. an example of face detection

Object	False Positive	False Negative	Images	Subwindows
Face	65	26	162	355,960,040
Face	918	14	162	355,960,040
Face	7542	1	162	355,960,040
Text	118	27	35	20,183,316
Text	1879	5	35	20,183,316

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: $q(W \to W'|\mathbf{I})$
- *Accept move* with probability

$$\alpha(W \to W') = \min(1, \frac{p(W'|\mathbf{I})}{p(W|\mathbf{I})} \cdot \frac{q(W' \to W|\mathbf{I})}{q(W \to W'|\mathbf{I})}).$$

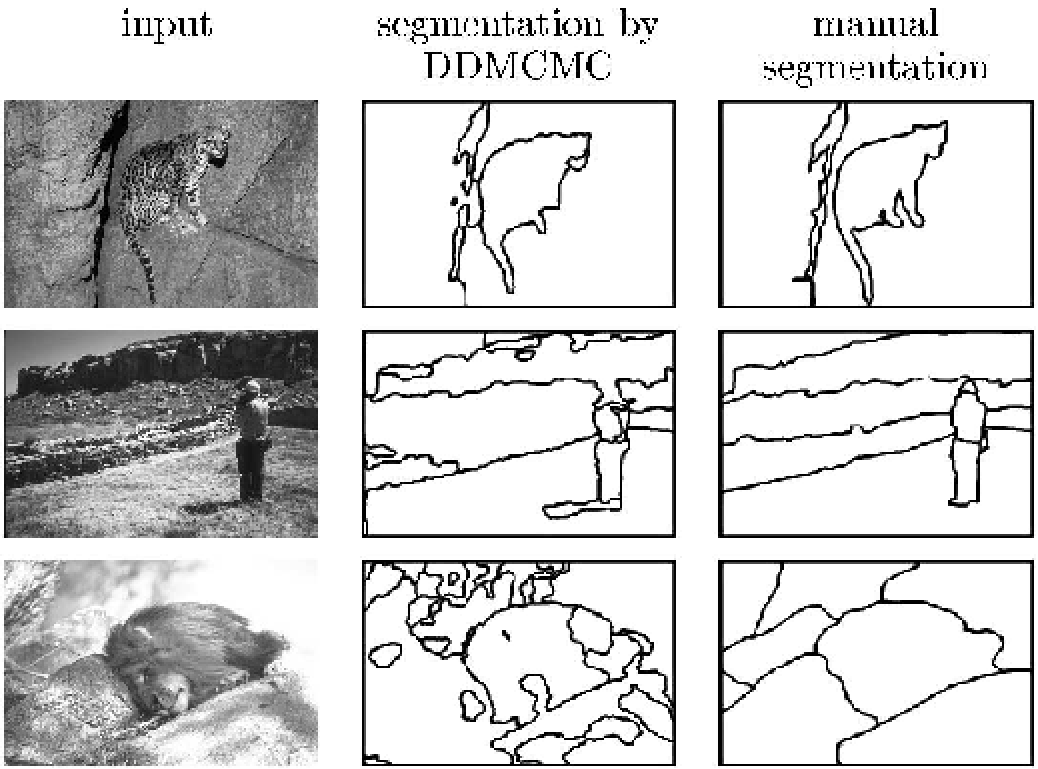
- The q's are low-level cues (heuristics) which *determine the speed* of the algorithm but *don't affect the final answer*.

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)

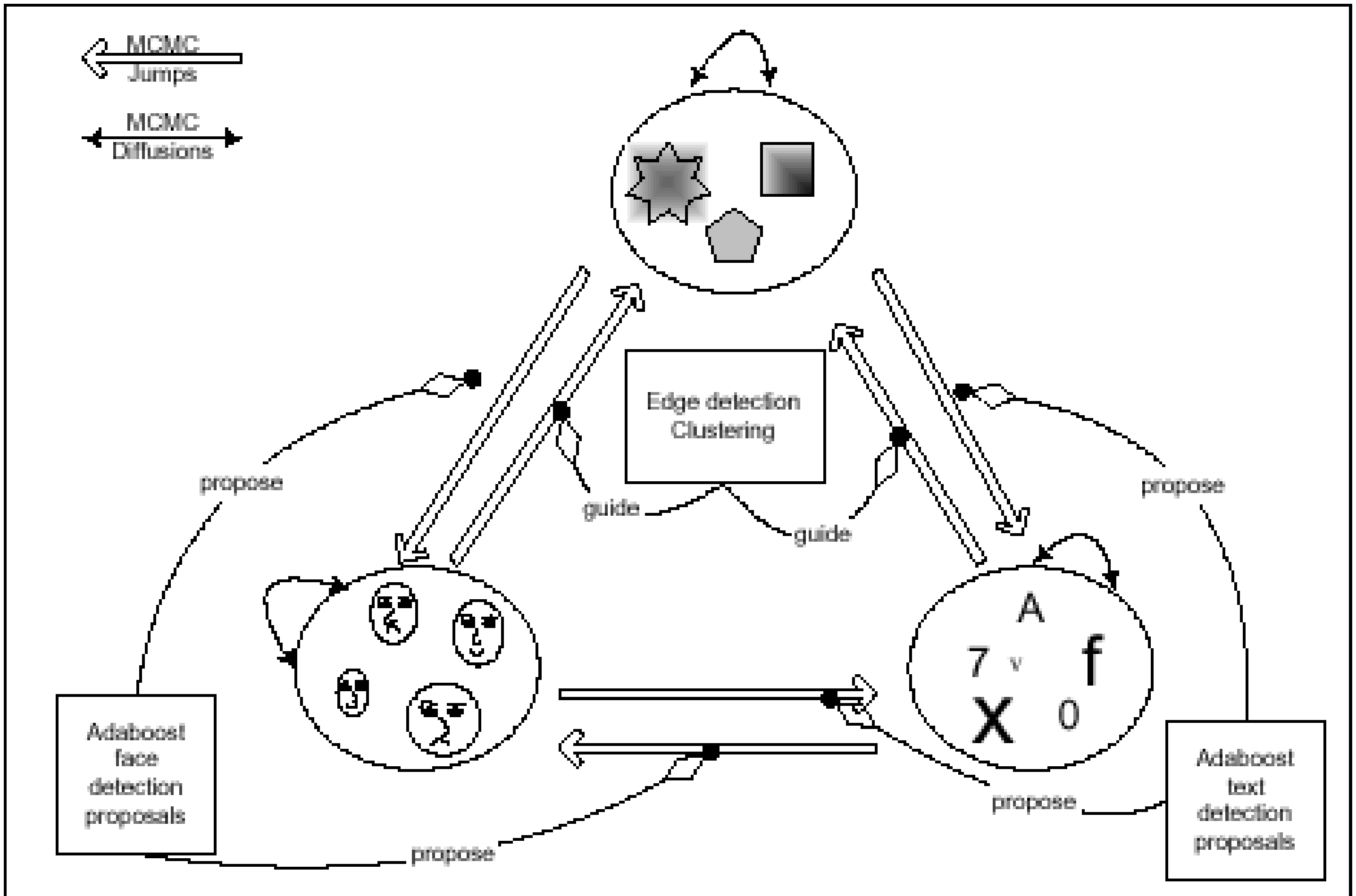
Evaluated on the Berkeley dataset. Ground truth from Berkeley students.

Errors often due to lack of knowledge of objects.

Inference: Image Parsing

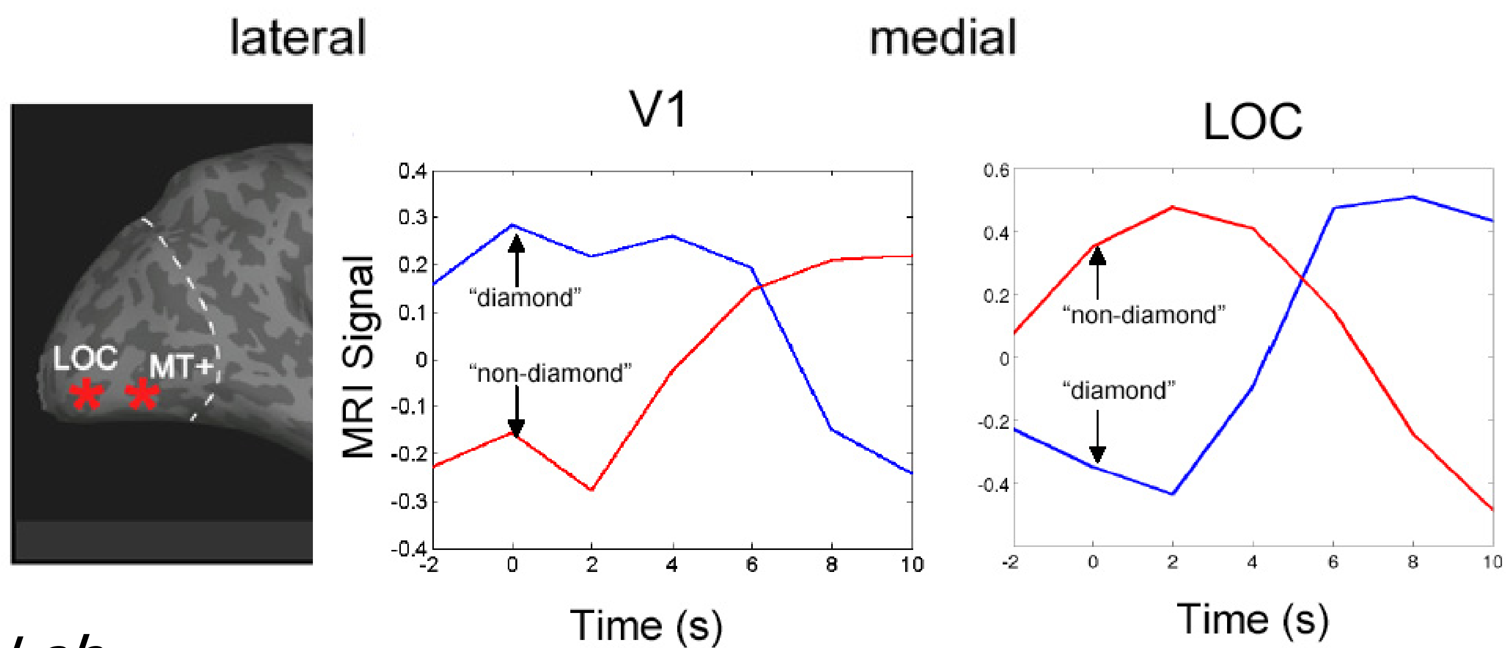
- Use DDMCMC algorithm (feedforward and feedback).
- Generative models of generic regions and objects (faces, text).
- Proposals for faces and text from AdaBoost learning algorithm.
- Proposals for generic regions as for segmentation (edges, clustering, etc.)

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".



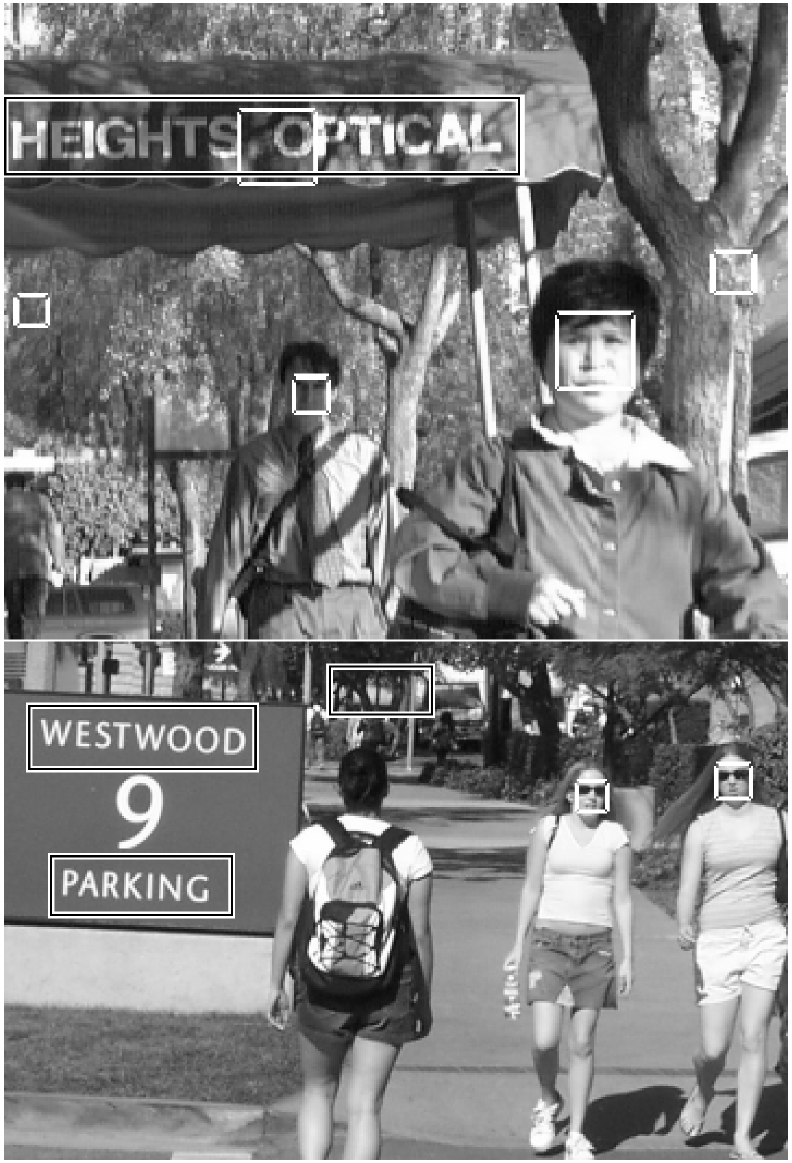
Kersten' Lab.

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.

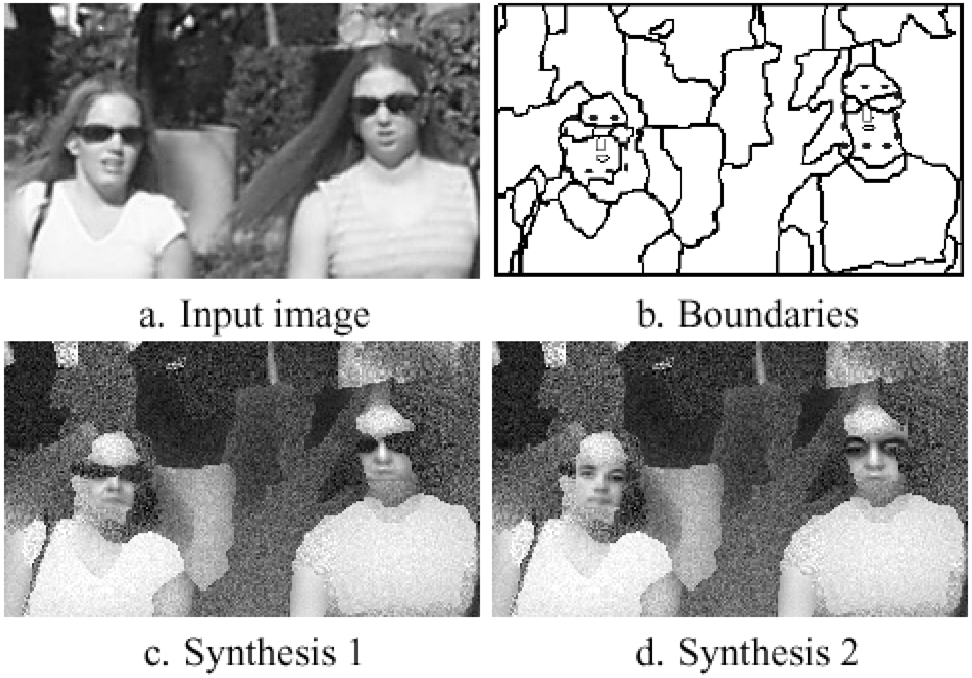


Results: cooperation/explain away

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.



a. Input image
b. Boundaries
c. Synthesis 1
d. Synthesis 2

Results: Scales, Cooperation.

Stop Sign.
Multiple scales.

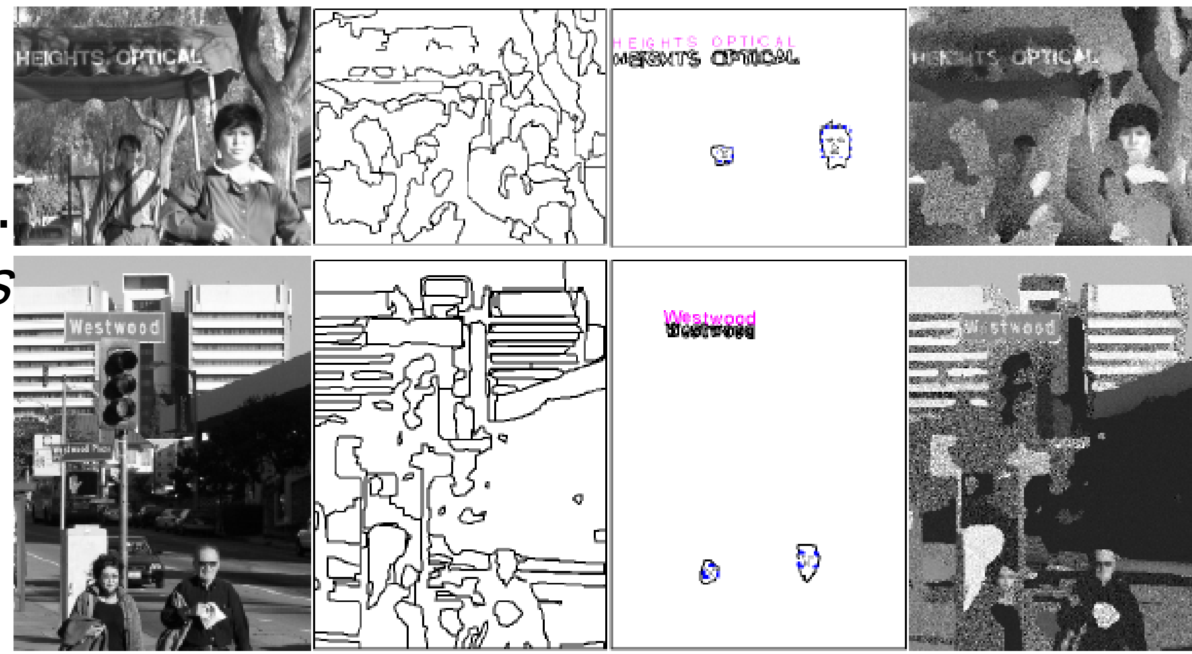
Soccer Image.

Parking Image.
Glasses/Shaded.
9 detected as a generic region. (cooperative).

Results. Reject and Explain away.

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.

- 1. Generative Models $P(I|W)$ (generic regions, faces, text…) and priors.

  Modeling the visual environment.
- 2. *Probabilistic Context Free Grammars.*
- 3. DDMCMC.
- 4. *Proposals – AdaBoost – smart heuristics.*

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.

- Image Parsing: Tu, Chen, Yuille, Zhu. International Conference on Computer Vision. 2003.
- Bayesian Theories of Object Recognition. Kersten, Mamassian, Yuille. Annual Review of Psychology. 2003 (to appear).
- The DDMCMC algorithm appears in Tu, Zhu 2002.
- To obtain: go to visciences.ucla.edu/people & access Yuille and Zhu webpages.