Image Parsing with Stochastic Grammar: LHI Dataset and Top-down/bottom-up Inference Scheme

Benjamin Yao, Xiong Yang and Tianfu Wu

Lotus hill institute of computer vision
Ezhou, China

Dept. of statistics, UCLA,
United States
Outline

- Introduction to the LHI (Lotus Hill Institute) dataset
  (See also [Benjamin Yao et al., EMMCVPR07] )
  - Features and discussions about the LHI dataset.
  - A few successful applications on the LHI dataset.

- Evaluating and scheduling Bottom-up and Top-down channels during inference of hierarchical model
  - To appear on ICCV09
What do you see in this picture?
i. Segmentation
i. Segmentation  ii. Semantic labels

- Sky
- Building
- Crane
- Traffic light/sign
- Lamp
- Palm trees
- Cars
- Road
i. Segmentation  
ii. Semantic labels  
iii. Object location
i. Segmentation  
ii. Semantic labels  
iii. Object location
i. Segmentation  ii. Semantic labels  iii. Object location
i. Segmentation  ii. Semantic labels  iii. Object location **iv. Recognize internal structure**

43 STREET  

(side view) car  

(rear view) Limo
i. Segmentation  
ii. Semantic labels  
iii. Object location  
iv. Recognize internal structure  

v. **3D geometry**

- Perspective ‘Parallel’ Lines
- Vanishing point
- 3D Coordinates
- Planes

Diagram showing building planes 1, 2, and 3, with ground plane and 3D coordinate axes (x, y, z).
i. Segmentation  
ii. Semantic labels  
iii. Object location  
iv. Recognize internal structure  
v. 3D geometry  
vi. Depth layer
i. Segmentation  ii. Semantic labels  iii. Object location  iv. Recognize internal structure  v. 3D geometry  vi. Depth layer  vii. Curve completion
i. Segmentation  ii. Semantic labels  iii. Object location  iv. Recognize internal structure  v. 3D geometry  vi. Depth layer  vii. Curve completion
Overview of functions and data structures

I. Segmentation
II. Semantic labels
III. Object location
IV. Sketch graph for representing internal structure
V. 3D geometry
VI. Depth layer
VII. Curve completion
VIII. Parsing
IX. And-Or graph
Overview of functions and data structures

I. Segmentation

II. Semantic labels

III. Object location

IV. Sketch graph for representing internal structure

V. 3D geometry

VI. Depth layer

VII. Curve completion

VIII. Parsing

IX. And-Or graph

Also in other datasets such as: LabelMe, Berkeley, MSRC, Caltech 101, PASCAL VOC

LHI dataset has better accuracy than others
Overview of functions and data structures

I. Segmentation
II. Semantic labels
III. Object location
IV. Sketch graph for representing internal structure
V. 3D geometry
VI. Depth layer
VII. Curve completion
VIII. Parsing
IX. And-Or graph

Novel designs that only appear in the LHI dataset
Segmentation and semantic annotation

Average time: 45 min/person

Average num of line-segment: 600/image
Aerial Images
Complex indoor scenes
Sketch Graph:
Object localization and internal structure
More examples
Sketch graph for face

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Expression</th>
<th>Lighting</th>
<th>Pose</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>![Age Image]</td>
<td>![Expression Images]</td>
<td>![Lighting Images]</td>
<td>![Pose Images]</td>
</tr>
</tbody>
</table>

The table above illustrates the differences in facial features due to age, expression, lighting, and pose.
Augmenting sketch graph w/ attributes

- Single-wing
- Surface (albedo)
- Boundary
- Shadow
- Bi-surface
Scene 3D Geometry

- Perspective ‘Parallel’ Lines
- Horizontal Lines
- 3D Coordinates
- Planes

Building Plane 1

Building Plane 2

Ground Plane
2.1 D Sketch Graph: Depth layer and curve completion
Parsing: parse tree and graph
Two more examples
Parse graph of street sign
And-Or graph: summarizing parse graphs

Parse graph 1  Parse graph 2  ...  Parse graph N

(a)
Data structure: And-Or graph compiled version of parse graphs.
And-Or graph summarizes all annotated examples
Statistics of And-Or graph

An interesting conclusion is that the And-Or graph is spindle-shaped. With smaller number of scene categories and primitives at the two ends, and large number of parts and objects in the middle.
Interactive Annotation Platform

Interface for Human Labeler

- Manual inputs
- Algorithm feedback
- Organization
- Top-down index

Interactive Image Parser

And-Or graph Data base
Video demo

- See a video demo of interactive parsing.
Top down graph match [Lin et al. CVPR07]
Please visit our website for more details
The LHI development and annotation team
Example Applications of LHI Dataset

- From Image Parsing to Painterly Rendering
  - [ M.Zhao et al., to appear on ACM Trans. on Graphics ]

- What, where and who? Classifying events by scene and object recognition
  - [ Fei-fei Li et al., ICCV07 ]

- Evaluating and scheduling Bottom-up and Top-down computing processes.
  - [ Tianfu Wu et al., to appear on ICCV09 ]

- and more....
From Image Parsing to Painterly Rendering

[ Kun Zeng, Mingtian Zhao and Song-chun Zhu, to appear on ACM Transaction on Graphics ]
What, where and who? Classifying events by scene and object recognition. [Fei-fei Li et al., ICCV07]

Eight categories
5,000 images in total
What, where and who? Classifying events by scene and object recognition

Input image

manual labeling

Import template

Templates of skeleton and face

And-Or graph knowledge base

Ground plane and horizon line

Object level segmentation and semantic annotation

Body/limb direction and landmarks on face

Output XML files, region mask and image patch of object
Annotation examples

Original Images
Segmentation
Skeleton & face
GroundPlane & horizon line
Evaluating and scheduling Bottom-up and Top-down computing processes.

Xiong Yang, Tianfu Wu, and Song-chun Zhu

Part of this work will appear in ICCV09
Outline

- Problem overview and motivation
- Evaluating the Information Contribution of bottom-up and top-down channels.
  - Low-middle level: junction detection
  - High level: human faces and cars
- A simple example of inference scheduling
Hierarchical modeling and $\alpha$, $\beta$ and $\gamma$ computing

1. Each node has its own $\alpha$, $\beta$ and $\gamma$ computing processes.

2. Bottom-up: $\alpha$, $\beta$, top-down: $\gamma$

3. All the three can contribute to recognize the node.

4. They contribute differently under different situations—how much?
Motivation

- **Modeling**: For object parsing, there are three categories of model in general
  - $\alpha$ – Flat models. [Viola & Jones, face detector], [Dalal & Triggs, pedestrian detector].
  - $\beta$ – Hierarchical compositional models [S. Geman et al., Composition machine] [Sinisa Tordorovic and Ahuja] [Zhu & Yuille Recursive compositional models] [Leonardis et al.] [Felzenswalb et al., Pictorial structure].
  - $\gamma$ – Contextual models [Top-down/bottom-up attribute grammar, Fen Hang and Song-chun] [Pietra et al., Mutual boosting]

- **Inference/Computing**: bottom-up v.s. top-down
  - Pure bottom-up feed-forward. [Riesenhuber & Poggio. Hierarchical models of object] [Geman et al., Composition machine]
  - Coarse to fine. [Felzenswalb et al., Generalized A*] [Iasonas and Yuille, HOP]
  - .....
In general: recursive $\alpha$, $\beta$ and $\gamma$ channels

1. Each node has its own $\alpha$, $\beta$ and $\gamma$ computing processes.

2. All the three can contribute to recognize the node.

1. On average, which channel is more informative for a given node?

2. When scale up to 100s object categories with 500s nodes, what are the optimal computing orders among different nodes?
Isolating $\alpha/\beta/\gamma$ channels by blocking

And-node  terminate-node

$\gamma$: channel

$\alpha$: channel

$\beta$: channel

C1  C2  C3  ...
Isolating $\alpha/\beta/\gamma$ channels by blocking
Isolating $\alpha/\beta/\gamma$ channels by blocking

- $\alpha$: channel
- $\beta$: channel
- $\gamma$: channel

And-node: C1 → C2 → C3 → …
Terminate-node: P
Example one: human faces

- Training $\alpha$, $\beta$ and $\gamma$ channels
Isolating $\alpha$-$\beta$-$\gamma$ channels: cropping and scaling

- Isolating $\alpha$-channel: only contain the compact image data for node A itself, and its parts can barely be recognized in isolation.
Isolating $\alpha$-$\beta$-$\gamma$ channels: cropping and scaling

- Isolating $\alpha$-channel: node A itself can not be recognized, but some of its parts can be recognized, say, their $\alpha$-channel is on.

- Isolating $\beta$-channel: node A itself can not be recognized, but some of its parts can be recognized, say, their $\alpha$-channel is on.
Isolating $\alpha$-$\beta$-$\gamma$ channels: cropping and scaling

- Isolating $\gamma$-channel: both node A and its parts can not be recognized, but some of parents or surrounding context can be recognized, say, their $\alpha$-channel is on.
Evaluating the information contribution of $\alpha/\beta/\gamma$

- Impurity function in CART decision tree.
Information contribution: $\alpha$ channels
Information contribution: $\beta$ channels

Combination of two children

Combination of three children

![Graphs showing information contribution](image-url)
As for face $\alpha$ channel is good enough

red for $\alpha$, blue for $\alpha+\beta$, green for $\alpha+\gamma$, cyan for $\alpha+\beta+\gamma$ channels
Example two: low-level primitives
Example two: low-level primitives
At low-middle level, $\beta$ channel dominate.

Red for $\alpha$, blue for $\beta$, green for $\gamma$.

(a) Information contributions (IC) evaluated for the five low-level primitives: $\alpha$ (in red), $\alpha+\beta$ (in blue), $\alpha+\beta+\gamma$ (in green).

Red for $\alpha$, blue for $\alpha+\beta$. 
Example three: car detection
For car detection, $\beta$ channel dominate.

Red for $\alpha$, blue for $\alpha+\beta$, green for $\alpha+\gamma$, cyan for $\alpha+\beta+\gamma$ channels.
An inference demo: How to integrate 3 channels?

Entails three kinds of computing processes (channels)

\[ \alpha \text{-channel} \quad \beta \text{-channel} \quad \gamma \text{-channel} \]
$\alpha$–channel: face
$\alpha$-channel: face
$\alpha$-channel: face
All $\alpha$ channels: things and stuffs

keep things and throw away stuffs by integrating $\alpha$, $\beta$ and $\gamma$ channels
Integrating $\alpha$, $\beta$ and $\gamma$ channels
Integrating $\alpha$, $\beta$ and $\gamma$ channels
Integrating $\alpha$, $\beta$ and $\gamma$ channels
Integrating $\alpha$, $\beta$ and $\gamma$ channels
Integrating $\alpha$, $\beta$ and $\gamma$ channels
Integrating $\alpha$, $\beta$ and $\gamma$ channels
Integrating $\alpha$, $\beta$ and $\gamma$ channels
Integrating $\alpha$, $\beta$ and $\gamma$ channels
Integrating $\alpha$, $\beta$, and $\gamma$ channels
Integrating $\alpha$, $\beta$ and $\gamma$ channels
Integrating $\alpha$, $\beta$ and $\gamma$ channels
Integrating $\alpha$, $\beta$ and $\gamma$ channels
Integrating $\alpha$, $\beta$ and $\gamma$ channels
Integrating $\alpha$, $\beta$ and $\gamma$ channels
Integrating $\alpha$, $\beta$ and $\gamma$ channels
Integrating $\alpha$, $\beta$ and $\gamma$ channels
Conclusion and future works

- Identify $\alpha$, $\beta$, and $\gamma$ channels in hierarchical models.
- Integrating $\alpha$, $\beta$, and $\gamma$ channels for performance improvements.
- At low-middle level, junctions often go through $\beta$ channels.
- At high level, human faces have a good $\alpha$ channel performance, cars is often better under $\beta$ channel.
- In future works, scheduling $\alpha$, $\beta$, and $\gamma$ channels for speeding up on-line computations, especially when scale up to a large scale vision system.
Thank you!