Formulating Structure for Vision Problems

Ziwei Liu

Department of Information Engineering
The Chinese University of Hong Kong
A mathematician is a person who can find analogies between theorems.

A better mathematician is one who can see analogies between proofs.

And the best mathematician can notice analogies between theories.

One can imagine that the ultimate mathematician is one who can see analogies between analogies.
Outline

• **Input Structure**: patch, image, video, multi-modality …

• **Model Structure**: information flow + regularization

• **Target Structure**: label, sequence, mask, multi-task …
Lessons Learned (Geometry + Semantics)

Random Thoughts (Inverse Thinking)

No Equations (Fast Forward)
Input Structure

• Case Study I —— low-level vision tasks
Input Structure

• Case Study I —— low-level vision tasks

noisy burst images  a clean image
Input Structure

• Case Study I —— low-level vision tasks

Data Alignment I: Geometry
Input Structure

• Case Study I —— low-level vision tasks

Data Alignment II:
Semantics
Input Structure

• Case Study I —— low-level vision tasks

High Dynamic Range

Auto Smiling
Input Structure

• Case Study I —— low-level vision tasks
Input Structure

• Case Study I —— low-level vision tasks

Graphics + Vision
Input Structure

- Case Study II —— high-level vision tasks

Data improvement? Model improvement?

X Zhu et al. Do We Need More Training Data? IJCV 2015

More data

We need both
Input Structure

• Case Study II —— high-level vision tasks
Input Structure

• Case Study II —— high-level vision tasks

Family of distributions of the form:

\[ f(x) = ax^k \]

• Frequency of tag words
• Content popularity

Most of the words are rare

Long tail

Limited vocabulary appears extremely large number of times

Power laws
Input Structure

• Case Study II — high-level vision tasks
  • User-generated content does not contain clean data
    – Non-visual texts / tags
    – Tags tend to have high precision, low recall
    – Frequency issue

• Hopefully, *large data-size* resolves issues
Input Structure

• Case Study II —— high-level vision tasks

Data Alignment I: Geometry

Data Alignment II: Semantics
Input Structure

- Case Study II —— high-level vision tasks
Input Structure

• Case Study II —— high-level vision tasks

Has-button

Similar Style Retrieval

Cloth Spotting in Video

Street-to-shop

Fashion Assistant
Model Structure

• Case Study —— facial attributes prediction
Model Structure

• Case Study I —— facial attributes prediction

Variance reduction
Model Structure

• Model Alignment I —— geometry

Attention to salient regions
Model Structure

- Model Alignment I ——— geometry

Pool features from salient regions
Model Structure

• Model Alignment I —— geometry

Transform features to canonical position
Model Structure

• Model Alignment II —— semantics

Abstract useful concepts
## Model Structure

- **Model Alignment II —— semantics**

<table>
<thead>
<tr>
<th>Test Image</th>
<th>Activations</th>
<th>Neurons</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td>Bangs, Brown Hair, Pale Skin</td>
<td>Narrow Eyes, High Cheek.</td>
</tr>
<tr>
<td><img src="image2.png" alt="Image" /></td>
<td>Eyeglasses, Mustache, Black Hair</td>
<td>Smiling, Big Nose</td>
</tr>
<tr>
<td><img src="image3.png" alt="Image" /></td>
<td>Wear. Hat, Blond Hair</td>
<td>Wear. Lipstick, Asian, Big Eyes</td>
</tr>
</tbody>
</table>

Combine to generalize
Model Structure

• Model Alignment II —— semantics

Squeeze to compress

Inverse Thinking
Model Structure

- Model Alignment II —— semantics

Model size comparisons

Running time comparisons
Target Structure

• Dependencies Among Target
Target Structure

• Message Passing

Classification  Localization  Segmentation
Target Structure

- Target Alignment I —— geometry

Location  Shape  Appearance
Target Structure

- Target Alignment II —— semantics

Group #1
- Mouth_Slightly_Open
- High_Cheekbones
- Smiling

Group #2
- Attractive
- Heavy_Makeup
- Bangs
- Wearing_Lipstick

Group #3
- Blond_Hair
- Gray_Hair
- Pale_Skin
- Blurry

Group #4
- Black_Hair
- Straight_Hair
- Wearing_Hat
- Eyeglasses

Group #5
- Male
- 5_o_Clock_Shadow
- Bald
- Sideburns

Group #6
- Wearing_Earrings
- Wearing_Necklace
- Wearing_Earrings
- Chubby_Double_Chip
- Arched_Eyebrows
- Narrow_Eyes
- Bags_Under_Eyes

Hierarchy
- Co-occurring
- Exclusive
- Unrelated

Semantics
- Young
- No_Beard
- Oval_Face
- Rosy_Cheeks
- Pointy_Nose
Target Structure

• Case Study I —— semantic segmentation
Target Structure

- Case Study I —— semantic segmentation
Target Structure

• Case Study I —— semantic segmentation

![Diagram of horse with different layers: Original Image, Ground Truth, Unary Term, Triple Penalty, Label Contexts, Joint Tuning]
Target Structure

- Original Image
- Ground Truth
- Unary Term
- Triple Penalty
- Label Contexts
- Joint Tuning
Target Structure

![Diagram showing the target structure with categories like background (bkg), area (areo), bike, bird, boat, bottle, bus, car, cat, chair, cow, table, dog, horse, mbike, person, plant, sheep, sofa, train, and tv, with a color gradient indicating favor and penalty.]
Target Structure

person : mbike

chair : person

favor

penalty
Target Structure

• Case Study II —— best pose for a selfie
Target Structure

• Case Study II —— best pose for a selfie
Reference

• Windows BLINK App
Reference

• SenseTime Fashion Eye
Reference

- More Details
Q & A