Scene Understanding
What more can we do with videos and text?

Karteek Alahari

Joint work with José Lezama, Guillaume Seguin, Ivan Laptev, Josef Sivic, Anand Mishra, C. V. Jawahar
Scene Understanding

Ladicky, Sturgess, Alahari, Russell, Torr, ECCV’10
Sturgess, Alahari, Ladicky, Torr, BMVC’09
Gould et al. NIPS’09
Li et al., CVPR’09
...
Scene Understanding: Are we done?
Scene Understanding: Are we done?
Scene Understanding: Are we done?
Scene Understanding: Are we done?

Judd et al., ICCV’09
Google Ground Truth

Source: BBC / Google
Scene Understanding: Are we done?

Hollywood dataset: Laptev et al., '08
Recognizing Street Text
Challenges
Simplifying the Problem

• Given a set of possible words

• Two datasets
  – Street View Text (SVT) dataset [Wang et al. ECCV’10, ICCV’11]
  – ICDAR ‘03 dataset
Our Approach

• Bottom-up cues
  – Individual character detections

• Top-down cues
  – Lexicon

Mishra, Alahari, Jawahar, CVPR’12
Character Detection

- Sliding window based
- HOG features + SVM

- Pruning based on
  - classifier confidence
  - character aspect ratio
Character Detection

- Not perfect

- Consider all detections and infer the *best* word
Recognizing Words

- Label set: characters + void/null

Inspired by Desai et al., ICCV’09
Recognizing Words

- Unary: From SVMs
- Pairwise: Occurrence of a pair of characters (~ language model)
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>SVT-WORD</th>
<th>ICDAR(50)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PICT</td>
<td>59.0*</td>
<td>_</td>
</tr>
<tr>
<td>PLEX + ICDAR</td>
<td>56</td>
<td>72</td>
</tr>
<tr>
<td>ABBYY</td>
<td>35</td>
<td>56</td>
</tr>
<tr>
<td>Proposed (Bi-gram)</td>
<td>70.03</td>
<td>76.96</td>
</tr>
<tr>
<td>Proposed (Node-specific)</td>
<td><strong>73.26</strong></td>
<td><strong>81.78</strong></td>
</tr>
</tbody>
</table>

PLEX + ICDAR: Wang, Babenko & Belongie, ICCV’11  
PICT: Wang & Belongie, ECCV’10
Results
Results

• Bi-gram vs Node-specific lexicon prior
Results: The Less Good Ones

• False negative detections

• Large rotation is not well-handled
Lexicon-driven v/s Lexicon-free Recognition

- Eg. application
  - Assist a visually impaired person to navigate in a grocery store

Lexicons = Grocery item list

Mishra, Alahari, Jawahar, CVPR’12
Wang et al., ICCV’11
Wang & Belongie, ECCV’10
Lexicon-driven v/s Lexicon-free Recognition

Recognize a cropped word

- Eg. application
  - Unconstrained word recognition

Mishra, Alahari, Jawahar, BMVC’12 (oral)
Lexicon free Recognition

<table>
<thead>
<tr>
<th>Datasets</th>
<th>ABBYY9.0</th>
<th>Pair-wise</th>
<th>Higher Order(=4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVT-Word</td>
<td>32.6</td>
<td>23.49</td>
<td>49.46</td>
</tr>
<tr>
<td>ICDAR2003</td>
<td>52</td>
<td>45</td>
<td>57.92</td>
</tr>
<tr>
<td>IIIT 5K-Word</td>
<td>14.60</td>
<td>20.25</td>
<td>43.30</td>
</tr>
</tbody>
</table>

- 0.5 Million dictionary words were used to compute priors

Mishra, Alahari, Jawahar, BMVC’12 (oral)
<table>
<thead>
<tr>
<th></th>
<th>Unary</th>
<th>Pair-wise</th>
<th>Higher Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOUR</td>
<td>YOUK</td>
<td>YOUK</td>
<td>YOUR</td>
</tr>
<tr>
<td>twilight</td>
<td>TWI1OHT</td>
<td>TWILIOHT</td>
<td>TWILIGHT</td>
</tr>
<tr>
<td>RESIST</td>
<td>KE5I5T</td>
<td>KESIST</td>
<td>RESIST</td>
</tr>
<tr>
<td>DOLCE</td>
<td>DOLCE</td>
<td>DOLCE</td>
<td>DOLCE</td>
</tr>
<tr>
<td>Beer</td>
<td>BEE1</td>
<td>BEEI</td>
<td>BEER</td>
</tr>
<tr>
<td>Srishti</td>
<td>SRISNTI</td>
<td>SRISNTI</td>
<td>SRISHTI</td>
</tr>
</tbody>
</table>
Video Scene Understanding
An Example

Hollywood dataset: Laptev et al., ‘08
Aim: Space-time video (over-) segmentation

• Consistent with objects

• Respect motion trajectories

• Associate object pixels over time

• Provide a low-level representation as building block for other tasks
Aim: Space-time video (over-) segmentation

- Consistent with objects
- Respect motion trajectories
- Associate object pixels over time
- Provide a low-level representation as building block for other tasks
Point-tracks to capture long-range motion

[Brox and Malik, ECCV’10]
Long Range Cues – Motion Similarity

[Brox & Malik ‘10]
[Costeira & Kanade ‘98]
[Shi & Malik ‘98]
[Sivic et al. ‘06]
...
Track clustering based on motion similarity
Long Range Cues – Depth Ordering (occlusion)
Long Range Cues – Depth Ordering (occlusion)

1. B occludes A, i.e. B is in front of A.
2. C occludes B, i.e. C is in front of B.

Hence:
3. C is in front of A (note that C has identical motion to A)
Clustering tracks with depth-ordering constraints

Formulate clustering as energy minimization:

\[
E(x) = \sum_{(i,j) \in \mathcal{E}} \left[ \alpha_{ij} \phi_1(x_i, x_j) + (1 - \alpha_{ij}) \phi_2(x_i, x_j) + \gamma_{ij} \phi_3(x_i, x_j) \right], \quad (1)
\]

Each track \(X_i\) takes a label \(x_i \in \{1, 2, \ldots, c\}\).

\(\mathcal{E}\) is the set of pairs of interacting tracks.

\(E(x)\) is the cost of label assignment.
Clustering tracks with depth-ordering constraints

\[
E(x) = \sum_{(i,j) \in \mathcal{E}} \left[ \alpha_{ij} \phi_1(x_i, x_j) + (1 - \alpha_{ij}) \phi_2(x_i, x_j) + \gamma_{ij} \phi_3(x_i, x_j) \right], \quad (1)
\]

Merges two tracks

\[
\phi_1(x_i, x_j) = \begin{cases} 
0 & \text{if } x_i = x_j, \\
1 & \text{otherwise.}
\end{cases}
\]
Clustering tracks with depth-ordering constraints

\[ E(x) = \sum_{(i,j) \in \mathcal{E}} \left[ \alpha_{ij} \phi_1(x_i, x_j) + (1 - \alpha_{ij}) \phi_2(x_i, x_j) + \gamma_{ij} \phi_3(x_i, x_j) \right], \quad (1) \]

Separates two tracks

\[ \phi_2(x_i, x_j) = \begin{cases} 
1 & \text{if } x_i = x_j, \\
0 & \text{otherwise.} 
\end{cases} \]
Clustering tracks with depth-ordering constraints

\[ E(x) = \sum_{(i,j) \in \mathcal{E}} \left[ \alpha_{ij} \phi_1(x_i, x_j) + (1 - \alpha_{ij}) \phi_2(x_i, x_j) + \gamma_{ij} \phi_3(x_i, x_j) \right], \quad (1) \]

\[ \alpha_{ij} = \exp \left( -\frac{(1 + \|a_i - a_j\|_2^2)\|v_i - v_j\|_2^2}{2l_{ij}\sigma_s^2} \right) \]
Clustering tracks with depth-ordering constraints

$$E(x) = \sum_{(i,j) \in \mathcal{E}} \left[ \alpha_{ij} \phi_1(x_i, x_j) + (1 - \alpha_{ij}) \phi_2(x_i, x_j) \right. $$

$$+ \left. \gamma_{ij} \phi_3(x_i, x_j) \right], \quad (1)$$

Captures relative depth-ordering of tracks

$$\gamma_{ij} \text{ is high if } i \text{ is occluded by } j$$
$$1 \ldots \text{ if } i \text{ is occluded by (is behind) } j$$
$$0 \ldots \text{ if } i \text{ occludes (is in front of) } j$$
Measuring occlusion

\[ \gamma_{ij} = 1 - \exp \left( -\frac{d||v_i - v_j||^2}{\sigma_0^2} \right) \]

\[ d = \exp(-D^2/\sigma_d^2) \]
Visualization of the occlusion score
The Example
Results

- Depth Order - Track Clustering

cluster 1 (front)
cluster 2
cluster 3
cluster 4
cluster 5 (back)

- Brox and Malik - Track Clustering
Video segmentation with long-range motion cues

• Build on the work of [Felzenszwalb & Huttenlocher 2004]
Video segmentation with long-range motion cues

- Build on the work of [Felzenszwalb & Huttenlocher 2004]
Results
Summary

• Partial ordering of point tracks
• Video over-segmentation consistent over many frames

• Object category video segmentation
• Parameter learning and better optimization
Merci!
Layered Segmentation: Ongoing work