Top-down and Bottom-up Cues for Scene Text Recognition

Anand Mishra1, Karteek Alahari2 and C.V. Jawahar1

1CVIT, IIIT Hyderabad, India 2INRIA-WILLOW, ENS, Paris, France

The Goal
- Recognize a cropped word
- LeXico
- CAPOGRO

Datasets
- Sign Evaluation (Weinman et al. PAMI’09)
- ICDAR 2003
- Street View Text [Wang et al., ECCV’10]
  - State-of-the-art commercial OCR: low accuracy
  - Sign Evaluation (60.5%), ICDAR (56%), Street View Text (35%)

Challenges
- Inter and intra character confusion
- Large number of classes
- Poor isolated character recognition

Need strong cues

Top-down and Bottom-up cues

- Top-down: Prior computed from lexicon
- Bottom-up: Sliding window based character detections
- The CRF model infers the true characters and the word as a whole.

Method Overview

Character detection
- Sliding window
- SVM classifier trained on ICDAR’03
- HoG features
- Some windows are pruned based on aspect ratio

Graph construction
- Character windows = nodes
- Unary cost = 1 - f(SVM Score)
- Pairwise cost = Lexicon + overlap based

The CRF Energy

- Set of labels k = \{0, 1, ..., 9, a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z, \}
- Minimize an energy of following form:
  \[ E(x) = \sum_{i=1}^{n} E_i(x_i) + E(x_i) \]

Unary cost: \( E_i(x_i = c) = 1 - P(v|x_i) \)

Unary cost of \( v \) is computed from SVM score and aspect ratio prior.

Pairwise cost:  
  Lexicon based: \( E_i(x_i = c_i, x_j = c_j) = \lambda (1 - P(v|x_i)) \)
  Overlap based: \( E_i(x_i = c_i, x_j = c_j) = \lambda \exp(-100 - \text{overlap}(x_i, x_j)) \)

Prior Computation

Toy examples: 28-gon prior vs node specific prior

\[
\begin{align*}
P(CV) &= 1/2 \quad 1/2 \quad 1/6 \quad 1/6 \\
P(I) &= 1/2 \quad 1/6 \quad 1/6 \quad 1/6 \\
P(CP) &= 1/6 \quad 1/6 \quad 1/6 \quad 16/0 \\
P(PR) &= 1/3 \quad 1/3 \quad 1/3 \quad 1/3 \\
\end{align*}
\]

Implementation Details

- Descriptor: Dense HOG, cell size = 4 x 4, bins = 10 bins, after resizing image to a 25x20.
- Influence: Tree-reweighted message passing (TRW-S) [Koegogorov, TPAMI’06].
- The method is is applicable to near frontal text datasets like Sign Evaluation data too.

Results

<table>
<thead>
<tr>
<th>Method</th>
<th>SVT-Word</th>
<th>ICDAR’05</th>
</tr>
</thead>
<tbody>
<tr>
<td>PICT</td>
<td>59</td>
<td>-</td>
</tr>
<tr>
<td>PLEX+WICAR</td>
<td>56</td>
<td>72</td>
</tr>
<tr>
<td>ARBYY  9.0</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>73.26</td>
<td>81.78</td>
</tr>
</tbody>
</table>

Many Applications
- Multi-media indexing
- Mobile apps
- Auto navigation
- Help for visually impaired

Center for Visual Information Technology
International Institute of Information Technology, Hyderabad, INDIA
http://cviit.iiit.ac.in/projects/SceneTextUnderstanding

Travel grant supported by Google Inc. India