Scene Text Recognition using Higher Order Language Priors

Anand Mishra\textsuperscript{1}, Karteek Alahari\textsuperscript{2} and C. V. Jawahar\textsuperscript{1}

\textsuperscript{1}IIIT Hyderabad, India
\textsuperscript{2}INRIA-WILLOW/ENS, France
Natural Scene Text: Why?

Text is everywhere!!

Many fundamental problems: Detection, Segmentation and Recognition

Many applications: mobile apps, auto navigations, multimedia indexing etc.
Natural Scene Text: Recent Interest

Detecting Text in Image

Stroke Width Transform based text detection
[Epshtein et al., CVPR’10]

End-to-end Scene Text Recognition
[Wang and Belongie, ICCV’ 11]

Real time localization and recognition
[Neumann and Matas, CVPR’12]
Natural Scene Text: Recent Interest

Text Recognition

- Exemplar Driven Character Recognition in the Wild [Sheshadri and Divvala, BMVC’12]
- PLEX and PICT [Wang and Belongie, ECCV’10, ICCV’11]
- Top-down and Bottom-up cues [Mishra et al., CVPR’12]
Scene Text Recognition

- Scene Text Recognition ≠ Optical Character Recognition (OCR)

- Good segmentation is tough

- Isolated character recognition accuracies are very low

- Not practical as a Classification problem
Scene Text Recognition

- Scene Text Recognition ≠ Optical Character Recognition (OCR)
- Good segmentation is tough
- Isolated character recognition accuracies are very low
- Not practical as a Classification problem

We need a better model
The Probabilistic Model

- Do not rely on “hard” segmentation
- HoG features
- Multi-class SVM trained on character level
- Sliding window based character detection
The Probabilistic Model
The Probabilistic Model

\[ L = \{0, 1, \ldots, 9, a, b, \ldots z, A, B, \ldots, Z, \epsilon\} \]
The Probabilistic Model

\[ L = \{0, 1, \ldots, 9, a, b, \ldots, z, A, B, \ldots, Z, \epsilon\} \]
The Probabilistic Model

$L = \{0, 1, \ldots, 9, a, b, \ldots z, A, B, \ldots, Z, \epsilon\}$
The Probabilistic Model

\[ L = \{0, 1, \ldots, 9, a, b, \ldots, z, A, B, \ldots, Z, \varepsilon\} \]

- Many-many labelings possible, which is the optimal?
- In general, the problem is NP-Hard
- We solve the approximate version of the problem in an energy minimization framework.

\[ E(x) = E_i(x_i) + E_{ij}(x_i, x_j) + E_{ij \ldots p}(x_i, x_j, \ldots, x_p) \]
The CRF Energy

- The unary term

\[ E_i(x_i = c_j) = 1 - P(c_j | x_i) \]

- The pair-wise term

  Lexicon based:

\[ E_{ij}(x_i = c_i, x_j = c_j) = \lambda_l (1 - P(c_i, c_j)) \]

  Overlap based:

\[ E_{ij}(x_i, x_j) = \lambda_o \exp(-100 - \text{overlap}(x_i, x_j)) \]

The cost of two highly overlapping nodes taking non-null label is high.
The CRF Energy

- Higher Order term
  = Unary + Pairwise

- Joint probability space of character sets occurring together is sparse

- Unary cost
  \[ E^a(x_i = L_i) = \lambda_a(1 - p(L_i)) \]

- Pairwise cost to prevent non-dictionary \( n \)-grams
Prior Computation

Pair-wise Priors
- Bi-gram Priors
  - Joint Probability \( P(c_i, c_j) \)
- Node Specific Priors
  - Joint Probability based on spatial position

Higher Order Priors
- \( n \)-gram Priors
  - Joint Probability \( P(c_i, c_j, \ldots, c_n) \)

Bigram Probability distribution

Tri-grams

THE AND ING DRO QUG

LOW

HIGH
Inference

- Minimize energy of following form:

\[
E(x) = E_i(x_i) + E_{ij}(x_i, x_j) + E_{ij...p}(x_i, x_j, ..., x_p)
\]

- We use Tree Re-weighted Message Passing (TRW-S) to minimize the energy

\$ Kolmogorov, Convergent Tree-Reweighted Message Passing for Energy Minimization, TPAMI'06 \$
Lexicon driven v/s Lexicon free Recognition

- Many Applications
  - e.g. assisting visually impaired person to navigate in a Grocery store
Lexicon driven v/s Lexicon free Recognition

- Many Applications
  - e.g. unconstrained word recognition
  - Word may or may not belong to dictionary
IIIT 5K-word dataset

- 5000 words: Street View and born-digital images
- At-least 6 times large than popular public datasets
- Wide Variations
- Up to character boundary level annotation
IIIT 5K-word dataset

- Train set: 2000 words, Test set: 3000 words
- Collected from total 1120 scene/born-digital images
- Grouped into easy/hard
## Lexicon driven Recognition

<table>
<thead>
<tr>
<th>Method</th>
<th>SVT-Word</th>
<th>ICDAR(50)</th>
<th>IIIT 5K word</th>
</tr>
</thead>
<tbody>
<tr>
<td>PICT [ECCV’10]</td>
<td>59</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PLEX [ICCV’11]</td>
<td>57</td>
<td>72</td>
<td>-</td>
</tr>
<tr>
<td>ABBYY 9.0</td>
<td>35</td>
<td>56</td>
<td>14.60</td>
</tr>
<tr>
<td>Proposed$ (Pair-wise)</td>
<td>73.26</td>
<td>81.78</td>
<td>66</td>
</tr>
<tr>
<td>Proposed (Higher Order)</td>
<td>73.27</td>
<td>80.28</td>
<td>68.25</td>
</tr>
</tbody>
</table>

Smaller lexicon: stronger context
Pairwise priors are powerful
Edit distance based corrections are possible

$ [Mishra et al., CVPR 2012]
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>43.3</td>
</tr>
</tbody>
</table>

- These experiments do not use any edit distance based correction
- 0.5 Million dictionary words are used to compute priors
- ICDAR2003 words with special characters are avoided.
## Qualitative Results

<table>
<thead>
<tr>
<th></th>
<th>Unary</th>
<th>Pair-wise</th>
<th>Higher Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOUR</td>
<td>Y0UK</td>
<td>YOUK</td>
<td>YOUR</td>
</tr>
<tr>
<td>twilight</td>
<td>TWI1IOHT</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>BEE1</td>
<td>BEER</td>
</tr>
<tr>
<td>Srishti</td>
<td>SRISNTI</td>
<td>SRISNTI</td>
<td>SRISHTI</td>
</tr>
</tbody>
</table>
Conclusions and On-going Work

- A principled framework
- Joint inference to detect true characters and recognize word as a whole
- A novel higher order potentials
- Largest dataset for Scene Text Recognition

**Ongoing Work**

- Better features and learning for scene character classification
Visit our project page for more detail of the work
http://cvit.iiit.ac.in/projects/SceneTextUnderstanding/

Supported by Microsoft Research India PhD Fellowship