Downtown Osaka Scene Text Dataset

Masakazu Iwamura, Takahiro Matsuda, Naoyuki Morimoto, Hitomi Sato, Yuki Ikeda and Koichi Kise

Osaka Prefecture University
Agenda

1. Introduction
2. Unique Features of DOST Dataset
3. Construction of DOST Dataset
4. Known Issues
5. Evaluation
6. Conclusion
Agenda

1. Introduction
2. Unique Features of DOST Dataset
3. Construction of DOST Dataset
4. Known Issues
5. Evaluation
6. Conclusion
Recent Improvement of Scene Text Recognition

Recent results are 80+% or even 90+%

This does not mean these methods can read a wide variety of text in the real environment.
Text in Real Environment

- We mean
- Text captured without intention (as much as possible)
- Text not screened so as to be easily read (with regard to resolution, capture angle and so on)
We present DOST Dataset
Agenda

1. Introduction
2. Unique Features of DOST Dataset
3. Construction of DOST Dataset
4. Known Issues
5. Evaluation
6. Conclusion
Unique Features of DOST Dataset

1. **Aim:** evaluation of methods in the real environment
   - Not aiming at training classifiers like MJSynth and SynthText datasets

2. **Completely not intentionally captured**
   - The most similar is ICDAR2015 Challenge 4 “incidental scene text” dataset captured with Google Glass
   - DOST is even free from face direction
Unique Features of DOST Dataset

3. Video dataset captured with omnidirectional camera
   - ICDAR 2013 & 2015 Challenge 3: single direction
   - YouTube Video (YTV) Dataset: YouTube Videos

4. Contains multiple images of single word
Unique Features of DOST Dataset

5. Large scale
   • Contains largest number of word Images
   • Excluding synthesized datasets (MJSynth and SynthText)
   • Excluding dataset containing numbers only (Google Streetview House Number dataset)
No. of Images Contained in Existing Datasets

Image DB

- ICDAR2003: 509
- ICDAR2013 Chal. 2: 462
- ICDAR2015 Chal. 4: 1,670
- NEOCR: 659
- KAIST: 3,000
- SVT: 349
- IIIT5K: 5,000
- COCO-Text: 63,686

Video DB

- ICDAR2013 Chal. 3: 15,277
- ICDAR2015 Chal. 3: 27,824
- YVT: 11,791
- DOST: 32,147

Almost double
No. of Word Images Contained in Existing Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Image DB</th>
<th>Video DB</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICDAR2003</td>
<td>2,268</td>
<td></td>
</tr>
<tr>
<td>ICDAR2013 Chal. 2</td>
<td>2,524</td>
<td></td>
</tr>
<tr>
<td>ICDAR2015 Chal. 4</td>
<td>17,548</td>
<td></td>
</tr>
<tr>
<td>NEOCR</td>
<td>5,238</td>
<td></td>
</tr>
<tr>
<td>KAIST</td>
<td>3,000</td>
<td></td>
</tr>
<tr>
<td>SVT</td>
<td>904</td>
<td></td>
</tr>
<tr>
<td>IIIT5K</td>
<td>5,000</td>
<td></td>
</tr>
<tr>
<td>COCO-Text</td>
<td>173,589</td>
<td></td>
</tr>
<tr>
<td>ICDAR2013 Chal. 3</td>
<td>93,598</td>
<td></td>
</tr>
<tr>
<td>ICDAR2015 Chal. 3</td>
<td>125,141</td>
<td></td>
</tr>
<tr>
<td>YVT</td>
<td>16,620</td>
<td></td>
</tr>
<tr>
<td>DOST</td>
<td>797,919</td>
<td></td>
</tr>
</tbody>
</table>

Images were captured in shopping streets where a lot of texts exist.
No. of Word Sequences in Existing Video Datasets

- ICDAR2013 Chal. 3: 1,962
- ICDAR2015 Chal. 3: 3,562
- YVT: 245
- DOST: 22,398

YVT has 6.3 times more word sequences than ICDAR2015 Chal. 3.
Unique Features of DOST Dataset

6. Contains Japanese characters

- On the other hand, a lot of non-Japanese words are contained

- OPEN
- _AMBA
- SECOM
- ROCKY
- Mickey
No. of Ground Truthed Characters per Category

- Alphabet: 837,489
- Kanji: 723,805
- Katakana: 696,697
- Hiragana: 355,158
- Digit: 324,742
- Symbol: 22,802
No. of Ground Truthed Characters per Category

- **Alphabet**: 837,489
- **Kanji**: 723,805
- **Katakana**: 696,697
- **Hiragana**: 324,742
- **Digit**: 22,802
- **Symbol**: 0

Japanese characters
7. Manually ground truthed
   • Amazon Mechanical Turk is not usable
   • Hiring students costed a lot!
Agenda

1. Introduction
2. Unique Features of DOST Dataset
3. Construction of DOST Dataset
4. Known Issues
5. Evaluation
6. Conclusion
Construction of DOST Dataset

1. Image capture
   - Point Grey Research LadyBug 3
   - 1,200x1,600 pixels, 6.5 fps

Completed in 2012
Place, time length, the number of images of capture

<table>
<thead>
<tr>
<th>Place</th>
<th>Length [h]</th>
<th>#Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sakai-Higashi</td>
<td>0.73</td>
<td>101,874</td>
</tr>
<tr>
<td>Namba</td>
<td>3.71</td>
<td>521,988</td>
</tr>
<tr>
<td>Shinsaibashi</td>
<td>0.25</td>
<td>35,100</td>
</tr>
<tr>
<td>Abiko</td>
<td>0.50</td>
<td>70,614</td>
</tr>
<tr>
<td>Tennoji</td>
<td>0.38</td>
<td>53,754</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>5.57</strong></td>
<td><strong>783,150</strong></td>
</tr>
</tbody>
</table>
Construction of DOST Dataset

2. Manual ground truthing
   - Most of GT policies are shared with ICDAR2013 & 2015 Challenge 3 datasets
   - GT software was developed
   - Reuse GT information in neighboring frames

3. Privacy preservation
   - Faces were blurred

We spent more than 1,500 man hours
Ground Truthing Policy

• **Basic unit**
  • Word or Bunsetsu (in Japanese)
    - Bunsetsu: the smallest unit of words that sounds natural in a spoken sentence
  • Proper noun is not divided

• **Bounding box**
  • Basic unit is represented by its four corners
Ground Truthing Policy

• **Transcription**
  • transcription consists of visible characters

• **Quality**
  • High, mid or low
  • Low corresponds to “Don’t care” regions

• **ID**
  • The same ID is assigned to a sequence of same basic units as long as it can be traced
  • Trace ends when a basic unit completely goes out from the frame
## Distribution of lengths of image sequences

<table>
<thead>
<tr>
<th>Length of sequence</th>
<th>#sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>5001</td>
<td>2</td>
</tr>
<tr>
<td>3181</td>
<td>1</td>
</tr>
<tr>
<td>2000 - 2009</td>
<td>4</td>
</tr>
<tr>
<td>1951</td>
<td>1</td>
</tr>
<tr>
<td>1500 - 1501</td>
<td>2</td>
</tr>
<tr>
<td>101-582</td>
<td>6</td>
</tr>
<tr>
<td>-100</td>
<td>9</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>27</strong></td>
</tr>
</tbody>
</table>
Agenda

1. Introduction
2. Unique Features of DOST Dataset
3. Construction of DOST Dataset
4. Known Issues
5. Evaluation
6. Conclusion
**Known Issues**

- **Ground truths are not perfect**
  - Bounding boxes of text regions are not tight enough
  - Ground trothing “Don’t care” is not comprehensive
  - Some word sequences are broken

- **Relationship between other cameras**
  - Word images in other cameras are not followed

We will improve them

“Don’t care” is marked in illegible regions
Agenda

1. Introduction
2. Unique Features of DOST Dataset
3. Construction of DOST Dataset
4. Known Issues
5. Evaluation
6. Conclusion
Evaluation: Methods

- **Text detection**
  - OpenCV API
  - Matsuda’s method based on NAT method

- **End-to-end text recognition**
  - Google Vision API
Evaluation: Datasets

- **Image datasets**
  - ICDAR2003
  - ICDAR2013 Chal. 2
  - ICDAR2015 Chal. 4
  - SVT
  - COCO-Text

- **Video datasets**
  - ICDAR2015 Chal. 3
  - YVT
  - DOST
  - DOST Latin

Subset of DOST which contain words consisting of alphabets and digits

Data were sampled
Text Detection by OpenCV API

<table>
<thead>
<tr>
<th>Database</th>
<th>F-measure [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>YVT</td>
<td>28.5</td>
</tr>
<tr>
<td>COCO-Text</td>
<td>11.9</td>
</tr>
<tr>
<td>SVT</td>
<td>19</td>
</tr>
<tr>
<td>ICDAR2015 Chal. 3</td>
<td>8.5</td>
</tr>
<tr>
<td>ICDAR2015 Chal. 4</td>
<td>13</td>
</tr>
<tr>
<td>ICDAR2013 Chal. 2</td>
<td>6.1</td>
</tr>
<tr>
<td>ICDAR2003</td>
<td>18.7</td>
</tr>
<tr>
<td>Image DB</td>
<td></td>
</tr>
<tr>
<td>Video DB</td>
<td></td>
</tr>
<tr>
<td>DOST Latin</td>
<td>2.4</td>
</tr>
<tr>
<td>DOST</td>
<td>1.2</td>
</tr>
</tbody>
</table>
Text Detection by Matsuda's method

Image DB

ICDAR2003: 47.5
ICDAR2013 Chal. 2: 4.8
ICDAR2015 Chal. 4: 6.3

Video DB

SVT: 29.1

COCO-Text

ICDAR2015 Chal. 3: 1.5
YVT: 3.9
DOST: 1.9
DOST Latin: 2.1

F-measure [%]
End-to-end Text Recognition by Google Vision API

- Image DB
  - ICDAR2003: 81.8%
  - ICDAR2013 Chal. 2: 71.3%
  - ICDAR2015 Chal. 4: 48.5%
  - SVT: 24.2%
  - COCO-Text: 17.1%
  - ICDAR2015 Chal. 3: 44.1%
  - YVT: 37.7%

- Video DB
  - DOST: 2.7%
  - DOST Latin: 11.2%

Recognized in Japanese mode.
Agenda

1. Introduction
2. Unique Features of DOST Dataset
3. Construction of DOST Dataset
4. Known Issues
5. Evaluation
6. Conclusion
Conclusion

- DOST dataset is presented
  - Has unique features
  - More challenging than existing datasets
Thank you for your attention!!