You Only Recognize Once: Towards Fast Video Text Spoting

Zhanzhan Cheng; Jing Lu; Yi Niu; Shiliang Pu
Hikvision Research Institute

Fei Wu
Zhejiang University

Shuigeng Zhou
Fudan University

lujing6@hikvision.com

Program ID: P2C-11
YORO-Fast Video Text Spotting

**Motivation**
- **Problems** of existing framework
  - unstable recognition results
  - excessive computational cost

**Advantages** of our framework
- more robust results
- faster by recognizing once

**Contribution**
- An unified two-stage framework YORO for fast video text spotting:
  - A self-attention based robust detector
  - A novel text recommender for fast text recognition
  - Fast and robust recognition:

Each frame detected text
- COCKTAIL
- servilletas
- COCKTAIL
- servilleta
- COCKTAIL
- servilleta

ResNet Backbone

Feature of each text region

Select for the best
- COCKTAIL
- servilletas

Track
- COCKTAIL
- servilleta
- servilleta

Score
- 0.53
- 0.75
- 0.90

Select
- servilletas

Recognize
- ’servilletas’
**YORO-Fast Video Text Spotting**

### Key component
- **Mechanism of quality scoring network**

\[ r^t = k\text{means}(r_{1}^{cor}, r_{2}^{cor}, \ldots, r_{k}^{cor}) \]
\[ s_i = \frac{r^t \odot r_i}{||r^t|| \times ||r_i||} \]

![Diagram showing the Teacher-Student Architecture](image)

1. **Teacher-Student Architecture**
   - Input: Pretrained Recognition Model
   - Output: Selected Text

### Experiments
- **Ablation**
  - Comparison with other frame selection methods
  - Effectiveness of each module on IC15
  - Performance of detection on IC13
  - Performance of end-to-end on IC15

### Comparison with SoTA

<table>
<thead>
<tr>
<th>Method</th>
<th>MOT(P_R)</th>
<th>MOT(A_R)</th>
<th>AT(A_R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stradvision [18]</td>
<td>0.69</td>
<td>0.57</td>
<td>0.29</td>
</tr>
<tr>
<td>Deep2Text [18]</td>
<td>0.62</td>
<td>0.35</td>
<td>0.19</td>
</tr>
<tr>
<td>Wang et al. [53]</td>
<td>0.70</td>
<td>0.69</td>
<td>0.60</td>
</tr>
<tr>
<td>Ours</td>
<td>0.76</td>
<td>0.69</td>
<td>0.63</td>
</tr>
</tbody>
</table>

### The Large Scale Video Text Dataset

<table>
<thead>
<tr>
<th>Datasets</th>
<th>#scenarios</th>
<th>#videos</th>
<th>#frames</th>
<th>#instances</th>
<th>quality?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Merino [28]</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Minetto [30]</td>
<td>-</td>
<td>5</td>
<td>3599</td>
<td>8706</td>
<td>-</td>
</tr>
<tr>
<td>IC13 [19]</td>
<td>7</td>
<td>28</td>
<td>15277</td>
<td>93934</td>
<td>-</td>
</tr>
<tr>
<td>YVT [33]</td>
<td>-</td>
<td>30</td>
<td>13500</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IC15 [18]</td>
<td>7</td>
<td>49</td>
<td>27824</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LSVTD</td>
<td>22</td>
<td>100</td>
<td>66700</td>
<td>569300</td>
<td>-</td>
</tr>
</tbody>
</table>

- Much larger scale (22 scenes)
- Multilingual text
- Release soon…