Mutually Guided Dual-Task Network for Scene Text Detection

Mengbiao Zhao\textsuperscript{1,2}, Wei Feng\textsuperscript{1,2}, Fei Yin\textsuperscript{1,2}, Xu-Yao Zhang\textsuperscript{1,2}, Cheng-Lin Liu\textsuperscript{1,2,3}

\textsuperscript{1}National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

\textsuperscript{2}University of Chinese Academy of Sciences

\textsuperscript{3}CAS Center for Excellence of Brain Science and Intelligence Technology
Outline

1. Introduction
2. System Overview
3. Mutual Guidance Strategy
4. Loss Function
5. Experiments
6. Conclusions
There are two mainstream annotation formats for scene text datasets: word-level and line-level annotations.

In previous works, word detection and text-line detection are usually treated separately.

Word-level and line-level detection are closely related.
Architecture

- A backbone network (ResNet-50 FPN) for feature extraction.
- Two detection heads for words and text-lines detection, respectively.
- Two novel modules for the mutual guidance of the two tasks.
Mutual Guidance Strategy

For ease of analysis, we divide the training process into two stages.
Mutual Guidance

- Stage 1

\[
\begin{align*}
O_W &= D_W(E_I), \\
O_L &= D_L(E_I).
\end{align*}
\]

- \( E \) : backbone network
- \( E_I \) : original features
- \( D_W \) : word detector
- \( D_L \) : text-line detector
- \( O_W \) : output results of word detector
- \( O_L \) : output results of text-line detector
Mutual Guidance

Stage 2

- Line filtering modules.

\[ E'_I = E_I \odot G_W + E_I, \]

- Word enhancing modules.

\[ E'_I = E_I + G_L, \]

\[ \begin{align*}
O'_W &= D_W(E_I, G_W), \\
O'_L &= D_L(E_I, G_L).
\end{align*} \]
Loss Function

- We use pair of datasets, one with word-level ground truth $Y_W$ and one with line-level ground truth $Y_L$.
- For a data batch with $Y_W$, we just compute the dice coefficient loss between $Y_W$ and its two stages’ outputs $O_W$ and $O'_W$.

\[
\mathcal{L} = \sum_{t \in \{W,L\}} \sum_{X \in \{O_t, O'_t\}} b_t \cdot \mathcal{L}_{\text{dice}}(X, Y_t),
\]

- Where $b_t$ represents the category of the current data batch.
- If a data batch has ground truth $Y_W$ only, then $b_W = 1$, $b_L = 0$ and vice versa.
Experiments

- **Datasets**
  - ICDAR2015: Word-level annotated dataset.
  - CTW1500: Line-level annotated dataset.
Experiments

- **Models**

  - **Baseline**: The basic detector trained with word-level and line-level annotated data separately.
  
  - **Baseline + joint**: The basic detector jointly trained with word-level and line-level annotated data.
  
  - **Dual-task**: Our proposed dual-task network jointly trained with word-level and line-level annotated data.
  
  - **Dual-task + guidance**: Our proposed dual-task network jointly trained with word-level and line-level annotated data, and the mutual guidance strategy added.
Experiments

- Ablation studies on ICDAR2015

<table>
<thead>
<tr>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline [3]</td>
<td>81.5</td>
<td>79.7</td>
<td>80.6</td>
</tr>
<tr>
<td>Baseline+joint [3]</td>
<td>82.32</td>
<td>76.45</td>
<td>79.28</td>
</tr>
<tr>
<td>Dual-task</td>
<td>87.41</td>
<td>74.91</td>
<td>80.68</td>
</tr>
<tr>
<td>Dual-task+guidance</td>
<td>82.08</td>
<td>80.98</td>
<td>81.53</td>
</tr>
</tbody>
</table>

- The basic detector jointly trained with two datasets yields deteriorated performance.

- The dual-task network leads to an improved performance.

- The dual-task network trained with mutual guidance yields the best detection performance.
Experiments

- Comparison with State-of-the-Art Methods.
  - Detection results on ICDAR2015.

<table>
<thead>
<tr>
<th>Method</th>
<th>Ext</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>EAST [27]</td>
<td></td>
<td>83.57</td>
<td>73.47</td>
<td>78.2</td>
</tr>
<tr>
<td>PixelLink [8]</td>
<td></td>
<td>82.9</td>
<td>81.7</td>
<td>82.3</td>
</tr>
<tr>
<td>TextBoxes++ [28]</td>
<td>√</td>
<td>87.2</td>
<td>76.7</td>
<td>81.7</td>
</tr>
<tr>
<td>DDR [1]</td>
<td></td>
<td>82.0</td>
<td>80.0</td>
<td>81.0</td>
</tr>
<tr>
<td>FOTS [29]</td>
<td>√</td>
<td>88.84</td>
<td>82.04</td>
<td>85.31</td>
</tr>
<tr>
<td>Mask TextSpotter [30]</td>
<td>√</td>
<td>91.6</td>
<td>81.0</td>
<td>86.0</td>
</tr>
<tr>
<td>TextField [31]</td>
<td>√</td>
<td>84.3</td>
<td>80.5</td>
<td>82.4</td>
</tr>
<tr>
<td>TextSnake [32]</td>
<td>√</td>
<td>84.9</td>
<td>80.4</td>
<td>82.6</td>
</tr>
<tr>
<td>PSENet [3]</td>
<td></td>
<td>81.5</td>
<td>79.7</td>
<td>80.6</td>
</tr>
<tr>
<td>PSENet [3]</td>
<td>√</td>
<td>86.9</td>
<td>84.5</td>
<td>85.7</td>
</tr>
<tr>
<td>Our Method</td>
<td></td>
<td>82.08</td>
<td>80.98</td>
<td>81.53</td>
</tr>
<tr>
<td>Our Method</td>
<td>√</td>
<td>88.60</td>
<td>84.54</td>
<td>86.52</td>
</tr>
</tbody>
</table>
Experiments

- Comparison with State-of-the-Art Methods.
  - Detection results on CTW1500.

<table>
<thead>
<tr>
<th>Method</th>
<th>Ext</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTPN [12]</td>
<td>-</td>
<td>60.4</td>
<td>53.8</td>
<td>56.9</td>
</tr>
<tr>
<td>SegLink [33]</td>
<td>-</td>
<td>42.3</td>
<td>40.0</td>
<td>40.8</td>
</tr>
<tr>
<td>CTD+TLOC [4]</td>
<td>-</td>
<td>77.4</td>
<td>69.8</td>
<td>73.4</td>
</tr>
<tr>
<td>TextSnake [32]</td>
<td>✓</td>
<td>67.9</td>
<td>85.3</td>
<td>75.6</td>
</tr>
<tr>
<td>Wang et al. [3]</td>
<td>-</td>
<td>80.1</td>
<td>80.2</td>
<td>80.1</td>
</tr>
<tr>
<td>TextField [31]</td>
<td>✓</td>
<td>83.0</td>
<td>79.8</td>
<td>81.4</td>
</tr>
<tr>
<td>PSENet [3]</td>
<td>-</td>
<td>80.57</td>
<td>75.55</td>
<td>78.0</td>
</tr>
<tr>
<td>PSENet [3]</td>
<td>✓</td>
<td>84.84</td>
<td>79.73</td>
<td>82.2</td>
</tr>
<tr>
<td>Our Method</td>
<td>-</td>
<td>81.48</td>
<td>78.42</td>
<td>79.92</td>
</tr>
<tr>
<td>Our Method</td>
<td>✓</td>
<td>85.59</td>
<td>80.21</td>
<td>82.81</td>
</tr>
</tbody>
</table>
Experiments

- Some examples of text detection.

ICDAR2015  CTW1500

- Each image can get two formats of detection results from two detection heads.
Conclusions

- **Propose a text detection method that can perform both word-level and line-level text detection.**
  - Dual-task network.

- **Propose two novel modules for the mutual guidance of the two tasks.**
  - Line filtering module.
  - Word enhancing module.

- **Proposed method has achieved competitive performance.**

- **Future works**
  - Weakly-supervised training.
  - Adding character-level detection.
Thanks & Question