Oriented Scene Text Detection Revisited

Xiang Bai
Huazhong University of Science and Technology
xbai@hust.edu.cn
http://mclab.eic.hust.edu.cn/~xbai/
Outline

- Problem Definition
- Review
- Our Works
- Benchmarks and Evaluation
- Applications
- Future Trends
Problem Definition

Text Detection
Word/line level

Text Recognition
Word/sequence classification

End-to-end Recognition
How do we perceive scene text?

Top-Down vs. Bottom-Up, which is better?
The Story of Oriented Scene Text Detection

- **Handcraft Features**
  - Component level. MSER, SWT...
  - Word / line level. Sliding Window

- **Deep Learning (2014-)**
  - Region Proposals
  - Segmentation
  - Hybrid Methods
The Story of Oriented Scene Text Detection

- **Handcraft Features**
  - Component level. MSER, SWT…
  - Word / line level. Sliding Window

- **Deep Learning (2014-)**
  - Region Proposals
  - Segmentation
  - Hybrid Methods
- Specially designed features.
- Two-level classification scheme.
- The 1st benchmark dataset for multi-oriented text detection: MSRA-TD 500
Detecting Texts of Arbitrary Orientations in Natural Images
[Yao et al., CVPR, 2012]

Full process of text detection

(a) Original image
(b) Edge detection
(c) SWT
(d) Association
(e) Component filtering

(f) Component verification
(g) Aggregation
(h) Chain verification
(i) Interpretation
(j) Detected texts
Two sets of rotation-invariant features that facilitate multi-oriented text detection:
- component level: estimate center, scale, and direction before feature computation…
- chain level: size variation, color self-similarity, structure self-similarity…
Orientation Robust Text Line Detection in Natural Images
[Kang et al., CVPR, 2014]

- Build a graph based on MSER components
- Higher-order correlation clustering (HOCC)
- Texton-based texture classifier to discriminate text and non-text regions
Morphology clustering: grouping characters candidates by the character appearances (Color, Stroke width and Compactness).

Orientation clustering: grouping character pairs by the character pair orientation.

Projection clustering: grouping character pairs by the character pair intercept.
The Story of Oriented Scene Text Detection

- Handcraft Features
  - Component level. MSER, SWT...
  - Word / line level. Sliding Window

- Deep Learning (2014-)
  - Region Proposals
  - Segmentation
  - Hybrid Methods
Reading Text in the Wild with Convolutional Neural Networks
[Jaderberg et al., IJCV, 2016]
Symmetry-based text line detection in natural scenes
[Zhang et al., CVPR, 2015]
Synthetic Data for Text Localisation in Natural Images [Gupta et al., CVPR, 2016]

- Synthesis text in the wild.
- Using synthetic text to train scene text detector.

Detecting Text in Natural Image with Connectionist Text Proposal Network
[Tian et al., ECCV, 2016]

- Dense sliding windows on feature maps to extract a feature vector of every location.
- BLSTM to capture the sequential context information.
- Fully-connected layer simultaneously predicts text/non-text scores, y-axis coordinates and side-refinement offsets of k anchors.
Detecting Text in Natural Image with Connectionist Text Proposal Network (Cont.)
[Tian et al., ECCV, 2016]

1. Fine-scale Proposals

Fig. 2: **Left:** RPN proposals. **Right:** Fine-scale text proposals.

2. Recurrent Connectionist Text Proposals

Fig. 3: **Top:** CTPN without RNN. **Bottom:** CTPN with RNN connection.

3. Side-refinement
TextBoxes: A fast text detector with a single deep neural network
[Liao et al., AAAI 2017]

- Fully convolutional network based on SSD[1].
- On every map location, a text-box layer predicts a 72-d vector (text presence scores (2-d) and offsets (4-d) for 12 default boxes)
- Longer convolutional filters
- Special designed default boxes

Detecting Oriented Text in Natural Images by Linking Segments
[Shi et al., CVPR 2017.]

- Fully convolutional network inspired by SSD
- Multi-stage outputs for segments and their links
- Solve the problem of CNN receptive field for long texts
Arbitrary-Oriented Scene Text Detection via Rotation Proposals [Jianqi Ma et al., arXiv:1703.01086, 2017.]

- Use the architecture of faster-rcnn
- RPN->Rotated RPN
- RoI->Rotated RoI

The implementation of Rotated RoI

- Use the architecture of SSD
- Use different matching strategy

Previous method

Our shared Monte-Carlo method

The ratio of overlapping points in total points multiplies the area of circumscribed rectangle is the overlapping area.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very high. The error is stable at about 1%.</td>
<td>Calculate overlapping area between polygons.</td>
</tr>
</tbody>
</table>

Low. The error increases rapidly as the relative angle increases.

Calculate overlapping area between rectangles.
The Story of Oriented Scene Text Detection

- **Handcraft Features**
  - Component level. MSER, SWT...
  - Word / line level. Sliding Window

- **Deep Learning (2014-)**
  - Region Proposals
  - Segmentation
  - Hybrid Methods
The Text-Block FCN is to predict the salient map of text block.

Multi-oriented text line hypotheses are generated by combining both global and local cues.

The Character-Centroid FCN is used to remove false positives.
Scene Text Detection Via Holistic, Multi-Channel Prediction
[Yao et al., arXiv:1606.09002, 2016]

- FCN based network.
- Multi task. Text region, individual characters and their relationship are estimated simultaneously.
Scene Text Detection Via Holistic, Multi-Channel Prediction
[Yao et al., arXiv:1606.09002, 2016]
Handcraft Features
- Component level: MSER, SWT...
- Word / line level: Sliding Window

Deep Learning (2014-)
- Region Proposals
- Segmentation
- Hybrid Methods
PVANet (faster than VGG16)

- Multi-channel:
  - Score map
  - Rotated bounding boxes
  - Quadrangle bounding boxes

- Refined NMS
Deep Direct Regression for Multi-Oriented Scene Text Detection
[He et al., arXiv:1703.08289, 2017.]

- Indirect regression
- Direct regression

- Multi-level feature fusion
- Up-sample to quarter size of the input image
- Multi-task learning for classification and regression
- Post Processing: Refined NMS
Outline

- Problem Definition
- Review
- Our work
- Benchmarks and Evaluation
- Applications
- Future Trends
Pipeline

- Fully convolutional network.
- On every map location, a text-box layer predicts a 72-d vector (text presence scores (2-d) and offsets (4-d) for 12 default boxes)
- Longer convolutional filters
- Special designed default boxes
Quantitative Results of Text Localization

<table>
<thead>
<tr>
<th>Datasets</th>
<th>ICDAR 2011</th>
<th></th>
<th>ICDAR 2013</th>
<th></th>
<th>Time/s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IC13 Eval</td>
<td>DetEval</td>
<td>IC13 Eval</td>
<td>DetEval</td>
<td></td>
</tr>
<tr>
<td>Methods</td>
<td>P R F</td>
<td>P R F</td>
<td>P R F</td>
<td>P R F</td>
<td></td>
</tr>
<tr>
<td>Jaderberg</td>
<td>– – –</td>
<td>– – –</td>
<td>– – –</td>
<td>– – –</td>
<td>7.3</td>
</tr>
<tr>
<td>MSERs-CNN (Yin et al. 2014)</td>
<td>0.88 0.71 0.78</td>
<td>– – –</td>
<td>– – –</td>
<td>– – –</td>
<td></td>
</tr>
<tr>
<td>MMser</td>
<td>– – –</td>
<td>– – –</td>
<td>0.86 0.70 0.77</td>
<td>– – –</td>
<td>0.75</td>
</tr>
<tr>
<td>TextFlow (Tian et al. 2015)</td>
<td>0.86 0.76 0.81</td>
<td>– – –</td>
<td>0.85 0.76 0.80</td>
<td>– – –</td>
<td>1.4</td>
</tr>
<tr>
<td>FCRNall+filts</td>
<td>– – –</td>
<td><strong>0.92</strong> 0.75 0.82</td>
<td>– – –</td>
<td><strong>0.92</strong> 0.76 0.83</td>
<td>&gt;1.27</td>
</tr>
<tr>
<td>Zhang (Zhang et al. 2016)</td>
<td>– – –</td>
<td>– – –</td>
<td>0.88 0.78 0.83</td>
<td>– – –</td>
<td>2.1</td>
</tr>
<tr>
<td>SSD (Liu et al. 2015)</td>
<td>– – –</td>
<td>– – –</td>
<td>0.80 0.60 0.68</td>
<td>0.80 0.60 0.69</td>
<td>0.1</td>
</tr>
<tr>
<td>Fast TextBoxes</td>
<td>0.86 0.74 0.80</td>
<td>0.86 0.74 0.80</td>
<td>0.86 0.74 0.80</td>
<td>0.88 0.74 0.81</td>
<td>0.09</td>
</tr>
<tr>
<td>TextBoxes</td>
<td><strong>0.88</strong> 0.82 0.85</td>
<td><strong>0.89</strong> 0.82 0.86</td>
<td><strong>0.88</strong> 0.83 0.85</td>
<td>0.89 0.83 0.86</td>
<td>0.73</td>
</tr>
</tbody>
</table>
CRNN: End-to-End Trainable Network for Scene Text Recognition
[Shi etc. PAMI 2017]

Network Structure
- Convolutional layers extract feature maps
- Convert feature maps into feature sequence
- Sequence labeling with LSTM
- Convert labeling into text
Combined with a recognition model (CRNN), we achieve state-of-the-art performance on ICDAR 2013.

<table>
<thead>
<tr>
<th>Method</th>
<th>End-to-End results</th>
<th>Word spotting Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>HUST_MCLAB</td>
<td>87.68 %</td>
<td>95.83 %</td>
</tr>
<tr>
<td>Adelaide_ConvLST...</td>
<td>79.50 %</td>
<td>96.68 %</td>
</tr>
<tr>
<td>SRC-B-TextProcess...</td>
<td>81.79 %</td>
<td>93.17 %</td>
</tr>
<tr>
<td>VGGMaxBBNet_095</td>
<td>82.12 %</td>
<td>91.05 %</td>
</tr>
<tr>
<td>VGGMaxBBNet_0...</td>
<td>82.99 %</td>
<td>89.63 %</td>
</tr>
<tr>
<td>Yunos_Robot1.0</td>
<td>75.57 %</td>
<td>95.06 %</td>
</tr>
<tr>
<td>Deep2Text II+</td>
<td>72.08 %</td>
<td>94.56 %</td>
</tr>
</tbody>
</table>
Detecting Oriented Text in Natural Images by Linking Segments
[Shi et al., CVPR 2017.]

- Fully convolutional network inspired by SSD
- Multi-stage outputs for segments and their links
- Solve the problem of CNN receptive field for long texts
Detecting Oriented Text in Natural Images by Linking Segments
[Shi et al., CVPR 2017.]

Linking segments

Long texts can be easily located
Results on ICDAR 2015 Incidental Text

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUST-MCLAB</td>
<td>47.5</td>
<td>34.8</td>
<td>40.2</td>
</tr>
<tr>
<td>NJU_Text</td>
<td>72.7</td>
<td>35.8</td>
<td>48.0</td>
</tr>
<tr>
<td>StradVision-2</td>
<td>77.5</td>
<td>36.7</td>
<td>49.8</td>
</tr>
<tr>
<td>MCLAB_FCN [30]</td>
<td>70.8</td>
<td>43.0</td>
<td>53.6</td>
</tr>
<tr>
<td>CTPN [22]</td>
<td>51.6</td>
<td>74.2</td>
<td>60.9</td>
</tr>
<tr>
<td>Megvii-Image++</td>
<td>72.4</td>
<td>57.0</td>
<td>63.8</td>
</tr>
<tr>
<td>Yao et al. [26]</td>
<td>72.3</td>
<td>58.7</td>
<td>64.8</td>
</tr>
<tr>
<td>SegLink</td>
<td>73.1</td>
<td>76.8</td>
<td>75.0</td>
</tr>
</tbody>
</table>

End-to-end results on ICDAR 2015 Incidental Text (combined with CRNN)
Outline

- Problem Definition
- Review
- Our work
- Benchmarks and Evaluation
- Applications
- Future Trends
ICDAR2015 - Incidental Scene Text dataset

- Focus on the incidental scene where text may appear in any orientation any location with small size or low resolution.
- Includes 1000 training images containing about 4500 readable words and 500 testing images.
MSRA-TD500

- Contains 500 natural images taken from indoor and outdoor.
- Texts in different languages (Chinese, English or mixture of both), fonts, sizes, colors and orientations.
- Annotated with text line bounding box.
- Ref. Detecting Texts of Arbitrary Orientations in Natural Images, CVPR12
RCTW-17 dataset

- Chinese Text in the Wild (12,034 images, 8034 images for training and 4000 images for testing)
- The text annotated in RCTW-17 consists of Chinese characters, digits, and English characters, with Chinese characters taking the largest portion.
- ICDAR2017 Competition on Reading Chinese Scene Text in the Wild (RCTW-17)
## Comparison on ICDAR 2015

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>Time/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhou et al. CVPR 2017</td>
<td>84</td>
<td>73</td>
<td>78</td>
<td>0.08</td>
</tr>
<tr>
<td>Shi et al. CVPR 2017</td>
<td>73</td>
<td>77</td>
<td>75</td>
<td>--</td>
</tr>
<tr>
<td>Ma et al. arxiv 2017</td>
<td>82</td>
<td>73</td>
<td>77</td>
<td>--</td>
</tr>
<tr>
<td>Liu et al. CVPR 2017</td>
<td>73</td>
<td>68</td>
<td>71</td>
<td>--</td>
</tr>
<tr>
<td>He et al. arxiv 2017</td>
<td>82</td>
<td>80</td>
<td>81</td>
<td>--</td>
</tr>
<tr>
<td>Tian et al. ECCV 2016</td>
<td>74</td>
<td>52</td>
<td>61</td>
<td>--</td>
</tr>
<tr>
<td>Zhang et al. CVPR 2016</td>
<td>71</td>
<td>43</td>
<td>54</td>
<td>2.1</td>
</tr>
</tbody>
</table>
## Comparison on MSRA-TD 500

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Time/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhou et al. CVPR 2017</td>
<td>87</td>
<td>67</td>
<td>76</td>
<td>0.08</td>
</tr>
<tr>
<td>Shi et al. CVPR 2017</td>
<td>86</td>
<td>70</td>
<td>77</td>
<td>0.11</td>
</tr>
<tr>
<td>Ma et al. arxiv 2017</td>
<td>82</td>
<td>68</td>
<td>74</td>
<td>0.3</td>
</tr>
<tr>
<td>He et al. arxiv 2017</td>
<td>77</td>
<td>70</td>
<td>74</td>
<td>--</td>
</tr>
<tr>
<td>Huang et al. ACM MM 2016</td>
<td>74</td>
<td>68</td>
<td>71</td>
<td>--</td>
</tr>
<tr>
<td>Yao et al. arxiv 2016</td>
<td>77</td>
<td>75</td>
<td>76</td>
<td>0.42</td>
</tr>
<tr>
<td>Zhang et al. CVPR 2016</td>
<td>83</td>
<td>67</td>
<td>74</td>
<td>--</td>
</tr>
<tr>
<td>Yin et al. PAMI 2015</td>
<td>81</td>
<td>63</td>
<td>71</td>
<td>1.4</td>
</tr>
<tr>
<td>Kang et al. CVPR 2014</td>
<td>71</td>
<td>62</td>
<td>66</td>
<td>--</td>
</tr>
<tr>
<td>Yao et al. CVPR 2012</td>
<td>63</td>
<td>63</td>
<td>60</td>
<td>--</td>
</tr>
</tbody>
</table>
The Drawback of IOU in Scene Text Detection
Outline

- Problem Definition
- Review
- Our work
- Benchmarks and Evaluation
- Applications
- Future Trends
Applications

- Fine-grained Classification
- Number
- Container
- Exercise search
- Word retrieval in the wild
Motivations

- Visual cues would group (a)-(b) whereas scene text reveals that and groups (b)-(c).
- Texts in images can improve the performance of fine-grained image classification.

Fine-Grained Image Classification with Text Information

Pipeline

GoogLeNet

Image Feature $f_r$

Classification

Restaurant

Text Feature $f_t$

Attended Text Feature $f_a$

Word Embedding

Word Spotting

Detection

Recognition

TRUCKEE DINER
Attention Model

Word Embedding

\[ \omega_i \propto \exp(f_v \cdot \bar{U} \cdot f_u) \]

Weighted Sumpooling

Attended Text Feature

word1, word2, word3, word4, ..., wordN
Results

(a) BARBERSHOP
BARBER: 1
SHOP: 7.8e-7
MENUS: 2.8e-8
ROOM: 1.2e-11
BARBS: 3.8e-18

(b) CAFE
COFFEE: 0.97
ESPRESSO: 0.03
CAPPUCCINO: 2.0e-10
ITALIAN: 2.2e-12

(c) BAKERY
CAKES: 0.57
PASTRIES: 0.43
OPEN: 5.5e-9
EGGO: 1.1e-10
DANISH: 3.1e-11

(d) CAFE
STARBUCKS: 1
SCOFF: 1.1e-8

(e) ROOTBEER
ROOT: 0.89
BEER: 0.11
BREWED: 1.3e-6
PURE: 2.4e-7
MICRO: 1.1e-9
MADE: 3.8e-10
NATURAL: 2.7e-11
RICH: 1.8e-11
EFL: 5.5e-12

(f) CHABLIS
CHABLIS: 0.99
FRANCE: 8.7e-12
FRANC: 1.1e-12
YIN: 2.4e-16
CON: 2.3e-18
CONTROL: 1.9e-18
BOUTIQUE: 2.5e-19
AFFILIATION: 6.2e-20

(g) BITTER
BITTER: 0.99
BROWN: 4.05e-5
PREMIUM: 3.5e-9
SPECIAL: 2.8e-9
ENGLISH: 9.4e-11
EXTRA: 6.11e-11

(h) GUINNESS
GUINNESS: 1
SPECIAL: 1.6e-25
EXPORT: 6.4e-27
QUINES: 1.3e-30
Person Re-identification with Numbers
检测特定的文字并识别

- Container

32.500KGS
71.650LBS
3.880KGS
8.580LBS

30.480KGS
67.200LBS
2.240KGS
4.940LBS

45G1
TGHU 953150 9

22G1
WHSU 207134 0

WHSU 2071340
no good persuading her to stopping smoking.

Children

21.已知直线与抛物线 $y = 2px (p > 0)$ 交于 A, B 两点, 且 $OA \perp OB$, $OD \perp AB$ 交 $AB$ 于点 D,

（1）词中所写的是什么季节？从哪里可以看出来？（3分）

3.下列语句有语病的一项是（ ）（2分）

计算下面机构的自由度，并并说明想使机构具有确定的运动，需要几个原动件
Word retrieval in the wild

以词搜图：
根据输入的关键词，系统返回数据库中包含该关键词的图片
Word retrieval in the wild

-绝大多数人眼清晰可辨的文字块均能被检测并正确识别
Word retrieval in the wild

• 相当比例的较小及模糊的文字块也能被检测并正确识别
Word retrieval in the wild

- 对于数据库中与检索词接近的词，系统将采用模糊匹配（按相似度排序显示）

当query为love时的部分检索结果（第一行：精准匹配，第二行：模糊匹配）
Future Trends

- End-to-end recognition.
- Retrieving Text in the wild
- Integrating Textual and Visual cues in many applications
Other resources (Datasets & Codes)

B. Shi, C. Yao, C. Zhang, X. Guo, F. Huang, X. Bai. Automatic script identification in the wild. ICDAR'15
Dataset: http://mc.eistar.net/~xbai/mspnProjectPage/

C. Zhang, C. Yao, B. Shi, X. Bai. Automatic discrimination of text and non-text natural images. ICDAR'15
Dataset&Code: http://mc.eistar.net/~xbai/textDis/textDis.html

Dataset: http://mclab.eic.hust.edu.cn/UpLoadFiles/dataset/HUST-TR400.zip

C. Yao, X. Bai, W. Liu, Y. Ma, Z. Tu. Detecting texts of arbitrary orientations in natural images. CVPR'12
Dataset: http://pages.ucsd.edu/~ztu/publication/MSRA-TD500.zip

Code: https://github.com/MhLiao/TextBoxes

B. Shi, X. Bai, C. Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. TPAMI'16
Code: http://mclab.eic.hust.edu.cn/~xbai/CRNN/crnn_code.zip

Code: https://github.com/stupidZZ/FCN_Text

Z. Zhang, W. Shen, C. Yao, X. Bai. Symmetry-based text line detection in natural scenes. CVPR'15
Code: https://github.com/stupidZZ/Symmetry_Text_Line_Detection

Code: http://mclab.eic.hust.edu.cn/~xbai/Strokelet_code/Strokelet_code.zip
END