Deep Neural Networks for Scene Text Reading

Xiang Bai

Huazhong University of Science and Technology
Problem definitions

- Definitions
  - Scene text detection
  - Scene text recognition

End-to-end recognition

Predicting the presence of text and localizing each instance (if any), usually at word or line level, in natural scenes

Converting text regions into computer readable and editable symbols

Summary Booklet
Outline

- Background
  - Scene Text Detection
  - Scene Text Recognition
  - Applications
  - Future Trends
Background

Xiang Bai, Kyoto, November 15

the intensity. We use this result to evaluate the quantity $N_{(a)} = \frac{1}{2} (a_{2} + a_{1})$, where $a_{1}$ is the transmitted power. Note that $N$ and $P$ are functions of $z$ but of $a_{2}$ because temporal dispersion and loss are assumed negligible. The coefficient $a_{2}^{1}$, defined by the relation $a_{2}^{1} = \frac{1}{2} (a_{2}^{1} + a_{2})$, is related to the coefficient $n_{2}^{1} = \frac{1}{2} (n_{2}^{1} + n_{2})$ from which it follows that $a_{2}^{1} = n_{2}^{1} P_{0}/P$. We use this, along with the definition of the critical power $P_{c} = 2a_{2}^{1}(a_{2}^{1} + a_{1})$, to determine the coefficient $n_{2}^{1}$ in Eq. (2) by $n_{2}^{1} P_{0}/P_{c}$, thus that $P_{c}$ is defined via (5). We substitute this result into Eq. (3) along with a new variable, $z = t - \lambda z$, to obtain

$$n_{2}^{1} P_{0}/P_{c} = \frac{1}{2} (a_{2}^{1} + a_{1})^{2}. \tag{5}$$

Now let us consider the hypothetical situation in which two beams of light with identical normalized amplitudes, $z_{1}$ and $z_{2}$, are two different samples, which we denote by the superindex $t = \tau$ (different sample and $t = \tau$ for the same sample). We let the samples have linear indices of refraction $n_{1}^{t}$ and $n_{2}^{t}$ and thicknesses $L_{1}$ and $L_{2}$, respectively. If the power is small enough that the last term in Eq. (2) is negligible, and if the sample lengths are chosen so that $L_{1}/n_{1}^{t} = L_{2}/n_{2}^{t}$, it follows from Eq. (4) that the normalized amplitudes are identical in the exit plane of the two samples. Furthermore, the normalized amplitudes will be nearly identical at the exit plane of the two samples if $L_{1}/n_{1}^{t} = L_{2}/n_{2}^{t}$ is zero, where $n_{0}$ is the Rayleigh range of the beam. If the input power is increased to some large values $P_{1}$ and $P_{2}$, and if the nonlinear index of refraction of the samples are $n_{1}^{n}$ and $n_{2}^{n}$, we see from Eq. (2) that to obtain the same $z_{1}$ in the exit plane of the two samples, we should adjust the power so that $[L_{1}/n_{1}^{t}][P_{1}/P_{c}] = [L_{2}/n_{2}^{t}][P_{2}/P_{c}]$. For two samples of the same thickness $L_{1} = L_{2}$, this condition is equivalent to $P_{1}/P_{c} = P_{2}/P_{c}$. With the sample thickness properly selected, the power properly adjusted, $z_{1}$ and $z_{2}$ will be the same for both samples of any given distance from the exit face, and therefore the measured normalized power is the same for both samples, $n_{2}^{n} = n_{2}^{n}$. Hence, a linear index of refraction of the samples is unique for the same distance from the exit face, and therefore the measured normalized power is the same for both samples, $n_{2}^{n} = n_{2}^{n}$. Hence, a linear index of refraction of the samples is unique for a distance of $z_{1}$ from the exit face, and therefore the measured normalized power is the same for both samples, $n_{2}^{n} = n_{2}^{n}$.
Background

Scene text detection methods before 2016

Proposals

- Generate candidates using hand-craft features

Filtering

- Text / non-text classification using CNN/Random forest

Regression

- Refine locations using CNN

References:


Xiang Bai, Kyoto, November 15
Background

Scene text detection methods after 2016

Segmentation-based method[1]

Proposal-based method[2]

Hybrid method[3]

Background

Scene text recognition methods

Word/Char Level\cite{1}
- Multi-class classification with one class per word/char

Sequence Level\cite{2}\cite{3}\cite{4}
- Text is a sequence of chars
- The whole sequence is recognized


Xiang Bai, Kyoto, November 15
Over 80% text detection papers focus on multi-oriented text detection.

- Scene text recognition and end-to-end recognition are paid less attention to.
- Most papers focus on English text.
Background

Latin text vs. Non-Latin text

**English**: there is always a blank space between neighbor words

**Chinese**: no vision cues for partition, while semantic information is needed.

Line-based detection and sequence labeling are appropriate for both Latin and Non-Latin text

Xiang Bai, Kyoto, November 15
Unlike general objects and English words, text lines have **larger aspect ratios**.

Given the fixed size of convolutional filters, text lines cannot be totally covered.
## Background

### Performance comparison on English / Chinese datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Language</th>
<th>Num. Train/Test</th>
<th>Best F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICDAR 2013</td>
<td>English</td>
<td>229/233</td>
<td>0.90</td>
</tr>
<tr>
<td>ICDAR 2015</td>
<td>English</td>
<td>1000/500</td>
<td>0.81</td>
</tr>
<tr>
<td>RCTW 2017</td>
<td>Mainly Chinese</td>
<td>8034/4229</td>
<td>0.66</td>
</tr>
</tbody>
</table>

The performance of Chinese dataset is much lower.

ICDAR 2017 Competition on Reading Chinese Text in the Wild

Possible solutions for Non-Latin text detection

- Long convolutional kernel.
- Inception convolutional kernels.
- Part detection and grouping.
Outline

- Background
- **Scene Text Detection**
  - Scene Text Recognition
  - Applications
- Future Trends
Scene Text Detection

- **Proposal-based method:**
  - Detecting text with a single deep neural network (TextBoxes)[1]

- **Part-based method:**
  - Detecting text with Segments and Links (SegLink)[2]

TextBoxes: Horizontal text detection

SSD: Single Shot MultiBox Detector

- Default boxes of different ratios and sizes
- Classify the default boxes
- Regress the matched default boxes

TextBoxes: Horizontal text detection

Long convolutional kernels and default boxes

- Use SSD as the backbone.
- Long default boxes.
- Long convolutional kernels.
**TextBoxes: Horizontal text detection**

### Experimental Results on ICDAR 2013

<table>
<thead>
<tr>
<th>Methods</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaderberg IJCV16</td>
<td>0.89</td>
<td>0.68</td>
<td>0.77</td>
</tr>
<tr>
<td>FCRN CVPR16</td>
<td>0.92</td>
<td>0.76</td>
<td>0.83</td>
</tr>
<tr>
<td>Zhang CVPR16</td>
<td>0.88</td>
<td>0.8</td>
<td>0.84</td>
</tr>
<tr>
<td>SSD</td>
<td>0.80</td>
<td>0.60</td>
<td>0.69</td>
</tr>
<tr>
<td><strong>TextBoxes</strong></td>
<td><strong>0.89</strong></td>
<td><strong>0.83</strong></td>
<td><strong>0.86</strong></td>
</tr>
</tbody>
</table>
TextBoxes++: Multi-oriented text detection

\[(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)\) denote coordinates of the bounding box

Text/Non-text Classification

\[(x, y)\]
TextBoxes++: Multi-oriented text detection

Text detection results on ICDAR 2015 Incidental Text

<table>
<thead>
<tr>
<th>Methods</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>SegLink CVPR17</td>
<td>0.768</td>
<td>0.731</td>
<td>0.75</td>
<td>8.9</td>
</tr>
<tr>
<td>EAST CVPR17</td>
<td>0.735</td>
<td>0.836</td>
<td>0.782</td>
<td>13.2</td>
</tr>
<tr>
<td>EAST multi-scale CVPR17</td>
<td>0.783</td>
<td>0.833</td>
<td>0.807</td>
<td>--</td>
</tr>
<tr>
<td>TextBoxes++</td>
<td>0.767</td>
<td>0.872</td>
<td>0.817</td>
<td>11.6</td>
</tr>
<tr>
<td>TextBoxes++_multi-scale*</td>
<td>0.785</td>
<td>0.878</td>
<td>0.829</td>
<td>--</td>
</tr>
</tbody>
</table>

* multi-scale: Testing image with multi-scale inputs
TextBoxes++: Long text line detection

Inception block for long text lines

- 3*3 conv
- 1*5 conv
- 5*1 conv

Inception block contains rich receptive fields

Concat
**TextBoxes++: Long text line detection**

A subset of RCTW which mainly consists of images with long text lines.

<table>
<thead>
<tr>
<th>Method</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.6902</td>
</tr>
<tr>
<td>Baseline + inception block</td>
<td><strong>0.7532</strong></td>
</tr>
</tbody>
</table>
## TextBoxes++: Long text line detection

### Comparison with competition winners

<table>
<thead>
<tr>
<th>Team Name</th>
<th>Max F-measure</th>
<th>FM-Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foo &amp; Bar</td>
<td>0.661054</td>
<td>1</td>
</tr>
<tr>
<td>NLPR_PAL</td>
<td>0.657598</td>
<td>2</td>
</tr>
<tr>
<td>gmh</td>
<td>0.636024</td>
<td>3</td>
</tr>
<tr>
<td><strong>TextBoxes++ with inception block</strong></td>
<td><strong>0.665295</strong></td>
<td>--</td>
</tr>
</tbody>
</table>
Scene Text Detection

- **Proposal-based method:**
  - Detecting text with a single deep neural network (TextBoxes)[1]

- **Part-based method:**
  - Detecting text with Segments and Links (SegLink)[2]

---

SegLink: Detect long text with segments and links

Large aspect ratio text lines can be detected using limited respective field with Segments and Links

Segments (yellow boxes) + Links (green edges) → Combined detection boxes

Xiang Bai, Kyoto, November 15
SegLink: Detect long text with segments and links

- Fully connected networks based on SSD and VGG16.
- Multiscale Segments and Links prediction
- Alternative solution to the limited respective field problem of long text lines
# SegLink: Detect long text with segments and links

## Results on MSRA-TD500

<table>
<thead>
<tr>
<th>Methods</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kang et al. (CVPR 2014)</td>
<td>71</td>
<td>62</td>
<td>66</td>
</tr>
<tr>
<td>Yin et al. (TPAMI 2015)</td>
<td>81</td>
<td>63</td>
<td>74</td>
</tr>
<tr>
<td>Zhang et al. (CVPR 2016)</td>
<td>83</td>
<td>67</td>
<td>74</td>
</tr>
<tr>
<td>SegLink</td>
<td><strong>86</strong></td>
<td><strong>70</strong></td>
<td><strong>77</strong></td>
</tr>
</tbody>
</table>

## Results on ICDAR2015

<table>
<thead>
<tr>
<th>Methods</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>StradVision-2</td>
<td><strong>77.5</strong></td>
<td>36.7</td>
<td>49.8</td>
</tr>
<tr>
<td>CTPN</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>SegLink</td>
<td>73.1</td>
<td><strong>76.8</strong></td>
<td><strong>75.0</strong></td>
</tr>
</tbody>
</table>

---

Xiang Bai, Kyoto, November 15
SegLink: Detect long text with segments and links

Seglink can detect text of curved shape
Outline

- Background
- Scene Text Detection
- Scene Text Recognition
- Applications
- Future Trends
Scene Text Recognition

- **CRNN** model for Regular Text Recognition
- **RARE** model for Irregular Text Recognition


**CRNN for Regular Text Recognition**

The Network Architecture

**Network Structure**
- Convolutional layers extract feature maps
- Convert feature maps into feature sequence
- Sequence labeling with LSTM
- Translate labels to text

**CTC**

**RNN**

**CNN**

Xiang Bai, Kyoto, November 15
CRNN for Regular Text Recognition

Sequence Modeling

Feature Sequence

Receptive field
CRNN for Regular Text Recognition

Comparisons

Advantages
- End-to-end trainable
- Free of char-level annotations
- Unconstrained to specific lexicon
- 40~50 times less parameters than mainstream models
- Better or comparable performance with state-of-the-arts

Results (lexicon-free)

<table>
<thead>
<tr>
<th>Method</th>
<th>IIIT5K</th>
<th>SVT</th>
<th>IC03</th>
<th>IC13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bissacco et al. (ICCV13)</td>
<td>-</td>
<td>78.0</td>
<td>-</td>
<td>87.6</td>
</tr>
<tr>
<td>Jaderberg et al. (IJCV15)*</td>
<td>-</td>
<td>80.7</td>
<td>93.1</td>
<td>90.8</td>
</tr>
<tr>
<td>Jaderberg et al. (ICLR15)</td>
<td>-</td>
<td>71.7</td>
<td>89.6</td>
<td>81.8</td>
</tr>
<tr>
<td>Proposed</td>
<td>81.2</td>
<td><strong>82.7</strong></td>
<td><strong>91.9</strong></td>
<td><strong>89.6</strong></td>
</tr>
</tbody>
</table>

*is not lexicon-free, as its outputs are constrained to a 90k dictionary

Xiang Bai, Kyoto, November 15
Scene Text Recognition

- **CRNN** model for Regular Text Recognition
- **RARE** model for Irregular Text Recognition


RARE for Irregular Text Recognition

Motivation

Perspective and curved texts are hard to recognize!

(a) Perspective texts

(b) Curved texts
RARE for Irregular Text Recognition

Attention-based Sequence Recognition

- **SRN**: an attention-based encoder-decoder framework
  - **Encoder**: ConvNet + Bi-LSTM
  - **Decoder**: Attention-based character generator

Results

<table>
<thead>
<tr>
<th>Method</th>
<th>IIIT5K</th>
<th>SVT</th>
<th>IC03</th>
<th>IC13</th>
<th>SVT-Per</th>
<th>CUTE80</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRN</td>
<td>83.6</td>
<td>84.9</td>
<td>93.6</td>
<td>91.8</td>
<td>68.2</td>
<td>62.5</td>
</tr>
</tbody>
</table>
RARE for Irregular Text Recognition

STN (Spatial Transform Network)\cite{1} for Text Rectification

- An end-to-end trainable network
  - **STN**: rectifies images with spatial transformation
  - **SRN**: an attention-based encoder-decoder framework

---


Xiang Bai, Kyoto, November 15
RARE for Irregular Text Recognition

Spatial Transformer Network (STN)

- **Localization Network**: A CNN that predicts the fiducial points.

RARE for Irregular Text Recognition

Spatial Transformer Network (STN)

- Grid Generator: Computes a Thin-Plate-Spline (TPS) transform, $T$, from the fiducial points $C$.
- Sampler: TPS-Transform input image $I$ into rectified $I'$.

Results

<table>
<thead>
<tr>
<th>Method</th>
<th>IIIT5K</th>
<th>SVT</th>
<th>IC03</th>
<th>IC13</th>
<th>SVT-Per</th>
<th>CUTE80</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRN</td>
<td>83.6</td>
<td>84.9</td>
<td>93.6</td>
<td>91.8</td>
<td>68.2</td>
<td>62.5</td>
</tr>
<tr>
<td>STN+SRN</td>
<td>88.2</td>
<td>86.7</td>
<td>93.4</td>
<td>92.7</td>
<td>76.8</td>
<td>76.7</td>
</tr>
</tbody>
</table>

RARE for Irregular Text Recognition

Supervised STN

- Localization Network
- Grid Generator
- Sampler
- Rectified Image $I'$

Input Image $I$

- Synthetic dataset with fiducial points $\tilde{C}$ to supervise the predicted $C$.

<table>
<thead>
<tr>
<th>Method</th>
<th>IIIT5K</th>
<th>SVT</th>
<th>IC03</th>
<th>IC13</th>
<th>SVT-Per</th>
<th>CUTE80</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRN</td>
<td>83.6</td>
<td>84.9</td>
<td>93.6</td>
<td>91.8</td>
<td>68.2</td>
<td>62.5</td>
</tr>
<tr>
<td>STN+SRN</td>
<td>88.2</td>
<td>86.7</td>
<td>93.4</td>
<td>92.7</td>
<td>76.8</td>
<td>76.7</td>
</tr>
<tr>
<td>STN(Supervised)+SRN</td>
<td>88.8</td>
<td>87.9</td>
<td>94.1</td>
<td>94.0</td>
<td>77.7</td>
<td>78.8</td>
</tr>
</tbody>
</table>

Xiang Bai, Kyoto, November 15
### RARE for Irregular Text Recognition

#### Rectification Visualization

<table>
<thead>
<tr>
<th>SVT-Perspective</th>
<th>Prediction Groundtruth</th>
<th>CUTE80</th>
<th>Prediction Groundtruth</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input</strong></td>
<td><strong>Rectified</strong></td>
<td><strong>Input</strong></td>
<td><strong>Rectified</strong></td>
</tr>
<tr>
<td><img src="image" alt="Restaurant" /></td>
<td>restaurant</td>
<td><img src="image" alt="Mercato" /></td>
<td>mercato</td>
</tr>
<tr>
<td><img src="image" alt="Quiznos" /></td>
<td>quiznos</td>
<td><img src="image" alt="Football" /></td>
<td>football</td>
</tr>
<tr>
<td><img src="image" alt="Sheraton" /></td>
<td>sheraton</td>
<td><img src="image" alt="Naval" /></td>
<td>naval</td>
</tr>
<tr>
<td><img src="image" alt="Mobil" /></td>
<td>mobil</td>
<td><img src="image" alt="Grove" /></td>
<td>grove</td>
</tr>
<tr>
<td><img src="image" alt="Jewelry" /></td>
<td>jewelry</td>
<td><img src="image" alt="Loka" /></td>
<td>loka</td>
</tr>
<tr>
<td><img src="image" alt="Public" /></td>
<td>public</td>
<td><img src="image" alt="Loka" /></td>
<td>loka</td>
</tr>
</tbody>
</table>
Recognition is helpful to detection

Detector

Recognizer

Detection Score

Recognition Score

Combined Score

Combine

Xiang Bai, Kyoto, November 15
Combination of TextBoxes++ and CRNN

- Detection and recognition are combined by

\[ S = \frac{2 \times \exp(S_d + S_r)}{\exp(S_d) + \exp(S_r)}, \]

Detection score | Recognition score

Text detection results on ICDAR 2015 Incidental Scene Text dataset

<table>
<thead>
<tr>
<th>Methods</th>
<th>Recall</th>
<th>Precision</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detection</td>
<td>0.785</td>
<td>0.878</td>
<td>0.829</td>
</tr>
<tr>
<td>Detection + Recognition</td>
<td>0.792</td>
<td>0.912</td>
<td>0.848</td>
</tr>
</tbody>
</table>

Xiang Bai, Kyoto, November 15
Outline

- Background
- Scene Text Detection
- Scene Text Recognition
- Applications
- Future Trends
Applications

- Fine-Grained Image Classification with Textual Cue
- Number-based Person Re-Identification
- From Text Recognition to Person Re-Identification
Visual cues would group (a)-(b) whereas scene would group (b)-(c).

Texts in images can improve the performance of fine-grained image classification.

Fine-Grained Image Classification with Textual Cue

Pipeline


Xiang Bai, Kyoto, November 15
Fine-Grained Image Classification with Textual Cue

Attention Model to Select Relevant Words

Repair shop

Hotel

➢ Some irrelevant words to this Category
Fine-Grained Image Classification with Textual Cue

Con-Text dataset[1]
- 28 categories of Scenes
- 24,255 images in total

Drink Bottledataset[2]
- Selected from ImageNet
- 20 categories of Drink Bottles
- 18,488 images in total


Xiang Bai, Kyoto, November 15
## Results: mAP(%) improvement on different datasets

<table>
<thead>
<tr>
<th>Method</th>
<th>Dataset</th>
<th>Con-Text</th>
<th>Drink Bottle</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoogLeNet[1]</td>
<td></td>
<td>61.3</td>
<td>63.1</td>
</tr>
<tr>
<td>GoogLeNet + Textual Cue</td>
<td></td>
<td>79.6 (+18.3)</td>
<td>72.8 (+9.7)</td>
</tr>
</tbody>
</table>

Visualization: learned weights of recognized words

- **BAKERY**
  - Cakes: 0.57
  - Pastries: 0.43
  - Open: 5.5e-9
  - ...

- **CAFE**
  - Starbucks: 1
  - Scoff: 1.1e-8
  - ...

- **ROOTBEER**
  - Root: 0.89
  - Beer: 0.11
  - Brewed: 1.3e-6
  - ...

- **CHABLIS**
  - Chablis: 0.99
  - France: 8.7e-12
  - ...

- **Filter** the incorrect recognized words
- **Select** more related words to the category
Fine-Grained Image Classification with Textual Cue

Results of Image Search

Visual cue only

GoogLeNet 48.0
GoogLeNet+Textual Cue 60.8 (+12.8)

Retrieval Results

Visual and Textual Cues

GoogLeNet 48.0
GoogLeNet+Textual Cue 60.8 (+12.8)
Applications

- Fine-Grained Image Classification with Textual Cue
- Number-based Person Re-Identification
- From Text Recognition to Person Re-Identification
Number-based Person Re-Identification

- **Problem:** hard to track and retrieve an athlete in a marathon game

Marathon athlete tracking

query retrieval

Marathon image database

- **Motivation:** every athlete has a unique racing bib number

Xiang Bai, Kyoto, November 15
Number-based Person Re-Identification

Proposed pipeline


Marathon Image → Text Detection → Localization → Text Recognition → Recognition → Result

Input Query: 2895

match
Number-based Person Re-Identification

Marathon Dataset

8706 training images, 1000 testing images

Experimental Results

Identification accuracy rate ($\text{Id\_acc}$): $85\%$

$$
\text{Id\_acc} = \frac{\text{Num(correctly recognized persons)}}{\text{Num(total persons)}}
$$

Xiang Bai, Kyoto, November 15
Applications

- Fine-Grained Image Classification with Textual Cue
- Number-based Person Re-Identification
- From Text Recognition to Person Re-Identification
From Text Recognition to Person Re-Identification

Sequence Modeling

Text Recognition (CRNN)

Person Re-Identification

Head
Upper body
Upper leg
Lower leg

Left-to-Right

Top to Down


Xiang Bai, Kyoto, November 15
From Text Recognition to Person Re-Identification

Model Architecture

CNN + LSTM

Results on Market1501[1]

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP(%)</th>
<th>R1(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>59.8</td>
<td>81.4</td>
</tr>
<tr>
<td>CNN + LSTM</td>
<td>65.5</td>
<td>85.8</td>
</tr>
</tbody>
</table>

R1: given a query, precision of the top-1 similar image from gallery discriminated by model.


Xiang Bai, Kyoto, November 15
From Text Recognition to Person Re-Identification

Retrival Results

CNN

query

CNN+LSTM

query
Outline

- Background
- Scene Text Detection
- Scene Text Recognition
- Applications
- Future Trends
Future Trends

- Irregular text detection (Curved & Perspective Text Lines)
- Multilingual End-to-end text recognition
- Semi-supervised or weakly supervised text detection and recognition
- Text image synthesis (GAN)
- Unified framework for OCR and NLP
- Integrating Scene text and Image/Videos for many applications.
Resources (Papers & Datasets & Codes)

- B. Shi, C. Yao, M. Liao, M Yang, P Xu, L Cui, S Belongie, S Lu, X Bai.
  ICDAR2017 Competition on Reading Chinese Text in the Wild (RCTW-17). ICDAR’17
  Dataset: http://mclab.eic.hust.edu.cn/icdar2017chinese

- B. Shi, X. Bai, S. Belongie.
  Detecting Oriented Text in Natural Images by Linking Segments. CVPR'17
  Code: https://github.com/bgshih/seglink

  TextBoxes: A fast text detector with a single deep neural network. AAAI'17
  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’17
  Code: http://mclab.eic.hust.edu.cn/~xbai/CRNN/crnn_code.zip

- B. Shi, X. Wang, P. Lyu, C. Yao, X. Bai.
  Robust scene text recognition with automatic rectification. CVPR’16


Xiang Bai, Kyoto, November 15
Literature review (Papers & PPTs)

- [Survey Paper]
  Scene text detection and recognition: Recent advances and future trends.
  Y Zhu, C Yao, X Bai.

- [Talk PPT in 2014]
  Representation in Scene Text Detection and Recognition.

- [Talk PPT in 2017]
  Oriented Scene Text Detection Revisited.
Refer to my homepage for more details

Thank you!

Xiang Bai, Kyoto, November 15