華中科技大學
Huazhong University of Science & Technology

Deep Neural Networks for Scene Text Reading

Xiang Bai

Huazhong University of Science and Technology

Problem definitions

Definitions

Summary Booklet

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

+

Scene text recognition: Converting text regions into computer readable and editable symbols

→ End-to-end recognition

Outline

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

Background

Cuvette Thickness (mm)	n_2(toluene)/n_2(glycerine)	n_2(methanol)/n_2(water)
1	14.1	1.05
2	14.6	1.07
5	14.4	1.06
10	14.2	1.07
20	14.0	1.07
Average	14.3	1.06

Document image

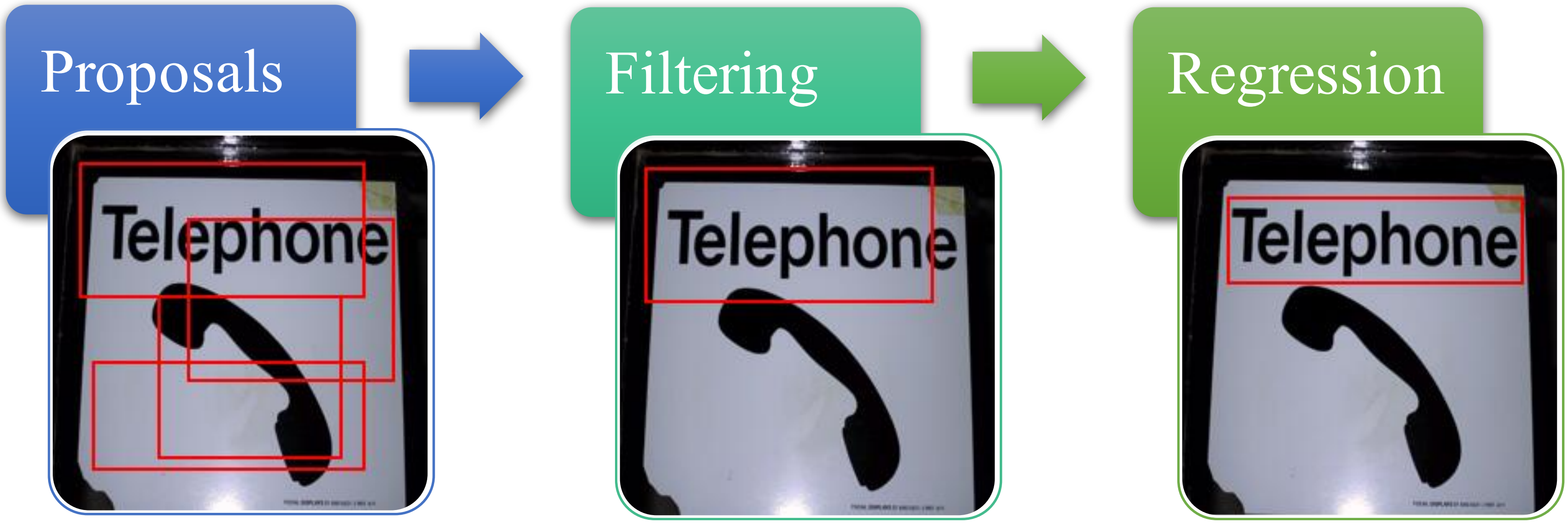
VS

Scene text image

- Scattered and sparse
- Multi-oriented
- Multi-lingual

Background

Scene text detection methods before 2016



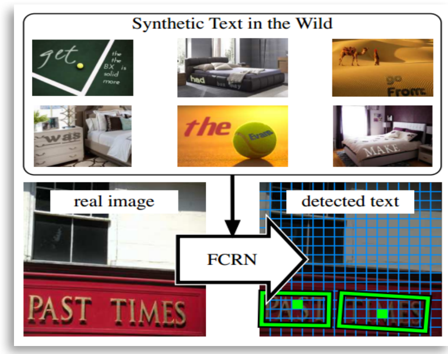
- Generate candidates using **hand-craft features**
- Text / non-text classification using **CNN/Random forest**
- Refine locations using **CNN**

[1] Jaderberg et al. Deep features for text spotting. ECCV, 2014.
[2] Jaderberg et al. Reading text in the wild with convolutional neural networks. IJCV, 2016.
[3] Huang et al. Robust scene text detection with convolution neural network induced mser trees. ECCV, 2014.
[4] Zhang et al. Symmetry-based text line detection in natural scenes. CVPPR, 2015.
[5] LGómez, D Karatzas. Textproposals: a text-specific selective search algorithm for word spotting in the wild. Pattern Recognition 70, 60-74

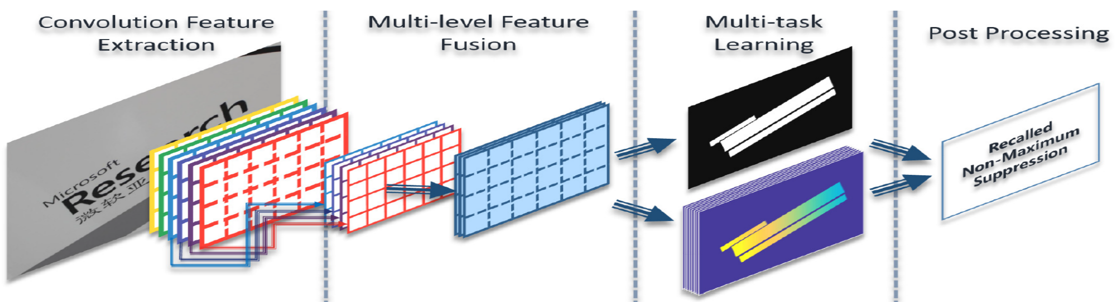
Background

Scene text detectionmethods after 2016

Segmentation-based method[1]



Proposal-based method[2]

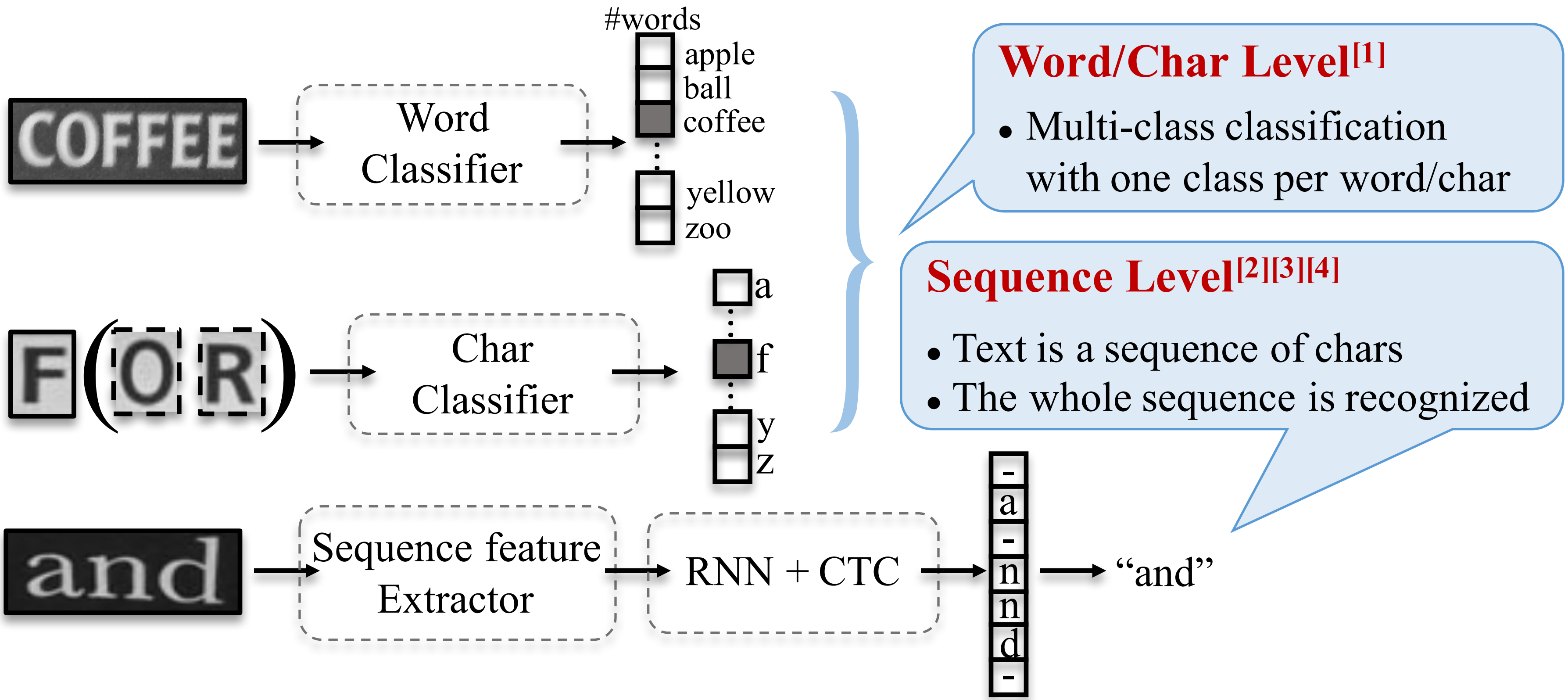


Hybrid method[3]

[1] Zhang Z, et al. Multi-oriented text detection with fully convolutional networks. CVPR, 2016.
[2] Gupta A, et al. Synthetic data for text localisation in natural images. CVPR, 2016.
[3] He W, et al. Deep Direct Regression for Multi-Oriented Scene Text Detection. ICCV, 2017
[4] Liao et al. TextBoxes: A fast text detector with a single deep neural network. AAAI, 2017.

Background

Scene text recognition methods



[1] M. Jaderberg et al. Reading text in the wild with convolutional neural networks. IJCV, 2016.
[2] B. Su et al. Accurate scene text recognition based on recurrent neural network. ACCV, 2014.
[3] He et al. Reading Scene Text in Deep Convolutional Sequences. AAAI, 2016.
[4] Shi B et al. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. TPAMI, 2017.

Background

Recent Trend

Statistics of related papers published in 2017 top conferences

Conference	Detection	Recognition	End-to-end recognition
AAAI-17	0	0	2
IJCAI-17	0	1	0
NIPS-17	0	1	0
ICCV-17	5	1	2
CVPR-17	3	0	0
ICDAR-17	8	2	1
TOTAL	**16**	**5**	**5**

- 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

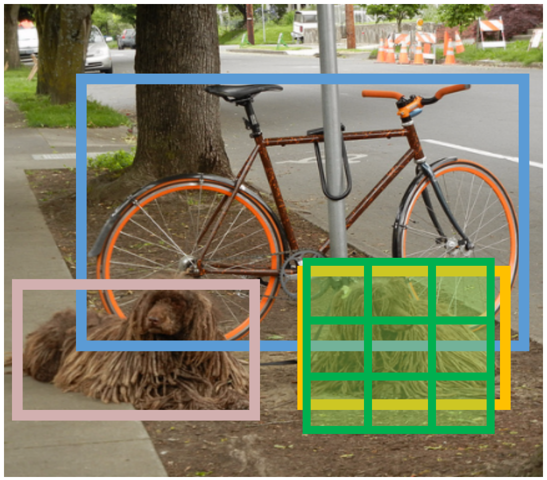
English: there is always a blank space between neighbor words

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

Background

Challenges in Non-Latin text detection



General objects

Words

Text Lines

- 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.

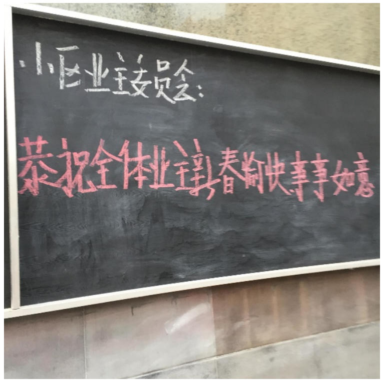
Background

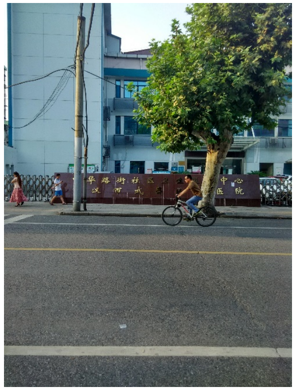
Performance comparison on English / Chinese datasets

Dataset	Language	Num. Train/Test	Best F-measure
ICDAR 2013	English	229/233	0.90
ICDAR 2015	English	1000/500	0.81
RCTW 2017	Mainly Chinese	8034/4229	**0.66**

The performance of Chinese dataset is much lower.

ICDAR 2017 Competition on Reading Chinese Text in the Wild



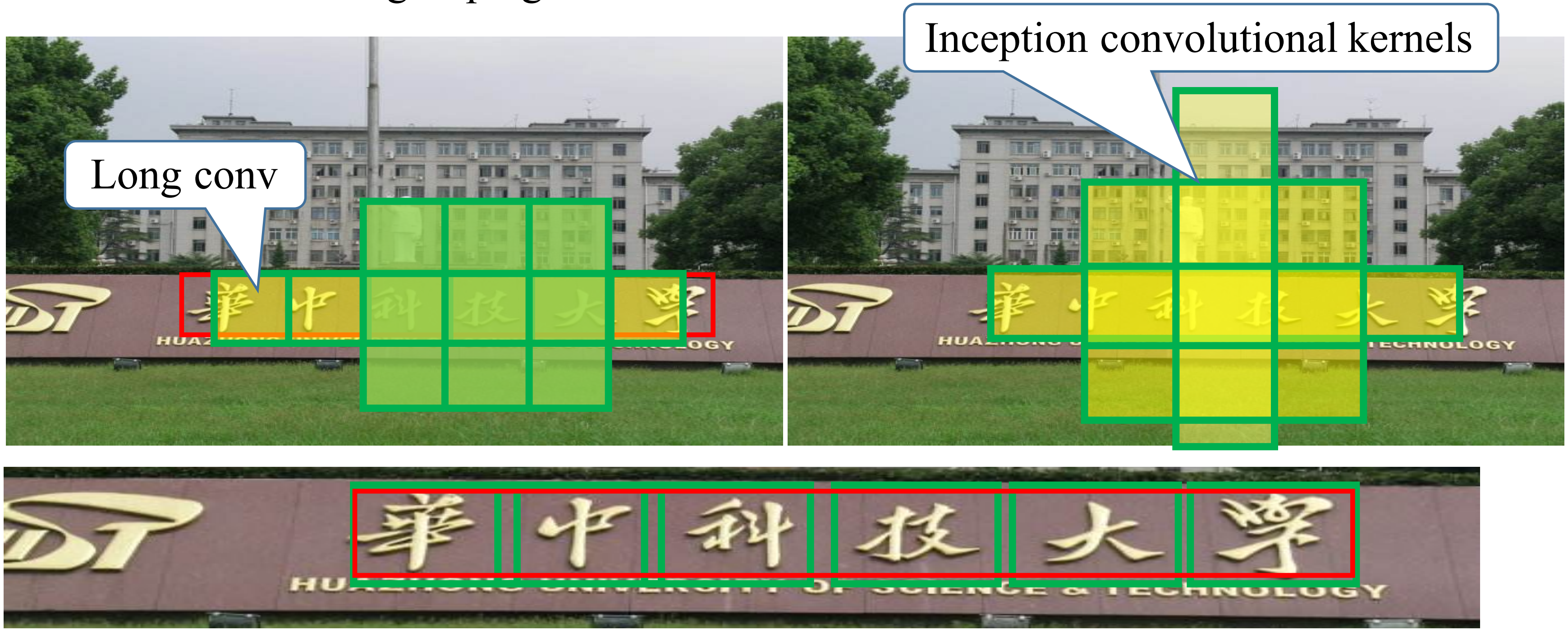


Link: http://mclab.eic.hust.edu.cn/icdar2017chinese/

Background

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]**

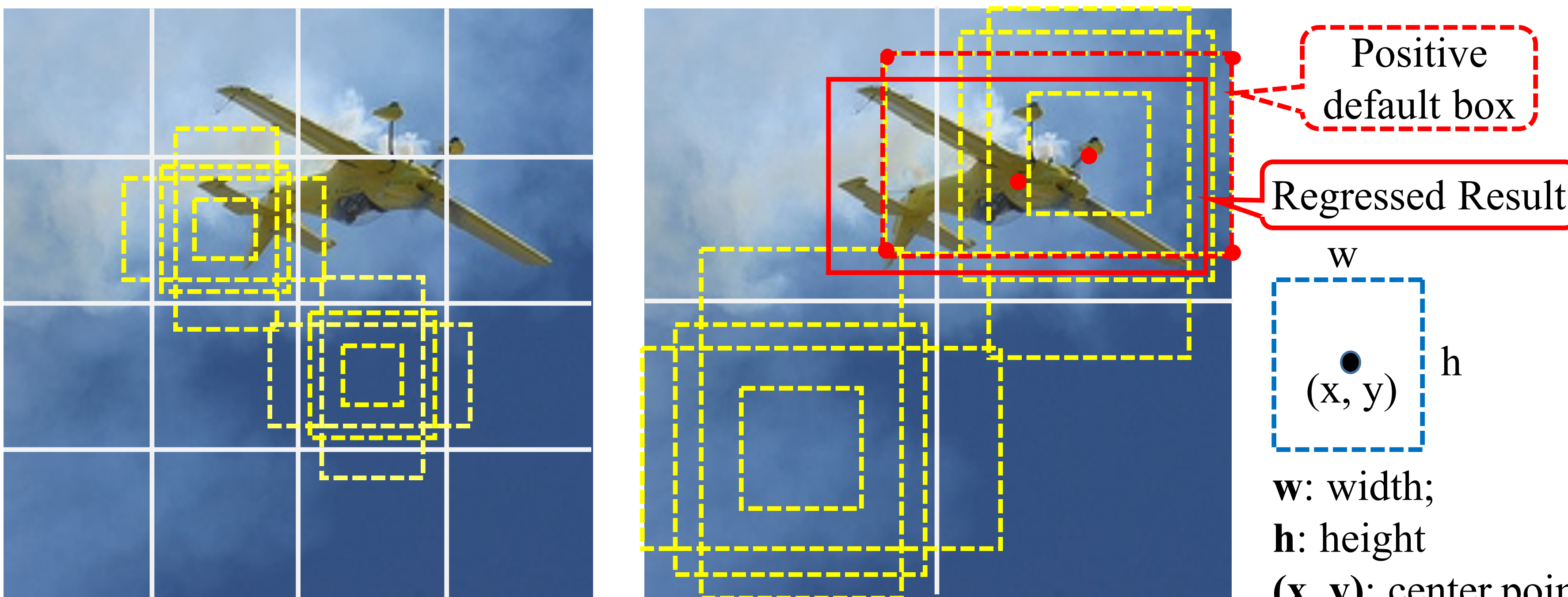
- **Part-based method:**
  - Detecting text with Segments and Links **(SegLink)[2]**

[1] M. Liao et al. TextBoxes: A Fast Text Detector with a Single Deep Neural Network. AAAI, 2017.
[2] B. Shi et al. Detecting Oriented Text in Natural Images by Linking Segments. IEEE CVPR, 2017.

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



w: width;
h: height
(x, y): center point

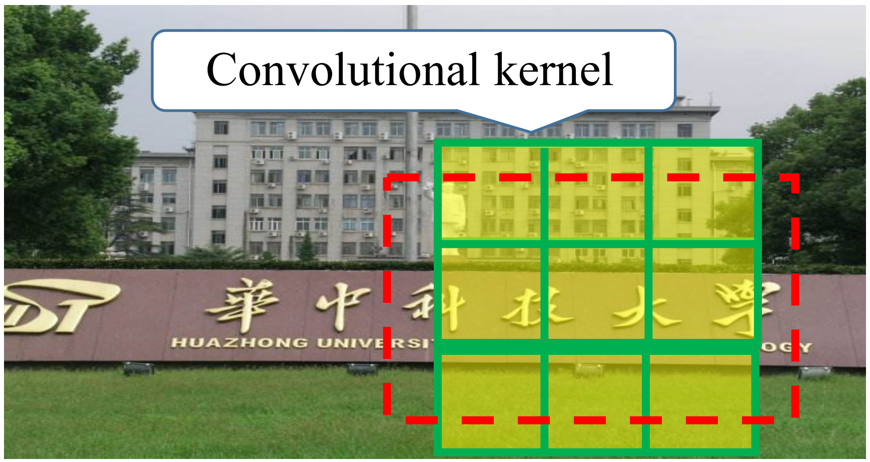
4×4 feature map

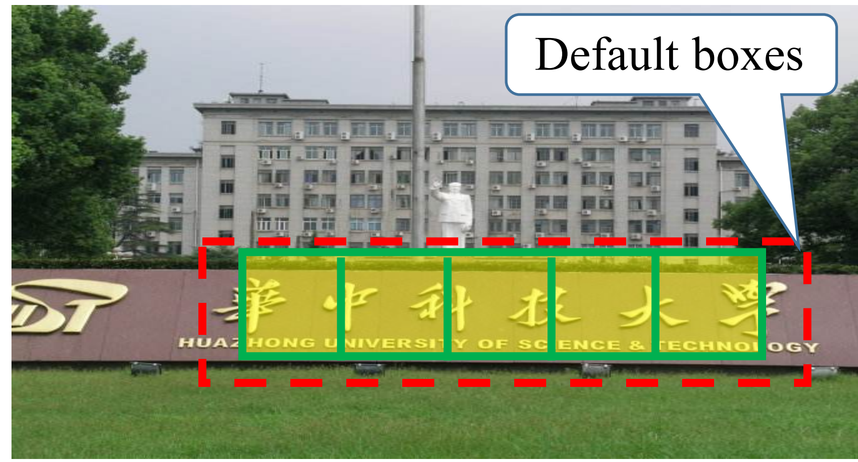
2×2 feature map

[1] W. Liu et al. SSD: Single Shot MultiBox Detector. ECCV, 2016.

TextBoxes: Horizontal text detection

Long convolutional kernels and default boxes





SSD: 3*3 conv → TextBoxes: 1*5 conv

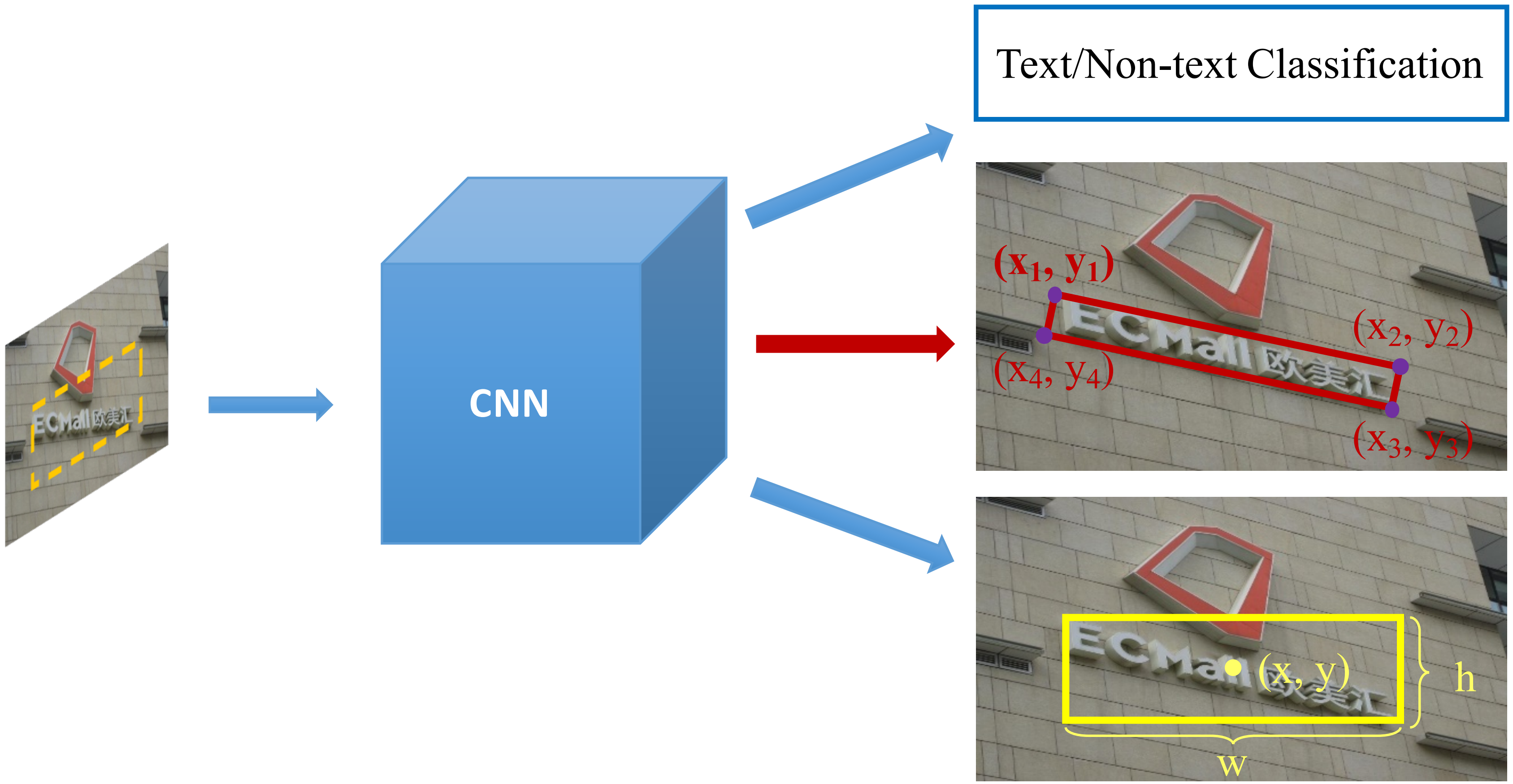
- Use SSD as the backbone.
- Long default boxes.
- Long convolutional kernels.

TextBoxes: Horizontal text detection

Experimental Results on ICDAR 2013

Methods	Precision	Recall	F-measure
Jaderberg IJCV16	0.89	0.68	0.77
FCRN CVPR16	0.92	0.76	0.83
Zhang CVPR16	0.88	0.8	0.84
SSD	0.80	0.60	0.69
TextBoxes	**0.89**	**0.83**	**0.86**

TextBoxes++: Multi-oriented text detection



(x_i, y_i) $(i = 1,2,3,4)$ denote coordinates of the bounding box

TextBoxes++: Multi-oriented text detection

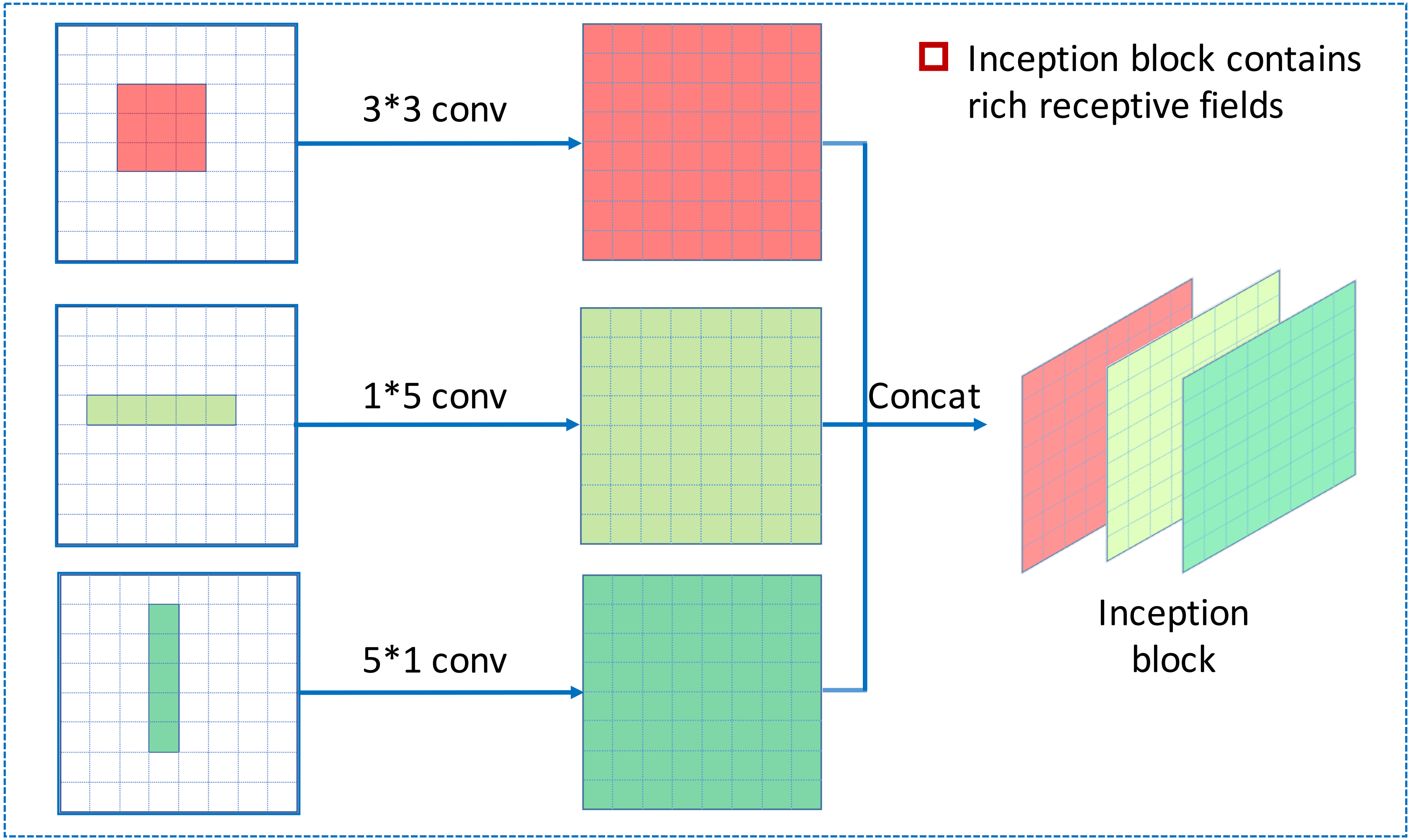
Text detection results on ICDAR 2015 Incidental Text

Methods	Recall	Precision	F-measure	FPS
SegLink CVPR17	0.768	0.731	0.75	8.9
EAST CVPR17	0.735	0.836	0.782	13.2
EAST multi-scale CVPR17	0.783	0.833	0.807	--
TextBoxes++	0.767	0.872	0.817	11.6
TextBoxes++_multi-scale*	**0.785**	**0.878**	**0.829**	--

* multi-scale: Testing image with multi-scale inputs

TextBoxes++: Long text line detection

Inception block for long text lines



- Inception block contains rich receptive fields

TextBoxes++: Long text line detection

Without inception block

With inception block

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

Performances on RCTW (long)

Method	F-measure
Baseline	0.6902
Baseline **+ inception block**	**0.7532**

TextBoxes++: Long text line detection

Comparison with competition winners

Team Name	Max F-measure	FM-Rank
Foo & Bar	0.661054	1
NLPR_PAL	0.657598	2
gmh	0.636024	3
TextBoxes++ with inception block	**0.665295**	**--**

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]**

[1] M. Liao et al. TextBoxes: A Fast Text Detector with a Single Deep Neural Network. AAAI, 2017.
[2] B. Shi et al. Detecting Oriented Text in Natural Images by Linking Segments. IEEE CVPR, 2017.

SegLink: Detect long text with segments and links

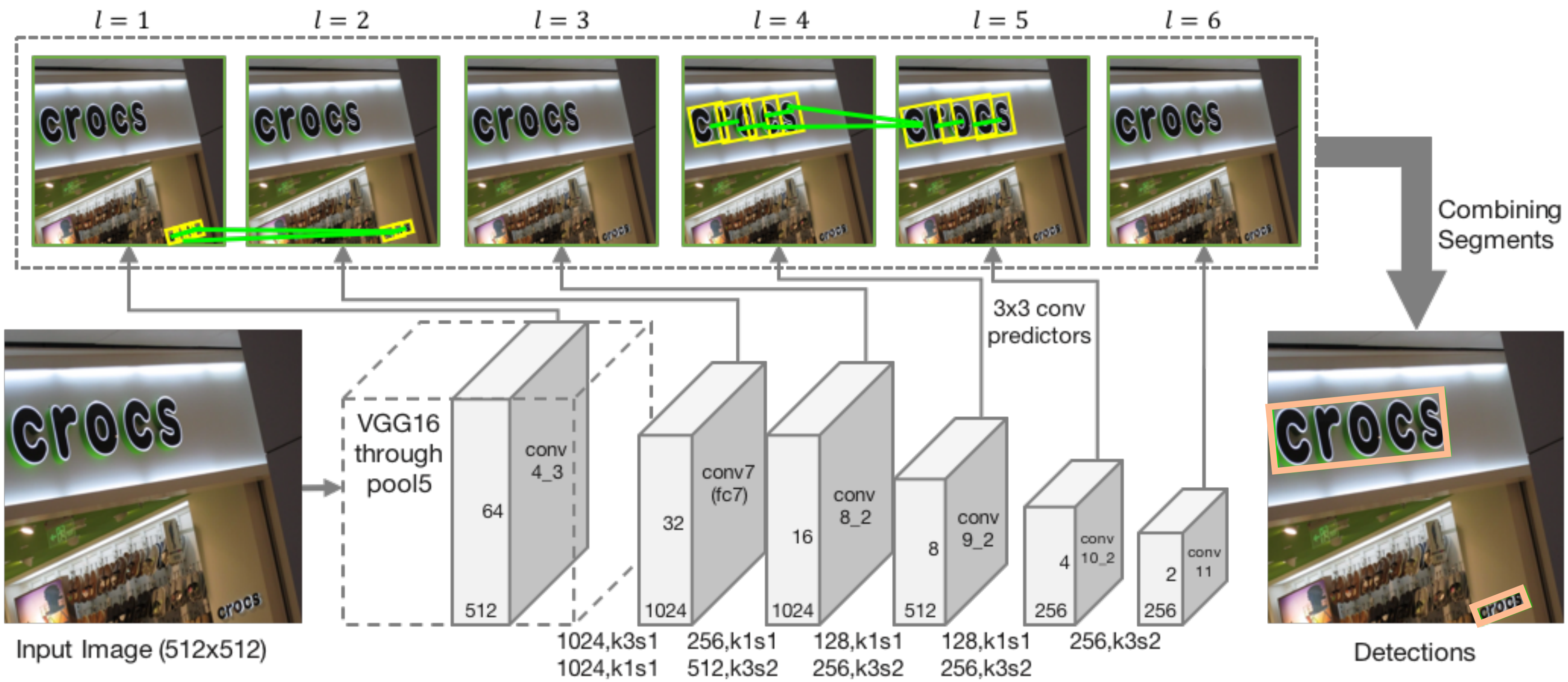
Large aspect ratio text lines can be detected using limited respective field with **Segments** and **Links**

Segments
(yellow boxes)

Links
(green edges)

Combined detection boxes

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

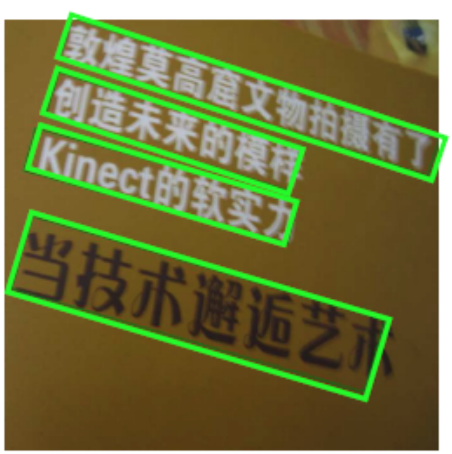
SegLink: Detect long text with segments and links

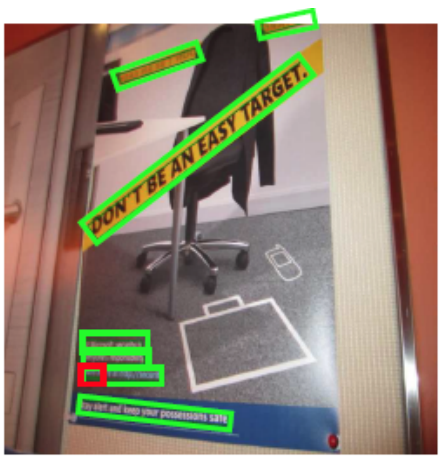
Results on MSRA-TD500

Methods	Precision	Recall	F-measure
Kang et al. (CVPR 2014)	71	62	66
Yin et al. (TPAMI 2015)	81	63	74
Zhang et al. (CVPR 2016)	83	67	74
SegLink	**86**	**70**	**77**

Results on ICDAR2015

Methods	Precision	Recall	F-measure
StradVision-2	**77.5**	36.7	49.8
CTPN	51.6	74.2	60.9
Megvii-Image++	72.4	57.0	63.8
SegLink	73.1	**76.8**	**75.0**





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

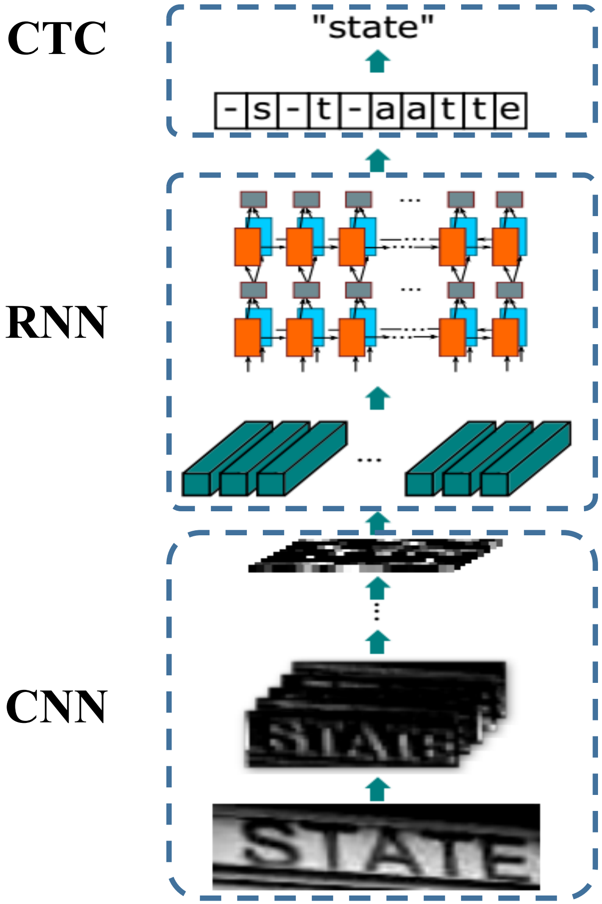
- **CRNN** model for Regular Text Recognition
- **RARE** model for Irregular Text Recognition

[1] CRNN: Shi B et al. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. TPAMI, 2017.

[2] RARE: Shi B et al. Robust scene text recognition with automatic rectification. CVPR, 2016.

CRNN for Regular Text Recognition

The Network Architecture

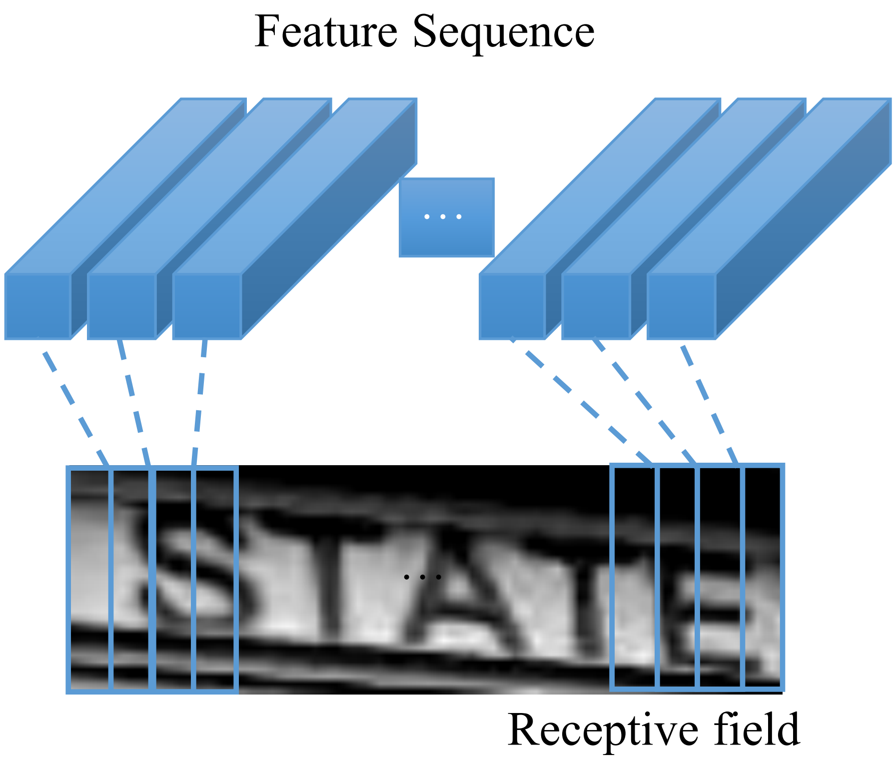


Network Structure

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

CRNN for Regular Text Recognition

Sequence Modeling



CRNN for Regular Text Recognition

Comparisons

Advantages

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

Results(lexicon-free)

Method	IIIT5K	SVT	IC03	IC13
Bissacco et al. (ICCV13)	-	78.0	-	87.6
Jaderberg et al. (IJCV15)*	-	80.7	93.1	90.8
Jaderberg et al. (ICLR15)	-	71.7	89.6	81.8
Proposed	81.2	**82.7**	**91.9**	**89.6**

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

Scene Text Recognition

- **CRNN** model for Regular Text Recognition

- **RARE** model for Irregular Text Recognition

[1] CRNN: Shi B et al. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. TPAMI, 2017.

[2] RARE: Shi B et al. Robust scene text recognition with automatic rectification. CVPR, 2016.

RARE for Irregular Text Recognition

Motivation

Perspective and curved texts are hard to recognize!

(a) Perspective texts

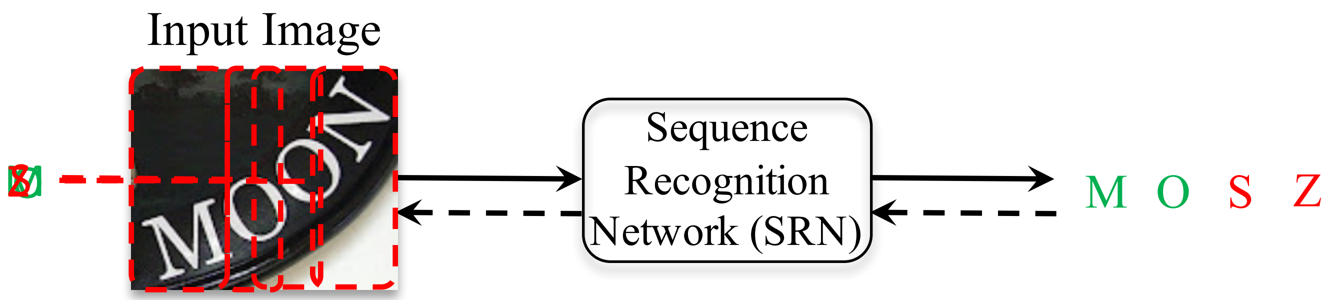
SVT-Perspective

(b) Curved texts

CUTE80

RARE for Irregular Text Recognition

Attention-based Sequence Recognition



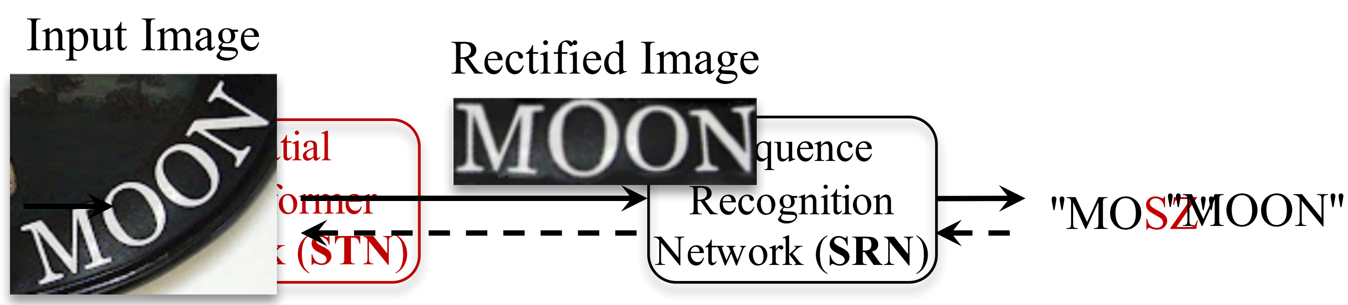
- **SRN**: an attention-based encoder-decoder framework
  - Encoder: ConvNet + Bi-LSTM
  - Decoder: Attention-based character generator

Results

Method	IIIT5K	SVT	IC03	IC13	SVT-Per	CUTE80
SRN	83.6	84.9	93.6	91.8	68.2	62.5

RARE for Irregular Text Recognition

STN (Spatial Transform Network)[1] for Text Rectification

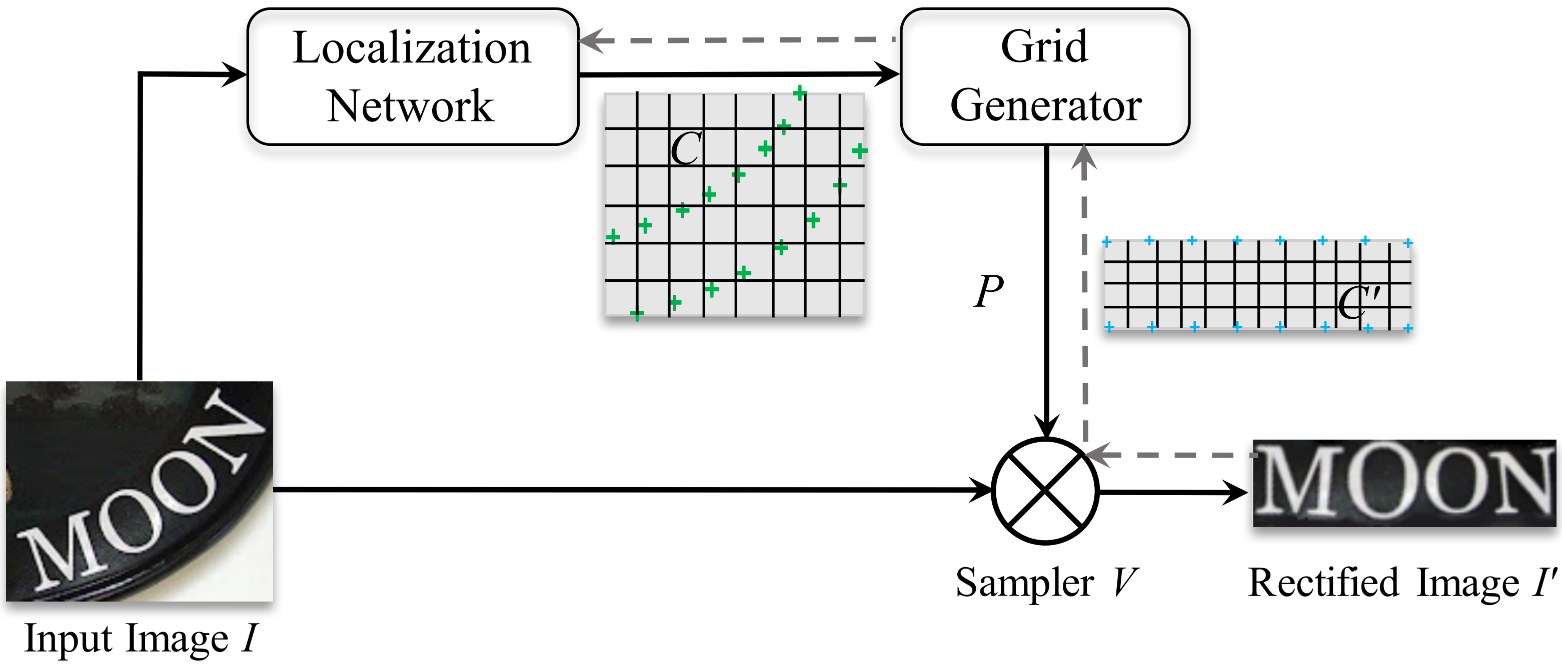


- An end-to-end trainable network
  - **STN**: rectifies images with spatial transformation
  - **SRN**: an attention-based encoder-decoder framework

[1] Jaderberg M et al. Spatial transformer networks. NIPS, 2015.

RARE for Irregular Text Recognition

Spatial Transformer Network (STN)

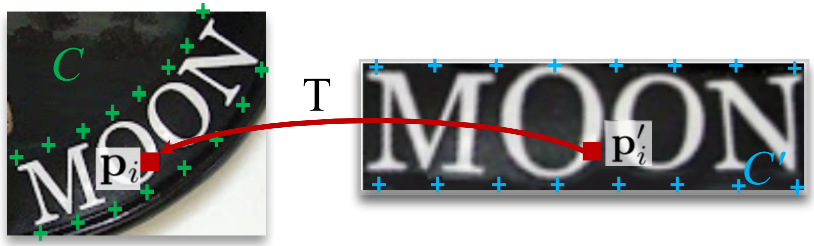


- **Localization Network**: A CNN that predicts the fiducial points.

[1] Jaderberg M et al. Spatial transformer networks. NIPS, 2015.

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'.

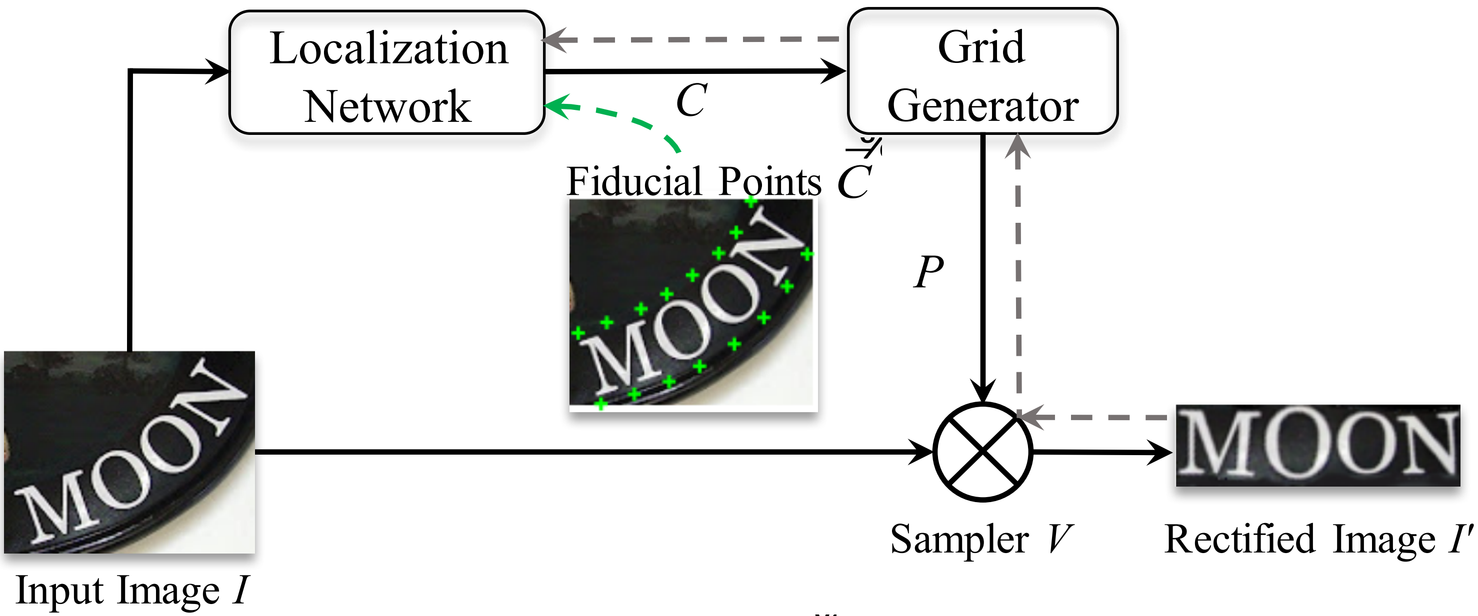
Results

Method	IIIT5K	SVT	IC03	IC13	SVT-Per	CUTE80
	Standard datasets				Deformable text datasets	
SRN	83.6	84.9	**93.6**	91.8	68.2	62.5
STN+SRN	**88.2**	**86.7**	93.4	**92.7**	**76.8**	**76.7**

[1] Jaderberg M et al. Spatial transformer networks. NIPS, 2015.

RARE for Irregular Text Recognition

Supervised STN



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

Method	IIIT5K	SVT	IC03	IC13	SVT-Per	CUTE80
SRN	83.6	84.9	93.6	91.8	68.2	62.5
STN+SRN	88.2	86.7	93.4	92.7	76.8	76.7
STN(Supervised)+SRN	**88.8**	**87.9**	**94.1**	**94.0**	**77.7**	**78.8**

RARE for Irregular Text Recognition

Rectification Visualization

SVT-Perspective

Input | Rectified

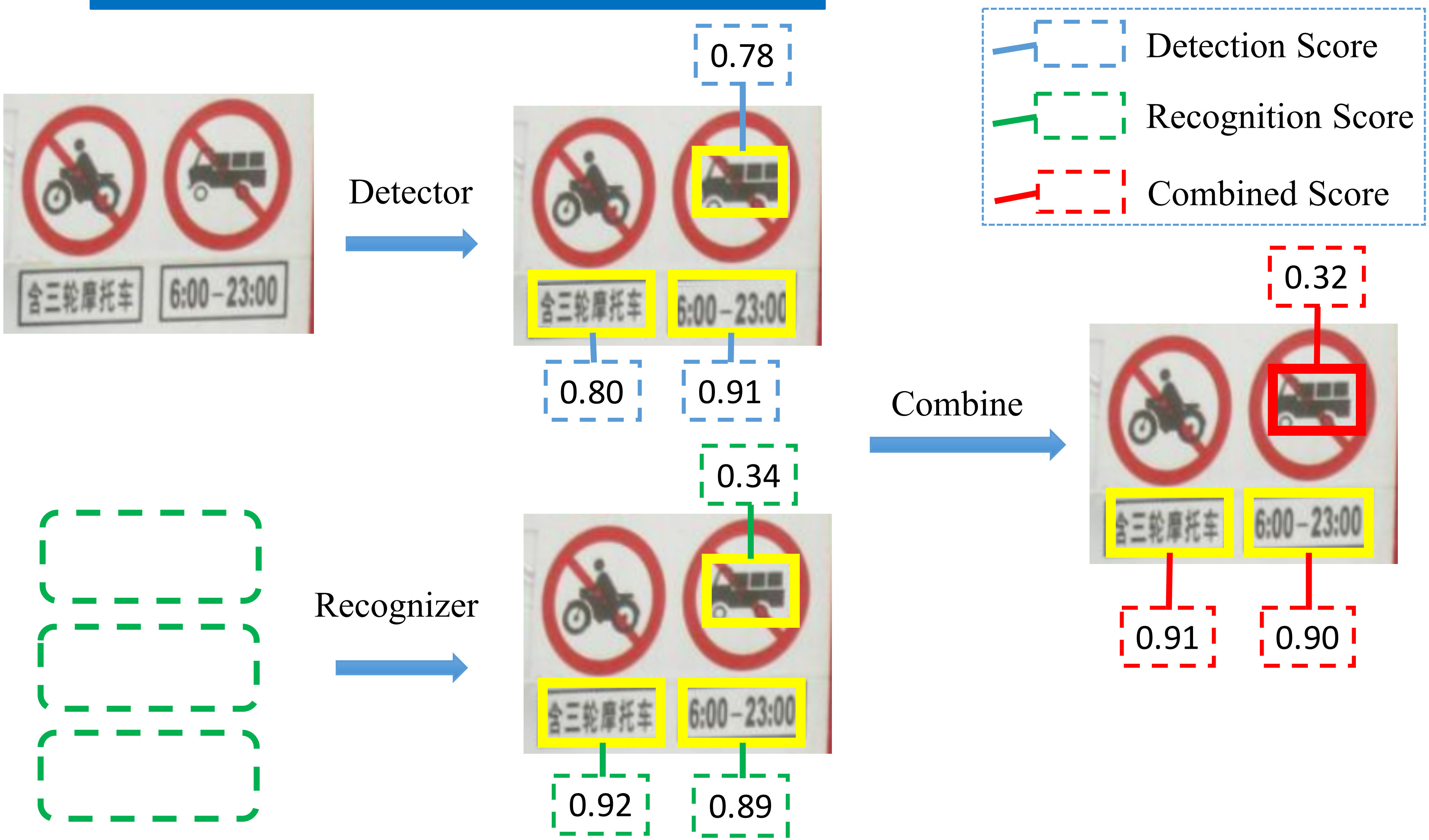
Prediction	Groundtruth
restaurant	restaurant
quiznos	quiznos
sheraton	sheraton
mobil	mobil
jewelry	jewelry
public	public

CUTE80

Input | Rectified

Prediction	Groundtruth
mercato	marcato
football	football
naval	naval
grove	grove
loka	loka

Recognition is helpful to detection



Combination of TextBoxes++ and CRNN

- Detection and recognition are combined by

$$S = \frac{2 * \exp(S_d + S_r)}{\exp(S_d) + \exp(S_r)},$$

Detection score

Recognition score

Text detection results on ICDAR 2015 Incidental Scene Text dataset

Methods	Recall	Precision	F-score
Detection	0.785	0.878	0.829
Detection + Recognition	**0.792**	**0.912**	**0.848**

Outline

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

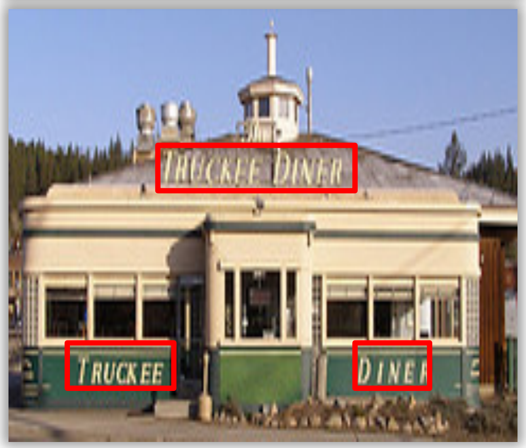
Applications

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

Fine-Grained Image Classification with Textual Cue

Motivations

TRUCKEE DINER



(a)

CAFE

(b)

COFFEE

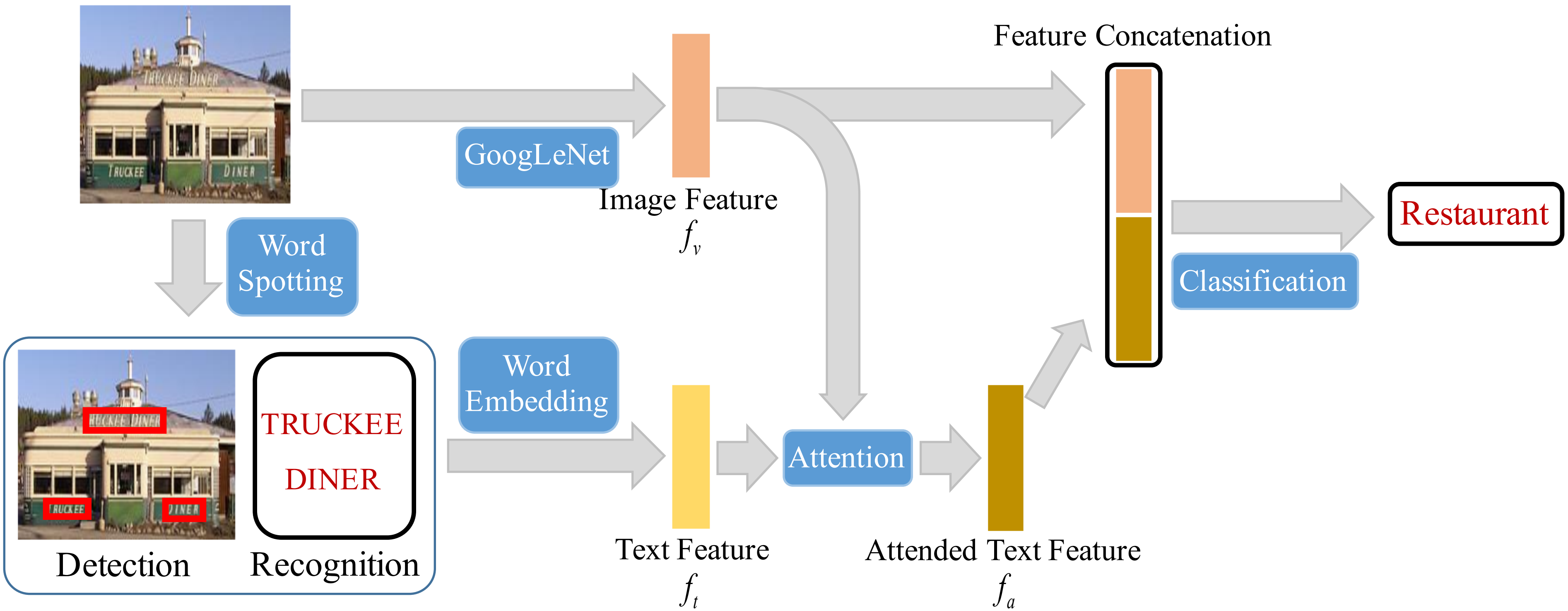
(c)

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

[1] Bai X. et al. Integrating Scene Text and Visual Appearance for Fine-Grained Image Classification with Convolutional Neural Networks[J]. arXiv:1704.04613, 2017.

Fine-Grained Image Classification with Textual Cue

Pipeline



[1] Bai X. et al. Integrating Scene Text and Visual Appearance for Fine-Grained Image Classification with Convolutional Neural Networks[J]. arXiv:1704.04613, 2017.

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

[1] S. Karaoglu. et al. Con-text: text detection using background connectivity for fine-grained object classification. ACM2013
[2] Bai X. et al. Integrating Scene Text and Visual Appearance for Fine-Grained Image Classification with Convolutional Neural Networks[J]. arXiv2017.

Fine-Grained Image Classification with Textual Cue

Results: mAP(%) improvement on different datasets

Method	Dataset	
	Con-Text	Drink Bottle
GoogLeNet[1]	61.3	63.1
GoogLeNet + Textual Cue	**79.6 (+18.3)**	**72.8 (+9.7)**

[1] C. Szegedy, et al. Going deeper with convolutions. CVPR2015

Fine-Grained Image Classification with Textual Cue

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

Root Beer Cream Soda Guinness Slivovitz Ginger ale

Visual and Textual Cues

Root Beer Sarsaparilla

Retrieval Results

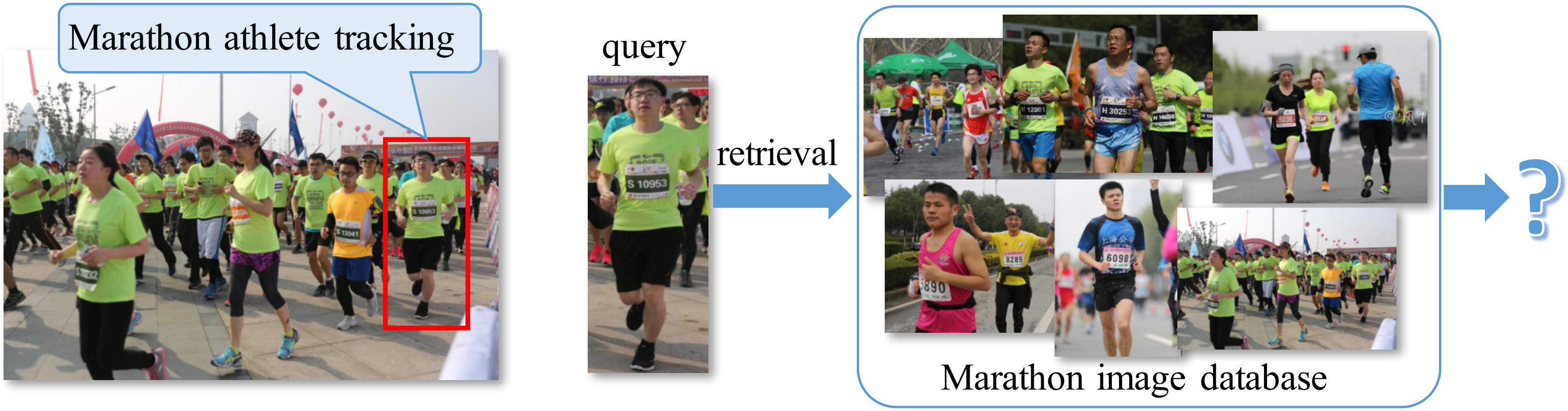
Method	mAP(%)
GoogLeNet	48.0
GoogLeNet+Textual Cue	**60.8 (+12.8)**

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**



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

Number-based Person Re-Identification

Proposed pipeline

Marathon Image | Localization | Recognization | Result

[1] M. Liao et al. TextBoxes: A Fast Text Detector with a Single Deep Neural Network. AAAI, 2017.
[2] Shi B, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. TPAMI, 2017.

Number-based Person Re-Identification

Marathon Dataset

8706 training images, 1000 testing images

Experimental Results

Identification accuracy rate(*Id_acc*): **85%**

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

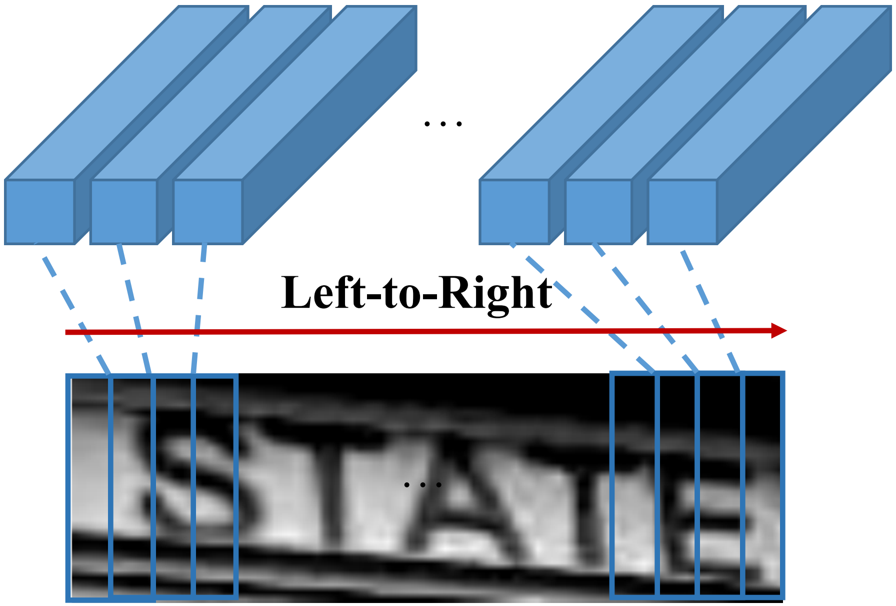
Applications

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

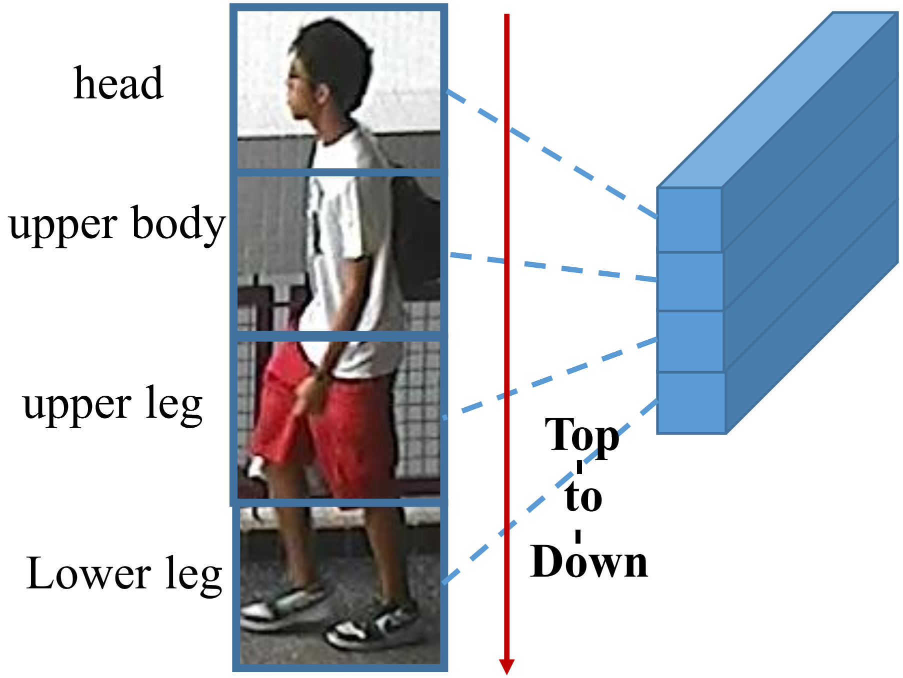
From Text Recognition to Person Re-Identification

Sequence Modeling

Text Recognition (**CRNN**)



Person Re-Idntification

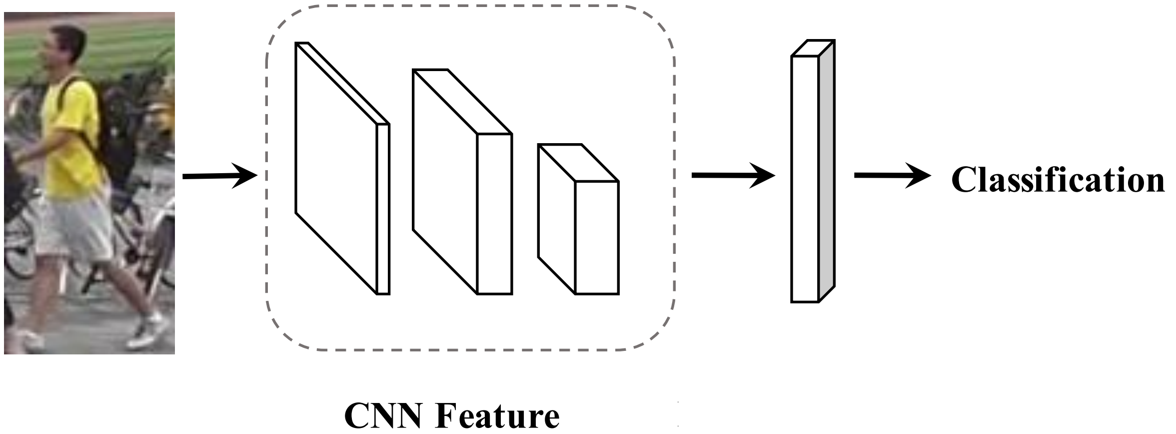


[1] CRNN: Shi B et al. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. TPAMI, 2017.

From Text Recognition to Person Re-Identification

Model Architecture

CNN + LSTM



Results on Market1501[1]

Method	mAP(%)	R1(%)
CNN	59.8	81.4
CNN + LSTM	**65.5**	**85.8**

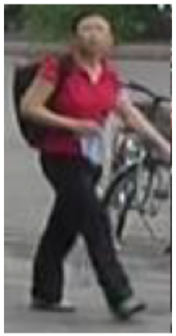
R1: given a query, precision of the top-1 similar image from gallery discriminated by model.

[1] Zheng et al. Scalable Person Re-identification: A Benchmark. ICCV 2015

From Text Recognition to Person Re-Identification

Retrival Results

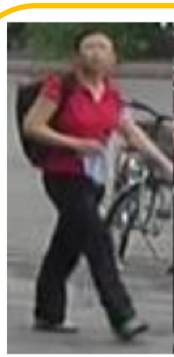
CNN



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

- M. Liao, B. Shi, X. Bai, X. Wang, W. Liu.
  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

- X. Bai, M. Yang, P. Lyu, et al.
  Integrating Scene Text and Visual Appearance for Fine-Grained Image Classification arXiv2017.

Literature review (Papers & PPTs)

- [Survey Paper]
  Scene text detection and recognition: Recent advances and future trends.
  Y Zhu, C Yao, X Bai.
  Frontiers of Computer Science 10 (1), 19–36, 2016.
  http://mclab.eic.hust.edu.cn/UpLoadFiles/Papers/FCS_TextSurvey_2015.pdf

- [Talk PPT in 2014]
  Representation in Scene Text Detection and Recognition.
  http://mclab.eic.hust.edu.cn/~xbai/Talk_slice/Representation%20in%20Scene%20Text%20Detection%20and%20Recognition_20150207.pdf

- [Talk PPT in 2017]
  Oriented Scene Text Detection Revisited.
  http://mclab.eic.hust.edu.cn/~xbai/Talk_slice/Oriented-Scene-Text-Detection-Revisited_VALSE2017.pdf

Collaborators

Cong Yao.
Cloud Team chief, Megvii Inc.

Baoguang Shi.
PHD Candidate, HUST

Zheng Zhang.
Associate Researcher, MSRA

Chengquan Zhang.
Researcher, Baidu IDL

Minghui Liao.
PHD student, HUST

Mingkun Yang.
Master student, HUST

Serge Belongie.
Professor, Cornell

Refer to my homepage for more details

Thank you !