

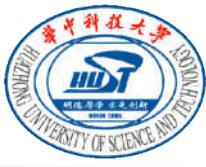
Oriented Scene Text Detection Revisited

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



Summary Booklet

Text Detection
Word/line level

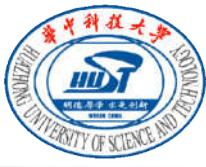
Text Recognition
Word/sequence classification

End-to-end Recognition

How do we perceive scene text?

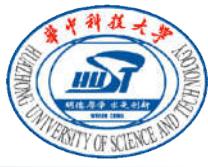
Top-Down vs. Bottom-Up,
which is better?





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



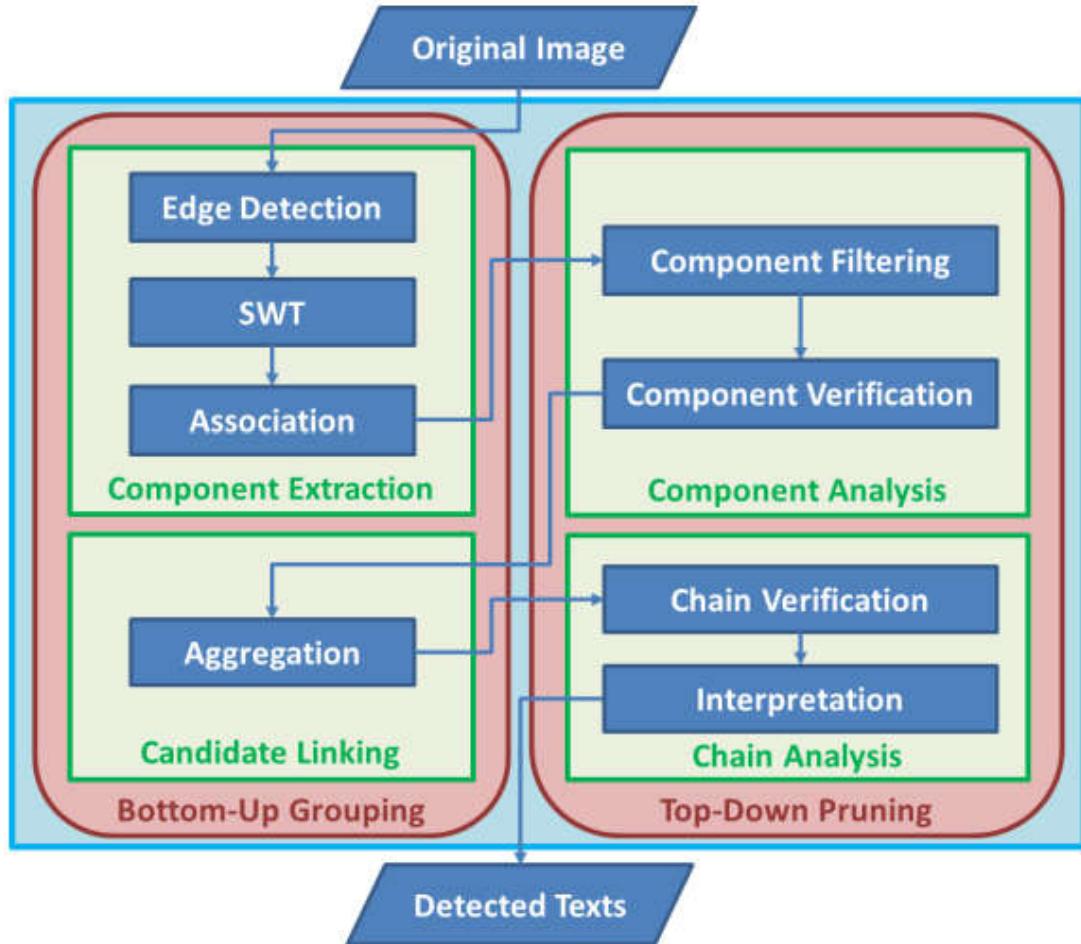
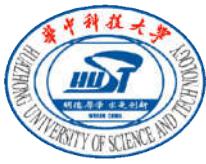
The Story of Oriented Scene Text Detection

➤ Handcraft Features

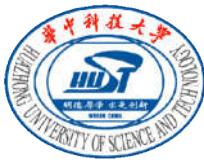
- Component level. MSER, SWT...
- Word / line level. Sliding Window

➤ Deep Learning (2014-)

- Region Proposals
- Segmentation
- Hybrid Methods



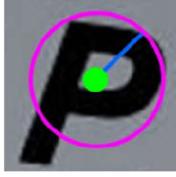
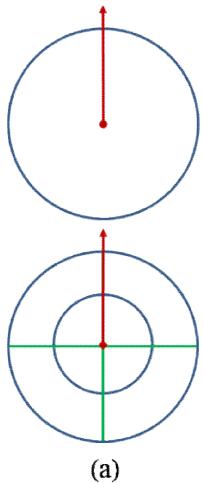
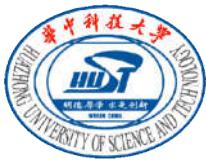
- Specially designed features.
- Two-level classification scheme.
- The 1st benchmark dataset for multi-oriented text detection: MSRA-TD 500



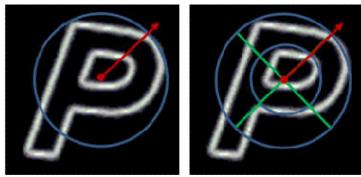
Full process of text detection



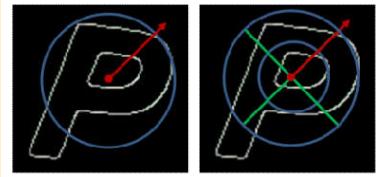
Detecting Texts of Arbitrary Orientations in Natural Images [Yao et al., CVPR, 2012]



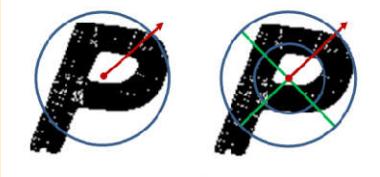
(b)



(d)

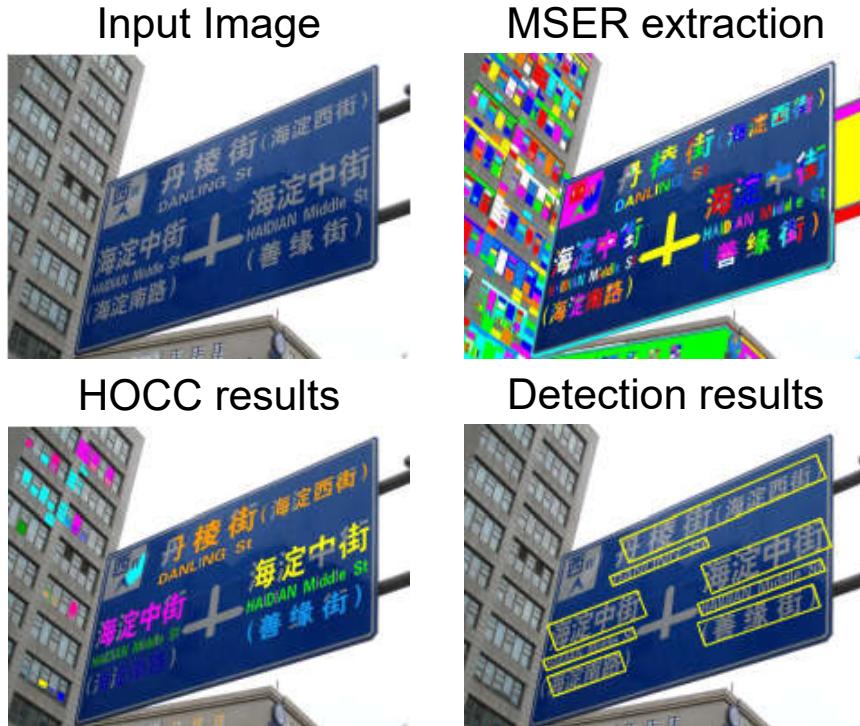
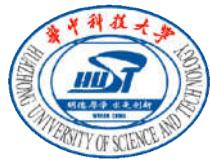


(c)



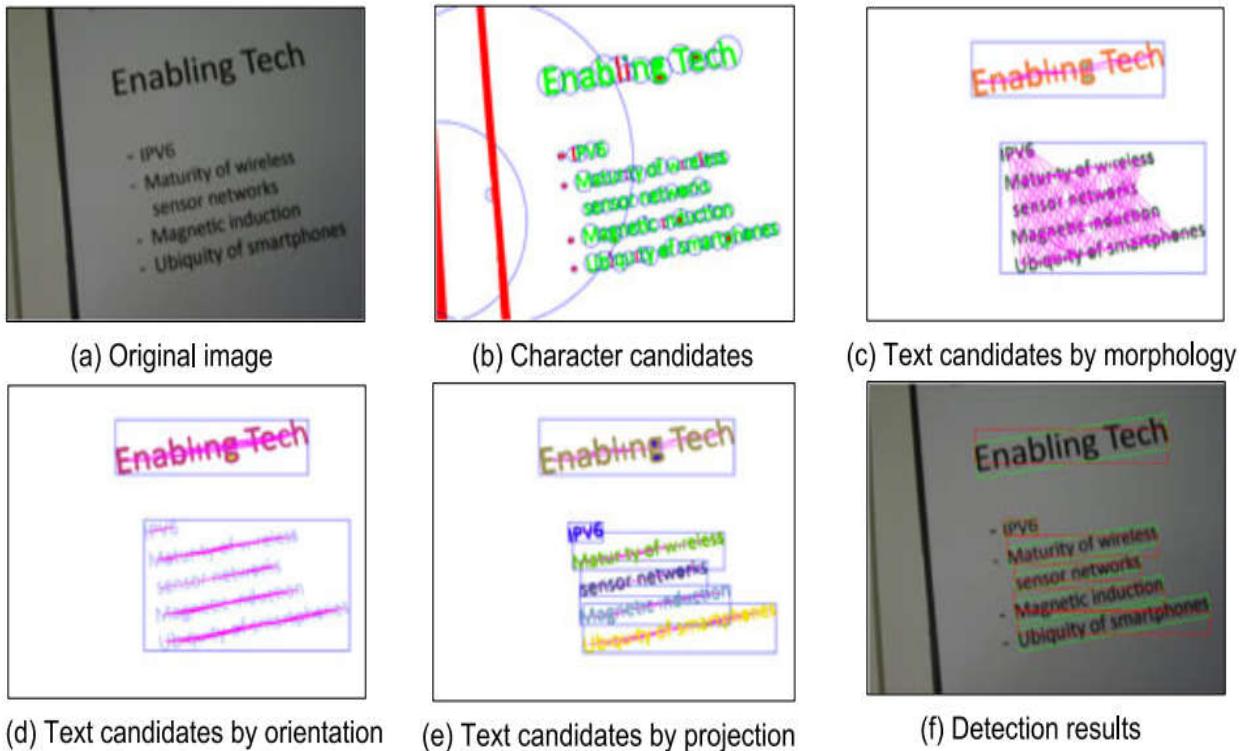
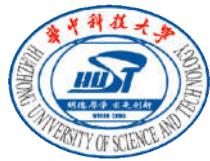
(e)

- 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...

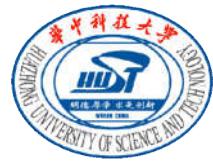


- Build a graph based on MSER components
- Higher-order correlation clustering (HOCC)
- Texton-based texture classifier to discriminate text and non-text regions

Multi-Orientation Scene Text Detection with Adaptive Clustering [Yin et al., PAMI, 2015]



- 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.

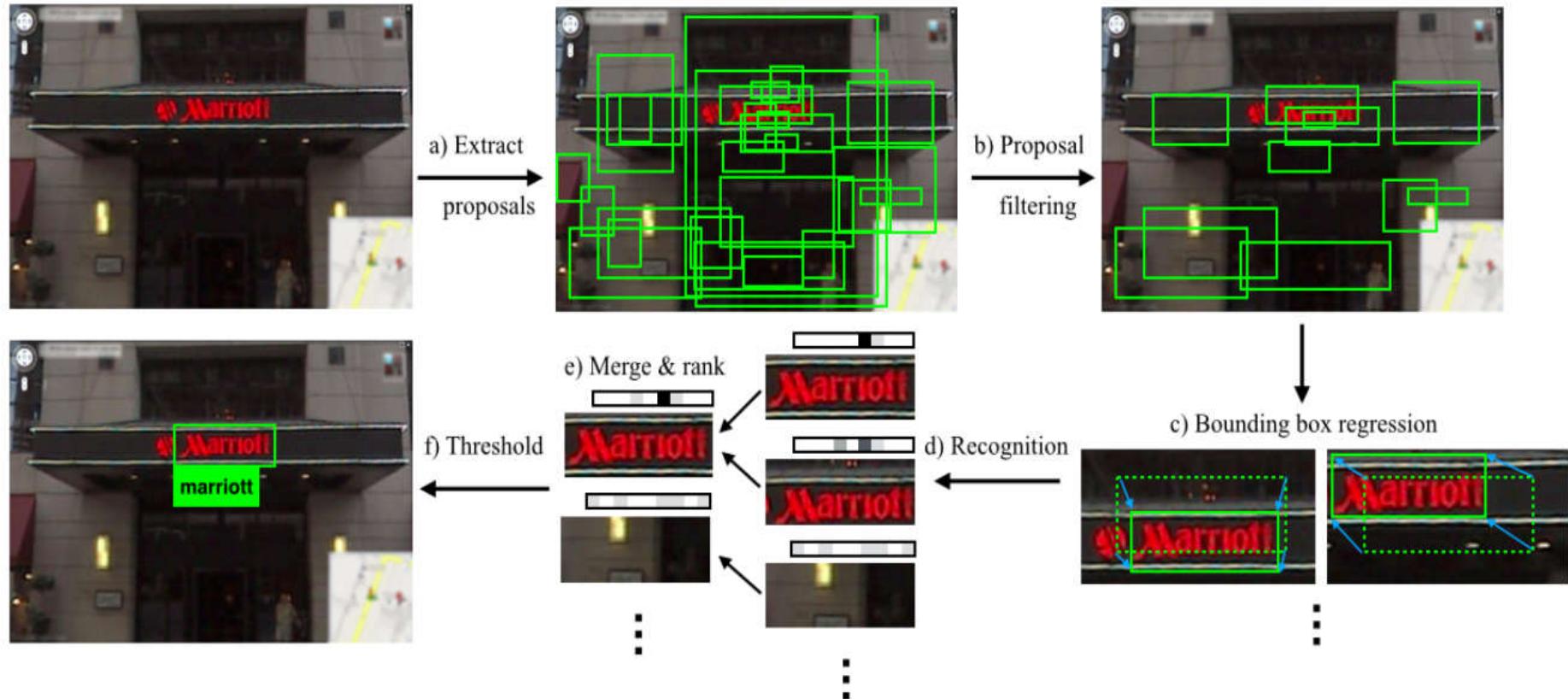
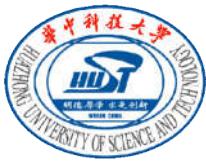


The Story of Oriented Scene Text Detection

- Handcraft Features
 - Component level. MSER, SWT...
 - Word / line level. Sliding Window
- Deep Learning (2014-)
 - Region Proposals
 - Segmentation
 - Hybrid Methods

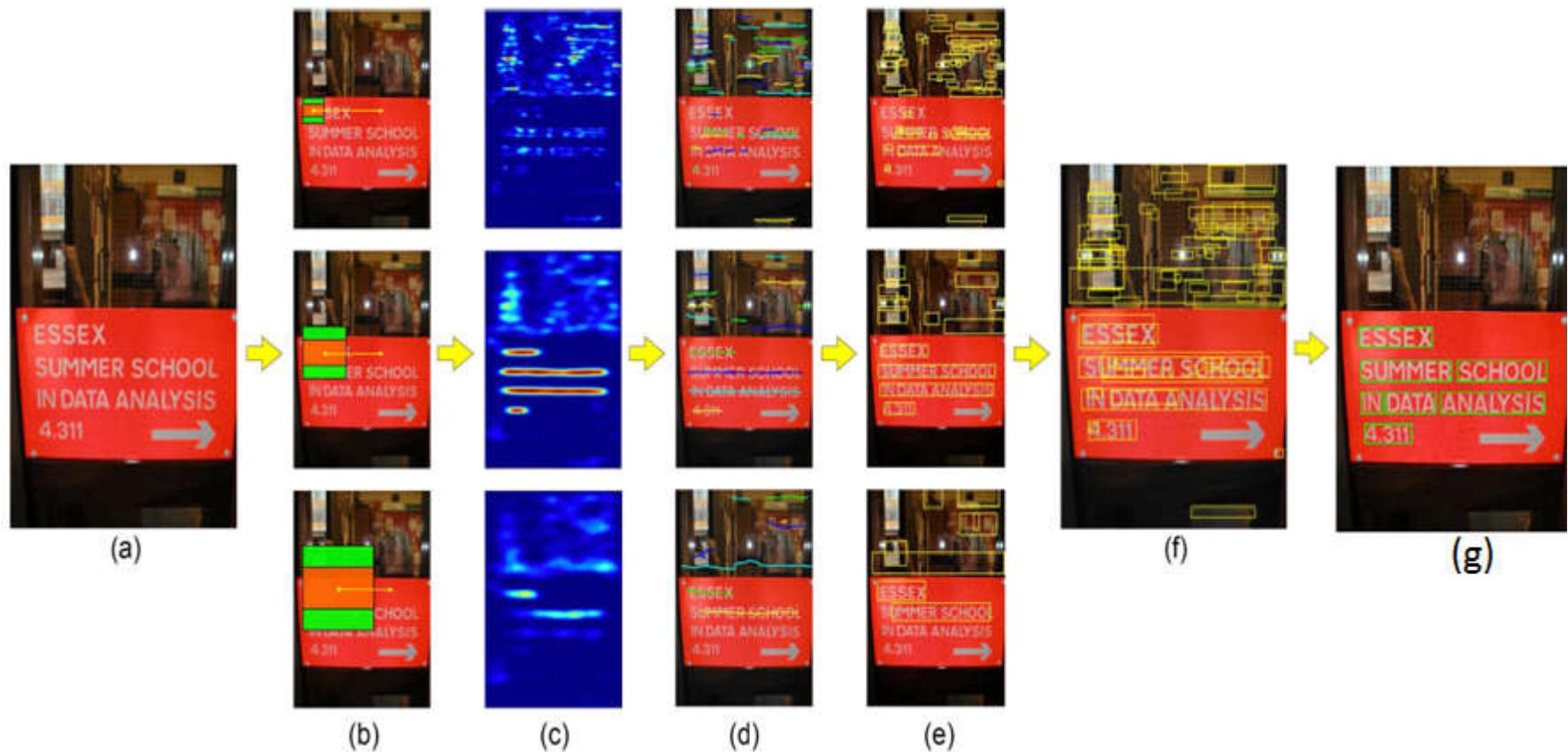
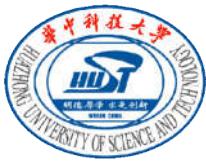
Reading Text in the Wild with Convolutional Neural Networks

[Jaderberg et al., IJCV, 2016]



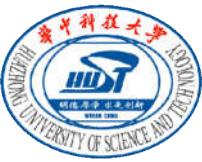
Symmetry-based text line detection in natural scenes

[Zhang et al., CVPR, 2015]



Synthetic Data for Text Localisation in Natural Images

[Gupta et al., CVPR, 2016]



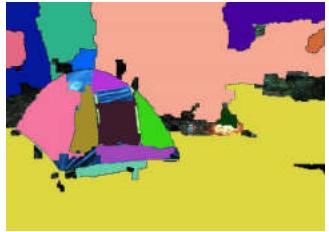
(a) RGB



(b) Depth



(c) Segmentation



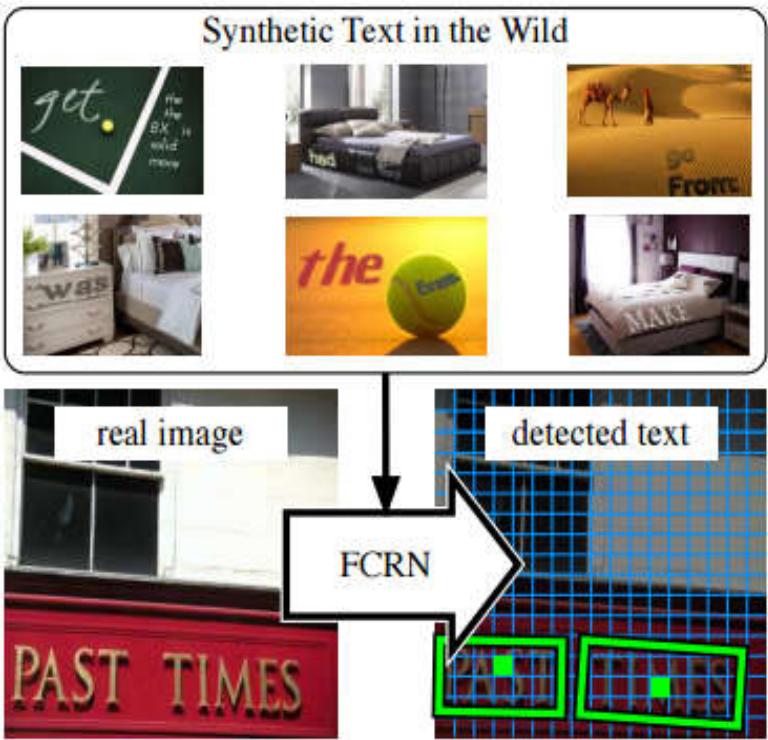
(d) Text Regions

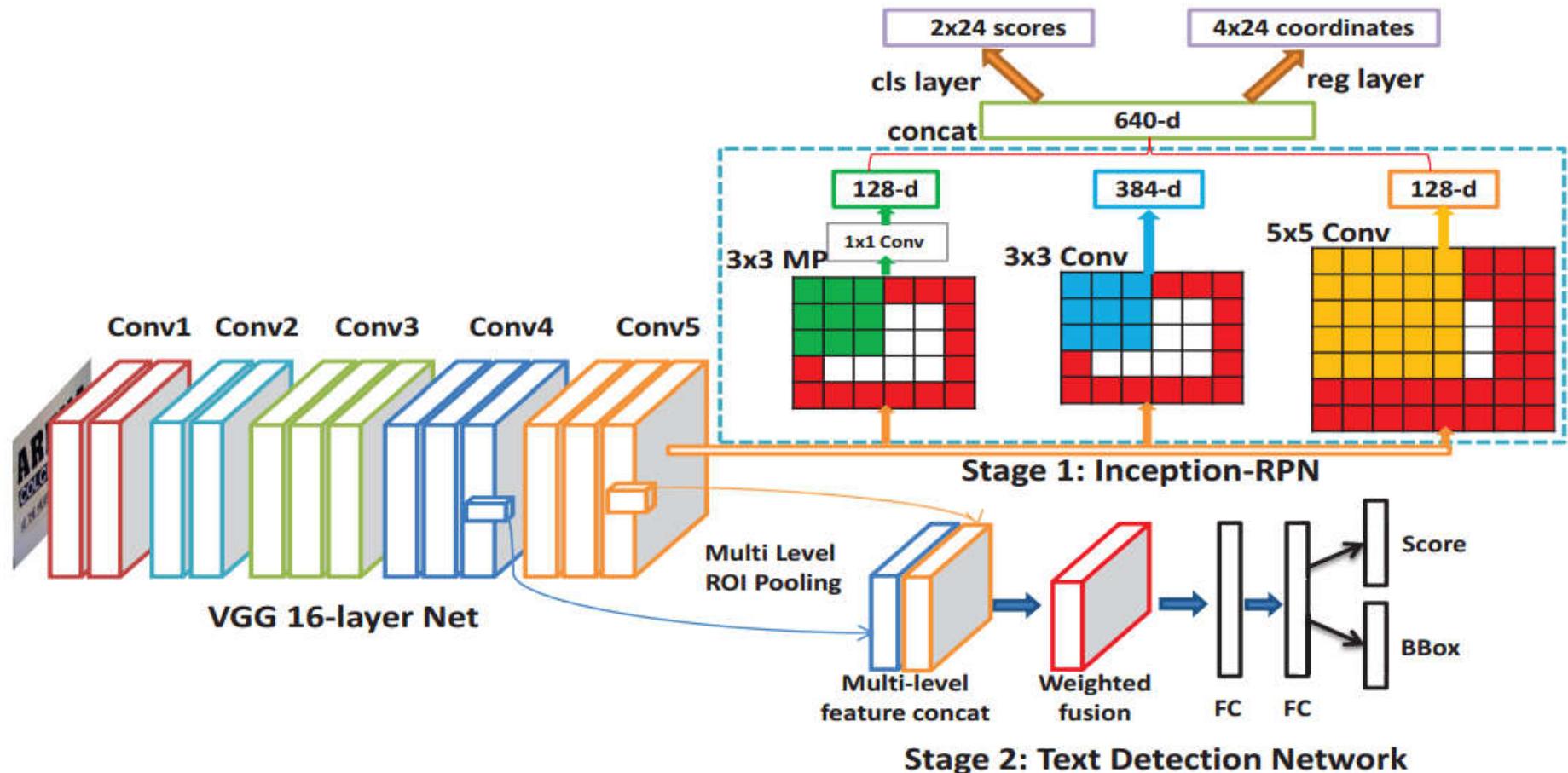
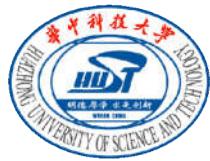


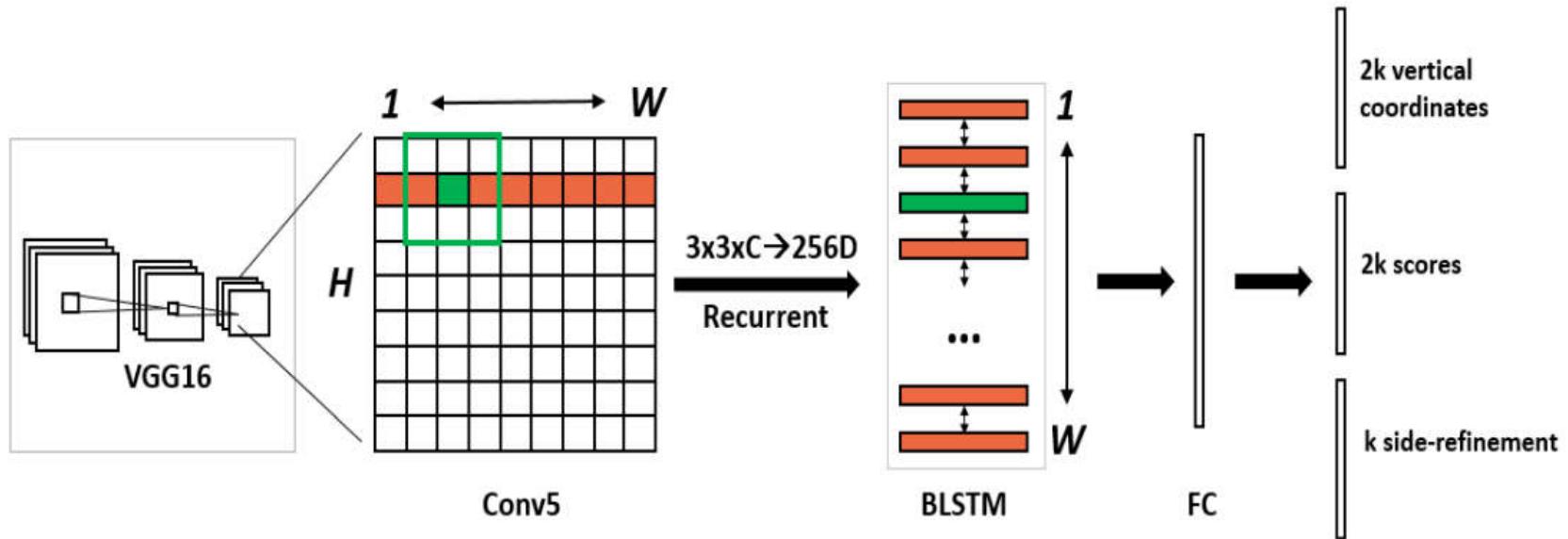
(e) Synthetic Text

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

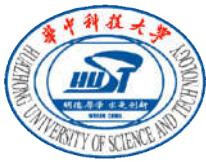
[1] Redmon et al., You Only Look Once: Unified, Real-Time Object Detection, CVPR 2016







- 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.



1. Fine-scale Proposals

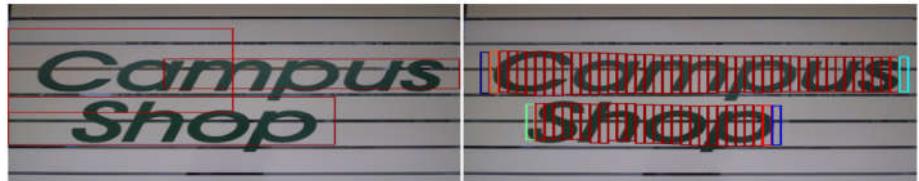


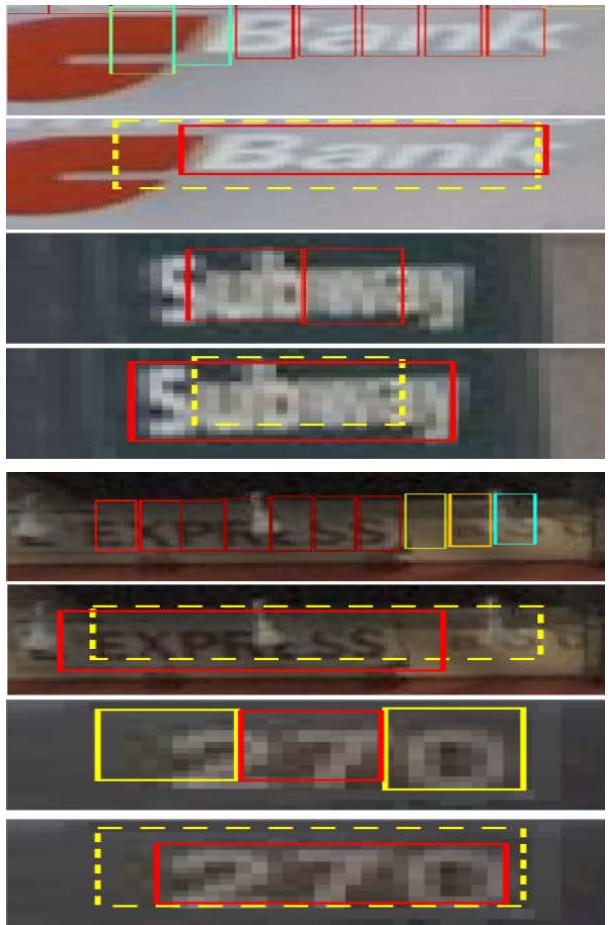
Fig. 2: Left: RPN proposals. Right: Fine-scale text proposals.

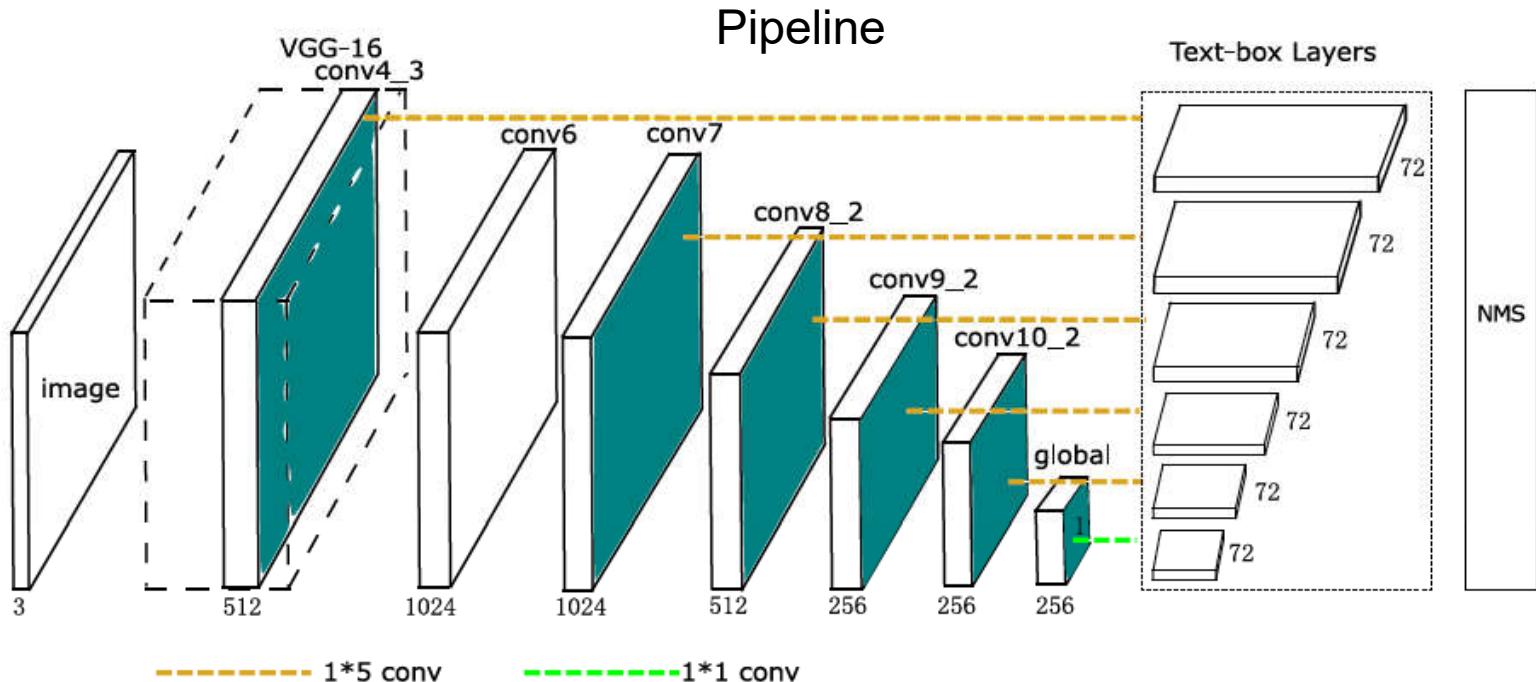
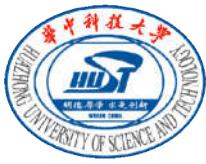
2. Recurrent Connectionist Text Proposals



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

3. Side-refinement

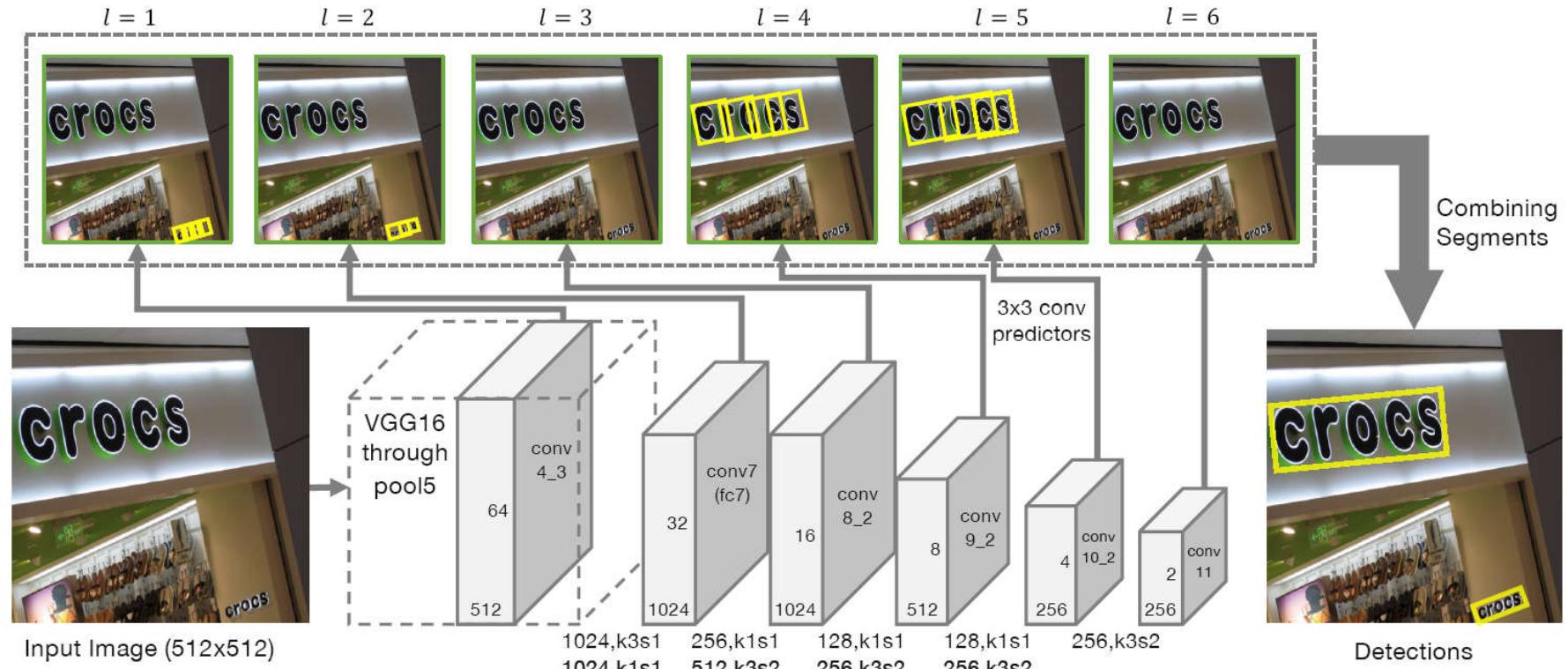
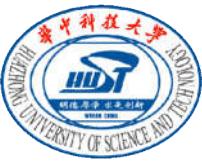




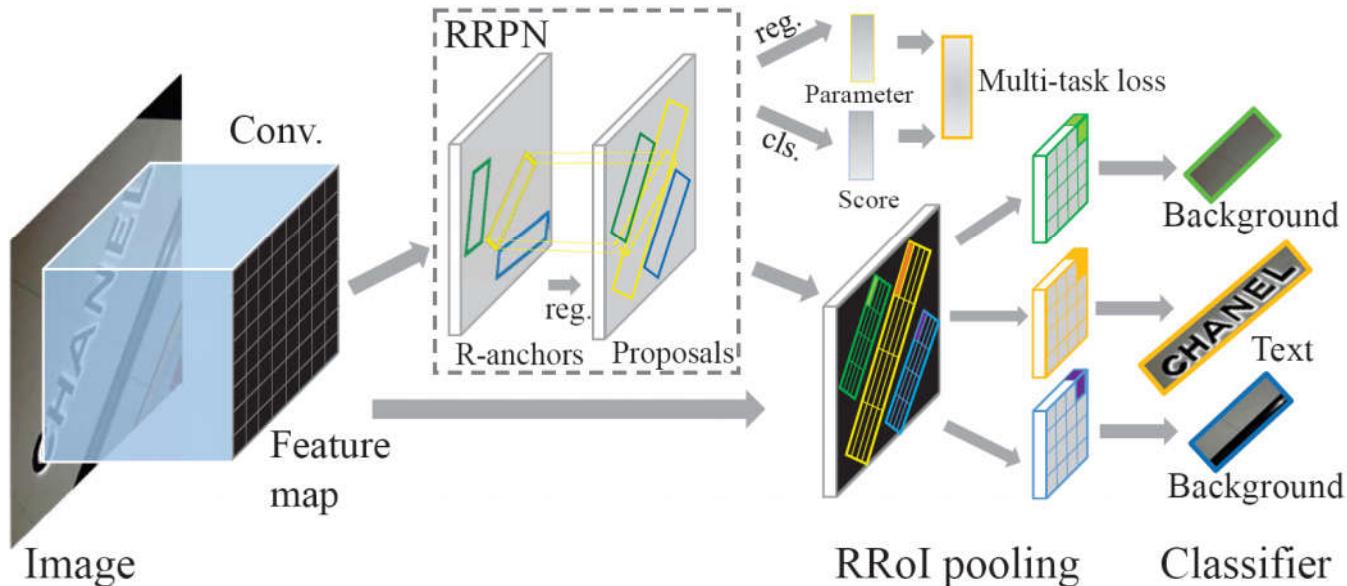
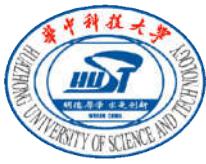
- Fully convolutional network based on SSD[1].
- On every map location, a text-box layer predicts a 72-d vector(text presence scores (2-d) and offsets (4-d) for 12 default boxes)
- Longer convolutional filters
- Special designed default boxes

[1] Liu et al., SSD: a single shot detector ECCV 2016

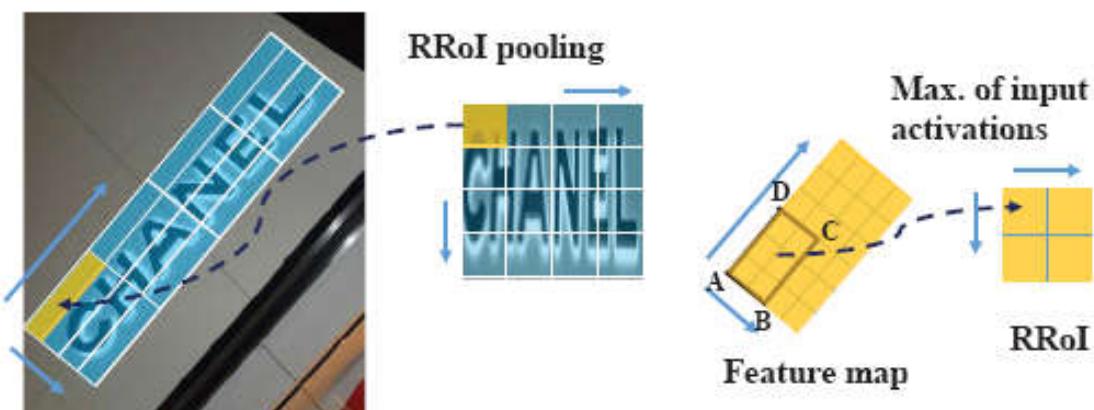
Detecting Oriented Text in Natural Images by Linking Segments [Shi et al., CVPR 2017.]



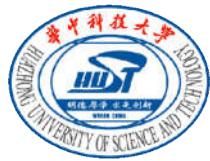
- Fully convolutional network inspired by SSD
- Multi-stage outputs for segments and their links
- Solve the problem of CNN receptive field for long texts



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

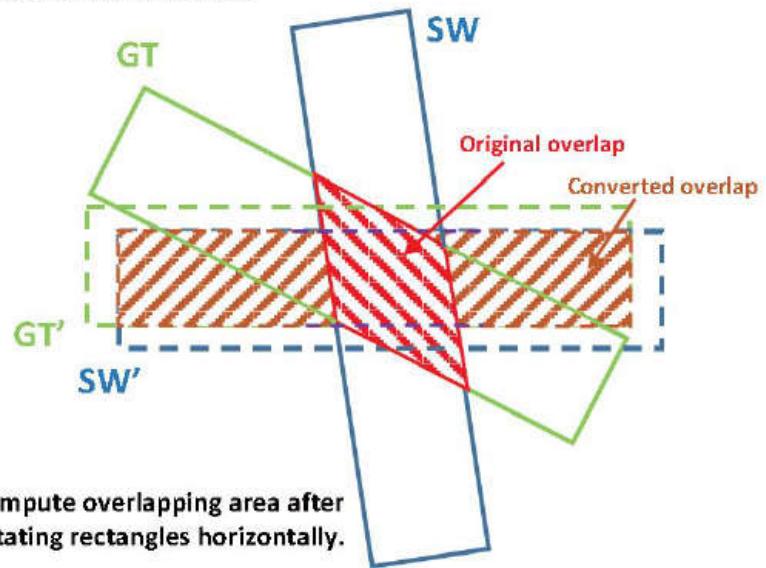


The implementation of
Rotated RoI



- Use the architecture of SSD
- Use different matching strategy

Previous method



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

Accuracy

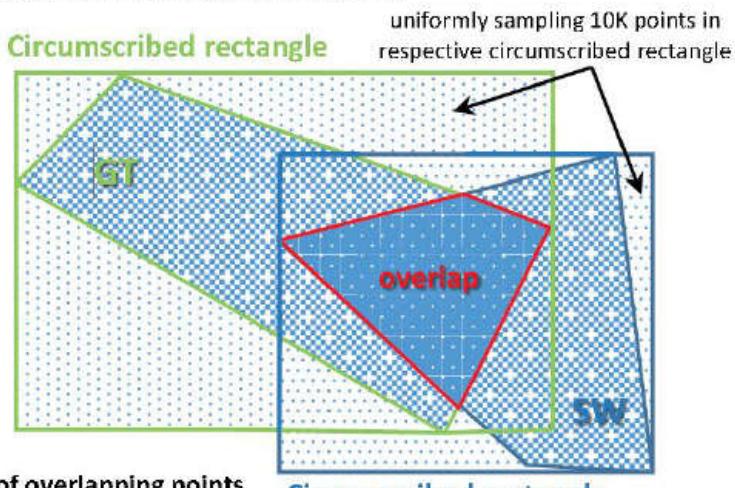
Very high. The error is stable at about 1%.

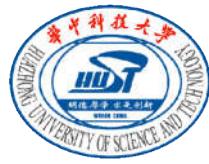
Calculate overlapping area between **rectangles**.

Function

Calculate overlapping area between **polygons**.

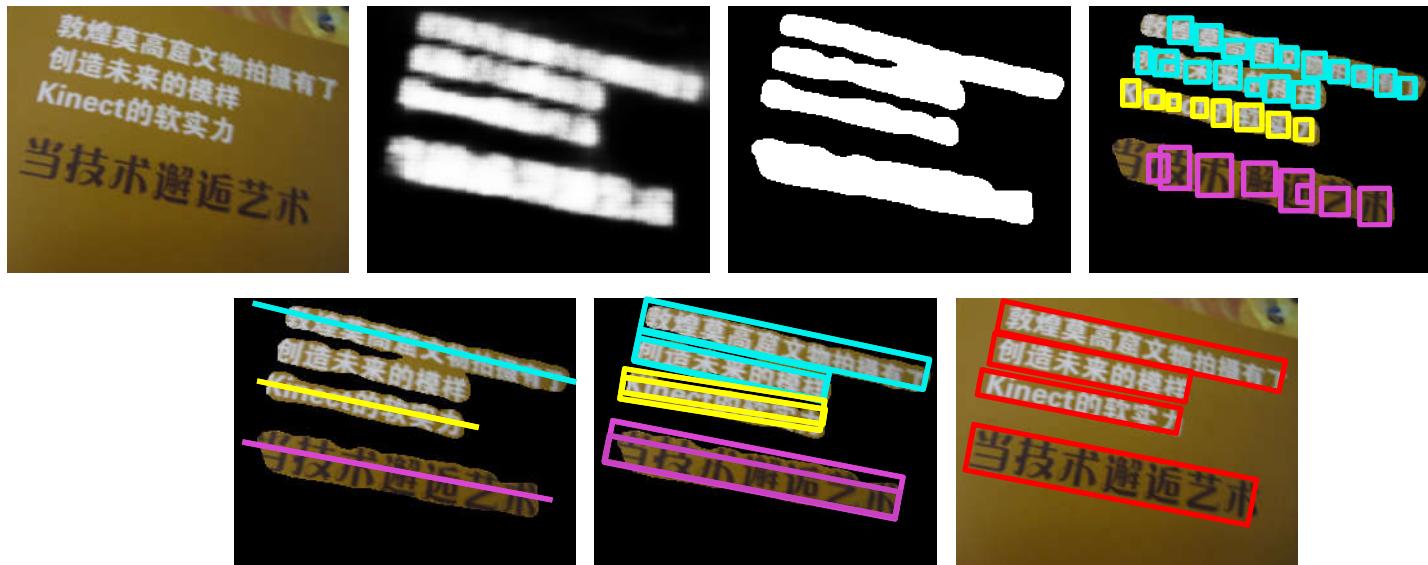
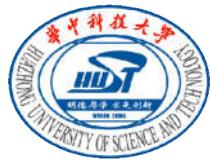
Our shared Monte-Carlo method



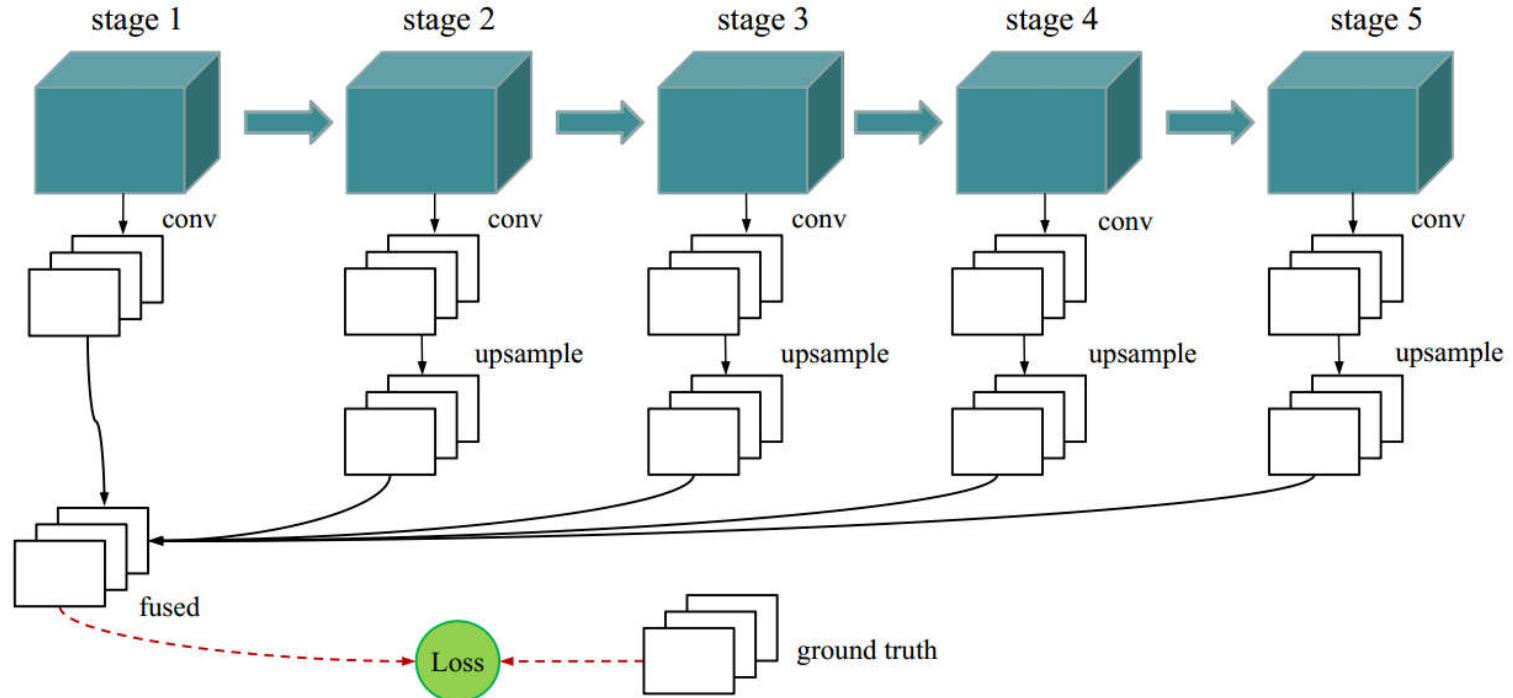
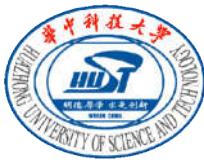


The Story of Oriented Scene Text Detection

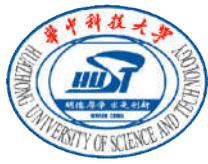
- Handcraft Features
 - Component level. MSER, SWT...
 - Word / line level. Sliding Window
- Deep Learning (2014-)
 - Region Proposals
 - Segmentation
 - Hybrid Methods



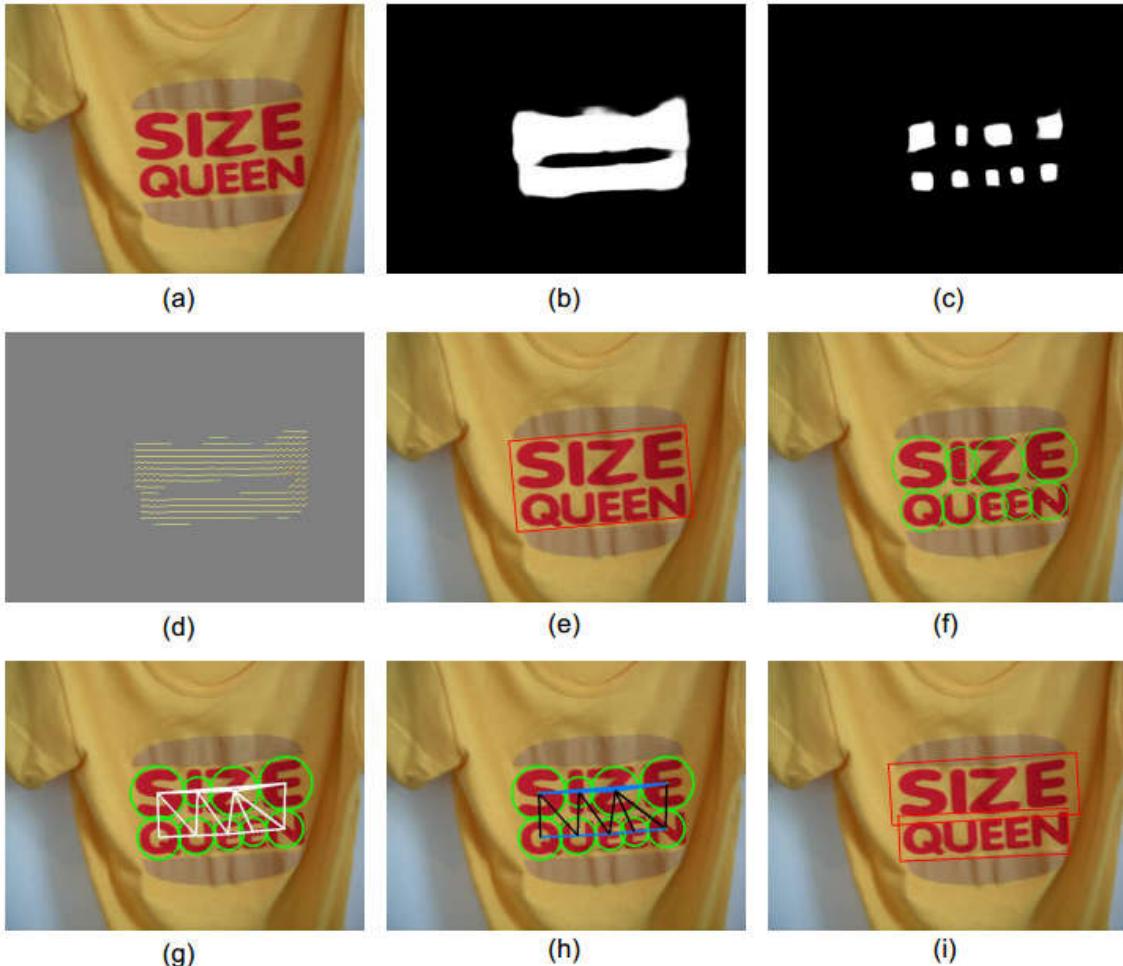
- The Text-Block FCN is to predict the salient map of text block.
- Multi-oriented text line hypotheses are generated by combining both global and local cues.
- The Character-Centroid FCN is used to remove false positives.

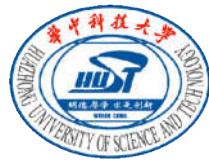


- FCN based network.
- Multi task. Text region, individual characters and their relationship are estimated simultaneously.



Process of text detection





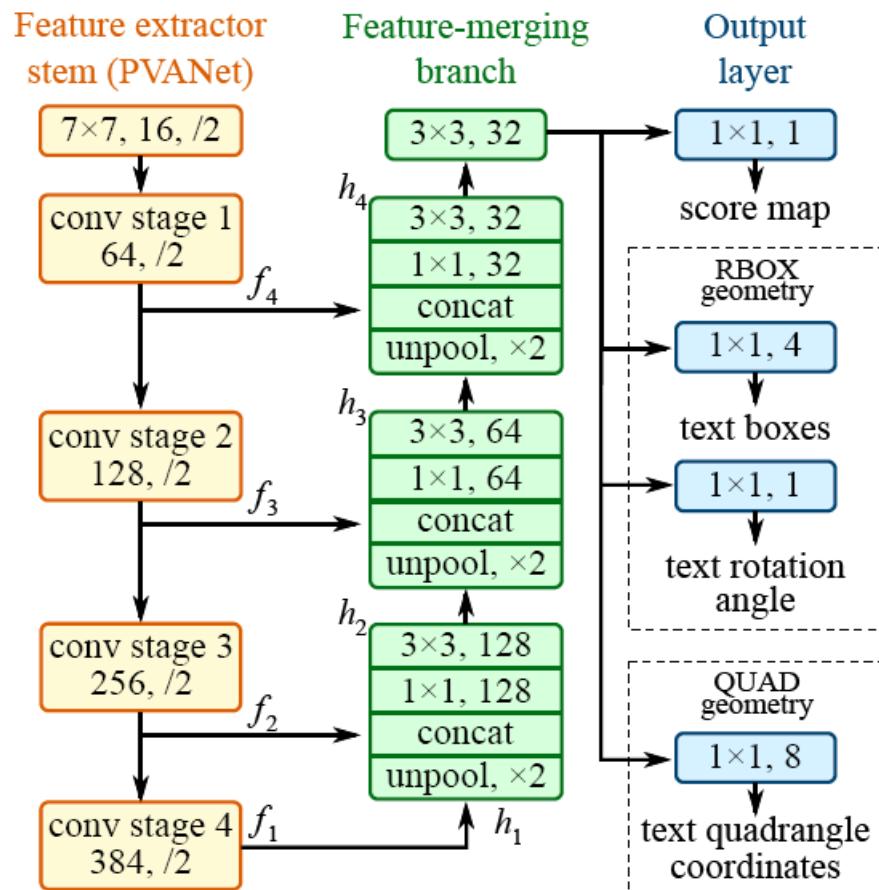
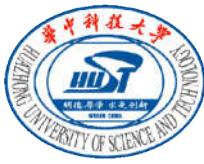
The Story of Oriented Scene Text Detection

➤ Handcraft Features

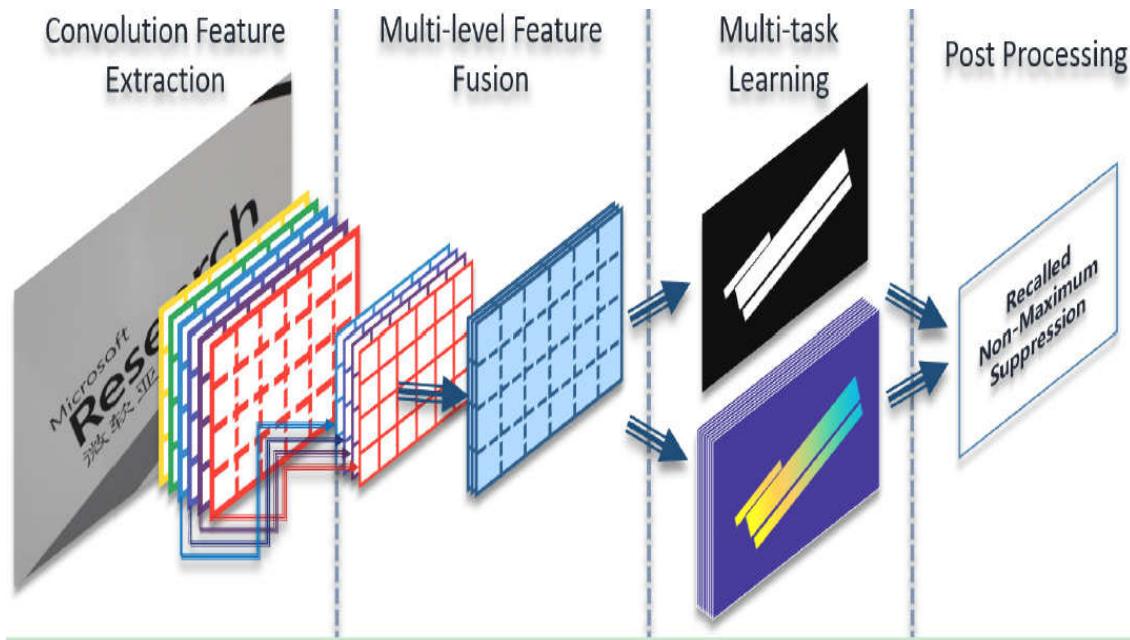
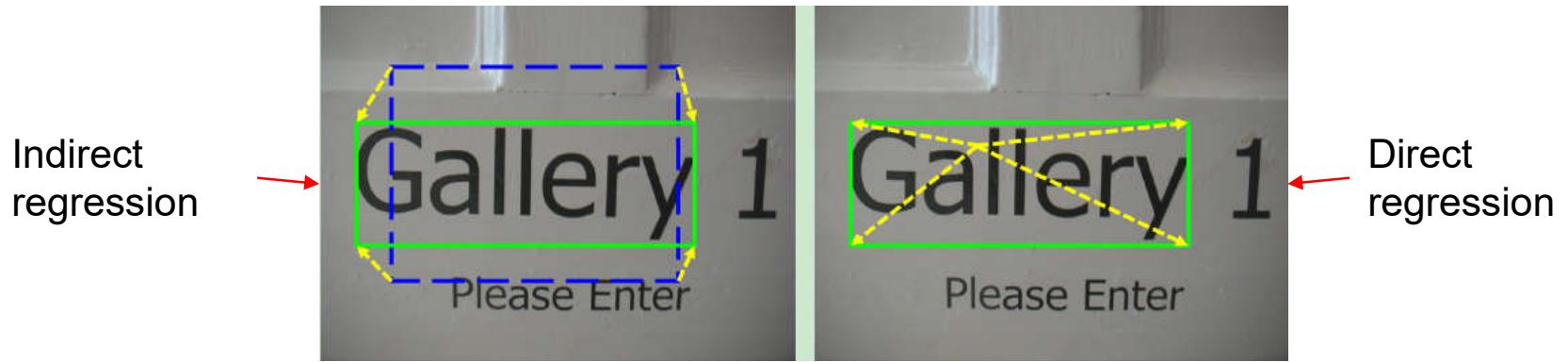
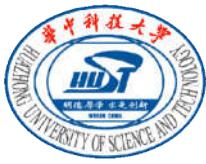
- Component level. MSER, SWT...
- Word / line level. Sliding Window

➤ Deep Learning (2014-)

- Region Proposals
- Segmentation
- Hybrid Methods

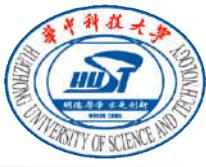


- PVANet(faster than VGG16)
- Multi-channel :
 - Score map
 - Rotated bounding boxes
 - Quadrangle bounding boxes
- Refined NMS



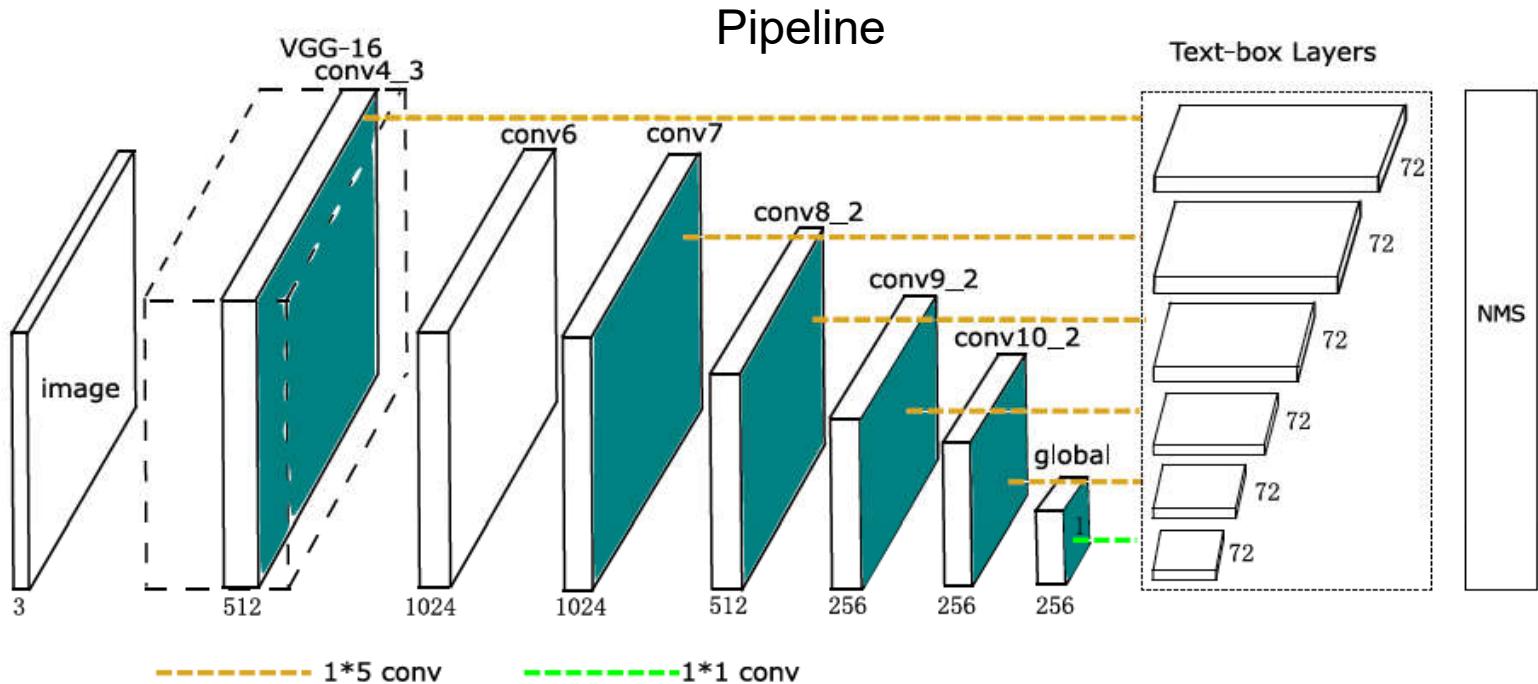
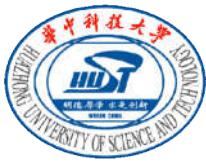
- Multi-level feature fusion
- Up-sample to quarter size of the input image
- Multi-task learning for classification and regression
- Post Processing:
Refined NMS

Architecture

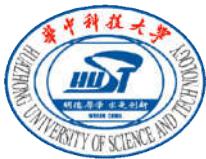


Outline

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

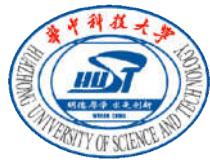


- Fully convolutional network.
- On every map location, a text-box layer predicts a 72-d vector(text presence scores (2-d) and offsets (4-d) for 12 default boxes)
- Longer convolutional filters
- Special designed default boxes



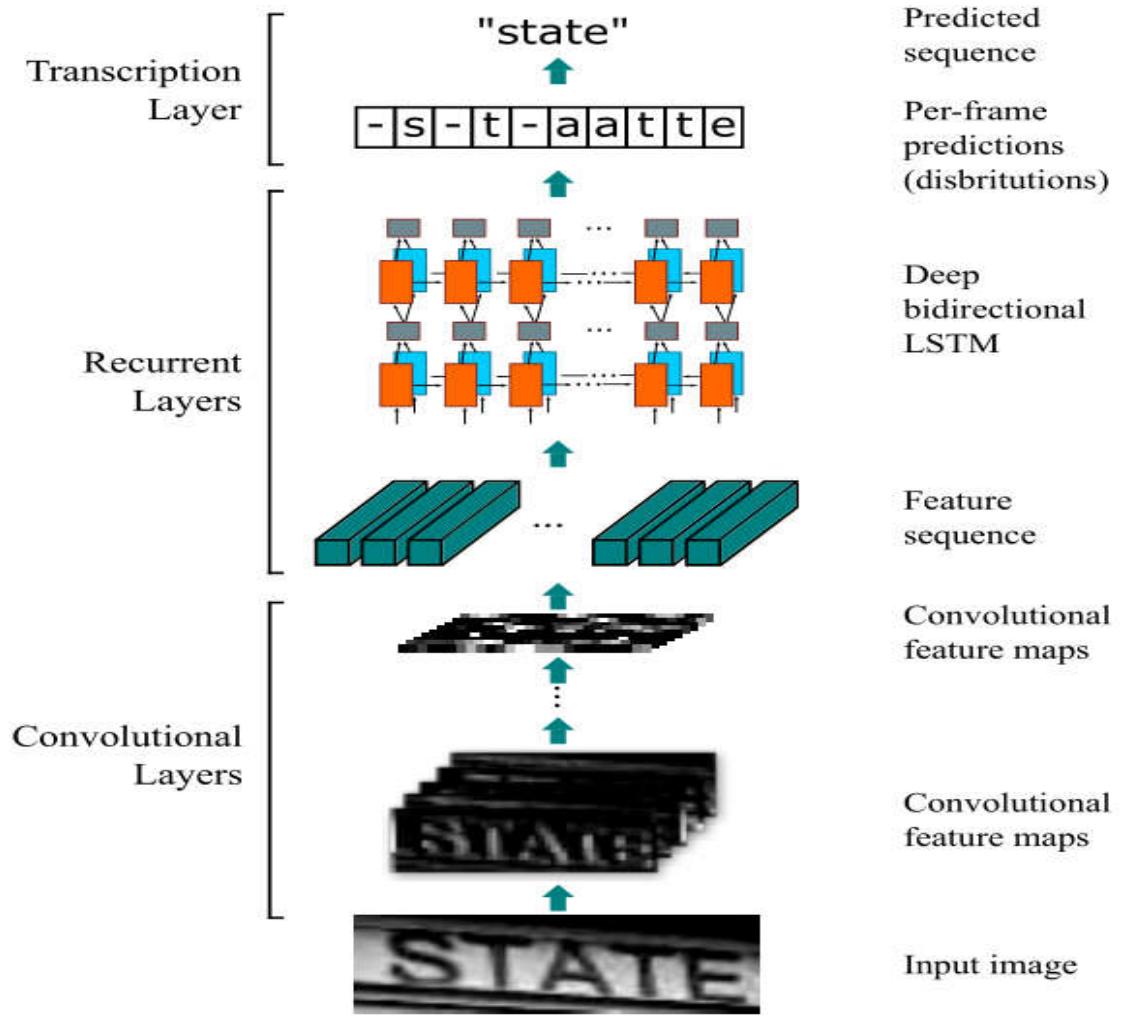
Quantitative Results of Text Localization

Datasets	ICDAR 2011						ICDAR 2013						Time/s	
Evaluation protocol	IC13 Eval			DetEval			IC13 Eval			DetEval				
Methods	P	R	F	P	R	F	P	R	F	P	R	F		
Jaderberg (Jaderberg et al. 2016)	–	–	–	–	–	–	–	–	–	–	–	–	7.3	
MSERs-CNN (Yin et al. 2014)	0.88	0.71	0.78	–	–	–	–	–	–	–	–	–	–	
MMser (Zamberletti, Noce, and Gallo 2014)	–	–	–	–	–	–	0.86	0.70	0.77	–	–	–	0.75	
TextFlow (Tian et al. 2015)	0.86	0.76	0.81	–	–	–	0.85	0.76	0.80	–	–	–	1.4	
FCRNall+filters (Gupta, Vedaldi, and Zisserman 2016)	–	–	–	0.92	0.75	0.82	–	–	–	0.92	0.76	0.83	>1.27	
Zhang (Zhang et al. 2016)	–	–	–	–	–	–	0.88	0.78	0.83	–	–	–	2.1	
SSD (Liu et al. 2015)	–	–	–	–	–	–	0.80	0.60	0.68	0.80	0.60	0.69	0.1	
Fast TextBoxes	0.86	0.74	0.80	0.88	0.74	0.80	0.86	0.74	0.80	0.88	0.74	0.81	0.09	
TextBoxes	0.88	0.82	0.85	0.89	0.82	0.86	0.88	0.83	0.85	0.89	0.83	0.86	0.73	



Network Structure

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



TextBoxes: A fast text detector with a single deep neural network

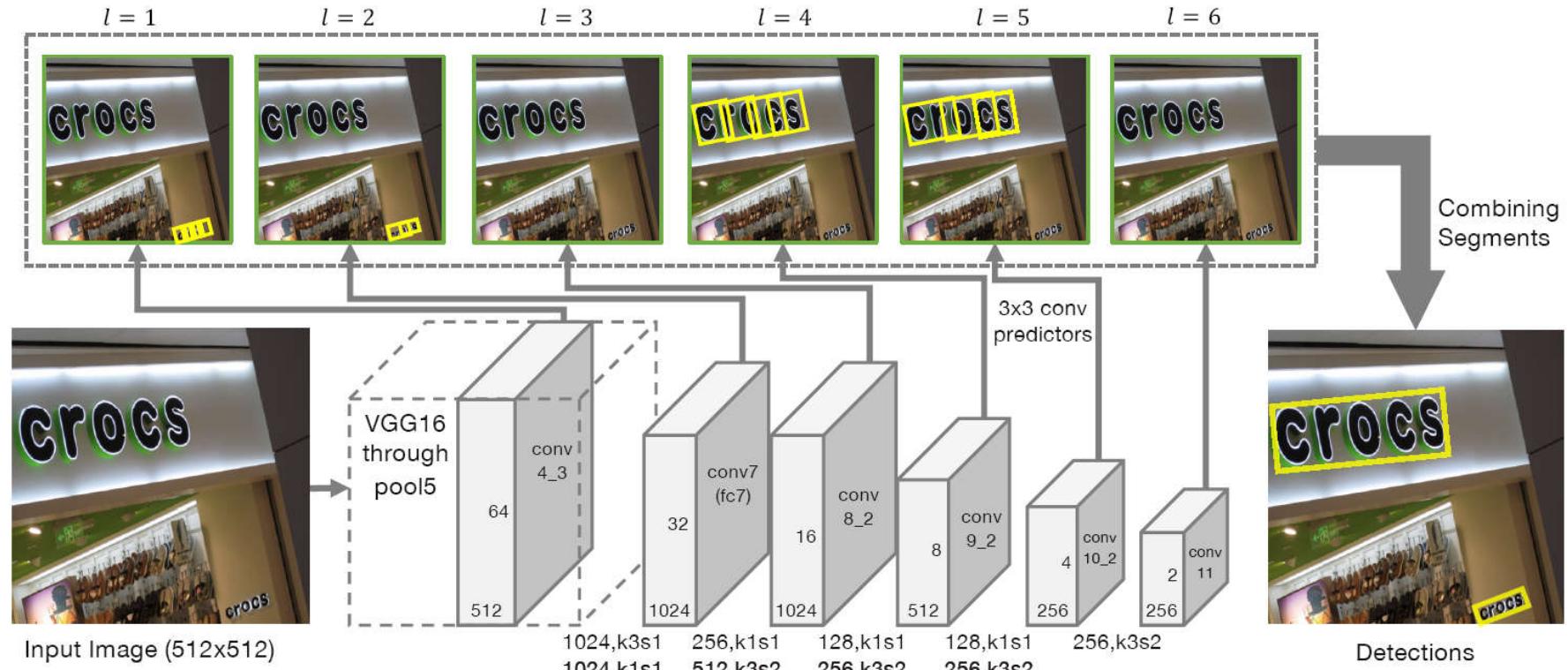
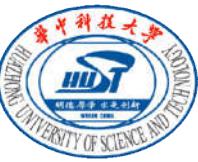
[Liao et al., AAAI 2017]

Combined with a recognition model(CRNN), we achieve state-of-the-art performance on ICDAR 2013.

Method	End-to-End results			Word spotting Results		
	Recall	Precision	Hmean	Recall	Precision	Hmean
HUST_MCLAB	87.68 %	95.83 %	91.57 %	90.77 %	97.25 %	93.90 %
Adelaide_ConvLST...	79.50 %	96.68 %	87.25 %	85.05 %	98.91 %	91.46 %
SRC-B-TextProces...	81.79 %	93.17 %	87.11 %	84.58 %	95.14 %	89.55 %
VGGMaxBBNet_095	82.12 %	91.05 %	86.35 %	86.68 %	94.64 %	90.49 %
VGGMaxBBNet (0...)	82.99 %	89.63 %	86.18 %	87.62 %	93.05 %	90.25 %
Yunos_Robot1.0	75.57 %	95.06 %	84.20 %	78.97 %	96.30 %	86.78 %
Deep2Text II+	72.08 %	94.56 %	81.81 %	75.82 %	96.29 %	84.84 %



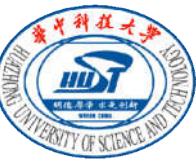
Detecting Oriented Text in Natural Images by Linking Segments [Shi et al., CVPR 2017.]



- 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

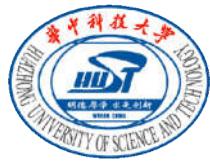
[Shi et al., CVPR 2017.]



Linking segments



Long texts can be easily located

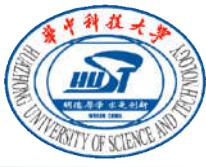


Results on ICDAR 2015 Incidental Text

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

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

Method	strongly			End-to-End results			Word spotting Results		
	Recall	Precision	Hmean	Recall	Precision	Hmean			
HUST_MCLAB	52.00 %	97.65 %	67.86 %	55.16 %	97.91 %	70.57 %			
Baidu IDL	60.81 %	67.54 %	64.00 %	62.82 %	69.02 %	65.78 %			
TextProposals + D...	37.89 %	89.84 %	53.30 %	40.50 %	90.73 %	56.00 %			
SRC-B-TextProces...	40.11 %	76.42 %	52.61 %	41.42 %	77.91 %	54.08 %			
Yunos_Robot1.0	36.30 %	67.81 %	47.29 %	38.30 %	69.83 %	49.47 %			
Megvii-Image++	39.38 %	57.48 %	46.74 %	42.29 %	61.02 %	49.95 %			

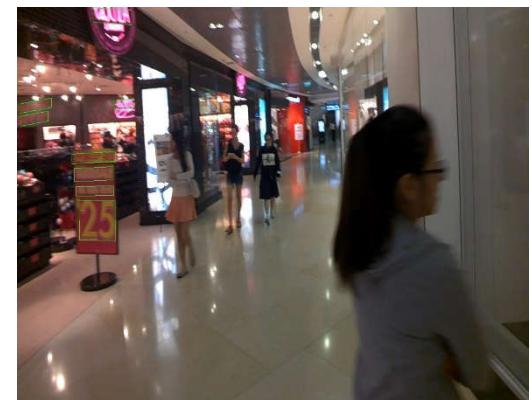


Outline

- 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

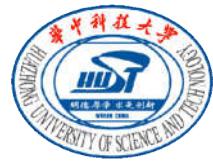
- Contains 500 natural images taken from indoor and outdoor.
- Texts in different languages (Chinese, English or mixture of both), fonts, sizes, colors and orientations.
- Annotated with text line bounding box.
- Ref. Detecting Texts of Arbitrary Orientations in Natural Images, CVPR12



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)
- Link: <http://mclab.eic.hust.edu.cn/icdar2017chinese/>

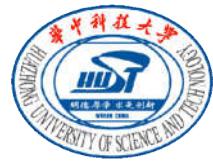




Comparison on ICDAR 2015

ICDAR 2015

Method	Precision	Recall	F-Measure	Time/s
Zhou et al. CVPR 2017	84	73	78	0.08
Shi et al. CVPR 2017	73	77	75	--
Ma et al. arxiv 2017	82	73	77	--
Liu et al. CVPR 2017	73	68	71	--
He et al. arxiv 2017	82	80	81	--
Tian et al. ECCV 2016	74	52	61	--
Zhang et al. CVPR 2016	71	43	54	2.1

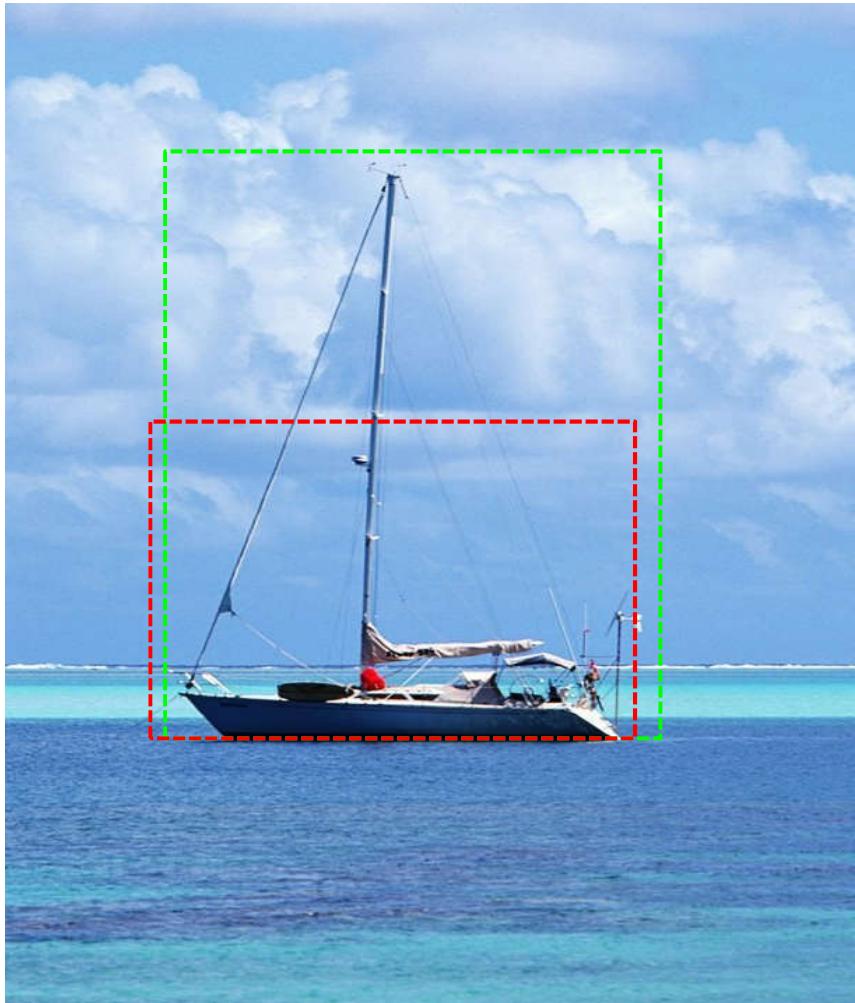


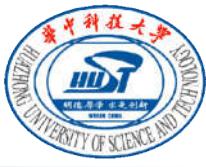
Comparison on MSRA-TD 500

MSRA-TD 500

Method	Precision	Recall	F-measure	Time/s
Zhou et al. CVPR 2017	87	67	76	0.08
Shi et al. CVPR 2017	86	70	77	0.11
Ma et al. arxiv 2017	82	68	74	0.3
He et al. arxiv 2017	77	70	74	--
Huang et al. ACM MM 2016	74	68	71	--
Yao et al. arxiv 2016	77	75	76	0.42
Zhang et al. CVPR 2016	83	67	74	--
Yin et al. PAMI 2015	81	63	71	1.4
Kang et al. CVPR 2014	71	62	66	--
Yao et al. CVPR 2012	63	63	60	--

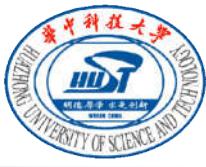
The Drawback of IOU in Scene Text Detection





Outline

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

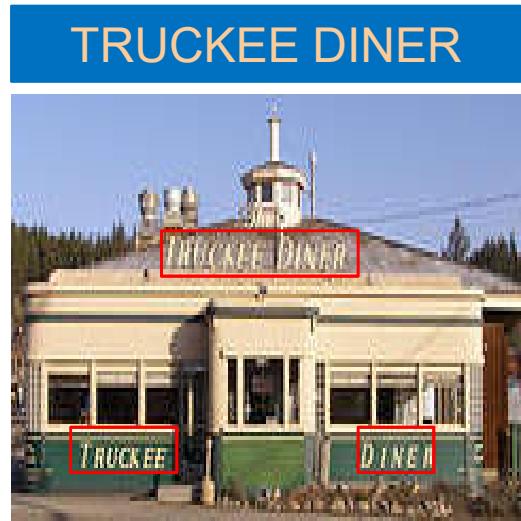


Applications

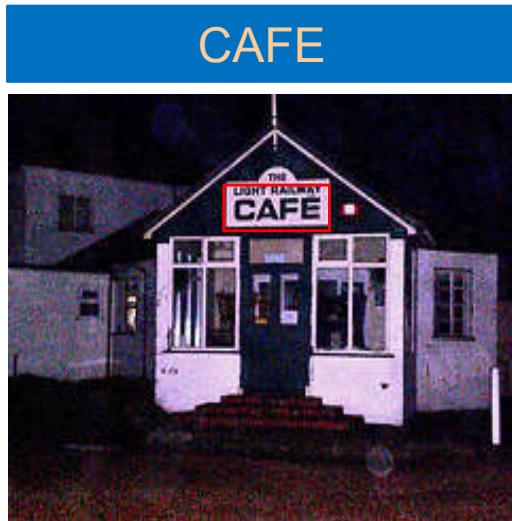
- Fine-grained Classification
- Number
- Container
- Exercise search
- Word retrieval in the wild

Fine-Grained Image Classification with Text Information

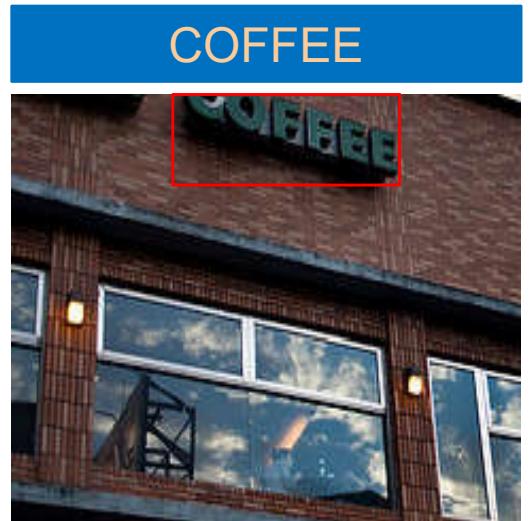
Motivations



(a)



(b)



(c)

- 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.

Integrating Scene Text and Visual Appearance for Fine-Grained Image Classification with Convolutional Neural Networks. [arXiv: 1704.04613](https://arxiv.org/abs/1704.04613)

Fine-Grained Image Classification with Text Information

Pipeline



GoogLeNet

Image Feature
 f_v



Word Spotting



Detection

TRUCKEE
DINER

Recognition



Word
Embedding

Text Feature
 f_t

Image Feature
 f_v

Text Feature
 f_t

Attention

Attended Text Feature
 f_a

Attention

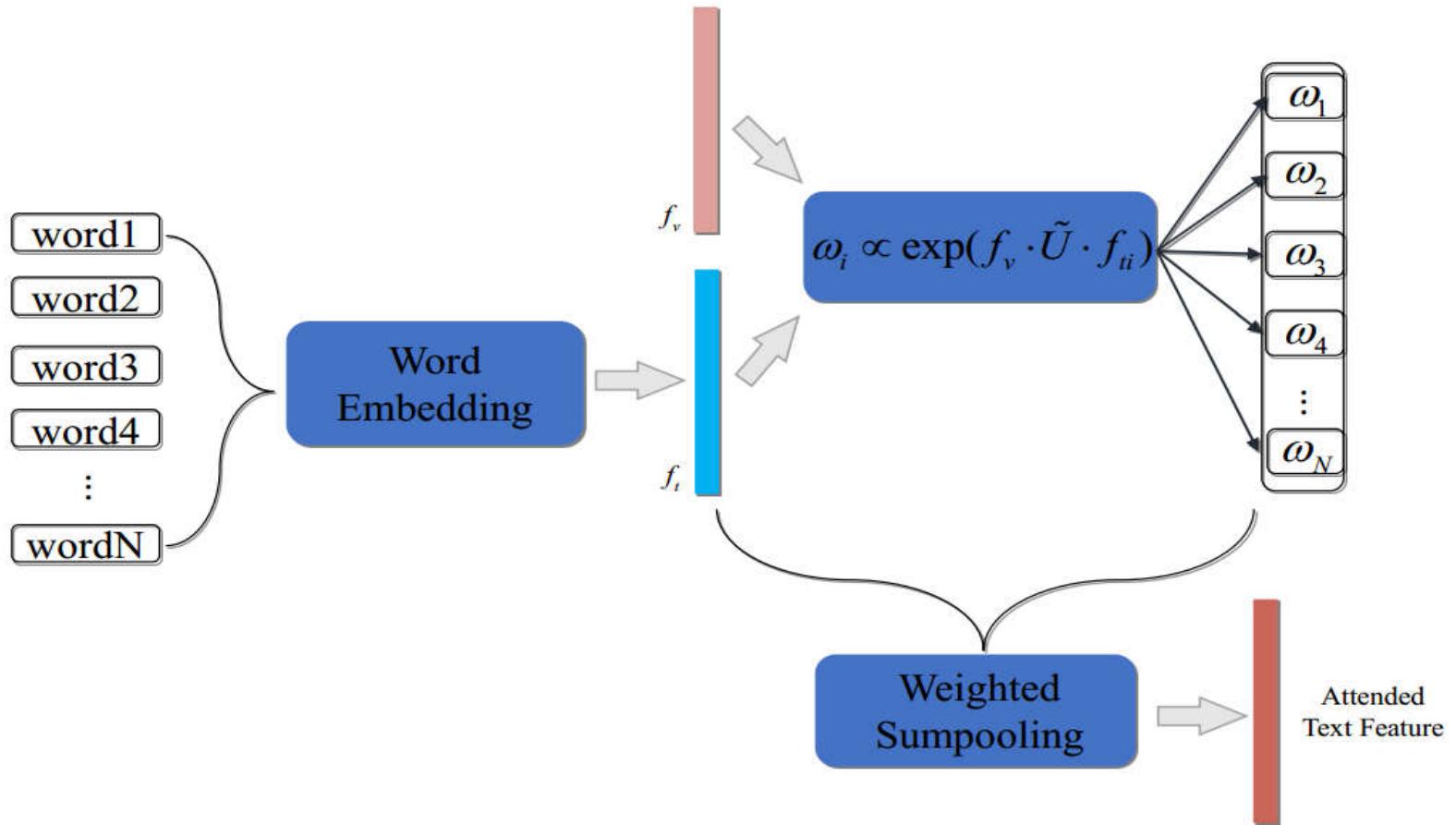
Attended Text Feature
 f_a

Classification

Restaurant

Fine-Grained Image Classification with Text Information

Attention Model



Fine-Grained Image Classification with Text Information

Results



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



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



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



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



(e) ROOTBEER **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**
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

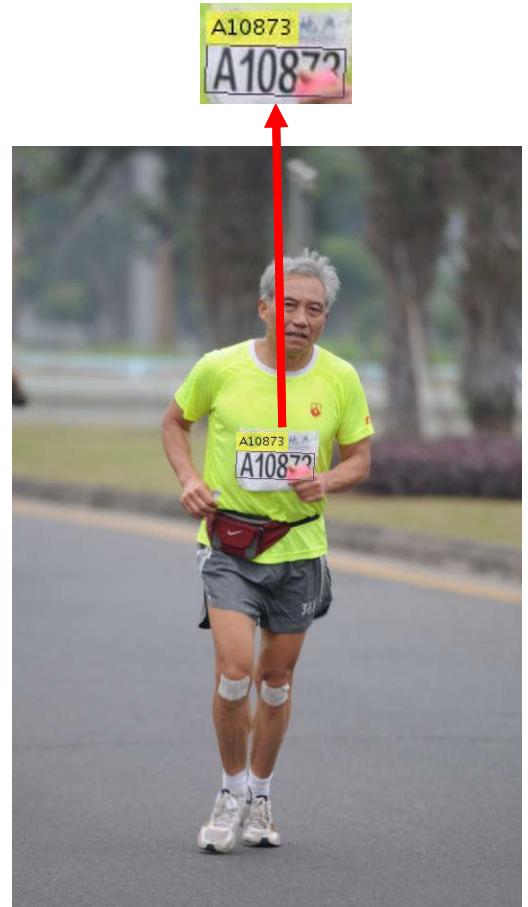
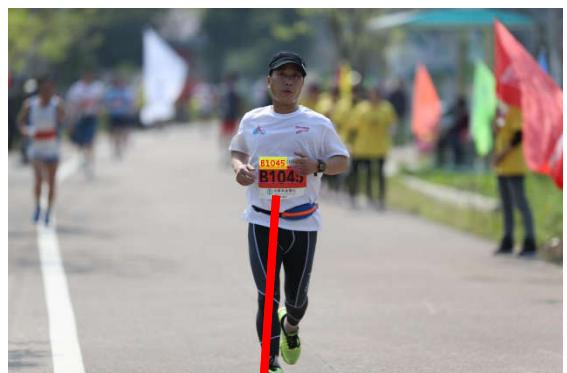


(g) BITTER **BITTER**
BITTER: 0.99
BROWN: 4.05e-5
PREMIUM: 3.5e-9
SPECIAL: 2.8e-9
ENGLISH: 9.4e-11
EXTRA: 6.11e-11



(h) GUINNESS **GUINNESS**
GUINNESS: 1
SPECIAL: 1.6e-25
EXPORT: 6.4e-27
QUINES: 1.3e-30

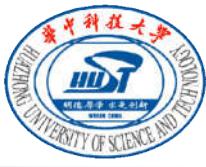
Person Re-identification with Numbers



Container

检测特定的文字并识别





Exercise

no good persuading her to stopping smoking. Children

no good persuading her to stopping smoking. Children

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

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)

(1) 词中所写的是什么学节? 从哪里可以看出来? (3分)

3. 下列语句有语病的一项是() (2 分)

3. 下列语句有语病的一项是 () (2分)

计算下面机构的自由度, 并说明想使机构具有确定的运动, 需要几个原动件

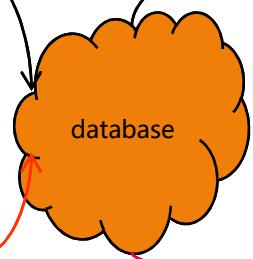
计算下面机构的自由度, 并说明想使机构具有确定的运动, 需要几个原动件

Word retrieval in the wild

检索关键词

SHOP

ATM



检索结果

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

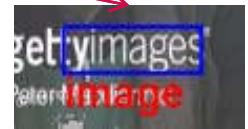
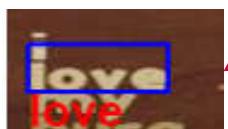
Word retrieval in the wild

- 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

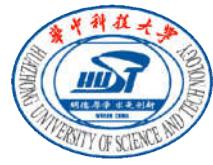


Word retrieval in the wild

- 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）

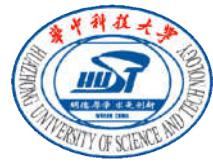


当query为love时的部分检索结果（第一行：精准匹配，第二行：模糊匹配）



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>

C. Yao, X. Bai, W. Liu. A unified framework for multi-oriented text detection and recognition. TIP'14
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>

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>

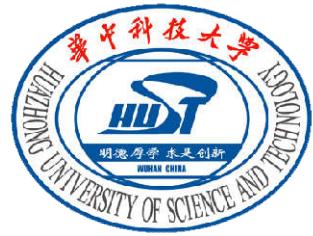
B. Shi, X. Bai, C. Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. TPAMI'16
Code: http://mclab.eic.hust.edu.cn/~xbai/CRNN/crnn_code.zip

Z. Zhang, C. Zhang, W. Shen, C. Yao, X. Bai. Multi-oriented text detection with fully convolutional networks. CVPR'16
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

C. Yao, X. Bai, B. Shi, W. Liu. Strokelets: A learned multi-scale representation for scene text recognition. CVPR'14
Code: http://mclab.eic.hust.edu.cn/~xbai/Strokelet_code/Strokelet_code.zip

The Invited Talk in Vision and Learning Seminar (VALSE)
Xiamen, 2017-4-22



END
