场景文字检测的进展与挑战

刘学博

SenseTime Researcher

liuxuebo@sensetime.com

Outline

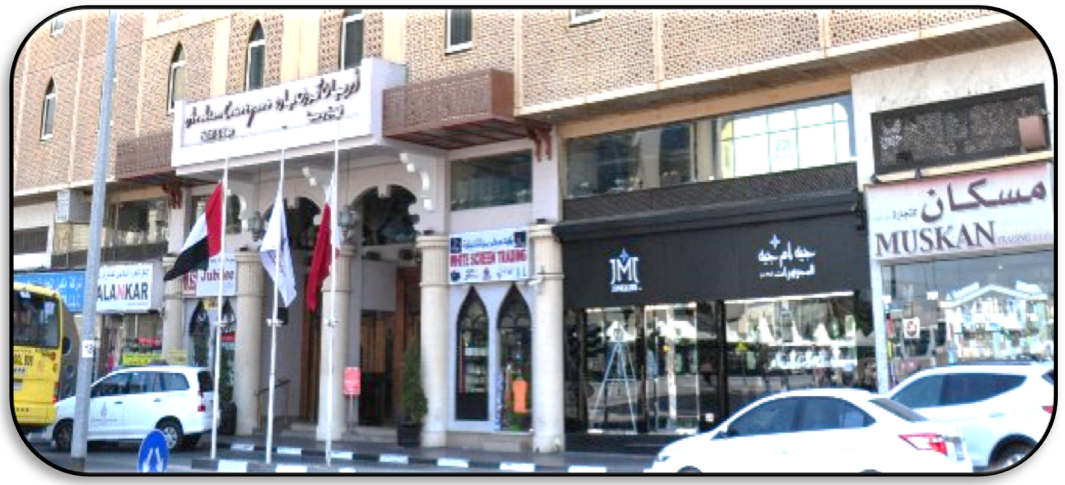
- Background
- Text detection
- From text detection to text spotting
- What’s next?

Outline

- Background
- Text detection
- From text detection to text spotting
- What’s next?

从物体识别到语义理解


Hospitality
FROSTED
Shredded
Wheat
BITE SIZE CEREAL
Good Source of 10 Vitamins and Minerals
NET WT 7 OZ (0.44 LB) 198g


MUSKAN


춘천집 먹거리
김치닭갈비 7,000
낙지닭갈비 7,000
뼈없는닭갈비 6,500
치즈닭갈비 7,500
매운닭갈비 6,500
낙지두루치기 7,500
쟁반국수 8,000
막국수 4,000
오징사리 3,500
비빔공기밥 1,500
점심특선메뉴
낙지정식 5,500
닭야채볶음밥 4,000
음료수 1,500
소주 3,000
맥주 3,500
청하 4,000
매화수 4,000
백세주 7,000
산사춘 8,000
막걸리 3,000
저희춘천집은 100% 국내산 하림 생고기 닭다리 살을 사용합니다.


万物之鼎：奇异博士用来观测世界的道具
你竟然靠在我的“万物之鼎”上?
Are you seriously leaning on the Cauldron of the Cosmos?

应用广泛

- 智慧交通：车牌识别，行驶证驾驶证识别，街景识别
- 金融：身份认证，智能核保
- 教育：智能阅卷，题目搜索，图像翻译
- 医疗：票据识别，病例电子化
- 互联网：图片审核，图片推荐

Outline

- Background
- Text detection
- From text detection to text spotting
- What’s next?

Problem Definition

Text detection is the process of predicting the presence of text and localizing each instance, usually at character, word or line level.

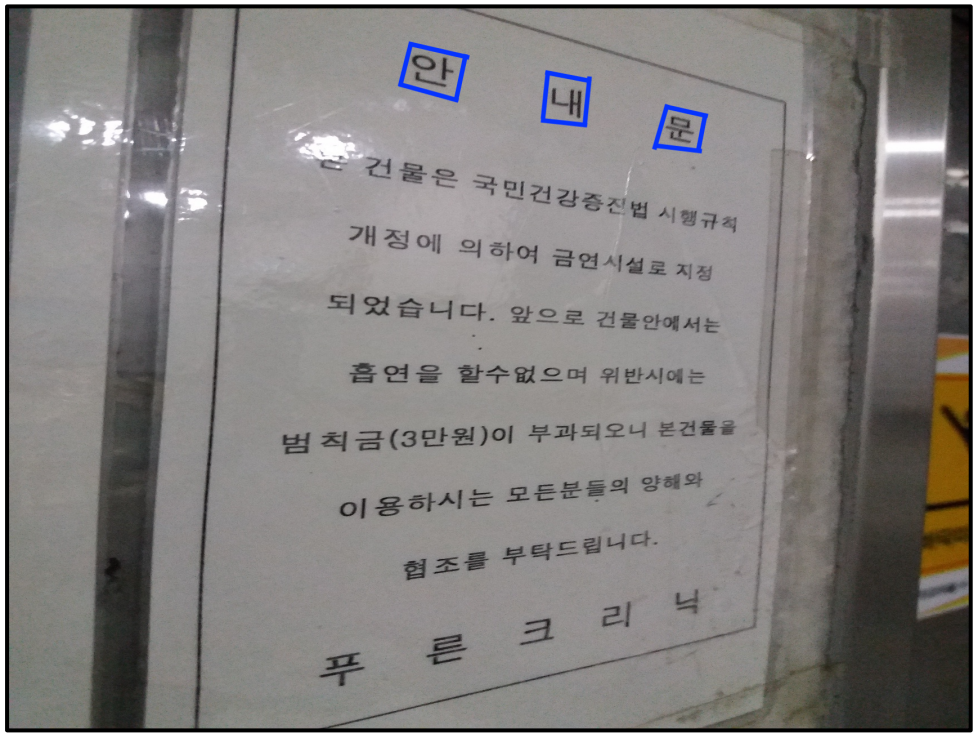
Horizontal text
(represented by two points)

Oriented text
(represented by four points)

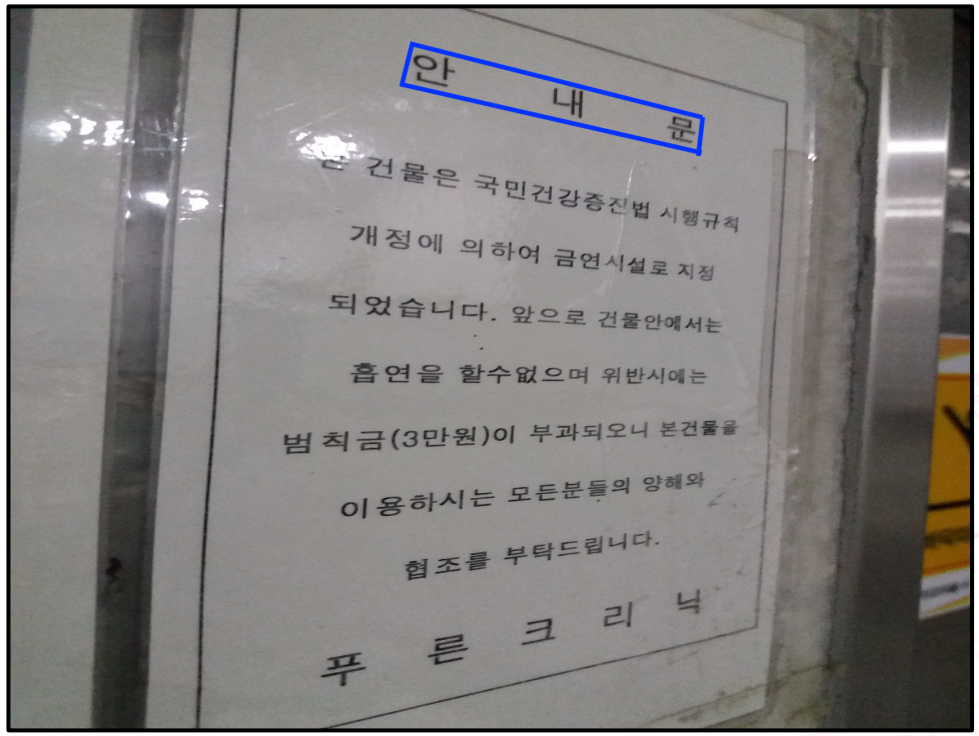
Curved text
(represented by several points or mask)

Challenges

- 难以定义和学习什么是文本行



VS



Challenges

- 巨大的长宽比

- 形状多变（曲线，倾斜）

Methods: Segmentation based

- 能较好地处理长文本行
- 较难区分相邻或有重合的文本行

Shi et al.. Detecting Oriented Text in Natural Images by Linking Segments. (SegLink) CVPR 2017

Li et al.. Shape Robust Text Detection with Progressive Scale Expansion Network. (PSENet) CVPR 2019

Baek et al.. Character Region Awareness for Text Detection. (CRAFT) CVPR 2019

Methods: Detection based

- 相对能更好地区分文本行
- 长文本行回归不准

Liao et al.. TextBoxes: A Fast Text Detector with a Single Deep Neural Network. AAAI, 2017

Ma et al.. Arbitrary-Oriented Scene Text Detection via Rotation Proposals. (RRPN) arxiv, 2017

Zhou et al.. EAST: An Efficient and Accurate Scene Text Detector. CVPR, 2017

Xie et al.. Scene Text Detection with Supervised Pyramid Context Network. (SPCNET) AAAI, 2019

Liu et al.. Pyramid Mask Text Detector. (PMTD) arxiv, 2019

PMTD: Overview

Pyramid Mask Text Detector (https://arxiv.org/abs/1903.11800)

2018年以来多个基于Mask-RCNN的方法取得了较好的效果

Method	Precision	Recall	F-measure
SPCNET	73.40%	66.90%	70.00%
SPCNET (multi-scale)	80.60%	68.60%	74.10%
PMTD	85.15%	72.77%	78.48%
PMTD (multi-scale)	84.42%	76.25%	80.13%

PMTD: Baseline

训练

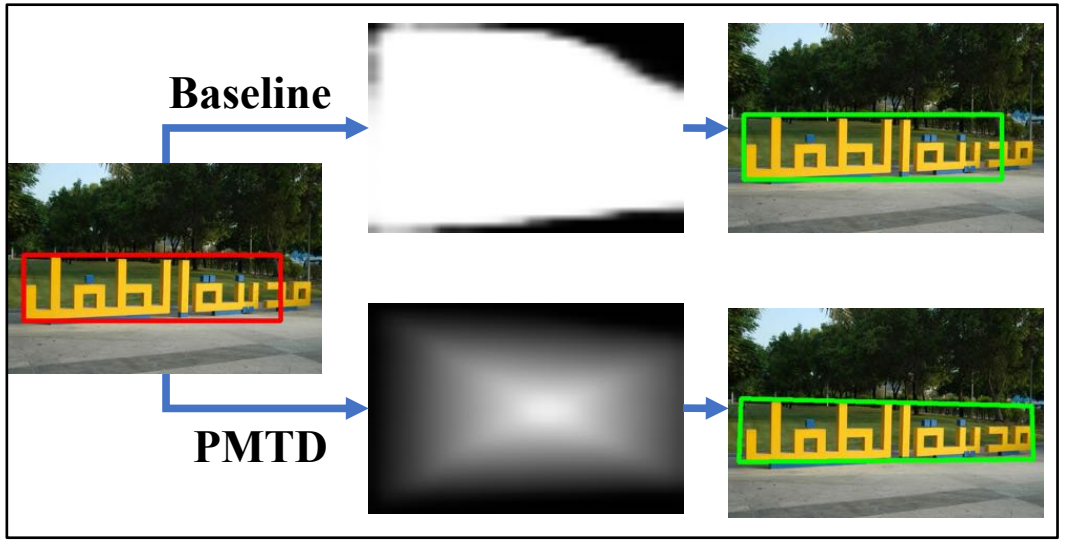
- 文字框内作为mask，文字框外接水平矩形作为bounding box
- 数据增强，filp + resize + crop (FOTS, TextNet, TextMountain, PMTD)
- 根据数据增强后的bounding box尺度和长宽比的统计设计anchor
- OHEM
- SyncBN

测试

- 在每个预测mask中寻找面积最大的连通区域
- 寻找该连通区域的最小外接矩形
- NMS得到最终输出

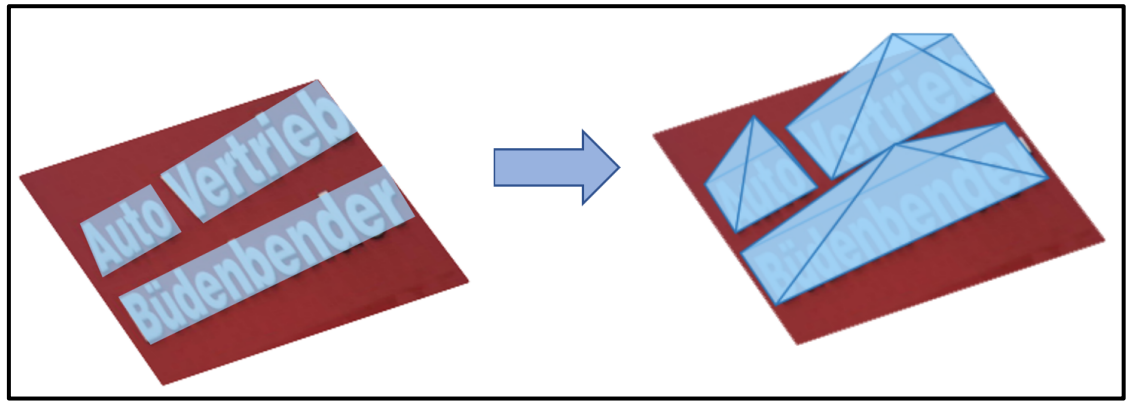
PMTD: Motivation

- 简化的监督信息
  - Baseline进行像素级分类，忽略了文字形状信息，inference时只利用了边缘信息
- 不精确的标签
- 误差传播



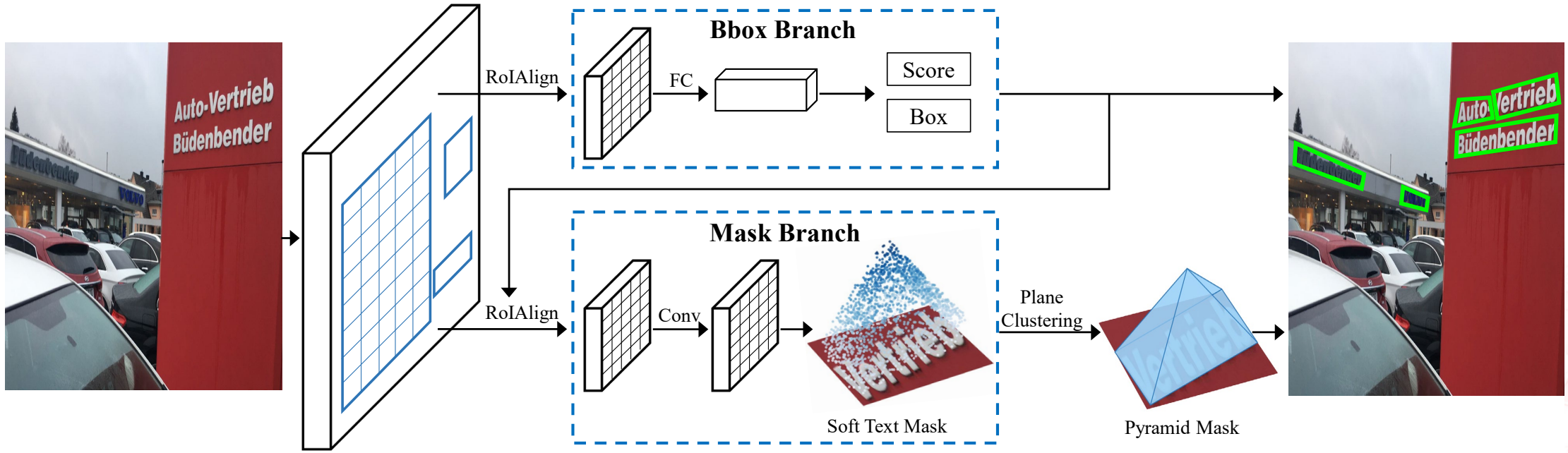
PMTD: Method

Pyramid mask将文字的形状和像素的位置信息编码到mask中



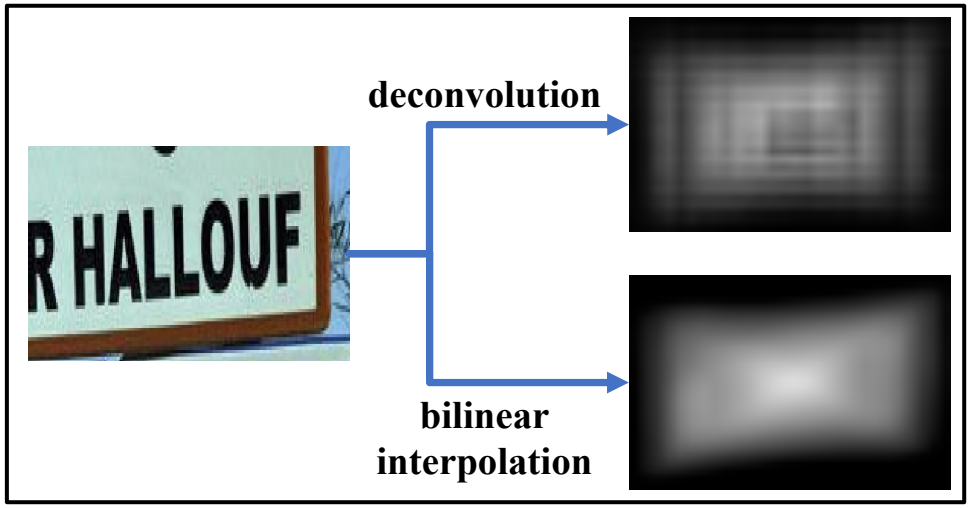
- 使用了更丰富的监督信息并在inference时利用了预测的soft mask
- 缓解了标签不精确问题
- 缓解了误差传播

PMTD: Pipeline



PMTD: Details

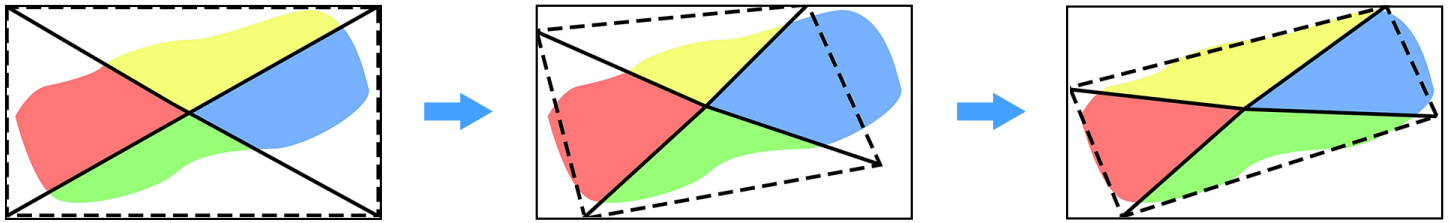
- Loss: $L = L_{\text{rpn}} + \lambda_1 L_{\text{cls}} + \lambda_2 L_{\text{box}} + \lambda_3 L_{\text{pyramid\_mask}}$
- Deconv -> bilinear Sample



- 将RCNN的四个卷积替换为空洞卷积增加感受野

PMTD: Plane Clustering

- 目的：将预测的soft mask转为pyramid，从而得到text box
- 方法：一种迭代更新的聚类算法
  - 四棱锥由四个侧面和底面表示，我们只需找到四个侧面，每个面可以表示为$Ax + By + Cz + D = 0, C = 1$
  - 先初始化四个平面，将soft mask每个点归为四个平面中距离最近的一个，形成四个点集
  - 得到四个点集后，使用鲁棒最小二乘(*RLS*)为每个点集找到距离最近的一个平面，更新四个平面
  - 迭代此过程，直到每个点距离所属平面的距离足够小为止



PMTD: Experiment

在ICDAR 2017 MLT, ICDAR 2015, ICDAR 2013取得了领先的结果

Method	Precision	Recall	F-measure
FOTS [26]	80.95	57.51	67.25
FOTS* [26]	81.86	62.30	70.75
Lyu *et al.* [31]	83.80	55.60	66.80
Lyu *et al.* * [31]	74.30	70.60	72.40
PSENet [20]	77.01	68.40	72.45
Pixel-Anchor [21]	79.54	59.54	68.10
Pixel-Anchor* [21]	83.90	65.80	73.76
SPCNET [41]	66.90	73.40	70.00
SPCNET* [41]	68.60	80.60	74.10
Huang *et al.* [14]	80.00	69.80	74.30
Baseline	84.72	70.37	76.88
PMTD	85.15	72.77	78.48
PMTD*	84.42	76.25	**80.13**

Table 1: Comparison with other results on ICDAR 2017 MLT. * means multi scale testing.

Method	Precision	Recall	F-measure
SegLink [37]	73.10	76.80	75.00
SSTD [8]	80.00	73.00	77.00
WordSup [12]	79.33	77.03	78.16
EAST* [46]	83.27	78.33	80.72
R2CNN [17]	85.62	79.68	82.54
DDR [10]	82.00	80.00	81.00
Lyu *et al.* * [31]	89.50	79.70	84.30
RRD* [24]	88.00	80.00	83.80
TextBoxes++* [22]	87.80	78.50	82.90
PixelLink [3]	85.50	82.00	83.70
FOTS [26]	91.00	85.17	87.99
IncepText* [42]	89.40	84.30	86.80
TextSnake [29]	84.90	80.40	82.60
FTSN [2]	88.60	80.00	84.10
SPCNET [41]	88.70	85.80	87.20
PSENet [20]	89.30	85.22	87.21
Baseline	85.84	90.55	88.14
PMTD	91.30	87.43	**89.33**

Table 2: Comparison with other results on ICDAR 2015. * means multi scale testing. For PMTD, we only report single scale testing result.

Method	ICDAR13 Eval	DetEval
CTPN [39]	85.00	86.00
SegLink [37]	-	85.30
TextBoxes* [23]	85.00	86.00
SSTD [8]	87.00	88.00
WordSup [12]	-	90.34
R2CNN [17]	87.73	-
DDR [10]	-	86.00
MCN [27]	88.00	-
Lyu *et al.* * [31]	88.00	-
RRD* [24]	89.00	-
TextBoxes++* [22]	88.00	89.00
PixelLink* [3]	-	88.10
FEN* [45]	91.60	92.30
FOTS* [26]	92.50	92.82
SPCNET [41]	92.10	-
Baseline	91.73	92.25
PMTD	**93.40**	**93.59**

Table 3: Comparison with other results on ICDAR 2013. * means multi scale testing. For PMTD, we only report single scale testing result.

PMTD: Experiment

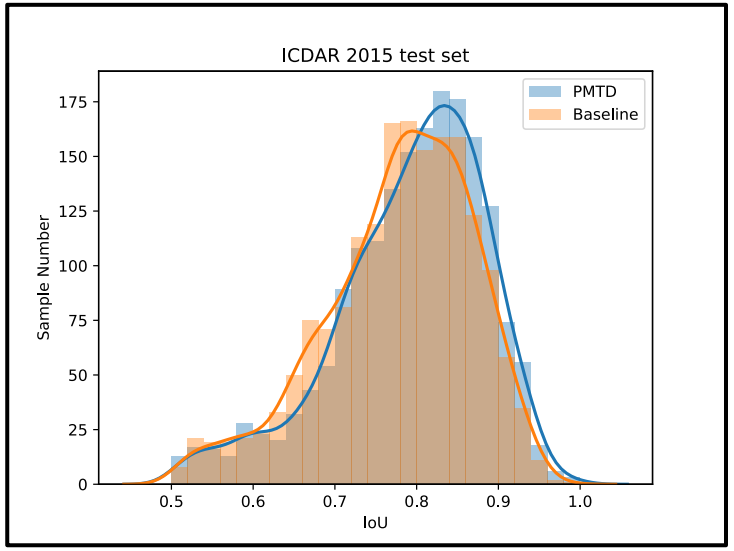
借助于包含更多信息的soft text mask，plane clustering算法可以：

- 预测更精准的文字边界
- 对预测不准确的bounding box更加鲁棒

然而现在的评价标准往往比较宽松(IoU > 0.5)，难以体现这个优势

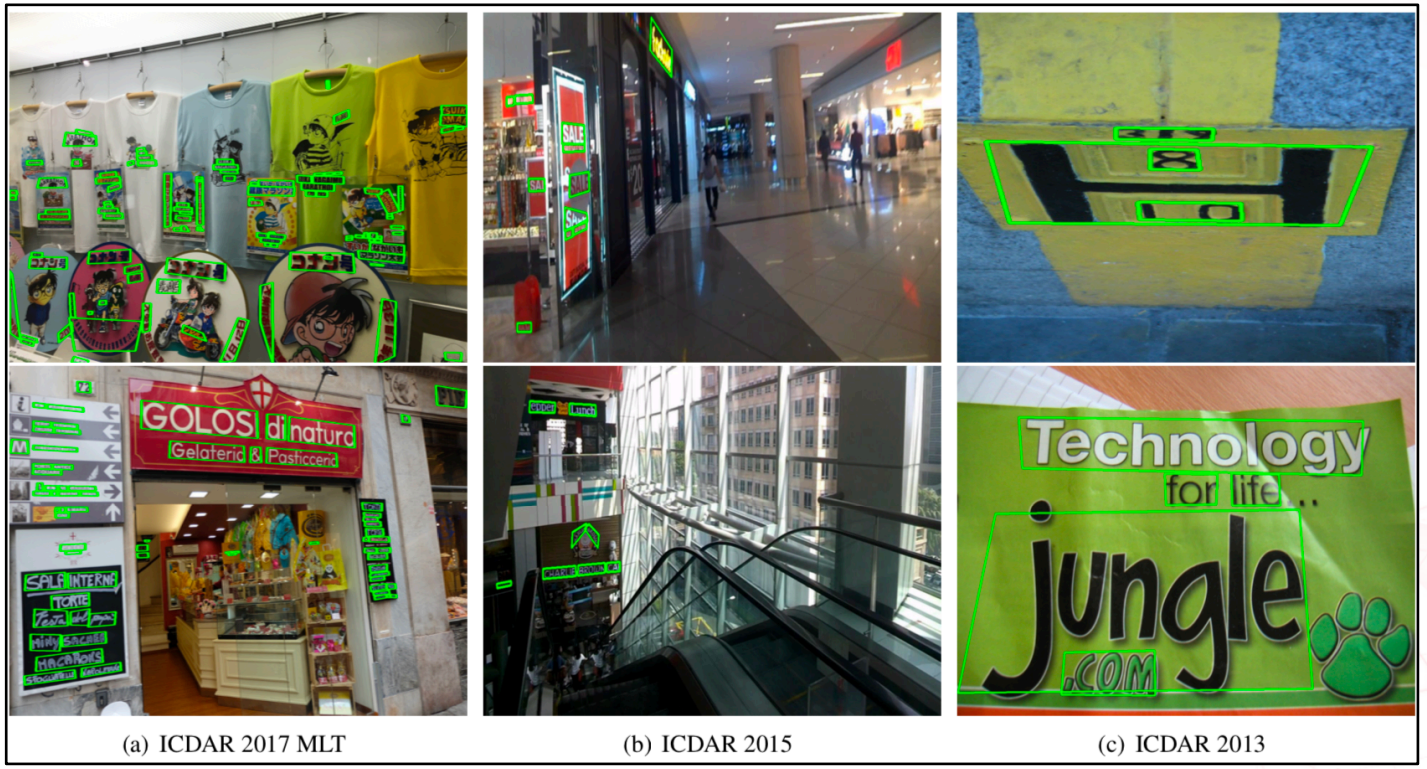
IoU	Matched number			F-measure		
	Baseline	PMTD	Relative improve	Baseline	PMTD	Relative improve
0.5	1784	1816	1.79%	88.14%	89.33%	1.35%
0.6	1696	1729	1.95%	83.60%	84.79%	1.42%
0.7	1443	1556	7.83%	70.44%	75.31%	6.91%
0.8	799	962	**20.40%**	38.36%	45.32%	**18.14%**
0.9	107	157	**46.73%**	5.14%	6.73%	**30.93%**

Table 5: Number of true positives and F-measure under different IoU threshold on ICDAR 2015. PMTD outperforms baseline significantly when IoU threshold is high.



PMTD: Experiment

PMTD对于不精确预测边界框的鲁棒性。从左到右依次为：不精确的边界框，预测的soft text mask，回归的text box。



(a) ICDAR 2017 MLT (b) ICDAR 2015 (c) ICDAR 2013

Qualitative results

PMTD: Summary

- Pyramid mask 相比 binary mask 是一种可以提供更丰富的信息的文本区域表示方式
- 我们提出了聚类算法Plane Clustering
- Pyramid mask 配合 Plane Clustering 可以得到更好的performance，更准确的文字框

Source code will be available: https://github.com/STVIR/PMTD

Outline

- Background
- Text detection
- From text detection to text spotting
- What’s next?

Why we need end-to-end text spotting

text detection + text recognition → text spotting

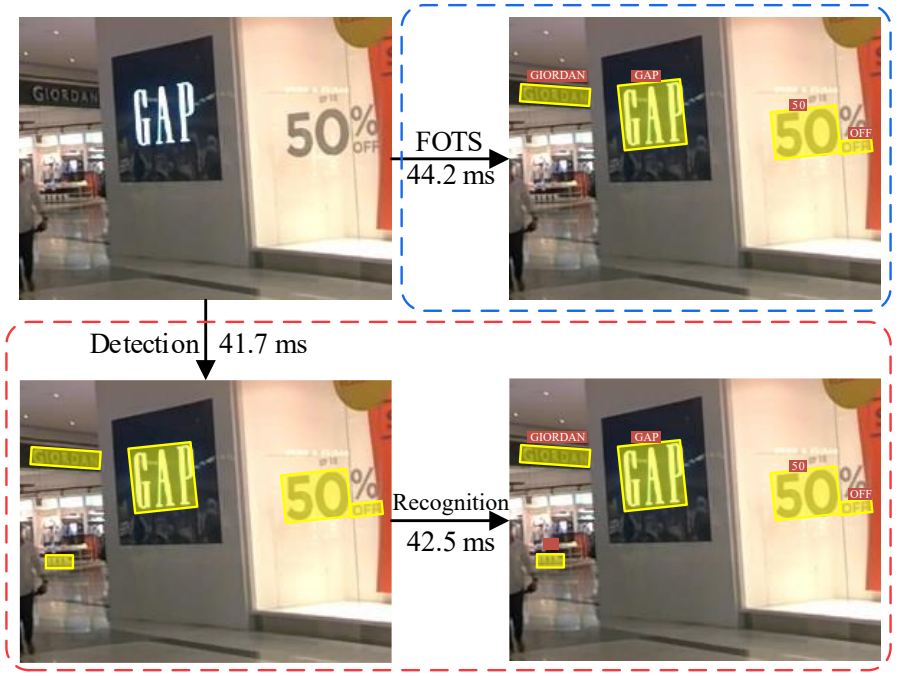
- 更快：检测与识别的特征有相关性，可以共享同一个特征提取网络
- 更好：检测和识别共同监督训练可以提升效果

Methods

- Li et al.. Towards End-to-end Text Spotting with Convolutional Recurrent Neural Networks. ICCV, 2017
- Liu et al.. FOTS: Fast Oriented Text Spotting with a Unified Network. CVPR, 2018
- Lyu et al.. Mask TextSpotter: An End-to-End Trainable Neural Network for Spotting Text with Arbitrary Shapes. ECCV, 2018

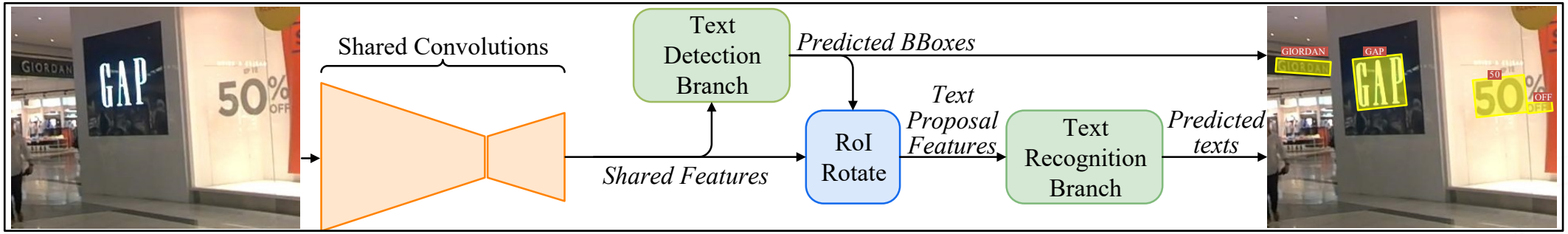
FOTS: Overview

- 首次将倾斜文本端到端识别做work
- 实验表明端到端训练可以较大提升检测效果
- 实验表明端到端方法相比传统的两步方法速度更快

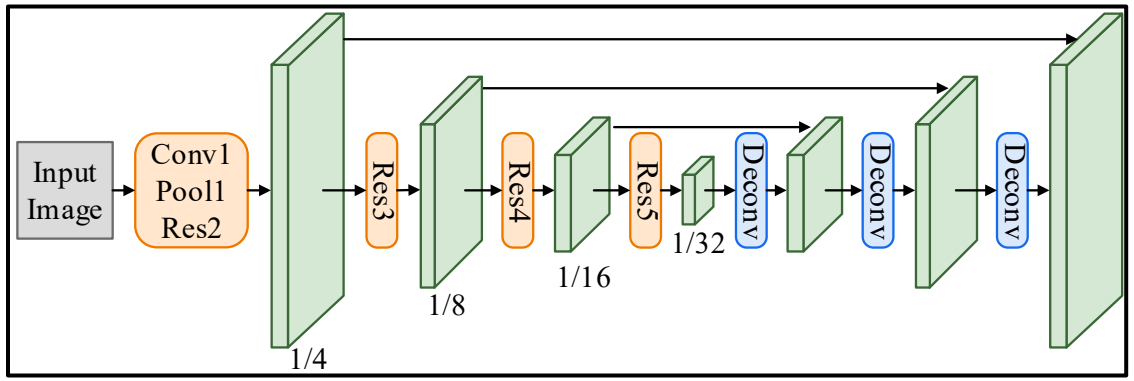


FOTS: Pipeline

- 共享的特征提取模块
- 基于EAST检测分支
- 基于CRNN的识别分支
- 提出RoI Rotate进行特征alignment



FOTS: Method



Architecture of shared convolutions

Type	Kernel [size, stride]	Out Channels
conv_bn_relu	[3, 1]	64
conv_bn_relu	[3, 1]	64
height-max-pool	[(2, 1), (2, 1)]	64
conv_bn_relu	[3, 1]	128
conv_bn_relu	[3, 1]	128
height-max-pool	[(2, 1), (2, 1)]	128
conv_bn_relu	[3, 1]	256
conv_bn_relu	[3, 1]	256
height-max-pool	[(2, 1), (2, 1)]	256
bi-directional_lstm		256
fully-connected		$\|S\|$

The detailed structure of the text recognition branch

FOTS: RoI Rotate

RoI Rotate将四边形区域的feature map投影到高度固定的水平矩形区域

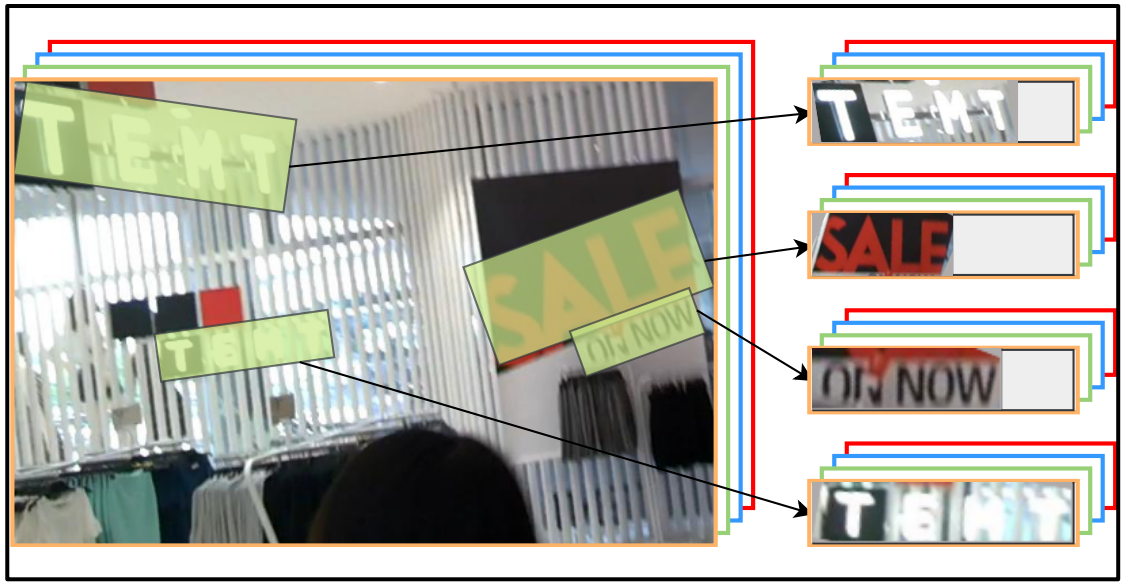


Illustration of RoIRotate

$$t_x = l * \cos\theta - t * \sin\theta - x \tag{4}$$

$$t_y = t * \cos\theta + l * \sin\theta - y \tag{5}$$

$$s = \frac{h_t}{t + b} \tag{6}$$

$$w_t = s * (l + r) \tag{7}$$

$$\begin{aligned} \mathbf{M} &= \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix} \\ &= s \begin{bmatrix} \cos\theta & -\sin\theta & t_x\cos\theta - t_y\sin\theta \\ \sin\theta & \cos\theta & t_x\sin\theta + t_y\cos\theta \\ 0 & 0 & \frac{1}{s} \end{bmatrix} \end{aligned} \tag{8}$$

FOTS: Experiment

端到端训练提升了检测效果，因为文字识别促使网络学习字符级的细节特征，而检测网络往往只能学到文字区域级的特征

Dataset	ICDAR 2015	ICDAR 2013	MLT
Our detection	85.31%	87.32%	66.69%
FOTS	87.99% (+2.68%)	88.30% (+0.98%)	67.25% (+0.56%)

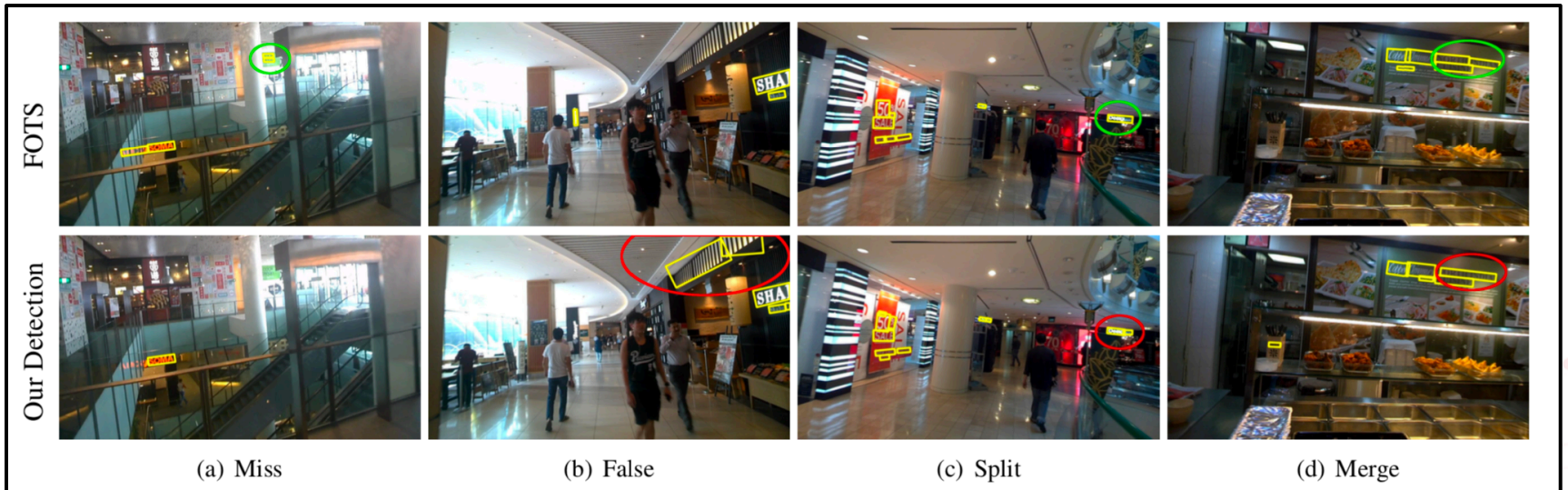
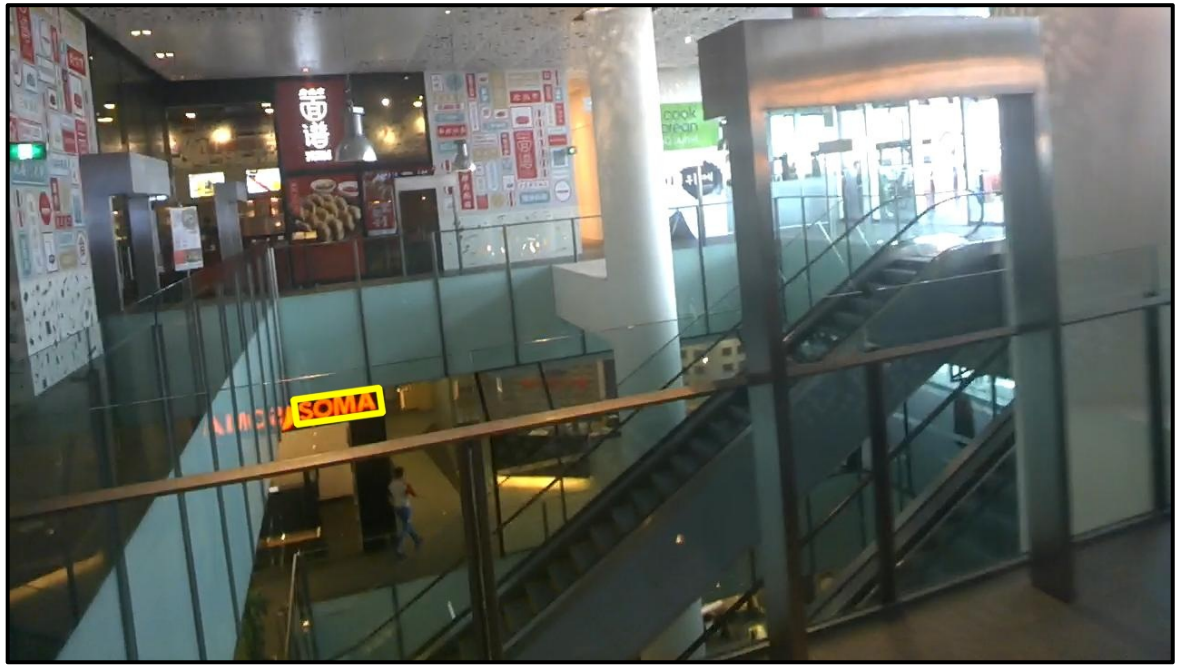
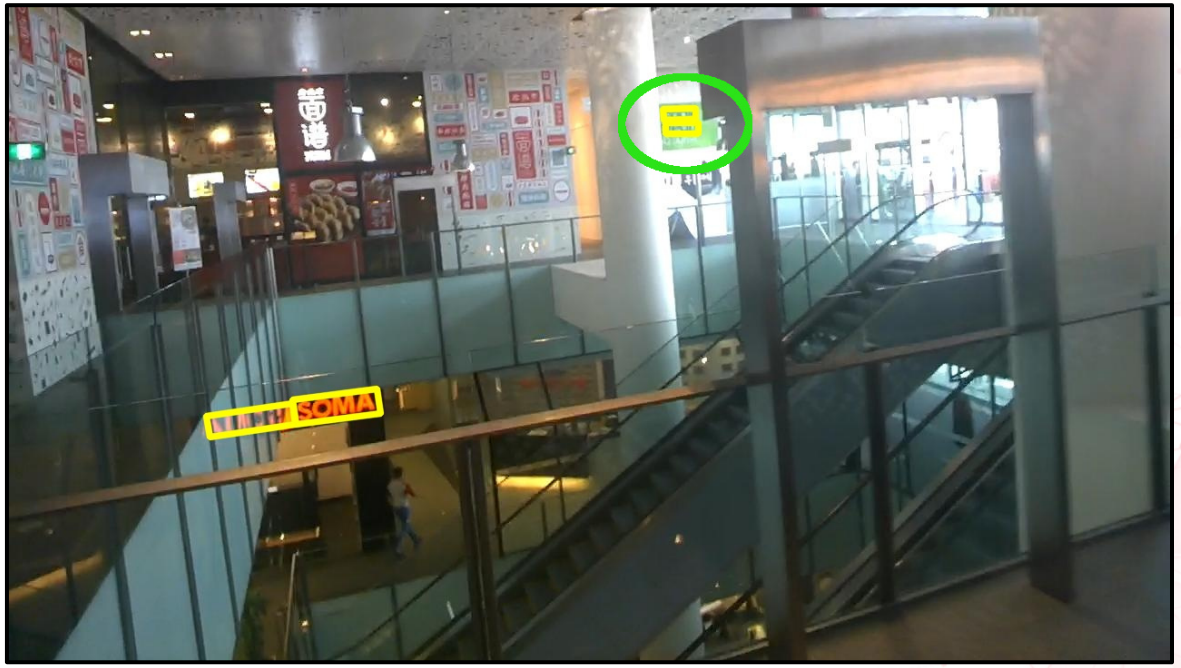


Figure 5: FOTS reduces Miss, False, Split and Merge errors in detection. Bounding boxes in green ellipses represent correct text regions detected by FOTS, and those in red ellipses represent wrong text regions detected by "Our Detection" method. Best view in color.

FOTS: Experiment



Our Detection

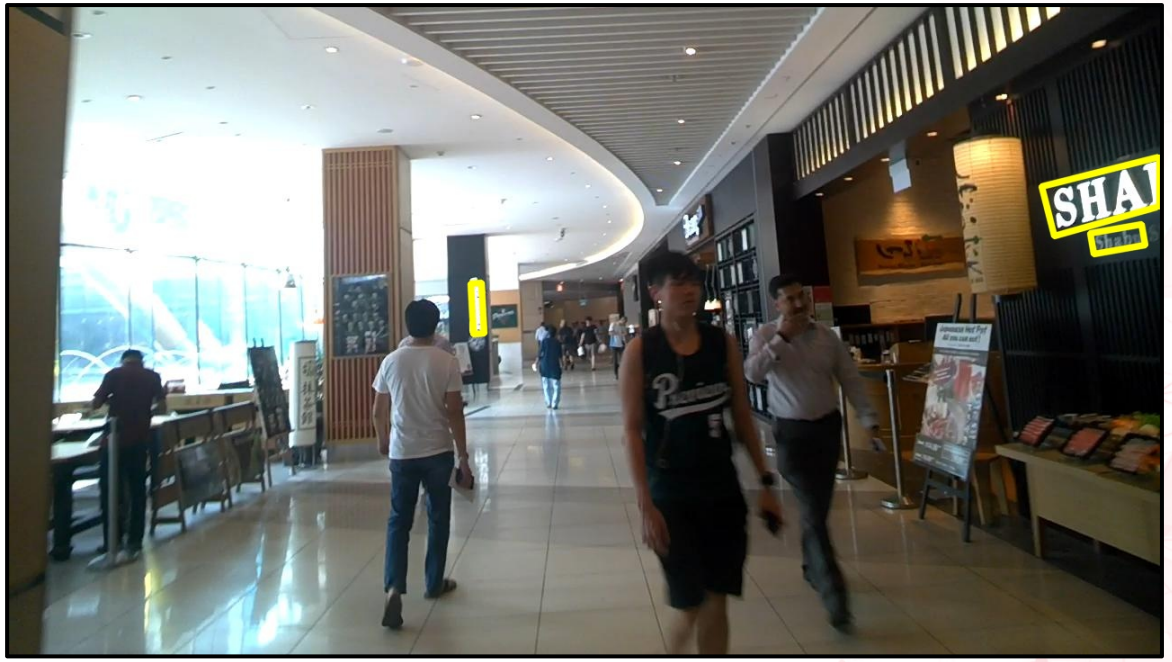


FOTS

Miss

FOTS: Experiment

Our Detection



FOTS

False

FOTS: Experiment

Our Detection

FOTS

Split

FOTS: Experiment

Our Detection

FOTS

Merge

FOTS: Experiment

FOTS相比两步方法速度提升接近两倍

Dataset	Method	Speed		Params
		Detection	End-to-End	
IC15	Our Two-Stage	7.8 fps	3.7 fps	63.90 M
	FOTS	7.8 fps	7.5 fps	34.98 M
	FOTS RT	24.0 fps	22.6 fps	28.79 M
IC13	Our Two-Stage	23.9 fps	11.2 fps	63.90 M
	FOTS	23.9 fps	22.0 fps	34.98 M

Table 5: Speed and model size compared on different methods. "Our Two-Stage" consists of a detection model with 28.67M parameters and a recognition model with 35.23M parameters.

FOTS: Experiment

在ICDAR 2015，ICDAR 2013取得了较好的结果

Results on ICDAR 2015

Method	Detection			Method	End-to-End			Word Spotting		
	P	R	F		S	W	G	S	W	G
SegLink [43]	74.74	76.50	75.61	Baseline OpenCV3.0+Tesseract [26]	13.84	12.01	8.01	14.65	12.63	8.43
SSTD [13]	80.23	73.86	76.91	Deep2Text-MO [51, 50, 20]	16.77	16.77	16.77	17.58	17.58	17.58
WordSup [17]	79.33	77.03	78.16	Beam search CUNI+S [26]	22.14	19.80	17.46	23.37	21.07	18.38
RRPN [39]	83.52	77.13	80.20	NJU Text (Version3) [26]	32.63	-	-	34.10	-	-
EAST [53]	83.27	78.33	80.72	StradVision_v1 [26]	33.21	-	-	34.65	-	-
NLPR-CASIA [15]	82	80	81	Stradvision-2 [26]	43.70	-	-	45.87	-	-
R^2CNN [25]	85.62	79.68	82.54	TextProposals+DictNet [7, 19]	53.30	49.61	47.18	56.00	52.26	49.73
CCFLAB_FTSN [4]	88.65	80.07	84.14	HUST_MCLAB [43, 44]	67.86	-	-	70.57	-	-
Our Detection	88.84	82.04	85.31	Our Two-Stage	77.11	74.54	58.36	80.38	77.66	58.19
FOTS	91.0	85.17	87.99	FOTS	81.09	75.90	60.80	84.68	79.32	63.29
FOTS RT	85.95	79.83	82.78	FOTS RT	73.45	66.31	51.40	76.74	69.23	53.50
FOTS MS	**91.85**	**87.92**	**89.84**	FOTS MS	**83.55**	**79.11**	**65.33**	**87.01**	**82.39**	**67.97**

Results on ICDAR 2013

Method	Detection		Method	End-to-End			Word Spotting		
	IC13	DetEval		S	W	G	S	W	G
TextBoxes [34]	85	86	NJU Text (Version3) [27]	74.42	-	-	77.89	-	-
CTPN [49]	82.15	87.69	StradVision-1 [27]	81.28	78.51	67.15	85.82	82.84	70.19
R^2CNN [25]	79.68	87.73	Deep2Text II+ [51, 20]	81.81	79.47	76.99	84.84	83.43	78.90
NLPR-CASIA [15]	86	-	VGGMaxBBNet(055) [20, 19]	86.35	-	-	90.49	-	76
SSTD [13]	87	88	FCRNall+multi-filt [10]	-	-	-	-	-	84.7
WordSup [17]	-	90.34	Adelaide_ConvLSTMs [32]	87.19	86.39	80.12	91.39	90.16	82.91
RRPN [39]	-	91	TextBoxes [34]	91.57	89.65	83.89	93.90	91.95	85.92
Jiang *et al.* [24]	89.54	91.85	Li *et al.* [33]	91.08	89.81	84.59	94.16	92.42	**88.20**
Our Detection	86.96	87.32	Our Two-Stage	87.84	86.96	80.79	91.70	90.68	82.97
FOTS	88.23	88.30	FOTS	88.81	87.11	80.81	92.73	90.72	83.51
FOTS MS	**92.50**	**92.82**	FOTS MS	**91.99**	**90.11**	**84.77**	**95.94**	**93.90**	87.76

FOTS: Summary

- 提出RoI Rotate 解决了四边形feature map的align问题
- 验证了end-to-end text spotting在检测效果和速度上的优势

待解决问题：

- 如何提高识别的performance
- 如何利用不完备的数据

Outline

- Background
- Text detection
- From text detection to text spotting
- What’s next?

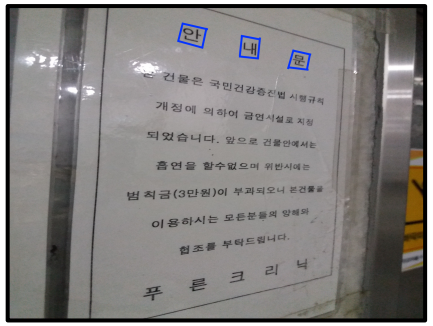
What’s next?

- Evaluation
- Problem definition
- Speed

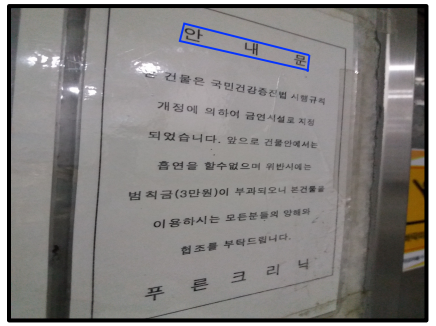
Evaluation

- IoU > 0.5 但是无法识别完整

- 难以定义ground truth



VS



- 错误split或merge文本行

  不一定影响后续使用

VS

Solutions

- 新的评价方法
  - Liu et al.. Tightness-aware Evaluation Protocol for Scene Text Detection. CVPR 2019
  - MTWI dataset
- 端到端评价
  - 检测识别端到端评价
  - 结合任务端到端评价
- 多种ground truth
  - ICDAR 2019 ReCTS

Problem definition

文字检测的粒度：字符，单词，文本行

为什么目前主流检测数据集和方法以单词或文本行为基本单位？

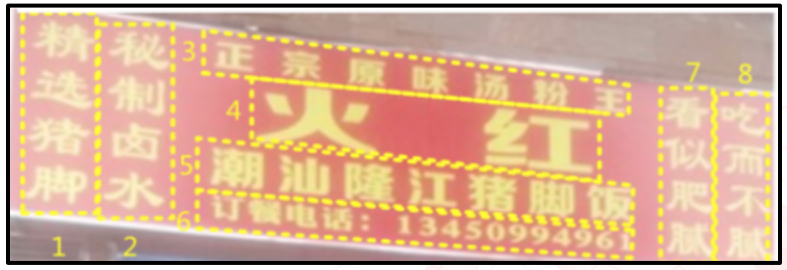
- 方便后续使用
- 标注成本低

Problem definition

缺点：

- 标注的歧义性
- 文本行连接/分离难以学习
- 形成很多长文本，难以回归
- 方便后续使用？
- 视觉足够吗？

VS



Problem definition

字符级检测？

优点：

- 字符是多种语言的基本单位，低歧义性
- Variance小，易于检测

缺点：

- 标注成本高
- 怎么group？

Problem definition

Methods:

- Liu et al.. Detecting Text in the Wild with Deep Character Embedding Network. arxiv 2019
- Baek et al.. Character Region Awareness for Text Detection. (CRAFT) CVPR 2019

Datasets:

- Synth800k
- CTW
- ICDAR 2019 ReCTS

Speed

- Object Recognition: MobileNet, ShuffleNet
- Object Detection: Tiny-DSOD, Pelee, ThunderNet
- Text Detection needs fast methods.

Speed

我们在水平文字检测上验证小模型可以在精度下降不多的情况下提速数十倍

网络结构	F-score(Deteval)	mFLOPs(resize long side to 1280)
ResNet50	94.01%	72731
ShuffleNet V2	90.79%	2546
ShuffleNet V2 0.5	89.75%	1055

Results on ICDAR2013

Speed

将FOTS的backbone从ResNet50替换为ShuffleNet，ICDAR 2015的F-score明显下降

以文本行为单位进行文字检测难以做快：

- 文本行定义的歧义性
- 文本行连接/分离难以学习
- 长文本难以回归

Summary

- 更合适的文字检测评价标准
- 文本行是否是一个好的粒度？字符级检测值得尝试
- 快速的文字检测和识别方法

Thank you!