场景文本识别与检测

–Overview

霍超凡

SHANDONG UNIVERSITY

2020 年 1 月 12 日

目录

目录

场景文本识别

Figure: 场景文本识别的一些例子[1]

[1] 来自《Deep Structured Output Learning for Unconstrained Text Recognition》

场景文本检测

Figure: 场景文本检测的一些例子[2]

[2] 来自《EAST：An Efficient and Accurate Scene Text Detector》

目录

STR 中较难处理的例子

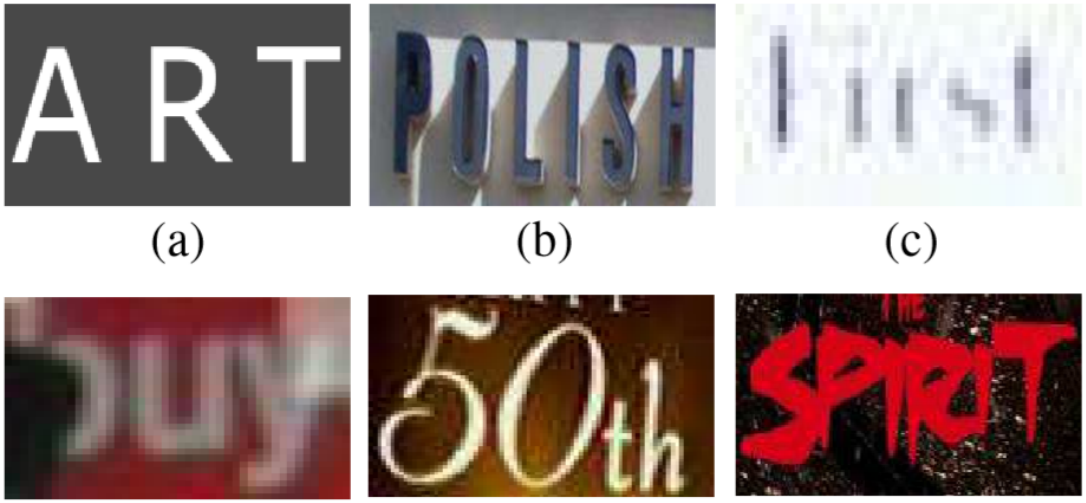


Figure 1. Examples of complicated / low-quality images. Subfigures (a) - (f) represent normal, complex background, blur, incomplete, different-size and abnormal font images, respectively.

3

[3] 来自《Focusing Attention: Towards Accurate Text Recognition in Natural Images》

STD 中较难处理的例子

Figure: 富有挑战性的例子[4]

[4] 来自 ICDAR 2019 – ArT

目录

四个阶段

- 矫正图片
  - STN、None
- 特征提取
  - VGG、ResNet、RCNN(Recurrent CNN，not R-CNN)
- 序列建模
  - RNN、LSTM、GRU
- 预测输出
  - CTC、Attention

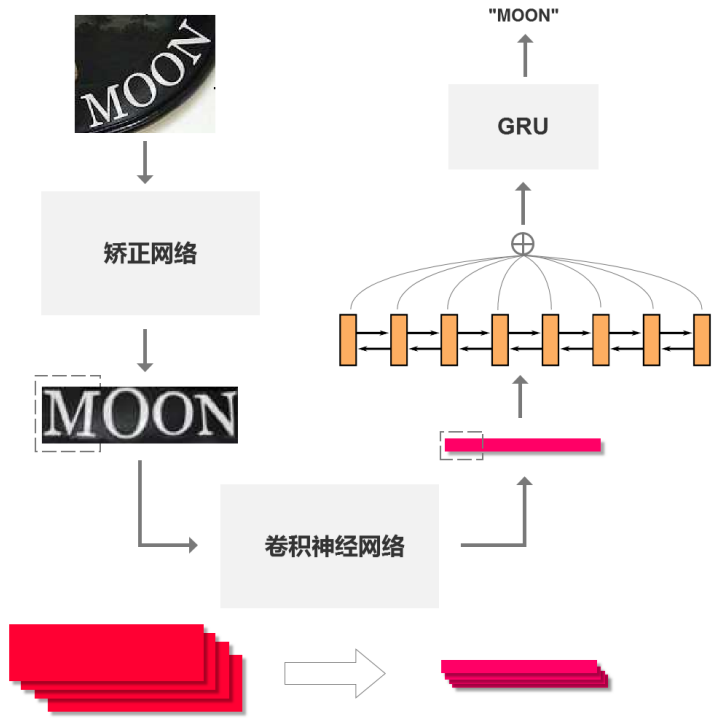


Figure: STR pipeline[a]

[a] 参考《Robust Scene Text Recognition with Automatic Rectification》

目录

CRNN

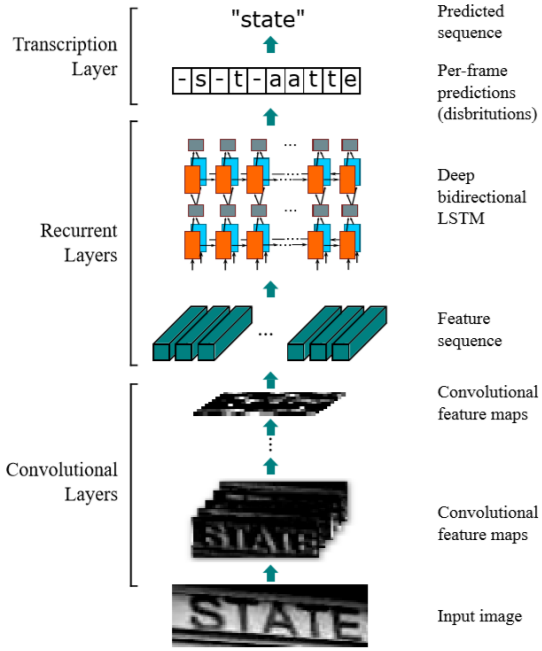


Figure 1. The network architecture. The architecture consists of three parts: 1) convolutional layers, which extract a feature sequence from the input image; 2) recurrent layers, which predict a label distribution for each frame; 3) transcription layer, which translates the per-frame predictions into the final label sequence.

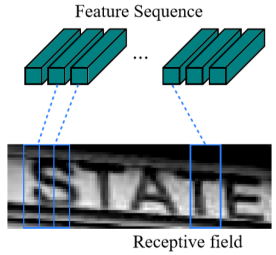


Figure 2. The receptive field. Each vector in the extracted feature sequence is associated with a receptive field on the input image, and can be considered as the feature vector of that field.

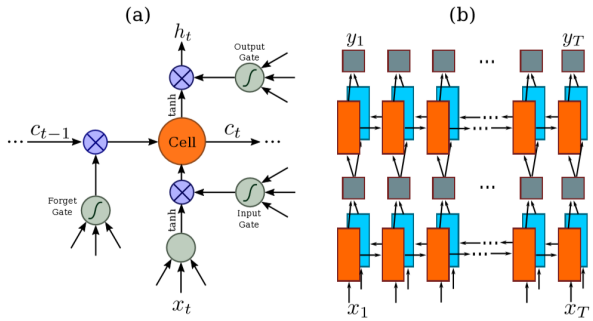


Figure 3. (a) The structure of a basic LSTM unit. An LSTM consists of a cell module and three gates, namely the input gate, the output gate and the forget gate. (b) The structure of deep bidirectional LSTM we use in our paper. Combining a forward (left to right) and a backward (right to left) LSTMs results in a bidirectional LSTM. Stacking multiple bidirectional LSTM results in a deep bidirectional LSTM.

Figure: CRNN 结构 [5]

[5] 来自《An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition》

CRNN 特点及优势

- CNN 提取特征 +RNN 序列建模
- 采用 RNN 编码，支持处理任意长度的文本。
- 将语言模型融入识别，自带纠错功能。

CRNN 之后的其它工作

- 引入 Attention 机制，替代 CTC。
- 解决 Attention Drift 问题。
- 引入 SAM（Spacial Attention Mechanism）处理任意排列的文本。
- 采用监督式训练方式，缓解 Attention 难以训练的问题。

目录

- R-CNN 系列
  - R-CNN: 传统方法提取候选框 + CNN 提取特征 + SVM 分类
  - Fast R-CNN: RoI pooling + 回归和分类结合
  - Faster R-CNN: RPN 提取候选框
  - Mask R-CNN: 像素级别
- 语义分割
  - FCN
  - FPN

FPN

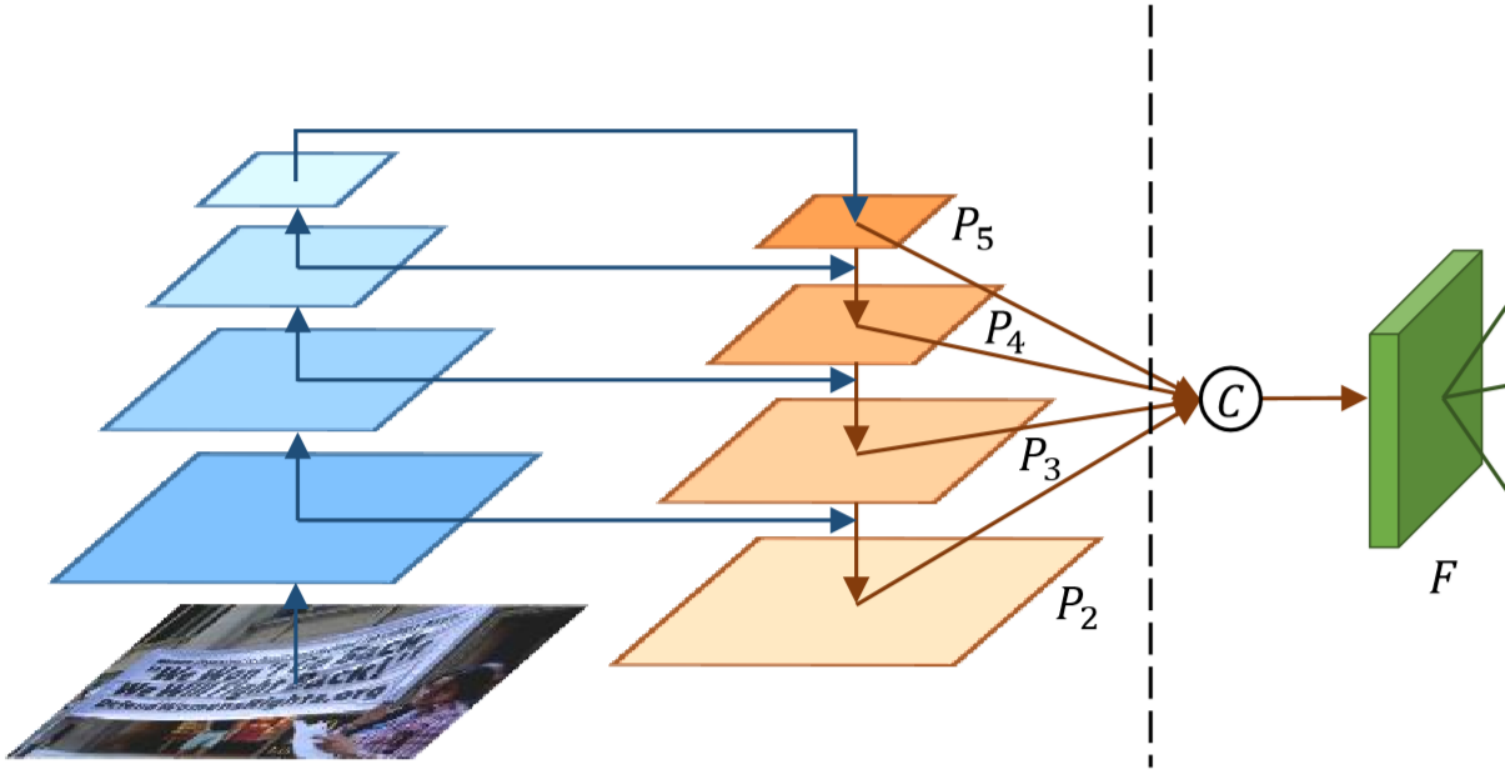


Figure: FPN 网络结构 [6]

[6] 来自《Shape Robust Text Detection with Progressive Scale Expansion Network》

目录

结合 R-CNN

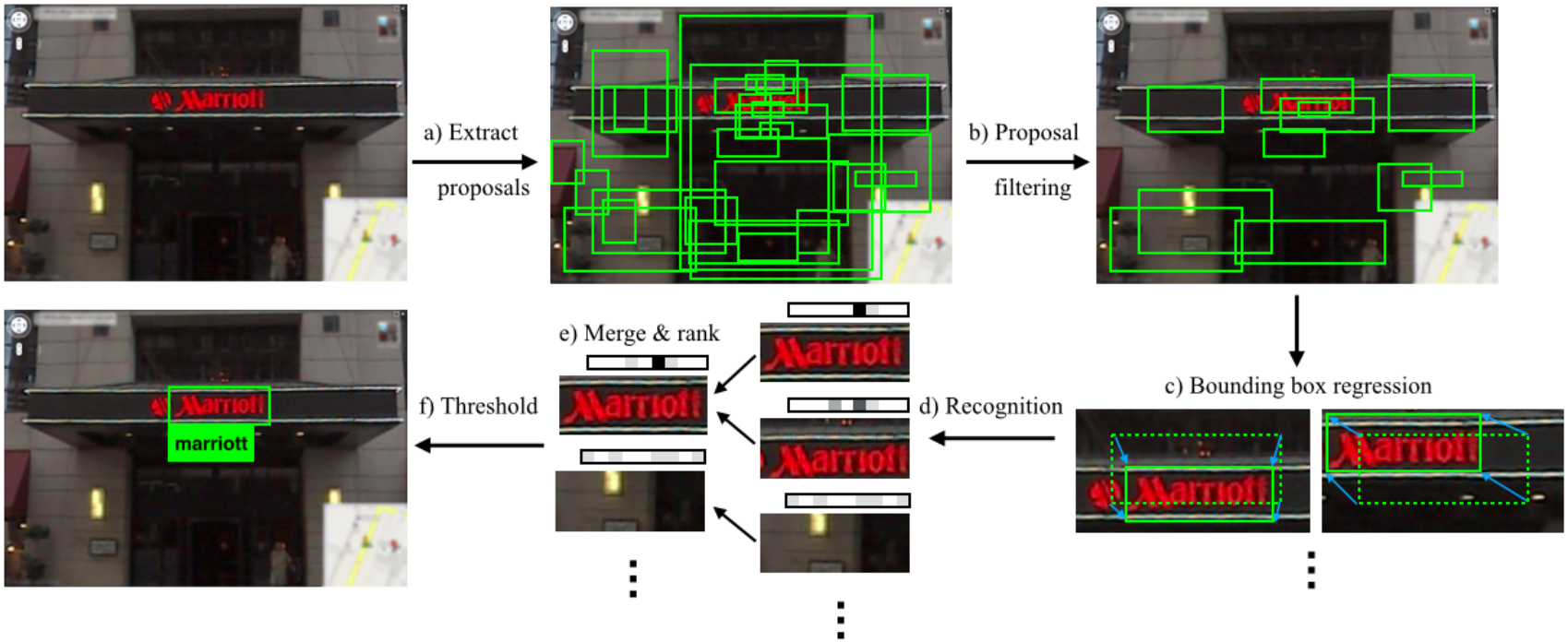


Figure: 网络结构 [7]

[7] 来自《Reading Text in the Wild with Convolutional Neural Networks》

结合 Faster R-CNN

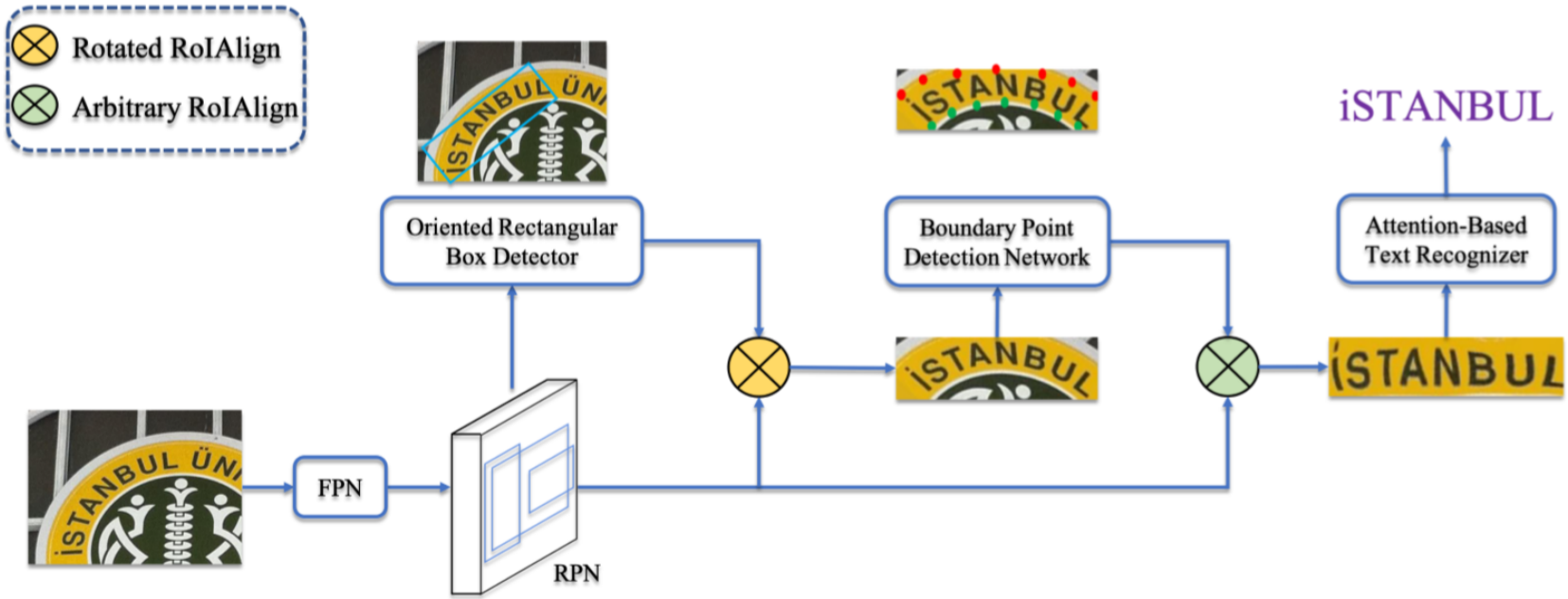


Figure: Boundary 网络结构 [8]

[8] 来自《All You Need Is Boundary：Toward Arbitrary-Shaped Text Spotting》

结合 Mask R-CNN

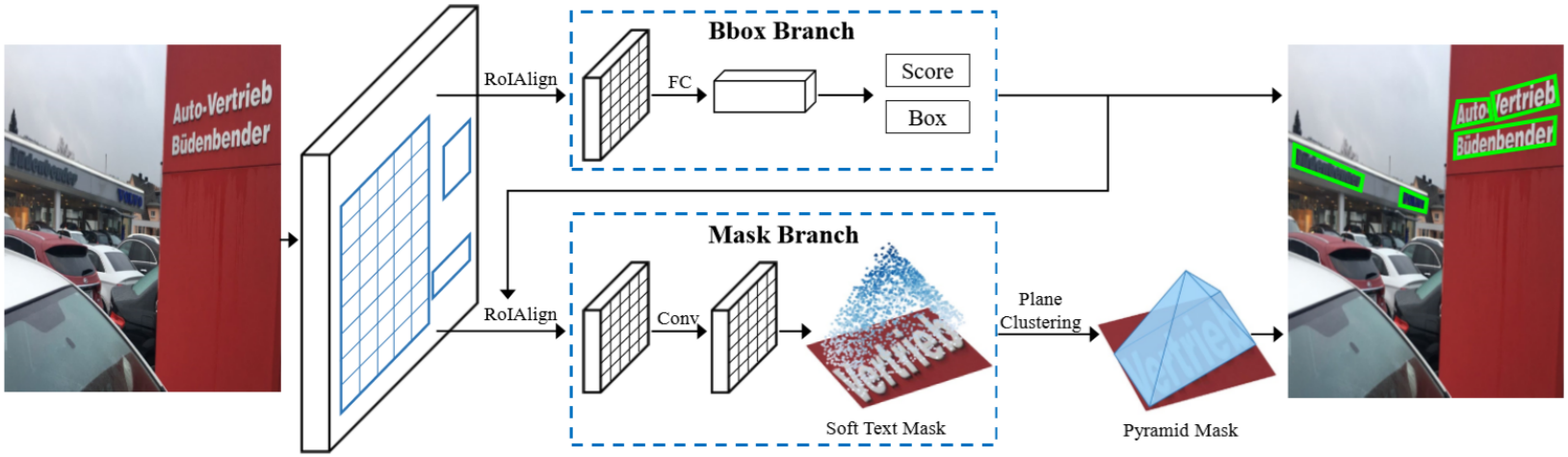


Figure: PMTD 网络结构 [9]

[9] 来自《Pyramid Mask Text Detector》

目录

矩形框

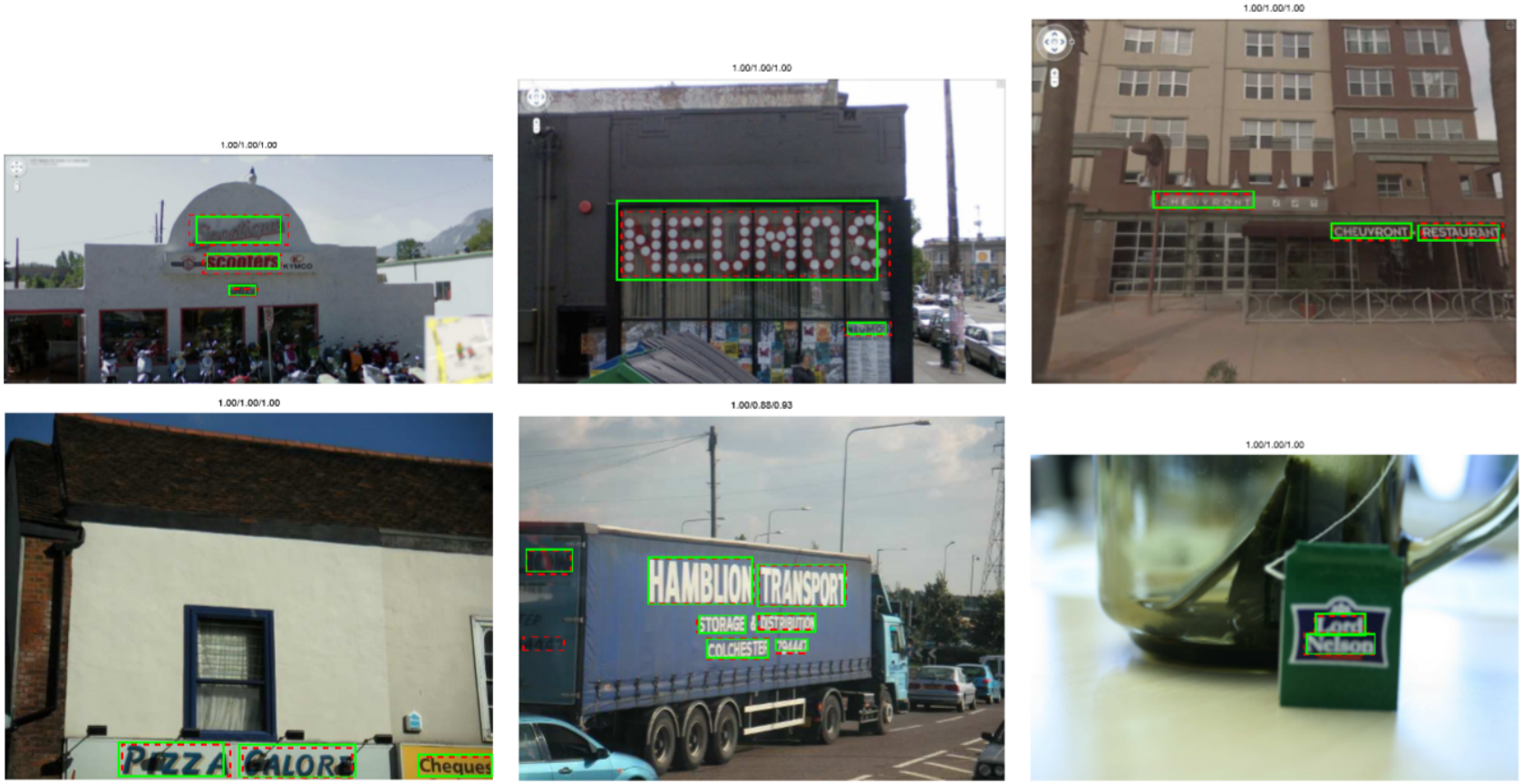


Figure: 矩形框描述区域 [10]

[10] 来自《Reading Text in the Wild with Convolutional Neural Networks》

旋转矩形框

Figure: 旋转矩形框描述区域 [11]

[11] 来自《EAST:An Efficient and Accurate Scene Text Detector》

边缘点

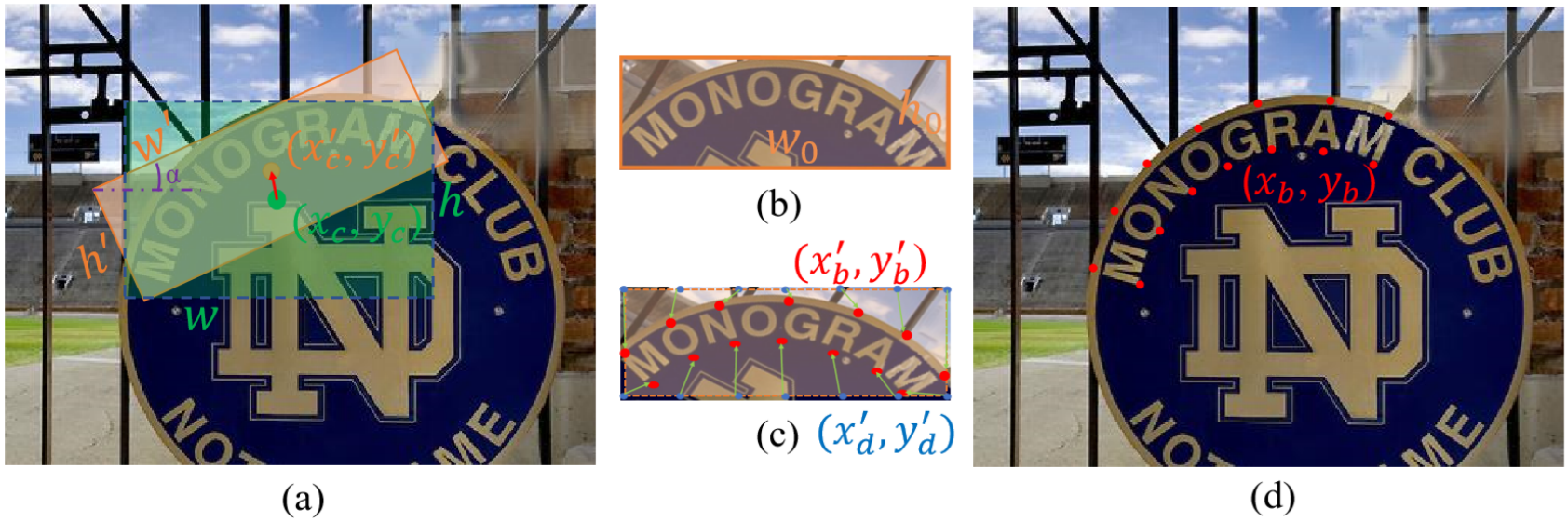


Figure: 边缘点描述区域 [12]

[12] 来自《All You Need Is Boundary：Toward Arbitrary-Shaped Text Spotting》

边缘点

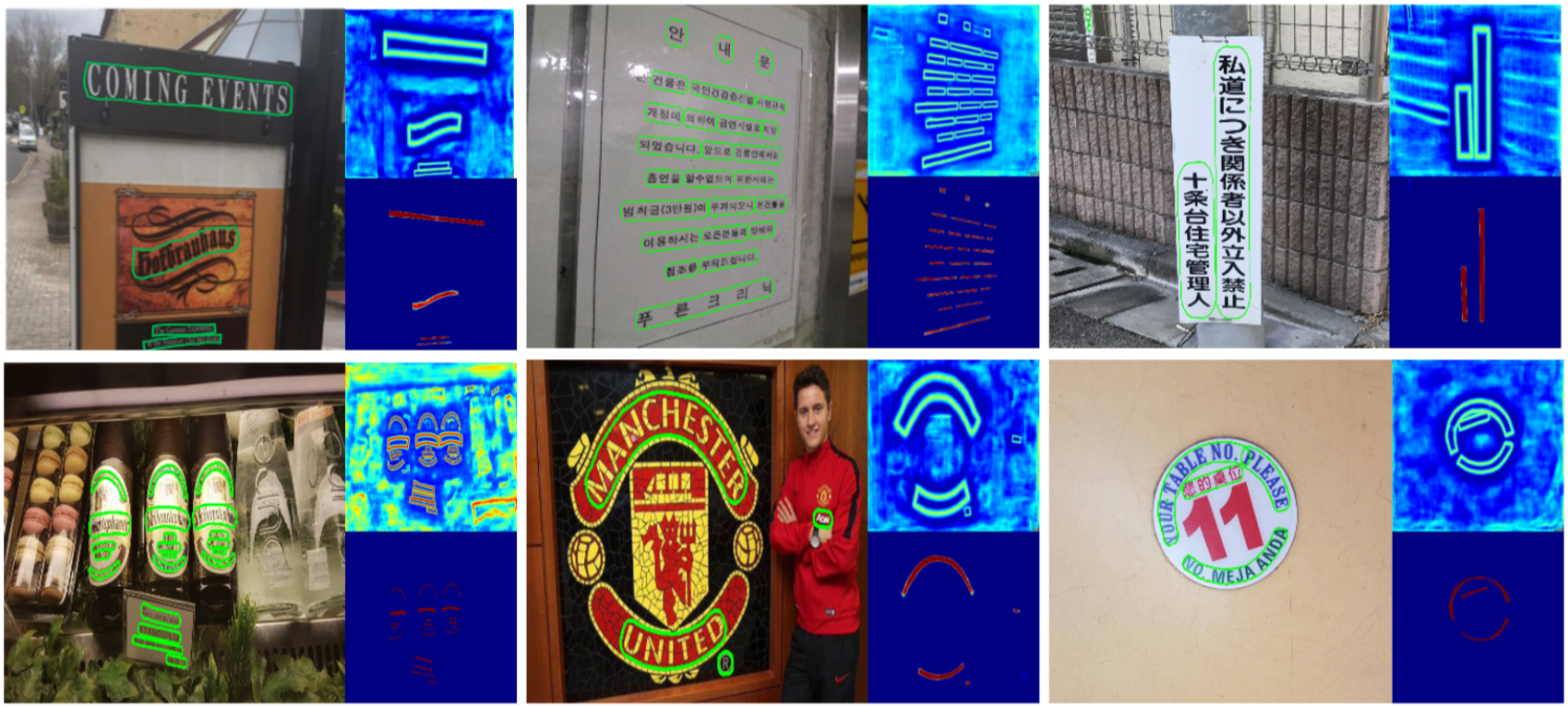


Figure: 边缘点描述区域 [13]

[13] 来自《Real-time Scene Text Detection with Differentiable Binarization》

像素级别

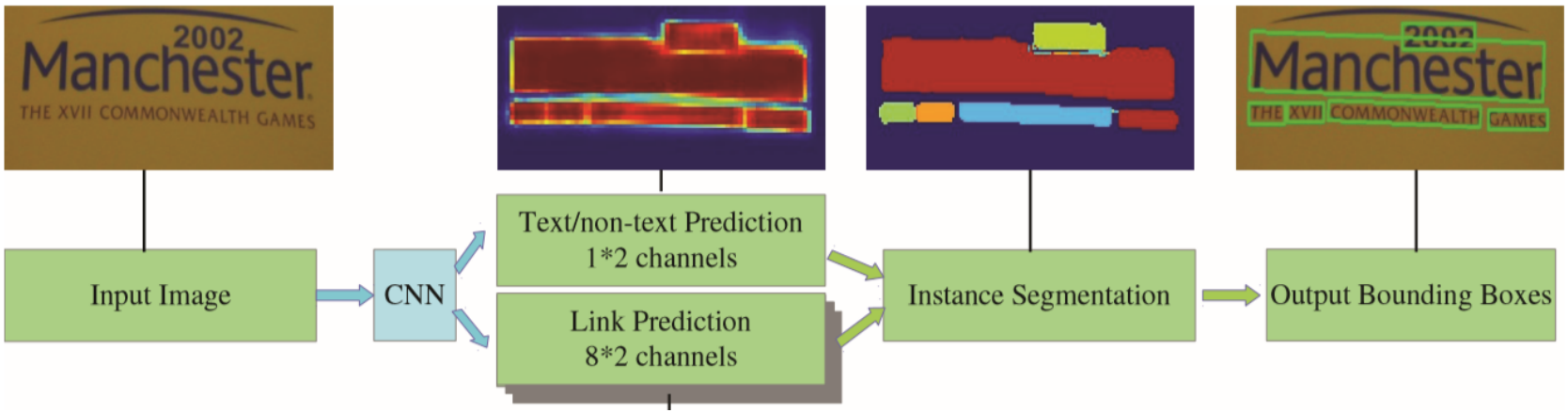


Figure: Pixel-Link 例子 [14]

[14] 来自《Pixel Link: Detecting Scene Text via Instance Segmentation》

一个新颖的方法

Figure: TextSnake 描述区域 [15]

[15] 来自《TextSnake: A Flexible Representation for Detecting Text of Arbitrary Shapes》

目录

总结

- 处理任意方式排列的文本（尤其是曲线）
- 采用恰当的方式描述文本区域，来减少后处理的时间
- 字符级别的检测，字符表示难以获取？半监督

一个例子

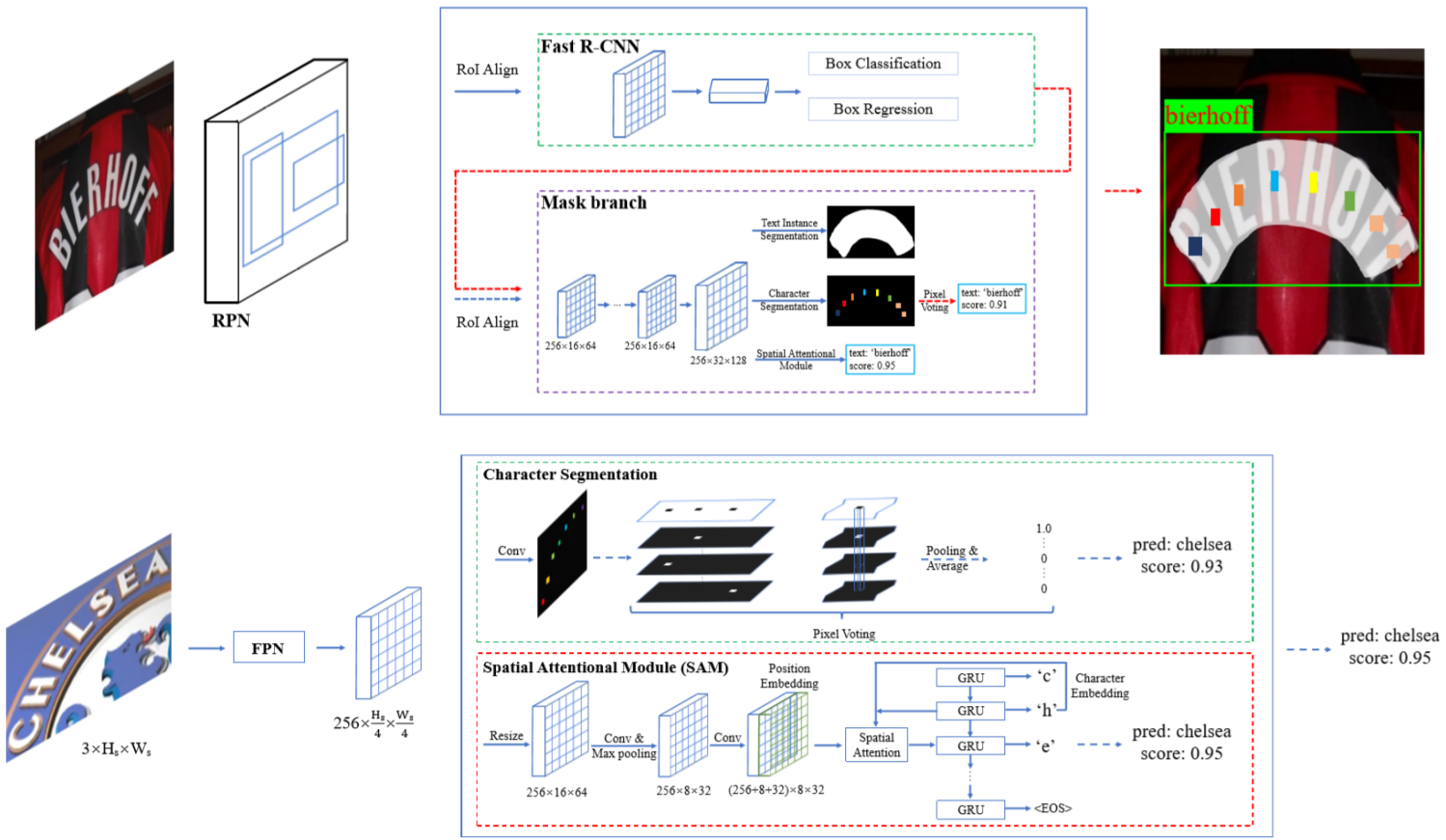


Figure: Mask TextSpotter [16]

[16] 来自《Mask TextSpotter: An End-to-End Trainable Neural Network for Spotting Text with Arbitrary Shapes》

不能很好的衔接

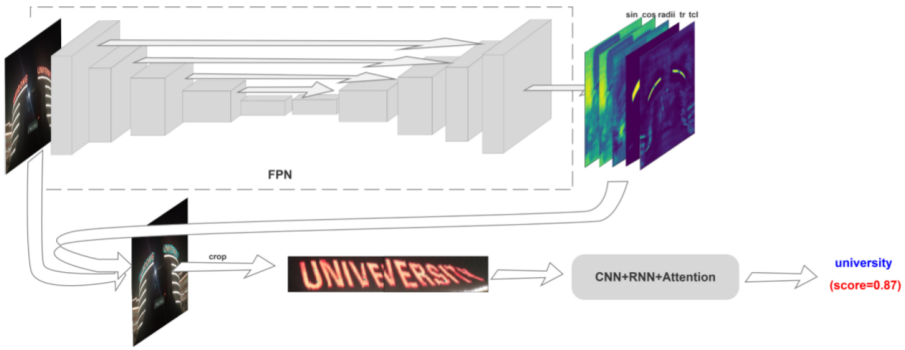
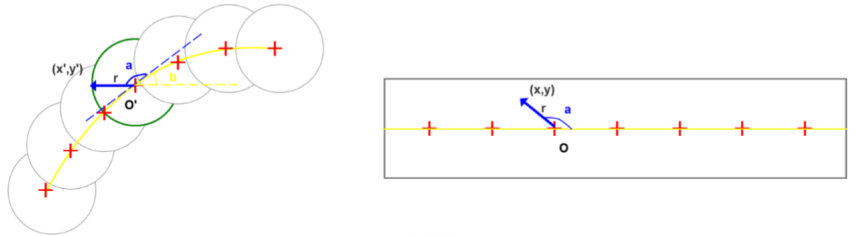
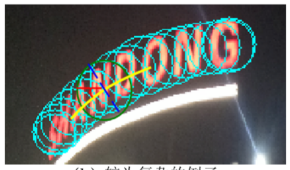


图 2: TextSnake+CRNN网络结构图



(a) 坐标转换图解



(b) 较为复杂的例子

(c) 能够胜任的例子

(d) 矫正结果

Figure: TextSnake + CRNN

展望

- STR 和 STD 融合
- 其它非英语系语言的识别