Cascaded Segmentation-Detection Networks for Word-Level Text Spotting

主要思想： FCN+YOLO

一个全卷机网络，得出比较粗糙的但比较稳定的分割；
一个类似 YOLO 的检测框架模型，在上述的基础上，进行区域回归，得到每个单字符的区域；

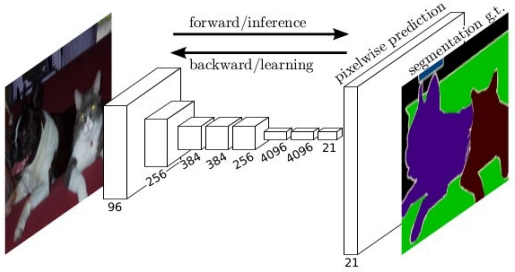


Figure 1. Fully convolutional networks can efficiently learn to make dense predictions for per-pixel tasks like semantic segmentation.

FCN 的缺点：

1. 得到的结果不够精确。进行 8 倍上采样虽然比 32 倍的效果好很多，但是上采样的结果还是比较模糊和平滑，对图像中的细节不够敏感；

2. 对各个像素进行分类，没有充分考虑像素和像素之间的关系。忽略了在通常的基于像素分类的分割方法中的使用空间规整 (spatial regularization) 步骤，缺乏空间一致性。

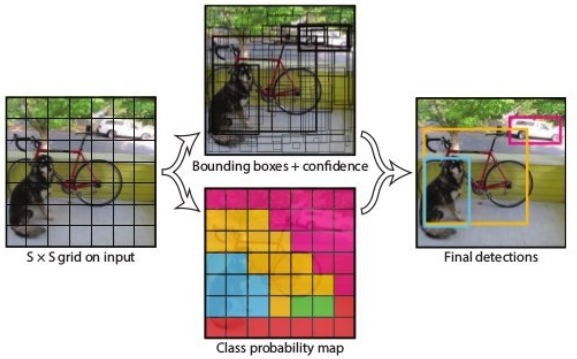


Figure 2: The Model. Our system models detection as a regression problem. It divides the image into an $S \times S$ grid and for each grid cell predicts B bounding boxes, confidence for those boxes, and C class probabilities. These predictions are encoded as an $S \times S \times (B * 5 + C)$ tensor.

YOLO 的缺点：

针对字符检测

1.yolo 每个 cell 每次只产生两个 anchor(两个区域)，如果存在多余 2 个的字符出现在一个 cell 中，则会出像漏检的情况；

2. 如果增加 anchor 的数量或者改变 cell 的尺寸，会导致计算量增加；

汽车之家 autohome.com.cn

Cascaded Segmentation-Detection Networks for Word-Level Text Spotting

主要思想： FCN+YOLO

一个全卷机网络，得出比较粗糙的但比较稳定的分割；
一个类似 YOLO 的检测框架模型，在上述的基础上，进行区域回归，得到每个单字符的区域；

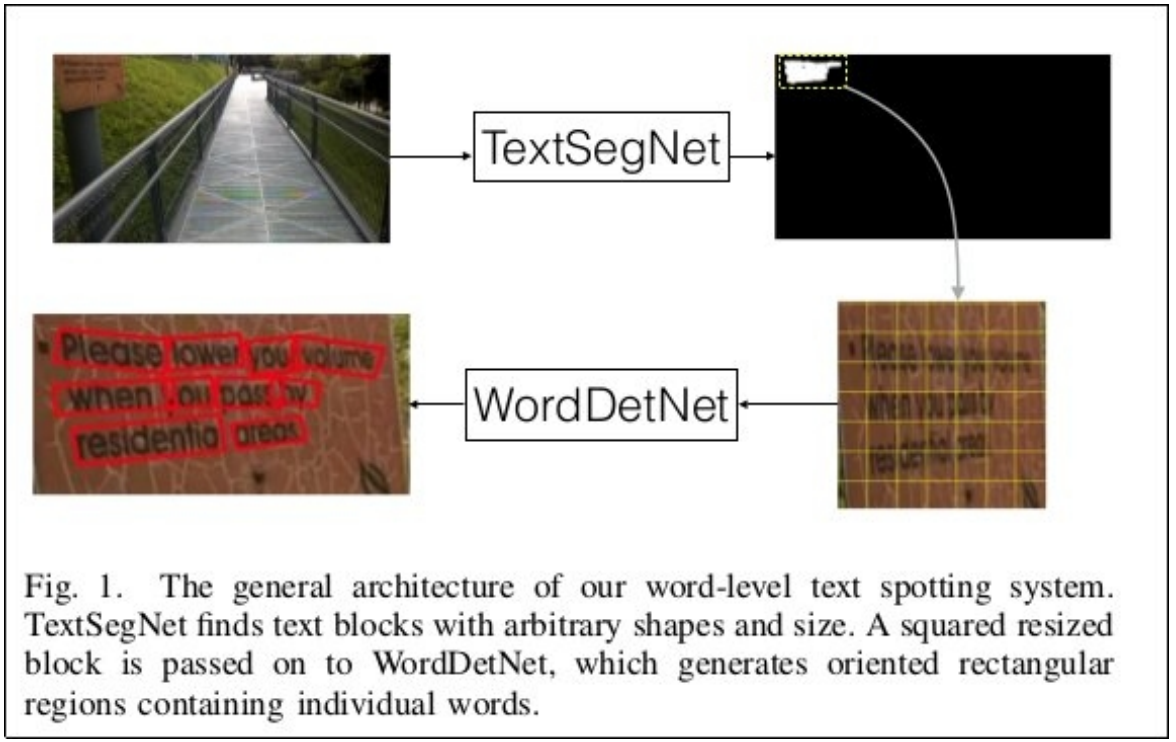


Fig. 1. The general architecture of our word-level text spotting system. TextSegNet finds text blocks with arbitrary shapes and size. A squared resized block is passed on to WordDetNet, which generates oriented rectangular regions containing individual words.

汽车之家
autohome.com.cn

Cascaded Segmentation-Detection Networks for Word-Level Text Spotting

主要思想： FCN+YOLO

一个全卷机网络，得出比较粗糙的但比较稳定的分割；
一个类似 YOLO 的检测框架模型，在上述的基础上，进行区域回归，得到每个单字符的区域；

Input image FCN segmentation

Fig. 2. FCN is in general unable to separate individual words.

Fig. 3. By cascading segmentation (TextSegNet) with detection (WordDetNet), our algorithm can detect both large and small words. If only WordDetNet is used, as trained on whole images, detection will fail in the case of too small words. Detected words by TextSegNet + WordDetNet are shown by red rectangles, the result of WordDetNet only are shown by yellow rectangles.

汽车之家
autohome.com.cn

Cascaded Segmentation-Detection Networks for Word-Level Text Spotting

主要思想： FCN+YOLO

TextSegNet: 一个全卷机网络，得出比较粗糙的但比较稳定的分割；
TextDetNet: 一个类似 YOLO 的检测框架模型，在上述的基础上，进行区域回归，得到每个单字符的区域；

检测部分的

$$L(\mathbf{P}) = \sum_{i=1}^{N\times N} \delta_i^{obj} \sum_{j=1}^{B} \left[\delta_i^j L^{obj}(\mathbf{p}_i^j) + (1-\delta_i^j)(C_i^j)^2\right] + \lambda_{noobj} \sum_{i=1}^{N\times N} (1-\delta_i^{obj}) \sum_{j=1}^{B} (C_i^j)^2 \quad (1)$$

$$L^{obj}(\mathbf{p}_i^j) = \left(1 - C_i^j\right)^2 + \lambda_{ang}\left(\hat{\theta}_i - \theta_i^j\right)^2 + \lambda_{coord} \cdot \left[(\hat{x}_i - x_i^j)^2 + (\hat{y}_i - y_i^j)^2 + (\sqrt{\hat{w}_i} - \sqrt{w_i^j})^2 + (\sqrt{\hat{h}_i} - \sqrt{h_i^j})^2\right]$$

汽车之家 autohome.com.cn

Cascaded Segmentation-Detection Networks for Word-Level Text Spotting

主要思想： FCN+YOLO

一个全卷机网络，得出比较粗糙的但比较稳定的分割；
一个类似 YOLO 的检测框架模型，在上述的基础上，进行区域回归，得到每个单字符的区域；

TABLE II
RESULTS ON THE ICDAR 2013 FOCUSED DATASET

Method	Precision (%)	Recall (%)	F-score (%)
Neumann *et al.* [27]	85	68	75
Yin *et al.* [28]	86	68	76
FASText [29]	84	69	77
Huang *et al.* [11]	88	71	78
Zhang *et al.* [12]	88	74	80
TextFlow [30]	85	76	80
He *et al.* [31]	93	73	82
Qin *et al.* [13]	88	77	82
Zhang *et al.* [6]	88	78	83
Gupta *et al.* [3]	92	76	83
Yao *et al.* [7]	88	80	84
TextBoxes [32]	88	**83**	85
Zhu *et al.* [33]	**93**	81	87
CTPN [4]	**93**	**83**	**88**
Proposed	90	**83**	86

TABLE I
RESULTS ON THE ICDAR 2015 INCIDENTAL DATASET

Method	Precision (%)	Recall (%)	F-score (%)
HUST [19]	44	38	41
AJON [19]	47	47	47
NJU-Text [19]	70	36	47
StradVision [19]	53	46	50
Zhang [6]	71	43	54
Google Vision API [1]	68	53	59
CTPN [4]	74	52	61
Megvii-image++ [7]	72	58	64
Proposed (Seg+Det)	**79**	**65**	**71**
Proposed (Det only)	61	40	48

汽车之家
autohome.com.cn

R2CNN:Rotational Region CNN for Orientation Robust Scene Text Detection

主要思想： Faster RCNN

1. 利用 RPN(Region Proposal Network) 产生 bounding box ，其中包含不同方向的文本；
2. 对上述产生的 proposal 提取特征：利用不同的 pooled size 和级联特征，来预测该区域是否为样本， box 信息，最小外接区域 box;
3. 利用倾斜 NMS 进行处理；

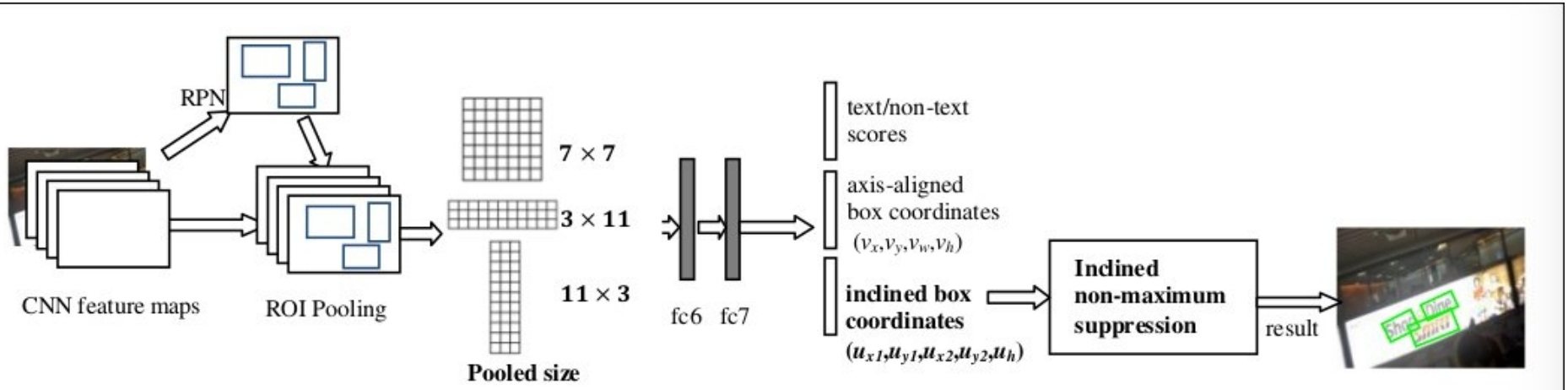


Fig.2. The network architecture of Rotational Region CNN (R^2CNN). The RPN is used for proposing axis-aligned bounding boxes that enclose the arbitrary-oriented texts. For each box generated by RPN, three ROIPoolings with different pooled sizes are performed and the pooled features are concatenated for predicting the text scores, axis-aligned box (v_x,v_y,v_w,v_h) and inclined minimum area box $(u_{x1},u_{y1},u_{x2},u_{y2},u_h)$. Then an inclined non-maximum suppression is conducted on the inclined boxes to get the final result.

汽车之家 autohome.com.cn

R2CNN:Rotational Region CNN for Orientation Robust Scene Text Detection

主要贡献：

1. 提出了一种基于 Faster-RCNN 的针对任意方向的文本检测框架；

2. 将任意方向的文本检测视为一种多任务问题。包括：text score, axis-aligned boxes, inclined minimum area boxes;

3. 在特征提取过程中，采用不同的 pool size(7x7, 11x3, 3x11):

对每一个 RPN Proposal 区域，采用 7x7 进行 pooling 。但由于文本的长度比宽度大很多的特点，增加了两个不同的尺寸 size:

a)11x3: 针对竖直方向的文本，其高度比宽度大很多的情况，提取特征，进行检测

b)3x11: 针对水平方向的文本，其宽度比高度大很多的情况，特区特征，进行检测

4. 增加了小区域的 anchor:

原始的 anchor scales 是 (8,16,32) ，本文引入了两个策略：

a) 直接采用缩小的 anchors (4,8,16) 进行预测

b) 在原有基础上，增加新的小 anchor ， 采用 (4,8,16,32) 进行预测

5. 利用倾斜非极大值抑制方法进行后处理 (incliend non-maximum suppression)

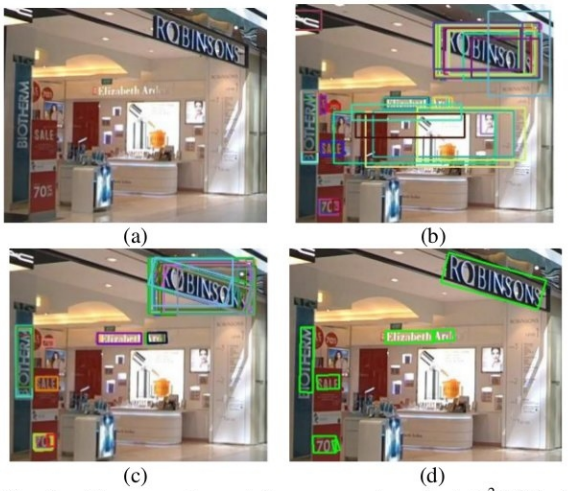
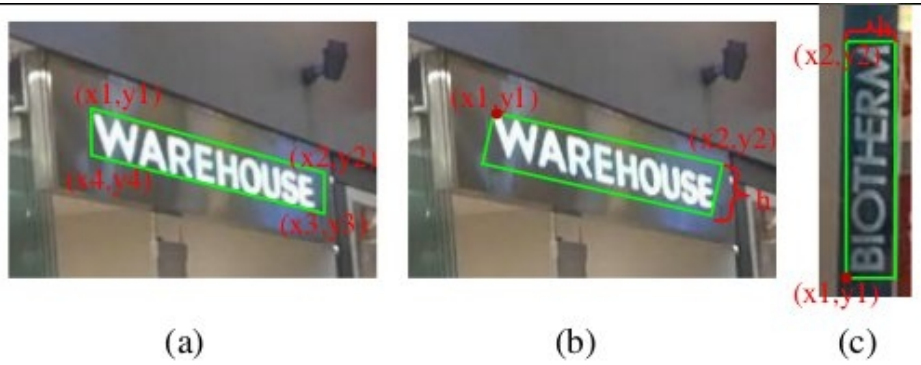


Fig. 1. The procedure of the proposed method R^2CNN. (a) Original input image; (b) text regions (axis-aligned bounding boxes) generated by RPN; (c) predicted axis-aligned boxes and inclined minimum area boxes (each inclined box is associated with an axis-aligned box, and the associated box pair is indicated by the same color); (d) detection result after inclined non-maximum suppression.

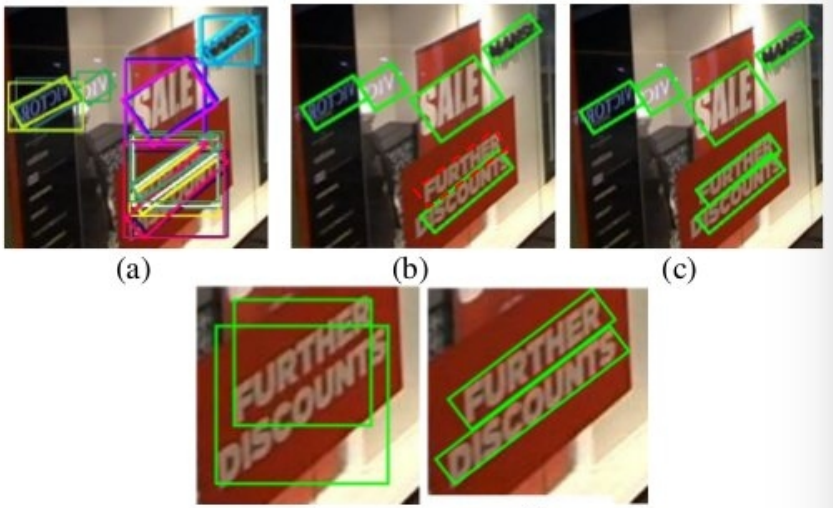
汽车之家 autohome.com.cn

R2CNN:Rotational Region CNN for Orientation Robust Scene Text Detection



(a) (b) (c)

Fig.3. Detection targets of arbitrary-oriented scene text detection. a) ICDAR 2015 labels the incidental scene texts in the form of four points in clockwise; b) the inclined minimum area rectangle is adopted as the detection targets in our approach; c) another example of the inclined rectangle.



(a) (b) (c) (d) (e)

Fig.4. Inclined non-maximum suppression. (a) The candidate axis-aligned boxes and inclined boxes; (b) the detection results based on normal NMS on axis-aligned boxes (the green boxes are the correct detections, and the red dashed box is the box that is not detected); (c) the detection results based on inclined NMS on inclined boxes; (d) an example of two axis-aligned boxes; (e) an example of two inclined boxes.

汽车之家
autohome.com.cn

R2CNN:Rotational Region CNN for Orientation Robust Scene Text Detection

Loss:

$$L(p,t,v,v^*,u,u^*) = L_{\text{cls}}(p,t) + \lambda_1 t \sum_{i\in\{x,y,w,h\}} L_{reg}(v_i,v_i^*) + \lambda_2 t \sum_{i\in\{x1,y1,x2,y2,h\}} L_{reg}(u_i,u_i^*) \quad (1)$$

λ_1 and λ_2 are the balancing parameters that control the trade-off between three terms.

Let (w,w^*) indicates (v_i,v_i^*) or (u_i,u_i^*), $L_{reg}(w,w^*)$ is defined as:

$$L_{reg}(w,w^*) = \text{smooth}_{L1}(w-w^*) \quad (2)$$

$$\text{smooth}_{L1}(x) = \begin{cases} 0.5x^2 & \text{if } |x|<1 \\ |x|-0.5 & \text{otherwise} \end{cases} \quad (3)$$

汽车之家 autohome.com.cn

R2CNN:Rotational Region CNN for Orientation Robust Scene Text Detection

Table 2. Comparison with state-of-the-art on ICDAR2015.

Approaches	Recall	Precision	F-measure
R^{2}CNN	**79.68 %**	**85.62 %**	**82.54 %**
Deep direct regression[34]	80.00%	82.00%	81.00%
EAST[32]	78.33%	83.27%	80.72%
RRPN[15]	82.17%	73.23%	77.44%
SegLink[31]	76.80%	73.10%	75.00%
DMPNet[33]	68.22%	73.23%	70.64%
CTPN[9]	51.56%	74.22%	60.85%
MCLAB_FCN[16]	43.09%	70.81%	53.58%

Table 3. Comparison with state-of-the-art on ICDAR2013.

Approaches	Recall	Precision	F-measure
CTPN [9]	83.00%	93.00%	88.00%
R^{2}CNN	**82.59%**	**93.55%**	**87.73%**
Deep direct regression[34]	81.00%	92.00%	86.00%
SegLink[31]	83.00%	87.70%	85.30%
TextBoxes [8]	82.59%	87.73%	85.08%
DeepText [7]	82.79%	87.17%	84.93%
R^{2}CNN-720	**79.73%**	**86.90%**	**83.16%**
MCLAB_FCN[16]	77.81%	88.14%	82.65%
TextFlow[10]	75.89%	85.15%	80.25%
RRPN [15]	71.89%	90.22%	80.02%
Faster R-CNN	74.52%	82.83%	78.45%

汽车之家 autohome.com.cn

EAST: An Efficient and Accurate Scene Text Detector

主要思想： FCN+NMS

本文主要采用全卷机网络直接处理字符或文本层面的预测，取出了之前很多方法中的中间过程。该方法简单有效，可以有效的链接 loss 和网络结构。并且能处理旋转的矩形框或者矩形框。

主要贡献：

1. 提出了一种文本检测方法，包括： FCN+NMS
2. 该方法可以变化，提供字符方面或者文本方面的检测，且对集合形状没有要求，可以是旋转的矩形框
3. 该方法在精度和速度上都达到了非常好的效果

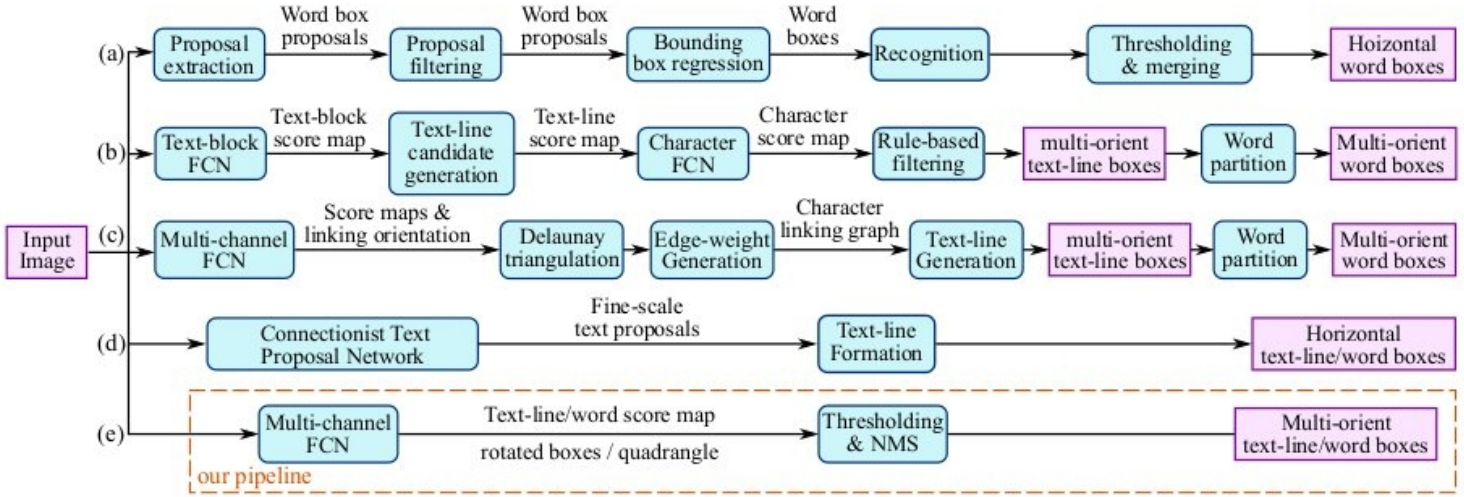


Figure 2. Comparison of pipelines of several recent works on scene text detection: (a) Horizontal word detection and recognition pipeline proposed by Jaderberg *et al.* [12]; (b) Multi-orient text detection pipeline proposed by Zhang *et al.* [48]; (c) Multi-orient text detection pipeline proposed by Yao *et al.* [41]; (d) Horizontal text detection using CTPN, proposed by Tian *et al.* [34]; (e) Our pipeline, which eliminates most intermediate steps, consists of only two stages and is much simpler than previous solutions.

汽车之家 autohome.com.cn

EAST: An Efficient and Accurate Scene Text Detector

网络结构

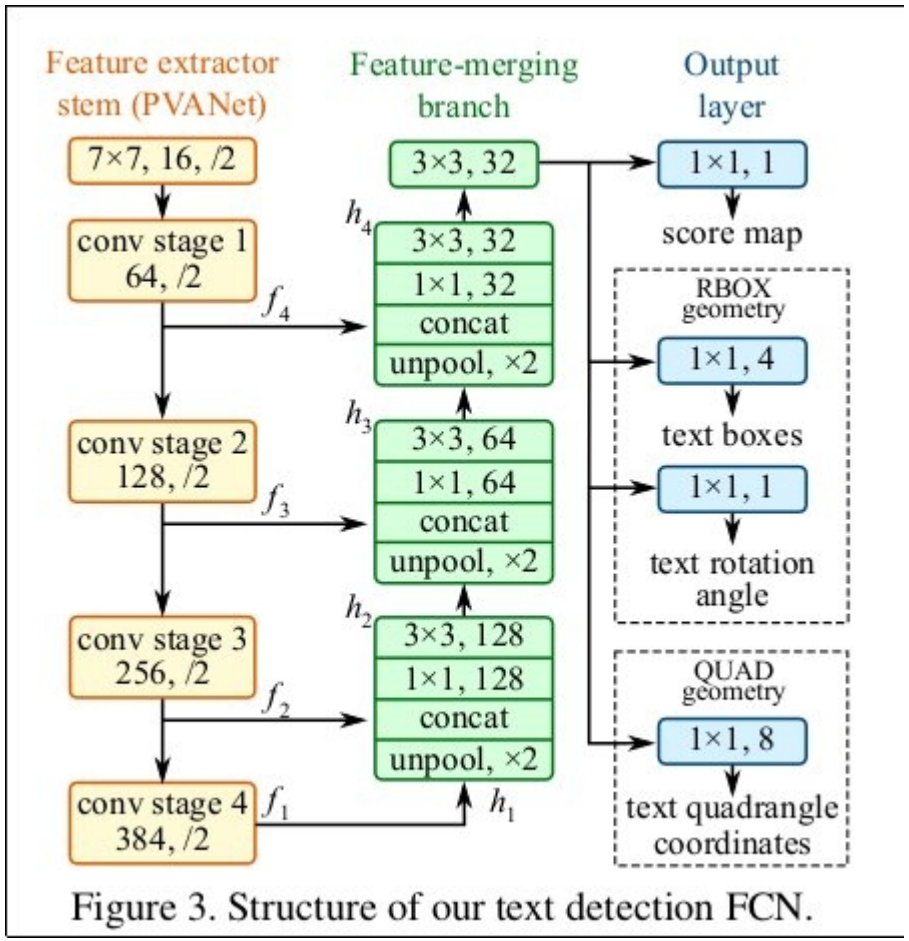


Figure 3. Structure of our text detection FCN.

汽车之家 autohome.com.cn

EAST: An Efficient and Accurate Scene Text Detector

Loss

The loss can be formulated as

$$L = L_s + \lambda_g L_g \tag{4}$$

where L_s and L_g represents the losses for the score map and the geometry, respectively, and λ_g weighs the importance between two losses. In our experiment, we set λ_g to 1.

$$\begin{aligned} L_s &= \text{balanced-xent}(\hat{\mathbf{Y}}, \mathbf{Y}^*) \\ &= -\beta \mathbf{Y}^* \log \hat{\mathbf{Y}} - (1-\beta)(1-\mathbf{Y}^*) \log(1-\hat{\mathbf{Y}}) \end{aligned} \tag{5}$$

where $\hat{\mathbf{Y}} = F_s$ is the prediction of the score map, and $\mathbf{Y}^*$ is the ground truth. The parameter β is the balancing factor between positive and negative samples, given by

$$\beta = 1 - \frac{\sum_{y^* \in \mathbf{Y}^*} y^*}{|\mathbf{Y}^*|}. \tag{6}$$

RBOX For the AABB part, we adopt IoU loss in [46], since it is invariant against objects of different scales.

$$L_{\text{AABB}} = -\log \text{IoU}(\hat{\mathbf{R}}, \mathbf{R}^*) = -\log \frac{|\hat{\mathbf{R}} \cap \mathbf{R}^*|}{|\hat{\mathbf{R}} \cup \mathbf{R}^*|} \tag{7}$$

where $\hat{\mathbf{R}}$ represents the predicted AABB geometry and $\mathbf{R}^*$ is its corresponding ground truth. It is easy to see that the width and height of the intersected rectangle $|\hat{\mathbf{R}} \cap \mathbf{R}^*|$ are

$$\begin{aligned} w_{\mathbf{i}} &= \min(\hat{d}_2, d_2^*) + \min(\hat{d}_4, d_4^*) \\ h_{\mathbf{i}} &= \min(\hat{d}_1, d_1^*) + \min(\hat{d}_3, d_3^*) \end{aligned} \tag{8}$$

where d_1, d_2, d_3 and d_4 represents the distance from a pixel to the top, right, bottom and left boundary of its corresponding rectangle, respectively. The union area is given by

$$|\hat{\mathbf{R}} \cup \mathbf{R}^*| = |\hat{\mathbf{R}}| + |\mathbf{R}^*| - |\hat{\mathbf{R}} \cap \mathbf{R}^*|. \tag{9}$$

Therefore, both the intersection/union area can be computed easily. Next, the loss of rotation angle is computed as

$$L_\theta(\hat{\theta}, \theta^*) = 1 - \cos(\hat{\theta} - \theta^*). \tag{10}$$

where $\hat{\theta}$ is the prediction to the rotation angle and θ^* represents the ground truth. Finally, the overall geometry loss is the weighted sum of AABB loss and angle loss, given by

$$L_g = L_{\text{AABB}} + \lambda_\theta L_\theta. \tag{11}$$

Where λ_θ is set to 10 in our experiments.

Note that we compute L_{AABB} regardless of rotation angle. This can be seen as an approximation of quadrangle IoU when the angle is perfectly predicted. Although it is not the case during training, it could still impose the correct gradient for the network to learn to predict $\hat{\mathbf{R}}$.

汽车之家 autohome.com.cn

EAST: An Efficient and Accurate Scene Text Detector

结果

Algorithm	Recall	Precision	F-score
Ours + PVANET2x RBOX MS*	**0.7833**	0.8327	**0.8072**
Ours + PVANET2x RBOX	0.7347	**0.8357**	**0.7820**
Ours + PVANET2x QUAD	0.7419	0.8018	0.7707
Ours + VGG16 RBOX	0.7275	0.8046	0.7641
Ours + PVANET RBOX	0.7135	0.8063	0.7571
Ours + PVANET QUAD	0.6856	0.8119	0.7434
Ours + VGG16 QUAD	0.6895	0.7987	0.7401
Yao *et al.* [41]	0.5869	0.7226	0.6477
Tian *et al.* [34]	0.5156	0.7422	0.6085
Zhang *et al.* [48]	0.4309	0.7081	0.5358
StradVision2 [15]	0.3674	0.7746	0.4984
StradVision1 [15]	0.4627	0.5339	0.4957
NJU [15]	0.3625	0.7044	0.4787
AJOU [20]	0.4694	0.4726	0.4710
Deep2Text-MO [45] [44]	0.3211	0.4959	0.3898
CNN MSER [15]	0.3442	0.3471	0.3457

Table 3. Results on ICDAR 2015 Challenge 4 Incidental Scene Text Localization task. MS means multi-scale testing.

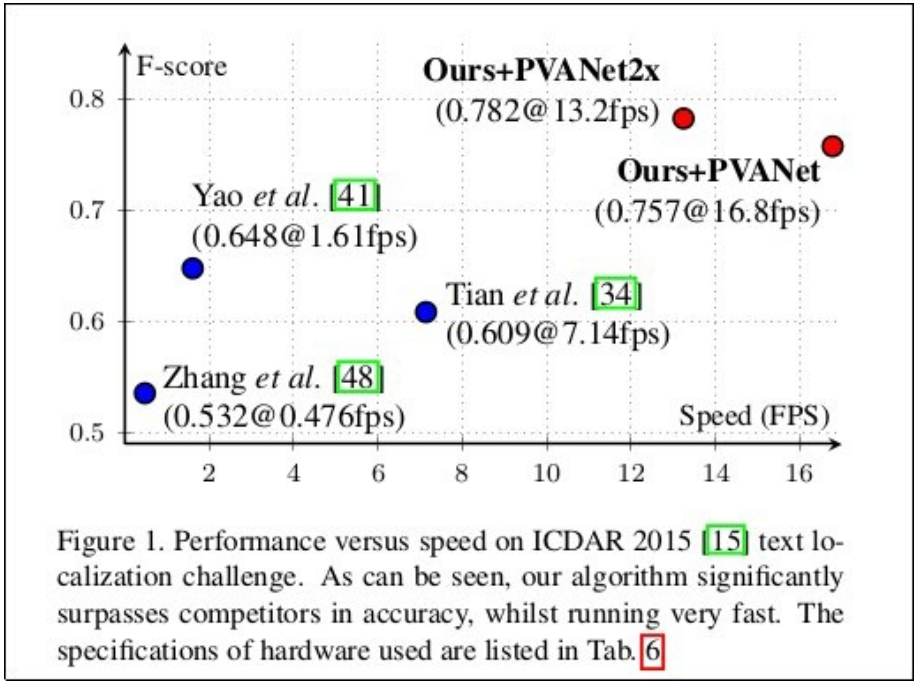


Figure 1. Performance versus speed on ICDAR 2015 [15] text localization challenge. As can be seen, our algorithm significantly surpasses competitors in accuracy, whilst running very fast. The specifications of hardware used are listed in Tab. 6

汽车之家
autohome.com.cn