Text Detection with Selected Anchors

Anna Zhu, Hang Du, Shengwu Xiong

Wuhan University of Technology

annakkk@live.com

Outline

1. Task
2. Motivation
3. Method
4. Experiment

Task

Scene text detection and recognition can be widely used in:

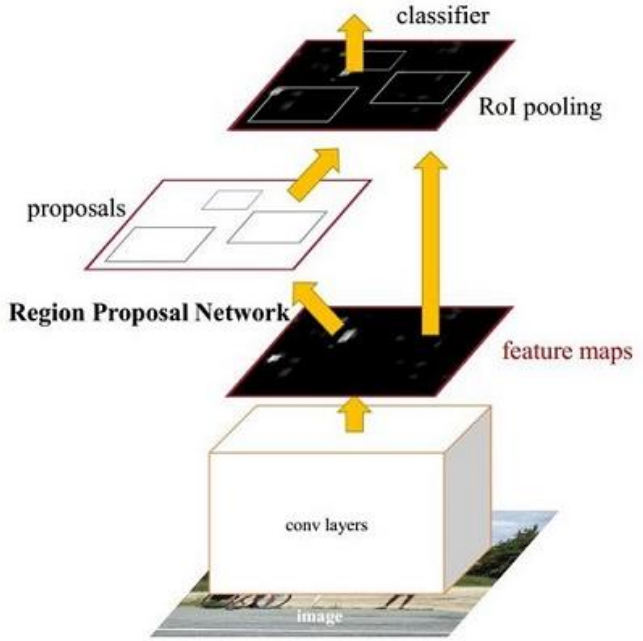
- License plate recognition
- Automatic translation
- Robot navigation
- Information extraction

(a) Image without text annotation

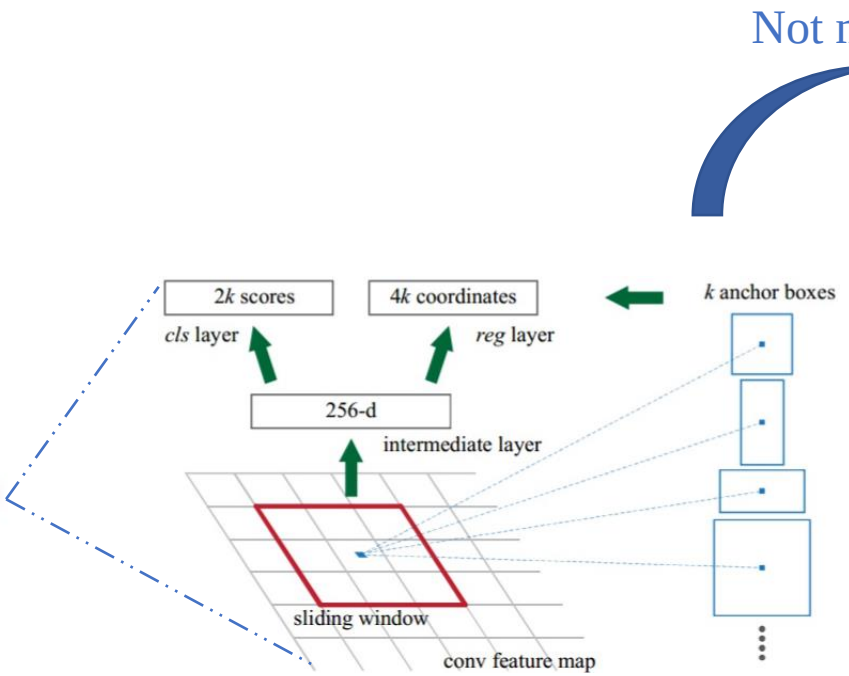
(b) Image with text annotation

(c) Image with text detection

Motivation



(d) Architecture of Faster RCNN

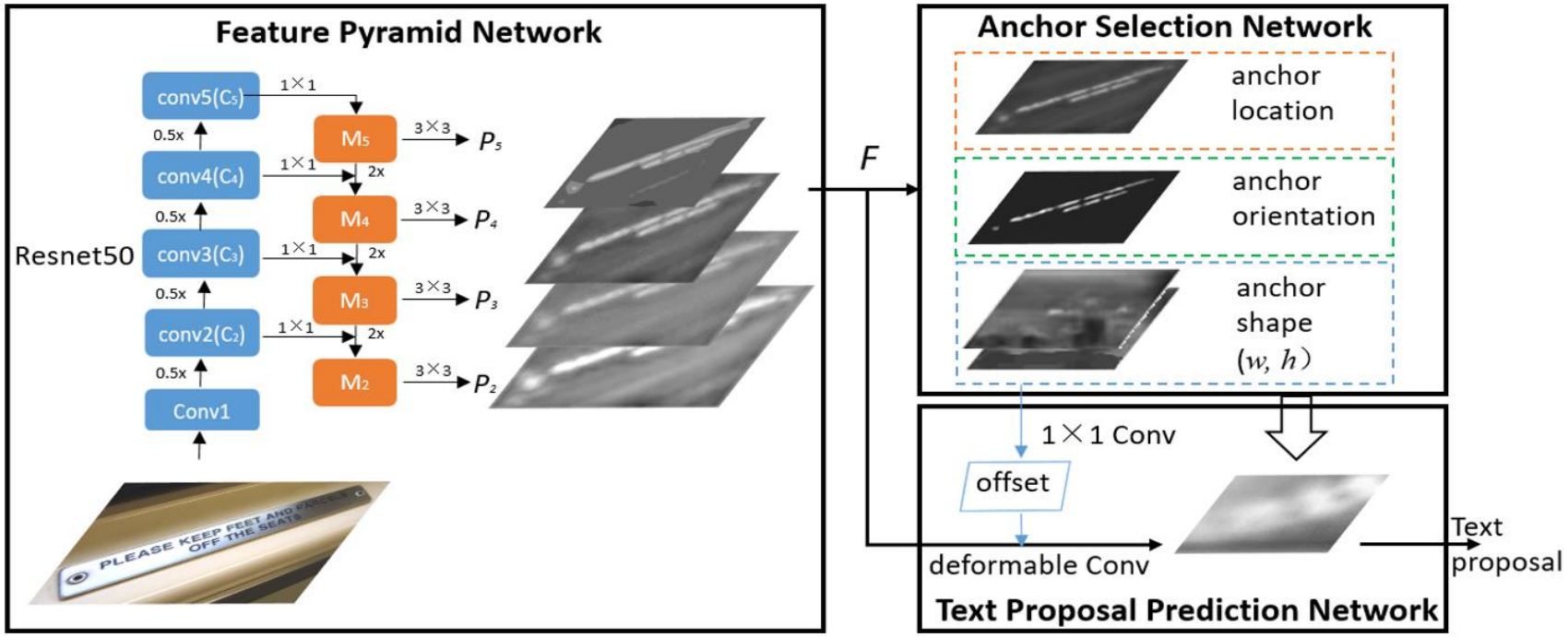


(e) Predefined dense anchors

(f) Instance with different scales and orientations

Can we lead to a higher recall and reduce numbers of anchors by using learnable anchors instead of fixing them ?

Method



(1) Schematic framework of AS-RPN

- Feature Pyramid Network , FPN
- Anchor Selection Network , AS-RPN
- Text Prediction Network

Method

Optimization function

$$\mathrm{L} = L_{conf} + L_{reg} + \alpha L_{loc} + \beta L_{angle} + \lambda L_{shape}$$

- Location loss

$$L_{loc} = \begin{cases} -\alpha(1-y')^{\gamma}\log y', y = 1 \\ -(1-\alpha)y'^{\gamma}\log(1-y'), y = 0 \end{cases}$$

Where y' is the output of the anchor location branch with a sigmoid function

- Angle loss

$$L_{angle} = 1 - \cos(\hat{\theta} - \theta_g)$$

Where $\hat{\theta}$ is the prediction of the orientation branch and θ_g is the angle target.

- Shape loss

$$L_{shape} = L_1\left(1 - min\left(\frac{w}{w_g}, \frac{w_g}{w}\right)\right) + L_1\left(1 - min\left(\frac{h}{h_g}, \frac{hg}{h}\right)\right)$$

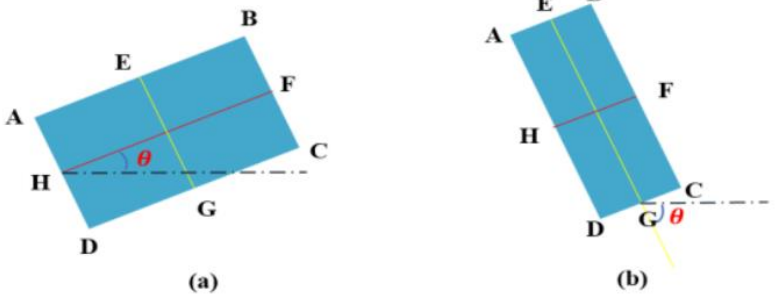
Where (w, h) denote the predicted anchor shape

α, β, λ are parameters to balance the location, orientation and shape prediction branches which are set to $\alpha = \beta = 1$; $\lambda = 0.1$;

$*_g$ donates the *(angle and shape) target

Method

- Label generation



- Angle normalization

$$\theta_t = \frac{\theta_g}{\pi} + \frac{1}{2}$$

Algorithm1 angle label generation

Input: original gt $O(x_1, y_1, \ldots, x_4, y_4)$, $A(x_1, y_1)$, $B(x_2, y_2)$, $C(x_3, y_3)$, $D(x_4, y_4)$ as shown in Fig.2

Output: output gt $F(x, y, w, h, \theta)$

for each line in O, do

  $(x, y) = (\frac{x_1+\cdots+x_4}{4}, \frac{y_1+\cdots+y_4}{4})$

  $w = \max(AB, AD), h = min(AB, AD)$

  Calculate the middle point{E, F,G,H} of AB, BC, CD and DE

  $\theta_1 = \arctan k_{EG}, \theta_2 = \arctan k_{HF}$

  if len (EG) > len (HF)

    $F = (x, y, w, h, \theta_1)$

    else $F = (x, y, w, h, \theta_2)$

  end if

end for

Experiments

Measure \ Method		RPN	FPN-RPN	AF-RPN	AS-RPN
IoU_0.5	TR_{50}	67.2	67.5	73.3	**74.5**
	TR_{100}	76.9	77.2	81.8	**82.9**
	TR_{300}	86.6	87.4	**89.3**	88.6
IoU_0.75	TR_{50}	22.8	28.8	35.0	**36.2**
	TR_{100}	27.9	36.0	41.3	**44.6**
	TR_{300}	33.8	47.2	48.2	**48.8**
IoU_Avg	TR_{50}	30.6	33.5	38.2	**38.8**
	TR_{100}	35.9	39.8	43.6	**44.9**
	TR_{300}	41.7	48.0	49.2	**50.0**

REGION PROPOSAL QUALITY EVALUATION ON COCO-TEXT VALIDATION SET (%)

Fig. 4. Examples of FPN-RPN text proposals (top row) and AS-RPN text proposals(bottom row).

Approach	P (%)	R (%)	F (%)
CPTN[35]	74.22	51.56	60.85
Seg Link[27]	74.74	76.50	75.61
SSTD[41]	80.23	73.86	76.91
RRPN[26]	82.02	73.00	77.05
EAST*[38]	84.36	81.27	82.79
R2CNN[42]	85.62	79.68	82.54
Text boxes++[25]	87.80	78.50	82.90
Ours	**83.34**	**79.99**	**81.63**

DETECTION RESULTS COMPARE WITH RELAVENT APPROACHES ON ICDAR 2015

Approach	P (%)	R (%)	F (%)
Baseline	57.40	54.50	55.90
He et al[37]	76.40	61.42	68.76
EAST*[38]	81.23	63.27	75.54
RRPN[26]	68.00	82.00	74.00
TextSnake[39]	83.20	73.90	78.30
Pixel Link[17]	83.00	73.20	77.82
Lyu et al[28]	87.60	76.20	81.50
Ours	**84.67**	**80.37**	**82.49**

DETECTION RESULTS COMPARE WITH RELAVENT APPROACHES ON MSRA-TD500

Experiments

(a) Detection results in ICDAR2013

(b) Detection results in ICDAR2015

(d) Detection results in MSRA-TD500

Thank you!