Supplementary of Mask TextSpotter v3

Minghui Liao[1], Guan Pang[2], Jing Huang[2], Tal Hassner[2], and Xiang Bai[1]

[1] Huazhong University of Science and Technology, China
{mhliao,xbai}@hust.edu.cn
[2] Facebook, USA
{gpang,jinghuang,thassner}@fb.com

1 Methodology details

Table 1. Illustration of our SPN segmentation prediction module. "Conv": convolution operator; "BN": batch normalization; "DeConv": de-convolution operator. "k", "s", and "p" are short for kernel size, stride, and padding respectively.

Tpye	Configurations	Input/output channel
Conv	k: 3; s: 1; p: 1	256/64
BN	momentum: 0.1	64/64
ReLU	-	64/64
DeConv	k: 2, s: 2, p: 0	64/64
BN	momentum: 0.1	64/64
ReLU	-	64/64
DeConv	k: 2, s: 2, p: 0	64/1
Sigmoid	-	1/1

2 Rotation robustness

More qualitative amd quantitative results on the Rotated ICDAR 2013 dataset are shown in Fig. 1, Tab. 2, and Tab. 3.

3 Ablation study

There are two attributions for the RoI masking operator: "direct/indirect" and "soft/hard". "direct/indirect" means using the segmentation/binary map directly or through additional layers; "soft/hard" indicates a soft probability mask map whose values are from $[0, 1]$ or a binary polygon mask map whose values are 0 or 1. We conduct experiments with the following settings:
(1) Baseline: Using the original RoI feature. **(2) Direct-soft**: It is similar to the RoI masking proposed in Qin et al. [1], applying element-wise multiplication

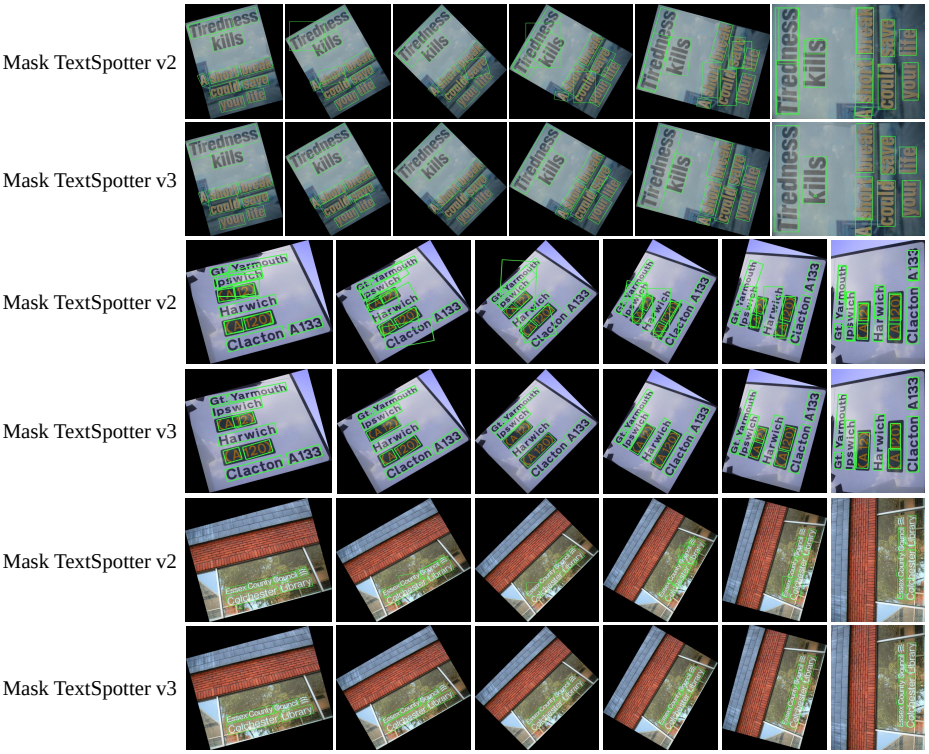

Fig. 1. Qualitative results on the Rotated ICDAR 2013 dataset. The rotating angles are 15°, 30°, 45°, 60° , 75°, and 90° for the columns from left to right.

between the corresponding segmentation probability map and the RoI feature. **(3) Direct-hard**: Our proposed hard RoI masking, applying element-wise multiplication between the corresponding binary polygon mask map and the RoI feature. **(4) Indirect-soft**: The corresponding segmentation probability map and the RoI feature are concatenated and then a mask prediction module consisting of two convolutional layers is applied to predict a new mask map. Element-wise multiplication is then applied between the new mask map and RoI feature. **(5) Indirect-hard**: First, a masked RoI feature is obtained by the hard RoI masking. Then, we concatenate the masked RoI feature and the original RoI feature. Finally, the concatenated feature is classified, choosing whether the masked RoI feature or the original RoI feature is used as the output feature.

The experimental results in Tab. 4 show that "direct" is better than "indirect" and "hard" is better than "soft". The reason is the "direct" and "hard" strategies provide the most strict mask, fully blocking background noise and neighboring text instances. Our proposed hard RoI masking is simple yet achieves the best performance.

Table 2. Quantitative detection results on the Rotated ICDAR 2013 dataset. The evaluation protocol is the same as the one in ICDAR 2015 dataset. *CharNet is tested with the official released pre-trained model; Mask TextSpotter v2 is trained with the same rotating augmentation as Mask TextSpotter v3. "RA" is short for rotating angles. "P", "R", and "F" indicate precision, recall and F-measure respectively.

RA (°)	CharNet			Mask TextSpotter v2			Mask TextSpotter v3		
	P	R	F	P	R	F	P	R	F
0	82.3	81.7	82	89.9	85.3	87.5	90.5	84.4	87.4
15	88.1	82.2	85.1	84.6	77.4	80.7	91.8	82.3	86.8
30	85.7	79.4	82.5	75.2	66	70.3	91.3	78.9	84.6
45	57.8	56.6	57.2	64.8	59.9	62.2	91.6	77.9	84.2
60	65.5	53.3	58.8	70.5	61.2	65.5	90.7	79.4	84.7
75	58.4	41.1	48.3	77.1	77.7	77.4	89.3	80.8	84.8
90	63.0	40.4	49.2	89.8	76.8	82.8	89.8	77.2	83.0

Table 3. Quantitative end-to-end recognition results (without lexicon) on the Rotated ICDAR 2013 dataset. The evaluation protocol is the same as the one in ICDAR 2015 dataset. *CharNet is tested with the official released pre-trained model; Mask TextSpotter v2 is trained with the same rotating augmentation as Mask TextSpotter v3. "RA" is short for rotating angles. "P", "R", and "F" indicate precision, recall and F-measure respectively.

RA (°)	CharNet			Mask TextSpotter v2			Mask TextSpotter v3		
	P	R	F	P	R	F	P	R	F
0	61.7	61.2	61.4	86.3	75.2	80.3	89.0	73	80.2
15	66.3	61.9	64	78.4	53.5	63.6	87.2	69.8	77.5
30	60.9	56.5	58.6	73.9	54.7	62.9	87.8	67.5	76.3
45	34.2	33.5	33.9	66.4	45.8	54.2	88.5	66.8	76.1
60	10.3	8.4	9.3	68.2	48.3	56.6	88.5	67.6	76.6
75	0.3	0.2	0.2	77.0	59.2	67.0	86.9	67.6	76.0
90	0.0	0.0	0.0	82.0	56.9	67.1	85.9	57.9	69.1

Table 4. Ablation study on the hard RoI masking. "Direct-hard" indicates our proposed hard RoI masking.

Method	Total-Text		IC15
	None	Full	Strong
Baseline	67.3	76.2	81.0
Direct-soft	69.1	76.0	82.2
Direct-hard	**71.2**	**78.4**	**83.3**
Indirect-soft	65.8	75.6	81.2
Indirect-hard	68.4	76.2	81.4

References

1. Qin, S., Bissacco, A., Raptis, M., Fujii, Y., Xiao, Y.: Towards unconstrained end-to-end text spotting. In: Proc. Int. Conf. Comput. Vision (2019)