Reading Arbitrary-Shaped Scene Text from Images Through Spline Regression and Rectification

— Supplementary Material

Long Chen, Feng Su, Jiahao Shi, and Ye Qian

Nanjing University

1 Network Configuration

Tables 1, 2, and 3 show the configuration of the text feature rectification module, the text recognition network, and the text region regression network, respectively. In the text region regression network, the ResNet [2] and FPN [3] backbones adopt the configurations in original literatures. The Cascade R-CNN [1] module comprises three stages, using IoU thresholds $\{0.5, 0.55, 0.6\}$ and loss weights $\{1, 0.5, 0.25\}$ in each stage respectively. As all the stages of the Cascade R-CNN have the same network configuration, only the configuration of one stage is shown. Code is available at: https://github.com/ChenLLong/SPRNet.

Table 1. Configuration of the text feature rectification module. The output is the predicted parameters for feature deformation. 'maps', 'k', and 's' denote the number of filters, kernel size, and stride.

Layer	Configuration	Output Size
Input	-	$16 \times 16 \times 256$
Conv Layer	maps:256, k:1×1, s:1×1	$16 \times 16 \times 256$
Conv Layer	maps:256, k:3×3, s:1×1	$16 \times 16 \times 256$
Conv Layer	maps:256, k:3×3, s:1×1	$16 \times 16 \times 256$
Max Pool	maps:256, k:2×2, s:2×2	$8 \times 8 \times 256$
Conv Layer	maps:512, k:3×3, s:1×1	$8 \times 8 \times 512$
Max Pool	maps:512, k:2×2, s:2×2	$4 \times 4 \times 512$
Reshape & Concat.	-	8192
FC	hidden units: 256	256
FC	hidden units: 14	14

2 Limitation

Figure 1 shows some examples of the failure cases of the proposed text spotting method. Most of the detection errors were caused by low/uneven illumination,

Table 2. Configuration of the text recognition network. 'maps', 'k', and 's' denote the number of filters, kernel size, and stride.

Layer	Configuration	Output Size
Input	-	$256 \times 8 \times 32$
Conv Layer $\times 2$	$maps: 256,\ k: 3,\ s: 1$	$256 \times 8 \times 32$
Max Pool	$k: (2,1),\ s: (2,1)$	$256 \times 4 \times 32$
Conv Layer $\times 2$	$maps: 256,\ k: 3,\ s: 1$	$256 \times 4 \times 32$
Max Pool	$k: (2,1),\ s: (2,1)$	$256 \times 2 \times 32$
Conv Layer $\times 2$	$maps: 256,\ k: 3,\ s: 1$	$256 \times 2 \times 32$
Avg Pool	$k: (2,1),\ s: (2,1)$	$256 \times 1 \times 32$
BiLSTM	hidden units: 256	$256 \times 1 \times 32$
Att. GRU	hidden units: 256	-

blurring, interfering object, or ambiguous orientation of text. On the other hand, as a light-weight text recognition module is employed in the text spotting network, it failed to recognize text in some challenging cases such as obscure text with stuck-together characters and upside-down ambiguous character sequences in some ring-shaped or mirrored text. To improve the performance of the proposed text spotting network in these challenging cases, an enhanced detection backbone and data augmentation measures can be employed, and some effective language models can be introduced to improve the accuracy of the text recognition network.

(a) Detection errors

(b) Recognition errors

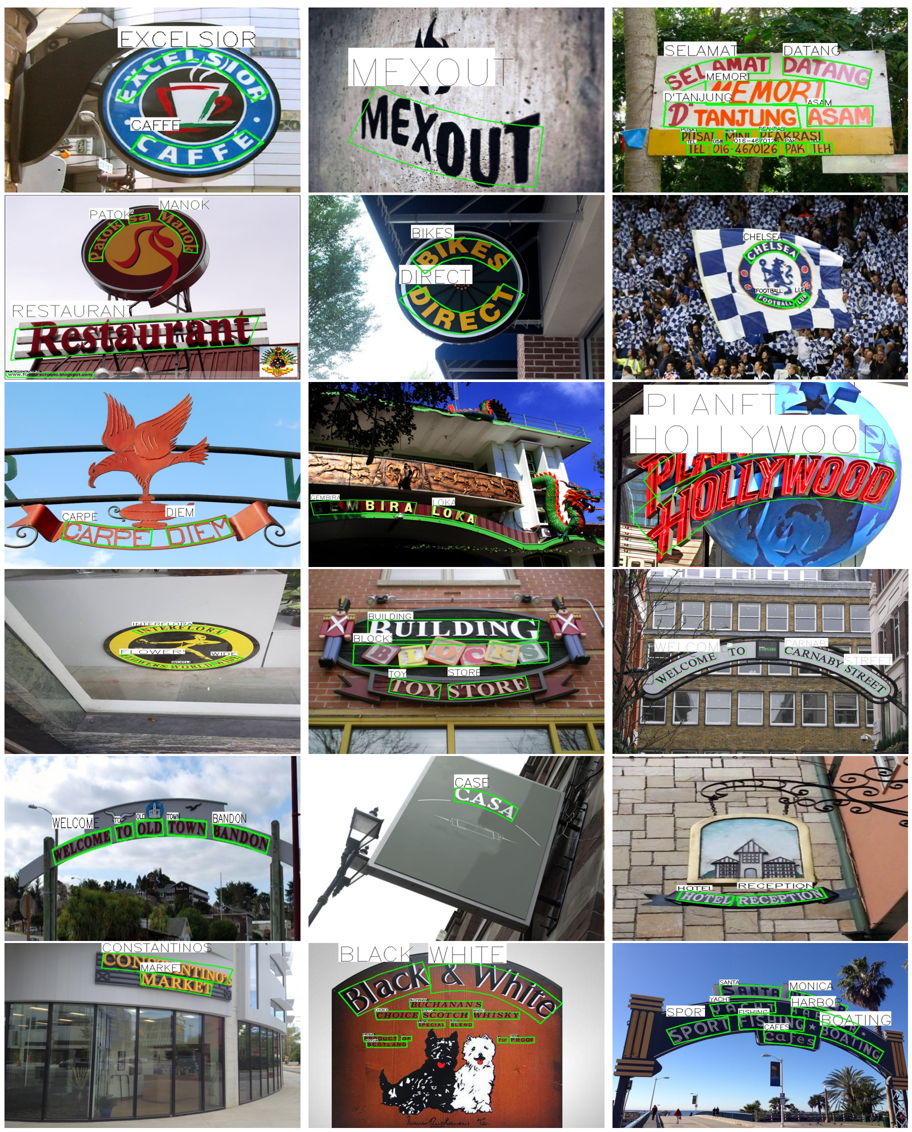
Fig. 1. Examples of failure cases of the proposed text spotting method. Green boxes indicate detected text instances with corresponding recognition results shown nearby. Blue boxes indicate ground-truth text instances. Red boxes indicate text instances that are not considered by the performance evaluation protocol.

Table 3. Configuration of the text region regression network. FC denotes fully connected layer. 'maps', 'k', and 's' denote the number of filters, kernel size, and stride respectively. w and h denote the width and height of the input image.

Module		Layer	Configuration	Output Size
ResNet & FPN	Same to [2] and [3]			
RPN	level 2	Input	-	$\frac{w}{4} \times \frac{h}{4} \times 256$
		Conv (class)	maps:3, k:1 × 1, s:1 × 1	$\frac{w}{4} \times \frac{h}{4} \times 3$
		Conv (box)	maps:12, k:1 × 1, s:1 × 1	$\frac{w}{4} \times \frac{h}{4} \times 12$
	level 3	Input	-	$\frac{w}{8} \times \frac{h}{8} \times 256$
		Conv (class)	maps:3, k:1 × 1, s:1 × 1	$\frac{w}{8} \times \frac{h}{8} \times 3$
		Conv (box)	maps:12, k:1 × 1, s:1 × 1	$\frac{w}{8} \times \frac{h}{8} \times 12$
	level 4	Input	-	$\frac{w}{16} \times \frac{h}{16} \times 256$
		Conv (class)	maps:3, k:1 × 1, s:1 × 1	$\frac{w}{16} \times \frac{h}{16} \times 3$
		Conv (box)	maps:12, k:1 × 1, s:1 × 1	$\frac{w}{16} \times \frac{h}{16} \times 12$
	level 5	Input	-	$\frac{w}{32} \times \frac{h}{32} \times 256$
		Conv (class)	maps:3, k:1 × 1, s:1 × 1	$\frac{w}{32} \times \frac{h}{32} \times 3$
		Conv (box)	maps:12, k:1 × 1, s:1 × 1	$\frac{w}{32} \times \frac{h}{32} \times 12$
RoIAlign				7 × 7 × 256
Cascade R-CNN		AvgPool	maps:256, k:7 × 7, s:1 × 1	1 × 1 × 256
		Reshape	-	256
		FC	hidden units: 1024	1024
		FC	hidden units: 1024	1024
	class branch	FC	hidden units: 2	2
	box branch	FC	hidden units: 8	8
RoIAlign				16 × 16 × 256
Text Region Regression	text direction branch	MaxPool	maps:256, k:16 × 16, s:1 × 1	1 × 1 × 256
		Reshape	-	256
		FC	hidden units: 2	2
	shape parameter branches	Conv	maps:256, k:1 × 1, s:1 × 1	16 × 16 × 256
		Conv	maps:256, k:3 × 3, s:1 × 1	16 × 16 × 256
		Conv	maps:256, k:3 × 3, s:1 × 1	16 × 16 × 256
		MaxPool	maps:256, k:2 × 2, s:2 × 2	8 × 8 × 256
		Conv	maps:512, k:3 × 3, s:1 × 1	8 × 8 × 512
		MaxPool	maps:512, k:2 × 2, s:2 × 2	4 × 4 × 512
		Reshape & Concat.	-	8192
		FC	hidden units: 512 (centerline), 512 (boundary)	512, 512
		FC	hidden units: 27 (centerline), 72 (boundary)	27, 72

3 Scene Text Spotting Results

The following are some examples of scene text spotting results of the proposed method. Text instances detected are marked with green boxes with the recognized text shown nearby. The results demonstrate the proposed method's capability to robustly localize and recognize scene text with various curved shapes, orientations, sizes, and tight spacing with other text instances.




BEST SELECTION
FRESH TAKIMAN MEAT
Sheffield
PIE SHOP
DALEWOOD
TOWNHOMES
MALIBU
27 MILES OF SCENIC BEAUTY
PARLIAMENT
SINGAPORE
PHOTO TOUR
Pryde's Old Westport
Kitchen & Home Accessories
LONDON
BAKERS
Silver Cross
Downstairs Bar
TEA HOUSE
WINE TASTING
GASTRONOMY
INDIAN HARBOR
YACHT CLUB
1889
KOFFEE KUP
FAMILY RESTAURANT
PENDERGRASS
FLEA MARKET
Sea Breeze
MOTEL
YUCA INN
POCO A POCO
LOS CABOS MEXICO
HALFWAY
CLUB
PIZZA
SOUTHERNMOST
HOTEL
IN THE USA
RECEPTION
CASEY
MOVERS
BAKERSFIELD
MOVING TO BAKERSFIELD?
WOODBINE
HOTEL
BODY REPAIR
WELDING POLISH
OVEN
Arby's
ROAST BEEF
Sandwich
IS DELICIOUS


grandarmy
GRAND ARMY
OF THE
REPUBLIC
highway
HIGHWAY
certified
CERTIFIED
ANGUS
BEEF
steves
STEVE'S
prime
PRIME
westside
WESTSIDE
wayout
WAYOUT
club
CLUB
village
VILLAGE
of
greentree
GREENTREE
winton
woods
WINTON WOODS
HARBOR POINT
pavilion
PAVILION
hotelpost
HOTEL POST
JOHN LITTLE
GEOX
ESPRIT
ROBINSONS

References

1. Cai, Z., Vasconcelos, N.: Cascade R-CNN: Delving into high quality object detection. In: CVPR. pp. 6154–6162 (2018)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
3. Lin, T., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. In: CVPR. pp. 936–944 (2017)