Robust Scene Text Recognition Through Adaptive Image Enhancement

– Supplementary Materials

Ye Qian
mf1833053@smail.nju.edu.cn
Yuyang Wang
mf1733066@smail.nju.edu.cn
Feng Su
suf@nju.edu.cn

State Key Laboratory for Novel Software Technology
Nanjing University
Nanjing 210023, China

1 Network Configuration

Table 1 and 2 show the parameter configurations of the spatial rectification network (STN) and the recognition network in the proposed scene text recognition model respectively.

Figure 1 shows the structure and parameter configurations of the MGN and MRN networks, which have the same U-shape structure, in the residual enhancement network (REN).

2 Effects of Text Image Enhancement on Scene Text Recognition

Figure 2 shows some examples of text images from irregular text datasets IC15, SVT-P, and CT80 and their enhanced versions attained by the proposed image enhancement model, along with their corresponding recognition results and the ground truth text.

It can be seen that the original text images, in which the text may be quite obscure or irregular so that sometimes we have to watch carefully to find out what it is, become easier to be recognized after passing through the proposed enhancement model, which effectively enhances the text cues and meanwhile reduces the interference of the background.

Table 1: Parameter configuration of the localization network in the spatial rectification network (STN). FC denotes fully connected layer. 'maps', 'k', and 's' denote number of filters, kernel size, and stride respectively. In each convolutional block, a batch normalization (BN) and a rectified linear unit (ReLU) are applied after the convolution operation. The network outputs finally 20×2 parameters of a thin plate spline transformation that is used to generate the sampling grid for spatially rectifying the input image.

Layer	Configuration
Input	-
Convolution	$maps: 32, k: 3 \times 3, s: 1 \times 1$
MaxPooling	$k: 2 \times 2, s: 2 \times 2$
Convolution	$maps: 64, k: 3 \times 3, s: 1 \times 1$
MaxPooling	$k: 2 \times 2, s: 2 \times 2$
Convolution	$maps: 128, k: 3 \times 3, s: 1 \times 1$
MaxPooling	$k: 2 \times 2, s: 2 \times 2$
Convolution	$maps: 256, k: 3 \times 3, s: 1 \times 1$
MaxPooling	$k: 2 \times 2, s: 2 \times 2$
Convolution	$maps: 256, k: 3 \times 3, s: 1 \times 1$
MaxPooling	$k: 2 \times 2, s: 2 \times 2$
Convolution	$maps: 256, k: 3 \times 3, s: 1 \times 1$
FC	hidden units: 512
FC	hidden units: 20×2

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

Layer	Configuration
ConvBlock 0	$maps: 32, k: 3 \times 3, s: 1 \times 1$
ConvBlock 1	$\begin{bmatrix} maps: 32, k: 1 \times 1 \\ maps: 32, k: 3 \times 3 \end{bmatrix} \times 3, s: 2 \times 2$
ConvBlock 2	$\begin{bmatrix} maps: 64, k: 1 \times 1 \\ maps: 64, k: 3 \times 3 \end{bmatrix} \times 4, s: 2 \times 2$
ConvBlock 3	$\begin{bmatrix} maps: 128, k: 1 \times 1 \\ maps: 128, k: 3 \times 3 \end{bmatrix} \times 6, s: 2 \times 1$
ConvBlock 4	$\begin{bmatrix} maps: 256, k: 1 \times 1 \\ maps: 256, k: 3 \times 3 \end{bmatrix} \times 6, s: 2 \times 1$
ConvBlock 5	$\begin{bmatrix} maps: 512, k: 1 \times 1 \\ maps: 512, k: 3 \times 3 \end{bmatrix} \times 3, s: 2 \times 1$
BiLSTM	hidden units: 256
BiLSTM	hidden units: 256
Att. GRU	hidden units: 512

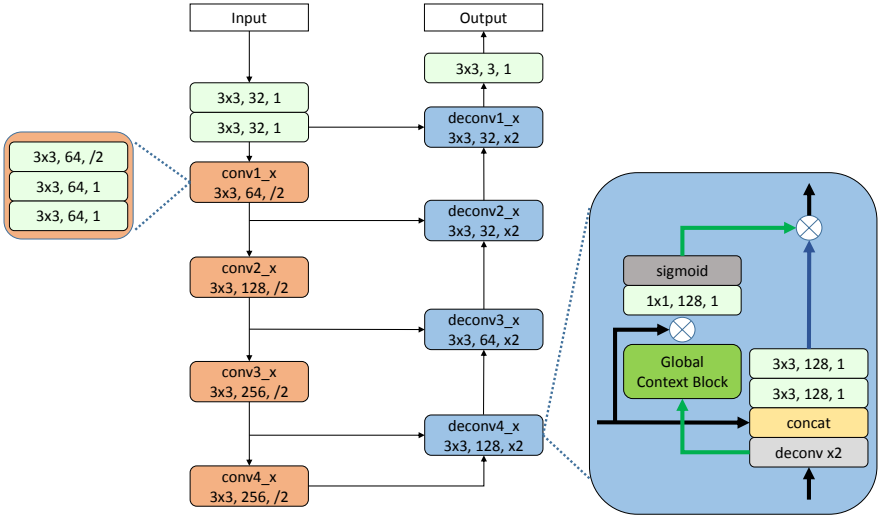

Figure 1: The architecture of the MGN and MRN networks in the residual enhancement network.

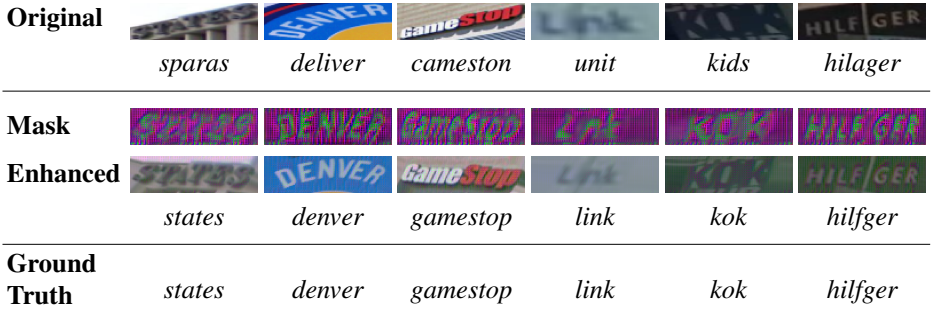

Figure 2: Qualitative evaluation of the effectiveness of the proposed text image enhancement model to improve recognition results. For each sample, the original text image, the enhanced image by the proposed enhancement model along with the predicted enhancement mask, and corresponding recognition results (the text under the images) are shown. The color representations of the enhanced text image and the enhancement mask are generated by taking the three output maps of the enhancement network (with values normalized to the range $[0, 1]$) as the R, G, and B color channels of the image and the mask.

Table 3: Text recognition accuracy with variant enhancement models. 'Hist Eq' denotes histogram equalization and 'Homo Filt' denotes homomorphic filtering, which replace the residual enhancement network (REN) as a separate pre-processing module in the text recognition model. 'STN' denotes the enhancement model comprising the spatial rectification network only, and 'Proposed' denotes the proposed end-to-end trainable adaptive enhancement model, as described in the paper.

Method	Regular Text				Irregular Text		
	IIIT5K	SVT	IC03	IC13	SVT-P	CT80	IC15
Hist Eq	91.8	88.9	94.5	91.4	81.2	78.5	74.3
Homo Filt	91.9	86.2	94.1	91.4	80.2	77.1	72.7
Homo Filt + Hist Eq	91.8	87.2	93.3	91.3	79.1	79.9	73.2
STN	93.5	89.5	95.2	93.2	82.2	79.5	75.7
Proposed	93.9	90.4	96.0	95.1	83.6	83.7	77.7

3 Text Recognition Performance with Traditional Image Enhancement Techniques

Table 3 shows the text recognition accuracy of employing two common traditional image enhancement techniques – histogram equalization and homomorphic filtering as a separate pre-processing module (i.e., not end-to-end optimized) in the text recognition model, combined with the retrained spatial rectification and recognition networks. Specifically, for the homomorphic filtering, a Butterworth high-pass filter with a cut-off frequency of 30 and a filter order of 2 is used, which is further wrapped in a high-frequency emphasis filter with parameters $\alpha = 0.75$ and $\beta = 1.25$. The histogram equalization requires no specific parameter setting.