An Adaptive Rectification Model for Arbitrary-Shaped Scene Text Recognition — Supplementary Material

Ye Qian
mf1833053@smail.nju.edu.cn
Long Chen
mg1933003@smail.nju.edu.cn
Feng Su
suf@nju.edu.cn

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

1 Estimation of 2D Projective Transformation Matrix

Given four pairs of corresponding control points (i.e., the vertices of a pair of quadrilateral source and target patches), the homogeneous deformation matrix **H** of a projective transformation can be formulated as follows:

$$\mathbf{H} = reshape([\mathbf{b} \quad 1])_{3\times 3} \tag{1}$$

where $reshape(\cdot)_{3\times 3}$ denotes a function reshaping the input tensor to a 3×3 view of it, and **b** is a 1×8 vector computed as follows:

$$\mathbf{b} = \mathcal{A}^{-1}\mathbf{x} \tag{2}$$

where **x** is an 8×1 vector of the coordinates of the four target control points. $\mathcal{A}$ is an 8×8 matrix formulated as:

$$\begin{bmatrix} r_x^{(0)} & r_y^{(0)} & 1 & 0 & 0 & 0 & -r_x^{(0)} * t_x^{(0)} & -r_y^{(0)} * t_x^{(0)} \\ 0 & 0 & 0 & r_x^{(0)} & r_y^{(0)} & 1 & -r_x^{(0)} * t_y^{(0)} & -r_y^{(0)} * t_y^{(0)} \\ r_x^{(1)} & r_y^{(1)} & 1 & 0 & 0 & 0 & -r_x^{(1)} * t_x^{(1)} & -r_y^{(1)} * t_x^{(1)} \\ 0 & 0 & 0 & r_x^{(1)} & r_y^{(1)} & 1 & -r_x^{(1)} * t_y^{(1)} & -r_y^{(1)} * t_y^{(1)} \\ r_x^{(2)} & r_y^{(2)} & 1 & 0 & 0 & 0 & -r_x^{(2)} * t_x^{(2)} & -r_y^{(2)} * t_x^{(2)} \\ 0 & 0 & 0 & r_x^{(2)} & r_y^{(2)} & 1 & -r_x^{(2)} * t_y^{(2)} & -r_y^{(2)} * t_y^{(2)} \\ r_x^{(3)} & r_y^{(3)} & 1 & 0 & 0 & 0 & -r_x^{(3)} * t_x^{(3)} & -r_y^{(3)} * t_x^{(3)} \\ 0 & 0 & 0 & r_x^{(3)} & r_y^{(3)} & 1 & -r_x^{(3)} * t_y^{(3)} & -r_y^{(3)} * t_y^{(3)} \end{bmatrix} \tag{3}$$

where $r_x^{(i)}$ and $r_y^{(i)}$ are x and y coordinates of the ith source control point respectively, while $t_x^{(i)}$ and $t_y^{(i)}$ are coordinates of the corresponding predefined target control point.

2 Network Configuration

Table 1 shows the configuration of the LN subnetwork for control/sampling points prediction in the proposed text rectification model. Table 2 shows the configuration of the attention-based text recognition network.

Table 1: Configuration of the LN subnetwork for control/sampling points prediction in the proposed text rectification model. FC denotes fully connected layer. 'maps', 'k', 's', and 'p' denote the number of filters, kernel size, stride, and padding size respectively. $K+1$ denotes the number of predicted pairs of sampling points on the Bezier curves depicting the upper and lower edges of a text region, while in the slicing scheme without Bezier sampling, it denotes the number of predicted pairs of control points of the slicing grid.

Layer	Configuration	Output Size
Input	-	$1 \times 32 \times 64$
Convolution	$maps: 32,\ k: 3,\ s: 1,\ p: 1$	$32 \times 32 \times 64$
MaxPooling	$k: 2 \times 2,\ s: 2 \times 2$	$32 \times 16 \times 32$
Convolution	$maps: 64,\ k: 3,\ s: 1,\ p: 1$	$64 \times 16 \times 32$
MaxPooling	$k: 2 \times 2,\ s: 2 \times 2$	$64 \times 8 \times 16$
Convolution	$maps: 128,\ k: 3,\ s: 1,\ p: 1$	$128 \times 8 \times 16$
MaxPooling	$k: 2 \times 2,\ s: 2 \times 2$	$128 \times 4 \times 8$
Convolution	$maps: 256,\ k: 3,\ s: 1,\ p: 1$	$256 \times 4 \times 8$
MaxPooling	$k: 2 \times 1,\ s: 2 \times 1$	$256 \times 2 \times 8$
Convolution	$maps: 256,\ k: 3,\ s: 1,\ p: 1$	$256 \times 2 \times 8$
MaxPooling	$k: 2 \times 1,\ s: 2 \times 1$	$256 \times 1 \times 8$
Convolution	$maps: 256,\ k: 3,\ s: 1,\ p: 1$	$256 \times 1 \times 8$
FC	hidden units: 512	512
FC	hidden units: $4 \times (K+1)$	$4 \times (K+1)$

3 Examples of Scene Text Rectification and Recognition Results

Tables 3, 4, and 5 show some text rectification results by the proposed rectification model on the three irregular scene text datasets, along with corresponding recognition results (shown under the images) on both the original and the rectified images. It can be seen that, through end-to-end training with the recognition network, the proposed rectification model effectively learns to rectify the text image in a way leading to improved recognition accuracy.

Table 2: Configuration of the attention-based text recognition network. 'maps', 'k', 's', and 'p' denote the number of filters, kernel size, stride, and padding size respectively.

Layer	Configuration	Output Size
Input	-	$1 \times 32 \times 100$
ConvBlock 0	$maps: 32,\ k: 3,\ s: 1,\ p: 1$	$32 \times 32 \times 100$
ConvBlock 1	$\begin{bmatrix} maps: 32, k: 1 \\ maps: 32, k: 3 \end{bmatrix} \times 3,\ s: 2 \times 2$	$32 \times 16 \times 50$
ConvBlock 2	$\begin{bmatrix} maps: 64, k: 1 \\ maps: 64, k: 3 \end{bmatrix} \times 4,\ s: 2 \times 2$	$64 \times 8 \times 25$
ConvBlock 3	$\begin{bmatrix} maps: 128, k: 1 \\ maps: 128, k: 3 \end{bmatrix} \times 6,\ s: 2 \times 1$	$128 \times 4 \times 25$
ConvBlock 4	$\begin{bmatrix} maps: 256, k: 1 \\ maps: 256, k: 3 \end{bmatrix} \times 6,\ s: 2 \times 1$	$256 \times 2 \times 25$
ConvBlock 5	$\begin{bmatrix} maps: 512, k: 1 \\ maps: 512, k: 3 \end{bmatrix} \times 3,\ s: 2 \times 1$	$512 \times 1 \times 25$
BiLSTM	hidden units: 256	$256 \times 1 \times 25$
BiLSTM	hidden units: 256	$256 \times 1 \times 25$
Att. GRU	hidden units: 256	-

Table 3: Results on SVT-P dataset

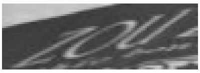
Original	 *restallizes*	 *camestud*	 *incated*	 *manehouse*
Rectified	 *restaurant*	 *gamestop*	 *theater*	 *warehouse*
Original	 *lines*	 *you*	 *straights*	 *alamount*
Rectified	 *mint*	 *zou*	 *marlboro*	 *paramount*

Table 4: Results on IC15 dataset

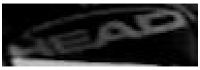
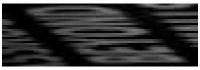
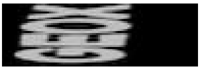
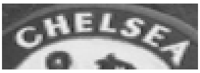
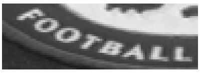
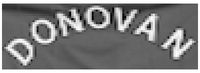
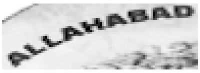
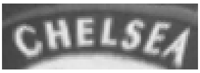
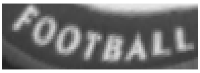
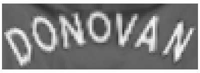
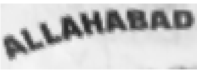
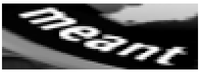
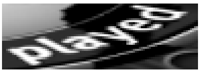
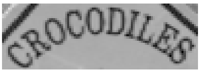
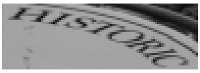
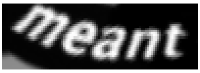
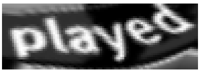
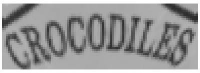
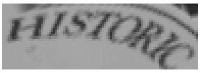
Original	 *kapa*	 *obhouse*	 *read*	 *sensure*
Rectified	 *kappa*	 *qbhouse*	 *head*	 *samsung*
Original	 *sender*	 *the*	 *const*	 *throcomn*
Rectified	 *beyond*	 *taken*	 *geox*	 *infocomm*
Original	 *galary*	 *lianga*	 *sende*	 *surger*
Rectified	 *galaxy*	 *manga*	 *sale*	 *burger*
Original	 *shangok*	 *kolinsony*	 *siccos*	 *contring*
Rectified	 *selangor*	 *robinsons*	 *jacobs*	 *pontian*

Table 5: Results on CT80 dataset

Original				
	chelse	*rooterize*	*conventi*	*prearrants*
Rectified				
	chelsea	*football*	*donovan*	*allahabad*
Original				
	dont	*please*	*crogodiles*	*mission*
Rectified				
	meant	*played*	*crocodiles*	*historic*