文章编号：2095-2163(2020)12-0080-06 中图分类号：TP391.1 文献标志码：A

基于矫正网络的场景文本识别应用与研究

赵高照，丁学明
（上海理工大学 光电信息与计算机工程学院，上海 200093）

摘 要： 场景文本在文字识别（Optical Character Recognition，OCR）领域一直是个难题，因此受到学术界的广泛关注。场景文本通常包括透视文本、弯曲文本、定向文本等。目前大多深度学习方法都不能够很好的识别这些不规则的文本，特别是严重变形的文本。针对上述问题，本文提出了一种迭代思想的矫正网络用于场景文本的识别，这种网络是一种端到端无需额外字符级注释的可训练网络。该矫正网络通过迭代细化的方式，逐步达到最优矫正。其中参数变换采用薄板样条（Thin Plate Spline，TPS）参数变换，自适应的进行图像变换，进而提高后序识别网络的识别性能。通过在大量公共数据集上进行的实验，证明了本文方法的有效性，特别是在不规则文本上的实验，证明了该方法有着较好的鲁棒性和准确性。

关键词： 场景文本；迭代；端到端；图像变换；TPS；不规则文本

Application and research of scene text recognition based on correction network

ZHAO Gaozhao，DING Xueming
（School of Optical-Electrical and Computer Engineering，University of Shanghai for Science and Technology，Shanghai 200093，China）

【Abstract】 Scene text has always been a difficult problem in the field of Optical Character Recognition（OCR），so it has been paid much attention in academic circles. Scene text usually includes perspective text，curved text，oriented text，etc. At present，most deep learning methods are not able to recognize these irregular texts，especially severely distorted texts. To solve the above problems，this paper proposes an iterative correction network for scene text recognition，which is an end-to-end trainable network without additional character level annotation. The correction network reaches the optimal correction step by step through the iterative refinement method. Thin Plate Spline（TPS）parameter transformation is used to transform the image adaptively，thus improving the recognition performance of the sequential recognition network. Experiments on a large number of public data sets demonstrate the effectiveness of the proposed method，especially on irregular text. It is proved that the proposed method has good robustness and accuracy.

【Key words】 Scene text；Iteration；End to end；Image transformation；Thin Plate Spline；Irregular text

0 引 言

随着科技的进步，文字识别技术逐渐应用于人们生活的各个领域。现实生活中的相关应用包括了通用场景文字识别、卡证文字识别、汽车场景文字识别、票据文字识别、教育场景文字识别，其他场景文字识别等等，其中场景文本占了相当大的部分。随着深度学习的发展，尽管在文字识别领域取得了很大的进展，但在场景文本识别方面仍然是一个不小的挑战[1]。

对于场景文本中出现的透视、弯曲等不可预测的变化问题，使得识别场景文本中任意形状的文本是一项极其困难的任务。神经网络的出现很大程度上推动了文字识别的发展，现如今的方法通常是利用卷积神经网络对输入文本进行特征的提取和分类[2]，以及将递归神经网络（RNN）应用于文本的序列识别，将 CNN 和 RNN 联合起来[3]，对文本图像的特征进行编码。这些方法主要是针对规则文本的识别。对于场景文本中的严重变形的文本，识别是比较困难的。解决这种问题通常的方法，是把不规则文本调整为易于识别的文本，最后送入识别网络进行识别。文献[4]中提出，在处理场景文本的识别问题时，把处理大量场景文本的识别问题分为矫正和识别两部分，使其在处理一些不规则文本的问题上有着明显的改善。但是，在处理一些严重变形的文本时仍然面临着各种问题。

针对上述问题，本文提出一种基于迭代思想的

基金项目： 国家自然科学基金（61673277）。

作者简介： 赵高照（1995-），男，硕士研究生，主要研究方向：深度学习、图像处理；丁学明（1971-），男，博士，副教授，主要研究方向：智能控制、系统辩识、嵌入式系统。

通讯作者： 丁学明 Email：xuemingding@ usst.edu.cn

收稿日期： 2020-10-27

矫正网络。对于严重变形的场景文本，一次矫正很难达到较好的识别性能指标，通过多次矫正可以很好的解决场景文本中的不规则性，每一次矫正都是对输入图像进行更好的矫正优化。本文把网络分为矫正和识别两部分。矫正网络部分的训练是通过识别网络向后传播的更好的场景文本识别来操作的，这是一种端到端的训练网络，无需人工中间注释，是一种自适应的网络模型。

在直接迭代矫正的过程中，每一次矫正都是基于之前的输入图像，预测出控制点的位置，使之接近于最优值。而每一次迭代将输出的矫正图像用于下一步的控制点预测估计。在这里对于输入图像，使用空间变换网络（Spatial Transformer Networks，STN）[5]进行操作，并且本文使用的 STN 网络是基于 TPS[6]参数变换的内核驱动。

1　迭代式矫正网络

1.1　迭代矫正

由于单次矫正不能很好地解决严重变形文本的透视、弯曲等问题，进而导致后序识别网络不能进行准确的识别。所以本文提出了采用迭代的方式对变形的文本进行矫正，在此将矫正过程分成多次进行，逐渐迭代细化，最终达到最优矫正。在第一次迭代时，把原始的场景文本作为输入。在空间变换的操作下，先经过定位网络预测其控制点在原图像上的位置，然后通过薄板样条插值（TPS），计算其转换参数 θ_n，最后输出的矫正图像，根据空间变换操作和转换参数 θ_n 从原图像中采样。第二次迭代矫正时，输入图像为第一次迭代矫正输出的矫正图像，接着送入同样的网络中进行参数估计和矫正。第三次、第四次……重复此过程。

假设在迭代矫正过程中，原始的输入图像为 I_0，第一次迭代的输出矫正图像为 I_1，第二次为 I_2，第 n 次为 I_n。空间变换操作为 T，其 TPS 转换参数为 θ_n，则每一次迭代矫正图像的采样公式为：

$$I_n = T(I_{n-1}, \theta_n). \tag{1}$$

如图 1 所示，本文设计的网络可以迭代式的进行空间变换，中间不需要额外的参数方面的注释，而且整个过程都是可微的，允许反向传播求导进行参数训练。在整个矫正过程完成后，可以很好的关注图片中利于识别方面的像素信息，并且将其放大，进而提供一个提高后续识别性能的矫正图片。

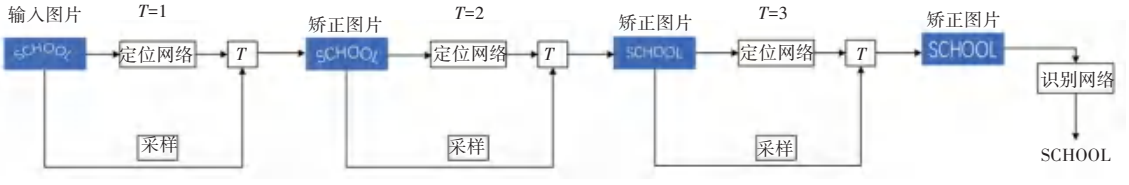

图 1　迭代矫正模型

Fig. 1　Iterative correction model

迭代矫正网络是以空间变换网络为基础的，空间变换网络主要分为 3 个部分：定位网络（Localisation net）、网格生成器（Grid generator）和采样器（Samper）。对于定位网络是由几个卷积层和全连接层组成，并且在每个卷积后面插入池化层，定位网络的作用是预测控制点的位置。在本文中将参数控制点的个数设为 20 个。

1.1.1　TPS 参数估计

在输入图像经过定位网络预测控制点之后，利用 TPS 参数变换生成一个网格采样器，将输入图像的像素一一对应到矫正后的图像上。TPS 插值函数是常用的 2D 插值方法，这种插值函数使图片弯曲的能量最小，这已经得到证明。

在第 n 次迭代时，定位网络预测的控制点位置归一化后设为 $A' = \{a_1', a_2', \cdots, a_t'\}$。这些控制点的位置表明了网络所关注的信息，是图片中文字的位置和姿态，并且这些控制点的位置由文字的边界线所决定。在矫正图像上，规定一个规范的控制点位置信息 $A = \{a_1, a_2, \cdots, a_t\}$，用来计算 TPS 变换的参数。式（2）即为矫正后图像的控制点位置计算公式。

$$a_i = \begin{cases} \left(\dfrac{2i}{t-2}, 0\right), 0 \leqslant i < \dfrac{t}{2}; \\ \left(\dfrac{2i-t}{t-2}, 1\right), \dfrac{t}{2} \leqslant i < t. \end{cases} \tag{2}$$

式中，t 为控制点的个数。

已知 t 个控制点，可以使用径向基函数，把控制点转换到另一个坐标系中去：

$$T(x) = \sum_{i=1}^{t} h_i \sigma(\| x - a_i \|). \tag{3}$$

式中，径向基函数 $\sigma(s) = s^2 \log s$。

其中，x 为标量，在二维空间上坐标有两个值，所以这里的插值函数应该为两个 x 和 y，因此向量形

式的插值函数为：

$$a' = T(a). \tag{4}$$

插值函数 T(a)可以表述为：

$$T(a) = m + n^T a + h^T s(a), \tag{5}$$

式中 m 为标量；n 为 2×1 维的向量；h 为 $t \times 1$ 维的向量；$s(a)$ 的计算公式为：

$$s(a) = (\sigma(\| a - a_1 \|), \sigma(\| a - a_2 \|), \cdots, \sigma(\| a - a_t \|)). \tag{6}$$

插值函数受到边界的约束条件为：

$$\begin{aligned} \sum_{i=1}^{t} h_i &= 0, \\ \sum_{i=1}^{t} a_i h_i &= 0. \end{aligned} \tag{7}$$

根据式(5)和式(7)，可将所有的条件写成矩阵形式为：

$$\begin{bmatrix} 1_t & A & S \\ 0 & 0 & 1_t^T \\ 0 & 0 & A^T \end{bmatrix} \begin{bmatrix} m \\ n \\ h \end{bmatrix} = \begin{bmatrix} A' \\ 0 \\ 0 \end{bmatrix}. \tag{8}$$

其中，S 为一个 $t \times t$ 的矩阵，$S_{ij} = \| a_i - a_j \|^2 \log(\| a_i - a_j \|)$。设 $\theta = [m \quad n \quad h]^T$，$\theta$ 即为 TPS 变换所求的参数。

$$\theta = \begin{bmatrix} 1_t & A & S \\ 0 & 0 & 1_t^T \\ 0 & 0 & A^T \end{bmatrix}^{-1} \begin{bmatrix} A' \\ 0 \\ 0 \end{bmatrix}. \tag{9}$$

给定一个矫正图像的点 a，对应的在输入图像上的点为 a'，a' 的计算公式为：

$$a' = \begin{bmatrix} 1 & a & \tilde{S} \end{bmatrix} \theta. \tag{10}$$

其中，$\tilde{S}_i = \|a - a_i\|^2 \log(\|a - a_i\|)$.

1.2.2 双线性插值采样

在式(10)中已经给出矫正图像与输入图像的对应关系，可根据矫正图像的坐标点取得对应输入图像的坐标点的像素值进行填充。但是在计算坐标点时，可能得出的坐标是带有小数的，若将其值四舍五入，可能在做反向传播时梯度难以下降。因此，本文在采样时采用双线性插值的方法对像素值进行填充。假设矫正图像的随意一个点为 (a_x, a_y)，经过式(10)可以得到其对应的原图像上的点 (a'_x, a'_y)。根据得到的坐标，按照以下公式计算像素值，进而填充到矫正图像上。

$$V = \sum_n \sum_m U_{nm} \max(0, 1 - |a'_x - m|) \max(0, 1 - |a'_y|). \tag{11}$$

在给出采样点坐标计算其像素值时，首先采集其周围点的像素值，然后根据式(11)计算出采样点的像素值，将其填充到矫正图像上。

1.2 识别网络

本文的识别网络采用带有注意力机制的序列到序列(Sequence - to - Sequence, seq2seq)编解码网络[7]，这个网络是由编码部分和解码部分组成。

1.2.1 编码网络

在编码部分，先对输入的矫正图像进行卷积操作，提取其特征信息。随着网络的加深，能够获取的信息就越多，而且提取的特征也就越丰富。但是伴随出现的问题就是，网络的加深使得优化效果越差，最后的识别准确率也随之降低。这是因为，网络的加深会造成梯度爆炸和梯度消失的问题。对此，本文在卷积部分加入了残差单元(ResNet)[8]来解决梯度爆炸和消失问题。

尽管在加入了残差单元的卷积网络能够提取较为丰富的特征信息，但是这些特征都是在矫正图像的区域提取的。为了提取丰富的特征，本文使用了双向长短期记忆网络(BLSTM, Bi-directional)[9]，从特征序列中提取文字在正反两个方向上的依赖关系，利用更为丰富的上下文关系提高文本的识别性能。

1.2.2 解码网络

在解码部分，本文使用了带有注意力机制的长短期记忆网络(Long Short-Term Memory, LSTM)来进行解码操作。对于编码网络生成的特征序列，需要生成正确顺序的序列，因此使用单向的 LSTM 网络。对于 seq2seq 模型，如果特征向量太长，在编码网络的后面部分就会被逐渐遗忘，而解码网络接收到的特征也就不完整。加入注意力机制后，解码网络每次更新状态时都会再次访问编码网络的所有状态，并且还会告诉 decoder 网络更要关注哪些部分。图 2 为加入 attention 后的效果曲线，图 3 为基于注意力机制的 seq2seq 模型。

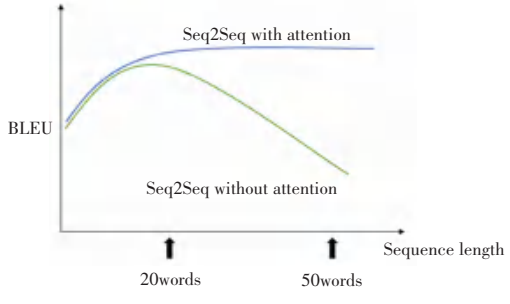

图 2 注意力机制效果曲线

Fig. 2 Attention mechanism effect curve

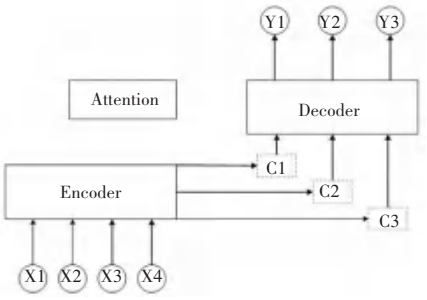

图 3　基于注意力机制的 seq2seq 模型

Fig. 3　Seq2seq model based on attention mechanism

具体做法为：解码器在每次迭代解码的过程中都会查询编码器的隐藏状态；计算输入序列特征的每一个位置相对于当前解码位置内容的相关程度，即权重。再根据权重，对各输入位置的隐藏状态进行加权平均，得到上下文向量 c，该向量包含了与当前解码部分的内容最相关的输入序列信息。在下一步解码过程中，上下文向量 c 将作为额外的信息输入 LSTM 网络中。这种情况下 LSTM 网络的每一时刻都可以读取到输入序列的信息，而不仅仅是上一时刻隐藏状态的信息。其中，第 j 时刻的上下文向量计算公式为：

$$\alpha_{ij}=\frac{\exp(e(h_i,s_i))}{\sum_i \exp(e(h_i,s_i))}, \tag{12}$$

$$e(h,s)=U\tanh(Vh+Ws), \tag{13}$$

$$c_j=\sum_i \alpha_{ij}h_i. \tag{14}$$

式中，U、V、W 为模型参数；h_i 表示编码器在第 i 个字符上的输出；s_i 为编码器预测第 j 个字符的状态；α 为通过 Softmax 计算的权值；$e(h,s)$ 为计算原文各单词与当前解码器状态的相关度函数，其构成了包含一个隐藏层的全连接神经网络。

2　实验研究

针对提出的迭代式矫正网络，分别在 3 个规则数据集上和 4 个不规则数据集上进行了该方法准确性评估，从而验证其有效性。规则数据集包括：IIIT5k[10]、ICDAR2003[11]、ICDA2013[12]；不规则数据集包括：Street View Text[13]、ICDAR2015[14]、Street View Text Perspective[15]、CUTE80[16]。本文在上述数据集上进行了迭代次数的对比实验，来验证迭代思想对识别性能的影响。图 4 为部分 SVTP 数据集的可视化，以及迭代一次后的矫正效果。本文中的实验均是在无字典后处理的训练方法下进行的。同时，本文与当前针对场景文本识别性能较好的主流网络进行对比实验，对比实验中还包括其他较早的一些场景文本的识别方法。

图 4　场景文本的矫正效果图

Fig. 4　Correction effect of scene text

本文将 Synth90k[17] 与 SynthText[18] 两个公共数据集作为训练样本，来训练神经网络模型。实验中使用的训练集和测试集都是来自于生活中的场景文本，并且数据集没有进行任何修改调整。大量的实验表明，本文基于矫正迭代的网络，相对于其他方法可以获得更好的识别效果。通过在规则文本和不规则文本的实验结果表明了本文方法的有效性。即使对于场景文本中出现的严重形变文本，仍然具有较好的鲁棒性。

2.1　迭代实验分析

为了实现本文提出的迭代矫正识别，首先对网络模型进行参数的训练。使用两个公共合成文本数据集，作为训练样本训练模型。在模型的训练过程中，分别采用了两种不同的方法：一种是基于 unknown 的方法，另外一种训练方法是不带 unknown 的。在正常训练文本识别网络时，如识别 26 个小写英文字母，全连接层输出 27 类，多出来的一类是 Eos 终止符。但是，在场景文本的训练集中，有些图片存在一些非法字符，在此把这些非法字符归结为 unknown 字符，这种训练方法就是基于 unknown 训练。表 1、表 2 中给出了两种方法训练的实验结果。

表 1 无 unknown 迭代实验

Tab. 1 No unknown iteration experiments

训练方法	测试集						
	IIIT5k	IC03	IC13	SVT	SVTP	CUTE	IC15
filter_unknown_1	86.90	90.49	91.82	80.68	73.33	77.35	72.74
filter_unknown_2	88.17	91.05	91.53	82.53	74.11	79.79	74.25

表 2 unknown 迭代实验

Tab. 2 Unknown iteration experiments

训练方法	测试集						
	IIIT5k	IC03	IC13	SVT	SVTP	CUTE	IC15
unknown_1	93.51	92.63	91.87	89.51	79.12	79.68	75.54
unknown_2	94.30	93.72	93.15	90.88	82.64	84.38	76.94

从表 1、表 2 的对比实验可以看出,加入 unknown 的类别后,本文的网络模型能够获得更好的识别性能。表中的 filter_unknown_1 和 filter_unknown_2,分别表示无 unknown 类别训练迭代一次和两次。本文后面的实验均是基于 unknown 训练方法得出的实验结果。

本文中的方法是对场景文本进行迭代矫正,逐渐细化输入图像,得到更好的矫正图像,从而更容易识别。因此可以得知,迭代次数是影响识别性能的重要指标。表 3 中给出了在规则数据集和不规则数据集上的文本识别结果。从表中可以看出,随着迭代次数的增加,文本的识别性能逐渐增加。从前三个规则文本的迭代实验与后四个不规则文本的迭代实验的对比中,不规则文本迭代实验的识别性能的改善要比规则文本的改善更加明显。这表明本文的迭代网络针对不规则文本的识别具有明显的作用。另外,本文的迭代实验收敛速度具有很大的优势,在第二次迭代时就获得了较好的识别性能。在第三次迭代时,文本的识别能力逐渐稳定。因此,本文提出的迭代矫正网络可以让模型训练更加稳定便捷。

表 3 迭代实验结果

Tab. 3 Iteration experiments

迭代次数	测试集						
	IIIT5k	IC03	IC13	SVT	SVTP	CUTE	IC15
0	90.95	91.45	90.12	86.24	75.23	74.53	72.87
1	93.51	92.63	91.87	89.51	79.12	79.68	75.54
2	94.30	93.72	93.15	90.88	82.64	84.38	76.94
3	94.29	93.85	93.21	90.56	81.98	84.42	76.92

矫正迭代次数的增加,会逐渐使不规则文本逐渐迭代细化更有助于后续对矫正文本的识别。矫正网络不仅会使输入图片朝着更有利于识别的方向改变,还会逐渐去除图片中的背景噪声。

2.2 对比实验分析

为了验证本文提出方法的有效性,在 7 个包含规则和不规则的数据集上进行评估,这些数据集中既有正常的场景文本,又有具有各种透视、弯曲等形变的场景文本。将本文提出的方法与当前主流的文本识别方法进行对比实验。表 4 和图 5 为该对比实验的识别结果。

表 4 对比实验结果

Tab. 4 Contrast experiment

方法	测试集						
	IIIT5k	IC03	IC13	SVT	SVTP	CUTE	IC15
RARE[19]	81.9	90.1	88.6	81.9	71.8	59.2	–
ASTER[4]	93.4	94.5	91.8	89.5	78.5	79.5	76.1
ESIR[20]	93.3	–	91.3	90.2	79.6	83.3	76.9
PRN[21]	94.3	94.0	93.3	88.7	81.2	88.2	76.8
Ours	94.30	93.72	93.15	90.88	82.64	84.38	76.94

由表 4 的对比实验结果可以看出,本文方法在 7 个场景文本的数据集上都取得了较高的识别性能。与当前最新的文字识别方法 PRN 的对比可以看出,本文的方法基本上与之不相上下。值得一提的是,在对比实验中本文的方法采用两次迭代的识别结果,而 PRN 采用的是四次迭代的识别结果。通过观察可以看出,本文的方法在 III5k、SVT、SVTP、IC15 上均取得了最好的识别效果,而在 IC13、CUTE 取得了第二的识别效果。但在 CUTE 数据集上识别效果要比 PRN 低近 4 个百分点。

相比之下,本文的方法能够有效的矫正严重变形的文本,并且取得不错的识别结果。该矫正网络在场景文本不规则的文字处理上,能够明显的改善其识别性能。

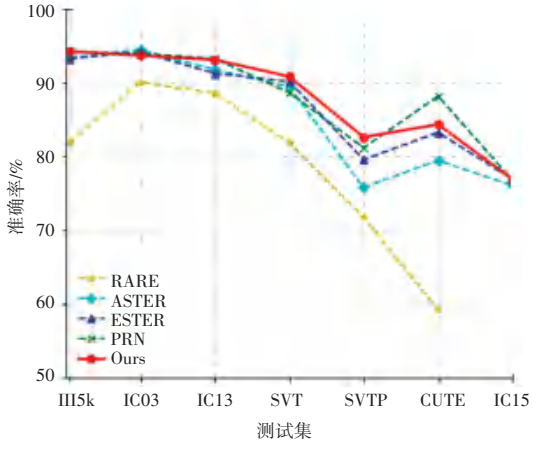

图 5 数据集识别准确率曲线

Fig. 5 Data set recognition accuracy curve

3　结束语

对于场景文本的识别效果主要取决于输入图像的质量和后续识别网络的性能。本文通过分析输入图像中文本的变形程度对识别性能的影响，提出一种迭代式的矫正网络，对输入图片不断进行迭代细化矫正，将对图片中文本的注意力通过矫正网络逐渐放大；同时去除图片中的背景噪声，进而改善输入图片的质量，以达到更好的识别性能。在 7 个包含大量场景文本数据集上的实验证明，本文提出的方法能够有效的提高场景文本的识别精度。通过对比实验可以看出，本文的方法较其他先进方法均获得了更好的识别结果。该方法是一种端到端的训练网络，无需中间的人工注释。通过在常规的场景文本数据集和不规则的场景文本数据集上的对比迭代实验，可以看出本文方法能够有效的改善不规则文本的识别性能，在透视和弯曲等变形本文中仍然具有较强的鲁棒性，在场景文本识别方面的有效性。

参考文献

[1] 张琳. 自然场景中任意形状文字提取关键技术研究[D]. 西安：西安理工大学，2020.

[2] JADERBERG M, SIMONYAN K, VEDALDI A, et al. Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition[J]. Eprint Arxiv, 2014.

[3] SHI B, BAI X, YAO C. An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 39(11): 2298-2304.

[4] BAOGUANG S, MINGKUN Y, XINGGANG W, et al. ASTER: An Attentional Scene Text Recognizer with Flexible Rectification[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2018, 1-1.

[5] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial Transformer Networks[J]. 2015.

[6] BOOKSTEIN F L. Principal warps: thin-plate splines and the decomposition of deformations[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002, 11(6): 567-585.

[7] SUTSKEVER I, VINYALS O, LE Q V. Sequence to Sequence Learning with Neural Networks[C]// NIPS. MIT Press, 2014.

[8] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society, 2016.

[9] GRAVES A, LIWICKI M, SANTIAGO FERNáNDEZ, et al. A Novel Connectionist System for Unconstrained Handwriting Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2009, 31(5): 855-868.

[10] GRAVES, ALEX, LIWICKI, et al. A Novel Connectionist System for Unconstrained Handwriting Recognition. [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2009.

[11] LUCAS S M, PANARETOS A, SOSA L, et al. ICDAR 2003 Robust Reading Competitions: Entries, Results and Future Directions[J]. Document Analysis and Recognition, 2005, 7(2-3): 105-122.

[12] KARATZAS D, SHAFAIT F, UCHIDA S, et al. ICDAR 2013 robust reading competition [C]// Document Analysis and Recognition (ICDAR), 2013 12th International Conference on. 2013.

[13] WANG K, BABENKO B, BELONGIE S. End-to-end scene text recognition[C]// IEEE International Conference on Computer Vision. IEEE, 2012.

[14] YAO C, WU J, ZHOU X, et al. Incidental Scene Text Understanding: Recent Progresses on ICDAR 2015 Robust Reading Competition Challenge 4[J]. Computer ence, 2015.

[15] PHAN T Q, SHIVAKUMARA P, TIAN S, et al. Recognizing Text with Perspective Distortion in Natural Scenes[C]// IEEE International Conference on Computer Vision. IEEE, 2014.

[16] RISNUMAWAN A, SHIVAKUMARA P, CHAN C S, et al. A robust arbitrary text detection system for natural scene images[J]. Expert Systems with Applications, 2014, 41(18): 8027-8048.

[17] JADERBERG M, SIMONYAN K, VEDALDI A, et al. Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition[J]. Eprint Arxiv, 2014.

[18] GUPTA A, VEDALDI A, ZISSERMAN A. Synthetic Data for Text Localisation in Natural Images[C]// IEEE International Conference on Computer Vision. IEEE, 2016.

[19] SHI B, WANG X, LYU P, et al. Robust Scene Text Recognition with Automatic Rectification[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[20] ZHAN F, LU S. ESIR: End-To-End Scene Text Recognition via Iterative Image Rectification[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[21] GAO Y, CHEN Y, WANG J, et al. Progressive rectification network for irregular text recognition[J]. 中国科学, 2020(2): 7-20.