# OSTER: An Orientation Sensitive Scene Text Recognizer with CenterLine Rectification

Zipeng Feng[1,2], Chen Du[1,2], Yanna Wang[1], and Baihua Xiao[1(✉)]

[1] The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation Chinese Academy of Sciences, Beijing, China
{fengzipeng2017,duchen2016,wangyanna2013,baihua.xiao}@ia.ac.cn
[2] University of Chinese Academy of Sciences, Beijing, China

**Abstract.** Scene texts in China are always arbitrarily arranged in two forms: horizontally and vertically. These two forms of texts exhibit distinctive features, making it difficult to recognize them simultaneously. Besides, recognizing irregular scene texts is still a challenging task due to their various shapes and distorted patterns. In this paper, we propose an orientation sensitive network aiming at distinguishing between Chinese horizontal and vertical texts. The learned orientation is then passed into an attention selective network to adjust the attention maps of the sequence recognition model, leading it working for each type of texts respectively. In addition, a lightweight centerline rectification network is adopted, which enables the irregular texts more readable while no redundant labels are needed. A synthetic dataset named SCTD is released to support our training and evaluate the proposed model. Extensive experiments show that the proposed method is capable of recognizing arbitrarily-aligned scene texts accurately and efficiently, achieving state-of-the-art performance over a number of public datasets.

AQ1

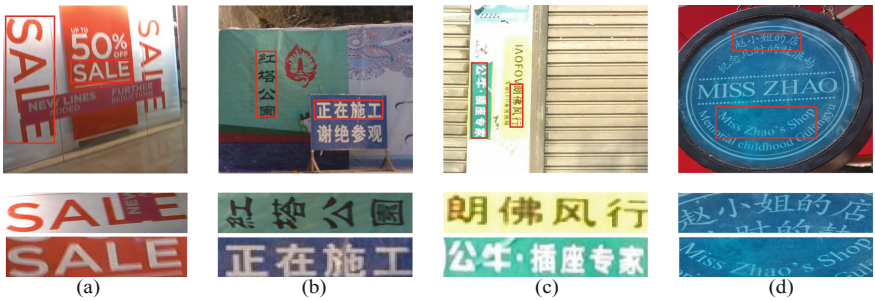
**Keywords:** Scene text recognition · Arbitrarily-aligned Chinese texts · Attention selective network · Centerline rectification · Sequence-to-sequence

## 1 Introduction

Scene text recognition has attracted much interest in the computer vision field because of its various applications, such as intelligent driving and goods identification. Nowadays, text recognition methods based on convolutional neural networks [9] have gained large success, especially integrating convolutional neural networks with recurrent neural networks [20] and attention mechanisms [28].

However, the majority of the scene text recognition algorithms only cover English texts. According to our statistics, English vertical texts always have the same arrangement mode as horizontal ones, as Fig. 1(a) shows. But in China, horizontal and vertical texts usually have completely different arrangement modes.


S. Palaiahnakote et al. (Eds.): ACPR 2019, LNCS 12046, pp. 1–15, 2020.
https://doi.org/10.1007/978-3-030-41404-7_34




(a) (b) (c) (d)

**Fig. 1.** Examples of arbitrarily-aligned scene texts. (a) Horizontal and vertical English texts generally have the same arrangement mode. (b) But Chinese scene texts have two different arrangement modes. (c) Identifying the arrangement modes of Chinese texts just by their aspect ratios may lead to terrible mistakes. (d) Curved and distorted patterns are popular in both Chinese and English scene texts.

If training them as a whole, the recognition network will learn entirely different features with the same label, which could puzzle the network, causing the training process slowly and inaccurately. In practice, we usually identify the orientation of texts depending on heuristic rules such as aspect ratio, but it is not enough to deal with scene texts which are arbitrarily-aligned. As Fig. 1(c) shows, some vertical texts have the same arrangement mode as horizontal texts, but they may be misclassified by merely aspect ratio. For these reasons, learning the orientation of texts is a more general approach.

Besides, irregular texts appear in natural scenes frequently owing to curved character placement, perspective distortion, etc. Recognizing texts with arbitrary shapes is an extremely difficult task because of unpredictable changeful text layouts. As illustrated in Fig. 1(d), both English and Chinese texts suffer from various irregular texts, causing additional challenges in recognition.

In this paper, we propose an orientation sensitive scene text recognition network (OSTER) to learn the orientation of texts automatically and deal with horizontal and vertical texts simultaneously. It consists of an orientation sensitive network (OSN), a centerline rectification network (CRN) and an attention selective sequence-to-sequence recognition network (ASN). The OSN judges the orientation of the input text images. As a result, the features of horizontal and vertical images could be learned respectively, helping the network to learn quickly and accurately. The CRN models the centerline of scene texts and corrects the distorted text to a regular one by fitting the equidistant sampling points with the centerline. The ASN generates entirely different attention maps for horizontal and vertical texts and predicts different character sequence according to the learned orientation weights. The whole network can be trained end-to-end by a multi-task learning manner.

The contributions of this paper can be summarized as follows:

- We propose an orientation sensitive network to learn the orientation of the input texts automatically, and an attention selective network to generate entirely different attention maps for horizontal and vertical texts, which improve the performance of the recognition network.
- We propose a centerline rectification network to correct the distorted text to a regular one by fitting the equidistant sampling points with the centerline.
- We establish a synthetic Chinese scene text dataset to train the proposed model, which includes half of horizontal texts and half of vertical texts.
- We develop an end-to-end trainable system that is robust to parameter initialization and achieves superior scene text recognition performance over a number of public datasets. To our knowledge, this paper is the first work to deal with different arrangement modes of Chinese texts.

## 2 Related Works

Recent years, scene text recognition has been widely researched and a variety of recognition methods have been proposed. The traditional character-level scene text recognition methods first generate multiple candidate character positions, and then applies the character classifier for recognition [2,26]. With the successful application of recurrent neural network (RNN) in sequence recognition, Shi [20] integrated convolutional neural networks (CNN) with RNN and proposed an end-to-end trainable network named CRNN. After that, Lee [13] proposed an attention mechanism, which can detect more discriminative regions in images, thus improving recognition performance. Facing the attention drift problem, Cheng [3] proposed the focusing attention network (FAN) to automatically adjust the attention weights. Liu [15] presented a binary convolutional encoder-decoder network (B-CEDNet). Combined with a bidirectional recurrent neural network, it can achieve significant speed-up. Although these approaches have shown attractive results, the irregular texts still can not be dealed with effectively. The main reason is that environments in scene texts are various, such as complicated backgrounds, perspective distortion and arbitrary orientation.

Due to these difficult cases, irregular text recognition has attracted more and more attention from researchers. Shi [21] applied the spatial transformer network (STN) [10] for text rectification, then put the rectified text images into the sequence-to-sequence recognition network to get the final result. To rectify the distorted text image better, Zhan [29] developed an iterative rectification framework, which can estimate and correct perspective distortion and text line curvature iteratively. Different from rectifying the entire distorted text image, Liao [14] presented a method called Character Attention FCN (CA-FCN), which modeled the irregular text images in a two-dimensional fashion. Although considerably improving the performance for irregular text recognition, it is still difficult to precisely locate the fiducial points which tightly bound the text regions, especially for severely distorted texts. This leads to errors in parameters estimation of the STN and causes the deformation of scene texts. Different from the existing methods, we model the centerline of texts and correct the distorted texts to straight lines, which is robust and flexible in scene text rectification.

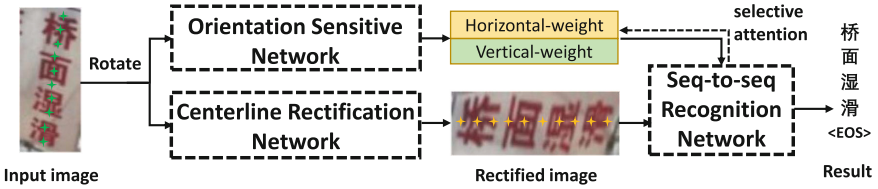



**Fig. 2.** Overview of the proposed OSTER. The centerline rectification network corrects the distorted text to a regular one; the orientation sensitive network perceives the arrangement mode of input images; the attention selective recognition network generate corresponding attention map at each time step for better recognition results.

At present, some methods process recognition task from the perspective of the orientation of texts. Cheng [4] devised an arbitrary orientation network (AON) to extract visual features of texts in four directions. Weighting the four-direction sequences of features, it can predict the orientation of rotated texts in an unsupervised learning method. However, the essence of AON is dealing with irregular texts caused by rotation, which is extremely different from learning arrangement modes of Chinese scene texts in this paper. Besides, the strategy of scaling word images to a square in AON will destroy the information of text lines, especially for Chinese texts whose aspect ratio are relatively large. For these reasons, we propose an orientation sensitive network to learn the orientation of the input texts automatically, and then pass the learned orientation into an attention selective network to deal with horizontal and vertical texts respectively.

## 3 Methodology

This section will present the proposed model including CRN, OSN and ASN. The overview of our proposed network is as Fig. 2.

### 3.1 Centerline Rectification Network

Recently, the STN based rectification methods have achieved great success in recognizing irregular texts, but there are still some problems to deal with. Firstly, it is still difficult to precisely locate the fiducial points which tightly bound the text region, especially for severely distorted text. Once the prediction of fiducial points deviates, the texts will often be incomplete, which will have a great impact on the recognition results. Our experiment shows that the background left at the text edges does not disturb the recognition results too much, but the real reason lies in that the character features in the distorted text line are significantly different from those of the horizontal text. Therefore, modeling the pose of texts and correcting the curved texts to straight lines should be paid more attention.

Some recent works [29] model the centerline of texts using the polynomial fitting method and have made great progress. But compared with predicting offsets of fiducial points, fitting polynomials is a more abstract task which needs

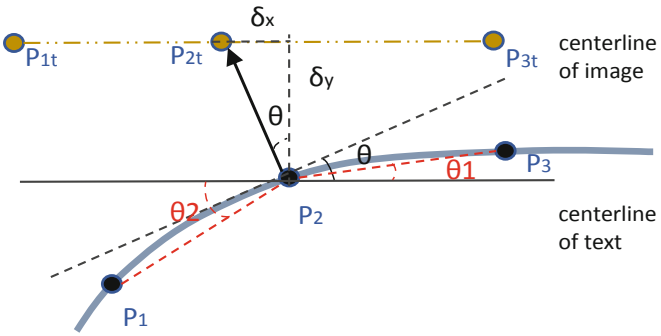



**Fig. 3.** Schematic of curved text correction.

greater learning cost. As a contrast, CRN predicts the offset from the centerline of texts to the centerline of the input images directly, which can correct the irregular texts at a small cost.

The CRN works as follows: firstly, it models the centerline of the texts and takes several points on it with equal sampling. As Fig. 3 shows, we predict N equal sampling points $(P_1, P_2, P_3, \ldots, P_N)$ on the centerline of the text. Their corresponding points on the centerline of image are $(P_{1t}, P_{2t}, P_{3t}, \ldots, P_{Nt})$. Each point $P_i(i = 1, \ldots, N)$ can be represented by $(x_i, y_i)$.

In order to distribute the characters as evenly as possible, avoiding deformation due to compression and stretching, areas with more severe curved should be mapped for longer horizontal distances. Therefore, mapping along the normal direction is an ideal choice. Apparently, the $y$-coordinate of $P_{it}$ is $h/2$, where $h$ is the height of input image. The $x$-coordinate of $P_{it}$ can be computed by

$$x_{it} = x_i - \Delta x_i \tag{1}$$

$$\Delta x_i = \Delta y_i \times \tan\theta_i \tag{2}$$

$$\Delta y_i = y_i - \frac{h}{2} \tag{3}$$

As can be seen from Fig. 3, $\theta_i$ can be approximated to $(\theta_{i1} + \theta_{i2})/2$, where

$$\tan\theta_{i1} = |\frac{y_{i+1} - y_i}{x_{i+1} - x_i}| \tag{4}$$

$$\tan\theta_{i2} = |\frac{y_i - y_{i-1}}{x_i - x_{i-1}}| \tag{5}$$

The CRN estimate the location of points by employing a network described in Table 1. Once the origin points and target points are determined, scene text distortions can then be corrected by a thin plate spline transformation (TPS) [22]. A grid can be generated within the distorted scene text, and a sampler is implemented to produce the rectified scene text image by using the determined

**Table 1.** Architecture of the CRN and OSN

| Module | Layers | Out size | Configurations |
|---|---|---|---|
| CRN | Conv1 | $15 \times 49$ | $3 \times 3conv$, 32, $2 \times 2pooling$, $0padding$ |
| | Conv2 | $6 \times 23$ | $3 \times 3conv$, 64, $2 \times 2pooling$, $0padding$ |
| | Conv3 | $2 \times 10$ | $3 \times 3conv$, 128, $2 \times 2pooling$, $0padding$ |
| | Conv4 | $1 \times 10$ | $3 \times 3conv$, 64, $2 \times 1pooling$, $1padding$ |
| | Conv5 | $1 \times 10$ | $1 \times 1conv$, 2, $0padding$, Tanh activate |
| OSN | Conv1 | $16 \times 50$ | $3 \times 3conv$, 32, $2 \times 2pooling$, $1padding$ |
| | Conv2 | $8 \times 25$ | $3 \times 3conv$, 64, $2 \times 2pooling$, $1padding$ |
| | Conv3 | $4 \times 12$ | $3 \times 3conv$, 128, $2 \times 2pooling$, $1padding$ |
| | Conv4 | $1 \times 5$ | $3 \times 3conv$, 256, $2 \times 2pooling$, $0padding$ |
| | FC1 | 64 | Sigmoid activate |
| | FC2 | 1 | Sigmoid activate |

grid, where the value of the pixel $p_t$ is bilinearly interpolated from the pixels near $p$ within the distorted scene text image. It should be noted that the training of the localization network does not require any extra annotation but is completely driven by the gradients that are back-propagated from the recognition network. Only estimating the $x$-coordinate and $y$-coordinate of points, the total parameter number is just $2N$. In our trained network, 10 points are employed for scene text pose estimation.

### 3.2 Orientation Sensitive Network

The OSN is designed to judge the orientation of input text images, and pass the confidence of orientation to the subsequent recognition network. As a result, the horizontal image features and vertical image features could be learned respectively, which is beneficial to achieve better recognition performance.

Similar to the binary classification problem, the OSN can learn the orientation attributes of texts guided by the orientation label in the training stage. For every input scene text, the OSN can identify the orientation of it precisely. Compared with training all texts as a whole, it provides more discriminative features for recognition network, improving the recognition ability of the network. Compared with training horizontal and vertical texts respectively, it simplifies the recognition process and can handle abnormal cases as Fig. 1(c) shows, improving the robustness and convenience of the recognition network. The structure of OSN can be viewed as Table 1 too.

### 3.3 Attention Selective Sequence-to-Sequence Network

The recognition network employs a sequence-to-sequence model with an attention mechanism. It consists of a convolutional-recurrent neural network encoder and an attention-based recurrent neural network decoder.

**Table 2.** Architecture of the recognition network

| Layers | Out size | Configurations |
|---|---|---|
| Block0 | $32 \times 100$ | $3 \times 3 conv, 32, 1 \times 1 stride$ |
| Block1 | $16 \times 50$ | $\begin{bmatrix} 1 \times 1 conv, 32 \\ 3 \times 3 conv, 32 \end{bmatrix} \times 3, 2 \times 2 stride$ |
| Block2 | $8 \times 25$ | $\begin{bmatrix} 1 \times 1 conv, 64 \\ 3 \times 3 conv, 64 \end{bmatrix} \times 4, 2 \times 2 stride$ |
| Block3 | $4 \times 25$ | $\begin{bmatrix} 1 \times 1 conv, 128 \\ 3 \times 3 conv, 128 \end{bmatrix} \times 6, 2 \times 1 stride$ |
| Block4 | $2 \times 25$ | $\begin{bmatrix} 1 \times 1 conv, 256 \\ 3 \times 3 conv, 256 \end{bmatrix} \times 6, 2 \times 1 stride$ |
| Block5 | $1 \times 25$ | $\begin{bmatrix} 1 \times 1 conv, 512 \\ 3 \times 3 conv, 512 \end{bmatrix} \times 3, 2 \times 1 stride$ |
| BiLSTM | 25 | 256 hidden units |
| BiLSTM | 25 | 256 hidden units |
| AttLSTM | - | 256 hidden units, 256 attention units |
| AttLSTM | - | 256 hidden units, 256 attention units |

We employ a 31-layer ResNet to extract the sequence feature from the input image. A $2 \times 2$ stride convolution is implemented to down-sample feature maps in the first two residual blocks, which is changed to $2 \times 1$ stride in all following residual blocks. This helps to reserve more information along the horizontal direction and is very useful for distinguishing neighbor characters. To enlarge the feature context, we adopt a multi-layer bidirectional LSTM (BLSTM) network [5] over the feature sequence.

An attention-based decoder directly generates the target sequence from an input feature sequence. We trade $T$ for the largest number of steps generated by the decoder and $L$ for the length of the feature sequence. At time step $t$, the vector of attention weights $\alpha_{t,i}$ is as follows:

$$\alpha_{t,i} = exp(c_{t,i}) / \sum_{j=1}^{L} (exp(c_{t,j})) \tag{6}$$

$$c_{t,i} = w_1 \times e^h_{t,i} + w_2 \times e^v_{t,i} \tag{7}$$

The weights $w_1$ and $w_2$ indicate the possibilities of each orientation, which are predicted by OSN. Apparently, we have $w_1 + w_2 = 1$. The $e^h_{t,i}$ and $e^v_{t,i}$ denote the intermediate attention states of horizontal and vertical sequence separately. They are generated by two distinct learnable nonlinear transformations of activated value $e_{t,i}$. Each of the nolinear transformation is a full connection layer followed by a $1 \times 1$ convolution. And $e_{t,i}$ can be described as follow:

$$e_{t,i} = Tanh(W_s s_{t-1} + W_h h_i + b) \tag{8}$$

where $s_{t-1}$ is the hidden state at previous time step $t-1$, and $h_i$ indicates the sequential feature vectors generated by the convolutional-recurrent encoder. $W_s$, $W_h$ and $b$ are learnable parameters.

The target sequence can be described as $(y_1, ..., y_T)$. At time step $t$, the output $y_t$ is:

$$y_t = softmax(W_{out}^T s_t + b_{out}) \tag{9}$$

where $W_{out}$ and $b_{out}$ are learnable parameters, too. $s_t$ is the hidden state at time step $t$, which can be computed by:

$$s_t = RNN((f(y_{t-1}), g_t), s_{t-1}) \tag{10}$$

Normally, the RNN function above represents an LSTM network. $(f(y_{t-1}, g_t)$ indicates the concatenation of $g_t$ and one-hot embedding of $y_{t-1}$. $g_t$ can be computed by:

$$g_t = \sum_{i=1}^{L} (\alpha_{t,i} h_i) \tag{11}$$

From the description above, the attention value of decoder is bound to the orientation confidence at the time step scale. Thus the recognition network deals with horizontal and vertical texts respectively, avoiding the confusion of samples with different arrangements.

The decoder stops processing when it predicts an end-of-sequence token "EOS" [24]. We adopt beam search [22] while testing to achieve better performance. The architecture of the recognition network can be viewed as Table 2.

### 3.4 Training

We integrate the OSN, CRN and ASN into one network. Using word-level annotations and orientation labels, the whole network can be trained end-to-end by a multi-task learning manner. It's worth noting that random initialization will not affect the performance of this network. The loss consists of two partly losses:

$$Loss = L_w + \lambda L_o \tag{12}$$

where $L_w$ and $L_o$ denote the word prediction loss and orientation classification loss respectively. The word prediction loss function can be described as follows:

$$L_w = -\frac{1}{2} \sum_{l=1}^{L} (logp_1(y_l|I) + logp_2(y_l|I)) \tag{13}$$

where $p_1$ and $p_2$ are the predict possibility from left-to-right decoder and right-to-right decoder respectively. $I$ denotes the input image, $y_l$ indicates the *l-th* character of its text groundtruth whose length is $L$.

And $L_o$ is binary cross entropy loss, whose labels are "horizontal" and "vertical". It can be described as follows:

$$L_o = -(ylog(p) + (1-y)log(1-p)) \tag{14}$$

where $y$ indicates the true label of input text and $p$ indicates the output of OSN. We set $\lambda = 1$ in our experiments.


**Fig. 4.** Examples of recognition results of vertical texts. On one hand, the single model with ASN achieves much better results than the model without ASN. On the other hand, using "two models" method, we tend to select the vertical model via their aspect ratios information, which may return the totally wrong results.

## 4 Experiments

In this section we describe extensive experiments conducted on various datasets.

### 4.1 Datasets Setup

**RCTW2017 and LSVT2019.** They are representative Chinese scene text datasets [8,23], most images of which are natural images collected by phone cameras. In LSVT2019, we only use the fully annotated part. Because annotations of the test set are not public, so we random split the training set into two subsets following the ratio 3:1, and segment text strings from the original images based on annotations. Texts which are marked by the "difficult" flag are excluded.

It's worth noting that the orientation labels of public datasets are easy to generate. A simple approach is based on their aspect ratio, which is the most commonly used to distinguish the horizontally and vertically texts when training two independent models. Obviously, it can't handle the abnormal cases as Fig. 1 in our paper. The other approach is using our orientation-classification model pretrained on our dataset. It has been proved that the accuracy of direction discrimination can reach 98% on real scene texts.

**SCTD.** Compared with various English public datasets, existing Chinese scene text datasets do not provide sufficient data for deep network training. To overcome the problem, we generate a synthetic Chinese scene text dataset (SCTD) for training and evaluating the performance of the proposed method. It comprises half of the vertical texts and half of the horizontal texts. We adapt the framework proposed by Gupta et al. [6] to a multi-oriented setup.

Different from [6], our task does not involve text detection, thus we adopt a faster and simpler approach to establish our dataset. We generate bounding boxes randomly on the scene pictures and calculate the standard deviation of the

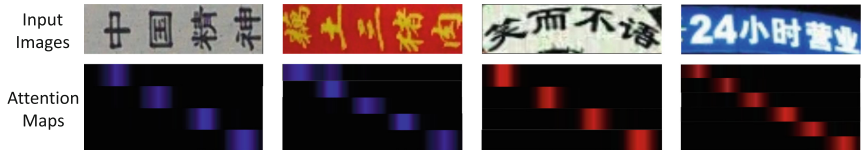



**Fig. 5.** Attention maps of arbitrarily-aligned texts. The maps in red color denote horizontal attention while maps in blue color denote vertical attention. (Color figure online)

**Table 3.** Performance of the ASN

| Methods | PM on SCTD | PM on RCTW2017 | PM on LSVT2019 |
|---|---|---|---|
| Two models | 70.4 | 53.1 | 51.9 |
| OSTER (without ASN) | 62.3 | 50.2 | 48.7 |
| OSTER (with ASN) | 69.5 | 52.5 | 51.3 |

Lab values of the regions, and select the regions whose standard deviation are less than the threshold for clipping (smaller standard deviation means the flatter color distribution). To increase sample diversity, we random add degrading strategies such as rotating, perspective transformation, brightness transformation. It includes 5 million training word images and 1,0000 testing word images. 5990 classes of Chinese characters are covered.

### 4.2 Implementation Details

Details about OSTER architecture are given in Tables 1 and 2 respectively. The number of hidden units of BLSTM in the decoder is 256. The recognition model outputs 5991 classes, including Chinese characters, digits, symbols, and a flag standing for "EOS". For vertical text images, we rotate them by 90°. After that, all images are resized to $32 \times 100$.

We adopt ADADELTA as the optimizer. The model is trained by batches of 64 examples. We set the learning rate to 1.0 at the beginning and decrease it to 0.1 after its convergence. We implement our method under the framework of PyTorch. Our model is GPU-accelerated with an NVIDIA GTX-1080Ti GPU, and CUDA 8.0 and CuDNN v7 backends are used.

### 4.3 Performance of the ASN

This part of the experiment is based on Chinese dataset. The OSTER is firstly trained on SCTD, then fine-tuned on the training subset of RCTW2017 and LSVT2019. We take perfect matching (PM) as the evaluation metric, which means the recognition results must be totally the same as the GT characters.

The accuracy of orientation discriminant can reach 99.8% on SCTD while 99.2% on RCTW2017 and 98.9% on LSVT2019. The recognition results are shown as Table 3. Compared with training all texts as a whole, the ASN improves

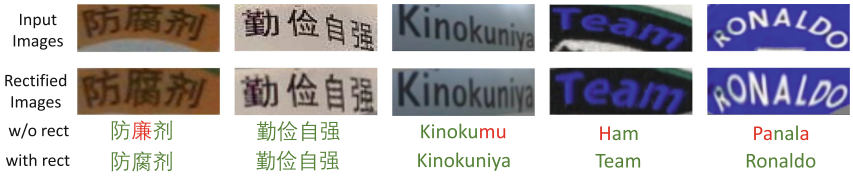



**Fig. 6.** Visualization of the source images, the rectified images and the recognition results. The results of rectified images is much better than that of source images.

**Table 4.** Performance of the CRN

| Methods | SCTD | RCTW2017 | LSVT2019 | IC15 | CUTE | IIIT5K | SVT | SVTP |
|---|---|---|---|---|---|---|---|---|
| Without CRN | 65.7 | 49.4 | 48.7 | 74.2 | 76.2 | 91.2 | 84.7 | 74.2 |
| With CRN | 69.5 | 52.5 | 51.3 | 77.1 | 81.8 | 92.2 | 89.2 | 80.4 |

the recognition accuracy greatly. Compared with training two models separately, the ASN achieves similar performance with only a single model, which is a more general and simple approach. Figure 4 shows an example where OSTER can deal with arbitrarily-aligned texts easily while severe mistakes are made by the separately trained models.

Examples of attention heat maps when decoding individual characters are visualized in Fig. 5. The maps in red color denote horizontal attention while maps in blue color denote vertical attention, and they are quite different from each other. Although learned in a weakly supervised manner, the attention module can still localize characters being decoded precisely.

### 4.4 Performances of the CRN

We visualize the outcomes of the CRN to verify the ability of its distortion correction. We perform this part of the experiment on both the Chinese datasets and the English datasets. The test English datasets include ICDAR2015 [12],

**Table 5.** Comparison on public Chinese benchmarks

| Methods | RCTW2017 | | LSVT2019 | |
|---|---|---|---|---|
| | MED score | PM score | MED score | PM score |
| Shi et al. [20] | 0.338 | 45.1 | 0.357 | 41.7 |
| Lee et al. [13] | 0.341 | 44.3 | 0.349 | 42.3 |
| Wang et al. [25] | 0.344 | 43.5 | 0.366 | 40.3 |
| Luo et al. [16] | 0.296 | 51.3 | 0.318 | 48.9 |
| Shi et al. [20] | 0.292 | 51.8 | 0.315 | 49.3 |
| OSTER (without OSN) | 0.293 | 51.7 | 0.318 | 49.2 |
| OSTER (with OSN) | **0.270** | **53.3** | **0.296** | **50.9** |

CUTE [18], IIIT5k-Words [17] and SVT [26]. As Fig. 6 shows, whether tilted, curved or partially distorted images can be effectively corrected. The recognition results are also improved dramatically.

Table 4 lists the results of the two variants. As can be seen, the model with rectification outperforms the one without on all datasets, particularly on SVTP (+6.2%) and CUTE (+5.6%). Since these two datasets both consist of irregular text, the rectification shows a significant effect.

### 4.5 Performances on Public Datasets

To prove the excellent performance of OSTER, We implement different open-source algorithms of text recognition on Chinese benchmarks. We keep the same aspect ratio as other methods ($32 \times 100$) for fair comparison, rather than resizing the Chinese texts to a larger aspect ratio ($32 \times 160$ for example) for higher accuracy. All images of the test subset of RCTW2017 are used.

The Mean Edit Distance (MED) [7] is also adopted as the evaluation metric. Lower MED and higher PM mean better performance. Table 5 lists the results of these methods. As can be seen, the OSTER outperforms the traditional methods on Chinese benchmarks either on MED or on PM metric.

To prove the generalization ability of OSTER, we conducted extensive experiments on various English benchmarks, including regular and irregular datasets. The training data consists of 8-million synthetic images released by Jaderberg [9] and 6-million synthetic images released by Gupta [6]. No extra data is used.

**Table 6.** Comparison on public English datasets

| Method | IC13 | IC15 | CUTE | IIIT5k | SVT | SVTP |
|---|---|---|---|---|---|---|
| Jaderberg et al. [11] | 81.8 | - | - | - | 71.7 | - |
| Rodriguez et al. [19] | - | - | - | - | 70.0 | - |
| Lee et al. [13] | 90.0 | - | - | 78.4 | 80.7 | - |
| Shi et al. [20] | 86.7 | - | - | 78.2 | 80.8 | - |
| Shi et al. [21] | 88.6 | - | 59.2 | 81.9 | 81.9 | 71.8 |
| Cheng et al. [3] | 89.4 | 66.2 | 63.9 | 83.7 | 82.2 | 71.5 |
| Yang et al. [28] | - | - | 69.3 | - | - | 75.8 |
| Cheng et al. [4] | - | 68.2 | 76.8 | 87.0 | 82.8 | 73.0 |
| Bai et al. [1] | 94.4 | - | - | 88.3 | 87.5 | - |
| Liao et al. [14] | 91.4 | - | 78.1 | 92.0 | 82.1 | - |
| Luo et al. [16] | **92.4** | 68.8 | 77.4 | 91.2 | 88.3 | 76.1 |
| Shi et al. [20] | 91.8 | 76.1 | 79.5 | 93.4 | **93.6** | 78.5 |
| Zhan et al. [29] | 91.3 | 76.9 | 83.3 | 93.3 | 90.2 | 79.6 |
| Wang et al. [27] | 91.3 | 74.0 | **85.1** | 93.3 | 88.1 | 80.2 |
| OSTER (ours) | **92.4** | **77.1** | 81.8 | 92.2 | 89.2 | **80.4** |

Because English datasets have rarely vertical texts, we set the orientation to be "horizontal" for every image. We only test lexicon-free case for SVT and IIIT5k. The training details are the same as training SCTD. Table 6 shows that OSTER achieves state-of-the-art performance over a number of public datasets.

## 5 Conclusion

In this paper, we propose an orientation sensitive scene text recognition network that is capable of recognizing horizontal and vertical scene texts simultaneously. The centerline rectification network corrects the distorted text to a regular one, the orientation sensitive network perceives the arrangement mode of input images, the attention selective recognition network generate corresponding attention map at each time step for better recognition results. The proposed network can be trained end-to-end and is robust to parameter initialization. Experiments over a number of public datasets demonstrate its superior performance in arbitrarily-aligned scene text recognition. The SCTD we established will be released to the public.

**Acknowledgment.** This work is supported by the Key Programs of the Chinese Academy of Sciences under Grant No. ZDBS-SSWJSC003, No. ZDBS-SSW-JSC004, and No. ZDBS-SSWJSC005, and the National Natural Science Foundation of China (NSFC) under Grant No. 61601462, No. 61531019, and No. 71621002.

## References

1. Bai, F., Cheng, Z., Niu, Y., Pu, S., Zhou, S.: Edit probability for scene text recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1508–1516 (2018)
2. Bissacco, A., Cummins, M., Netzer, Y., Neven, H.: PhotoOCR: reading text in uncontrolled conditions. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 785–792 (2013)
3. Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., Zhou, S.: Focusing attention: towards accurate text recognition in natural images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5076–5084 (2017)
4. Cheng, Z., Xu, Y., Bai, F., Niu, Y., Pu, S., Zhou, S.: AON: towards arbitrarily-oriented text recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5571–5579 (2018)
5. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. **18**(5–6), 602–610 (2005)
6. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2315–2324 (2016)
7. He, M., et al.: ICPR 2018 contest on robust reading for multi-type web images. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 7–12. IEEE (2018)

8. ICDAR2019: ICDAR 2019 robust reading challenge on large-scale street view text with partial labeling. https://rrc.cvc.uab.es/?ch=16. Accessed 20 Apr 2019
9. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Synthetic data and artificial neural networks for natural scene text recognition. arXiv preprint arXiv:1406.2227 (2014)
10. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems, pp. 2017–2025 (2015)
11. Jaderberg, M., Vedaldi, A., Zisserman, A.: Deep features for text spotting. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 512–528. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_34
12. Karatzas, D., et al.: ICDAR 2015 competition on robust reading. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 1156–1160. IEEE (2015)
13. Lee, C.Y., Osindero, S.: Recursive recurrent nets with attention modeling for OCR in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2231–2239 (2016)
14. Liao, M., et al.: Scene text recognition from two-dimensional perspective. arXiv preprint arXiv:1809.06508 (2018)
15. Liu, Z., Li, Y., Ren, F., Yu, H.: A binary convolutional encoder-decoder network for real-time natural scene text processing. arXiv preprint arXiv:1612.03630 (2016)
16. Luo, C., Jin, L., Sun, Z.: MORAN: a multi-object rectified attention network for scene text recognition. Pattern Recogn. **90**, 109–118 (2019)
17. Mishra, A., Alahari, K., Jawahar, C.: Scene text recognition using higher order language priors. In: BMVC-British Machine Vision Conference. BMVA (2012)
18. Risnumawan, A., Shivakumara, P., Chan, C.S., Tan, C.L.: A robust arbitrary text detection system for natural scene images. Expert Syst. Appl. **41**(18), 8027–8048 (2014)
19. Rodriguez-Serrano, J.A., Gordo, A., Perronnin, F.: Label embedding: a frugal baseline for text recognition. Int. J. Comput. Vis. **113**(3), 193–207 (2015)
20. 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. IEEE Trans. Pattern Anal. Mach. Intell. **39**(11), 2298–2304 (2017)
21. Shi, B., Wang, X., Lyu, P., Yao, C., Bai, X.: Robust scene text recognition with automatic rectification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4168–4176 (2016)
22. Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., Bai, X.: ASTER: an attentional scene text recognizer with flexible rectification. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 2035–2048 (2018)
23. Shi, B., et al.: ICDAR 2017 competition on reading Chinese text in the wild (RCTW-17). In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 1429–1434. IEEE (2017)
24. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, pp. 3104–3112 (2014)
25. Wang, J., Hu, X.: Gated recurrent convolution neural network for OCR. In: Advances in Neural Information Processing Systems, pp. 335–344 (2017)
26. Wang, K., Babenko, B., Belongie, S.: End-to-end scene text recognition. In: 2011 International Conference on Computer Vision, pp. 1457–1464. IEEE (2011)

27. Wang, P., Yang, L., Li, H., Deng, Y., Shen, C., Zhang, Y.: A simple and robust convolutional-attention network for irregular text recognition. arXiv preprint arXiv:1904.01375 (2019)
28. Yang, X., He, D., Zhou, Z., Kifer, D., Giles, C.L.: Learning to read irregular text with attention mechanisms. In: IJCAI, pp. 3280–3286 (2017)
29. Zhan, F., Lu, S.: ESIR: end-to-end scene text recognition via iterative image rectification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2059–2068 (2019)