Research on Natural Scene Text Detection and Recognition

Jieming Wang¹, Wei Wu²,*

¹Department of information Engineering, Wuhan University of Technology, Wuhan, Hubei Province, 430070, China
²Department of information Engineering, Wuhan University of Technology, Wuhan, Hubei Province, 430070, China
*Corresponding author’s e-mail: wuwei@whut.edu.cn

Abstract. In recent years, as the importance of scene text detection and recognition in real life has gradually become prominent, its research methods have become mature, and more attention has been placed on improving accuracy. Due to the complex background in natural scenes, different font arrangements, lighting, etc., various methods are difficult to meet the requirements of the application. In this paper, the detection and recognition of curved text are studied, and the research method based on the ESIR framework [1] is improved, specifically in the following two aspects: (1) TPS transform fiducial points are predicted by the positioning network, and the positioning network uses CNN to return to the required fiducial points [2] The x and y coordinates. (2) Using ACE (aggregated cross entropy) method for sequence recognition, its loss function is realized faster and it occupies less memory than CTC and attention mechanism, and it performs well in the one-dimensional prediction of scene text recognition. In this paper, the irregular data set is sampled by reference points, and the parameter matrix T of TPS is calculated to transform the points in the picture, and cyclically transform until the image text is level, and then use the CNN-BLSTM model to perform one-dimensional prediction and output the recognition results. The selected data set is irregular text data set: for example, SVT-Perspective, CUTE80.

1. Introduction
The text image of the natural scene contains a wealth of information, and the text information can help us understand the content of the image. Scene text recognition is to convert the text information in the image into a text sequence and output it under the condition of a complex background and diverse fonts. The emergence of this technology also caters to some emerging applications in recent years, the most common is to identify various street signs in the street view.

In recent years, there has been a lot of research in this area, and there have been major breakthroughs in horizontal and general oblique text, but the accuracy of curved and irregular text recognition still needs to be improved.

There are roughly four types of traditional text detection methods: based on edge features, based on connected regions, based on texture features, and based on machine learning deep learning methods. Among the traditional text recognition methods, the more popular one is to apply the CNN-BLSTM-CTC model or attention mechanism to solve the problem. The detection and recognition methods used in this article will be detailed later.

Since the research object of this paper is curved text in natural scenes, it is expected that the irregular shape text in the original image can be converted into horizontal regular text before...
recognition. This deformation process requires the use of Spatial Transformer Networks, which requires the use of Thin Plate Spline (TPS transformation). On the whole, the entire network is composed of two parts, one is the ESIR model with improved detection method to detect the shape of text and correcting the text, and the other is the sequence recognition network that uses the ACE method to recognize the corrected image.

2. Materials and Methods

2.1. STN Network

![Spatial transformation network](image1)

As shown by the above structure, the input of STN is U and the output is V. The network is divided into three steps: (1) Define a positioning network, whose input is U, and output variable parameters θ. This parameter is used to map the coordinate relationship between U and V. (2) According to the coordinate points and change parameters in V in the grid generator θ to calculate the coordinate points in U. This is because the size of V is defined by myself. Of course, all coordinate points of V can be obtained. When filling the pixel value of each coordinate point in V, it must be taken from U, so according to each coordinate in V Point and change parameters θ are calculated to get a coordinate. In the sampler, the pixel value is found in U according to this coordinate, and V is filled in this way. (3) What needs to be done in the sampler is to fill V, according to the series of coordinates obtained by the grid generator and the original image U (because the pixel value needs to be taken from U), because the calculated coordinates may be decimals. Use another method to fill, such as bilinear interpolation.

2.1.1. Location Network

In the scene text localization network, we use CNN to perform regression instead of classification. Similar to the traditional convolutional neural network structure, it is composed of a convolutional layer, a pooling layer, and a fully connected layer. In the classification task, the classification probability of the Softmax layer output prediction is usually added at the end. When used in the regression task, the regression analysis is used on the neural network, which is an approximate prediction of the true value. The final output layer sets the number of output nodes to 2K. Due to the insufficient expression of the linear model, it is necessary to use the activation function to add non-linear factors. Since the tanh function is more superior than the sigmoid function and is better in practical applications, the activation function is set to tanh (\(\gamma\)) to make the output vector All are between (-1,1).

The entire positioning network locates the position information of the reference point by returning to their x, y coordinates, where the Kth coordinate can be expressed as \((x_k, y_k)\). Among them is a normalized coordinate system constructed with the image center as the origin, so the value range of \((x_k, y_k)\) is (-1, 1). The final coordinate set can be expressed as:

\[
c = [c_1, c_2, ..., c_K] \in \mathbb{R}^{2 \times K}
\]

Among them, the kth column \(c_k = [x_k, y_k]^T\) is the coordinate of the Kth datum point.
2.1.2. TPS Transformation

TPS (Thin Plate Spline), namely thin plate spline interpolation[3]. Among them, the interpolation function refers to an expression that approximates the true function by a set of observations. The observations are represented by a set of coordinate points \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\) means. The functional form of TPS is expressed as follows:

\[
\phi(x) = c + a^T + \omega^T s(x) \\
s(x) = \left(\sigma(x - x_1), \sigma(x - x_2), \ldots, \sigma(x - x_n)\right)^T \\
\sigma(x) = \|x_i\|^2 \log \|x_i\|^2
\]

Among them, \(c \in R^{1 \times 1}, a \in R^{D \times 1}, w \in R^{N \times 1}\) except for the above boundary End point coordinates, also need to define another set of basic datum points expressed as \(c' = [c'_1, c'_2, \ldots, c'_K] \in R^{2 \times K}\), these basic datum points are evenly distributed up and down along the text. Since K is a constant and the entire coordinate points are normalized, the transformed point set \(c'\) is also a constant.

\(P\) is the matrix formed by the reference point set of the original image, and \(P'\) is the matrix formed by the reference points of the transformed image. Then the relationship between the reference point of the curved image sample and the defined basic reference point after transformation can be expressed as: \(P' = T \cdot P\). The thin-plate spline interpolation function is used to fit the approximate transformation. The value of the pixel \(P'\) is bi-linearly interpolated from the pixel \(P\) close to the distorted scene text image. This is the initial process of using TPS to convert curved samples.

2.2. Iterative Transformation

Figure 2. Perform N iterations to correct the original image through TPS transformation.

As shown in the flowchart, the initial curved text is first confirmed by the positioning network and the reference point is marked, and the corrected text image is obtained through the calculated TPS transformation parameter T. If the transformation result is still not ideal, repeat the above process, N is the number of pre-iterations. In the process of each iteration, due to the existence of edge effects, the TPS transformation is added to the original input image during each iteration, and the corrected image is used to predict control points. The final basically horizontal corrected image will be sent to the sequence recognition network for scene text recognition.
2.3 Sequence Recognition Network

2.3.1 CNN-BLSTM-CTC

The CNN-BLSTM-CTC structure is a mainstream text sequence recognition algorithm\textsuperscript{[4]}. In this architecture, first use CNN to encode the character level information of the above modified text information into its character level representation, and then we will The character level representation and the vocabulary level representation are connected, and they are placed in the BLSTM to model the context information of each vocabulary.

BLSTM or bidirectional LSTM (Long Short-term Memory Network) is an improved recurrent neural network, which can save the long-term state by adding a unit state c, which solves the problem of RNN's inability to handle long-distance dependence. Due to the original RNN The hidden layer has only one state h, which is very sensitive to short-term input.

This recognition network uses a sequence-to-sequence model with attention mechanism. It consists of an encoder and a decoder. In the encoder, the input is a rectified scene text image, and CNN performs a convolution operation, using multiple convolution kernels to convolve the matrix, and convolve to obtain convolution features; then convolve through pooling operations Part of the area of the accumulated features takes the average or maximum value to obtain the feature image of the convolution stage. In this step, the 53-layer residual network (ResNet) is used to extract the features. The residual network is followed by two layers with 256 The bidirectional long-term short-term memory (BLSTM) of the hidden unit, which inputs the feature image generated by the convolution stage into the BLSTM neural network in the form of time steps, and outputs a fixed-length vector. The decoder consists of a dual attention LSTM with 256 hidden units and 256 attention units.

However, the forward and backward algorithm of CTC\textsuperscript{[5]} is very complicated, resulting in a lot of computational consumption. Therefore, the method of replacing CTC with ACE is used in the article to realize sequence recognition more easily.

2.3.2 ACE loss function

This aggregate cross entropy loss function includes three stages: first, calculate the probability distribution of each category along the time dimension; then normalize the cumulative result and label to the probability distribution of all categories; finally use the cross entropy to compare the two probability distributions. This method can infer and back-propagate faster than CTC, and has less memory requirements\textsuperscript{[6]}.

This loss function doesn't require character order information, only the classes and their number used for supervision. If a class (that is, the letter type of the letter sequence in the image) appears n times in the annotation, it is required that the cumulative predicted probability on the T time step is also n. In order to minimize the general loss function, the network can be required to accurately mark the number of characters contained in each class:

\[
L(\omega) = - \sum_{(I,S) \in Q} \sum_{t=1}^{L} \log P(S|I; \omega) \approx - \sum_{(I,S) \in Q} \sum_{t=1}^{L} \log P(N_k|I; \omega) \quad (5)
\]

Where \( L(\omega) \) is given the input image I from the training set Q and its sequence label S, the general loss function evaluation of the sequence recognition problem is the length L of the image I as the condition under the model parameter \( \omega \) Label the probability of S. \( N_k \) represents the number of times the character appears in the sequence label S.

Due to the inconvenience of regression-based ACE loss\textsuperscript{[7]} in backpropagation, the gradient will be scaled to a very small size by \( y'_k \) and \( y''_k \), resulting in a gradient The problem disappeared. In order to prevent the disappearance of the gradient, the influence of the term \( y'_k \) introduced by the Softmax function needs to be offset. In information theory, cross entropy is used to measure the distance between two probability distributions. Accordingly, the cumulative probability of \( y'_k \) for the k-th
character can be standardized as $y_k = y_k / T$, and the number of occurrences of characters $N_k$ can be standardized as $N_k = N_k / T$. Finally, calculate the cross entropy between $\bar{y}$ and $\bar{N}$.

3. Results & Discussion

3.1. Data
The improved model is trained using Synth90K and SynthText datasets, among which Synth90K is a horizontal English scene text dataset, consisting of approximately 90k images; SynthText dataset is mainly used for text detection in natural scenes, the dataset consists of 800,000 Image composition, each text instance is annotated with text string, word-level and character-level bounding boxes. The trained model evaluates 6 public data sets, including 3 common data sets ICDAR2013, IIIT5K and SVT, most of which are almost horizontal, and 3 distortion data sets ICDAR2015, SVTP and CUTE80, among which a large number The scene text samples all have irregularities.

3.2. Training Details
In this experiment, the proposed scene text recognition network uses the Tensorflow framework. In formal training, we use ADAM as the optimizer. The network training is performed on GTX1080-gpu. The network is trained in 1 million iterations with a batch size of 64. In the course of the experiment, we tried to increase the rectification iteration number $N$, we found that the model performance improves consistently when a larger number of rectification iterations is implemented. Usually, when the number $N$ is set to 5, which enables the best visual image restoration.

3.3. Test of our Network
In the experiment, as shown in Table 1, we compared the improved model with the recognition accuracy of the RARE model and the ESIR model on the classic data set. The use of the ACE loss function makes the recognition accuracy on the data set It has been improved. And as shown in Figure 3-5, our experimental method also makes it possible to obtain better corrected images for easy recognition.

Table 1. Comparison of recognition accuracy of various recognition models.

| Methods   | ICDAR2013 | ICDAR2015 | IIIT5K | SVT | CUTE80 |
|-----------|------------|------------|--------|-----|--------|
|           | None       | None       | 50     | 1k  | None   |
| RARE      | 88.6       | -          | 96.2   | 93.8| 81.9   | 95.5  | 81.9 | -    |
| ESIR[ResNet,SK] | 89.1     | 70.1       | 97.8   | 96.1| 82.9   | 97.1  | 85.9 | 72.1 |
| Ours      | 91.2       | 78.1       | 97.4   | 97.3| 85.7   | 98.2  | 87.7 | 82.9 |

Figure 3. Irregular scene text image selected from dataset CUTE80.
4. Conclusions
In this article, the method of locating the network prediction reference point is improved and combined with the iterative ideas provided in the ESIR model, so that irregular text can be corrected better. In the text recognition process, by combining the new ACE loss function to improve the efficiency and accuracy of the traditional recognition network, the improved detection and recognition network is more efficient, and the recognition accuracy has been improved.

Acknowledgments
This article is one of the phased results of the research on the visual perception enhancement method of unmanned surface ships in the harsh navigable environment, a project of the National Natural Science Foundation of China.

References
[1] Zhan, F., & Lu, S.. (2020). ESIR: End-To-End Scene Text Recognition via Iterative Image Rectification. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.
[2] Shi, B., Wang, X., Lyu, P., Yao, C., & Bai, X.. (2016). Robust scene text recognition with automatic rectification.
[3] Donato, G., & Belongie, S.. (2002). Approximate Thin Plate Spline Mappings. Proceedings of the 7th European Conference on Computer Vision-Part III. Springer, Berlin, Heidelberg.
[4] Liao M, Wan Z, Yao C, et al.(2019) Real-time Scene Text Detection with Differentiable Binarization.
[5] Ch'Ng C K, Chan C S. (2017) Total-Text: A Comprehensive Dataset for Scene Text Detection and Recognition.2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE.
[6] Wang, K., Babenko, B., & Belongie, S.. (2012). End-to-end scene text recognition. IEEE International Conference on Computer Vision. IEEE.
[7] Xie, Z., Huang, Y., Zhu, Y., Jin, L., Liu, Y., & Xie, L.. (2020). Aggregation Cross-Entropy for Sequence Recognition. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.