Transactions on Computational and Scientific Methods | Vol. 5, No. 1, 2025
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Precision Recognition of Irregular Scene Text Leveraging Advanced Attention Mechanisms

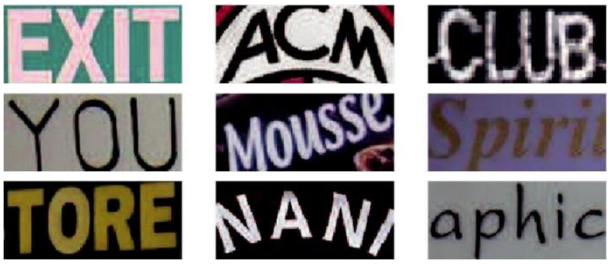
Averick Holleran
School of Computer and Data Sciences, Oregon State University, Corvallis, USA
averick84@osu.edu

Abstract: Scene text recognition has emerged as a crucial research area due to its wide-ranging applications in domains such as traffic sign recognition, driverless cars, and product packaging. Recognizing irregular scene text—characterized by curved, distorted, or low-resolution features—presents significant challenges for existing methods. This paper introduces a novel Multi-Scale Feature Fusion Attention Recognition Network (MSFARN) to address these challenges effectively. MSFARN comprises two core components: a Multi-Scale Feature Fusion Network (MSFN) for text rectification and an Attention-Based Recognition Network (ARN) for precise text recognition. MSFN applies multi-scale feature extraction and fusion to rectify irregular text images, enhancing readability. ARN combines channel and spatial attention mechanisms to focus on critical feature regions, ensuring robust recognition performance. Extensive experiments conducted on multiple datasets, including IIIT5K, ICDAR2003, ICDAR2013, ICDAR2015, SVT-Perspective, and CUTE80, validate the framework's superiority. The results demonstrate MSFARN's ability to achieve state-of-the-art recognition accuracy. Future research will explore text detection and recognition in more complex scenes and further generalize the approach to any font type and orientation.

Keywords: Scene Text Recognition; Deep Learning; Irregular Text; Image Processing.

1. Introduction

In recent years, due to the increasing use of text in natural images, scene text recognition has gradually become a research hotspot in academia and industry. Scene text is also practically used in traffic sign recognition Bulan et al (2017), driverless cars Zhu et al (2016), product packaging recognition and image search Jegou et al (2008) and other fields, which makes scene text recognition an open and challenging research topic.



(a) (b) (c)

Fig 1. Examples of regular and irregular scene text. Where (a) is regular text, (b),(c) are irregular text

Nowadays, regular text recognition Shi et al (2016a); Su and Lu (2017); Wang et al (2012) has achieved excellent results. In addition, methods based on convolutional neural networks Wang et al

(2012); Bissacco et al (2013); Jaderberg et al (2016)have been widely used. Many researchers have introduced recurrent neural networks Shi et al (2016a); He et al (2016) and attention mechanisms Liu et al (2016); Cheng et al (2017, 2018); Li et al (2019); Lee et al (2020) into recognition models to produce better performance. However, most of the current models are not stable enough to deal with various disturbance factors in natural images in time. Therefore, recognizing irregular text with various shapes or distorted fonts in images will be a major challenge. As shown in Figure 1, the irregular scene text contains various perspective or curved characters, which are still difficult to identify.

Therefore, this paper proposes a multi-scale feature fusion attention recog-nition network MSFARN, which can read rotated or curved or low-resolution irregular characters in scene text. MSFARN consists of a multi-scale feature fusion network (MSFN) for rectifying text in images and an attention-based recognition network (ARN) for recognizing rectified text. The recognition task consists of two parts. First, MSFN is used as a spatial transformer to rectify the irregular text in the image. As shown in Figure 2, the inclined and curved text in the image becomes regular and easier to read after rectification. Second, the rectified image is entered into the ARN to output the predicted characters.

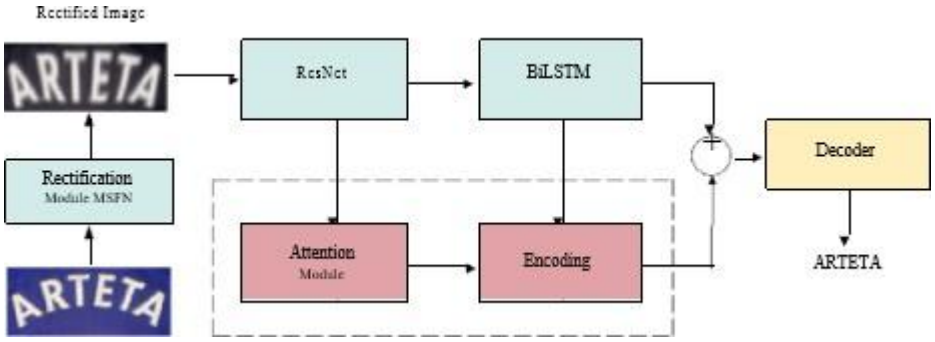


Fig 2. Overview of MSFARN

The current text rectification uses a multi-scale feature fusion network to rectify irregular text. The training of MSFN is guided by the ARN, only text labels are required, and the MSFN is trained in a weakly supervised form. The overall process of MSFN is shown in Figure 3 below. Since each pixel in the image has its own position coordinates, a coordinate is initialized first, MSFN learns the offset position through training, and then samples its pixel values to rectify the image. Finally, the rectified image is sent to ARN for recognition. Since there are many curved, inclined or blurred texts in irregular text images, in the training network of MSFN, this paper focuses on the multi-scale feature module method to increase the different scales of the extracted features, and fuse them to obtain more features, which is convenient for feature extraction of more complex irregular characters. In the recognition network, a method that integrates the channel attention mechanism and the spatial attention mechanism is adopted to obtain the precise alignment between the feature region and the target. Due to the enhanced attention, ARN is more robust to the variation of context. In summary, the research contributions of this paper are as follows:

This paper proposes the MSFARN framework to recognize irregular scene text. The framework is trained in a weakly supervised form, only text labels are provided, and no extensive annotations are required. The framework contains two processes of rectification and recognition.

This paper proposes a multi-scale feature fusion network MSFN for rectify-ing irregular scene text. The network uses a multi-scale feature module to extract multi-directional information and fuses it into the recognition net-work. Irregular text images rectified with MSFN are easier for subsequent recognition.

This paper proposes an attention-based recognition network ARN to recog-nize irregular texts. This mechanism combines channel attention and spatial attention, effectively concentrating computing power on the most mean-ingful and informative parts, and significantly improving the recognition

performance.

2. Related Work

2.1 Text Recognition

In recent years, due to the rapid development of neural networks Zhu et al (2016); Shi et al (2016a); He et al (2016); Ma et al (2017); Santos et al (2021), the ability of regular text recognition has been greatly improved. ReferencesYe and Doermann (2014); Zhu et al (2016) reviews the detection and recognition of scene text in recent years. Relative to hand-crafted features such as connected components Neumann and Matas (2012), semi-Markov conditional random fields Seok and Kim (2015), and generative shape models Lou et al (2016), etc., features extracted by neural networks have dominated. For example, Yin et al. Yin et al (2017) proposed methods for unconstrained recognition based on neural network structure.Lee et al. Lee et al (2020) proposed a method for 2D attention.

With the wide application of RNN(Recurrent Neural Network), methods based on the combination of CNN(Convolutional Neural Network) and RNN can better extract text information. The CRNN(Convolutional Recurrent Neu-ral Network) proposed by Shi et al. Shi et al (2016a) used CNN to extract the features in the pictures, and then sent them to the RNN for sequence predic-tion, and finally obtained the final result through the translation of the CTC loss function. CRNN transformed the text recognition problem into a sequence learning problem, and the recognition performance was significantly improved. Furthermore, the attention-based recursive recurrent neural network proposed by Liu et al. Liu et al (2016) paid more attention to information-rich regions for good text recognition performance. Ye et al. Ye and Doermann (2014) proposed a two-dimensional attention mechanism. These methods first encode the input image into a sequence representation, and then output predictions through a decoder, which usually consists of two techniques: the connectionist temporal classification (CTC) Shi et al (2016a); He et al (2016) and attention mechanism Lee and Osindero (2016); Cheng et al (2017). However, attention methods may lead to inconsistent sequence and target text mappings. To solve the problem of attention drift,Cheng et al. Cheng et al (2017) used Focused Attention Network (FAN) to deal with the problem of attention mechanism transfer and achieve more accurate localization. After this, CNN and RNN were integrated for text recognition, such as Shi et al. Shi et al (2016a) and He et al. He et al (2016) extended this field with gated recurrent CNN. Later, Lee et al. Lee et al (2020) used the method of self-attention to recognize text. Li et al. Li et al (2022) also used the dual relation module to extract text features.

The approach we propose is an attention-based sequence-to-sequence model Bahdanau et al (2014); Chorowski et al (2015). This model was originally proposed for machine translation and speech recognition, predicts an output sequence from an input sequence. This work further extends the bidirectional decoder.

2.2 Text Rectification

In recent years, irregular texts exist in more and more images, and the curved, slanted and blurred texts limit people's ability to recognize texts, which makes it more difficult to recognize irregular texts. Researchers have proposed some methods to solve irregular text recognition. One is a bottom-up approach Yang et al (2017); Cheng et al (2018), which searches for the position of each character and then concatenates them together. For example, Yang et al. Yang et al (2017) proposed a two-dimensional attention mechanism to address irregular text. Based on character-level annotations, the model can accurately extract text characters while ignoring redundant background. Li et al. Li et al (2019) utilized a simpler 2D attention mechanism and achieved better performance without character-level annotations.

The other is a top-down approach Liu et al (2016); Shi et al (2016b), which recognizes text directly from the input image without detecting and recogniz-ing individual characters, reducing the recognition difficulty. For example, Liu et al. Liu et al (2016) transformed rotated irregular text into

regular, easily recognizable text by using an affine transformation network. Later, the ResNet network gradually entered people's field of vision. This network can not only use forward propagation to obtain feature reuse, but also alleviate the prob-lem of gradient signal disappearance during backward propagation, making it easier to extract image features. But the deeper the network, the longer the training time and the more difficult the training. Later, Jaderberg et al. Jaderberg et al (2014a) utilized the CNN-RNN framework to encode the input image into four orientations, and designed an orientation attention mecha-nism to ablate these four directional features, which can represent scene text in arbitrary orientations. Later, Li et al. Li et al (2021) proposed a method for character-aware sampling and correction for scene text recognition. Wu et al. Wu et al (2022) proposed a two-level rectified attention network for scene text recognition to recognize irregular text. Although this method can rectify irregular characters in any direction to a certain extent, most scene characters in the real world will be affected by illumination and blur. At this time, it is particularly important to be able to accurately extract these irregular features and recognize them completely. Therefore, we propose a multi-scale feature fusion attention recognition network to recognize irregular scene text, which is not only effective for distorted, slanted text, but also can better recognize low-resolution scene text.

3. Methodology

The MSFARN proposed in this paper adopts a top-down approach and consists of two parts. One is MSFN, which trains the irregular text dataset in a weakly supervised form to learn the offsets of parts in irregular text images. The rectified image position is obtained from these offsets. The other is ARN, a CNN and CTC framework followed by an attention-mechanism decoder that finally predicts the characters in the image.

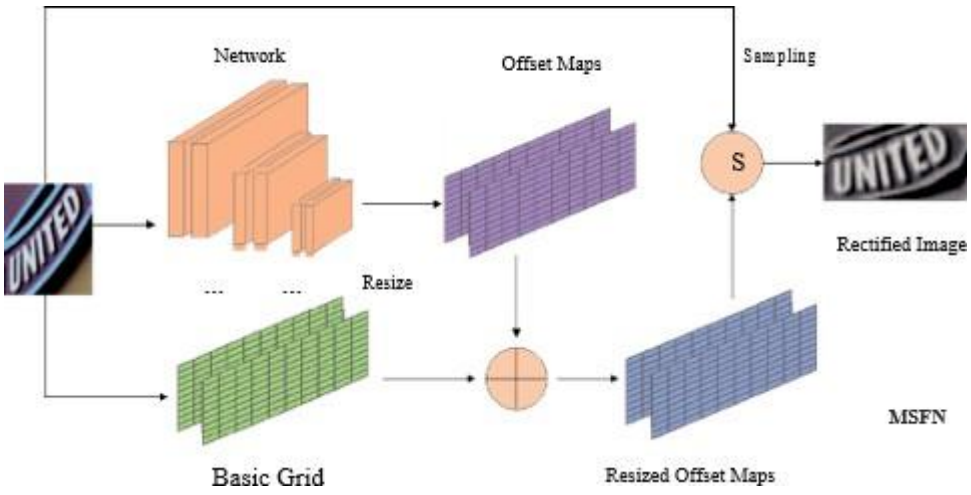


Fig 3. The overall process of MSFN

3.1 Multi-scale Feature Fusion Network

Commonly used rectification modes such as affine transformation methods are affected by factors other than rotation, translation and scaling. However, irreg-ular scene text can also be influenced by lighting, blurring or deformation. This makes recognition more difficult. Another approach is to utilize deformable convolution kernels in deformable convolutional neural networks to extract informative features. Although the above methods are used, the network is still failed to converge, and the recognition of irregular characters is greatly challenged. Therefore, this paper will further improve the accuracy and com-pleteness of recognition based on the ideas of rectification. As shown in Figure 3, the MSFN network structure can rectify the distorted image.

Since MSFN only predicts the position offset of characters, a pooling layer is firstly placed to reduce noise.The overall process of MSFN is shown in the Figure 3. MSFN first divides the input image with the input size of 32×100 into 3×11 parts and calculates the offset of each part. For faster convergence, use the Tanh() activation function and get values in the range (-1,1). The offset image consists of two

parts, the x-coordinate and the y-coordinate, and then the size of the offset image is smoothly modified to 32×100 using bilinear interpolation, which is consistent with the input image size.

3.2 Multi-scale Feature Module

As mentioned above, most of the recognition networks are based on the superposition and connection of multiple convolutional layers to form a deep convolutional neural network, which is then used to extract features from regular or irregular text images, and provide feature information of dif-ferent dimensions for recognition in subsequent networks.The stability and effectiveness of the extracted features directly determine the performance of recognition. However, it is more difficult to extract irregular text features than regular text. Including: edge feature, transformation feature, penetra-tion feature, grid feature, feature point feature, direction line feature and other features extraction. Based on the feature extraction network proposed in the past, this paper further improves the breadth and dimension of feature extraction. First, the network structure is changed from multiple convo-lutional neural networks in series to multiple parallel convolutional neural networks with multi-layer convolutional layers for feature extraction. The specific convolutional neural network structure is shown in Figure 4.

Secondly, In the original convolution layer, we can only use a fixed-size convolution kernel, now, we can use different size convolution kernels in differ-ent parallel networks for feature extraction. This parallel convolutional neural network structure is called is a multi-scale feature module. In this block, the number of channels is first reduced by 1×1 convolution to aggregate the infor-mation, so as to prevent some important features from being lost with the deepening of feature extraction during the convolution process. Then we use two parallel convolutional structures with a depth of 2 for feature extraction at different scales. Among them, using a 3×3 convolution kernel can better extract text features with smaller height and width in irregular text images, the edge features and feature point information of small text can be more fully extracted, and will not be ignored with the deepening of the convolution oper-ation. The purpose of using a 5×5 convolution kernel is to obtain richer spatial information, grid features and other features when extracting features from larger irregular texts. At the same time, 1×1 convolution is added in front of 3×3 convolution and 5×5 convolution, and after 3×3 pooling, which can aggregate information and effectively reduce the amount of parameters. In this way, In this way, each layer in the network can learn features with different characters, which increases the width of the network and the applicability of the network to scale. Finally, the nonlinear properties are obtained by stack-ing each multi-scale feature module and then fusing the features extracted by the parallel network in each block,which provides a more stable and efficient feature of irregular text for the rectifying part.

First, the position coordinates of the original pixels in the input image are set as a base grid, which is represented by x and y. This base grid normalizes the coordinates of each pixel to [-1, 1]. The coordinates of the upper left corner are (-1, -1), and the coordinates of the lower right corner are (1, 1). Sum the original offset and base grid as follows,

$$\text{offset}' = \text{offset}_{(c,i,j)} + \text{basic}_{(c,i,j)},\ c = 1,\ 2$$

where (i, j) represents the position coordinates of the i-th row and j-th column in the grid. c=1, 2 represent the x-coordinate and y-coordinate, respectively.

Before sampling, the x- and y-coordinates on the offset image are first normalized to [0,W][0,H], where H×W is the size of the input image. The pixel values of the rectified image I′ are as follows:

where I is the input image, i′ and j′ represent the values of c=1 and 2 in the above, and both are real numbers. Sample rectified image I′ using bilinear interpolation. Since equation (2) is differentiable, MSFN can backpropagate gradients.The training algorithm for MSFN is as follows Algorithm 1:

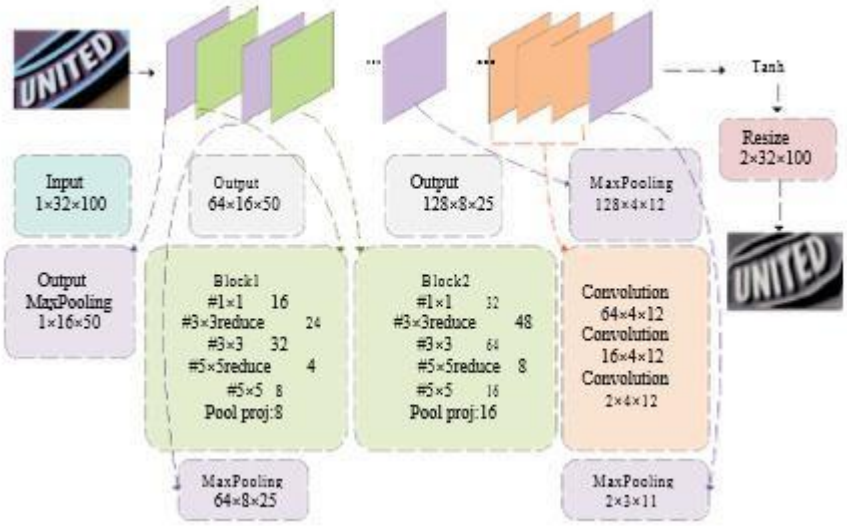


Fig 4. Multi-scale feature fusion convolutional neural network

3.3 Networks based on Attention Recognition

As shown in Figure 5, ARN is mainly composed of CNN-BLSTM-GRU frame-work. The encoder adopts the CNN-BLSTM framework. The decoder converts the feature sequence of the previous encoder into a character sequence. The decoder adopts the "focusing model" Bahdanau et al (2014), which is a one-way recurrent neural network, the maximum number of steps generated is T, and the output label is a letter or a special symbol "End-Of-Sequence" (EOS), used as a prompt message at the end of the sequence.

Specifically, at step t, the focus model calculates the attention weight, and this weight is denoted by $\alpha_{t,i}$:

$$\alpha_{t,i} = \exp(e_{t,i})/(\exp(e_{t,i}') = 1$$
$$e_{t,i} = \tanh(U\mathbf{R}_{T-1} + V\mathbf{s}_I + b)$$

where U , V are trainable weight vectors. The focus weights represent the importance of each item in the encoder sequence over the previous item. At this point, the focus vector W_t can be calculated:

$$y_{pre} = \text{Embedding}(y_{t-1})$$
$$W_t = \sum_{i=1}^{x} (\alpha_{t,i}, s_i)$$

Where s_i denotes the sequence feature vector and L denotes the feature map size. while y_{pre} refers to the embedding vector of the output y_{t-1} of the previous time period. R_t is specified as the hidden layer state at time step t. Instead, R_t is updated by the GRU and R_t is calculated as follows:

$$R_t = \text{GRU}\ (y_{pre},\ W_t,\ R_{t-1})$$

Table 1. Algorithm 1 MSFN training algorithm

Algorithm 1 MSFN training algorithm
Require: Img Grid=g_0.....g_{32} Offset=o(0,0),.....o(H,W)
Ensure: Img is the input image Grid is a grid of images divided into 3×11 parts Offset is a offset of each pixel taken by a grid for a part
1: if requirements are true then
2: Rectif ied$_1$ $\leftarrow$ $g_0 + o_{(i,j)}$, i $(0, \overline{11}^{W}), j$ $(0, \overline{3}^{H})$
3: for iteration=1,2,... do
4: for k=1,2,...,33 do
5: Recitf ied$_k$ $\leftarrow$ $g_k + o_{(i,j)}$, i $[\overline{11}^{W}(k-1), \overline{11}^{W}_{k}], j$ $[\overline{3}^{H}(k-1), \overline{3}^{H}_{k}]$
6: end for
7: Rectif ied $\leftarrow$ grid sample(Rectif ied$_{(1,2,...,33)}$)
8: end for
9: end if

At this time, when the time step is t, the output result y_t is:

$$y_t = Soft+max(\theta_{oRT}+b_o)$$

where θ_o , b_o are trainable parameters.

3.4 Attention Block

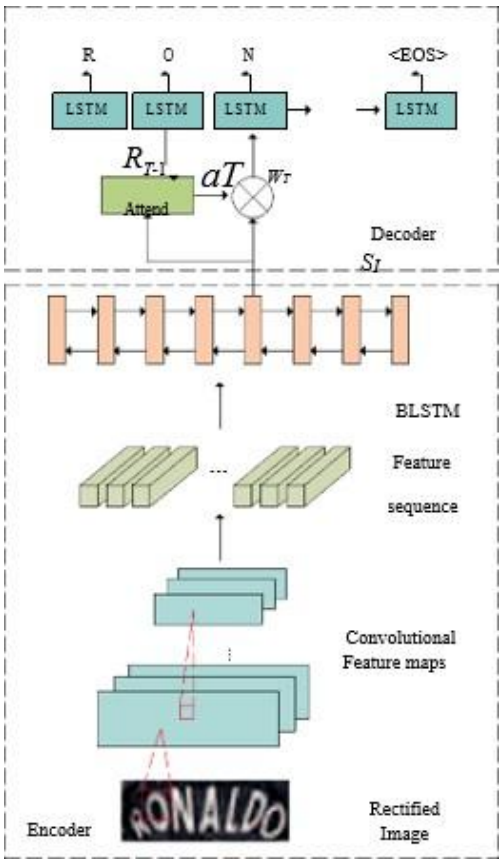


Fig 5. Detailed structure of ARN

After rectifying the irregular text by the rectification network MSFN, the dif-ficulty of the recognition part is reduced, but the information extracted by the deep convolutional neural network is rich and diverse, ordinary attention mechanism models cannot accurately align feature regions with target regions due to the complexity and low resolution of images.In order to focus on the information that is more critical to the current task in the redundant feature information , reduce the attention to other information and improve the accu-racy of recognition, so as to introduce an attention mechanism based on the combination of channel attention and spatial attention, which can solve the problem of information overload and improve the efficiency and accuracy of task processing.

The Channel Attention Module is to compress the feature map in the spa-tial dimension to obtain a one-dimensional vector and then operate. When compressing in the spatial dimension,we have to consider not only Average Pooling, but also Max Pooling. The Spatial Attention Module compresses the channel, and performs Average pooling and Max pooling respectively in the channel dimension. Average pooling and Max pooling can be used to aggre-gate the spatial information of the feature map, send it to a shared network, compress the spatial dimension of the input feature map, and sum and merge element-wise to generate a channel attention map, thereby improving the accu-racy of recognition. The detailed attention mechanism block is shown in Figure 6, As can be seen from the figure, ARN integrates the channel attention mech-anism and the spatial attention mechanism. Given a feature map F $R^{C\times H\times W}$ as input, The attention mechanism will in turn derive a one-dimensional channel attention map M_c $R^{C\times 1\times 1}$ and a two-dimensional spatial attention map M_s $R^{1\times H\times W}$, The complete attention process is as follows:

$$F' = M_c F + F$$
$$F'' = M_s F' + F'$$

Where represents element-wise multiplication.

First, set the average pooling and max pooling to aggregate the spa-tial information of the feature map and obtain F^c_A and F^c_M to represent the average pooling feature and the max pooling feature, respectively. These two features are then forwarded into a shared network consisting of a multi-layer perceptron (MLP) and a hidden layer to generate M_c . Briefly, channel attention is calculated as follows:

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F)))$$
$$= \sigma(W_1(W_0(F^c_A)) + W_1(W_0(F^c_M)))$$

where σ represents the sigmoid function,where W_0 $R^{C/r\times C}$,W_1 $R^{C\times C/r}$,The two weights W_0 , W_1 of the MLP are shared for both inputs, and W_0 after the ReLU activation function.

Similarly, use F^s_A $R^{1\times H\times W}$,F^s_M $R^{1\times H\times W}$ to represent the average-pooled and max-pooled features over channels and generate spatial attention:

$$M_s(F) = \sigma(f^{7\times 7}([AvgPool(F); MaxPool(F)]))$$
$$= \sigma(f^{7\times 7}([F^s_A; F^s_M]))$$

where $f^{7\times 7}$ is the convolution operation with a filter size of 7 × 7.

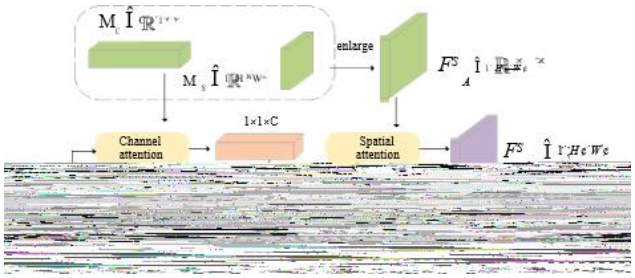


Fig 6. Attention mechanism module of ARN

MSFARN also achieves good results on irregular text datasets such as SVTP, CUTE80 and IC15. As shown in Table 8, MSFARN achieves better results.

In the SVTP dataset, most of the text images are perspective text with extremely low resolution. Whereas our method outperforms other methods with a 50-word lexicon and a combined dictionary of full images.

The CUTE80 dataset contains most of the curved texts, and the MSFARN method can rectify most of the curved texts in CUTE80 and can correctly recognize them. Therefore, it is robust enough to rectify image text with small bending angles.

Most of the pictures in the IC15 dataset are dim and blurry, and there are even some vertical text images, which are more difficult to identify. It can be seen from the table that the recognition accuracy of this method on this dataset only reached 71.8%, but it is still the highest. Therefore, this method outperforms other methods even on the IC15 dataset with greater difficulty.

Table 7 Results on the regular scene text datasets. Note: “50” and “1k” are lexicon sizes. “Full” indicates the combined lexicon of all images in the benchmarks. “None” means lexicon-free.“-” indicates that the method cannot be applied to recognition without a lexicon, or that the recognition accuracy cannot be reported without constraints

Table 8 Results on the irregular scene text datasets. Note: “50” and “1k” are lexicon sizes. “Full” indicates the combined lexicon of all images in the benchmarks. “None” means lexicon-free.“-” indicates that the method cannot be applied to recognition without a lexicon, or that the recognition accuracy cannot be reported without constraints

Table 8. Results on the irregular scene text datasets

		SVT-Perspective		CUTE80	IC15
Method	50	Full	None	None	None
ABBYY et al. Wang et al (2011)	40.5	26.1	-	-	-
Mishra et al. Mishra et al (2012)	45.7	24.7	-	-	-
Wang et al. Wang et al (2012)	40.2	32.4	-	-	-
Phan et al. Phan et al (2013)	75.6	67.0	-	-	-
Shi et al. Shi et al (2016b)	91.2	77.4	71.8	59.2	-
Yang et al. Yang et al (2017)	93.0	80.2	75.8	69.3	-
Liu et al. Liu et al (2016)	94.3	83.6	73.5	-	-
Cheng et al. Cheng et al (2017)	92.6	81.6	71.5	63.9	66.2
Shi et al. Shi et al (2018)	-	-	68.9	75.4	67.4
Ours	94.6	84.1	73.8	75.6	71.8

4.8 Limitations of MSFARN

Although MSFARN achieves good results on regular text datasets and irreg-ular text datasets, a problem is found in this paper during a large number of experiments. From Table 8, it is found that the MSFARN method does not achieve the highest recognition accuracy on lexicon-free in the SVT-Perspective dataset, but the method proposed by Yang et al.Yang et al (2017) achieves the highest accuracy. Analysis of the reasons is mainly due to the attention mech-anism proposed by Yang et al.

focuses on visually distorted or over-curved irregular texts, At the same time, this paper finds in practical tests that when the curved angle of the text in the image is too large, the method cannot correctly predict the text, which affects the final result. Figure 9 below shows the recognition results of some overly curved image texts.

Moreover, the MSFARN proposed in this paper is more aimed at the irregu-lar scene text in the horizontal direction, and most of the images in the dataset are obtained after cropping. Therefore, in the next stage, we will start with the vertical text and the addition of a text detector to frame the text in the scene. In conclusion, scene text recognition is still a very challenging problem.

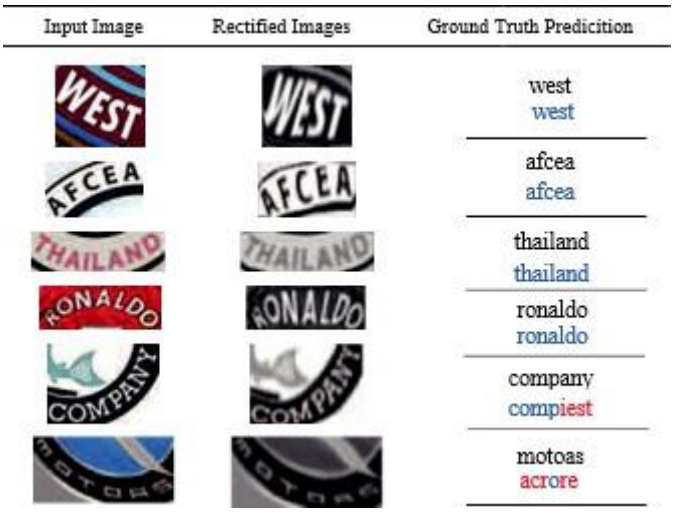


Fig 9. Effect of different curve angles in scene text

5. Conclusion

In this paper, we propose a multi-scale feature fusion attention recognition network (MSFARN) for text recognition in irregular scenes. The proposed framework consists of two parts, rectification and recognition. The rectification part uses the multi-scale feature fusion network MSFN, the purpose is to rec-tify the irregular (such as bending,deformation, blur) texts, convert them into the images with better readability, and then use the attention-based recogni-tion network ARN for recognition to obtain Final result. This paper conducts multiple experiments on a large number of regular and irregular text datasets (including IIIT5K, ICDAR2003, ICDAR2013, ICDAR2015, SVT-Perspective and CUTE80, etc.), all of which shows excellent performance, proving the superiority of the proposed MSFARN framework.

In the future, we will focus on text detection and recognition in more complex scenes. At the same time, the research of scene text in any direction and any font type is also very meaningful. In the next stage, more in-depth research on these aspects will be carried out.

References

[1] Almaz´an J, Gordo A, Forn´es A, et al (2014) Word spotting and recogni-tion with embedded attributes. IEEE transactions on pattern analysis and machine intelligence 36(12):2552–2566. https:// doi.org/ 10.1109/ TPAMI. 2014.2339814.

[2] Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:14090473 https://doi. org/10.48550/arXiv.1409.0473.

[3] Bissacco A, Cummins M, Netzer Y, et al (2013) Photoocr: Reading text in uncontrolled conditions. In: Proceedings of the IEEE international conference on computer vision, pp 785–792.

[4] Bulan O, Kozitsky V, Ramesh P, et al (2017) Segmentation- and annotation-free license plate recognition with deep localization and failure identification. IEEE Transactions on Intelligent Transportation Systems 18(9): 2351–2363. https://doi.org/10.1109/TITS.2016.2639020.

[5] Castro JDB, Canchumuni SWA, Villalobos CEM, et al (2021) Improvement optical character recognition for structured documents using generative adversarial networks. In: 2021 21st International Conference on Computa-tional Science and Its Applications (ICCSA), pp 285–292, https://doi.org/ 10. 1109/ ICCSA 54496.2021.00046.

[6] Cheng Z, Bai F, Xu Y, et al (2017) Focusing attention: Towards accurate text recognition in natural images. In: Proceedings of the IEEE international conference on computer vision, pp 5076–5084.

[7] Cheng Z, Xu Y, Bai F, et al (2018) Aon: Towards arbitrarily-oriented text recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5571–5579.

[8] Chorowski JK, Bahdanau D, Serdyuk D, et al (2015) Attention-based models for speech recognition. Advances in neural information processing systems https://doi.org/10.1016/0167-739X(94)90007-8.

[9] Gordo A (2015) Supervised mid-level features for word image representation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[10] Gupta A, Vedaldi A, Zisserman A (2016) Synthetic data for text localisation in natural images. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2315–2324.

[11] He P, Huang W, Qiao Y, et al (2016) Reading scene text in deep convolutional sequences. In: Thirtieth AAAI conference on artificial intelligence.

[12] Jaderberg M, Simonyan K, Vedaldi A, et al (2014a) Deep structured output learning for unconstrained text recognition. arXiv preprint arXiv:14125903 https://doi.org/10.1111/j.1365-277X.2011.01209.x.

[13] Jaderberg M, Simonyan K, Vedaldi A, et al (2014b) Synthetic data and arti-ficial neural networks for natural scene text recognition. arXiv preprint arXiv:14062227 https://doi.org/10.48550/arXiv.1406.2227.

[14] Jaderberg M, Vedaldi A, Zisserman A (2014c) Deep features for text spotting. In: European conference on computer vision, Springer, pp 512–528.

[15] Jaderberg M, Simonyan K, Vedaldi A, et al (2016) Reading text in the wild with convolutional neural networks. International journal of computer vision 116(1):1–20. https://doi.org/10.1007/s11263-015-0823-z.

[16] Jegou H, Douze M, Schmid C (2008) Hamming embedding and weak geo-metric consistency for large scale image search. In: European conference on computer vision, Springer, pp 304–317, https:// doi.org/ 10. 1007/978-3-540-88682-2 24.

[17] Karatzas D, Shafait F, Uchida S, et al (2013) Icdar 2013 robust reading com-petition. In: 2013 12th International Conference on Document Analysis and Recognition, IEEE, pp 1484–1493.

[18] Karatzas D, Gomez-Bigorda L, Nicolaou A, et al (2015) Icdar 2015 competi-tion on robust reading. In: 2015 13th international conference on document analysis and recognition (ICDAR), IEEE, pp 1156–1160.

[19] Lee CY, Osindero S (2016) 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.

[20] Lee J, Park S, Baek J, et al (2020) On recognizing texts of arbitrary shapes with 2d self-attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp 546–547.

[21] Li H, Wang P, Shen C, et al (2019) Show, attend and read: A simple and strong baseline for irregular text recognition. In: Proceedings of the AAAI Confer-ence on Artificial Intelligence, pp 8610–8617, https://doi.org/10.1609/aaai. v33i01.33018610.

[22] Li M, Fu B, Zhang Z, et al (2021) Character-aware sampling and rectification for scene text recognition. IEEE Transactions on Multimedia.

[23] Li M, Fu B, Chen H, et al (2022) Dual relation network for scene text recognition. IEEE Transactions on Multimedia.

[24] Liu W, Chen C, Wong KYK, et al (2016) Star-net: a spatial attention residue network for scene text recognition. In: BMVC, p 7.

[25] Liu Z, Li Y, Ren F, et al (2018) Squeezedtext: A real-time scene text recog-nition by binary convolutional encoder-decoder network. In: Proceedings of the AAAI Conference on Artificial Intelligence.

[26] Lou X, Kansky K, Lehrach W, et al (2016) Generative shape models: Joint text recognition and segmentation with very little training data. advances in neural information processing systems 29.

[27] Lucas SM, Panaretos A, Sosa L, et al (2005) Icdar 2003 robust reading com-petitions: entries, results, and future directions. International Journal of Document Analysis and Recognition (IJDAR) 7(2):105–122.

https: //doi. org/10.1007/s10032-004-0134-3.

[28] Mishra A, Alahari K, Jawahar C (2012) Scene text recognition using higher order language priors. In: BMVC-British machine vision conference, BMVA, https://doi.org/10.5244/C.26.127.

[29] Neumann L, Matas J (2012) Real-time scene text localization and recognition. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 3538–3545.

[30] Phan TQ, Shivakumara P, Tian S, et al (2013) Recognizing text with perspec-tive distortion in natural scenes. In: Proceedings of the IEEE International Conference on Computer Vision, pp 569–576.

[31] Raisi Z, Naiel MA, Younes G, et al (2021) 2lspe: 2d learnable sinusoidal posi-tional encoding using transformer for scene text recognition. In: 2021 18th Conference on Robots and Vision (CRV), IEEE, pp 119–126.

[32] Risnumawan A, Shivakumara P, Chan CS, et al (2014) A robust arbitrary text detection system for natural scene images. Expert Systems with Applications 41(18):8027–8048. https://doi.org/ 10.1016/ j. eswa. 2014. 07.008.

[33] Santos I, Castro L, Rodriguez-Fernandez N, et al (2021) Artificial neural net-works and deep learning in the visual arts: A review. Neural Computing and Applications 33(1):121–157. https://doi.org/ 10.1007/ s00521-020-05565-4.

[34] Seok JH, Kim JH (2015) Scene text recognition using a hough forest implicit shape model and semi-markov conditional random fields. Pattern Recogni-tion 48(11):3584–3599. https://doi.org/ 10.1016/ j. patcog. 2015.05.004.

[35] Shang M, Gao J, Sun J (2020) Character region awareness network for scene text recognition. In: 2020 IEEE International Conference on Multimedia and Expo (ICME), IEEE, pp 1–6.

[36] Shi B, Bai X, Yao C (2016a) An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE transactions on pattern analysis and machine intelligence 39(11):2298– 2304. https://doi.org/10.1109/TPAMI.2016.2646371.

[37] Shi B, Wang X, Lyu P, et al (2016b) Robust scene text recognition with automatic rectification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4168–4176.

[38] Shi B, Yang M, Wang X, et al (2018) Aster: An attentional scene text recog-nizer with flexible rectification. IEEE transactions on pattern analysis and machine intelligence 41(9):2035–2048.

[39] Su B, Lu S (2014) Accurate scene text recognition based on recurrent neural network. In: Asian Conference on Computer Vision, Springer, pp 35–48.

[40] Su B, Lu S (2017) Accurate recognition of words in scenes without character segmentation using recurrent neural network. Pattern Recognition 63:397–https://doi.org/10.1016/j.patcog.2016.10.016.

[41] Wang K, Babenko B, Belongie S (2011) End-to-end scene text recognition. In: 2011 International conference on computer vision, IEEE, pp 1457–1464.

[42] Wang T, Wu DJ, Coates A, et al (2012) End-to-end text recognition with convolutional neural networks. In: Proceedings of the 21st international conference on pattern recognition (ICPR2012), IEEE, pp 3304–3308.

[43] Wang Y, Ha JE (2021) Scene text recognition with multi-decoders. In: 2021 21st International Conference on Control, Automation and Systems (ICCAS), IEEE, pp 1523–1528.

[44] Wu L, Xu Y, Hou J, et al (2022) A two-level rectification attention network for scene text recognition. IEEE Transactions on Multimedia.

[45] Yang X, He D, Zhou Z, et al (2017) Learning to read irregular text with attention mechanisms. In: IJCAI, p 3.

[46] Yao C, Bai X, Shi B, et al (2014) Strokelets: A learned multi-scale represen-tation for scene text recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4042–4049.

[47] Ye Q, Doermann D (2014) Text detection and recognition in imagery: A survey. IEEE transactions on pattern analysis and machine intelligence 37(7):1480– 1500. https://doi. org/10. 1109/ TPAMI. 2014. 2366765.

[48] Yin F, Wu YC, Zhang XY, et al (2017) Scene text recognition with sliding convolutional character models. arXiv preprint arXiv:170901727 https://doi. org/10.48550/arXiv.1709.01727.

[49] Zhu X, Zhang Z (2021) Transformer-based end-to-end scene text recognition. In: 2021 IEEE 16th

Conference on Industrial Electronics and Applications (ICIEA), IEEE, pp 1691–1695.

[50] Zhu Y, Yao C, Bai X (2016) Scene text detection and recognition: Recent advances and future trends. Frontiers of Computer Science 10(1):19–36. https://doi.org/10.1007/s11704-015-4488-0.