International Core Journal of Engineering Volume 11 Issue 6, 2025
ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0042

CRNN Enhancement Architecture Integrating Linear Deformable Convolution and Multi-head Attention Mechanism

Xing Chen*

School of Automation and Electrical Engineering University of Jinan, Jinan 250200, China

*Corresponding author: cx1286700191@126.com

Abstract

This paper studies the challenges faced by the recognition of handwritten text on work orders in manufacturing factories, especially in complex scenarios where handwritten text is dense and connected in strokes in industrial work orders, making it difficult to achieve the desired effect. In view of the particularity of the application scenarios in this paper, this paper proposes the Enhanced Convolutional Recurrent Neural Network (MCRNN). In the MCRNN architecture, the linear variable convolution (LD-Conv) and the multi-head attention mechanism (MHA) are integrated, effectively enhancing the modeling ability of the model for local variations of fonts and time-dependent features. Based on this enhanced architecture, handwritten data is trained to construct a more adaptable handwritten text recognition model. In addition, compared with other advanced methods, our method shows better text recognition performance and improves the recognition accuracy of handwritten text.

Keywords

Handwritten Text; Text Recognition; Convolutional Recurrent Neural Network; Linear Variable Convolution; Multi-Head Attention Mechanism.

1. Introduction

Writing was the way and tool by which early humans recorded and expressed information through symbols. Chinese characters are one of the oldest writing systems in the world. As an important carrier of Chinese culture, their existence holds profound significance. With the development of technology, digital images are almost present in every aspect of life. For instance, in the manufacturing process of factories, most of the production orders for parts, equipment inspection records, and quality inspection reports of workpieces are still kept in paper form, among which there are many that contain a large amount of printed and handwritten information. At present, converting such paper data into electronic document form mainly relies on manual entry and management, which leads to low data processing efficiency, high error rate and difficult information traceability. Therefore, OCR technology [1] has emerged in response to the trend. Therefore, the research on the identification technology of manufacturing parts work orders will greatly enhance the digitalization and automation level of production data.

Before the popularization of deep learning, traditional handwriting recognition methods mainly relied on steps such as image preprocessing [2], feature extraction [3], and classification recognition [4]. Usually, it was necessary to locate, denoise, and correct the tilt of the text region, and then use manual features such as directional gradient histograms in combination with classifiers for recognition [5]. However, when dealing with complex scenarios (such as the existence of inconsistent handwritten text styles in production work orders), traditional methods have problems of low detection accuracy and poor recognition robustness. In contrast, OCR technology based on deep learning can detect and

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0042

recognize handwritten information in work orders more accurately, significantly improving the recognition accuracy and generalization ability.

This paper proposes a CRNN [8] enhanced architecture (MCRNN) that integrates linear deformable convolution [6] and multi-head attention [7] mechanism to solve the above problems. The main contributions are summarized as follows:

● The design introduces linear variable convolution in the convolutional layer to replace part of the traditional convolution, enhancing the feature extraction of tilted and cursive text.

● The design introduces a multi-head attention mechanism module in the loop layer to capture richer multi-scale context dependencies.

● Construct a dataset of handwritten work orders, conduct model training and experimental verification to verify the performance of the method proposed in this paper.

The remaining parts of this article unfold successively. Section 2 introduces the MCRNN method, Section 3 presents the experimental setup and results, and finally Section 4 summarizes the full text..

2. Method Research

2.1 Model Overview

The improved CRNN model proposed in this paper mainly consists of three parts: the feature extraction network based on linear deformable convolution, the Bidirectional LSTM (BiLSTM) [9] time series modeling module integrated with MHA, and the CTC decoder [10]. In the feature extraction stage, the LD-Conv module uses learnable affine matrices for sampling position transformation, and combines SVD [11] decomposition to constrain its transformation range to achieve robust modeling of local deformations. Meanwhile, a dynamic weighting mechanism is introduced to enhance the characteristic response of the deformed area.

The extracted spatial features are expanded into sequences through linear mapping and then input into the BiLSTM network to model the bidirectional temporal dependence between characters. To further enhance the perception ability of fragmented time series information, MHA is introduced between the BiLSTM layers. With the help of the sliding window and the gated attention mechanism, local details and context information are dynamically fused to improve the continuity and accuracy of character recognition. Ultimately, the model outputs the character probability distribution through the CTC layer and is trained and optimized using the CTC loss function. MCRNN structure is shown in Figure 1.

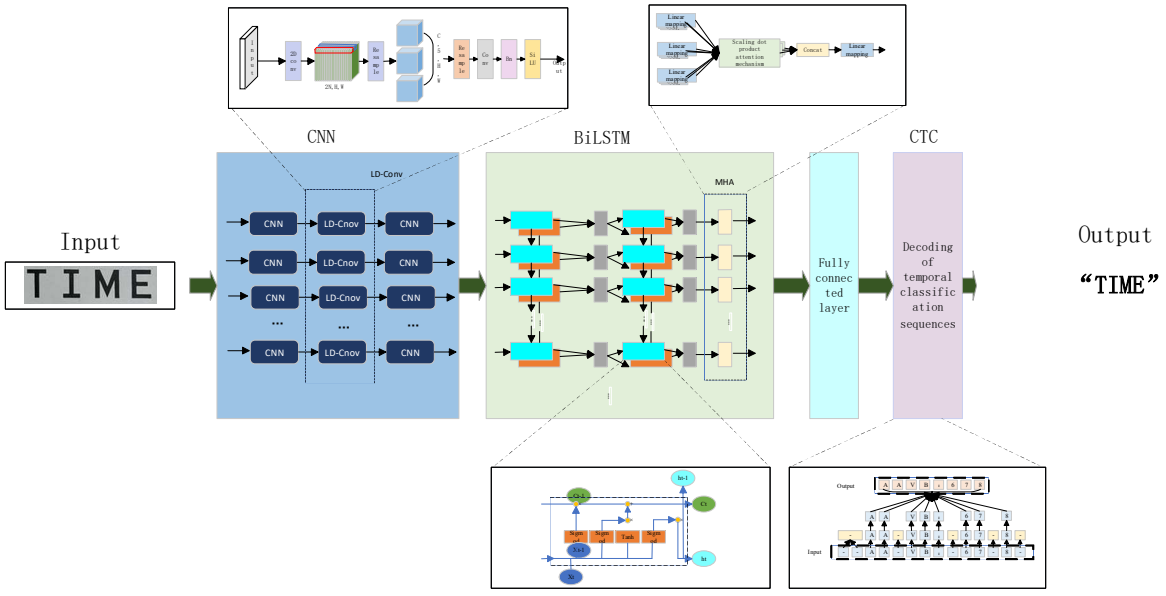


Figure 1. MCRNN model architecture diagram

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0042

2.2 Linear Deformable Convolution

The original CRNN model has limitations in feature extraction for handwritten text. Especially when dealing with complex and diverse handwritten text images, there are problems such as weak adaptability to font deformations like tilting and cursive strokes, and low recognition accuracy. To overcome these limitations and further enhance the recognition ability of the model, LD-Conv is introduced in this paper to improve the feature extraction network and adapt to the local geometric deformation of handwritten text. The structure of LD-Conv is shown in Figure 2.

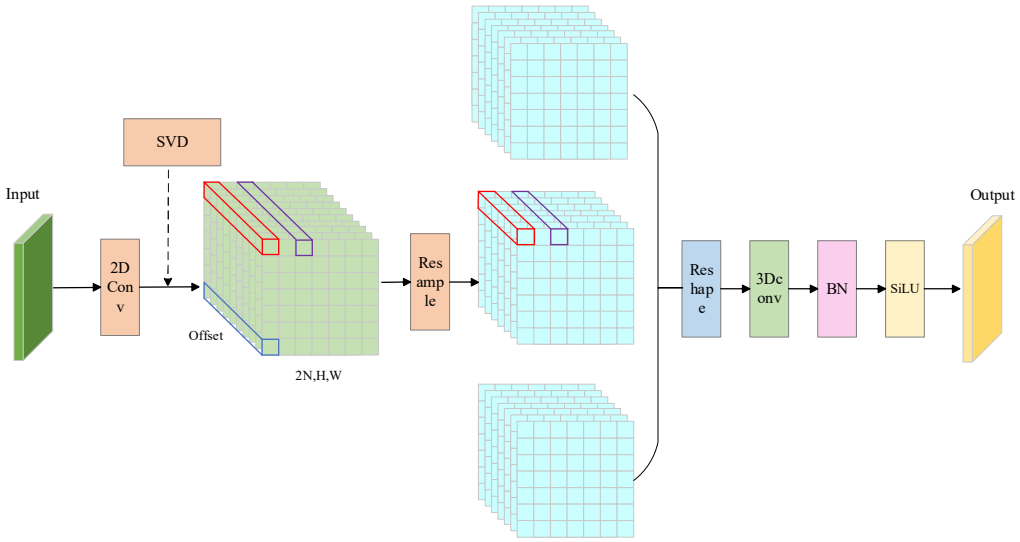


Figure 2. LD-Conv structure diagram

LD-Conv is a new type of convolution operation, which is particularly suitable for handwritten text recognition tasks. It can use different numbers of convolution kernel parameters (such as 1,2,3,4,5,6,7, etc.) to extract features from handwritten text images. This flexible parameter selection enables LD-Conv to better adapt to the diversity of handwriting than standard convolution or deformable convolution, especially when dealing with complex handwritten text images, such as character tilting, cursive deformations, etc.

Since the cursive features of different handwriting styles vary, this chapter introduces a dynamic weight scaling mechanism to adjust the weight distribution of subsequent convolutional layers using the mean value of the heat map. Specifically, this chapter first calculates the mean value of the heat map:

$$\bar{H} = \frac{1}{HW} \sum H_{i,j} \tag{1}$$

Dynamically adjust the kernel weights of the subsequent convolutional layers:

$$W_{ldconv} = W_{base} \cdot (1 + \alpha \cdot \bar{H}) \tag{2}$$

Among them, α is a learnable parameter, with an initial value set at 0.5 and optimized through backpropagation. The physical significance of this mechanism lies in that when the degree of convolution in the image is relatively high (i.e., large), the network will enhance the weight of the local dynamic convolution to adapt to the deformation area; When the number of connected strokes is relatively small, the basic convolutional feature extraction ability is maintained.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0042

Furthermore, the design concept of LD-Conv is scalable. It can customize specific sampling shapes based on prior knowledge and automatically adapt to the changes of the target shape through dynamic offsets, thereby enhancing the flexibility and adaptability of the model.

2.3 Modeling of the Attention Mechanism Sequence

Based on the traditional BiLSTM network, this paper designs a sequence modeling structure integrating MHA. This structure aims to enhance the model's modeling ability for complex relationships between characters and improve the recognition robustness in handwritten text scenarios such as irregular character arrangements, deformations, and cursive strokes.

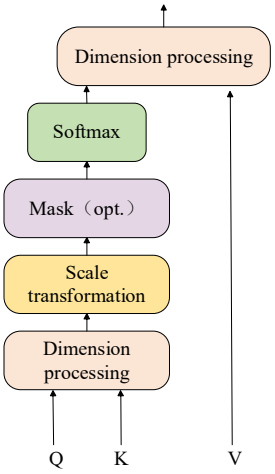


Figure 3. Self-attention mechanism structure

The core idea of MHA is to map the input features into multiple different representation subspaces and independently perform the self-attention mechanism within each subspace, thereby capturing the context dependencies between different positions in the input sequence. Each attention head can focus on learning the relationships between characters in different subspaces, and thus has significant advantages in modeling complex character structures (such as nonlinear stroke connections in handwriting, character deformations of upper and lower lines, etc.). The basic structure of the self-attention mechanism is shown in Figure 3.

It maps the input features to queries (Q), keys (K), and values (V) respectively, calculates the attention weights based on the dot product of Q and K, and performs weighted summation on V to output the context representation of the current time step. Its calculation process is as follows:

$$\begin{aligned} Q &= W^{Q} X \\ K &= W^{K} X \\ V &= W^{V} X \end{aligned} \tag{3}$$

Among them, W^{Q}, W^{K} and W^{V} are learnable weight matrices, and X is the input feature. Then, the attention score is calculated by performing dot product on the transposes of the query matrix and the key matrix and scaling by the square root of the key vector dimension. In each head, its calculation formula is as follows.

$$Attention(Q,K,V) = softmax\left(\frac{QK^{T}}{\sqrt{d_{K}}}\right)V \tag{4}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0042

Among them, d_k is the dimension of the key vector.

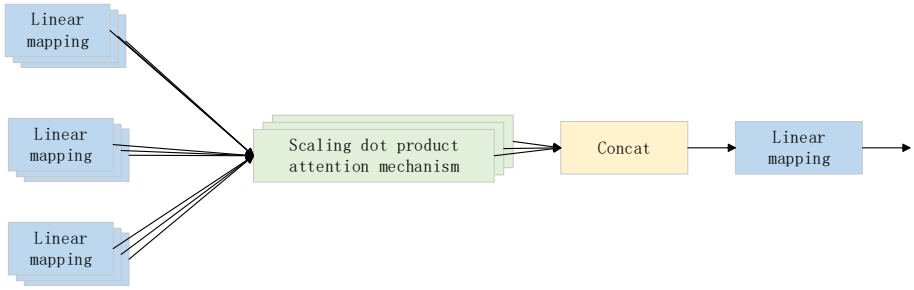


Figure 4. Multi-head attention mechanism

The multi-head attention structure is shown in Figure 4. The input features are passed in parallel to multiple attention heads, and the attention representations are calculated respectively. Through concatenation operations, the outputs of each head are fused and further integrated through linear transformation to obtain a global representation with richer context dependencies. This mechanism significantly enhances the model's modeling ability for the interaction between characters, and is particularly suitable for handwritten text recognition tasks with uneven character spacing or structural deformation.

3. Experiments and Results

3.1 Model Training

This paper trains the model on the NVIDIA GeForce RTX 3080Ti*4 (24G) GPU. The training parameters are set as image size [3,64,128], maximum text length 100, optimizer Adam [12], batch size 256, learning rate 0.0005, and weight attenuation coefficient 0.0005. The number of training rounds is 1000 For the handwritten text recognition model, since handwritten texts are more uncertain in structure and style, the learning rate is appropriately reduced to 0.0005 during training, and the weight attenuation coefficient is increased to 0.0005 to enhance the generalization ability and robustness of the model and avoid overfitting. The remaining training parameters are consistent with those of the printed model to ensure the fairness of the comparison. The training process of the handwritten model is shown in Figure 5. The training loss enters a stable decline stage after 200 rounds, and the accuracy rate of the validation set reaches the 96% saturation point around 300 rounds.

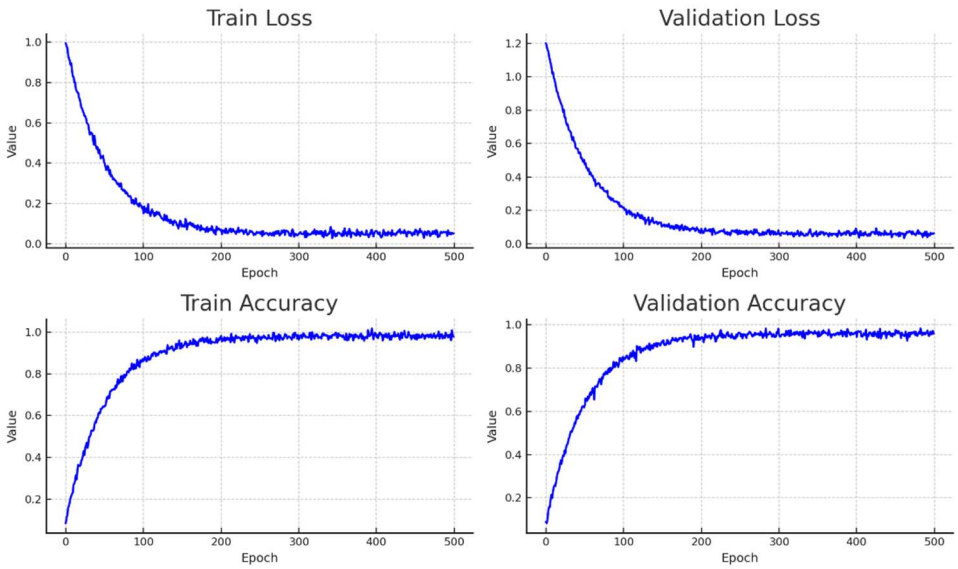


Figure 5. Visualization diagram of the training process of the handwritten recognition model

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0042

3.2 Dataset Construction

The datasets adopted for handwritten text recognition are the SCUT-EPT dataset [13] and the self-built manufacturing parts work order dataset. The SCUT-EPT dataset was released by the Laboratory of Deep Learning and Visual Computing of South China University of Technology and is specifically designed for the research of offline handwritten Chinese text recognition. This dataset contains 50,000 text-line images. Considering the diversity of data sources and the adaptability requirements of the model to actual scenarios, in this paper, 10,000 representative handwritten text images were randomly selected from the SCUT-EPT dataset and jointly constructed with 5,589 images from the self-built manufacturing parts work order dataset to build the training set. A total of 15,589 samples. This training set includes both widely distributed general handwriting styles and typical work order writing styles in real industrial scenarios, which is helpful for the model to learn both general text features and specific scene semantics simultaneously. The two types of data are mixed and used in the training at a ratio of approximately 2:1 to enhance the generalization ability and practical application effect of the model. Specific examples of the two datasets are shown in Figures 6 and 7.

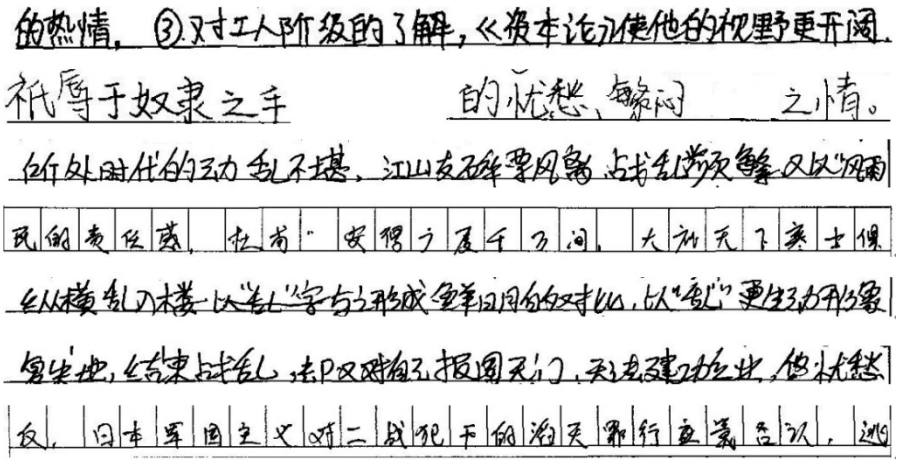


Figure 6. Part of the HwChinese dataset is presented

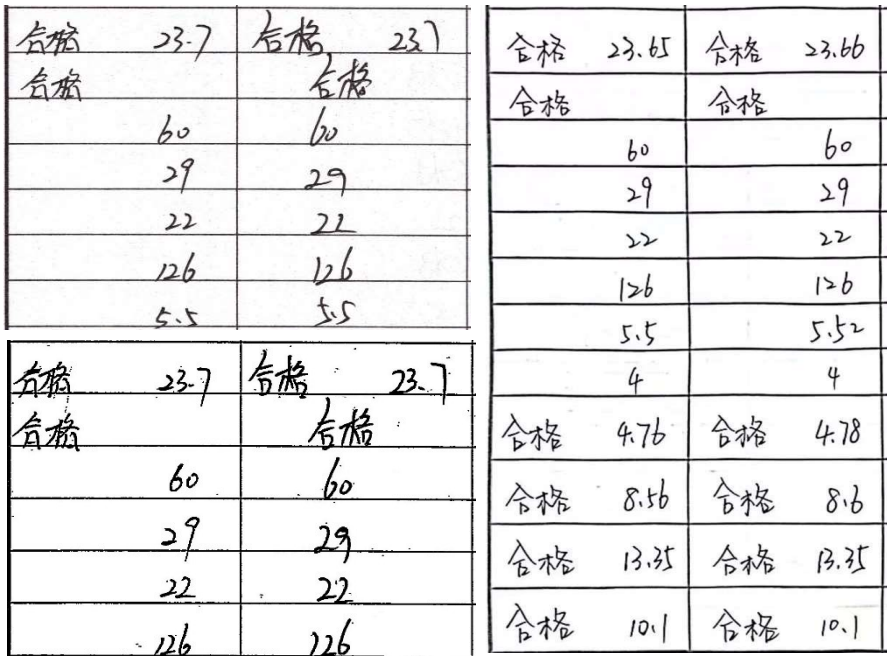


Figure 7. The handwritten part of the self-built data set of manufacturing parts work order texts

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0042

3.3 Results

This section conducts an experimental comparison of the recognition results of handwritten characters by the enhanced CRNN network. In order to verify the improvement of the model effect after the introduction of each LD-Conv and MHA, this section first conducted ablation experiments for each enhancement module, and the results are shown in Table 1.

Table 1. Comparison of enhanced module ablation experiments

Model	Accuracy (%)	Parameter (M)
CRNN(Base)	75.7	8.8
CRNN+LD-Conv	76.1	9.0
CRNN+MHA	78.5	9.8
CRNN+LD-Conv+MHA	83.9	10.4

Experiments show that the spatial deformation ability of LD-Conv improves the accuracy to some extent, but the modeling of cursive lines still relies on the original RNN, so the performance improvement is limited. After strengthening the MHA modeling, the accuracy was improved by 2.8%, indicating that the cursive error is strongly correlated with the long-range dependency. The accuracy of the method proposed in this paper reaches 83.9% after the collaborative optimization of the two modules, verifying the complementarity between spatial local deformation and temporal global correction. After adding MHA, the 0.3M parameter count brought a significant improvement of 2.8% in accuracy, reflecting the high efficiency of the timing module. The method proposed in this paper achieves a comprehensive performance breakthrough with a 1.6M model increment, proving the feasibility of the joint design of the two enhancement modules.

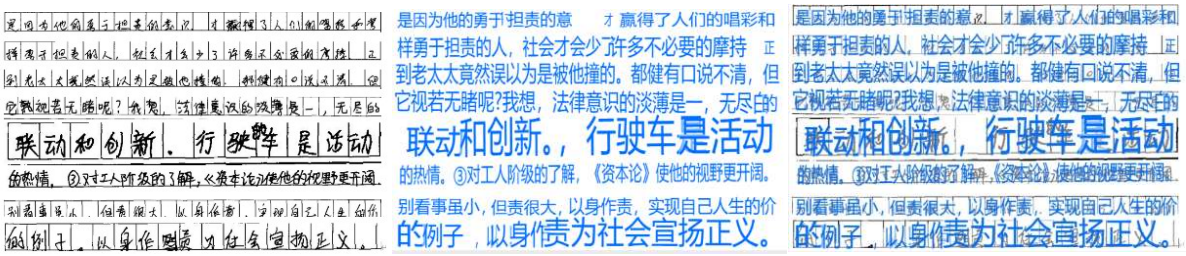
Experimental analysis was conducted for the recognition of handwritten text. In this section, the same handwritten text dataset was trained using different mainstream text recognition networks and a comparison of handwritten text recognition was made. The experimental results are shown in Table 2.

Table 2. Comparison of CER experiments in different models

Model	Whole CER(%)	Lianbi CER(%)	Parameter(M)
CRNN(Base)	16.7	22.3	8.8
DenseRNN	16.5	22.8	11.6
ABINet	14.2	18.1	19.9
MCRNN(Ours)	12.7	15.7	10.4

It can be seen from Table 2 that in the overall CER, the method proposed in this paper decreased by 24.0% (16.7%→12.7%) compared with the baseline and further decreased by 4.9% compared with ABINet, verifying the effectiveness of the collaborative optimization of the enhanced module. Meanwhile, in the continuous CER, it decreased by 24.1% compared with the baseline, which was significantly better than DenseRNN and ABINet, proving the complementarity between LD-Conv deformation modeling and multi-head attention. In addition, the parameter count of the enhanced CRNN is only 1.4M, which is only 1.6M and 1.5M higher than that of the original CRNN, and it is less than both DenseRNN and ABINet, indicating that it requires fewer computing resources. It can not only prove the effectiveness of the handwritten text recognition method proposed in this paper. The recognition results of the handwritten part of the method in this paper are shown in Figures 8 and 9.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0042



(a) Original image (b) Identification result diagram (c) Result restoration image

Figure 8. This paper presents the recognition results of handwritten text by method

(a) Original image (b) Identification result diagram (c) Result restoration image

Figure 9. The method proposed in this paper is aimed at the text recognition results of the handwritten part of the work order for manufacturing parts

4. Conclusion

This paper studies the challenges of recognizing handwritten text on industrial work orders, improves the original CRNN model, introduces linear variable convolution to replace the feature extraction part in the original CRNN network, and enhances the model's feature extraction ability for irregular and cursive areas of handwritten text. Furthermore, the MHA enhanced sequence modeling method is introduced, enabling the model to effectively handle character sequences with complex character arrangements and inconsistent character spacings. Experiments have proved that after introducing two enhancement modules, the recognition accuracy of handwritten text has been further improved. This comprehensive improvement method has also significantly reduced the false recognition rate of the model for handwritten text on industrial work orders.

References

[1] Nguyen T T H, Jatowt A, Coustaty M, et al. Survey of post-OCR processing approaches[J]. ACM Computing Surveys (CSUR), 2021, 54(6): 1-37.

[2] Liao M, Wan Z, Yao C, et al. Real-time scene text detection with differentiable binarization[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 11474-11481.

[3] Mutlag W K, Ali S K, Aydam Z M, et al. Feature extraction methods: a review[C]//Journal of Physics: Conference Series. IOP Publishing, 2020, 1591(1): 012028.

[4] Ahlawat S, Rishi R. A genetic algorithm based feature selection for handwritten digit recognition[J]. Recent Patents on Computer Science, 2019, 12(4): 304-316.

[5] Tatar G. Design Aspects of Machine Learning Algorithms for the Hardware Implementation of Advanced Driver Assistance Systems (A/DAS)[D]. Marmara Universitesi (Turkey), 2024.

[6] Wang, Z., et al. Linear deformable convolution with dynamic kernel scaling for efficient vision transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence[J], 2024, 46(3), 1452-1466.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0042

[7] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[8] Shi B, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(11): 2298-2304.

[9] Chen T, Xu R, He Y, et al. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN[J]. Expert Systems with Applications, 2017, 72: 221-230.

[10] Tsunoo E, Futami H, Kashiwagi Y, et al. Decoder-only architecture for speech recognition with ctc prompts and text data augmentation[J]. arxiv preprint arxiv:2309.08876, 2023.

[11] Levinson J, Esteves C, Chen K, et al. An analysis of svd for deep rotation estimation[J]. Advances in Neural Information Processing Systems, 2020, 33: 22554-22565.

[12] Edwards D R, Handsley M M, Pennington C J. The ADAM metalloproteinases[J]. Molecular aspects of medicine, 2008, 29(5): 258-289.

[13] Wang Z R. Integrating Canonical Neural Units and Multi-Scale Training for Handwritten Text Recognition[J]. arxiv preprint arxiv:2410.18374, 2024.