SIN-CHN
SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2025, 22(4), 897.
https://doi.org/10.62617/mcb897

Article

Biomechanics-inspired analysis of the recognition function of recurrent neural networks in primary school math homework under low carbon background

Jing Xu[1], Ying Wang[1], Xuan Wang[2], Zheng Wang[1,*]

[1] Department of Primary Education, Baoding Preschool Teachers College, Baoding 071000, China
[2] Department of Preschool Education, Baoding Preschool Teachers College, Baoding 071000, China
*** Corresponding author:** Zheng Wang, wangzheng112602@126.com

CITATION

Xu J, Wang Y, Wang X, Wang Z. Biomechanics-inspired analysis of the recognition function of recurrent neural networks in primary school math homework under low carbon background. Molecular & Cellular Biomechanics. 2025; 22(4): 897. https://doi.org/10.62617/mcb897

ARTICLE INFO

Received: 24 November 2024
Accepted: 2 December 2024
Available online: 13 March 2025

Abstract: In the traditional education assessment landscape, the manual grading of subjective exam questions poses significant challenges. The labor-intensive nature of this process and the potential for human error can negatively impact teaching and learning outcomes. As society transitions towards a low-carbon future, there is a pressing need to reform educational evaluation methods, reduce paper-based exams, and leverage advanced intelligent technologies. Inspired by the principles of biomechanics, this research introduces a novel image-based handwritten text recognition algorithm powered by recurrent neural networks, specifically designed for the automated scoring of primary school mathematics subjective questions. Drawing insights from the human visual and cognitive systems, the proposed approach mimics the hierarchical and adaptive nature of biological information processing to tackle the complexities inherent in handwritten text detection, recognition, and understanding. The study first constructs a comprehensive dataset of real primary school math exam answer sheets, capturing the diverse range of handwriting styles and mathematical notations. This dataset serves as a robust training and evaluation platform, akin to the diverse sensory inputs that biological systems process. The recurrent neural network architecture employed in this work exhibits biomimetic properties, such as the ability to dynamically process sequential information and adaptively refine its internal representations, much like the human brain's neural networks. This allows the algorithm to effectively handle the contextual cues and structural patterns present in handwritten mathematical responses, enabling accurate recognition and interpretation. Rigorous comparative and ablation experiments were conducted to assess the performance of the proposed algorithm. The results demonstrate high accuracy in recognizing and interpreting handwritten subjective responses, showcasing the practical value of this biomechanics-inspired approach. These findings align with the study's overarching goal of developing resource-saving and environmentally-friendly education evaluation systems, paving the way for the widespread adoption of intelligent technologies in the assessment of subjective questions. By drawing inspiration from the elegant and efficient information processing mechanisms observed in biological systems, this research contributes to the advancement of intelligent handwritten text recognition, ultimately supporting the transition towards a more sustainable and equitable educational landscape.

Keywords: deep learning; text detection and recognition; handwritten text; text mining; biomechanics; handwriting analysis

1. Introduction

In the field of education, examination is an important tool to evaluate teaching quality and students' knowledge level. Traditionally, teachers ask questions, students answer, and teachers grade the answers. However, manual grading faces many challenges, such as error-prone correction, prolonged grading period and blocked teaching progress. Subjective factors, such as students' writing conditions and teachers' subjective inclination, affect the scoring results, which leads to differences in candidates' scoring. On the other hand, computerized marking provides convenience and fairness, which is especially valuable for large-scale marking.

With the development of the era of big data, China gradually attaches importance to the construction of a resource-saving and environment-friendly society, so it advocates the education department to do a good job in green examination reform and reduce the environmental loss of the examination. The primary school mathematics examination also needs to be involved. We should pay attention to the needs of low-carbon development, reduce the loss of homework paper and manpower, actively invest in advanced homework correction technology, and do a good job in examination. Among them, the education department needs to pay attention to the use of automatic grading technology in the process of correcting exam assignments, so as to reduce the waste of scoring statistics paper, so as to appeal to the principle of low carbon and environmental protection. Behind the "low carbon" is the harmony of the whole ecology. Therefore, in the process of educational evaluation reform, the implementation of green automatic identification technology is an important measure to build a low-carbon campus, and it is of great significance to promote ecological civilization. Moreover, the implementation of automatic identification technology is also in line with the general idea of low-carbon city construction. Low-carbon campus can reduce the consumption of paper resources, and use automatic identification technology to reduce resource consumption and actively promote the realization of low-carbon campus. In this way, adopting automatic grading technology is essentially a foundation of low-carbon consciousness in campus, and it is also an improvement of campus culture, which plays an important educational role in the education of students' spiritual level. By building a low-carbon scoring system, we can introduce and exchange our own experience to the local, national and even international community, improve our reputation and image in the region, and be conducive to the development of the school itself.

Although the automatic scoring of multiple-choice questions is very mature, due to the limitation of the automatic scoring system, the types of questions that need to be answered manually, such as fill-in-the-blank questions, calculation questions and application questions, are still dominated by manual scoring. At present, it is difficult for the system to accurately detect the position of handwritten text, identify the types of questions, and achieve a high level of text recognition accuracy, which hinders the understanding of candidates' answers and provides wise correction.

With the rapid increase of computing and data resources and the evolution of diversified network structures, deep learning has achieved promising results in various fields, and artificial intelligence has penetrated into the fields of voice and license plate recognition. The key technologies supporting intelligent scoring, especially handwritten text detection and recognition, have gained a lot of research attention, resulting in advanced methods conducive to comprehensive automatic scoring. Based on deep learning, this study studies a method to automatically identify the subjective content of primary school mathematics test paper images.

2. Literature review

The evolution of text detection methodologies has seen a significant shift from traditional feature-based approaches to the adoption of deep learning techniques. This section is organized by theme and methodology to provide a comprehensive overview of the state of the art in text detection.

2.1. Traditional text detection methods

Traditional methods mainly focus on extracting image features to identify text candidate regions. Notable feature extraction techniques include stroke width transform (SWT) [1], maximum stable extremum region (MSER) [2] and gradient direction histogram (HOG) [3]. These methods usually use sliding window or connected component method to select candidate regions, and then classify them according to a set of predefined rules. The overall workload is cumbersome and the operation is difficult, and the accuracy of the results is still lacking. For example, Li et al. [4] introduced a method to classify candidate windows by image decomposition and second moment calculation using the average value of image wavelet coefficients. Zhang et al. [5] used the structural similarity of text lines to detect the center point through multi-scale sliding windows, and applied constraints such as symmetry, character spacing and angle to improve the text region prediction. There are obvious deviations in the predicted results. That is to say, the traditional calculation method is not only difficult to calculate, but also may consume a lot of time and resources when the graph is very dense. And it can’t capture the local structure and depends on the initial conditions. Different initial conditions may lead to different results.

2.2. Deep learning in text detection

The advent of deep learning has revolutionized text detection by offering improved accuracy and adaptability. Early deep learning approaches, such as CTPN [6–8], introduced the concept of using anchor boxes to predict partial text line areas and bidirectional long-term memory networks (Bi-LSTM) [9] to learn contextual information. However, these methods were limited in their ability to handle curved and rotated text lines.

Subsequent advancements, such as TextBoxes [10] and its extension TextBoxes++ [11,12], adapted the convolutional neural network (CNN) architecture to better detect long text lines by changing the convolution kernel size. SegLink [13] improved upon this by predicting text candidate boxes and merging them through an area link algorithm, allowing for the detection of text lines at various angles.

2.3. End-to-end text line recognition

The current trend in text line recognition favors end-to-end methods that directly recognize complete text lines. DTRN [14] was a pioneer in this domain, employing a CNN-RNN combination to transform text recognition into a sequence recognition problem. The introduction of CTC [15] by Shi et al. [16] enabled end-to-end training for RNNs, accommodating word images of arbitrary lengths. Liao et al. [17] proposed an innovative end-to-end approach that combines region proposal with instance segmentation for character-level recognition, although it requires character-level annotations for training. ABCNet [18] stands out with its use of Bezier curves for parameterizing text shapes and its impressive speed, outperforming previous methods by more than tenfold.

2.4. Addressing research gaps

The aforementioned review highlights the progression from traditional methods to deep learning and end-to-end recognition. However, these advancements have not fully addressed the challenges of detecting text in complex scenarios, such as those with large character intervals or curved text lines. This study aims to bridge these gaps by exploring the integration of deep learning techniques with advanced feature extraction methods to improve the robustness and accuracy of text detection in diverse contexts.

3. Research design

3.1. Automatic identification algorithm for primary school mathematics subjective questions

This article mainly focuses on the needs for automatic correction of subjective questions in primary school mathematics examination scenarios, and designs and implements a set of algorithms for automatic correction. The overall flow chart of the content identification and automatic correction algorithm for primary school mathematics subjective questions is shown in **Figure 1**. It can be seen from the figure that during the entire correction process, this paper proposes a multi-question handwritten text line detection algorithm based on various key technologies. Handwritten text line recognition algorithm, and text similarity calculation.

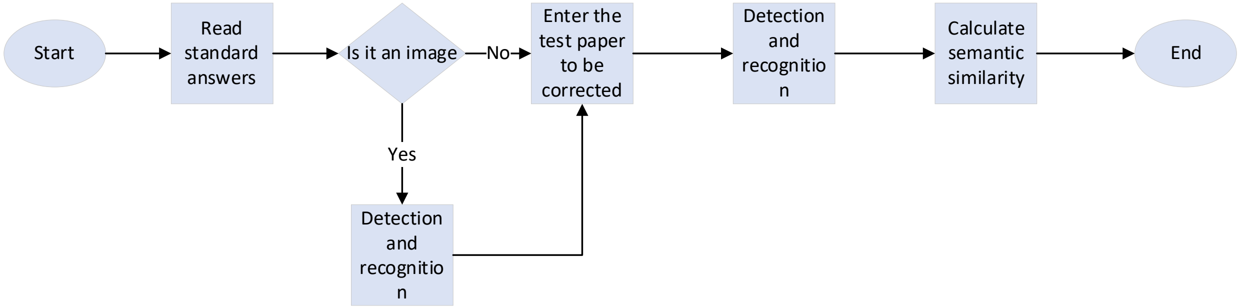
Start
Read standard answers
Is it an image
No
Yes
Enter the test paper to be corrected
Detection and recognitio n
Detection and recognitio n
Calculate semantic similarity
End

Figure 1. The overall process of the algorithm.

For the existing technology, the multi-topic handwritten text line detection algorithm can quickly process a large amount of data in a certain period of time, which is very important in the case of huge data. In the process of automatic identification of data and information, the algorithm can complete the task efficiently, while the traditional manual processing method will be very inefficient in the face of large-scale data. And it can also improve the accuracy of data processing, and the automatic identification algorithm can reduce the errors caused by human factors. It can achieve high accuracy when dealing with tasks such as character recognition.

3.2. Detection algorithm

Handwritten subjective questions in primary school mathematics examination include fill-in-the-blank questions, judgment questions, calculation questions, application questions and other questions. According to the actual writing form in the answer sheet image, this paper divides the text types into three categories: Fill-in-the-blank questions, judgment questions and calculation questions. In the text detection stage, it is the key premise for subsequent text recognition and automatic correction to accurately detect the position of text lines and correctly distinguish the types of questions they belong to. We try to use the existing text detection algorithms, such as CTPN [8], PSENet [19] and DB [20], to detect the text on the data set in this paper. We find that the above algorithms can get good detection results for text lines with clear outlines and no adjacency, but for two or more lines of text that are very similar or even overlap in the horizontal direction, DB algorithm can get good separation results, while other algorithms have no effect. For DB algorithm, there is no need to specify the number of clusters in advance like K-means algorithm. It automatically identifies natural clustering in data sets through the concept of density. At the same time, it maintains scalability and efficiency in dealing with large data sets, which makes DB algorithm a powerful tool in dealing with complex data distribution and large-scale data sets. Therefore, this paper uses DB algorithm as the detection algorithm of subjective questions [21].

3.3. Recognition algorithm

Based on the mainstream CRNN+CTC [16] end-to-end text recognition framework, the author improves it by combining the channel attention mechanism, and realizes the network structure of feature extraction by using multi-scale acceptance domain, thus improving the accuracy of handwritten text line recognition algorithm.

3.3.1. Overall network structure

In the conventional CRNN+CTC [16] structure, the input image is first scaled to $32 \times W \times 3$ while maintaining its original aspect ratio, and then passed through several layers (3 × 3) Convolution, average pooling, and maximum pooling are used to obtain a feature map with size $1 \times \frac{W}{4} \times C$. Then, each pixel in the feature map is serialize according to the channel dimension to obtain a sequence $T = \frac{W}{4}$ of length X, as shown in Equation (1):

$$X = \{x_1, x_2, x_3, x_4, \dots, x_T\} \tag{1}$$

where x_i represents the i element in the sequence, which is a vector in the $(1 \times C)$ dimension, and x_i can be represented as Equation (2):

$$x_i = \{x_i^1, x_i^2, x_i^3, x_i^4, \dots, x_i^C\} \tag{2}$$

Take sequence X as the input of the recurrent neural network BLSTM module and output a probability matrix with a scale of $n \times T$, where n represents the number of characters in the dataset. The output of the BLSTM module is influenced by the serialized X input value. Each vector element x_i in X is obtained from the feature map and can actually be mapped to a certain region in the original image, which can be understood as x_i receptive field in the original image.

CRNN+CTC [16] uses a single (3 × 3) convolution when doing feature extraction, and the receptive field corresponding to the convolution at this scale cannot be applied to each character. Therefore, this paper proposes a feature extraction network based on the channel attention mechanism. In the feature extraction network, convolution kernels of multiple scales are used for feature extraction, and the channel attention mechanism allows the network to autonomously learn which scale is more suitable for Parts of the characters in the image.

The overall network structure of the handwritten text line recognition algorithm proposed in this article is shown in **Figure 2**, which includes a feature extraction network based on the channel attention mechanism, a BLSTM module and a string transformation module. In the feature extraction network, convolution kernels of three scales (1 × 3), (3 × 3) and (3 × 1) are used for feature extraction and divided into three branches, and the three branches are combined in the channel attention module. The three feature maps obtained by the branches are spliced into the same feature map. The feature map of each branch corresponds to a channel in the spliced feature map, and then the channel attention mechanism is used to calculate the weight of each channel.

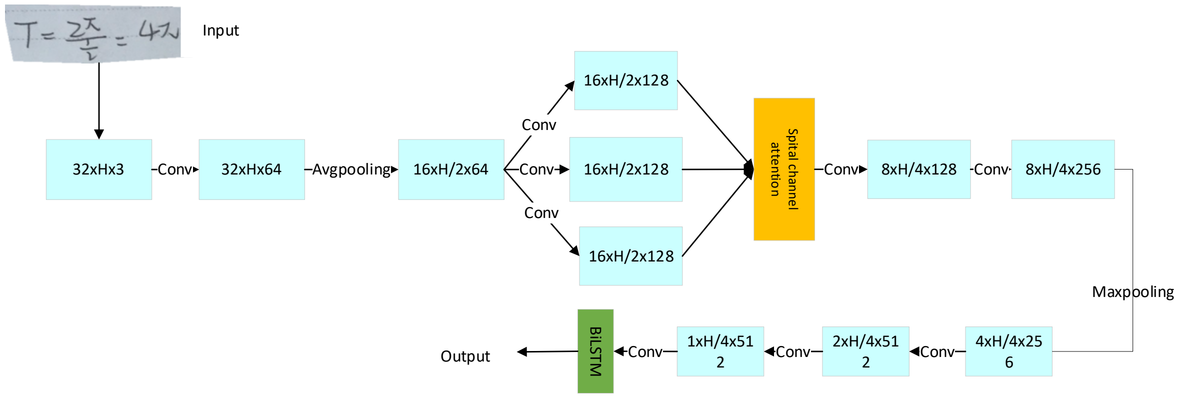

Figure 2. Network structure of handwritten text recognition algorithm.

As shown in **Figure 2**, while maintaining the aspect ratio, scale the input image to $32 \times W \times 3$. Through the feature extraction network, obtain a feature map $(1 \times \frac{W}{4} \times 512)$, which becomes 1/32 of the input scale in the height direction and 1/4 of the input scale in the width direction. The scale changes of width and height during the feature extraction process are inconsistent, which is different from the general feature extraction network. This is because in the partial pooling and convolution operations of the network, the step size in the width direction is set to 1 and the step size in the height direction is set to 2, which can preserve the detailed information in the horizontal direction as much as possible. The specific parameters for convolution, pooling, and other operations in the network are shown in **Table 1**.

Table 1. Feature extraction network parameters based on channel attention mechanism.

Types	Kenel	Stride	Padding	Channel
Conv	(3 × 3)	(1 × 1)	(1 × 1)	64
Avg pooling	(2 × 2)	(2 × 2)	(0 × 0)	/
Conv 3 × 3	(3 × 3)	(1 × 1)	(1 × 1)	128
Conv 3 × 1	(3 × 1)	(1 × 1)	(1 × 0)	128
Conv 1 × 3	(1 × 3)	(1 × 1)	(0 × 1)	128
attention	/	/	/	/
Conv 3 × 3	(3 × 3)	(1 × 1)	(1 × 1)	256
Conv 3 × 3	(3 × 3)	(1 × 1)	(1 × 1)	256
Max Pooling	(2 × 2)	(2 × 1)	(0 × 0)	/
Conv 3 × 3	(3 × 3)	(1 × 1)	(1 × 1)	512
Conv 3x1	(3 × 1)	(1 × 1)	(1 × 0)	512
Conv 1 × 3	(1 × 3)	(1 × 1)	(0 × 1)	512
attention	/	/	/	/
Conv 2 × 3	(2 × 3)	(1 × 1)	(0 × 1)	512

3.3.2. Multi scale channel attention module

The multi-scale receptive field channel attention module proposed in this article is shown in **Figure 3**. This module receives three feature maps of the same size ($H \times W \times C$) output by different branches, first averages each feature map in the channel direction, compresses it into a single-channel feature map ($H \times W \times 1$), and then the three feature maps are spliced in the channel dimension to form a feature map of $H \times W \times 3$ size, denoted as F. Then each channel in F corresponds to the feature extraction result of the corresponding size convolution kernel in the network. Through the channel attention by using the force mechanism, the weight of the feature map extracted by the convolution kernel of each scale can be learned. The feature map F' of $H \times W \times 1$ is obtained by weighting and summing the values of each pixel in different channels in F, and then using (1 × 1) convolution to restore it to the input size ($H \times W \times C$). At the end of this module, the feature map needs to be processed by maximum pooling. The first attention module of the feature extraction network in this article uses a step size of (2 × 2) to halve both the width and height. The second the attention module uses a step size of (2 × 1), so that the height is halved but the width remains unchanged.

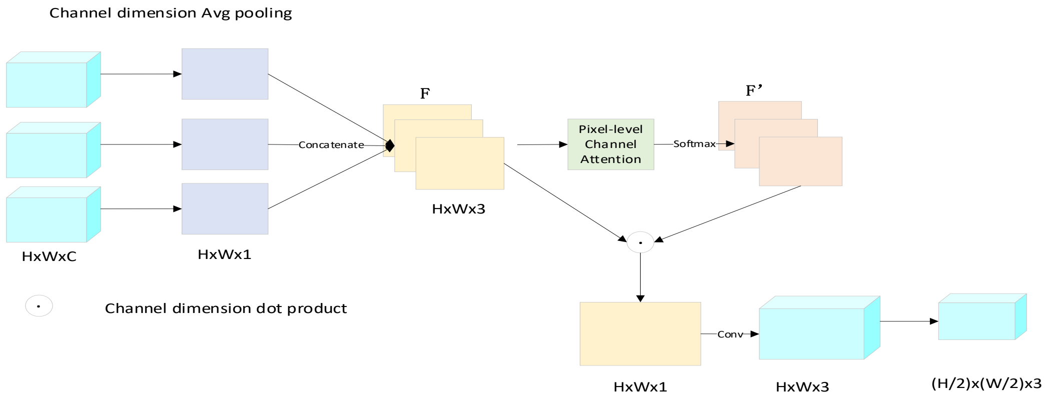
Channel dimension Avg pooling
HxWxC
HxWx1
Concatenate
F
HxWx3
Pixel-level Channel Attention
Softmax
F'
Channel dimension dot product
HxWx1
Conv
HxWx3
(H/2)x(W/2)x3

Figure 3. Multi scale channel attention module structure.

As shown in **Figure 3**, the channel attention module of this article is modified with reference to the channel attention module in the CBAM [22] model, replacing its Sigmoid module with a Softmax module, and replacing the parameters of the multi-layer perceptron module, so that the output of the middle layer After the result is passed through Softmax, a weight matrix A of size $H \times W \times 3$ is obtained. Each value in the matrix corresponds to the weight of each pixel in the feature map F. Therefore, the weight matrix output by the channel attention module is at the pixel level. The weight matrix A can be obtained according to Equation (3).

$$A = S(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{3}$$

where S represents the Softmax function. To make the pixel-level weight act on the feature map F, the dot product operation defining the channel dimension is as shown in Equation (4):

$$F_{i,j}^{'} = \sum_{c=0}^{2} F_{i,j,c} \times A_{i,j,c} \tag{4}$$

where C represents the sequence number of the channel dimension. Since F is a three-channel feature map, the value of C ranges from 0 to 2.

4. Experiment and conclusion

4.1. Dataset

In this study, we have constructed a specialized dataset for the recognition of handwritten text lines in primary school mathematics exams and compared various text recognition methods on this dataset. The images in the dataset are standardized to ensure that all images have a fixed width of 32 pixels, with the length of the images varying dynamically based on the number of characters contained within the image. As shown in **Figure 4**, our annotated dataset includes 4000 images of handwritten text lines, with 3600 images allocated for the training set to train the model to recognize various styles of handwriting; the remaining 400 images constitute the test set, used to evaluate the model's recognition performance. To enhance the model's generalization capabilities, we have additionally generated 10,000 images, all of which are used for the pre-training phase of the model. This dataset design not only provides ample learning material but also ensures the model's adaptability to different writing styles, text layouts, and character densities by including diverse handwriting samples and dynamically adjusting image lengths, thereby achieving higher recognition accuracy and robustness in practical applications.

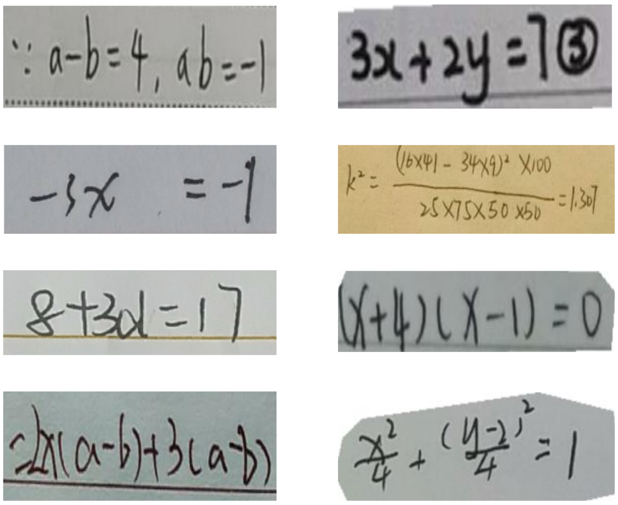

Figure 4. Data set example.

4.2. Evaluation index

The edit distance of a string refers to the minimum number of operations required to delete, insert, or replace characters in string *A* to transform it into string *B*. It can measure the similarity between two strings. The edit distance can be calculated by Equation (5).

$$L\_\{str1, str2\} = I + D + R \tag{5}$$

Among them, *I* represents the number of characters that need to be inserted during the transformation process, *D* represents the number of characters that need to be deleted during the transformation process, and *R* represents the number of characters that need to be replaced during the transformation process. This article uses two indicators to evaluate the recognition effect of the handwritten text line recognition algorithm, namely the character error rate CRE defined by Equation (6) and the recognition accuracy ACC defined by Equation (7).

$$CRE = \frac{\Sigma_{i=I}^{N} \frac{L_i}{Len_i}}{N} \tag{6}$$

$$ACC = \frac{\Sigma_{i=1}^{N} F(L_i)}{N} \tag{7}$$

$$F(L_i) = \begin{cases} 1, & \text{if } L_i = 0 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Among them, N represents the total number of pictures in the test set, L_i represents the edit distance between the picture recognition result and the label string, and Len_i represents the length of the picture label string. The character accuracy of a single picture cannot reflect the overall effect of the recognition algorithm, so the CRE indicator is Take the average of the accuracy of all picture characters in the entire test set. This article believes that when the edit distance between the recognition string and the label string is 0, that is, when the two strings are exactly the same, the recognition result is correct, otherwise it is a recognition error. The recognition accuracy is the ratio of the number of correctly recognized pictures to the total number of pictures.

4.3. Experimental and results

In terms of parameter settings, we employ batch normalization techniques to accelerate training and reduce the risk of overfitting. The learning rate of the model is set to 0.001, and parameters are efficiently updated using the Adam optimizer. This structural and parameter configuration is aimed at improving the accuracy and generalization capability of the handwritten text line recognition.

This article compares the classic algorithms in the field of text recognition proposed by Shi et al. [16], Mask textspotter [17], ABCNet [18] and the handwritten text recognition algorithm proposed in this article. The comparison results of various indicators are shown in **Table 2**. It shows that the handwritten text line recognition algorithm proposed in this article achieved the lowest character error rate on 400 test images compared with other methods. The algorithm proposed by Shi et al. [16] had a total of 74 picture recognition errors, 38 of which were adjacent to the same character recognition errors, and of these 38 pictures, 29 were correctly recognized by this method, proving that many of the methods proposed in this article were incorrect. The scale receptive field channel attention module plays a positive role in the recognition of adjacent identical characters. ABCNet [18] is a method that integrates text detection and text recognition, and can simultaneously achieve detection and recognition tasks in a network. From **Table 2**, it can be found that the algorithm proposed in this article not only has the highest recognition accuracy, but also has the highest character recognition accuracy. The error rate is also the lowest.

Table 2. Comparison of text line recognition algorithm indicators.

Methods	**ACC**	**CRE**
Shi et al. [16]	81.50%	9.25%
Mask textspotter [17]	85.00%	6.45%
ABCNet [18]	89.50%	5.87%
CRNN+CTC [15]	90.24%	5.46%
Ours	91.24%	5.04%

The algorithm proposed in this paper has been verified by experiments. By setting specific evaluation indicators and comparing with other relevant algorithms, it has proved that the algorithm proposed in this paper has certain progressiveness and feasibility [23].

To further validate the effectiveness of the proposed algorithm, we conducted a series of ablation experiments. These experiments were based on the CRNN+CTC model and progressively introduced additional feature extraction and attention mechanisms [24]. Here are the designs and results of the ablation experiments:

- CRNN+CTC: This serves as our baseline model, combining Convolutional Neural Networks (CNN) with Recurrent Neural Networks (RNN) and Connectionist Temporal Classification (CTC) loss function for end-to-end handwriting text recognition.
- CRNN+CTC+3 convolutional layers: In this variant, we added three convolutional layers on top of the CRNN+CTC to enhance the model's capability for feature extraction from images. This helps the model to better capture the details and structure of handwritten text.
- CRNN+CTC+3 convolutional layers + multi-scale channel attention: Building upon the second variant, we further introduced a multi-scale channel attention module. This module adaptively adjusts the importance of different channels, thereby improving the model's sensitivity to key features, especially when dealing with adjacent identical characters.

As shown in the **Table 3**, the introduction of additional convolutional layers and the multi-scale channel attention module led to an increase in the model's accuracy (ACC) and a corresponding decrease in the character error rate (CRE). The significant performance improvement after incorporating the multi-scale channel attention module demonstrates its effectiveness in handwriting text recognition. These ablation experiment results further confirm the advancement and practicality of our proposed algorithm [25].

Table 3. Ablation experiment.

Methods	**ACC (%)**	**CRE (%)**
CRNN+CTC	90.24%	5.46%
CRNN+CTC+3 Conv Layers	90.42%	5.35%
CRNN+CTC+3 Conv Layers + Multi-Scale Channel Attention	91.24%	5.04%

5. Conclusion

In conclusion, the proposed handwritten text line recognition algorithm has demonstrated significant effectiveness through comparative experiments, outperforming existing classic algorithms in terms of recognition accuracy and character error rate. The algorithm's particular proficiency in recognizing adjacent identical characters highlights its advanced capabilities within the scope of this study.

However, there are limitations to the current research that warrant attention. The algorithm's performance is primarily tailored to the domain of primary school mathematics exams, which may limit its generalizability to other academic subjects or linguistic contexts. Additionally, the dataset, while robust for this study, may not capture the full spectrum of handwriting variations and text layouts, potentially impacting the model's robustness in more diverse real-world applications.

Looking ahead, we will expand the data set to include a wider range of handwriting styles, and explore the integration of our algorithm with advanced machine learning technology to enhance its adaptability and robustness.In the future, it will tend to realize multimodal interaction, which can not only combine handwriting, voice, but also touch and other input methods to improve the accuracy of recognition and user experience. At the same time, it is necessary to pay attention to the joint use of deep learning and other artificial intelligence technologies in the future, so as to deal with complex fonts and different writing styles more effectively, achieve high accuracy and preserve spelling mistakes.We also aim to optimize the model for complex text structures and diverse real-world scenes to ensure its applicability in various educational and professional environments. This future work will focus on improving the performance of the algorithm and promoting the progress of document analysis and information retrieval technology.

Author contributions: Conceptualization, ZW and JX; methodology, JX and ZW; software, JX; validation, YW and XW; formal analysis, JX; investigation, JX and YW; resources, XW; data curation, JX; writing—original draft preparation, JX and YW; writing—review and editing, ZW, YW and XW; visualization, XW; supervision, ZW; project administration, ZW; funding acquisition, ZW. All authors have read and agreed to the published version of the manuscript.

Consent for publication: Written consent to publish this information was obtained from study participants.

Availability of data and material: The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Ethical approval: Not applicable.

Conflict of interest: The authors declare no conflict of interest.

References

1. Epshtein B, Ofek E, Wexler Y. Detecting text in natural scenes with stroke width transform, 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 2010, pp. 2963-2970.
2. Matas J, Chum O, Urban M, Pajdla T. Robust wide-baseline stereo from maximally stable extremal regions, Image and Vision Computing, 2004, 22(10), 761-767.
3. Dalal N, Triggs B. Histograms of oriented gradients for human detection, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005, pp. 886-893
4. Li H, Doermann D, Kia O. Automatic text detection and tracking in digital video, in IEEE Transactions on Image Processing, 2000, 9(1), 147-156.
5. Zhang Z, Shen W, Yao C, Bai X. Symmetry-based text line detection in natural scenes, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 2558-2567.
6. Huang W, Lin Z, Yang J, Wang J. Text localization in natural images using stroke feature transform and text covariance descriptors, 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 2013, pp. 1241-1248.
7. Lukas N, Matas J. A method for text localization and recognition in real-world images. In: Kimmel, R., Klette, R., Sugimoto, A. (eds) Computer Vision – ACCV 2010. ACCV 2010. Lecture Notes in Computer Science, 2010, vol. 6494. Springer, Berlin, Heidelberg.
8. Zhi T, Huang W, He T, Qiao Y. Detecting text in natural image with connectionist text proposal network. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. Lecture Notes in Computer Science, 2016, vol. 9912. Springer, Cham.
9. Alex G, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 2005, 18(5-6), 602-610.
10. Liao M, Shi B, Bai X, et al. Textboxes: A fast text detector with a single deep neural network. Proceedings of the AAAI Conference on Artificial Intelligence, 2017, 31(1), 4161-4167.
11. Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. Lecture Notes in Computer Science, 2016, vol. 9905. Springer, Cham.

12. Liao M, Shi B, Bai X. Textboxes++: A single-shot oriented scene text detector. IEEE transactions on image processing, 2018, 27(8), 3676-3690.
13. Shi B, Bai X, Belongie S. Detecting oriented text in natural images by linking segments. 017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 3482-3490.
14. He P, Huang W, Qiao Y, et al. Reading scene text in deep convolutional sequences. Proceedings of the AAAI conference on artificial intelligence, 2016, 30(1), 3501-3508.
15. Graves S, Fernández F, Gomez JS. Connectionist temporal classification. In Proceedings of the 23rd International Conference on Machine Learning. 2006, 369-376
16. Shi X, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition, in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(11), 2298-2304.
17. Liao M, Pang G, Huang J, et al. Mask TextSpotter v3: Segmentation proposal network for robust scene text spotting. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, JM. (eds) Computer Vision – ECCV 2020. Lecture Notes in Computer Science, 2020, 12356. Springer, Cham.
18. Liu Y, Chen H, Shen C, et al. Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.
19. Wang W, Xie E, Li X, et al. Shape robust text detection with progressive scale expansion network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.
20. Liao M, Zou Z, Wan Z, et al. Real-time scene text detection with differentiable binarization and adaptive scale fusion, In IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1), 919-931.
21. Woo S, Park J, Lee J, Kweon IS. CBAM: Convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, 2018, 11211. Springer, Cham.
22. Long, S., He, X. & Yao, C. Scene Text Detection and Recognition: The Deep Learning Era. Int J Comput Vis 129, 161–184 (2021). https://doi.org/10.1007/s11263-020-01369-0
23. RSCA: Real-Time Segmentation-Based Context-Aware Scene Text Detection Jiachen Li, Yuan Lin, Rongrong Liu, Chiu Man Ho, Humphrey Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2021, pp. 2349-2358
24. Geetha, R.; Thilagam, T.; Padmavathy, T. Effective offline handwritten text recognition model based on a sequence-to-sequence approach with CNN–RNN networks. Neural Computing & Applications, 2021, Vol 33, Issue 17, p10923
25. Pandey, D., Pandey, B.K. & Wairya, S. Hybrid deep neural network with adaptive galactic swarm optimization for text extraction from scene images. Soft Comput 25, 1563–1580 (2021). https://doi.org/10.1007/s00500-020-05245-4