





The Construction Of "Internet+5G" Vocal Music Teaching Mode in Universities Based on Embedded Systems

Li Han^{1*} and Mingyue Liu²

^{1,2}Zhengzhou Vocational College of Finance and Taxation, Zhengzhou, Henan 450000, China
¹18203693167@163.com, ²lmy123321lmy2022@163.com

Corresponding author: Li Han, 18203693167@163.com

Abstract. The national strategy of "Internet+5G" is deeply rooted in various fields of life, facing many changes in the new situation. The new modes of vocal music teaching under the background of "Internet+5G" such as catechism, mobile APP and WeChat are emerging. They have broken the traditional classroom vocal teaching mode and reorganized the time and space relationship of vocal teaching. Optical score recognition, as a key technology for the symbolization of score content, helps to store and edit music files, and has important application values in the fields of music information retrieval and computer-aided teaching. The optical score recognition algorithm based on the general framework has the problems of complicated processing steps and low accuracy, while the algorithm based on deep learning effectively simplifies the general framework, but its recognition accuracy needs further improvement, and the recognition error for difficult notes is large and the model training is time-consuming. The experimental results show that the average symbol error rate of the improved network model decreases to 0.3234%, and the training time of the model is about one-third of that of the traditional convolutional recurrent neural network, which is optimized in terms of recognition accuracy and training time.

Keywords: "Internet+5G"; new mode of vocal music teaching; optical score recognition; convolutional neural network.

DOI: <https://doi.org/10.14733/cadaps.2024.S8.228-242>

1 INTRODUCTION

In the 1960s, the Internet originated from a research conducted by the US federal government to build a robust and fault-tolerant computer communication network [12]. In the early 1980s, the Internet was mainly used in the direction of regional academic exchanges and military communications. Since then, the National Science Foundation and some private foundations for commercial use have contributed to the widespread development of new network technologies

around the world. Beginning in the 1990s, the combination of business networks and enterprises ushered in a new era of the Internet. Since then, the Internet has entered a stage of rapid development. Traditional media, such as television and radio, have developed into Internet TV and online radio in the Internet age. Newspapers, books, music, etc. also have new forms of presentation due to the Internet, and become part of blogs and network resources [17]. Driven by the Internet, music education has also entered a new era of development. Traditional music education adopts face-to-face teaching and guidance, while music education in the Internet age breaks some of the limitations of traditional music education. It uses the Internet as an educational platform. On the online platform, the rich and diverse learning resources also make students' learning more interesting, and their enthusiasm for learning becomes even higher [6]. We don't need to spend a lot of time looking for relevant information, and then filter them one by one [21]. In the process of learning, we only need to enter the keywords of the information we need in various search engines on the Internet, and then we can get a variety of related information. We are looking for information related to the materials, which greatly improves the spatiality of autonomous learning for students in the process of learning, and if we can make good use of Internet resources, we can also get more authoritative and first-hand information [16].

The new mode of vocal music teaching in the context of "Internet+5G" makes it difficult to identify and record music scores [19][14]. This has led to the emergence of Optical Music Recognition (OMR) technology, Its function is to convert the printed music into characters. This is not only an innovation in music, but also an achievement of computers in reading digital files and images. [5][4]. Although some commercial recognition software exists, the recognition methods still have problems such as poor noise immunity, so there is an urgent need to continue to research OMR algorithms with better robustness and higher accuracy [20].

OMR is similar to Optical Character Recognition (OCR), but the problems faced in the actual recognition process are quite different. First of all, for most characters, different characters differ greatly in shape and the recognition process is mainly based on contour recognition, while notes are almost similar in shape, basically a combination of vertical lines and circles or near circles, and their recognition depends mainly on subtle differences [3]. The difference in note position alone makes the four notes represent different information, because the vertical direction of the notes determines the pitch of the notes. Thus, it can be seen that the focus of music score recognition is on note recognition, and the focus of note recognition is different from that of OCR [11].

In this paper, we will briefly explain the research status of vocal music teaching inside and outside of China, clarify the purpose of this paper, and analyze the problems of vocal music teaching in universities in the era of "Internet+5G". In order to improve the vocal music teaching mode in colleges and universities by combining CNN and SUR, we will verify and analyze the problems through corresponding experiments.

2 THE RELATED WORKS

The new mode of vocal music teaching under the background of "Internet+5G" is in the modern network catechism, cell phone APP, WeChat and other ways to present. Modern network catechism, cell phone APP, WeChat, etc. have reconstructed the teaching mode and teacher-student relationship, broken the limitation of space, and allowed the massive vocal teaching resources to be widely disseminated in a timely manner.

Due to the special nature of vocal music teaching, it is not possible to rely on the simple operation of the Internet and computers alone, but also needs to be done through face-to-face teaching by teachers. Vocal music is an art of sound, and its requirements for the quality of music score are very high. The new mode of vocal music teaching under the background of "Internet+5G" is usually

conducted through the Internet. Due to various reasons such as technology, equipment, network speed and surrounding environment, the transmission and recording of music can be lost.

At present, note recognition adopts the idea of layered recognition. The first step requires the detection of the score and the separation of the score from the pentatonic score; the second step implements the segmentation operation based on the note features, obtaining the corresponding note primitives and classifying them according to the categories determined by the recognition [9]. Although the pentatonic information is the auxiliary information of notes, i.e., the position of notes in the pentatonic spectrum determines their pitch and other information, to a certain extent, the presence of pentatonic spectrum will interfere with the recognition of notes, so the accuracy of pentatonic detection and deletion will directly affect the accuracy of note recognition. The cross-over characteristics of the pentatonic spectrum and the notes make it difficult to delete the pentatonic spectrum. When the degree of pentatonic deletion is not enough, the residual pentatonic score will affect the recognition of notes [2]. On the contrary, when the pentatonic score is deleted excessively, the note shape will be broken, which will also make the note recognition more difficult.

Usually, the detection of pentatonic scores is mainly divided into methods based on statistical transformation and methods based on structural search [13]. Statistical transform-based methods use horizontal projection, Hough transform, wavelet transform and other methods for detection of pentatonic spectra, which have strong noise immunity but the score images are susceptible to deformation [10]; Based on this, Miyao et al. used a structural search algorithm in order to improve the resistance to variation, which achieves the purpose by reducing the level of noise resistance; in terms of detection correctness, Cardoso et al. adopted a detection strategy of stable paths, which breaks the limitation of domain knowledge and increases the correct rate to 98.6% [1]. Since the recognition of musical score images is difficult to control the quality of the data, it is difficult to achieve a balance between these two methods and guarantee better results. Dalit et al. provided many algorithms in separating pentatonic scores, such as skeleton method, line adjacency graph method, vector line method, etc.; there is also T.L. Wu et al. implemented the operation of pentatonic separation for XGBoost handwritten scores by multi-dimensional local binary model, and the method has a low training requirement, which can improve the accuracy by reaching only 0.05%. pugin et al. directly skipped the pentatonic separation step and directly used Hidden Markov Model (HMM) for notes recognition, which increased the difficulty factor of model computation and data modeling, and obtained recognition results that differed more from the expected results [18].

Several research works have applied neural networks to subtasks in a generic framework, i.e., experimenting with Convolutional Neural Networks (CNN) in stages such as pentatonic deletion, note recognition, and note classification [8]. Calvo-Zaragoza et al. viewed the detection of pentatonic scores as a classification task and implemented pentatonic detection using CNNs. Each pixel is labeled as a pentatonic or note and even without applying post-processing, the results are still better than most traditional methods [15].

Although image processing technology has been developing at a high speed, optical music score recognition is still slow after years of efforts. However, the following problems still need to be solved in the process of OMR algorithm research: first, the calculation and recognition of most framework algorithms are generally complex and highly difficult; it is difficult to balance between noise resistance and deformation resistance in the detection process; spectral line separation affects the difficulty level of recognition of notes with dots to a certain extent; note features are diverse, and it is extremely difficult to achieve recognition and feature classification operations by one algorithm, and there is great variability in the results obtained by different classification algorithms. These problems make the overall recognition accuracy of the OMR task not high enough; the end-to-end training of the deep neural network algorithm simplifies the complexity of the general framework, no longer analyzes and studies the key steps in the OMR task separately, and reduces the error transmission in multiple steps. However, the OMR task is more sensitive to detailed information,

especially for the recognition of difficult notes, the insufficient will seriously affect accuracy; the note recognition method using the BiLSTM model has a slow convergence speed, and when the parameters increase, the model training costs time increases, so that each parameter modification will take a lot of time to retrain. An optical music score recognition algorithm, which minimizes the steps of the whole process and improves the recognition accuracy of the algorithm. By optimizing the convergence speed of the model, the training time of the model is reduced.

It can be seen that although image processing technology is widely used and rapidly developed in many fields, its research in the field of music score recognition is relatively slow and there are still many gaps. Therefore, this paper improves the algorithm of music score recognition, simplifies the recognition process as much as possible, reduces the recognition error rate, improves the model convergence speed, and achieves the purpose of shortening the training time.

3 COMBINING MULTI-SCALE RESIDUAL CNN AND SRU FOR OPTICAL MUSIC SCORE RECOGNITION

With the rapid development of technology, cell phone is not only a communication device, but also a high-tech product that affects every aspect of our life. Mobile APPs for various purposes have emerged, bringing great convenience to people's lives. APPs about vocal teaching also came into being, and many of them can be searched in the app store in any cell phone to choose from. The most representative one with the largest number of users is the APP music software "Niu Ban" developed by the team led by famous singer and musician Robin Hu in 2015. The star music classroom in the software has already gained good results in its first month of broadcasting, with a single episode of "Don't break my heart" already reaching a million views within 36 hours, showing a great impact.

In 2011, WeChat software quickly attracted a large number of young customers with its free, real-time communication, border and easy operation features, driving its market share, and five years later, its smartphone users already account for more than 90% of smartphone users in the market. WeChat's inclusion in education would be invaluable to the development of education. First, WeChat connects schools with students, allowing both to fully interact on the platform, which directly affects the quality of teaching and learning. Secondly, WeChat has diversity and speed in education. Both sides of teaching can make full use of voice, video as well as graphics to express the questions that teachers want to express, and students can ask the questions they want to ask and get the right answers according to their actual needs, thus learning new knowledge. With the development of WeChat, more and more users are willing to accept it, and it has become an indispensable social platform in the lives of most people.

OMR mainstream algorithms are divided into general framework-based and deep learning-based algorithms. The OMR algorithm based on the general framework divides the entire task into a single sub-task, and each step is a complex task. Although the recognition accuracy in each step can be improved by optimizing the algorithm, the wrong transfer characteristics will directly affect the deletion of the staves. The identification of subsequent notes, and the lack of a unified method in note identification, needs to be identified according to the characteristics of the notes. The deep learning-based algorithm effectively avoids such problems. The entire image is processed as input, and a universally applicable method can be obtained through parameter learning and model training. However, there are insufficient recognition accuracy and slow model convergence and time-consuming The problem. Therefore, based on deep learning, this paper proposes an algorithm with strong feature recognition ability and fast model convergence.

Bi-directional Simple Recurrent Network (BiSRU) combination of optical sheet music recognition method based on a deep learning framework, which constitutes an MF-RC-BiSRU network, mainly divided into three parts: image pre-processing, feature extraction and note recognition. In the first

part, the height and width of the score are set uniformly as needed, while the simulation of the real environment is realized by the incorporation of additive Gaussian white noise, additive Berlin noise and elastic deformation to increase the interference factors of the score image. In the second part, the note feature information in the image is obtained by the model in this paper, and the deep and shallow feature information are fused in the same feature map. In the third part, the dimensions of the feature sequences are transformed and output, and note recognition is achieved by BiSRU, and its classification is realized by connectionist temporal classification (CTC), as shown in Figure 1 for common notes.



Figure 1: Example of Notes.

The optimization goal of the whole training process is achieved by the loss function. The loss function such as mean squared deviation loss function and cross entropy loss function are commonly used loss functions, and the model is usually trained with custom loss functions according to the actual requirements.

The mean squared regression problems, mainly for prediction problems with specific data, and it is defined as follows.

$$\text{MSE}(y, y') = \frac{\sum_{i=1}^n (y_i - y'_i)^2}{n} \quad (1)$$

Where, y_i denotes the correct result corresponding to the first i data in a batch, y'_i denotes the predicted value of the model for the data y_i , and n denotes the number of data in a batch.

Cross entropy is often used to judge how close the model output is to the expected value, and is widely used in classification tasks. For a given probability distribution P and Q , the cross-entropy of P is represented by Q :

$$H(p, q) = -\sum_x p(x) \log q(x) \quad (2)$$

Cross-entropy is a description of the distance between two probability distributions, while probability distribution is a description of the probability of occurrence of different events, events in finite cases, with probability $p(X = x)$ satisfied:

$$\forall x \quad p(X = x) \in [0,1], \sum_x p(X = x) = 1 \quad (3)$$

That is, the probability of any event occurring ranges from $0 \sim 1$. For the case where the output of the neural network is not a probability, this can be achieved by Softmax transformation. Suppose the original neural network output is y_1, y_2, \dots, y_n , and after Softmax the output is :

$$\text{softmax}(y)_i = y'_i = \frac{e^{y_i}}{\sum_{j=1}^n e^{y_j}} \quad (4)$$

The output of the original neural network is used as the confidence level to generate the new output, which satisfies the probability distribution. When the cross entropy is used as the loss function of the neural network, P denotes the correct result of the corresponding input, and Q denotes the predicted value of the corresponding neural network.

Suppose θ is a parameter in the model, and the corresponding loss function value is $J(\theta)$ when the parameter takes the value determined, the gradient descent method updates the parameter θ by iteratively, and keeps the parameter updated in the direction of smaller total loss along the opposite direction of the gradient. Where, the x axis represents the value of the parameter θ , the y axis represents the value of the loss function corresponding to the parameter $J(\theta)$, the gradient descent method shifts the parameter θ to the left along the x axis, and the parameter update equation is:

$$\theta_{n+1} = \theta_n - \eta \frac{\partial}{\partial \theta_n} J(\theta_n) \quad (5)$$

where η denotes the learning rate, and $\frac{\partial}{\partial \theta} J(\theta)$ denotes the gradient of the parameter θ .

Parameter regularization means adding a penalty term to the loss function to constrain the coefficients so that the model cannot be fitted arbitrarily to the random noise in the training data. The commonly used formula for L2 regularization is as follows:

$$Loss_{reg} = Loss + \frac{\lambda}{2n} \sum_w w^2 \quad (6)$$

Where, n is the number of training samples and λ is the regularization hyperparameter. The loss function with penalty term is biased against the weights w and the bias b to obtain :

$$\begin{aligned}\frac{\partial \text{Loss}_{\text{reg}}}{\partial w} &= \frac{\partial \text{Loss}}{\partial w} + \frac{\lambda}{n} w \\ \frac{\partial \text{Loss}_{\text{reg}}}{\partial b} &= \frac{\partial \text{Loss}}{\partial b}\end{aligned}\quad (7)$$

It can be seen that the bias for the weights w is increased by one, while for the bias b remains unchanged, so that in gradient descent w is updated to:

$$w = w - \eta \frac{\partial \text{Loss}_{\text{reg}}}{\partial w} = w - \eta \left(\frac{\partial \text{Loss}}{\partial w} + \frac{\lambda}{n} w \right) = \left(1 - \frac{\eta \lambda}{n} \right) w - \eta \frac{\partial \text{Loss}}{\partial w}\quad (8)$$

The coefficients of the weight w will be changed by the change of λ , so as to achieve the limitation of the weights and avoid the overfitting problem. Another parameter regularization is the L1 regularization, which is calculated as follows:

$$\text{Loss}_{\text{reg}} = \text{Loss} + \frac{\lambda}{n} \sum_w |w|\quad (9)$$

To solve the potential non-convergence problem of the model, the residual formula CNN is used in this paper. The traditional CNN input data x is passed through a convolutional layer and a nonlinear activation function layer to obtain the output $y = H(x)$. In this paper, we design the residual formula CNN as shown in Figure 2. The "dead zone" state can be expressed as follows

$$f(x) = \begin{cases} 0.01x & x \leq 0 \\ x & x > 0 \end{cases}\quad (10)$$

Where x is the activation function input, and $f(x)$ is the activation function output.

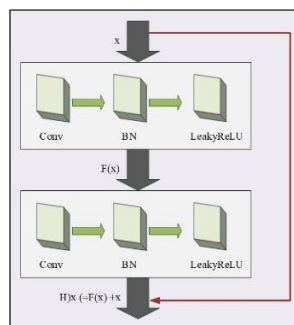


Figure 2: Schematic Diagram of the Residual Structure.

In this paper, we use the SRU module, which removes the strong constraint between consecutive moment states and uses the weaker cyclic nature and higher parallelism to make the computation of gate states dependent.

The current moment t under the forgotten door of the hidden cell f_t , the cell status c_t is defined as:

$$f_t = \sigma(w_f x_t + b_f) \quad (11)$$

$$c_t = f_t * c_{t-1} + (1 - f_t) * g_t \quad (12)$$

where σ denotes the Sigmoid activation function, w_f and b_f are the parameter matrix and bias of the forget gate, respectively, $*$ denotes the dot product operation of the corresponding elements between the matrices, and g_t denotes the linear transformation of the input x_t at the current moment:

$$g_t = w x_t \quad (13)$$

$$\tilde{c}_t = g(c_t) \quad (14)$$

$$r_t = \sigma(w_r x_t + b_r) \quad (15)$$

$$h_t = r_t * g(c_t) + (1 - r_t) * x_t \quad (16)$$

By comparing SRU and LSTM networks from input to output, we can see that the previous output vector h_{t-1} and the current input vector x_t as input. The optimization of SRU is to learn the inter-sequence correlation by the previous moment cell state c_{t-1} , which does not depend on the previous moment output h_{t-1} , so that the computation of gate state can be synchronized at all moments, without waiting for the previous moment output h_{t-1} to finish the computation before starting the next moment forgetting the gate f_t , which provides theoretical support for the parallel implementation of a large number of operations.

When training the score image data with the powerful temporal modeling capability of RNN, the network needs to provide the expected output, i.e., the corresponding label, for each note in the sequence. However, RNN requires strict alignment of the corresponding labels with the original image pixels during the loss computation. Therefore, this paper presents the optimal choice of RNN model for processing serialized data. Unlike the other loss functions, this function is used to train the model and learn the parameters for the end-aligned data set.

For the input x of length T , define RNN with m dimensional real vector input $(R^m)^T$, n dimensional real vector output $(R^n)^T$ and its weight vector w , and obtain a continuous mapping from input space to target space $\mathcal{N}_w : (R^m)^T \mapsto (R^n)^T$, the output sequence of the network is $y = \mathcal{N}_w(x)$, and define y_t^k as the probability that the output is k at the time step of t . The network outputs at different moments are independent of each other, the distribution is defined on the set L^T :

$$p(\pi | x) = \prod_{t=1}^T y_{\pi_t}^t, \forall \pi \in L^T \quad (17)$$

Where L^T is called a path and labeled by π , $L^{\leq T}$ denotes the set of possible labels, i.e., the original set of label characters L with length less than or equal to the T sequence, i.e., $L^{\leq T}$ is obtained by removing the recurring labels and gaps in the path. Also, define the many-to-one mapping $\mathcal{B} : L^T \mapsto L^{\leq T}$, then the probability of the input x for the label sequence is $l \in L^{\leq T}$ $p(l | x)$ defined as:

$$p(l | x) = \sum_{\pi \in \mathcal{B}^{-1}(l)} p(\pi | x) \quad (18)$$

$$h(x) = \arg \max_{l \in L^{\leq T}} p(l | x) \quad (19)$$

4 ANALYSIS OF SIMULATION RESULTS

In this paper, we test the performance of the proposed optical score recognition algorithm combining multi-scale residual CNN and SRU. The data used for the experiments are from the open dataset The PRIMus Dataset (Printed Images of Music Staves), which contains about 87687 real score examples, each composed of a single line of pentatonic score and divided into about 4-7 measures by bar lines. Most of the examples contain not only simple sequences of notes, such as leans and dots. The data set itself shows good clarity and stationarity, which needs to be extended by introducing disturbance factors in some of the samples and dividing them into 10 parts, with the validation set and test set taking any one part and the rest being the training set.

There is no unified specification for the evaluation metrics of OMR algorithms, and different researchers have their own evaluation metrics. Some studies use three evaluation metrics, namely, pitch accuracy, note accuracy and pitch length accuracy, to measure the accuracy, but considering that there may be inconsistency between note correctness and pitch correctness of a note, and in real note recognition, an error in either of the evaluation metrics should be judged as a recognition error. Therefore, in this paper, the algorithm is evaluated using common metrics such as Sequence Error Rate and Symbol Error Rate. Symbol Error Rate (SER) refers to the ratio of the average number of editing operations such as insertion, modification or deletion required to generate label sequences

from the predicted sequences to the current sequence length. There is no absolute correlation between the above sequence error rate and symbol error rate. The sequence error rate describes the error ratio of the concentrated test cases, while the symbol error rate summarizes the error situation of the notes in the cases, so the measurement of the accuracy of note recognition in this paper focuses more on the evaluation index of symbol error rate, but the sequence error rate still has guiding significance in many fields of application.

In the convolutional recurrent neural network C-BiLSTM algorithm, the validation of the effect of the augmented dataset is achieved by adding interference factors for dataset expansion, and the model will be trained on the original and augmented datasets and tested on the same test set, respectively, as shown in Figure 3 for the comparison of recognition error rate results.

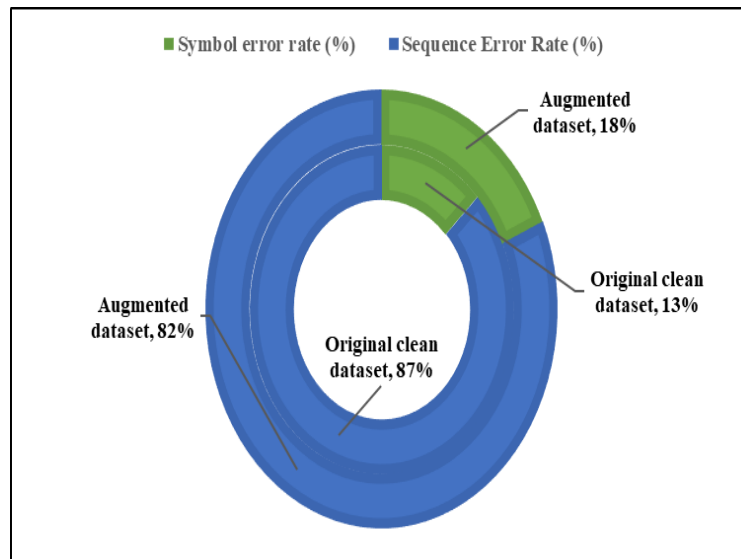


Figure 3: Comparison of Test Results in Different Models.

As can be seen in Figure 3, the sequence error rate of the model trained by the enhanced dataset decreases more significantly, and the experiment shows that the expansion of the dataset by data enhancement effectively identifies images with different light sources and different print quality. However, the symbol error rate does not improve or even the recognition accuracy decreases, because the addition of noise does cause some distortion and interference to the data, and since the notes themselves are small, the recognition of individual notes will be affected if some pixels have a large offset. The CNN structure of the algorithm is simple and cannot learn the deviations effectively, so the feature capability of the model can be improved by subsequent optimization to reduce the symbol error rate.

The CNNs in the C-BiLSTM network are improved into residual CNNs to form a residual convolutional recurrent neural network (RC-BiLSTM), and the performance of the models before and after the improved CNNs are compared in the same experimental environment, and the sign error rates of the C-BiLSTM network and RC-BiLSTM are trained separately. Figure. 4 which reflects that the loss value of RC-BiLSTM network is lower than that of C-BiLSTM network in each iteration, and the loss value of RC-BiLSTM network has been reduced to 5 and stabilized after training, while the loss value of C-BiLSTM network is only reduced to about 10 and always has large fluctuations.

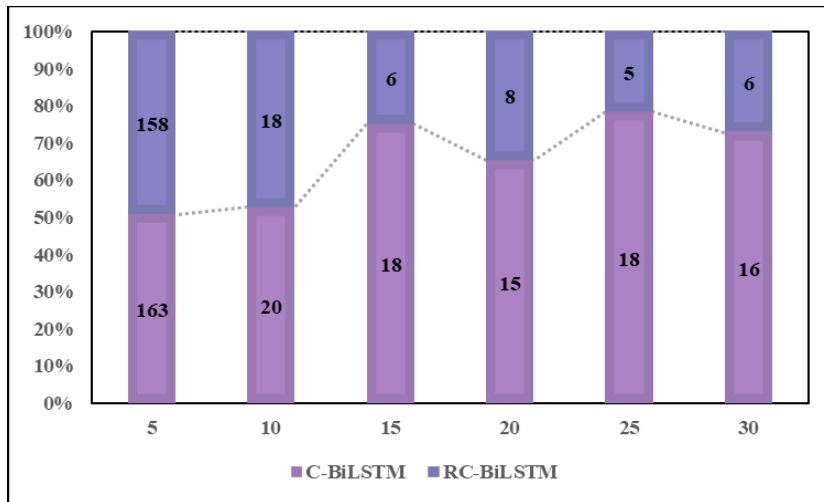


Figure 4: Comparison of the Training Loss of C-BiLSTM and RC-BiLSTM Networks.

Meanwhile, the symbolic error rates in both algorithms were compared in the validation set after every 1000 iterations, and the results are shown in Figure 5. During the whole training process, the symbol error rate of C-BiLSTM network decreases significantly, but during the iteration process, the symbol error rate appears to increase significantly, which shows that the model does not converge, while the symbol error rate of RC-BiLSTM network can be reduced steadily with less volatility, which shows that the note recognition accuracy of RC-BiLSTM network is significantly improved. This indicates that the improve the model accuracy, but also solve the model degradation problem and enhance the model generalization ability.

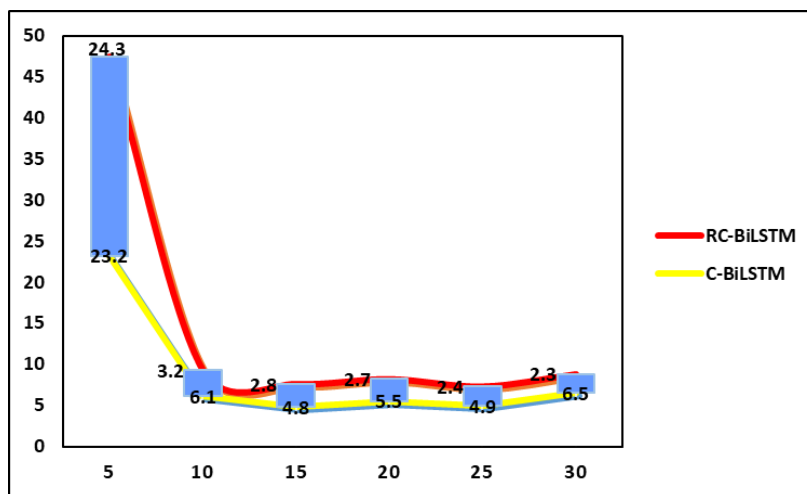


Figure 5: Comparison of Symbol Error Rate of C-BiLSTM and RC-BiLSTM Networks.

Figure 6 shows the comparison of the symbol error rate of RC-BiLSTM network and MF-RC-BiLSTM network on the validation set, from which it can be seen that the symbol error rate of MF-RC-BiLSTM

network is significantly reduced to less than 0.5%, which shows that the multi-scale feature fusion can effectively improve the extraction ability of the model for note features and further improve the correct recognition rate of notes.

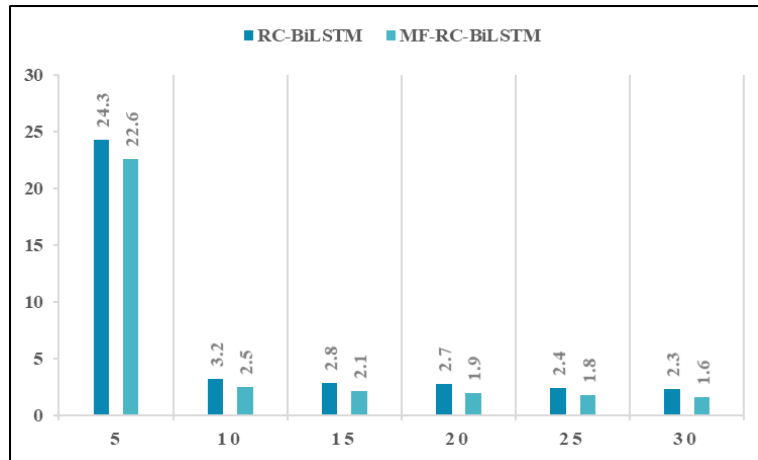


Figure 6: Comparison of Symbolic Error Rates of RC-BiLSTM and MF-RC-BiLSTM Networks.

Comparing the performance of the four network models on the same test set, in this paper is optimal in terms of both sequence error rate and symbol error rate, with a sequence error rate of about 1.4571% and a symbol error rate of about 0.3219%, while the sequence error rate of the C-BiLSTM network in the same data set is about 14.3498% and the symbol error rate is about 3.2480%.

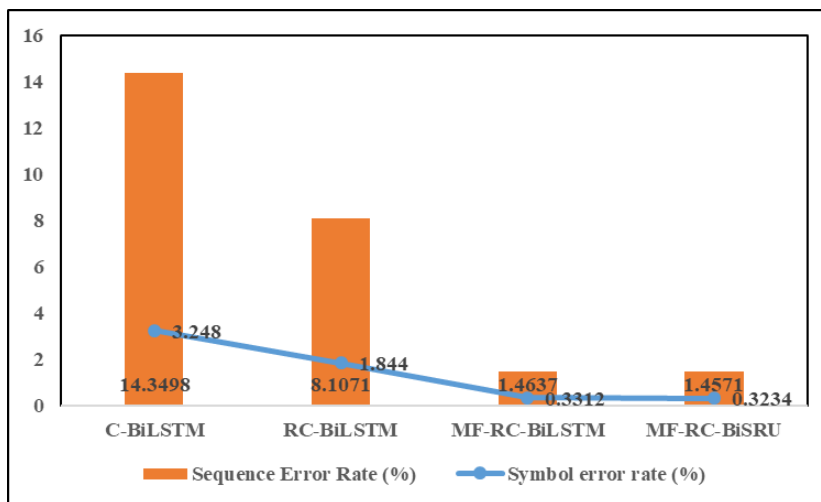


Figure 7: Comparison of the Accuracy of Different Networks.

In this paper, MF-RC-BiSRU is compared with other methods in the existing literature for experiments. The experimental results are shown in Figure. 8. Firstly, it can be seen that the MF-

RC-BiSRU method achieves better results in both symbol error rate and sequence error rate. In the absence of a priori knowledge, the CNN-STN algorithm note head as the feature basis for training can achieve the recognition of variant notes with a low error rate. The DWD algorithm can recognize all types of notes, but its accuracy is affected by the differences between different types of notes, and the recognition error rate reaches 20%, and the recognition rate of six consecutive notes only reaches The overall recognition accuracy rate is low.

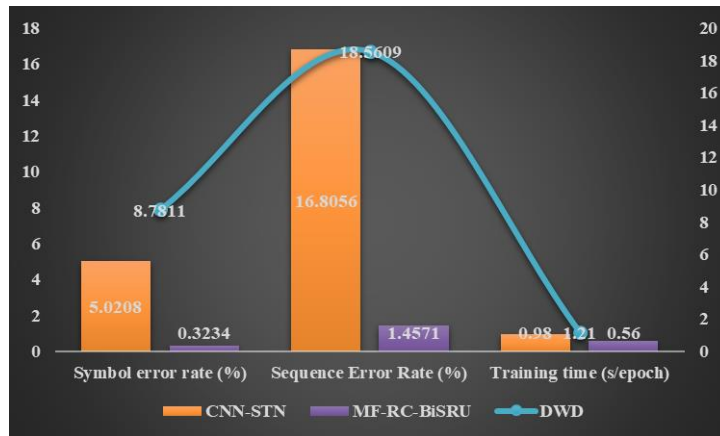


Figure 8: Performance Comparison of Different Methods.

The algorithm is compared with the recognition results of the existing commercial software (Capella-scan, PhotoScore and SmartScore) on 50 scores. 1132 notes are found in 50 scores, of which 5 scores are image enhanced scores with 107 notes. The experimental results are shown in Figure. 9. The proposed algorithm achieves the optimal symbol error rate and sequence error rate, and only 5 notes are incorrectly identified, which are mainly distributed in 2 scores. Among the three major commercial software, Capella-scan has the lowest sequence error rate and symbol error rate, while SmartScore is prone to a large number of note recognition errors due to the error in dividing bar lines.

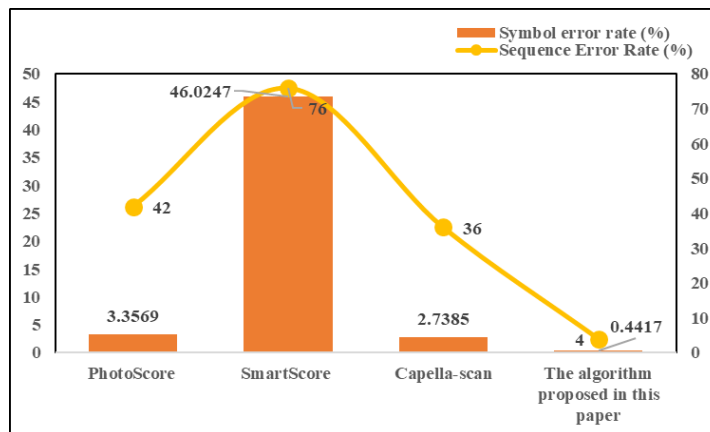


Figure 9: Comparison of the Recognition Accuracy of Different Commercial Software.

5 SUMMARY AND OUTLOOK

The purpose of this paper is to study an optical music recognition method with a simple algorithm process, high recognition accuracy and fast model convergence. This paper focuses on optimizing the model in four aspects: image pre-processing, feature extraction capability, model convergence speed and loss function. In the image preprocessing, the data set is expanded by three image enhancement methods to enhance the robustness of the model.

The ultimate goal of OMR is to convert any musical score image into symbolic form with high accuracy. Although the OMR algorithm proposed in this paper has obtained better results in comparative experiments, it has not yet addressed more complex scores, such as chords and polyphonic scores. In addition, the data set used in this paper are all single-line pentatonic scores, and the recognition of notes is limited to the information of notes on one line of pentatonic score, while the whole score may contain more information such as text and richer difficult notes.

In response to the above problems, further research can be carried out from the following two aspects: (1) The PrIMus Dataset data set used in this paper is all single music scores, that is, composed of a single line of staff and note information, not all information. The entire score of the data does not appear in the data, such as complex music information such as chords and multiple voices. The algorithm proposed in this paper works well in the dataset, but the recognizable range is relatively small for real-world OMR tasks. Therefore, OMR follow-up research can expand the data set so that the data set includes not only single-line scores, but also whole scores and more difficult chords and other scores. Algorithms are tuned for complex datasets. (2) The symbol error rate of the algorithm in this paper for the identification of difficult notes has been reduced, but there are still identification errors, such as inaccurate identification of some key signatures. It may be due to the fact that the proportion of difficult notes in the actual score image is very small, and the features with a small amount of data are insignificant for the amount of model parameters, which leads to the poor learning effect of the model on such features, even if the model feature dimension is high, for Data that appears infrequently is also prone to overfitting, which leads to errors in the testing process. Therefore, when the amount of data is small but the note recognition rate is still required, some prior knowledge constraints need to be added, such as cross-complementing the time signature, key signature and note information of each measure, and further learning through parameters with constraints Improve the feature expression ability of the model.

Li Han, <https://orcid.org/0009-0009-8905-2846>

Mingyue Liu, <https://orcid.org/0009-0001-6550-1207>

REFERENCES

- [1] Chen, C.: Design and Research of Intelligent Score-Reading Guitar Playing Robot, Guangdong University of Technology, 2021. <https://doi.org/10.27029/d.cnki.ggdgu.2021.001027>.
- [2] Fan, Y.: Research on Music Score Correction Algorithm and Generation of Universal Music Score set, Beijing University of Posts and Telecommunications, 2020. <https://doi.org/10.26969/d.cnki.qbydu.2020.003225>.
- [3] Guo, C.: Research on Humming Audio Sheet Music Recognition Technology Based on Deep Learning, Wuhan University, 2018.
- [4] Guo, L.W.; Guan, X.; Li Mang.: Difficulty Level Recognition of Piano Scores Based on Measure Learning Support Vector Machine, Journal of Intelligent Systems, 13(02), 2018, 196-201.
- [5] Huang, Z.Q.; Jia Xiang.; Guo, Y.F.; Zhang J.: End-to-End Musical Note Recognition Based on Deep Learning, Journal of Tianjin University (Natural Science and Engineering Technology Edition), 53(06), 2020, 653-660.

- [6] Jia, J.: Analysis of the Reform Practice Of Vocal Music Teaching System Of Senior Teachers in the era of "Internet+", China Ethnic Expo, (09), 2021, 103-104+114.
- [7] Jia, W.: Music Score Image Recognition System Based on Embedded Platform, Beijing University of Posts and Telecommunications, 2018.
- [8] Jia, X.: Research on Printed Music Score Recognition Method Based on Deep Learning, Beijing University of Technology, 2020. <https://doi.org/10.1109/ACCTCS52002.2021.00042>
- [9] Jing, Y.: Deep Learning Algorithm Composition System based on Music Score Recognition, Nanjing Arts Institute, 2020. <https://doi.org/10.27250/d.cnki.gnjyc.2020.000005>.
- [10] Ma, X.: Design and Implementation of Music Score Recognition Software, Nanjing University of Technology, 2020. <https://doi.org/10.27241/d.cnki.gnjgu.2020.001928>.
- [11] Meng, F.: Research on Detection and Deletion Algorithm of Printed Music Score Lines, Tianjin University, 2017.
- [12] Peng, Y.: Exploration of Vocal Music Teaching Mode in Local Colleges and Universities Under the Background of "Internet+", Northern Music, (24), 2020, 125-127.
- [13] Qiong, W.: Research on Optical Music Score Recognition Method Combining Multi-Scale Residual Convolutional Neural Network and Simple Recurrent unit, Tianjin University, 2019. <https://doi.org/10.27356/d.cnki.gtjdu.2019.003751>.
- [14] Qiong, W.; Medaka, L.; Xin, G.: Optical Music Score Recognition Based on Multiscale Residual Convolutional Neural Networks with Bidirectional Simple Recurrent Units, Advances in Lasers and Optoelectronics, 57(08), 2020, 67-76.
- [15] Qiong, W.; Medaka, L.; Xin, G.: Combining Multi-Scale Residual CNN and SRU for Optical Music Score Recognition, Advances in Lasers and Optoelectronics, 1-17. <http://kns.cnki.net/kcms/detail/31.1690.TN.20190916.1613.008.html>
- [16] Sun, J.: Reform of Vocal Music Teaching in Colleges and Universities Under the Background of Internet+, Art View, (04), 2021, 110-111.
- [17] Wu, L.: Opportunities and Challenges of Vocal Music Teaching Under the Background of "Internet+", Drama House, (29), 2020, 92-93.
- [18] Wu, T.L.: Research on Handwritten Music Score Image score Line Removal Algorithm Based on Machine Learning, Tianjin University, 2019. <https://doi.org/10.27356/d.cnki.gtjdu.2019.001061>.
- [19] Yang, J.: Research on the Fundamental Frequency Identification Method of Polyphonic Music Notes Based on HMM model, Science and Technology Bulletin, 35(11), 2019. [https://doi.org/10.1016/S0958-2118\(19\)30206-X](https://doi.org/10.1016/S0958-2118(19)30206-X)
- [20] Yin, Z.: Exploring the Development of Folk Vocal Music Teaching in the Internet Era, Chinese Literary Artists, (07), 2021, 155-156.
- [21] Zuo, D.: The integration of Vocal Music Teaching and Information Technology Under the Thinking of "Internet+" , Popular Color, (07), 2020, 146-147.