Research on automatic recognition algorithm of piano music based on convolution neural network

Ran Wu
Shandong University of Arts, Jinan, Shandong Province, 250012, China

Corresponding author and e-mail: Ran Wu, Z00294@sdca.edu.cn

Abstract. Convolutional neural network is a typical deep neural network, which combines the advantages of deep learning and image processing, improving the accuracy of feature recognition and greatly reducing the calculation of neural network. Window function is a signal with finite width in time domain. In signal processing, the use of window function can reduce the leakage of spectrum. Combining the window function and convolution neural network, the effective input in the model can be increased, and the accuracy and speed can be improved. In this paper, the music score of piano music is processed by window function as the input image of convolutional neural network. Through the analysis of the results, it is found that the algorithm has a 25.8% increase in rate compared with the traditional recognition method in rate, and the average value of F value reaches 89.24% in the accuracy test. The test results show that the piano music recognition rate of the algorithm has been greatly improved, and good results have been achieved in accuracy.

1. Introduction
With the improvement of the quality of life, more and more people are willing to learn some skills in their spare time. According to the Internet media survey, one of the skills that adults aged 25 to 60 want to learn most is piano, so piano music plays an important role in modern society. Nowadays, piano music recognition method is the focus of people's research, the first recognition method is carried out by professional music staff using manual recognition, which will not only consume time and energy, but also consume material resources. Therefore, automatic piano music recognition technology is the focus of piano music research nowadays, which not only liberates the hands of music staff, but also speeds up the recognition speed. In studying the development of piano music creation in China, Wittier pointed out that the first fully mature Chinese piano music was "shepherd boy Piccolo" written by He Luting in 1934. In the history of Chinese piano music creation, this is a work of great significance. Then, in the 1970s, Moore RL of Stanford University firstly tried to recognize multi tone music automatically [1-2]. In 2012, Bonitos et al. proposed the NMF based piano music automatic score recording system for automatic Piano Teaching [3]. In 2016, Cheng Tian et al. Proposed the attack / decay model, which modeled the piano notes as the attack stage of striking the strings and the decay stage of harmonic structure. They believed that the energy of each overtone in the decay stage was exponentially decaying, and used sparse NMF to estimate the parameters [4]. In 2018, the F value of the piano music automatic score recording algorithm proposed by Yang Nan in the first 30 seconds of map sensed is 86.80% (the allowable deviation range of note starting point is ± 100ms) [5].
However, there are few research articles on the topic of piano music automatic recognition, and the accuracy rate is not very high. At present, the mainstream piano music recognition still adopts manual recognition, and some automatic recognition algorithms using deep learning method also have the problems of low input effective value and slow recognition rate. In this paper, on the basis of predecessors, it tries to establish a convolution neural network model based on deep learning to realize automatic recognition of piano music, combined with window function to improve the speed and accuracy in the process of automatic recognition of piano music.

2. Algorithm and principle
In recent years, deep learning (as shown in Figure 1) method has achieved great success in many fields, among which convolutional neural network has achieved good results in the field of speech recognition and image recognition. At present, the best algorithm of note starting point detection is CNN model [6]. This paper studies and realizes the task of piano music automatic recognition based on convolution neural network model. The detected note starting point in the time-frequency representation of music sound signal is similar to the edge detection of convolution neural network applied in the field of computer vision. The recognition feature of music is that the spectrum content changes along the time axis, and the note starting point can be seen clearly in the time-frequency representation the boundary at the point. The edge detection in the image only needs local information, and no matter where the edge is in the image, it can be recognized with the same set of weights, so it is usually based on CNN model, so this paper uses convolutional neural network model to study and analyze piano music.

![Figure 1. Deep network model.](image)

CNN consists of an input layer, several hidden layers and an output layer. The hidden layer consists of a convolution layer, a pooling layer and a fully connected layer. The function of the hidden layer is to calculate the input data and take the output result as the input of the next layer. Each node of the full connection layer is connected to each node of the upper layer with a certain weight:

\[
    h_{l+1} = f \left( W_l h_l + b_l \right)
\]  

Where, \( h_l \), \( W_l \) and \( b_l \) represent the output, weight matrix and bias of layer \( L \) respectively, the input layer is represented as \( h0 = x \), \( x \) is the input of the neural network, and \( f \) is the activation function that operates on the input data element by element. The convolution layer in CNN has three characteristics: sparse connection, translation equivariant and weight sharing. Generally, there is a
pooling layer after the convolution layer, in which the pooling function changes the input elements into the overall statistics of adjacent elements. According to this characteristic, when the input is shifted, most of the output after the pooling function will not change. In the classification problem, the output of neural network usually represents the posterior probability distribution $P(Y | x, \theta)$, that is, when the parameter of neural network model is $\{W_l, b_l\}_{l=0}^{L-1}$, 0 < $l$ < $L$, and the input data is $x$, the probability of label $y$ is predicted. In network training, by using back propagation algorithm and optimization algorithm, such as random gradient descent, to find a group of parameters $\theta$, the probability distribution of network output is close to the real distribution of labels, so that in practical application, the network can accurately predict the corresponding labels for the unknown input data.

3. Experimental design

Because convolutional neural network is often used in the field of image recognition and has achieved good results, this paper takes the spectrum of different piano music as the input image of CNN, and realizes the recognition of piano music indirectly through image recognition.

As shown in Figure 2, the main process can be roughly divided into three steps: first, the piano music file is divided into test samples and training samples, and then the audio data is segmented to generate a spectrum feature map, then the spectrum of the training samples is put into the convolution neural network model for training, and finally the trained convolution neural network model is used to identify the tested piano music data.

The first step is to collect the signal of piano music. The signal of piano music changes with time. To collect it, we need to pay attention to the frequency of collection. Different music has different requirements for frequency [7]. Traditional automatic piano music recognition methods mainly use sensors to collect piano music, this method can only provide a unified collection frequency, lack of pertinence, the change of music will not make it change. In order to better collect the piano music signal, this paper uses the echo device to collect the piano music signal:

$$p(t, w) = s(t, w) \times f(t, w)$$  \hspace{1cm} (2)
Among them, P (T, w) represents the acquisition frequency of music signal; t represents the acquisition time; W represents the echo window function; s represents the length of echo window function; F represents the fundamental frequency of music signal.

According to the frequency determined above, the music signal is collected:

$$X = P \sum_{i=1}^{n} x_i \ast m$$

(3)

Among them, X represents the collection of music signals; X represents the its music signal in the collection; n represents the total number of music signals; m represents the collection parameters of music signals. In the first step, the acquisition of piano music signal provides data support for the following preprocessing of piano music signal. The main purpose of this step is to prepare for the realization of automatic piano music recognition.

In the second step, we take the piano music signal set obtained in the first step as the basis, and preprocess the set, mainly for music signal fundamental frequency feature extraction. Because the piano music signal collection process will be affected by the interference of the environment voice, resulting in impure signal, it cannot directly extract the fundamental frequency features of the collected piano music signal, otherwise there is a big deviation from the real situation. Therefore, in order to ensure the effectiveness of the piano music signal, this paper preprocesses the collected piano music signal. In the preprocessing of piano music signal, there are generally three kinds of window functions used by predecessors, which are Hamming window, rectangular window, triangular window, etc. different window functions have different effects on the short-term characteristics of piano music signal analysis frame. According to the research purpose, this paper finally selects Hamming window to preprocess the waveform image of piano music signal. The processing formula of Hamming window function music signal waveform image is (where n represents the total number of windows added):

$$F(x) = 0.54X - 0.46 \cos \frac{n}{N-1}$$

(4)

The third step is the output, because the input piano music signal is multi frame time-frequency representation, the output of the note start point detection module represents whether the input intermediate frame contains the start point of the music signal. The network output layer detected by the starting point of the music signal only contains one output unit. We select the sigmoid activation function, and the output value represents the probability that the intermediate frame represented by the input time-frequency contains the starting point of the music signal.

4. Results and discussion

The results of piano music signal waveform image processing obtained in the experimental design part are stored in MS SQL Server software. (MS SQL Server software has the characteristics of powerful function and reliable performance, so it has great advantages to select MS SQL Server software to store the score of piano music.) In order to get the experimental results, we select 9 piano pieces, and use our convolution neural network based piano music recognition method and traditional recognition method for comparative experiments. The results are as follows: 3
Figure 3. Recognition rate of convolution neural network method and traditional method in different piano music.

As shown in the figure, it can be clearly seen that the recognition rate of convolution neural network method is significantly higher than that of traditional methods. In the convolution neural network method, the maximum recognition rate is 93%, and the average recognition rate is 75.6%, while in the control group, the maximum recognition rate is 60%, and the average recognition rate is 49.8%. The average recognition rate of convolution neural network method is 25.8% higher than that of traditional methods, it shows that the automatic recognition of piano music based on convolution neural network method is highly effective.

Next, we evaluate the model, we use the common evaluation indicators in the field of music recognition and detection: F-measure (F1 score), recall (recall, also known as recall) and accuracy (precision, also known as precision). The accuracy is the correct proportion of the detected piano music signal starting point, the recall is the proportion of the detected piano music signal starting point in the standard result (ground truth), and the F value is the harmonic average of the accuracy and recall. The calculation formulas are as follows:

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (5)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (6)
\]

\[
F\text{-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (7)
\]

Among them, TP is true positive, that is, the number of piano music signal start points correctly detected in the standard results; FP is false positive, that is, the number of more detected; FN is false negative, that is, the number of missed detections.

This paper uses Python library Mir_ Eval [8] calculated the value of each evaluation index, Mir_. The eval library is specially used to calculate the general evaluation indexes of information retrieval and signal processing tasks of all kinds of music or audio.

The evaluation index used in this paper is the same as MIREX (music information retrieval evaluation exchange), the top international competition in the field of music information retrieval. MIREX competition provides a unified and regular evaluation platform for various tasks in the field of music information retrieval. Since 2007, the automatic piano music recognition subtask has been
Figure 4 shows the highest annual evaluation result of the piano music automatic recognition algorithm in the MIREX competition (MIREX data set, only considering the correctness of the starting point of the note, not considering the ending time of the note), in which the highest F value is 80.67%. In recent years, the accuracy of AMT algorithm seems to have reached a stable level.

![The F Value](image)

**Figure 4.** The highest annual evaluation results of MIREX piano music automatic score recording task.

The evaluation results of the piano music automatic recognition algorithm implemented in this paper on 9 different piano audios are shown in Table 1. It can be seen from the table that the F value of the algorithm in this paper is more than 90% on six test audio, but less than 80% on only two test audios. Among them, comparing the piano music with the highest F value of 95.48% and the piano music with the lowest F value of 73.61%, it is found that the music with relatively slow rhythm and relatively small number of notes pressed has a higher F value.

| Piano music | F value | Recall | Accuracy |
|-------------|---------|--------|----------|
| 1           | 0.9548  | 0.9075 | 0.9345   |
| 2           | 0.9236  | 0.9295 | 0.9039   |
| 3           | 0.9208  | 0.9295 | 0.9487   |
| 4           | 0.8903  | 0.8963 | 0.8834   |
| 5           | 0.9103  | 0.9456 | 0.9631   |
| 6           | 0.7963  | 0.7966 | 0.8023   |
| 7           | 0.7865  | 0.7440 | 0.7635   |
| 8           | 0.9325  | 0.9295 | 0.9457   |
| 9           | 0.9168  | 0.9400 | 0.9326   |

By averaging the evaluation results of all the selected piano music in the table, the F value, recall rate and accuracy rate of the algorithm based on convolutional neural network are 89.24%, 89.09% and 89.75% respectively. That is to say, the algorithm can detect most of the notes correctly.

### 5. Conclusions

The method of piano music automatic recognition based on convolutional neural network has greatly improved the rate of automatic recognition, the recognition rate of the algorithm is superior to the traditional method in every piano music. From the experimental results, the average recognition rate of
A convolutional neural network is 25.8% higher than that of the traditional method. In the process of our model evaluation, we find that the F value, recall rate and accuracy rate of the automatic recognition algorithm for piano music based on convolution divine network are found in the whole test set the results are 89.24%, 89.09% and 89.75%, which is a good result. The accuracy of the model is apparent. However, the piano music used in this paper is composed by software. The performance of the model in real piano music may be worse than that obtained by us. Because the piano music synthesized by software cannot fully express the sound details of the real piano. Finally, we find that the rhythm is relatively slow and the number of notes pressed is relatively small, and the F value of the tracks is higher. Therefore, the recognition method of piano music based on convolutional neural network needs to be improved. Because the setting of external environment parameters in the experiment ignores the interference of some factors in real life, which leads to the reality. Though the results of the test will have some deviation, it does not affect the general trend. Therefore, it is necessary to further study and analyze the automatic recognition method of piano music based on convolutional neural network.

References

[1] Wei tinge. The development of piano music creation in China [J]. Music research, 1983 (02): 41-52
[2] Moorers J A. On the transcription of musical sound by computer [J]. Computer Music Journal, 1977: 32-38.
[3] Bonitos E, Kalpeni A, Dixon S. Score-informed transcription for automatic piano tutoring[C]. IEEE, 2012: 2153-2157.
[4] Cheng T, Much M, Bonitos E, et Alan attack/decay model for piano transcription[C]. ISMIR, 2016.
[5] Yang Nan. Research and implementation of automatic score recording algorithm for piano music based on CNN[D]. Huazhong University of science and technology, 2018
[6] Liu Qi, Zhao Rong, Pu Yanni. Research on the practice of western music score Cataloging -- Taking the library of Central Conservatory of music as an example [J]. Library theory and practice, 2017 (08): 34-38
[7] Zheng Longyin. An analysis of the ancient scores of Kalu songs in the Tibetan court -- a comparative study based on two versions of Kalu scores more than 300 years apart [J]. Journal of Central Conservatory of music, 2017 (04): 69-79
[8] Raffiel C, McFee B, Humphrey E J, et al. mir_eval: A transparent implementation of common Miller metrics[C]/In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR. 2014.