2018 11th International Symposium on Computational Intelligence and Design

Optical Music Notes Recognition for Printed Music Score

Chuanzhen Li, Jiaqi Zhao
Information and Communication Engineering School
Communication University of China
Beijing, China
email: lichuanzhen@cuc.edu.cn
meetjiaqi@163.com

Juanjuan Cai, Hui Wang*, Huaichang Du
Key Laboratory of Media Audio & Video, Ministry of Education
Communication University of China
Beijing, China
email: {caijuanjuan, hwang}@cuc.edu.cn,
cucdu@163.com

***Abstract*—To convert printed music score into a machine-readable format, a system that can automatically decode the symbolic image and play the music is proposed. The system takes a music score image as input, segments music symbols after preprocessing the image, then recognizes their pitch and duration. Finally, MIDI files are generated. The experiments on Rebelo Database shows that the proposed method obtains superior recognition accuracy against other methods.**

Keywords- music score; segmentation; recognition; music symbols; MIDI

I. INTRODUCTION

The transcription of sheet music into machine readable format can be carried out manually. However, the complexity of music notation inevitably leads to burdensome software for music score editing, which makes the whole process very prone to errors. Consequently, Optical Music Recognition (OMR) system, which is dedicated to solving the automatic transcription of music score image, has been proposed and has become a hot research area. Typically, OMR system takes music score image as input and automatically converts its content into some symbolic structure, such as MIDI [1]. Generally, an OMR system is composed of three sub-modules: image preprocessing, symbol segmentation, and symbol recognition.

Most of the recent researches focuses on staff line detection, staff line removal, music symbol segmentation, and note recognition.

Removing the staff lines is one of the most important job. It will make the symbols segmentation, recognition and representation process much smoother. Hough Transform [2] and Stable path method [3][4] were applied to detect staff lines. However, how to remove staff lines without destroying the music notes has become another difficult problem. Chen and Xia [5] used 'open' morphological operation to process the music score image. The staff lines were removed after detecting rectangular note boundaries. But the operation corrupts symbols resting on staff lines more or less, so Mehta [6] and Vo [7] preferred to carry out segmentation process without removing stave lines.

The segmentation of music symbols is the operation following the staff lines removal. Segmentation is most important part of the entire process, as music notes can be recognized if and only if it has been segmented correctly. The most usual approach for symbol segmentation is hierarchical decomposition [8]. A music sheet is firstly analyzed and split by staff lines and then the elementary graphic symbols are extracted. Template matching [9] and projection method [6][10] were applied to segment the symbols as well. In addition, various musical symbols have been selected according to different goals. There are kinds of basic musical symbols with respective graphic representations: note heads, rests, dots, stems, flags, etc. Wen [11] segmented all the basic symbols. Mehta [6] and Blanes [12] preferred to keep the note head, stem and flag connected, which can reduce the type of recognition.

The segmented symbols are to be identified as specific duration and pitches. Traditional SVM [13] and nearest neighbor method [14] were used for symbol recognition. Minimum Spanning Tree Algorithm [15] is also a popular method for note identification. Recently, with the blooming of neural networks, Recurrent Neural Networks [1] and Convolutional Neural Network [12][14] were used to recognize music notes. Rebelo et al. [16] carried out an investigation on four common classification methods: support vector machines (SVMs), neural networks (NNs), nearest neighbor (kNN) and Hidden Markov Models. It is worth mentioning that the above machine learning methods are usually adopted to recognize note duration, as the pitch of note is usually determined by the distance between the note heads and the staff line.

The proposed system is aimed to convert printed music score into MIDI files. Morphological operations [5] are firstly used to remove the staff lines and projection method [6] is used to segment the symbols. Then the music notes are recognized by neural networks [14]. The rest of the paper is organized as follows. Section II describes composition of an OMR system. In section III, the experiments and results are presented. Finally, the conclusion and future work are described in Section IV.

II. PROPOSED WORK

The proposed system is carried out in five steps: image preprocessing, symbol segmentation, pitch recognition, duration recognition and MIDI transformation. Specifically, the input image is gray-scaled and binarized in image preprocessing. Then, the image is segmented after the staff lines are removed. The pitch and duration of the symbol are

2473-3547/18/$31.00 ©2018 IEEE
DOI 10.1109/ISCID.2018.00071

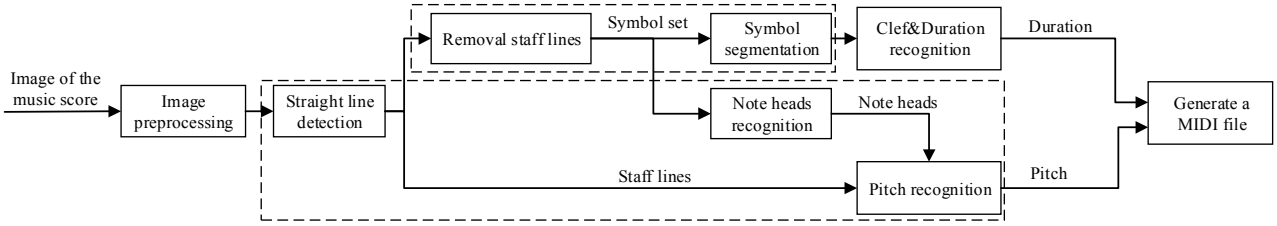


Figure 1. The proposed OMR system

recognized in the third and fourth steps, respectively. The last step is MIDI transcription. The illustration of the system is shown in Fig. 1.

A. Symbol Segmentation

The staff lines are firstly removed before the symbol is segmented. Morphological operations are used to accomplish this. The image was enhanced by using rectangle structuring elements, and only the musical symbols were retained. Once staff lines are removed, the music score sheet is left with only symbols of interest, as shown in Fig. 2. So, it will make the symbols segmentation, recognition and representation process much smoother.

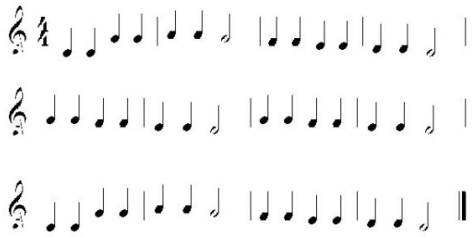
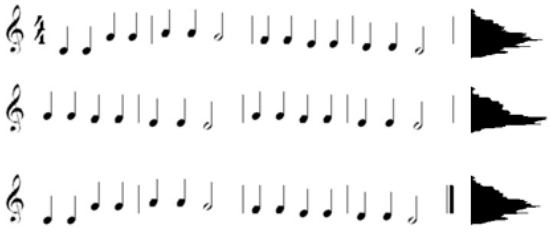
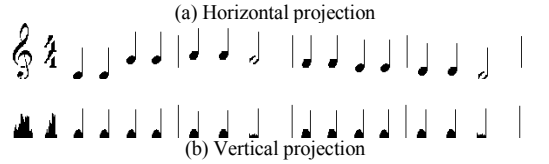


Figure 2. The image of removing the music line

We use the projection method to segment the music score image [6]. First, the image is horizontally projected to obtain a cumulative histogram of black pixels, which is then segmented at the trough position. Vertical projection of the resulting image also yields a cumulative histogram of black pixels and is segmented at the trough position. Eventually, single musical symbol is obtained. Fig. 3 shows the results of horizontal projection and vertical projection respectively. Algorithm 1 summarizes the overall algorithm as shown in Fig. 4.



(a) Horizontal projection



(b) Vertical projection

Figure 3. Projected image

Algorithm 1: Segmentation of symbols by projection

Input: Printed music score image after preprocessing
Output: Music symbols
1: Create a empty lists 1 to represent the number of black pixels in the image
2: **for** in row i do
3: Calculate the black pixel values of each row into the list, and compare the pixel maximum: gMax
4: **if** The element m in the list < 0.95× gMax
5: Segmented image horizontally from here
6: **end if**
7: **end for**
8: Create a empty lists 2 to represent the number of black pixels in the horizontal image
9: **for** in column j do
10: Calculate the black pixel values of each column into the list, and compare the pixel maximum: mMax
11: **if** The element n in the list < 0.95× mMax
12: Segmented image vertically from here
13: **end if**
14: **end for**

Figure 4. Algorithm of symbol segmentation

B. Pitch Recognition

The pitch is calculated based on the distance between the staff lines and the note heads. First of all, we detect the position of the staff line by using the Hough transform. Hollow notes are filled with solids by using morphological closing operations, and then all the note heads are found using the template matching method. Algorithm 2 summarizes the overall algorithm as shown in Fig. 5. Fig. 6(a) shows the filled notes. The position of the staff line and the note heads are shown in Fig. 6(b).

Algorithm 2: Pitch recognition

Input: Printed music score image
Output: Music notes pitch
1: Use the Hough transform to detect the position of the line and calculate the staff space
2: Use the template matching algorithm to detect the position of the note heads
3: Determine the pitch based on the distance between the position of staff line and the note heads
4: **end**

Figure 5. Algorithm of pitch recognition

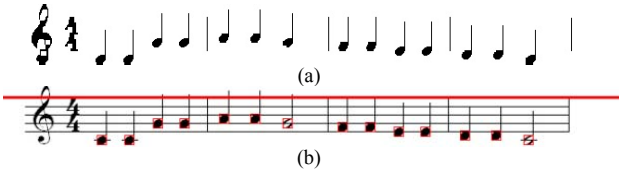


Figure 6. (a) Filled notes, (b) The position of the staff line and the note heads

C. Duration Recognition

Neural networks have successfully solved many practical problems that are difficult to solve in the fields of pattern recognition, intelligent robots, predictive estimation and so on. In this paper, two common neural network methods are used for symbol recognition.

CNNs are very popular models for machine vision applications. CNNs may consist of multiple convolutional layers, optionally with pooling layers in between, followed by fully connected perceptron layers. Typical CNNs learn through the use of convolutional layers to extract features using shared weights in each layer. The feature pooling layer (i.e., subsampling) generalizes the network by reducing the resolution of the dimensionality of intermediate representations (i.e., feature maps) as well as the sensitivity of the output to shifts and distortions. The extracted features, at the very last convolutional layer, are fed to fully connected perceptron model for dimensionality reduction of features and classification [17].

Long Short Term Memory (LSTM) networks are a class of Recurrent Neural Networks (RNNs) that are widely used for machine learning tasks involving sequences, including machine translation, text generation, and speech recognition. An LSTM is composed of cells, each of which contains a cell state along with multiple gating units that control the addition and removal of information from the state [18]. Let x_t, c_t, and h_t denote the input, cell, and hidden states, respectively, at iteration t. Given the current input x_t, previous cell state c_{t-1}, and previous hidden state h_{t-1}, the new cell state c_t and the new hidden state h_t are computed as:

$$[f, i, o, j]^{\mathrm{T}} = [\sigma, \sigma, \sigma, \tanh]^{\mathrm{T}}\left(W x_t + U h_{t-1} + b\right)$$

$$c_t = f \odot c_{t-1} + i \odot j$$

$$h_t = o \odot \tanh(c_t)$$

Where $\odot$ denotes element-wise multiplication, and b is the bias. The activation function σ is the sigmoid function $\sigma(x) = 1/(1 + \exp(-x))$. The output of an LSTM layer at iteration t is h_t [19].

D. MIDI Transformation

MIDI format is a digital music format to record sound. In this step, we use the MIDIUtil library[1] to generate the MIDI file. MIDIUtilis a pure Python library that allows one to write multi-track Musical Instrument Digital Interface (MIDI) files from within Python programs. The previously identified notes duration and pitch sequence are taken as input and a MIDI file is output as the result. The file can be played by music player, or it can be viewed and edited by using professional software. The generated file is shown in Fig.7.

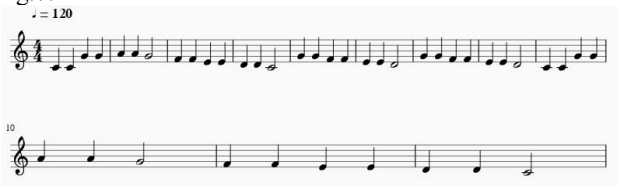


Figure 7. Generated MIDI file

III. Experimental Result

In this paper, a variant of the Rebelo Database has been adopted for training. There are many handwritten symbols and printed symbols in the database. Only printed music symbols are used for training in the paper. 14 classes are considered: notes and rests from whole to a sixteenth, clefs and time signatures. We put each note symbol onto a large enough block. The size of the blocks is set to 64×64, then they are converted as row vector into 1×4096. Finally, resized row vector is input into Neural networks for training. We use two common neural network methods to train the database.

The test data is derived from the segmented symbols in Section II(B). We segment some common music score images and obtain about 500 symbols. Using a 3-layer Convolutional Neural Network, after 12 iterations, the recognition accuracy of 98.47% has been achieved. Using a 3-layer LSTM, after 25 iterations, the recognition accuracy of 94.64% has been achieved. The results show the efficiency of the CNN for image classification. Fig. 8 shows the confusion matrix of CNN and the confusion matrix of LSTM is shown in Fig. 9.

According to the experimental results above, we can see that, the half note and the quarter note are often misjudged. And the eighth note and sixteenth note have the similar misjudgment problem. It may result from the high degree of appearance similarity between them.

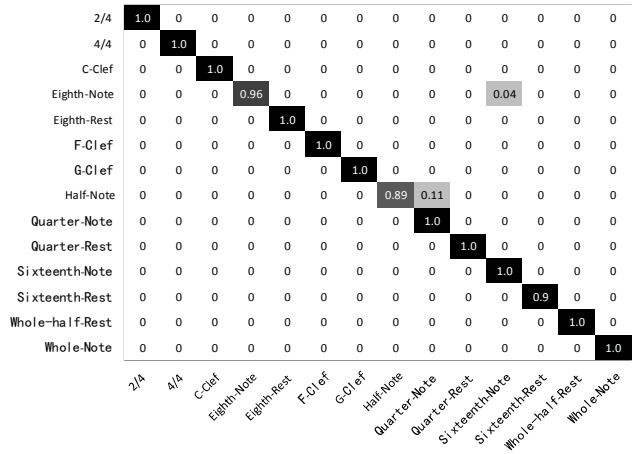


Figure 8. Confusion matrix of CNN

[1] https://pypi.org/project/MIDIUtil/

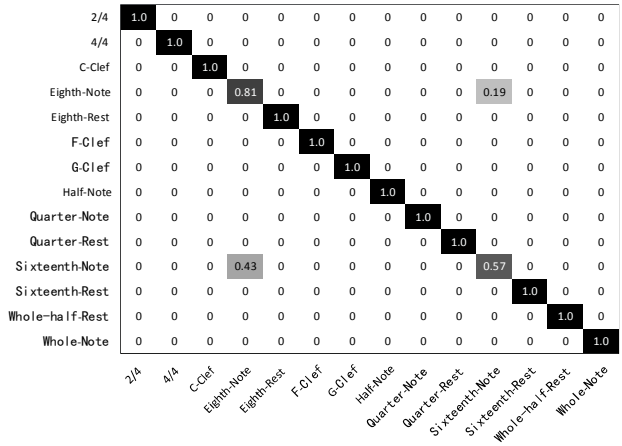


Figure 9. Confusion matrix of LSTM

In the paper, we only compare the results of notes recognition for printed music score. The comparative results against other methods are listed in Table 1. The results show that our method is superior to the general methods, but there is still room for improvement.

TABLE I. COMPARISON OF OUR METHOD TO OTHER REPORTED METHODS

Methods	Accuracy
Artificial Neural Network [6]	92.38%
Combined Neural Network [11]	98.82%
Minimum Spanning Tree [15]	97.9%
The proposed method of LSTM	94.64%
The proposed method of CNN	**98.47%**

IV. CONCLUSION AND FUTURE WORK

In this paper, an OMR system using CNN for printed music score is proposed. The music score image is first preprocessed and followed by staff lines removal with the morphological operation. The projection method is used to segment the note symbols and they are classified and recognized by Neural Networks to get note duration. The pitches of the notes are recognized according to the distance between the note heads and the staff lines. Finally, the duration and pitch sequence are converted to MIDI files by MIDIUtil. Experimental results show that our method has obtained superior results.

In the future, more notes will be increased to the database to test the accuracy of recognition. And the proposed method will be used for written scores recognition to validate the generality.

ACKNOWLEDGMENT

This research was supported by the NSFC grant (No.61501410 and No.61631016) and the Engineering Planning Project of Communication University of China (Grant No. 2017XNG1716 and 2018XNG1861).

* Hui Wang is the corresponding author.

REFERENCES

[1] A Baró, P Riba, J Calvo-Zaragoza, A Fornés, "Optical Music Recognition by Recurrent Neural Networks," Iapr International Conference on Document Analysis & Recognition , 2018 :25-26.

[2] H Miyao, Y Nakano, "Note symbol extraction for printed piano scores using neural networks," IEICE Transactions on Information and Systems, 1996: 548-554.

[3] SCJ Dos, A Capela, A Rebelo, C Guedes, dCJ Pinto, "Staff detection with stable paths," IEEE Transactions on Pattern Analysis & Machine Intelligence , 2009 , 31 (6) :1134.

[4] HN Bui, IS Na, GS Lee, HJ Yang, SH Kim, "Boosted Stable Path for Staff-line Detection Using Order Statistic Downscaling and Coarse-to-Fine Technique," International Conference on Pattern Recognition, 2014 :522-526

[5] G Chen, S Xia, "The study and prototype system of printed music recognition," International Conference on Neural Networks & Signal Processing, 2003, 2:1002-1008 Vol.2.

[6] AA Mehta, MS Bhatt, "Optical Music Notes Recognition for Printed Piano Music Score Sheet," International Conference on Computer Communication & Informatics, 2015 :1-6.

[7] QN Vo, T Nguyen, SH Kim, HJ Yang, GS Lee, "Distorted Music Score Recognition without Staffline Removal," International Conference on Pattern Recognition, 2014:2956-2960.

[8] A Rebelo, I Fujinaga, F Paszkiewicz, ARS Marcal, C Guedes, JS Cardoso, "Optical music recognition: state-of-the-art and open issues," International Journal of Multimedia Information Retrieval, 2012 , 1 (3) :173-190.

[9] S Dai, Y Tian, CW Lee, "Optical music recognition and playback on mobile devices", Stanford, 2012.

[10] DM Dhanalakshmy, HP Menon, V Vinaya, "Musical Notes to MIDI Conversion," International Conference on Advances in Computing , 2017 :799-804.

[11] C Wen, A Rebelo, J Zhang, J Cardoso, "Classification of Optical Music Symbols based on Combined Neural Network," International Conference on Mechatronics & Control , 2014 :419-423.

[12] AR Blanes, AF Bisquerra, "Camera-based Optical Music Recognition using a Convolutional Neural Network," Iapr International Conference on Document Analysis & Recognition , 2018.

[13] M Akbari, AT Targhi, MM Dehshibi, "TeMu-App: Music Characters Recognition Using HOG and SVM," Machine Vision & Image Processing, 2015.

[14] J Calvozaragoza, AJ Gallego, A Pertusa, "Recognition of Handwritten Music Symbols with Convolutional Neural Codes," IAPR International Conference on Document Analysis and Recognition, 2017:691-696.

[15] Y Sazaki, R Ayuni, S Kom, "Musical Note Recognition Using Minimum Spanning Tree Algorithm," International Conference on Telecommunication System Services and Applications , 2015 :1-5.

[16] A Rebelo, G Capela, JS Cardoso, "Optical recognition of music symbols: a comparative study," Springer-Verlag, 2010, 13 (1):19-31.

[17] H Salehinejad, S Sankar, J Barfett, E Colak, S Valaee, "Recent Advances in Recurrent Neural Networks," arXiv preprint arXiv, 2017:1801.01078.

[18] S Sen, A Raghunathan, "Approximate Computing for Long Short Term Memory (LSTM) Neural Networks," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2018, 1-1.

[19] W. Zaremba, I. Sutskever, and O. Vinyals. Recurrent neural network regularization. arXiv preprint arXiv:1409.2329, 2014. 2.