Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 2415857, 9 pages
https://doi.org/10.1155/2022/2415857

Research Article

Music Score Recognition and Composition Application Based on Deep Learning

Mingheng Liang

CITI University, Ulan Bator 999097, Mongolia

Correspondence should be addressed to Mingheng Liang; 41823038@xs.ustb.edu.cn

Received 7 April 2022; Revised 1 May 2022; Accepted 10 May 2022; Published 17 June 2022

Academic Editor: Vijay Kumar

Optical score recognition is a critical technology for retrieving music information, and note recognition is a critical component of score recognition. This article evaluates and discusses the current state of research on important technologies for score recognition. To address the issues of low note recognition accuracy and intricate steps in the present music score image, a deep learning-based music score recognition model is proposed. The model employs a deep network, accepts the entire score image as input, and outputs the note's time value and pitch directly. Experiments on music scores demonstrate that the method described in this study has a high note identification accuracy of 0.95 for time values and 0.97 for pitch, which is suitable for composition.

1. Introduction

OMR is a type of optical character recognition used in music to recognize scores in editable or playable formats such as MIDI (for playback) and MusicXML (for encoding and formatting) (for page layout). Pitch and duration are represented by notes, which, in comparison to other symbols in musical scores, contain significant semantic information. Notes comprise a sizable portion of the total symbol count in musical scores. As a result, note recognition lies at the heart of the musical score comprehension process [1, 2].

Most OMR technologies have been improved over the last few years using common frameworks, including image preprocessing, stave identification and deletion, symbol recognition and classification, as well as other uses. Some researchers have used the GMMRF model to achieve image binarization. Images of musical scores with complicated backgrounds can be effectively reduced to noise by using image binarization. Researchers have also developed a way to minimize the detection error rate to 1.4% by identifying spectral lines based on their stable paths. Multidimensional local binary patterns and XGBoost may be used to delete staves from handwritten musical scores, with only 0.05 percent of the training data needed to greatly improve model recognition performance [3–6].

Not only is the shape of notes always changing but it is also diverse and polymorphic, making it difficult to distinguish between them. We can divide the traditional note recognition system into two categories: segmentation grammar rules and image rules. When notes are divided into groups based on their shape, segmentation algorithms are used to perform note recognition within each group. Examples include the use of symbolic descriptors to organize and identify notes, as proposed by certain academics [7–9]. The second type of method defines a collection of grammar or rules that can be used to integrate different types of bases together. To identify notes, use metasymbols (such as note heads, beams, tails, and beams), connect primitive symbols according to a set of predetermined criteria, and then use tree diagrams to identify the notes. Finally, graph-based approaches make use of graphs to define the relationship between primitive symbols and to express the shape of notes in a variety of languages. To use the classic note recognition approach, the staff must be deleted first, followed by the extraction of primitive symbols, and finally the combination of primitive symbols to complete the note recognition

process. The entire procedure is extremely sophisticated, and each step has an impact on the accuracy of note recognition. The standard framework provides strong optimization of each individual phase, but the complexity is great, and the improvement in total accuracy is not immediately apparent [10].

Recent years have seen a sea change in MR processing, owing to the emergence of deep learning (DL) in the realm of computer vision. Each year, an increasing number of studies on DL for OMR problems are done. There are two primary groups of study methodologies that can be identified. They are object detection and sequence recognition. When symbols are discovered in a scoring image, the object detection method determines their location and categorizes them. When recognizing the score symbols, a region-based CNN is presented that divides the score image into several single-line staff images, which are then fed into the region-based CNN using an R-CNN detector. When it comes to detecting musical scores, numerous researchers have found success by combining semantic segmentation models with subsequent detectors. By collapsing the challenge of note recognition into a classification problem posed by a series of binary pixels and using a linked component detector, it is feasible to establish the category of the sign. Watershed detectors for musical notes have been proposed by some researchers. They propose training a CNN model to learn a custom energy function, which is then used to semantically partition the entire score and then identify the notes using the watershed transformation. In comparison, all approaches for detecting targets share some drawbacks [11–15]. For example, a model can only identify a symbol category but it cannot distinguish between a note's pitch and its duration. Note that the sequence recognition approach converts the score image into a sequence that is then fed into the RNN model, which predicts the output of the note recognition algorithm. In order to identify the pitch and duration of notes, some researchers divide the entire score image into several image segments, encode the segments into fixed-size sequences using CNNs and RNNs, and then decode the sequence using RNNs to determine the pitch and duration of notes. The drawback of this method is that the entire image cannot be simply submitted; rather, the image must be sliced into single-line staves and then input sequentially. Additionally, this method has the disadvantage of having extremely low recognition accuracy when dealing with multivoice musical scores [16–18].

DL has a number of advantages when it comes to OMR processing. For example, as compared to the traditional OMR technique, recognition accuracy is greatly improved, and recognition processes are significantly simplified. However, existing target detection algorithms are unable of distinguishing between note pitch and time value, whereas the sequence recognition method is capable of resolving challenges such as low recognition accuracy for multivoice musical scores. This paper proposes a DL-based musical note identification model for printed musical scores. This approach directly outputs the time value and pitch of the notes on the score. Additionally, the model is capable of detecting multivoice score visualizations and is fully integrated from start to finish.

2. Background

A music score is a static depiction of the temporal, auditory, and dynamic art form of music that can be recorded, preserved, and revisited. Historically and currently, the bulk of musical compositions has been preserved in the form of paper scores, which remain the principal medium for expressing, publishing, and disseminating musical works today. Computer music's progress has altered the way human musical activities are generated in a variety of ways. As a fundamental prerequisite for human musical activity, musical scores have been given a new carrier, namely digital musical scores in the computer music mode. As a result, the conversion of existing paper musical scores to digital musical scores is critical. At the moment, the digitalization of paper music errors is still dependent on manual reading, which is a time-consuming and inefficient process. This will invariably result in a conflict between the low-speed music information input and the high-speed music information processing, finally resulting in inefficient music information processing. The contradiction is resolved by utilizing the computer's automatic reading of music scores. That is, paper music scores are scanned into a computer, and all possible symbols and semantics in the image are identified automatically, resulting in the realization of a digital expression including the entirety of a piece of music [19, 20].

Many research findings have emerged as a result of decades of development in music score recognition technology, and commercial score recognition systems such as sharp eye, smart score, photo score, and Capella-Scan have also been brought to the market. Due to the unique structure and semantic properties of music scores, however, music score identification technology continues to face difficulties and obstacles in a variety of areas. The authors of a survey of current research on music score recognition and understanding noted that no system has yet demonstrated sufficient performance in terms of recognition accuracy, resilience, and the range of applications that can be used to execute the task [21–27].

This section examines the visual characteristics, representation characteristics, and associated characteristics of symbols in musical scores in accordance with the rules of notation and addition of musical scores and discusses and summarizes the three key technologies of musical score recognition, the current research status of staff line detection and deletion, note recognition, and global association analysis, as well as the problems that currently exist in these areas.

2.1. Representation Characteristics of Symbols in Musical Scores. In long-term music practice, human beings progressively develop a music information description language that comprises a set of notation standards that are derived from the reading habits of those who participate in the music. Because the notation technique was designed with manual reading in mind, it contains a number of characteristics that are incompatible with machine reading in terms of information expression. These characteristics include the following:

(1) Complexity: One of the most important characteristics of a music score is its complexity and changeability. A music score is a two-dimensional data collection made up of lines, blocks, symbols, sentences, and other morphological primitives.

(2) Intensive intersection and multiple adhesions: Various primitives are interlaced in the spectral lines, and the intersections and adhesions between them and the spectral lines are ubiquitous. There are also adhesions between the primitives that do not occur as a result of any external circumstances.

(3) Polymorphism: In addition, because the notation is very schematic, items having the same semantics (such as musical notes) might have radically different shapes or structures.

(4) Relevance: Relevance is the fourth point to consider. A sophisticated network of links exists between primitives, which manifests itself as spatial position associations at the syntactic level (e.g., the construction of notes) and global relationships at the semantic level (such as the semantic interpretation of notes).

(5) Implicit: The sentence implies that by combining and inferring low-level primitive information, it is necessary to extract a great deal of high-level semantic knowledge.

(6) Fuzziness: In some cases, the same rules cannot be applied to different scores, or even different parts, and effective judgments can only be made with the help of people's unique experience [28, 29].

Musical scores can be classified into two types based on the number of recorded parts they contain: single-part musical scores and multipart musical scores. Multipart music is an important component of today's musical culture, and it is frequently employed in a variety of musical activities such as piano performances, folk music performances, symphony orchestra performances, chorus performances, and other musical events. Multivoice musical compositions include complicated layouts, a wide variety of graphic components, and deep arrangements that make them difficult to recognize when performed by multiple voices. The amount of music information available is enormous, and the six qualities listed above are particularly apparent in multivoice musical scores, which provides significant benefits to computer automatic notation reading systems. There are a lot of difficulties.

2.2. Research Status of Key Technologies. The key technologies for musical score recognition are primarily divided into three categories: detection and elimination of spectral lines, identification of notes, and global correlation analysis.

Music scores are characterized by their complexity, which means that there is no universal way of recognizing primitives in musical scores. Following the approach of dividing and conquering according to local conditions, we must first separate distinct sorts of primitives in advance before picking the most effective method for identifying each type of primitive. The properties of dense intersection and multiadhesion make it impossible to separate musical score primitives, and spectral lines are the source of the majority of interference. As a result, the detection and deletion of spectral lines are regarded as the primary essential link in the recognition of musical scores by the majority of scholars.

Strictly speaking, the methods of spectral line identification that are commonly used can be split into two categories:

(1) Structural feature search methods are based on statistical transformation, such as horizontal projection, Hough transform, and wavelet transform

(2) Methods are based on run-length analysis, row-neighbor graph method, feature point DP matching method, and path search method, among others

The two sorts of approaches each have their own set of advantages and downsides, which are listed below. It is possible for the statistical transformation approach to fail when the spectral line has been deformed and does not conform to a rigid straight line shape; however, this is not likely to happen in most cases. Local details are easily influenced by noise interference, which might be difficult to detect [30]. When the level of interference reaches a specific threshold, the system will be forced to contend with the issue of insufficient local information collecting and no general direction.

The most important aspect of eliminating spectral lines is to ensure that the integrity of the primitives is not compromised during the deletion process. As a result, academics have developed a variety of ways for removing spectral lines, including vector line analysis, run-length analysis, the row-neighbor graph method, the skeletonization method, and other approaches. Because of the complicated intersection and adhesion of spectral lines and primitives, it is difficult to distinguish between them. Additionally, it is required to maintain the integrity of the primitives once the hidden lines have been removed, due to the interference of many phenomena that may be encountered in the real scanned image (such as tilt deformation, bending deformation, and fracture). It is quite challenging, and there is no spectral line deletion method that clearly outperforms the others in terms of effectiveness [31].

In order to extract and analyze musical score information automatically, note recognition is the foundation and fundamental to the process. Note is a polymorphic pattern object that has a lot of different variations [32]. The separation and extraction of its components (including the note head, stem, tail, and beam, commonly referred to as note primitives) are required before using the association between primitives to identify the note. Relationships and associated notation rules restructure the data in order to produce higher-level graphical representations.

Because of the wide variety of note primitives, there is no fixed or universal solution for extracting primitives from their corresponding notes. Methods for basic extraction now in use include the following: the projection method, skeletonization, run-length analysis, and template matching. Mathematical morphology, NN, Kalman filter, and other techniques are used. According to the actual effect of the existing methods, it is difficult to ensure the accuracy of the extraction results simply by relying on the underlying image processing methods and the primary source of the difficulty stems from the dense intersection and adhesion between primitives in the extracted image. The rational application of high-level information restrictions in structure and semantics to guide the extraction of primitives is a novel concept that deserves further investigation.

The algorithm description approach is the most commonly used method of primitive restructuring; that is, the association information and notation between primitives are hidden within the program's algorithm. The application of rules can be controlled in a flexible and efficient manner by this type of system, and the algorithm execution process is completely transparent. The drawback is that it is tough to keep up with and grow with over time. Researchers have developed a basic rearrangement approach based on structural grammar description in order to stress the scalability of the system. This method is intended to emphasize scalability. A preset language is used to specify the raw association information and recombination rules, and grammar parsing is used to complete the reconstruction from primitives to notes. This separation of data and control results in a system that is scalable and achieves good scalability. However, notation is not a grammatical system that is rigid and accurate in its application. Numerous regulations are implied or ambiguous in nature. As a result, the structural grammar method has difficulty in providing a comprehensive description of the principles when working with complex multivoice musical works. If one looks at the existing grammatical description methods, it appears that they are geared around the restricted material of Shanjibu Lepu Huotianjiewei, which similarly has pretty simple content.

For each item in a musical score, there exists some sort of logical or semantic global relationship that connects the individual musical information contained inside it to everything else. The ultimate goal of musical score recognition is to get and understand the entire high-level semantics of musical scores by analyzing the global links between various musical pieces, which is the primary focus of the research. These notes contain a wealth of information, including the interpretation of functional semantics (pitch and duration), the limitation of their temporal characteristics, the overall consistency of the bar duration, and their spatial arrangement and distribution characteristics (Note 1) (for example, the notes are arranged in a square).

The most common application of global association is to establish the functional meaning of a note as the final link of identification, which is the primary focus of current research. An essential technique for aiding in the identification and verification of musical scores has been discovered by some researchers in recent years: high-level information can be used to create a feedback function for the identification of a single object or of an entire music score! Researchers have developed a fuzzy set theory-based adaptive system for identifying music scores that can automatically check and correct recognition results by limiting the duration value of the time signature. The system's concepts, theories, and methods, on the other hand, can only be applied to single-part sheet music.

Multivoice scores and single-voice scores can both benefit from the recognition and feedback function of semantic information. Though not specifically stated in the score, it is assumed that all of the notes can be correctly identified by their respective voices. In the score, each voice is not expressly labeled as such. In order to separate the horizontal and vertical distribution of the full set of notes, one must analyze and reason. It is a difficult problem to tackle, yet high-level semantic information in multivoice musical scores necessitates solving the challenge of autonomously partitioning voices. As of right now, there are only a handful of related studies, and no comprehensive approach to the problem exists.

3. Method

This section introduces the DL model and its application to end-to-end score recognition. The model takes the preprocessed image of the score as input and outputs the time and pitch of each note. The recognition process is schematically represented in Figure 1.

The note recognition model goes through the following process: after performing a series of convolution, residual, and splicing operations on the score image, extract the feature map, classify, and output the note duration and pitch on the feature map, and then return to the note's bounding box to obtain the coordinates of the candidate frame, as illustrated in the following diagram.

The CNN + LSTM structure in this model is shown in Figure 2.

In this research, a single-stage target detection model is used, and the prior conditions of the candidate boxes are provided in a straightforward manner. Rather than manually selecting a prior candidate box, the experiment will employ k-means clustering on the area of the bounding box in the training set to automatically determine the width and height of a good previous candidate box. In the author's study, a total of seven different width and height measurements are chosen as the prior condition input. In this paper, the *k*-means algorithm is applied to seven distinct widths and heights, and the width and height that is closest to the centroid are used as an a priori candidate box input to the neural network in the following step. Because of the use of *k*-means to produce candidate boxes, the performance of the neural network model will be improved, and the model will be simpler to learn.

The loss functions of note pitch $loss^c$ and note time worth $loss^d$ are as follows:

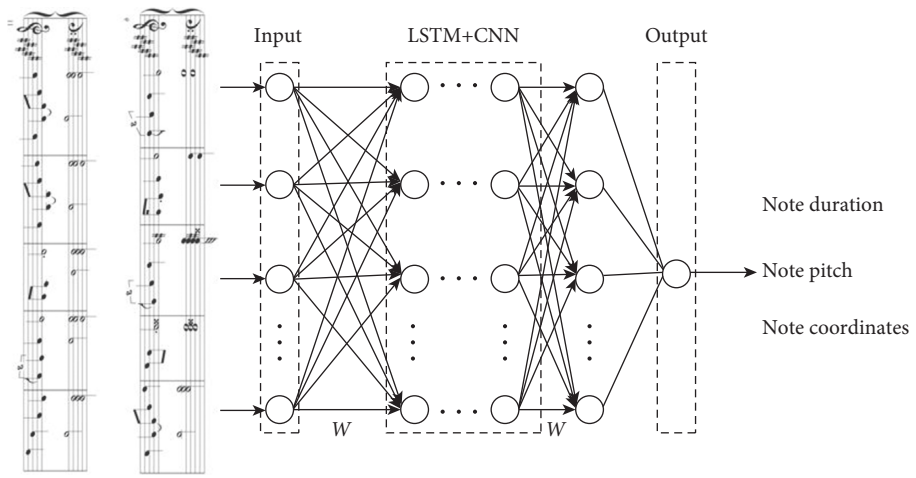

Figure 1: Structure of our method.

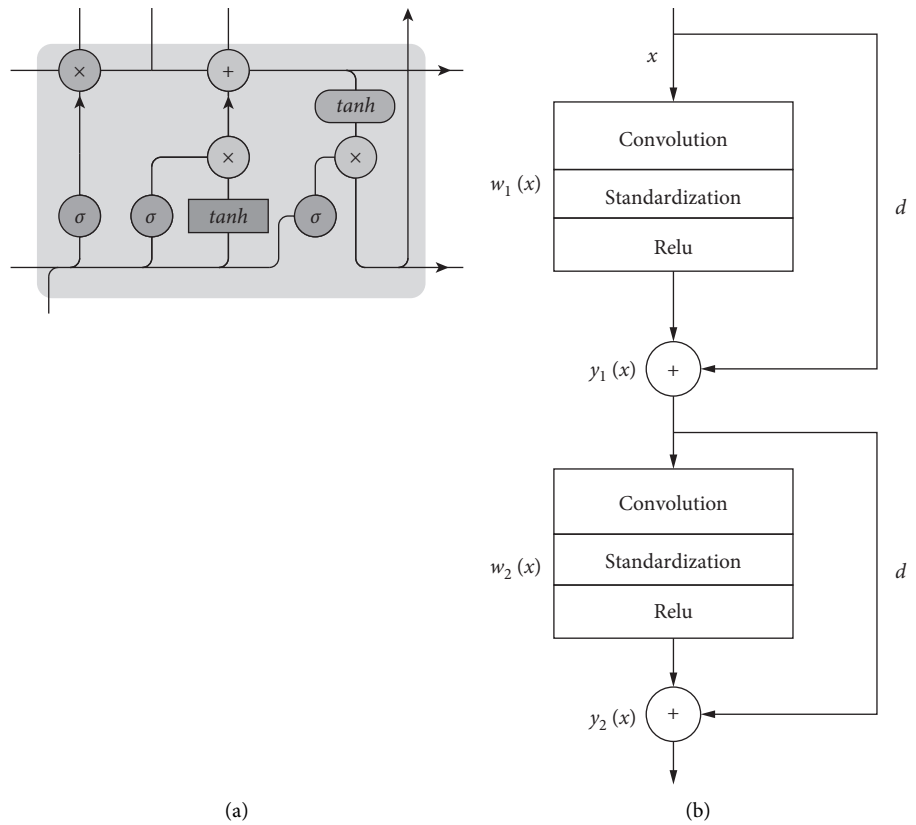

Figure 2: Structure of CNN + LSTM, (a) framework of LSTM, (b) framework of CNN.

$$\begin{aligned} \text{loss}^c &= y_n \ln \sigma(x_n) + (1 - y_n)\ln(1 - \sigma(x_n)), \\ \text{loss}^d &= y_k \ln \sigma(x_k) + (1 - y_k)\ln(1 - \sigma(x_k)). \end{aligned} \tag{1}$$

The feature vector returns 4 offsets t_x, t_y, t_w, t_h for each candidate frame, if the offset of the cell pixel from the upper left corner of the image is (C_x, C_y), and the candidate frame has prior information width and height (p_w, p_h), and the prediction result is

$$\begin{aligned} b_x &= \sigma(t_x) + c_x, \\ b_y &= \sigma(t_y) + c_y, \\ b_w &= p_w e^{t_w}, \\ b_h &= p_h e^{t_h}, \end{aligned} \tag{2}$$

where b_i is the coordinate of the candidate frame output by the model, and t_i is the offset predicted by the model.

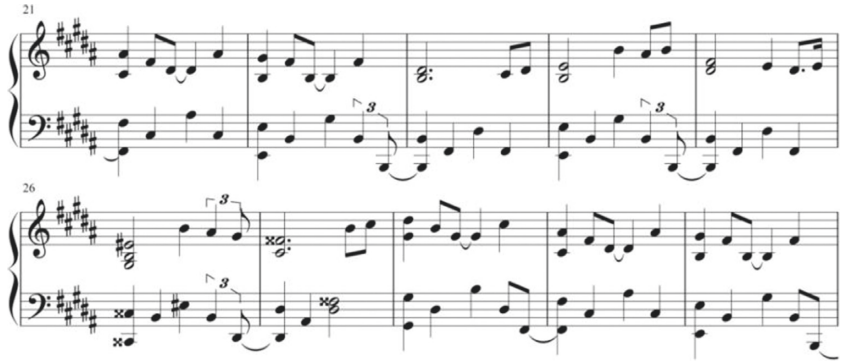

Figure 3: Scores generated by MusicXML.

Each pixel on the feature map predicts 7 feature vectors, and each feature vector uses the $\sigma(\cdot)$ activation function to regress the offset of the bounding. In the experiment, the MSE is used to calculate the loss function.

$$\text{loss}^{h} = \sqrt{\hat{t}_i - t_i}. \tag{3}$$

The loss function of the regression offset is loss^{b}, the loss function of note treble classification is loss^{c}, the loss function for note duration classification is loss^{d}, and the loss function of the bounding box is loss^{e}; then, we have

$$\text{loss}^{\text{sum}} = \text{loss}^{b} + \text{loss}^{c} + \text{loss}^{d} + \text{loss}^{e}. \tag{4}$$

A logistic regression model is used to predict the confidence of each bounding box in the data set. Initialize the predicted bounding box with the ground-truth bounding box at 0.6, and if this predicted bounding box has more overlap with the real regression box than any other predicted bounding box, select it as the best match and highest confidence level, and otherwise, select any other predicted bounding box and lowest confidence level. The degree is a one-digit number. Alternatively, if the overlapping section exceeds the threshold but is not the best bounding box, the projected bounding box is discarded, resulting in a bounding box loss^{sum} of zero (see below). According to the gradient descent method, the four sublosses will converge to the final value. If the overlap is less than the threshold, the confidence in the bounding box is equal to one hundred percent. Finally, the bisection crossover is used to construct the loss function of the confidence interval.

$$p^{\text{conf}} = \begin{cases} 1, & \text{Optimal matching}, \\ 0, & \text{Overlap} > \text{Threshold}, \\ \text{Miss}, & \text{Overlap} < \text{Threshold}, \end{cases} \tag{5}$$

where p^{conf} is the confidence of the candidate box.

4. Results

Specifically, the MusicXML files in the Muse Score collection were used to create the dataset that was used in this paper [9, 12]. A variety of musical scores created by users are included in the collection, each with a unique compositional style and structure. In order to train and evaluate the system, about 10,000 MusicXML files were picked, with these files being separated into three different groups. Sixty percent of the data were used for training, twenty percent for validation, and twenty percent for evaluation of the model.

A dataset of sheet music photos and related note annotations is built from a corpus of selected MusicXML files. The dataset is based on the selection of the MusicXML files. Using Muse Score, you can create a score image from the MusicXML file you downloaded. The image of the generated score is shown in Figure 3. The label for each note in the score is represented by a vector consisting of the note's pitch, duration, and bounding box position information. Pitch and length are used to represent each individual note. In this work, the pitch is recoded as the vertical distance, which is the distance between the note and the vertical axis of the staff (i.e., the distance between the note and the vertical axis of the staff). In Figure 4, the vertical distance between a note and the staff is shown to be responsible for determining the pitch value of a note.

During training, data augmentation is performed, and each time the network model is used, it is supplied with a unique set of training samples. The model is trained to find the optimal solution using a stochastic gradient descent optimizer with a batch size of 32 and an initial learning rate of 0.001, a gradual drop in the learning rate, and a halving of the learning rate every 10 epochs. The model approaches convergence after around 40 cycles.

The evaluation indicators used in this paper are the time value accuracy rate TA, the pitch accuracy rate PA, and the average note accuracy ANA. The calculation formula is as follows:

$$\begin{aligned} TA &= \frac{FS}{FS + NS}, \\ PA &= \frac{FS}{FS + NS}, \\ ANA &= \frac{FS}{FS + NS}. \end{aligned} \tag{6}$$

A total of 2000 Muse Score-transformed music score photographs were used in the model's testing and validation phases, with 1200 images used in the testing phase, 400 images used in the validation phase, and 400 images used in

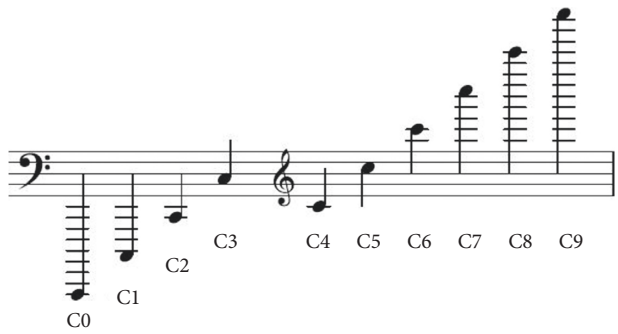

Figure 4: Pitch of notes.

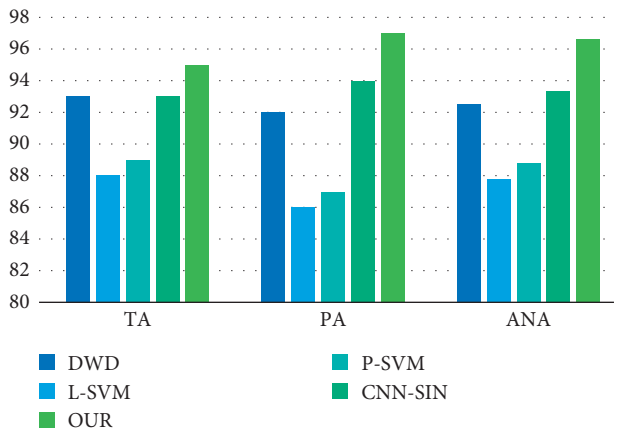

Figure 5: Comparison results on the training set.

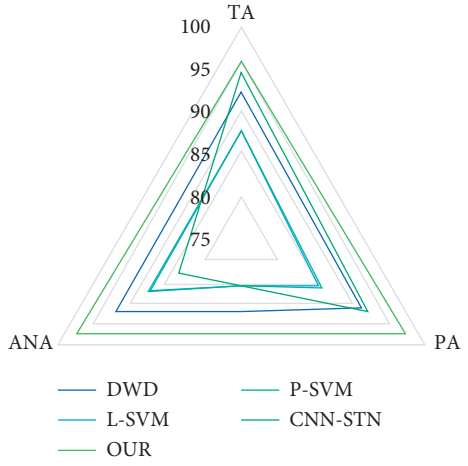

Figure 6: Comparison results on the validation set.

the evaluation phase. DWD, L-SVM, P-SVM, and CNN-STN are the algorithms utilized for comparison. The similarities are depicted in Figures 5–7.

As illustrated in the graph, the algorithm described in this study outperforms the comparison algorithm on three indicators.

Specifically, whether it is the training set, the validation set, or the test set, the performance of our algorithm is far superior to other algorithms. The L-SVM and P-SVM algorithms have the worst performance because their SVM generalization ability and representation ability are poor, while CNN-STN is slightly inferior to our method because of the strong generalization ability of CNN. In the future, we will incorporate more advanced graph neural network methods into our method and add techniques such as the attention mechanism.

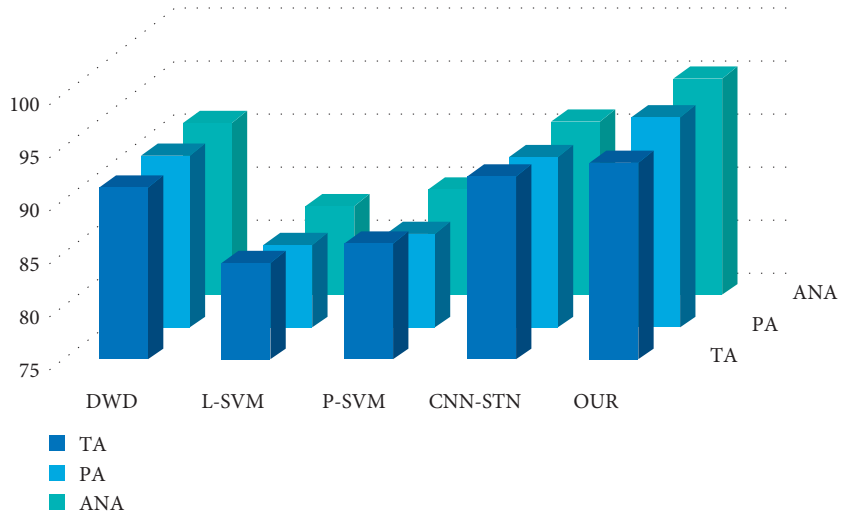

Figure 7: Comparison results on the testing set.

5. Conclusion

Note recognition is a fundamental component of optical score recognition, a technique used extensively in the field of musical information retrieval. We analyze and discuss the state of research on critical technologies for score recognition, as well as the challenges that occur as a result of these technologies. A deep learning-based algorithm for sheet music recognition is being developed in response to current concerns with note recognition accuracy and redundant phases in sheet music photography. A deep neural network is utilized to directly output the temporal value and pitch of the notes from a picture of the whole piece of music's sheet music. The testing conducted with Music Score revealed that this method achieves 0.96 time value and 0.98 pitch accuracy, making it an excellent option for use in music creation.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Pacha and H. Eidenberger, "Towards self-learning optical music recognition," in *Proceedings of the 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 795–800, IEEE, Cancun, Mexico, December 2017.

[2] A. Baró, P. Riba, J. Calvo-Zaragoza, and A. Fornés, "From optical music recognition to handwritten music recognition: a baseline," *Pattern Recognition Letters*, vol. 123, pp. 1–8, 2019.

[3] A. Baró, C. Badal, and A. Fornés, "Handwritten historical music recognition by sequence-to-sequence with attention mechanism," in *Proceedings of the 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 205–210, IEEE, Dortmund, Germany, September 2020.

[4] E. van Der Wel and K. Ullrich, "Optical music recognition with convolutional sequence-to-sequence models," 2017, https://arxiv.org/abs/1707.04877.

[5] A. Ríos-Vila, J. Calvo-Zaragoza, and J. M. Inesta, "Exploring the two-dimensional nature of music notation for score recognition with end-to-end approaches," in *Proceedings of the 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 193–198, IEEE, Dortmund, Germany, September 2020.

[6] J. Moon, M. Kim, Y. Lim, and K Kong, "Conversion program of music score chord using OpenCV and deep learning," *The Journal of the Institute of Internet, Broadcasting and Communication*, vol. 21, no. 1, pp. 69–77, 2021.

[7] A. Rico and A. Fornés, "Camera-based optical music recognition using a convolutional neural network," in *Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 2, pp. 27-28, Kyoto, Japan, November 2017.

[8] J. Calvo-Zaragoza, F. Castellanos, G. Vigliensoni, and I. Fujinaga, "Deep neural networks for document processing of music score images," *Applied Sciences*, vol. 8, no. 5, p. 654, 2018.

[9] H. Purwins, B. Li, T. Virtanen, and J. S.-Y. T. Schluter, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[10] G. F. Chen, "Music sheet score recognition of Chinese gongche notation based on deep learning," in *Proceedings of the 2021 international conference on big data analysis and computer science (BDACS)*, pp. 183–190, IEEE, Kunming, China, June 2021.

[11] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298–2304, 2017.

[12] Y. R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 2887–2905, 2021.

[13] R. M. Pinheiro Pereira, C. E. F. Matos, G. Braz Junior, D. S. de Almeida, and A. C. de Paiva, "A deep approach for handwritten musical symbols recognition," in *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, pp. 191–194, Teresina, Brazil, November 2016.

[14] L. Chen, R. Jin, S. Zhang, S. Lee, Z. Chen, and D. Crandall, "A hybrid HMM-RNN model for optical music recognition," in *Proceedings of the 17th International Society for Music Information Retrieval Conference*, vol. 33, New York City, USA, August 2016.

[15] M. Alfaro-Contreras, J. Calvo-Zaragoza, and J. M. Iñesta, "Approaching end-to-end optical music recognition for homophonic scores," *Pattern Recognition and Image Analysis*, Springer, vol. 11868, pp. 147–158, , Cham, 2019.

[16] A. Baró, P. Riba, J. Calvo-Zaragoza, and A Forn, "Optical music recognition by long short-term memory networks[C]," *International Workshop on Graphics Recognition*, pp. 81–95, Springer, Cham, 2017.

[17] S.N. J. Rajesh, "Recognition of musical instrument using deep learning techniques," *International Journal of Information Retrieval Research*, vol. 11, no. 4, pp. 41–60, 2021.

[18] F. Henkel and G. Widmer, "Multi-modal conditional bounding box regression for music score following," in *Proceedings of the 2021 29th European signal processing conference (EUSIPCO)*, pp. 356–360, IEEE, Dublin, Ireland, August 2021.

[19] P. Torras, A. Baró, L. Kang, and F Alicia, "On the integration of language models into sequence to sequence architectures for handwritten music recognition," *ICDAR (submitted)*, vol. 88, 2021.

[20] Y. R. Lai and A. W. Y. Su, "Deep learning based detection of GPR6 GTTM global feature rule of music scores," in *Proceedings of the 8th International Conference on New Music Concepts*, vol. 56, Treviso, Italy, 2021.

[21] A. Liu, L. Zhang, Y. Mei et al., "Residual recurrent CRNN for end-to-end optical music recognition on monophonic scores," in *Proceedings of the 2021 Workshop on Multi-Modal Pre-training for Multimedia Understanding*, vol. 41, pp. 23–27, Taipei, China, November 2021.

[22] J. Nam, K. Choi, J. Lee, and S.-Y. Y.-H. Chou, "Deep learning for audio-based music classification and tagging: teaching computers to distinguish rock from bach," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 41–51, 2019.

[23] D. Ghosal and M. H. Kolekar, "Music genre recognition using deep neural networks and transfer learning," *Interspeech*, vol. 90, pp. 2087–2091, 2018.

[24] J. Calvo-Zaragoza, G. Vigliensoni, and I. Fujinaga, "Document analysis for music scores via machine learning," in *Proceedings of the 3rd International Workshop on Digital Libraries for Musicology*, pp. 37–40, Springer, Dalian, 2016.

[25] B. L. Sturm, J. F. Santos, O. Ben-Tal, and I. Korshunova, "Music transcription modelling and composition using deep learning," 2016, https://arxiv.org/abs/1604.08723.

[26] F. J. Castellanos, J. Calvo-Zaragoza, and J. M. Inesta, "A neural approach for full-page optical music recognition of mensural documents," in *Proceedings of the 21th International Society for Music Information Retrieval Conference*, pp. 12–16, ISMIR, Montreal, Canada, October 2020.

[27] A. Pacha, K. Y. Choi, B. Coüasnon, Y Ricquebourg, R Zanibbi, and H Eidenberger, "Handwritten music object detection: open issues and baseline results," in *Proceedings of the 2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pp. 163–168, IEEE, Vienna, Austria, April 2018.

[28] F. J. Castellanos, A.-J. Gallego, and J. Calvo-Zaragoza, "Automatic scale estimation for music score images," *Expert Systems with Applications*, vol. 158, Article ID 113590, 2020.

[29] A. Paul, R. Pramanik, S. Malakar, and S Ram, "An ensemble of deep transfer learning models for handwritten music symbol recognition," *Neural Computing & Applications*, vol. 48, pp. 1–19, 2021.

[30] M.-T. Tran, Q.-N. Vo, and G.-S. Lee, "Binarization of music score with complex background by deep convolutional neural networks," *Multimedia Tools and Applications*, vol. 80, no. 7, pp. 11031–11047, 2021.

[31] S. Kittaka, "Making music score with deep autoencoder shinsaburo kittaka, yoko uwate, and yoshifumi nishio," in *Proceedings of the IEEE Workshop on Nonlinear Signal Processing*, Shanghai, China, March 2016.

[32] F. Zalkow and M. Müller, "Using weakly aligned score-audio pairs to train deep chroma models for cross-modal music retrieval," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 184–191, Montréal, Canada, October 2020.