Computer Assisted Chord Detection Using Deep Learning and YOLOV4 Neural Network Model

Yingdong Ru*
metropolitan college of data analyticsBoston UniversityBoston, United States

*Corresponding author: ydru@bu.edu

Abstract. Music symbol recognition is an important part of Optical Music Recognition (OMR). Chord recognition is one of the most important research contents in the field of music information retrieval. It plays an important role in information processing, music structure analysis, and recommendation systems. Aiming at the problem of low chord recognition accuracy in the OMR recognition model, the article proposes a chord recognition method based on the YOLOV4 neural network model. First, the YOLOV4 network model is used to train single-voice scores to obtain the best training model. Then, the scores containing chords are trained through neural network fine-tuning technology. The experimental results show that the method recognizes the chords with great results, the model was tested on the test set generated by MuseScore. The experimental results show that the accuracy of note recognition is high, which can reach the accuracy of duration value of 0.96 which is higher than the accuracy of note recognition of other score recognition models.

Keywords: Optical music recognition, chord recognition, deep learning, object detection.

1. Introduction

Optical score recognition (OMR) is an important way to realize the digitization of score images, and it has a wide range of application prospects in the fields of computer music, digital music libraries, and computer-assisted music teaching. The usual OMR system mainly includes two links: (1) Recognition of score music symbols, that is, converting score images into a collection of graphic objects with musical semantics (2) Semantic reconstruction of music symbols, and all kinds of symbols are organized according to the organization mechanism of the score Integrate organically. In music scores, music symbols can be divided into musical notes and various notation marks. Therefore, the recognition of notes is the recognition of music scores. However, the notes are usually densely arranged, diverse in shape, and small in size, which makes it difficult to recognize the notes.

The role of Optical Music Recognition (OMR) is to effectively convert the score image into the score information [1-4]. The score information is in an editable or playable form, such as MIDI (for playing) And MusicXML (for page layout). Pacha et al. [5] integrated all data sets in the OMR research field and extracted a classification model based on convolutional neural networks. This method is dedicated to training a general musical notation classifier. The experimental results show that the accuracy of the classifier reaches 95%. And Pacha et al. [6] used three different types of deep
learning models (namely, two-stage Faster R-CNN detector, one-stage Retinanet detector, and U-Net) and determined their performance according to a common evaluation standard. Their experimental results can be used as a baseline for general detectors in music symbol detection tasks. Tuggener et al. [7] connected ResNet-101 with the RefineNet up-sampling network as a score feature extractor, and combined with the border detection method to identify various symbols. The test results in the DeepScores and MUSCIMA++ [8] data sets show that for the middle and high scores, the recognition effect of high-frequency music symbols is good, but the recognition effect of non-high-frequency symbols is not ideal; Huang Zhiqing et al. [9] used darknet53 as the backbone network for note feature extraction and then combined feature fusion technology to propose a note recognition model. The recognition accuracy of is 0.87, but the model is limited to single notes. When there are chord notes in the score, the model will have problems such as missing recognition. Van der Wel [10] and others divide the score with staff as the unit, encode it into a sequence vector through a convolutional neural network, and combine it with a recurrent neural network to identify the pitch and time value of the notes in the sequence. The experimental results the accuracy of notes is 0.76. This method needs to divide the music score into rows before entering the model and cannot recognize dense symbols.

The current deep learning-based note recognition has achieved preliminary results. The above methods are only aimed at recognizing single-voice scores, and they cannot recognize dense music symbols, tuplets, etc. However, most of modern music scores are polyphonic scores, which contain a large number of chords. Therefore, this paper proposes a deep learning-based score recognition model for printed scores, that is, the model can accurately identify musical scores and chords.

2. Dataset
This paper uses DeepScoresV2[11] as the training data set, an extended and improved version of the original DeepScores[12] dataset that specifically addresses these issues and makes the following contributions: (a) add 20 formerly absent classes including symbols without fixed size or shape but are nonetheless fundamental to music notation, thereby increasing the list of musical symbols that can be detected by 23%; (b) add ground truth for oriented bounding boxes, thus enabling research into detectors with potentially much higher precision; (c) add ground truth for further higher-level musical semantics, therefore making the dataset valuable for tasks beyond pure music object detection downstream in OMR. We selected 20,000 sheet music pictures, of which 70% were used to train the model, 15% were used to verify the model, and the remaining 15% were used to test the model.

2.1. Data Augmentation
Since the score images collected by the DeepScores V2 dataset are all data under ideal conditions, the trained model lacks generalization ability. In order to make the trained model robust to score images in low-quality and multi-scenarios, this paper proposes four data enhancement methods to simulate noise in different situations. In this paper, Gaussian noise, Gaussian blur, color transformation, rotation and stretching and other elastic deformations are added to the original score image to simulate several undesirable situations that may appear in the score image. Figure 1 (a) simulates the impact of different lighting conditions on the score image by changing the hue, brightness and saturation of the original score image; adding Gaussian noise can simulate the low-quality printing or scanning of the score image, the results are as follows Figure 1(b) shows; Figure 1(c) performs elastic transformations such as stretching and rotation on the original image, which can simulate the folding and distortion of the music score during the scanning process; Figure 1(d) adds Gaussian blur to make The music score produces a partial color reduction effect, which simulates the deterioration of the layout of the music score when stored for a long time.
3. Implementation Details

In the music score image, the shape of the chord is shown as a dense head, that is, there are a large number of detection targets in a small area, which brings a huge challenge to the target detection algorithm. In order to identify small objects such as musical notes, we use CSPDarknet53 as the backbone to improve the size of the receptive field and we also use the FPN network structure for multi-channel feature fusion.

3.1. Detection Model

This paper uses the backbone network of cpsdarknet53[13], combined with feature fusion technology (SPP [14], show in Fig3) to detect chord. The intuitive diagram of the network is shown in Fig 2, the following is the overall process of a musical score data image passing through the network:

a: FPN is to use the pyramid form of CNN hierarchical features, while generating feature pyramids with strong semantic information at all scales. The network is designed with a top-down structure and horizontal connections to integrate a shallow layer with high resolution and a deep layer with rich semantic information. In this way, it is possible to quickly build a feature pyramid with strong semantic information at all scales from a single input image at a single scale. A complete sheet music picture is directly input into the FPN network to obtain multi-layer features.

b: High-level neurons mainly respond to the entire note, and other neurons are more inclined to respond to the local texture information of the note. This network structure enhances the positioning capability of the entire feature structure by propagating strong low-level responses. Because the strong response to the edge and part of the instance is a strong beacon for accurately positioning the instance. Therefore, the horizontal connection is used to connect the low-level and high-level information, so that the network can more accurately identify each note.
c: The previously extracted multi-layer features are adjusted to a feature map with the same scale through up and down sampling, and then feature fusion is performed on the channel to obtain the final feature map.
d: Regression, used to recognize note coordinates.
e: Fully-connection, use to calculate the confidence of the note.

3.2. Loss function
The loss function of the model has three parts: the coordinate loss function of the prediction box, the category loss function predicted by the prediction box, and the confidence loss function of the prediction box. Compared with the yolov3 network model, the yolov4 network optimizes the position loss function of the prediction box.

Yolov4 uses generalized IoU [14] loss, and ciou considers the scale information of the overlap, center distance and aspect ratio of the frame on the basis of iou, its loss function is as formula one.

\[
R_{C_{IoU}} = \frac{\rho^2(b_xb_{x'} + )}{c^2} + \alpha v
\]
\[
v = 4 \frac{4}{\pi^2} (\arctan\frac{w}{h} - \arctan\frac{w_{x'}}{h_{x'}})^2
\]
\[
\alpha = \frac{v}{(1 - IoU) + v}
\]
\[
L_{C_{IoU}} = 1 - IoU + \frac{d^2}{c^2} + \alpha v
\]

3.3. Evaluation Method
The evaluation criteria for chords have been re-established. In chords, if the score detection model can detect each note head, the model can accurately detect the chord. If there are one or more missed detection, the model is deemed to be unable to detect this chord. Since the chord contains a large number of note heads, detecting all the dense note heads is undoubtedly a huge challenge, and the final chord accuracy will also be low.

• detection: Indicates that the chord can be detected successfully.
• miss detection: indicates that there are undetected note-head.
• lack detection: indicates that no chords can be detected at all.

3.4. Network Training
The entire training process is completely end-to-end, directly inputting the score image, calculating the loss function through the model, and finally optimizing the model parameters through the loss function. Data enhancement is applied during training, and different training samples are presented to the
network model each time. The batch size of the training model using the stochastic gradient descent optimizer is 32, the initial learning rate is 0.001, the learning rate decays constantly, and the learning rate is halved every ten cycles. After about 40 cycles, the model began to converge, and the model was trained in about 6 hours.

3.5. Results
Table 1 shows the recognition results of chords. There are a total of 23,251 chords, 12,442 can be fully recognized.

| Evaluation results | Numbers | Instruction |
|--------------------|---------|-------------|
| Total number of chords | 23251 | The total number of chords in the test set |
| Detection | 12442 | Chord detected successfully |
| Miss Detection | 10444 | Detect part of the note head of a chord |
| Lack Detection | 365 | Undetectable chord's note head |

From the chord detection results, it can be seen that the accuracy of this model is 53.5%, the missed detection rate is 44.9%, and the missed detection rate is 1.5%. Since each note in the chord is too dense, the receptive field of the note will be very small, and an accuracy of 53.5% is an acceptable result. In the aspect of chord detection, more in-depth experiments are still needed to improve the detection effect of chords. Fig 6 shows the detection effect of our network.
4. Conclusion and Further Research
In this work, we proposed a complete printed score recognition model based on a deep convolutional neural network. The system only needs to input a complete score to be able to output the category of the symbols. In future work, we plan to use more advanced networks to detect chords.

References
[1] Casey, M; Veltkamp, R.; Goto, M; Leman, M; Rhodes, C; Slaney, M. Content-Based Music Information Retrieval: Current Directions and Future Challenges. Proc. IEEE 2008, 96, 668–696. doi:10.1109/JPROC.2008.916370.
[2] Roland, P. The Music Encoding Initiative (MEI). In Proceedings of the First International Conference on Musical Applications Using XML, Milan, Italy, 19–20 September 2002; pp. 55–59.
[3] Good, M.; Actor, G. Using MusicXML for file interchange. In Proceedings of the Third International Conference on WEB Delivering of Music, Leeds, UK, 15–17 September 2003; p. 153. doi:10.1109/WDM.2003.1233890.
[4] Bainbridge, D.; Bell, T. The challenge of optical music recognition. Comput. Humanit. 2001, 35, 95–121. doi:10.1023/A:1002485918032.
[5] Recognition of Handwritten Music Symbols with Convolutional Neural Codes - IEEE Conference Publication [EB/OL]./2020-11-10.http://ieeexplore.ieee.org/document/8270049.
[6] Pacha A, Hajic J, Calvo-Zaragoza J. A Baseline for General Music Object Detection with Deep Learning [J]. Applied Sciences, Multidisciplinary Digital Publishing Institute, 2018, 8(9): 1488.
[7] Tuggener L, Elezi I, Schmidhuber J, et al. Deep Watershed Detector for Music Object Recognition [J]. arXiv:1805.10548 [cs], 2018.
[8] Hajic J, Pecina P. The MUSCIMA++ Dataset for Handwritten Optical Music Recognition [A]. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)[C]. Kyoto: IEEE, 2017: 39–46.
[9] Z. Huang, X. Jia, and Y. Guo, “State-of-the-art model for music object recognition with deep learning,” Applied Sciences, vol. 9, no. 13, pp. 2645–2665, 2019.
[10] van der Wel E, Ullrich K. Optical Music Recognition with Convolutional Sequence-toSequence Models [J]. arXiv:1707.04877 [cs], 2017.
[11] Tuggener, L., Elezi, I., Schmidhuber, J., Pelillo, M., Stadelmann, T. (2018, August). DeepScores-a dataset for segmentation, detection and classification of tiny objects. In 2018 24th International Conference on Pattern Recognition (ICPR) (pp. 3704-3709). IEEE
[12] Tuggener, L., Satyawan, Y. P., Pacha, A., Schmidhuber, J., and Stadelmann, T. (2020). The DeepScoresV2 dataset and benchmark for music object detection. In 25th International Conference on Pattern Recognition 2020 (ICPR’20), Online, 10-15 January 2021. IAPR.
[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 37(9): 1904–1916, 2015. 2, 4, 7
[14] Zhao-hui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-IoU Loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2020. 3, 4