Optical Medieval Music Recognition

Subjects: Computer Science, Artificial Intelligence | Music
Contributor: Alexander Hartelt, Frank Puppe

Optical Music Recognition (OMR) is one of the key technologies to accelerate and simplify the transcription task in an automatic way. Typically, an OMR system takes an image or manuscript of a musical composition and transforms its content encoded in some digital format such as MEI or MusicXML.

Keywords: Optical Music Recognition ; historical document analysis ; medieval manuscripts

1. Introduction

As most musical compositions in the western tradition have been written rather than recorded, the preservation and digitization of those musical documents by hand is time-consuming and often error-prone. Optical Music Recognition (OMR) is one of the key technologies to accelerate and simplify this task in an automatic way. Since the process of automatically recognizing musical notations is rather difficult, the usual OMR approach divides the problem in several sub-steps: (1) Preprocessing and deskewing, (2) staff line detection, (3) symbol detection and (4) finally the reconstruction and generation of the target encoding.

On modern documents, the OMR process already obtains very good results. However, on historical handwritten documents the performance is much worse and requires expert knowledge in some cases. This is due to the fact that those documents are much more heterogeneous, often differ in their quality, and are affected from various degradations such as bleed-through. Furthermore, the notation of early music documents had been developed and thus changed several times, requiring different solutions. Formulating parts of this expert knowledge and incorporating it into a post-processing pipeline substantially improves the symbol detection task.

This research focuses on symbol detection and segmentation tasks and deals with manuscripts written in square notation, an early medieval notation, which was used from the 11th–12th century onwards. A staff usually contains four lines and three spaces and in most cases a clef marker at the start (**Figure 1**).

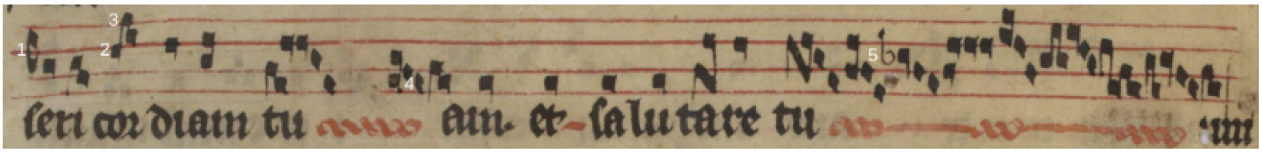

Figure 1. Staff image written in square notation of the Pa904 (Page 13) dataset. One instance for each class has been marked with a small white number in the picture. There are no instances for the classes F-clef, sharp- and natural-accidental in the example image.

The extraction of the symbols is addressed as a segmentation task. For this, a Fully Convolutional Network (FCN), which predicts pixel-based locations of music symbols, is used. Further post-processing steps extract the actual positions and symbol types. Furthermore, the pitch of the symbols can be derived from the position on the staff of the symbol relative to the position of the clef on the staff and other pitch alterations. The fundamental architecture used for this task resembles the U-Net [1] structure. The U-Net structure is symmetric and consists of two major parts: an encoder part consisting of general convolutions and a decoder part consisting of transposed convolutions (upsampling). For the encoder part, several architectures, which are often used for image classification tasks, are evaluated. The decoder part was expanded so that it fits the architecture of the encoder part. Furthermore, a post-processing pipeline is introduced that aims to improve the symbol recognition using background knowledge.

Background knowledge comprises implicit knowledge used by human experts to decide ambiguous situations in addition to the direct notation. The knowledge covers various sources, e.g., conventions, melodic knowledge, constraints based on the notation, knowledge about the scan process, and may be general- or source-document-specific, respectively.

2. Related Work of Optical Medieval Music Recognition

The use of Deep Neural Network techniques has reported outstanding results in many areas related to computer vision, including OMR. Pacha et al. treated this problem as an object detection task [2][3]. Their pipeline uses an existing deep convolutional neural network for object detection such as Faster R-CNN [4] or YOLO [5] and yields the bounding boxes of all detected objects along with their most likely class. Their dataset consists of 60 fully annotated pages from the 16th–18th century with 32 different classes written in mensural notation. Typically, for object detection methods, the mean Average Precision (mAP) was used as the evaluation metric. A bounding box is considered as correctly detected if it overlaps 60% of the ground-truth box. Their proposed algorithm yielded 66% mAP. Further observation showed that especially small objects such as dots or rests posed a significant challenge to the network.

Alternatively, the challenge can be approached as a sequence-to-sequence task [6][7][8][9]. Van der Wel et al. [6] used a Convolutional Sequence-to-Sequence model to segment musical symbols. A common problem for segmentation tasks is the lack of ground truth. Here, the model is trained with the Connectionist Temporal Classification (CTC) loss function [10], which has the advantage that it is not necessary to provide information about the location of the symbols in the image; just pairs of input scores and their corresponding transcripts are enough. The model was trained on user generated scores from the MuseScore Sheet Music Archive (Musescore: https://musescore.org/, accessed on 30 May 2022) to predict a sequence of pitches and durations. In total, the dataset consisted of about 17,000 MusicXML scores, of which 60% were used for training. Using data augmentation, a pitch and duration accuracy of 81% and 94% was achieved, respectively. The note accuracy, where both pitch and duration are correctly predicted, was 80%.

Calvo-Zaragoza et al. [11] used Convolutional Neural Networks for automatic document processing of historical music images. Similar to the task, the corpus used for training and evaluation was rather small, consisting of 10 pages of the Salzinnes Antiphonal (https://cantus.simssa.ca/manuscript/133/, accessed on 30 May 2022) and 10 pages of the Einsiedeln (http://www.e-codices.unifr.ch/en/sbe/0611/, accessed on 30 May 2022). Both of them are also written in square notation. The task was addressed as a classification task, where each pixel in the image was assigned one of four selected classes: text, symbol, staff line or background.

3. Results

Various types of background knowledge about overlapping notes and text, clefs, graphical connections (neumes) and their implications on the position in staff of the notes were used and evaluated. Moreover, the effect of different encoder/decoder architectures and of different datasets for training a mixed model and for document-specific fine-tuning based on an extended OMR pipeline with an additional post-processing step were evaluated. The use of background models improves all metrics in all architectures and in particular the melody accuracy rate (mAR), which is based on the insert, delete and replace operations necessary to convert the generated melody into the correct melody. When using a mixed model and evaluating on a different dataset, the best model achieves without fine-tuning and without post-processing a mAR of 90.4%, which is raised by nearly 30% to 93.2% mAR using background knowledge. With additional fine-tuning, the contribution of post-processing is even greater: the basic mAR of 90.5% is raised by more than 50% to 95.8% mAR.

In conclusion, the post-processing has greatly improved the accuracy and is also very robust. Nevertheless, the post-processing still has a lot of potential for further improvements. Promising approaches will decide whether a note is translucent or not based on local context, in order to use rules about conventions over neumes and to use full or partial transcriptions from similar songs, if available. Another topic is the adaption of background knowledge to different types of manuscripts and notation styles. An ambitious goal is to add a step in the post-processing pipeline that incorporates a language model with knowledge of the most common melodies and transition probabilities between notes in order to correct further errors. This would be even allowed to tackle previously uncorrectable errors, e.g., missing symbols caused by noise, and also to refine the existing post-processing steps with further background knowledge. Finally, such knowledge could be used not only to correct the automatic transcription but also to highlight ambiguous decisions for manual inspection and even to create a second opinion on already manually transcribed documents.

References

1. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv 2015, arXiv:1505.04597.

2. Pacha, A.; Calvo-Zaragoza, J. Optical Music Recognition in Mensural Notation with Region-Based Convolutional Neural Networks. In Proceedings of the 19th International Society for Music Information Retrieval Conference, Paris, France, 23–27 September 2018; pp. 240–247.
3. Pacha, A.; Choi, K.Y.; Coüasnon, B.; Ricquebourg, Y.; Zanibbi, R.; Eidenberger, H.M. Handwritten Music Object Detection: Open Issues and Baseline Results. In Proceedings of the 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), Vienna, Austria, 24–27 April 2018; pp. 163–168.
4. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv 2015, arXiv:1506.01497.
5. Redmon, J.; Divvala, S.K.; Girshick, R.B.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. arXiv 2015, arXiv:1506.02640.
6. Wel, E.; Ullrich, K. Optical Music Recognition with Convolutional Sequence-to-Sequence Models. arXiv 2017, arXiv:1707.04877.
7. Calvo-Zaragoza, J.; Hajič, J., Jr.; Pacha, A. Understanding Optical Music Recognition. ACM Comput. Surv. 2021, 53, 1–35.
8. Baró-Mas, A. Optical Music Recognition by Long Short-Term Memory Recurrent Neural Networks. Master's Thesis, Universitat Autònoma de Barcelona, Bellaterra, Spain, 2017.
9. Calvo-Zaragoza, J.; Rizo, D. End-to-End Neural Optical Music Recognition of Monophonic Scores. Appl. Sci. 2018, 8, 606.
10. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In Proceedings of the Proceedings of the 23rd International Conference on Machine Learning (ICML '06), Pittsburgh, PA, USA, 25–29 June 2006; Association for Computing Machinery: New York, NY, USA, 2006; pp. 369–376.
11. Calvo-Zaragoza, J.; Castellanos, F.J.; Vigliensoni, G.; Fujinaga, I. Deep Neural Networks for Document Processing of Music Score Images. Appl. Sci. 2018, 8, 654.
