



Optical Music Recognition: Recent Advances, Current Challenges, and Future Directions

Jorge Calvo-Zaragoza^(✉), Juan C. Martinez-Sevilla, Carlos Penarrubia,
and Antonio Rios-Vila

University Institute for Computing Research, University of Alicante, Alicante, Spain
jcalvo@dlsi.ua.es

Abstract. Optical Music Recognition (OMR) is an interdisciplinary field that aims to automate the process of transcribing sheet music into a digital format. Over the past few years, significant progress has been made in developing OMR systems that can recognize musical symbols with high accuracy. However, completing the pipeline of OMR remains a challenging endeavor due to the complexity and variability of music notation, and there are several open challenges that need to be addressed. In this position paper, we provide an overview of the current state-of-the-art in OMR through the two main lines of research. We include the problems that have been recently addressed and the techniques that have been considered. We then identify the key challenges that remain, such as learning to reconstruct the music notation, recognizing multiple voices, or dealing with artifacts such as lyrics. Finally, we suggest some possible directions for future research. We argue that addressing these challenges is crucial to making OMR a more practical and useful tool for musicians, scholars, and librarians alike.

Keyword: Optical Music Recognition

1 Introduction

Music is an important part of our cultural heritage. The digital humanities have played a crucial role in preserving and making music accessible to a wider audience. One area of research that has emerged in this context is Optical Music Recognition (OMR), which seeks to automate the process of transcribing written music sources into a digital format [5]. OMR represents an interdisciplinary field that draws on document image analysis, computer vision, and music theory to develop effective algorithms. The progress of this technology multiplies the

Work produced with the support of a 2021 Leonardo Grant for Researchers and Cultural Creators, BBVA Foundation. The Foundation takes no responsibility for the opinions, statements and contents of this project, which are entirely the responsibility of its authors.

options on which digital humanities can operate and, therein, lies its nowadays importance.

Recent advances in deep learning have led to significant improvements in OMR, and several state-of-the-art systems now rely on neural networks. In particular, deep learning has proven effective in addressing some of the challenges that have traditionally plagued OMR, such as staff-line removal [18] or music-object classification [23].

Despite these advances, OMR remains an open problem. In this paper, we provide a position statement on the state of the art in OMR and discuss the current use of deep learning techniques. We then identify the open challenges that need to be addressed and suggest some possible directions for future research. We argue that solving these challenges is crucial to advancing the field of OMR and making it a more practical tool for musicians, scholars, and enthusiasts alike.

For the sake of clarification, let us note that in the rest of the paper we will split the intended discussion into the two great paradigms that currently dominate OMR: the one based on a multi-step pipeline and the one focused on end-to-end formulations.

2 Background

Before elaborating on the aspects related to the state of the art in OMR, we introduce in this section the necessary background to understand the rest of the sections as regards how the task is approached from the two aforementioned paradigms.

2.1 Pipeline-Based Optical Music Recognition

The traditional approach to OMR involves a multi-stage workflow [12]. It consists of image preprocessing, music symbol identification, notation assembly, and encoding.

In the preprocessing step, the music score image is prepared for further analysis. This may include operations such as skew correction, binarization, and staff line removal. In the music symbol identification stage, individual symbols such as notes, rests, and accidentals are detected and classified. Then, notation assembly is performed, where the identified symbols are combined into larger structures such as compound symbols, measures, staves, and systems. Finally, in the encoding stage, the recognized notation is translated into a machine-readable format such as MusicXML or MIDI.

2.2 Holistic Optical Music Recognition

Holistic methods for OMR have been proposed that aim to transcribe entire sections of music notation at once, without explicitly identifying individual symbols. These methods typically involve a staff extraction step, where staves are identified within the whole page, followed by a staff-level end-to-end transcription, where an entire staff is transcribed as a single sequence of symbols.

This approach has the advantage of being more robust to variations in notation and layout, and can be applied to both printed and handwritten music. However, it also poses several challenges, such as the need to deal with overlapping staves and the difficulty of handling polyphonic music with multiple voices. Despite these challenges, holistic methods represent an active area of research in the field of OMR.

3 State of the Art

In this section, we outline the advances that have been taking over the publications in the OMR field for the last years.

3.1 Pipeline-Based Optical Music Recognition

Concerning the first stages of the pipeline, semantic segmentation has emerged as a promising method for OMR. This involves labeling each pixel of the image based on its layout category, such as staff, notes, rests, or lyrics. Recent works have shown success in this endeavor by considering deep learning models [9,33].

Direct music symbol identification, treated as an object detection task, has also received significant attention in recent years [16,21,22]. The idea is to detect and classify individual symbols directly from the music score image. Many researchers have proposed deep learning-based methods for this task. Furthermore, several datasets have been created to facilitate training and benchmarking of these methods [15,28,29].

Notation assembly, the process of combining identified symbols into larger structures such as compound symbols, measures, staves, and systems, has also seen a few learning-based approaches [20]. In these methods, the symbols are first identified, and then their relationships are modeled as a graph. Within this context, Graph Neural Networks (GNN) have also been used to learn the structure of the graph and assemble the symbols into the desired larger structures [3].

3.2 Holistic Optical Music Recognition

Staff detection, as a necessary preprocessing step for staff-line level recognition, has received much attention. This process involves identifying the location of staves as a specific region of the music score image. Staff detection is essential to achieving accurate results in subsequent stages, and many methods have been proposed to address this challenge [10].

Staff-level end-to-end recognition has been widely studied and achieved impressive results [31]. This approach involves recognizing the entire staff-level image and directly outputting the corresponding notation. This approach is particularly useful for old music, where monodic staves are common and music notation can be expressed as a plain sequence. Many publications have explored this topic, which can be categorized into two areas: Image-to-Sequence and CTC-based approaches. The former one takes the path of the sequence-to-sequence mechanisms proposed in the Machine Translation field, with works showing their

effectiveness for OMR [4,24,30]; the latter resembles the Handwritten Text Recognition field by considering the CTC loss function [14], which has been demonstrated to perform accurately for OMR as well [6].

Recently, a step forward has been taken to extend the end-to-end paradigm to more complex music layouts. Several efforts have been proposed to transcribe music scores at a full-page level, from systems that combine layout analysis and staff-level recognition processes [8] to networks that learn to perform transcription in a single step [26]. A representation on how currently state-of-the-art methods address full page transcription can be found in Fig. 1 Although these advances are promising, they only cover currently monophonic music scores.

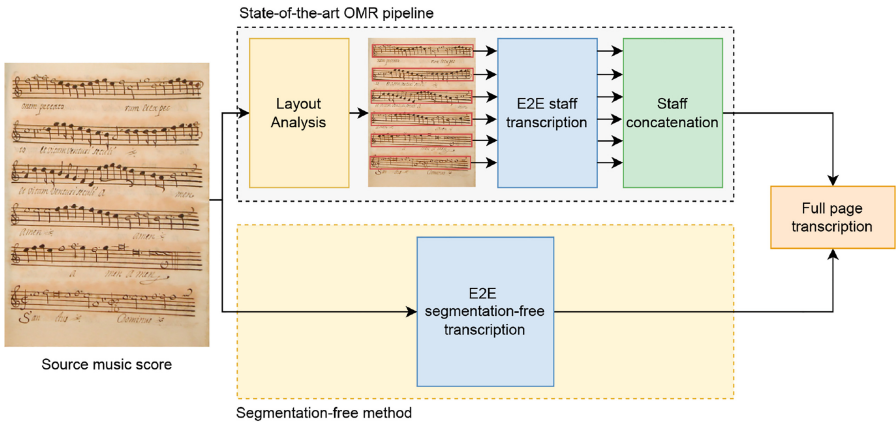


Fig. 1. General overview of the current OMR pipeline (top) in contrast to the holistic full-page approach (bottom), where a previous layout analysis is not needed to transcribe the music symbols in the score.

4 Current Challenges

As the field of OMR continues to evolve, researchers must face several new challenges. In this context, the open challenges differ greatly depending on the paradigm.

4.1 Pipeline-Based Optical Music Recognition

One of the current challenges in pipeline-based OMR is to improve the performance of music-object detection. Despite the advances mentioned above, it is still risky to assume that the algorithms work in any context, especially considering the less common musical symbols or the great variability in size between them as shown in Fig. 2. In many cases, the common object-detection metrics (e.g., mean Average Precision) do not necessarily reflect the goodness of the model for OMR.



Fig. 2. Variability in score formats and notational systems.

While the music-symbol detection stage can benefit from other advances in computer vision, both the notation assembly and the encoding stage are yet to be further developed because they are particular to OMR. For instance, it is not clear how the various music symbols should be generally related to one another, and how this information can be applied to generate a meaningful encoding. Additionally, the encoding stage itself has barely received attention from the literature, finding few works that address it [25]. One possible avenue is to keep on modeling the music notation as a graph and then use GNN to generate an encoding, but this has not been explored.

Furthermore, there is a lack of proper datasets that include the ground truth information required for both the notation assembly and encoding stages, which makes it difficult to train and evaluate full pipeline-based OMR systems.

4.2 Holistic Optical Music Recognition

The main challenge for holistic OMR is to move towards more complex music systems—such as quartets, orchestral scores or simultaneous lyric-accompanied music—as it is currently limited to single-staff and monophonic full-page recognition. Specifically, we refer to this challenge as OMR has to face scores where multiple melodic lines develop at the same time.

In this area, researchers have barely explored the use of language models to improve results [7,27]. These models can help predict the next note based on

the context of the previous symbols, and thereby enhance the accuracy of the output. However, the improvement brought by language models is still limited, and more work is needed to explore their full potential in OMR, especially given the inherent properties of music as a language.

While current methods have shown promise in recognizing single staves, extending these methods to handle multiple staves and complex musical structures is still an open research problem. This requires developing algorithms that can separate and recognize multiple voices, handle overlapping and intersecting staves, and recognize complex musical symbols such as dynamic markings and articulations.

Another pressing challenge is the recognition and alignment of music notation and lyrics, which is essential for cultural heritage, where vocal music is prevalent. This challenge is particularly interesting because it requires the intersection of graphical recognition of written elements and underlying language processing to relate them appropriately. Just a few approaches have been considered [32], but the challenge is quite open for further research.

Additionally, there is a need to move towards actual holistic transcription, skipping the staff detection step and recognizing the music notation directly from the image. This has already begun in text transcription, where end-to-end methods have shown promising results.

Some efforts have been recently proposed in OMR. Specifically, to pianoform scores and aligned music and lyrics transcription. The most recent approach treats these scores like full-page handwritten text recognition paragraphs—thanks to the properties of well-known music encodings—and uses advanced neural network architectures to achieve impressive results. An example is shown in Fig. 3. This represents a significant advancement in OMR research and promises to enhance the capability of automated music transcription.

Finally, there are certain limitations on the music documents that OMR can effectively handle. For instance, there are only a limited number of techniques available to deal with handwritten modern music notation, which is mainly due to the scarcity of datasets for conducting experiments. As a result, OMR is currently biased towards being a tool for historic manuscripts, where although the number of available datasets is still limited, more data can be obtained.

This creates a more extensive challenge, which is the lack of adequate datasets that incorporate the ground truth information necessary to evaluate OMR systems, including different types of musical scores. Creating such datasets is a crucial step in advancing the field.

5 Future Directions

There are several exciting directions for future research in OMR, that can help address the current challenges and push the field forward.


*clefF4	*clefG2	
*k[b-e-a-]	*k[b-e-a-]	
*M3/4	*M3/4	
4r	8.B-L	
.	16B-Jk	
=	=	
4B- 2E-	4g	
.	. b-q	
8cL.	8a-L	
8B-.	8g	
8A- 4E-	8f	
8GJ.	8e-J	
=	=	

Fig. 3. Example on how a pianoform music excerpt can be aligned with its current digital music notation representation, in such a way that it could be read like a text paragraph.

5.1 End-to-End Music Notation Graph Retrieval

One promising direction is to merge the two branches of research in OMR, namely the pipeline-based and holistic approaches, and move towards end-to-end music-notation graph retrieval. This involves developing algorithms that can simultaneously perform staff extraction, music symbol identification, notation assembly, and encoding, using deep learning techniques to model the relationships between different music primitives in a single step. End-to-end approaches have shown promise in other computer vision tasks, and applying similar techniques to OMR can potentially improve recognition accuracy and reduce error propagation. This has only been considered on a very preliminary setting of compound symbols [13], as illustrated in Fig. 4.

5.2 Synthetic Data Generation and Data Augmentation

Another avenue for addressing the limitations of OMR is to generate synthetic data to augment existing datasets [2]. This can help overcome the shortage of appropriate data for various stages of the OMR process. Additionally, data augmentation can enhance the model generalizability and resilience. This trend is evident in other research fields, where deep learning solutions are trained on synthetic databases to develop a basic understanding of the task and then fine-tuned on specific datasets, resulting in impressive outcomes [11, 17, 19].

Generating realistic synthetic data for music notation is a challenging task, and it requires the development of techniques that can capture the complexity and diversity of various music systems.

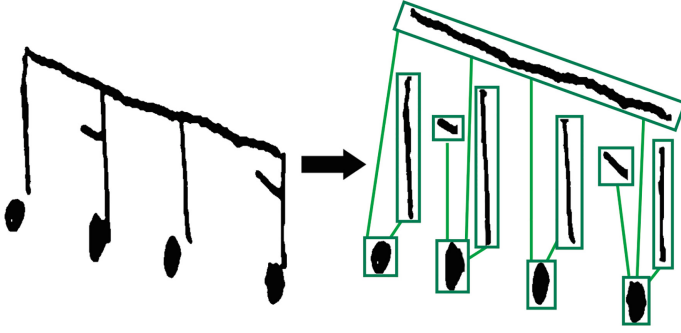


Fig. 4. Image to graph formulation for compound symbols.

5.3 Domain Adaptation Techniques

Domain adaptation techniques can complement the use of synthetic data by allowing models to generalize better to real-world data. Domain adaptation involves adapting models trained on synthetic or other sources of data to perform well on real-world data. This can be particularly useful in OMR, where the variations in notation style, font types, and scanning quality can significantly affect recognition accuracy.

5.4 Self-Supervised Learning

Self-Supervised Learning (SSL) is a machine learning technique that has gained attention in recent years due to its ability to learn useful representations from unlabeled data. Unlike supervised learning, where models are trained on labeled data, SSL involves training models on tasks that do not require explicit supervision. By leveraging large amounts of unlabeled data, SSL can help overcome the limitations of labeled datasets and improve model generalization and robustness.

In OMR, SSL has the potential to be a powerful technique for building more robust and generalizable models. By utilizing the vast amounts of unlabeled music data available on the internet, SSL can potentially help overcome the lack of labeled data for certain music systems or notation styles [1]. SSL also presents an opportunity to create general-purpose music transcription models that can transcribe music across various genres and styles, which can help capture the complexity and diversity of different music systems.

5.5 Foundational Models for OMR

As a crucial long-term objective, the creation of foundational models for OMR can establish a shared framework for the field, making it easier to compare different approaches. Foundational models would offer a cohesive representation of music notation that can be applied throughout various stages of the OMR pipeline and across diverse music systems or, indeed, produce general-purpose

holistic solutions to perform music scores transcription. This can enhance recognition accuracy, diminish the spread of errors, and propel the field toward more universal and resilient OMR systems.

References

1. Bardes, A., Ponce, J., LeCun, Y.: Vicreg: variance-invariance-covariance regularization for self-supervised learning. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, 25–29 April 2022 (2022)
2. Baró, A., Riba, P., Calvo-Zaragoza, J., Fornés, A.: From optical music recognition to handwritten music recognition: a baseline. *Pattern Recognit. Lett.* **123**, 1–8 (2019)
3. Baró, A., Riba, P., Fornés, A.: Musigraph: optical music recognition through object detection and graph neural network. In: Porwal, U., Fornés, A., Shafait, F. (eds.) *Frontiers in Handwriting Recognition - 18th International Conference, ICFHR 2022, Proceedings. Lecture Notes in Computer Science*, Hyderabad, India, 4–7 December 2022, vol. 13639, pp. 171–184. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-21648-0_12
4. Baró, A., Badal, C., Fornés, A.: Handwritten historical music recognition by sequence-to-sequence with attention mechanism. In: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 205–210 (2020)
5. Calvo-Zaragoza, J., Jr, J.H., Pacha, A.: Understanding optical music recognition. *ACM Comput. Surv. (CSUR)* **53**(4), 1–35 (2020)
6. Calvo-Zaragoza, J., Toselli, A.H., Vidal, E.: Handwritten music recognition for mensural notation with convolutional recurrent neural networks. *Pattern Recognit. Lett.* **128**, 115–121 (2019)
7. Calvo-Zaragoza, J., Toselli, A.H., Vidal, E.: Hybrid hidden Markov models and artificial neural networks for handwritten music recognition in mensural notation. *Pattern Anal. Appl.* **22**(4), 1573–1584 (2019)
8. Castellanos, F.J., Calvo-Zaragoza, J., Inesta, J.M.: A neural approach for full-page optical music recognition of mensural documents. In: *Proceedings of the 21st International Society for Music Information Retrieval Conference*, pp. 558–565. ISMIR, Montreal (2020)
9. Castellanos, F.J., Calvo-Zaragoza, J., Vigiensoni, G., Fujinaga, I.: Document analysis of music score images with selectional auto-encoders. In: *Proceedings of the 19th International Society for Music Information Retrieval Conference*, pp. 256–263 (2018)
10. Castellanos, F.J., Garrido-Munoz, C., Ríos-Vila, A., Calvo-Zaragoza, J.: Region-based layout analysis of music score images. *Expert Syst. Appl.* **209**, 118211 (2022)
11. Coquenat, D., Chatelain, C., Paquet, T.: Dan: a segmentation-free document attention network for handwritten document recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 8227–8243 (2023)
12. Fujinaga, I., Vigiensoni, G.: The art of teaching computers: the SIMSSA optical music recognition workflow system. In: *27th European Signal Processing Conference, EUSIPCO 2019, A Coruña, Spain, 2–6 September 2019*, pp. 1–5. IEEE (2019)
13. Garrido-Munoz, C., Ríos-Vila, A., Calvo-Zaragoza, J.: A holistic approach for image-to-graph: application to optical music recognition. *Int. J. Doc. Anal. Recognit.* **25**(4), 293–303 (2022)

14. Graves, A., Fernández, S., Gomez, F.J., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the Twenty-Third International Conference on Machine Learning, (ICML 2006), Pittsburgh, Pennsylvania, USA, 25–29 June 2006, pp. 369–376 (2006)
15. Hajic, J., Pecina, P.: The MUSCIMA++ dataset for handwritten optical music recognition. In: 14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, 9–15 November 2017, pp. 39–46. IEEE (2017)
16. Huang, Z., Jia, X., Guo, Y.: State-of-the-art model for music object recognition with deep learning. *Appl. Sci.* **9**(13), 2645 (2019)
17. Kang, L., Riba, P., Rusiñol, M., Fornés, A., Villegas, M.: Pay attention to what you read: non-recurrent handwritten text-line recognition. *Pattern Recogn.* **129**, 108766 (2022)
18. Konwer, A., et al.: Staff line removal using generative adversarial networks. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 1103–1108. IEEE (2018)
19. Li, M., et al.: Trocr: transformer-based optical character recognition with pre-trained models (2021). arXiv preprint [arXiv:2109.10282](https://arxiv.org/abs/2109.10282)
20. Pacha, A., Calvo-Zaragoza, J., Hajic Jr., J.: Learning notation graph construction for full-pipeline optical music recognition. In: Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, 4–8 November 2019, pp. 75–82 (2019)
21. Pacha, A., Choi, K.Y., Coiñason, B., Ricquebourg, Y., Zanibbi, R., Eidenberger, H.: Handwritten music object detection: open issues and baseline results. In: 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), pp. 163–168. IEEE (2018)
22. Pacha, A., Hajič, J., Jr., Calvo-Zaragoza, J.: A baseline for general music object detection with deep learning. *Appl. Sci.* **8**(9), 1488 (2018)
23. Paul, A., Pramanik, R., Malakar, S., Sarkar, R.: An ensemble of deep transfer learning models for handwritten music symbol recognition. *Neural Comput. Appl.* **34**(13), 10409–10427 (2022)
24. Ríos-Vila, A., Iñesta, J.M., Calvo-Zaragoza, J.: On the use of transformers for end-to-end optical music recognition. In: Pattern Recognition and Image Analysis: 10th Iberian Conference, IbPRIA 2022, Aveiro, Portugal, 4–6 May 2022, Proceedings, pp. 470–481. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-04881-4_37
25. Ríos-Vila, A., Esplà-Gomis, M., Rizo, D., Ponce de León, P.J., Iñesta, J.M.: Applying automatic translation for optical music recognition’s encoding step. *Appl. Sci.* **11**(9), 3890 (2021)
26. Ríos-Vila, A., Inesta, J.M., Calvo-Zaragoza, J.: End-to-end full-page optical music recognition for mensural notation. In: Proceedings of the 23rd International Society for Music Information Retrieval Conference, pp. 226–232. ISMIR, Bengaluru (2022)
27. Torras, P., Baró, A., Kang, L., Fornés, A.: On the integration of language models into sequence to sequence architectures for handwritten music recognition. In: Proceedings of the 22nd International Society for Music Information Retrieval Conference, pp. 690–696 (2021)
28. Tuggener, L., Elezi, I., Schmidhuber, J., Pelillo, M., Stadelmann, T.: Deepscores—a dataset for segmentation, detection and classification of tiny objects. In: 24th International Conference on Pattern Recognition, ICPR 2018, Beijing, China, 20–24 August 2018, pp. 3704–3709. IEEE Computer Society (2018)

29. Tuggener, L., Satyawan, Y.P., Pacha, A., Schmidhuber, J., Stadelmann, T.: The deepscoresv2 dataset and benchmark for music object detection. In: 25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event/Milan, Italy, 10–15 January 2021, pp. 9188–9195. IEEE (2020)
30. van der Wel, E., Ullrich, K.: Optical music recognition with convolutional sequence-to-sequence models. In: Cunningham, S.J., Duan, Z., Hu, X., Turnbull, D. (eds.) Proceedings of the 18th International Society for Music Information Retrieval Conference, pp. 731–737 (2017)
31. Wen, C., Zhu, L.: A sequence-to-sequence framework based on transformer with masked language model for optical music recognition. *IEEE Access* **10**, 118243–118252 (2022)
32. Wick, C., Puppe, F.: Experiments and detailed error-analysis of automatic square notation transcription of medieval music manuscripts using cnn/lstm-networks and a neume dictionary. *J. New Music Res.* **50**(1), 18–36 (2021)
33. Wick, C., Hartelt, A., Puppe, F.: Staff, symbol and melody detection of medieval manuscripts written in square notation using deep fully convolutional networks. *Appl. Sci.* **9**(13), 2646 (2019)