END-TO-END FULL-PAGE OPTICAL MUSIC RECOGNITION FOR MENSURAL NOTATION

Antonio Ríos-Vila
U.I for Computer Research
University of Alicante, Spain
arios@dlsi.ua.es

José M. Iñesta
U.I for Computer Research
University of Alicante, Spain
iñesta@dlsi.ua.es

Jorge Calvo-Zaragoza
U.I for Computer Research
University of Alicante, Spain
jcalvo@dlsi.ua.es

ABSTRACT

Optical Music Recognition (OMR) systems typically consider workflows that include several steps, such as staff detection, symbol recognition, and semantic reconstruction. However, fine-tuning these systems is costly due to the specific data labeling process that has to be performed to train models for each of these steps. In this paper, we present the first segmentation-free full-page OMR system that receives a page image and directly outputs the transcription in a single step. This model requires only the annotations of full score pages, which greatly alleviates the task of manual labeling. The model has been tested with early music written in mensural notation, for which the presented approach is especially beneficial. Results show that this methodology provides a solution with promising results and establishes a new line of research for holistic transcription of music score pages.

1. INTRODUCTION

The majority of music creations, both historical and modern, are only available through music scores. Unfortunately, most of them have never been stored in a structured digital format, such as MEI [1] or Humdrum **kern [2], that allows their indexing, retrieval, and digital processing. The high cost of performing a manual transcription of these documents demands processes that solve this task automatically.

Optical Music Recognition (OMR) is the research field that studies how to computationally read music scores [3]. Most OMR literature is framed within a multi-stage workflow, where several steps, such as image binarization [4], staff-line removal [5], symbol classification [6, 7], and notation assembly [8], are considered. These pipelines, although effective to some extent, are hard to manage in a production environment, as they imply the development of several models to obtain a complete OMR system.

With the advent of machine learning technologies, namely those related to deep neural networks, OMR systems have evolved towards alternatives that simplify these complex workflows. Specifically, two trends emerged from this advance. On the one hand, complex multi-stage pipelines for symbol isolation and retrieval have been replaced by object detection algorithms [9], where symbols are directly located in the image. On the other hand, holistic strategies—commonly referred to as end-to-end approaches [10, 11]—are also used to transcribe music scores. In this case, the system directly outputs the transcription of a single-staff image.

Although the advances brought about by these new technologies have simplified these pipelines, these systems still rely on other models to complete the transcription of scores. In the case of the object detection approach, models for notation reconstruction and sequence encoding need to be implemented to obtain a readable representation of the music score. End-to-end approaches, in turn, are only able to transcribe single music staves, so staff detection algorithms have to be implemented to retrieve all the information from the page and be recognized individually [12, 13]. The latter pipeline is represented in the upper workflow of Fig. 1.

Following the path of end-to-end approaches, in this paper we present the first segmentation-free strategy for full-page music transcription that unifies the two required steps—see the lower workflow in Fig. 1—based on evolving the current state of the art. For this model to be trained, only page-level annotations—which could be available in any music encoding—are required. This greatly alleviates the labeling of music scores, as additional OMR-specific data preprocessing steps are avoided.

The motivation for implementing this idea is to solve some issues of pipeline-based end-to-end approaches. The main inconvenience is that they are composed of two models that are trained specifically on independent tasks—that is, each document has to be manually labeled twice. First, bounding boxes and region classes have to be inserted to train the staff detection model. Once this information has been retrieved, the corresponding symbolic music representation must be written for each staff present in the dataset to train the staff-level recognition system. Therefore, labeling a corpus to train a production-ready OMR system for full-page transcription is costly—due to the need for manual segmentation and transcription of individual music staves—and time-consuming [14]. Another issue is that there has to be assumed an accumulated er-

© F. Author, S. Author, and T. Author. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

Attribution: F. Author, S. Author, and T. Author, "End-to-End Full-Page Optical Music Recognition for Mensural Notation", in *Proc. of the 23rd Int. Society for Music Information Retrieval Conf.,* Bengaluru, India, 2022.

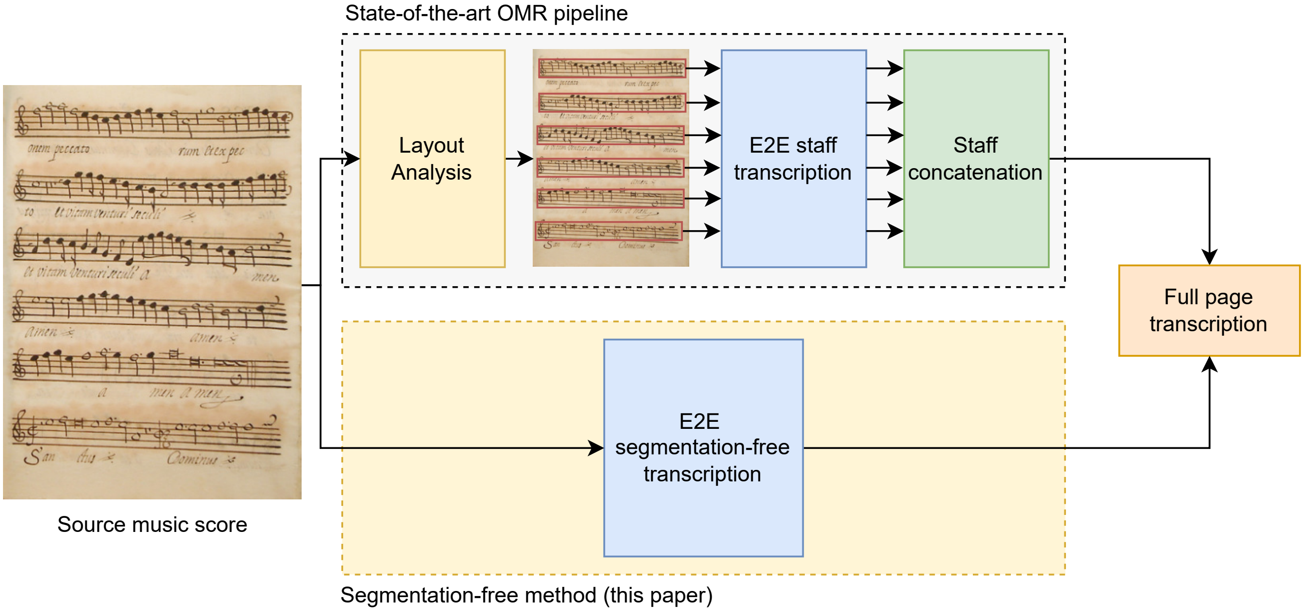


Figure 1: General overview of the current full page OMR pipeline (top) in contrast to this paper proposal (bottom), where a previous layout analysis is not needed to transcribe the music symbols in the score.

ror that is carried over from one step to the next, as some works have studied [12, 15]. That is, the performance of a specific step is highly related to the effectiveness of the previous one, which is also dependent on their predecessors, thus creating a snowball effect.

2. EXTENDING OMR TRANSCRIPTION SYSTEMS

In this section, we explain how state-of-the-art end-to-end OMR models work, what issues they present when transcribing a complete music document, and how they can be adapted to a full-page scenario.

2.1 End-to-end staff transcription systems

The state-of-the-art end-to-end music staff transcription systems are neural models that directly output the music notation sequence of a given single-staff image. These networks are based on two main blocks: first, there is an encoder block, which filters an input image to learn and establish its relevant features. Commonly, this part is approached with Convolutional Neural Networks (CNN). Then, the output of this module is processed by the decoder block, which models temporal dependencies in the obtained features—which are represented as a vector sequence—to enhance results. This second module is commonly implemented with Recurrent Neural Networks (RNN). The whole model, referred to in the literature as a Convolutional Recurrent Neural Network (CRNN), is trained using the Connectionist Temporal Classification (CTC) [16] loss strategy, which maximizes through expectation-maximization the probability of obtaining a music sequence in a given notation alphabet. That is, the CTC method forces the network to align the output sequence to the available information extracted from the image, retrieving the full transcription in a single step. One important step to understand the model is how the output of the encoder is adapted to be processed as a sequence in the decoder. State-of-the-art implementations [10] reshape this obtained feature map—which is a 3D structure of size (h, w, c), where h is the height of the feature map, w is its width, and c is the number of filters (*channels*) obtained from the last convolutional layer—, by concatenating all the consecutive frames on the height axis of the image. That is, we obtain a sequence-like 2D structure that now has a shape of $(w, c \times h)$. This methodology has proven to be effective to transcribe both printed music staves [17] and handwritten ones [10, 11].

2.2 The issues with full-page images

The methodology described in the previous section is ideal for single-staff images, as it solves the transcription problem by reading the obtained feature map from left to right. However, in the case of page images, this formulation cannot be applied. The main issue with this formulation is that a single frame—which is an image column—contains more than one symbol. This is troublesome in two ways. First, transcription and retrieval of the score would result in a costly task, as ground-truth should be written and read from top to bottom and then from left to right, introducing undesirable post-processing methods to obtain a readable score. The second problem is related to the training of the system, as the CTC method should maximize the probability of non-consecutive temporal symbols in a few time steps. That is, the network has to classify symbols from different staves in the same image columns. This implicit interpretation is wrong, as these symbols are located in farther positions of the sequence, as they belong to different staves.

2.3 From one staff to full page

To transcribe full-page music scores, a new interpretation of the obtained features map is considered. It is required to use an alternative reshape method from the output of the encoder block. In this paper, we apply a score unfolding reshape method. Instead of concatenating frame-wise elements along the height axis, as it is done for staff transcription, we reshape the image in the concatenation of all of its rows, obtaining then a $(h \times w, c)$ sequence. From a high-level perspective, this method can be understood as a staff concatenation process, as it is depicted in Fig. 2. This operation has to be done from top to bottom of the page, as it is crucial to correctly transcribe the music score.

Processing the feature maps this way prevents the aforementioned issues, as the score can be labeled as it is naturally read—from top to bottom and left to right—and the CTC method will not face the problem of collision of non-consecutive times, as now the used sequence does not have merged features from different staves.

This methodology, although theoretically a valid solution, presents some points that should be noted. The adaptation of the current music transcription models only needs the transcription of the full page for training, as the model directly outputs the music sequence from the input image. By avoiding the previously required object detection algorithm from the pipeline, now the system has to face two main challenges: (i) identifying all the staves in the score and (ii) transcribing them into an ordered sequence. The second challenge is solved by the reshape method, as we force the model to align and read all the staves in a specific order. In this case, the CTC *blank* token—which is an additional element introduced in the notation vocabulary to indicate time step separations—denotes both music element separations and staff breaks, since the single-staff-like produced sequence identifies the first symbol of the next staff as a consecutive timestep to the last symbol of the previous one. Challenge (i), however, is somewhat relegated to the network learning. The hypothesis is that, by not modifying the sequence structure in the decoder block, this alignment can be learned and mapped by the encoder. This hypothesis needs to be validated by the performance of the proposed models during experimentation.

2.4 Further considerations

This paper evolves the already established OMR staff recognition models by implementing a learned music score unfolding. Theoretically, this method is still a single-staff transcription system. All the staves of the pages can be understood as a single long staff that, due to physical constraints of paper, had to be divided into several staves. This methodology implicitly learns to reconstruct this original interpretation and transcribe the resulting long single-staff image in one step. The alignment process between the different staves is learned by the network during training.

It is important to note that, at this point, the proposed method is suitable only for monophonic staves, as polyphony requires additional information and processing to be transcribed holistically. Vocal music scores—such as those written in mensural notation—can benefit from this approach because polyphony is usually written through a series of independent monophonic voices. For this reason, the experiments will be carried out with early music scores written in mensural notation, as detailed in Section 3.2.

3. EXPERIMENTAL SETUP

In this section, we describe the proposed models to address full-page OMR, present the used corpora during experimentation, and define the metrics used to evaluate the performance of our proposals [1] .

3.1 Models

As mentioned throughout this paper, we are adapting state-of-the-art OMR models to full-page transcription. However, two additional variations have been included in the proposed model to extend the study on this topic. All the presented models have a fully convolutional block, which acts as an encoder of the input image features. This network is composed of stacked convolutional layers, which end up producing a feature map of size $(h/32, w/8, c, b)$, h and w being the height and the width of the input image, c the filters in the last convolutional layer, and b the batch size. The downscale of the original image size is produced by pooling operators. Then, the decoding architectures proposed for processing the sequence obtained after the reshaping procedure are described below.

3.1.1 Recurrent Neural Network

We follow the implementation of the original CRNN-CTC staff transcription model from [10], where the reshaped feature map is fed into a Bidirectional LSTM (BLSTM) and linearly projected onto the music notation dictionary. Specifically, we implemented a BLSTM with 256 units, whose output matches the output space of the fully convolutional network.

3.1.2 The Transformer

The base model of this work uses RNNs to process the reshaped feature map as a sequence. However, a recent recurrent-free model has gained popularity in the Natural Language Processing (NLP) field: the Transformer [18]. This model replaces the RNN architecture by implementing sequence modeling through attention mechanisms and position learning. This model solves some common issues that RNNs have—such as processing long sequences, training time effort, and contextual information retrieval—at the cost of needing more data to converge. As we have observed in the reshape step, the model would have to process significantly long sequences in one step, something that can have a negative impact on the performance of RNNs. For this reason, the first alternative model implemented in this paper replaces the recurrent layer of the CRNN model with a Transformer encoder (referred to as

[1] Source code for the implementation of the presented models can be found in `https://github.com/antoniorv6/ismir_fpomr.git`

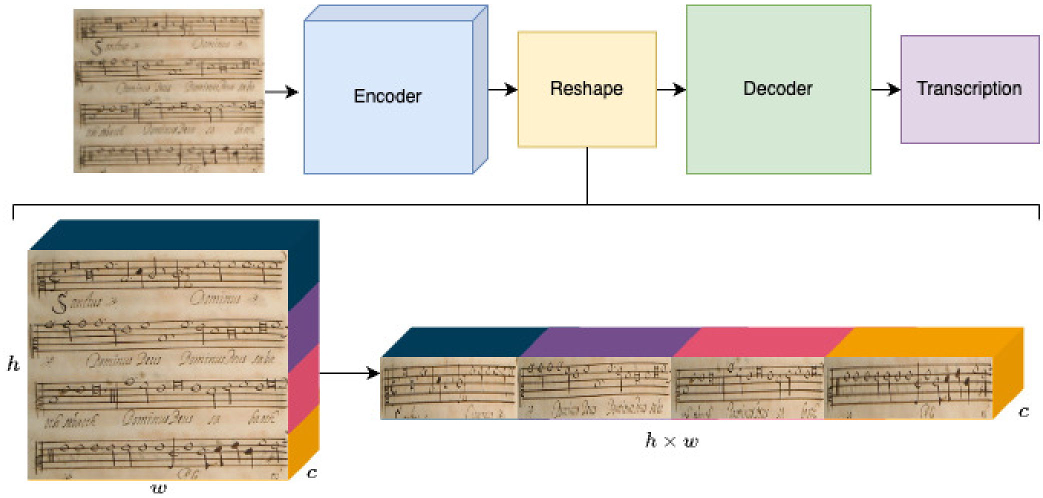


Figure 2: Graphic visualization of the reshape method to adapt current systems to full-page transcription. The reshape module learns how to separate and concatenate the staves on a page in a single line. Note that the original image has been used in the reshape for clarity of explanation, but the alignment is done in the *feature space* extracted by the encoder.

CNNT hereafter). In particular, we implemented one encoder layer with an embedding size of 512, a feed-forward dimension of 1024, and 8 attention heads.

3.1.3 Sequence-processing-free module

One of the challenges the model has to face during the score recognition is the alignment of the staves extracted from a 2D image to build a 1D sequence. This leads to the question of how a sequence processing module could impact learning this specific task while performing backpropagation. In analogous text recognition models, like [19], the solution lies in preserving the prediction space in 2 dimensions, applying backpropagation directly to the retrieved feature map before being reshaped. We have implemented this model to analyze the impact of the sequence processing module on the score page transcription. It is also based on fully convolutional layers, so it will be referred to as FCN in the results section.

3.2 Corpora

Two mensural-notation music datasets with different characteristics in engraving style were used, in order to represent the different challenges that the model can face in these real-case scenarios.

The first corpus is "Il Lauro Secco" [20] (denoted as SEILS), which corresponds to an anthology of 150 typeset printed images of the 16th-century Italian madrigals.

The second corpus is the CAPITAN dataset [10], which contains a complete ninety-six pages manuscript from the 17th-century containing a handwritten *missa*.

Specific details about the used corpora can be found in Table 1, and Fig. 3 depicts examples of these documents.

Table 1: Details of the corpora regarding the pages' features, such as sizes in pixels, number of samples and staves per page, number of symbols present and unique symbols per dataset (*vocabulary*).

	SEILS	CAPITAN
Num pages	150	123
Max page size	1200×813	1593×2126
Min page size	1200×813	1100×780
Avg staves per page	4	10
Max staves per page	9	12
Min staves per page	1	2
Avg symbols per page	222	136
Max symbols per page	331	220
Min symbols per page	110	23
Unique symbols	183	321

3.3 Data augmentation

One constraint that our proposed model can find observing Table 1 is the scarcity of data, as these kind of early music documents usually have few pages completely labeled. To address this potential issue, we applied a data augmentation process to increase the number of samples per corpus. This procedure is composed of several image distortion operations, such as reduction, erosion, dilation, or perspective modifications. These distortions are randomly applied for each batch sample, allowing us to obtain massively extended corpora that also have an added variability for the samples. The SEILS corpus is increased up to 29000 pages and the CAPITAN dataset to 24000.

3.4 Metrics

Currently, there are no OMR-specific metrics to evaluate the performance of the transcription systems. In this paper, we resort to the Symbol Error Rate (SER), which is the most commonly used metric in the OMR literature for end-

(a) SEILS (b) CAPITAN

Figure 3: Music page examples from the used corpora.

Table 2: Test SER (%) obtained for the studied models. Castellanos et al.'s work is included in the last row as a reference value to observe how current OMR pipelines work to transcribe the used corpora in this paper.

Model	Augmentation	SEILS	CAPITAN
CRNN	-	6.3	26.6
	✓	**4.3**	**15.5**
CNNT	-	12.9	28.4
	✓	7.2	18.2
FCN	-	23.3	89.5
	✓	13.3	22.5
Staff-based [12]	-	3.6	10.8

to-end approaches. This metric is computed as

$$\mathrm{SER}\,(\%) = 100\,\frac{\mathrm{ED}(\hat{S}, S)}{|S|}\,,$$

where ED is the edit distance between the transcription hypothesis $\hat{S}$ and its corresponding ground truth, S. $|S|$ is the length (in tokens) of S. We chose this metric because it represents accurately the recognition performance of the model and correlates with the effort that a user would have to invest to manually correct the output sequence.

The datasets have been split into fixed partitions, where 60% of the samples have been used for training, 20% have been used for validation, and 20% for testing.

4. RESULTS

Table 2 shows the results obtained by the studied methodology on the test set for each corpus. Results reported by Castellanos et al. [12] are included to establish a reference value from a state-of-the-art algorithm based on a standard OMR pipeline (the one depicted in Fig. 1-top). Note that this reference model is trained under more favorable conditions, as it addresses the full-page transcription in two separate tasks, which have specific data for training each. However, it requires much more labeling work to build the training sets than our segmentation-free implementation. Therefore, it should be understood only as a reference and not as a competing approach.

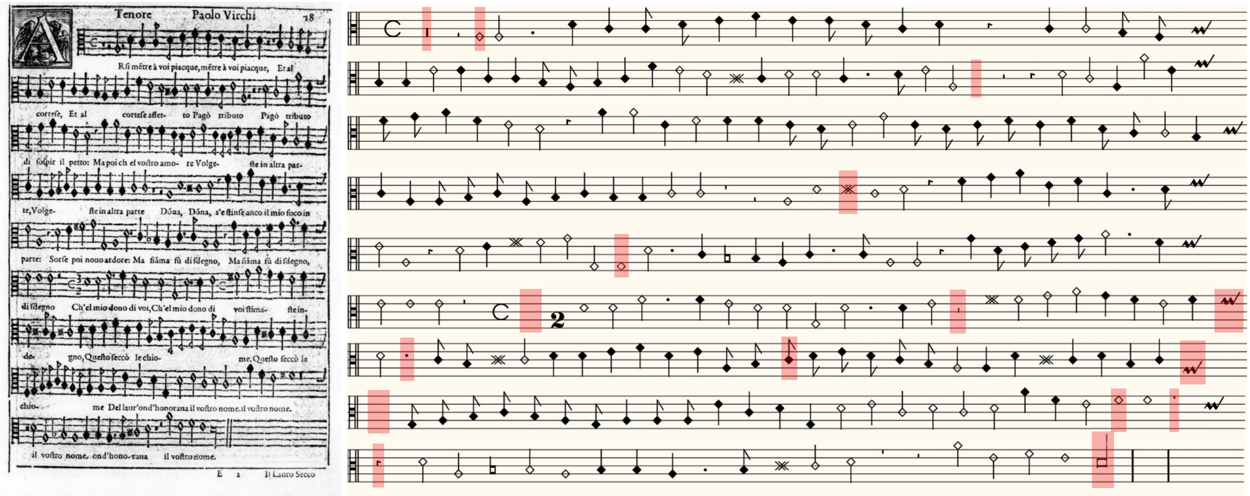
The results obtained show that the extended models were able to transcribe full-page scores with fair SER values, below 30% except for the FCN without data augmentation in the CAPITAN dataset. These error values also scale depending on the engraving complexity of each corpus, being the printed documents (SEILS) easier to transcribe than the handwritten ones (CAPITAN). The model that reported the best results was the combination of the convolutional network with the RNN decoder (CRNN), which obtained an error rate of 4.3% in the SEILS corpus and 15.5% in CAPITAN.

The models that use sequence processing decoders (CRNN and CNNT) performed better than the single FCN network. This was expected to some extent, as these architectures exploit and optimize sequential information to improve their performance, which is a considerable advantage over the FCN approach. In fact, they also seem to bring some robustness to the model against the scarcity of training data, if we compare their results with the high error rate of the FCN when no data augmentation was applied to the CAPITAN corpus.

Continuing the data dependency analysis, it can be observed that applying data augmentation, significantly improved the overall performance of all models. The most notable case of this dependency was found in the FCN network on the CAPITAN corpus, where overfitting issues seemed to be solved with the augmented database. For the other models, improvements were reported as well, reducing the SER by approximately 30%–40%. In other words, the models are able to work with few samples, but they require a considerable amount of data to obtain their best results.

Comparing the two sequence processing decoders, we observe that RNNs performed better than CNN Transformers in all cases. The reason for this is aligned with the results obtained in works that explore the use of these models on document transcription [22]. Looking at the Table 1, we observe that both models contain, on average, short sequences. Transformers, by replacing recurrence with self-attention and position encoding, improved computation time and accuracy at the cost of more data needed to converge. However, the Transformers literature has reported relevant improvements with long sequences, containing approximately 512 tokens, with which the RNNs have convergence problems. In this case, the Transformer model is in a disadvantageous scenario, where few data samples are available to train and the output sequences are relatively short, which is a much better scenario for RNN-based decoders.

For the sake of visualization, Fig. 4 presents the results obtained in a test set page from the SEILS dataset by the best model (CRNN). As can be seen, the system produces a good transcription, in which most of the symbols are correctly labeled and aligned within their corresponding



(a) Original page.

(b) Transcribed prediction with CRNN, rendered with the Verovio toolkit [21].

Figure 4: Visualization of the transcription produced by the CRNN model in a music page from the SEILS dataset. Errors are highlighted. Note that the output is actually produced on a single line, but a multi-line representation of the score has been reconstructed to facilitate comparison with the original document. In this particular case, the obtained SER is 6.1% (higher than the average on this corpus).

staves. If we analyze the produced errors, most of them are subtle—such as vertical position misplacement— and can be easily corrected with score editing software. This happens often with narrow symbols—such as rests—whose manual annotation would also have been difficult to perform.

5. CONCLUSION

In this paper, we present a first segmentation-free end-to-end approach for full-page OMR. This method is trained with weakly-annotated data: it only requires a set of page images with their corresponding transcription, in contrast to current state-of-the-art full-page OMR pipelines that require spatial information—such as bounding boxes or pixel-wise regions. Our methodology extends the current staff-level end-to-end systems to full-page transcription by applying a concatenation step that learns how to process the two-dimensional sequence document.

We evaluated three variants of this model with two early music collections written in mensural notation, which have been used in many other works as a benchmark for OMR. The reported results showed that the proposed system produces competitive results for full-page transcriptions. Although precision is slightly lower than a multi-stage OMR pipeline, the proposed segmentation-free approach stands as an interesting alternative for those models. The model provides a favorable trade-off between the cost of labeling and the system's accuracy.

Several avenues for future research arise from this work. First, it only covers the extent of historic vocal music in mensural notation, where only monophonic staves are found. In fact, the method can handle any page representing monophonic staves but, in modern music, the coverage is more restricted, since polyphony is much more common. This scenario could be an interesting case to analyze alternative architectures for OMR, such as attention-based systems [23] or non-CTC-based image-to-sequence architectures [24], which do not have to be constrained by the specific layout.

Another aspect to consider in the future is the need for data. Although one of the goals of this work is to reduce the labeling effort, the models still require a large number of fully annotated pages. We believe that further research needs to be done to study both transfer learning and self-supervised learning approaches to address this issue.

6. ACKNOWLEDGEMENTS

This paper is part of the project MultiScore (PID2020-118447RA-I00), funded by MCIN/AEI/10.13039/501100011033. The first author is supported by grant ACIF/2021/356 from "Programa I+D+i de la Generalitat Valenciana". Third author was supported with a 2021 Leonardo Grant for Researchers and Cultural Creators, BBVA Foundation. The Foundation takes no responsibility for the opinions, statements and contents of this paper, which are entirely the responsibility of its authors.

7. REFERENCES

[1] A. Hankinson, P. Roland, and I. Fujinaga, "The music encoding initiative as a document-encoding framework." in *Proc. of the 12th Int. Society for Music Information Retrieval Conference*, 2011, pp. 293–298.

[2] D. Huron, "Humdrum and kern: Selective feature encoding," *Beyond MIDI*, 1997.

[3] J. Calvo-Zaragoza, J. Hajič Jr., and A. Pacha, "Understanding optical music recognition," *ACM Computing Surveys*, vol. 53, no. 4, Jul. 2020.

[4] J. A. Burgoyne, L. Pugin, G. Eustace, and I. Fujinaga, “A comparative survey of image binarisation algorithms for optical recognition on degraded musical sources,” in *Proc. of the 8th Int. Conf. on Music Information Retrieval*, pp. 509–512.

[5] J. dos Santos Cardoso, A. Capela, A. Rebelo, C. Guedes, and J. P. da Costa, “Staff detection with stable paths,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 6, pp. 1134–1139, 2009.

[6] A. Pacha and H. Eidenberger, “Towards a universal music symbol classifier,” in *Proc. of 14th Int. Conf. on Document Analysis and Recognition*, 2017, pp. 35–36.

[7] A. Fornés, J. Lladós, and G. Sánchez, “Old handwritten musical symbol classification by a dynamic time warping based method,” in *Graphics Recognition. Recent Advances and New Opportunities*, W. Liu, J. Lladós, and J.-M. Ogier, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 51–60.

[8] F. Rossant and I. Bloch, “Robust and adaptive OMR system including fuzzy modeling, fusion of musical rules, and possible error detection,” *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, p. 081541, 2006.

[9] A. Pacha, J. Hajič, and J. Calvo-Zaragoza, “A baseline for general music object detection with deep learning,” *Applied Sciences*, vol. 8, no. 9, p. 1488, 2018.

[10] J. Calvo-Zaragoza, A. H. Toselli, and E. Vidal, “Handwritten music recognition for mensural notation with convolutional recurrent neural networks,” *Pattern Recognition Letters*, vol. 128, pp. 115–121, 2019.

[11] A. Baró, C. Badal, and A. Fornés, “Handwritten historical music recognition by sequence-to-sequence with attention mechanism,” in *Proc. of the 17th Int. Conf. on Frontiers in Handwriting Recognition*. IEEE, 2020, pp. 205–210.

[12] F. J. Castellanos, J. Calvo-Zaragoza, and J. M. Iñesta, “A neural approach for full-page optical music recognition of mensural documents,” in *Proc. of the 21th Int. Society for Music Information Retrieval Conference*, 2020, pp. 12–16.

[13] M. Kletz and A. Pacha, “Detecting staves and measures in music scores with deep learning,” in *Proc. of the 3rd Int. Workshop on Reading Music Systems*, J. Calvo-Zaragoza and A. Pacha, Eds., Alicante, Spain, 2021, pp. 8–12.

[14] M. Alfaro-Contreras, D. Rizo, J. M. Iñesta, and J. Calvo-Zaragoza, “OMR-assisted transcription: A case study with early prints,” in *Proc. of the 22nd Int. Society for Music Information Retrieval Conference*, 2021, pp. 35–41.

[15] A. Ríos-Vila, D. Rizo, and J. Calvo-Zaragoza, “Complete optical music recognition via agnostic transcription and machine translation,” in *Proc. of the 16th Int. Conf. on Document Analysis and Recognition*, J. Lladós, D. Lopresti, and S. Uchida, Eds., 2021, pp. 661–675.

[16] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *Proc. of the 23rd Int. Conf. on Machine Learning*, 2006, p. 369–376.

[17] L. Tuggener, I. Elezi, J. Schmidhuber, M. Pelillo, and T. Stadelmann, “DeepScores-A Dataset for Segmentation, Detection and Classification of Tiny Objects,” in *Proc. of the 24th Int. Conf. on Pattern Recognition*, 2018, pp. 3704–3709.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.

[19] D. Coquenet, C. Chatelain, and T. Paquet, “Span: A simple predict & align network for handwritten paragraph recognition,” in *Proc. of the 16th Int. Conf. on Document Analysis and Recognition*, 2021, pp. 70–84.

[20] E. Parada-Cabaleiro, A. Batliner, and B. W. Schuller, “A Diplomatic Edition of Il Lauro Secco: Ground Truth for OMR of White Mensural Notation,” in *Proc. of the 20th Int. Society for Music Information Retrieval Conference*, 2019, pp. 557–564.

[21] L. Pugin, R. Zitellini, and P. Roland, “Verovio: A library for engraving MEI music notation into SVG,” in *Proc. of the 15th Int. Society for Music Information Retrieval Conference*, 2014, pp. 107–112.

[22] L. Kang, P. Riba, M. Rusiñol, A. Fornés, and M. Villegas, “Pay attention to what you read: Non-recurrent handwritten text-line recognition,” *Pattern Recognition*, vol. 129, p. 108766, 2022.

[23] D. Coquenet, C. Chatelain, and T. Paquet, “End-to-end handwritten paragraph text recognition using a vertical attention network,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[24] S. S. Singh and S. Karayev, “Full page handwriting recognition via image to sequence extraction,” in *Proc. of the 16th Int. Conf. on Document Analysis and Recognition*, 2021, pp. 55–69.