TOWARDS UNIVERSAL OPTICAL MUSIC RECOGNITION: A CASE STUDY ON NOTATION TYPES

Juan C. Martinez-Sevilla[1] **David Rizo**[1,2] **Jorge Calvo-Zaragoza**[1]

[1] Pattern Recognition and Artificial Intelligence Group, University of Alicante, Spain
[2] Instituto Superior de Enseñanzas Artísticas de la Comunidad Valenciana, Spain
{jcmartinez.sevilla, drizo, jorge.calvo}@ua.es

ABSTRACT

Recent advances in Deep Learning have propelled the development of fields such as Optical Music Recognition (OMR), which is responsible for extracting the content from music score images. Despite progress in the field, existing literature scarcely addresses core issues like performance in real-world scenarios, user experience, maintainability of multiple pipelines, reusability of architectures and data, among others. These factors result in high costs for both users and developers of such systems. Furthermore, research has often been conducted under certain constraints, such as using a single musical texture or type of notation, which may not align with the end-user requirements of OMR systems. For the first time, our study involves a comprehensive and extensive experimental setup to explore new ideas towards the development of a *universal* OMR system—capable of transcribing all textures and notation types. Our investigation provides valuable insights into several aspects, such as the ability of a model to leverage knowledge from different domains despite significant differences in music notation types.

1. INTRODUCTION

Optical Music Recognition (OMR) is a field of research focused on converting written music documents into machine-readable formats, such as Humdrum `**kern`, MEI, or MusicXML [1–3]. This technology holds significant promise for digital musicology, libraries, and academia, facilitating the digitization of scores for further musical analysis, large-scale information retrieval, and making vast musical archives more accessible [4].

Historically, the development of OMR has evolved from relying on basic heuristic methods to a more dynamic application of Deep Learning (DL) techniques [5]. This shift brought new advances to the field, leading to substantial improvements in the accuracy of music score transcription [6–9]. However, despite these advances, OMR models still face significant challenges in generalization. The DL methodologies, while robust in specific contexts, often struggle to perform consistently across diverse data distributions [10]. This is particularly evident when dealing with a variety of music notations and textures, from ancient Neumic chants to modern polyphonic compositions. Most existing OMR works focus on a narrow range of music types (often just one), which limits their usability for more comprehensive archival tasks [11].

In response to these limitations, this paper proposes the conceptualization of a *universal* OMR system capable of processing *all* types of musical notations and textures. [1] The long-term objective is to develop a versatile technology that can adapt to any musical document, regardless of its historical period or stylistic characteristics.

This paper takes the first steps towards such a system by exploring a few alternatives to achieve this goal. In particular, we carry out a specific case study focused on diverse notation types, involving medieval square notation, Mensural notation, and Common Western Modern Notation (CWMN) corpora. We consider whether it is more feasible to develop separate OMR models for each notation or to create a single, all-encompassing model. This dichotomy has not been thoroughly studied before. Separate OMR models for each notation maximize accuracy by addressing specific characteristics, but require extensive resources and individual updates. Conversely, a single, all-encompassing model enhances scalability and maintenance efficiency, benefiting from shared knowledge across notations—a potential advantage in deep learning—although it may struggle with variability. Additionally, we include an intermediate case in which a part of the model is common and only one specialized module is created for each notation, thereby representing a trade-off between the previous pros and cons.

This paper is organized as follows: Section 2 offers background information on OMR. In Section 3, we outline our methodology for analyzing the question at hand and the different training scenarios to leverage the system's performance. Section 4 details the experimental setup, while

[1] In this work, we will focus on Western notations that share some fundamental characteristics, such as indicating duration with the shape of the music-notation symbols and pitch with their position over a set of staff lines. These also follow a left-to-right reading order.

Section 5 presents the work results and analysis. Finally, we conclude the paper in Section 6, along with potential avenues for future research.

2. RELATED WORK

Modern research in OMR using DL methodologies has led to several successful approaches [4,5,12]. Notably, one approach that stands out is the so-called "end-to-end" formulation. This approach provides a holistic method where images of music notation are directly inputted into the model, which then predicts their content. The end-to-end formulation represents the state of the art in related areas such as text or speech recognition and is now considered by several works in OMR [11, 13–15].

Some works have successfully addressed end-to-end OMR for monophonic staff images, likely because most ancient notations depict monophonic staves. Specific efforts are underway to address other textures such as homophonic scores [6], polyphonic music [7, 16], and vocal pieces [8, 17]. However, despite recent advances in the field, there is still no approach for building a *universal* OMR system capable of handling all this variability of music notation types and textures simultaneously.

The fundamental challenge lies in an unsolved problem in DL models: they perform well when there are regular statistics and abundant data to train on, allowing them to learn the regularities in the distribution properly [10]. This is not the case in the OMR problem, where rich labeled data is scarce and the graphical feature variability is extensive, making it a complex task.

Due to the inherent characteristics of DL methodologies, the existing literature work with analogous or highly similar train-test distributions [11]. Consequently, since there is a lack of research focusing on the development of *universal* OMR systems capable of processing any input score regardless of its content, we propose the first study aimed at developing, understanding, and evaluating a *universal* OMR system for dealing with different notation types simultaneously.

3. METHODOLOGY

Our objective is to explore an initial approach towards developing a *universal* OMR system. Specifically, we consider the case of accounting for different notation types. To achieve this, we consider three different scenarios: (i) a single model per dataset; (ii) a model leveraging all available data; and (iii) a hybrid model, for which some parts are common across all cases, but there are also specific layers tailored to each notation type. We opt for a deep end-to-end model as representative of the state-of-the-art in OMR. Below, we provide a detailed explanation of how this model works and then explain the different approaches selected to address the task.

3.1 Learning framework

The end-to-end OMR model seeks to directly retrieve the music notation from a single staff image. As in recent literature [11, 13, 14], we assume that a certain preprocessing stage has already separated the staves of the score [18].

Based on other works addressing the OMR challenge [13], a Convolutional Recurrent Neural Network (CRNN) scheme is considered for the end-to-end pipeline. The CRNN architecture incorporates an *encoder*: a block of convolutional layers that learns a set of features from the input image. Then, it includes a *decoder*: group of recurrent stages that model the temporal dependencies of the feature-learning block. Finally, a fully connected network with a softmax activation is used to retrieve a posteriogram, which is decoded to obtain the predicted *musical symbols* [2]. The Connectionist Temporal Classification (CTC) training procedure [19] is used to achieve an end-to-end scheme, as it allows training the network using unsegmented sequential data.

For training, let $\mathcal{T} \subset \mathcal{X} \times \Sigma^*$ be a set of data where sample $x_i \in \mathcal{X}$ of single staff image is related to symbol sequence $\mathbf{z}_i = \left(z_{i1}, z_{i2}, \ldots, z_{i|\mathbf{z}_i|}\right) \in \Sigma^*$, where Σ represents the symbol vocabulary used for encoding the music score. Note that the use of CTC to model the transcription task requires the inclusion of an additional "*blank*" symbol in the Σ vocabulary, i.e., $\Sigma' = \Sigma \cup \{blank\}$.

At prediction, for a given music staff image input $x_i \in \mathcal{X}$, the model outputs a posteriogram $p_i \in \mathbb{R}^{|\Sigma'| \times K}$, where K represents the number of frames provided by the recurrent stage. Finally, the predicted sequence $\hat{\mathbf{z}}_i$ is obtained resorting to a *greedy* policy that retrieves the most probable symbol per frame in p_i, later a subsequent mapping function merges consecutive repeated symbols and removes *blank* labels.

3.2 Approaches to OMR for different notation types

In order to explore diverse learning frameworks to assess the transcription performance, we pose three different scenarios that differ in how data is fed to the model and the training strategies for the model layers. An overview of these scenarios is described as follows (illustrated in Figure 1):

Only: In this scenario, one model is trained for each single dataset. This is the baseline of our experiments and it will allow us to compare properly the different approaches selected. It should be emphasized that this training scenario will employ a set of resources and time associated with each corpora. This methodology stands as the current state of the art, as recent research resorted to training individual models as described in Sec.2.

All: For this scenario, all available notation types in this work are merged to train a single model. As commented in the introduction, our long-term objective is to create a *universal* OMR capable of retrieving all types of notation and textures. This option allows us to explore the capabilities and drawbacks of integrating all possibilities in the same model.

[2] In this work, a *musical symbol* is represented as the conjunction of the glyph or shape and the position within the staff.

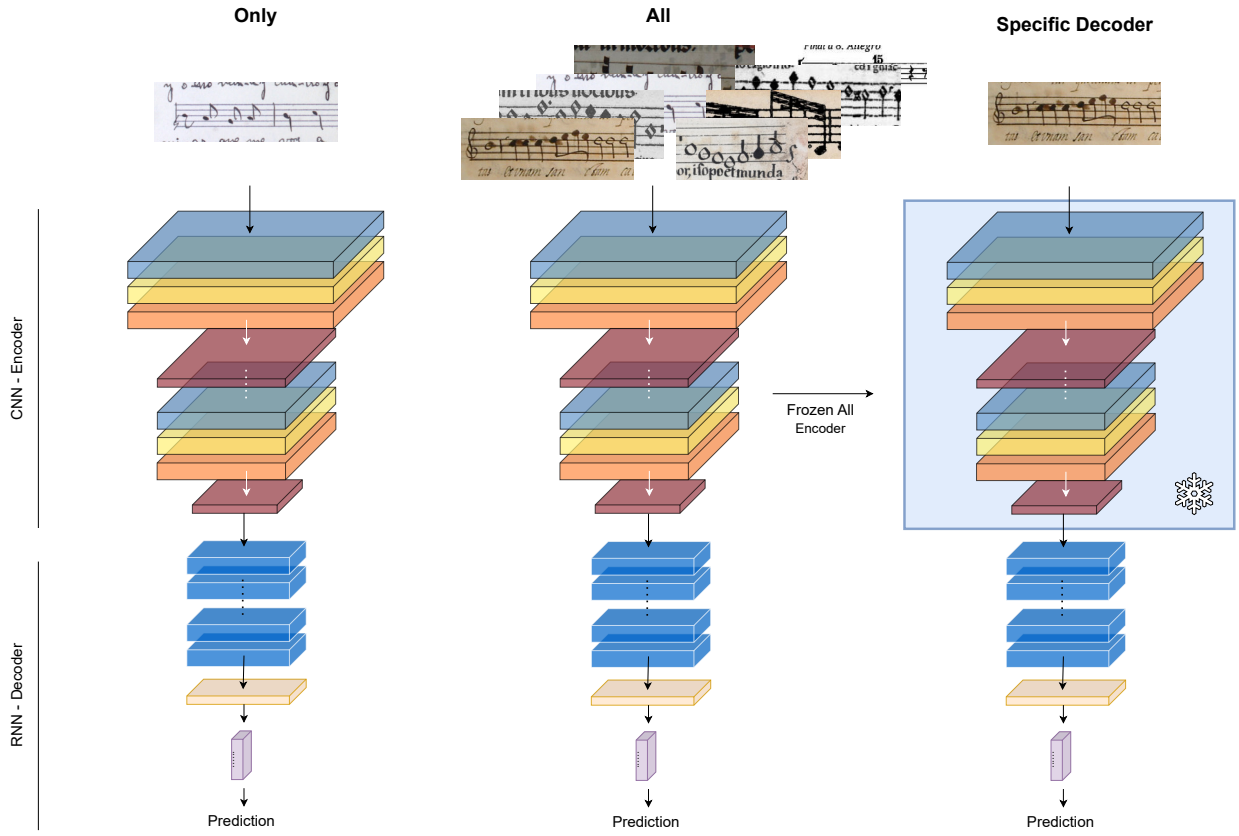


Figure 1: Graphical scheme of the three different approaches considered for this work using the CRNN architecture as the backbone. *Only*: a model trained per notation type individually. *All*: a model trained with all the notation types available for this work. *Specific Decoder*: once the *All* scenario is finished, the encoder (already trained) is frozen to train specific decoders for each notation type.

Specific Decoder: Recent DL approaches pictured the adequacy of learning via a general feature extractor (encoder) [20]. Similarly, we leverage the encoder block of the *All* approach weights as our starting point. By doing so, we establish an already-evaluated feature extractor shared across all corpora. Having the features extracted, we then fine-tune a notation-specific decoder block based on the unique underlying musical context.

The selection of these scenarios helps to study performance but also other important aspects such as maintainability, reusability, or resource leveraging, which are valuable for real-case systems and have been barely analyzed in OMR literature.

4. EXPERIMENTAL SETUP

According to the choices made for the experimental road map, we first introduce the studied evaluation metrics. Later we give further details about the learning model hyperparameters selected and the training techniques used. Eventually, we describe the data collections used for train and evaluation.

4.1 Evaluation

Current OMR systems are designed to serve as a tool. Bearing this in mind, it should be more than interesting to compute the amount of effort it would take a user to correct the errors made by the system. However, there is not a clear way of properly measuring this case. This is why when evaluating an OMR system we resort to the *Symbol Error Rate* (SER). Given a prediction $\hat{\mathbf{z}}_i$ and the ground truth musical symbol sequence $\mathbf{z}_i$, SER is calculated as the average number of elementary editing operations (insertions, deletions, or substitutions) required to convert prediction $\hat{\mathbf{z}}_i$ into reference $\mathbf{z}_i$, normalized by the length of the latter. Formally, this is expressed as:

$$\text{SER}\ (\%) = \frac{\sum_{i=1}^{|\mathcal{S}|} \text{ED}\left(\hat{\mathbf{z}}_i,\ \mathbf{z}_i\right)}{\sum_{i=1}^{|\mathcal{S}|} |\mathbf{z}_i|} \tag{1}$$

where $\mathcal{S}$ is a set of test data, $\text{ED} : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{N}_0$ denotes the string edit distance, and $\hat{\mathbf{z}}_i$ and $\mathbf{z}_i$ respectively represent the estimated and target sequences.

4.2 Neural model configuration

The CRNN hyperparameters used in this study are based on the ones used in previous works [11, 13]. Authors adopt a 4 Convolutional layer block with batch normalization 2D, Leaky ReLu activation, and max-pooling 2D downsampling. Feature maps extracted from the encoder, i.e. the Convolutional Neural Network (CNN) block, are introduced into 2 Bidirectional Long Short-Time Memory

(BLSTM) layers with 256 hidden units each and a dropout value of $d = 50\%$ followed by a fully connected network with $|\Sigma'|$ units. The architecture described results in a model with 5.3M parameters.

All the models were trained with a batch size of 16 samples—it is important to mention that given the different sizes of the datasets, all the generated batches had the same proportion of samples from each dataset so the network did not adjust to the bigger dataset, i.e., dataset interleaving. The ADAM [21] optimizer was considered, a fixed learning rate of 10^{-3}, and weight decay of 10^{-6}. We iterate over 200 epochs using image augmentation techniques (blur, rotation, contrast, erosion, brightness, etc.), ensuring the robustness of the model, keeping the weights of the model that minimize the SER evaluation metric in the validation partition. The early stopping technique is used with a patience of 20 epochs. Lastly, all experiments were run using the Python language (v3.10.13) with the PyTorch and PyTorch Lightning frameworks on a single NVIDIA GeForce RTX 4090 card with 24GB of GPU memory.

4.3 Datasets

As introduced in Sec.1, music manuscripts depict a great challenge for transcription methods. Their variety in content and appearance poses a still unsolved question. In order to study the adequacy of a *universal* OMR system, we gathered data sources taking into account their variability in terms of notation, graphical appearance, and musical context, aiming to reflect Western musical diversity. A set of 40 different works has been collected that have been grouped to simplify the experimentation and insights reported. Among them, we find square notation, white Mensural notation, and CWMN. A brief description of some dataset features and staves can be found in Table 1 and Fig. 2.

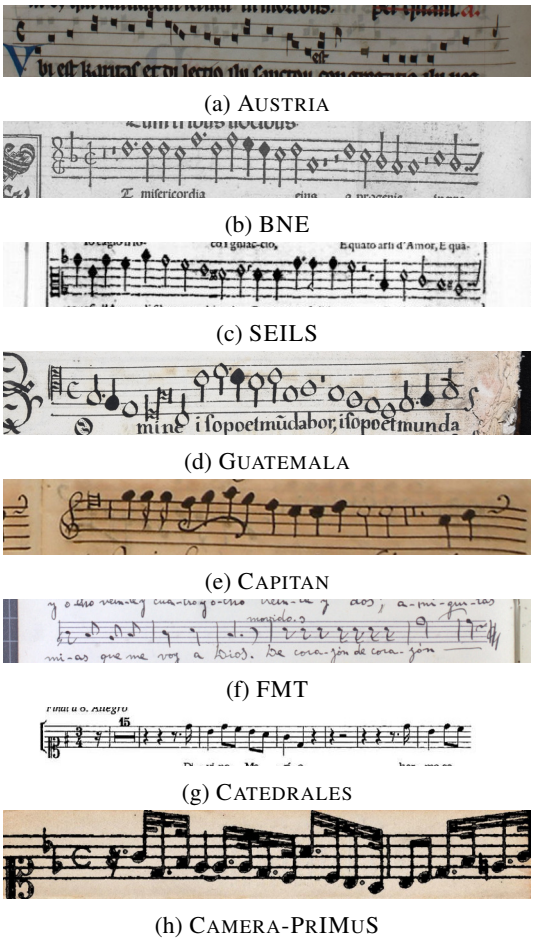
Table 1: Dataset descriptions in terms of notation type, pages, music fragments (staves), and vocabulary sizes.

Notation type	Dataset	Number of pages	Music fragments	Vocabulary size
Square	AUSTRIA	685	4 850	270
Mensural	BNE	4 125	27 746	709
	SEILS	151	1 136	206
	GUATEMALA	385	3 263	316
	CAPITAN	97	828	373
CWMN	FMT	348	1 305	425
	CATEDRALES	52	308	245
	CAMERA-PRIMUS	–	15 000	1 443

Diverse cases have been considered looking for different printers, copyists, authors, and periods considering the more variability the better. The list of datasets used is classified by notation type and ordered temporally below.

4.3.1 *Square Notation*

Square notation is written on a staff with four lines and three spaces. In this notation, ascending notes are shown as stacked squares, while descending notes are written with diamonds. This system of notation appears in liturgical chant books.



(a) AUSTRIA

(b) BNE

(c) SEILS

(d) GUATEMALA

(e) CAPITAN

(f) FMT

(g) CATEDRALES

(h) CAMERA-PRIMUS

Figure 2: Samples of staves of the different datasets employed in the experimentation.

AUSTRIA. The Austria dataset contains 685 printed pages of 15th-century manuscripts in German Gothic square notation. Provided by the Austrian Centre for Digital Humanities and Cultural Heritage. [3]

4.3.2 *White Mensural Notation*

Notation system used in polyphonic European vocal music. Mensural notation can use different note shapes to denote rhythmic durations. It is written on a staff with five lines and four spaces.

BNE. The "Biblioteca Nacional de España (BNE)" dataset corresponds to the pages from the corpus obtained from the collection of mensural books of the Biblioteca Digital Hispánica. [4] It comprises multiple authors and printers, e.g., F. Guerrero, H. of G. Scoto or Antonio Gardano, with a size of 4 125 pages. License: public.

SEILS. The "Second Edition of the Il Lauro Secco (SEILS)" dataset consists of 151 printed pages of the "Il

[3] https://www.oeaw.ac.at/ (accessed April 8th, 2024)

[4] https://www.bne.es/es/catalogos/biblioteca-digital-hispanica (accessed April 8th, 2024)

Lauro Secco" collection corresponding to an anthology of 16th-century Italian madrigals in white Mensural notation [22]. License: public.

GUATEMALA. The Guatemala dataset incorporates 383 handwritten pages from a polyphonic choir book, part of a larger collection held at the "Archivo Histórico Arquidiocesano de Guatemala" [23]. License: private.

CAPITAN. The Capitan dataset contains 100 handwritten pages of 17th-century manuscripts in late white Mensural notation extracted from collections found in the "Catedral del Pilar" in Zaragoza [24]. License: private.

4.3.3 Common Western Modern Notation

Current notation system, written in five lines and four spaces. It is capable of indicating to the musician all the parameters to properly interpret the piece, such as dynamics or tempo changes. [5]

FMT. This collection consists of four groups of handwritten score sheets of popular Spanish songs transcribed by musicologists between 1944 and 1960. taken from the "Fondo de Música Tradicional IMF-CSIC" [6], with a total of 348 images. License: public.

CATEDRALES. The Catedrales dataset contains 52 pages of printed liturgical examples from Málaga, Granada, and Sevilla cathedral archives [25]. License: public.

CAMERA-PRIMUS. The Printed Images of Music Staves (PrIMuS) dataset is a hybrid corpus, i.e., the musical content comprehends the RISM Database [7] but the images have been obtained using the digital engraver tool Verovio [26]. To the generated images multiple distortions and textures are applied to simulate the look and conditions of the real sources. Although the original dataset consists of almost 100 000 samples, we have randomly selected 15 000 to make it more suitable for our experimentation [27]. License: public.

All the datasets presented use an agnostic output encoding which represents a musical symbol as `glyph:position_in_staff`. This encoding helps transcribe the tokens given their graphical appearance rather than their musical meaning, which can be ambiguous in many situations for the model to learn, making it unsuitable for OMR. Additionally, the agnostic encoding facilitates a straightforward conversion to standard formats such as MusicXML, MEI, or Humdrum `**kern` [28].

5. RESULTS

Table 2 presents the test results obtained with the proposed experimental scheme in terms of the SER (%) metric.

The *Only* scenario acts as our baseline. Here training, validation, and testing splits comprise exclusively samples from each individual dataset. We observe varying performances across different datasets. Notably, the SER metric ranges from 1.54% for the CAMERA-PRIMUS dataset to 17.34% for the CATEDRALES dataset. This indicates significant variability in model performance depending on the dataset size, notation, and graphical features, being the higher values the ones associated with CWMN, where we find more complex musical symbols and context.

Table 2: Results in terms of the SER(%) metric for the training scenarios *Only*, *All* and *Specific Decoder*.

Dataset	Only (baseline)	All	Specific Decoder
AUSTRIA	**3.77**	3.87	3.78
BNE	**3.25**	3.67	3.31
SEILS	2.71	**1.88**	1.94
GUATEMALA	2.22	**1.87**	1.88
CAPITAN	8.60	**6.80**	7.91
FMT	8.98	**5.72**	7.11
CATEDRALES	17.34	**8.49**	17.94
CAMERA-PRIMUS	**1.54**	3.07	1.60

When training on the *All* scenario, the model demonstrates performance improvements compared to the *Only* scenario for most datasets. This proves the validity of unifying training pipelines for different notations as the model learns to extract more robust features from the images, which helps in datasets with fewer samples while sacrificing very little accuracy in other datasets, e.g., BNE or CAMERA-PRIMUS. It is worth highlighting the great improvement in the CATEDRALES dataset reducing the SER from 17.34% to 8.49%. On the other hand, we lose accuracy in datasets such as AUSTRIA (from 3.77% to 3.87%), BNE (from 3.25% to 3.67%), and CAMERA-PRIMUS (from 1.54% to 3.07%). This situation reports valuable insights given that on bigger datasets like BNE or CAMERA-PRIMUS with enough data to be trained individually we lose performance, but if we are willing to sacrifice that performance we improve in several datasets. AUSTRIA poses a different situation, due to being the only square notation dataset, the labeling is slightly different to the other corpora increasing the SER metric when merging it with the other datasets. [8]

After training on the *All* scenario, experiment outcomes show the adequacy of merging different datasets to better learn the data features. Thus, in the *Specific Decoder* scenario, the *All* encoder, or CNN, is frozen, and specific decoders were trained for each dataset individually. This approach is aimed at capturing dataset-specific features and learning the underlying musical language of each dataset.

While some datasets exhibited improved performance

[5] For evaluation, pitch, rhythm and articulation are considered.

[6] `https://musicatradicional.eu/es/home` (accessed April 8th, 2024)

[7] `https://rism.info/` (accessed April 8th, 2024).

[8] For the *All* scenario we have checked that the tokens predicted are present in the target vocabulary. Without performance variation, such a fact evinces the adequacy of using all data available to train a unique model, to better learn the image features and the inherent difficulty of music when applying OMR without focusing on a specific given dataset.

(e.g., SEILS with a 1.94% SER in the *Specific Decoder* compared to 2.71% SER in the *Only* setup), others experienced only marginal improvements or even a slight degradation in performance (BNE, GUATEMALA, CAPITAN, FMT, CAMERA-PRIMUS). Since this approach could be discarded at first glance for not being the best performing, we make an in-depth explanation of the results obtained in the latter scenario in Sec. 5.1.

5.1 Time-efficient model training

Another important factor to take into account when looking at the experiment results is the time consumption, which is a key factor to better understand the outcomes of this research. Given the datasets presented in this work, we employ a total of 54 436 monophonic staff images with different notation types and graphical features. In Fig. 3, we report the runtime of the experiments presented. When using the training scenario specified as *All* and the configuration explained, the time that took to train the model was 1D 20H 36M 49S. If we evaluate the performance obtained in the *All* scenario we could think that these are the best approaches, as the SER metric poses improvements even in datasets with few samples. However, in real scenarios, this approach would have to be retrained from scratch in case we want to integrate a new dataset [9]. That is why the *Specific Decoder* scenario—where a common CNN is trained and specific decoders, i.e, BLSTMs, are created for each dataset—emerges, given that once the encoder block (CNN) is trained the average time to integrate a new dataset, i.e., train its decoder block, is 1H 6M 11S. This time-efficient model training approach attends more accurately to the end-user requirements in conjunction with better resource management.

This analysis strengthens our proposal of building a *universal* OMR system, that leverages all the existent musical data and is capable of transcribing multiple notation types. In these experiments, we explore the end-to-end architecture for every notation type, which clearly helps as explained in Sec.5. This will allow creating a robust, maintainable, reusable system as a first step never done before towards *universal* OMR.

6. CONCLUSIONS

This work stands out as the first to introduce the *universal* OMR goal, which involves the design, construction, and evaluation of a system capable of retrieving musical content from a document, taking into account different notation types and textures, such as monophonic, homophonic, vocal, polyphonic, etc., and the end-user requirements in real-case scenarios. To achieve this, we studied and compared different settings of real and heterogeneous data corpora to provide invaluable insights into these first steps towards *universal* OMR.

The obtained results validate the capabilities of current OMR state-of-the-art model architectures to transcribe real documents with different notation types, as the SER(%) rates match those observed in works that exclusively address one notation (either square, Mensural, or CWMN). Moreover, the use of a frozen trained encoder block as a common feature extractor proves to be useful for saving resources, maintaining the system, and reducing training time, since in some cases it considerably improves the overall transcription performance when there are not enough samples.

Future work seeks to expand the presented assortment by considering other textures such as homophony, vocal, or polyphony, to provide further insights and analysis towards *universal* transcription pipelines. Fine-tuning all or certain layers of the encoder would also be relevant, given that differences among datasets manifest in their visual representation rather than in their output. Furthermore, given the results obtained, another promising avenue is to investigate adequate encoding formats to properly represent music from different centuries and textures.

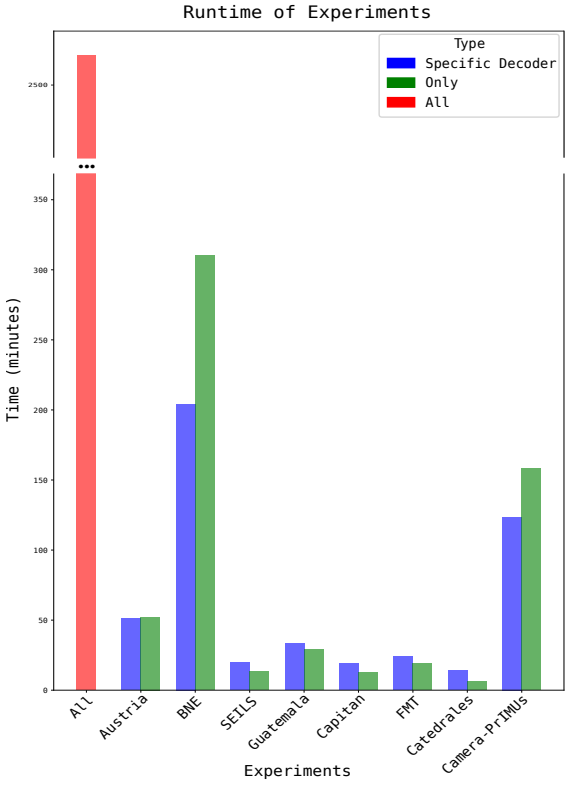


Figure 3: Runtime of experiments presented in this work in minutes for the *All*, *Only* (baseline) and *Specific Decoder* scenarios.

7. ACKNOWLEDGEMENTS

This paper is supported by grant CISEJI/2023/9 from "Programa para el apoyo a personas investigadoras con talento (Plan GenT) de la Generalitat Valenciana".

8. REFERENCES

[1] D. Huron, "Humdrum and Kern: Selective Feature Encoding BT - Beyond MIDI: The handbook of musi-

[9] Except if we use Continual Learning techniques [29], yet to be explored in OMR.

cal codes," in *Beyond MIDI: The handbook of musical codes*. Cambridge, MA, USA: MIT Press, jan 1997, pp. 375–401.

[2] A. Hankinson, P. Roland, and I. Fujinaga, "The music encoding initiative as a document-encoding framework," in *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011*. University of Miami, 2011, pp. 293–298.

[3] M. Good *et al.*, "Musicxml: An internet-friendly format for sheet music," in *Xml conference and expo*. Citeseer, 2001, pp. 03–04.

[4] M. Alfaro-Contreras, D. Rizo, J. M. Iñesta, and J. Calvo-Zaragoza, "OMR-assisted transcription: a case study with early prints," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference*. Online: ISMIR, Nov. 2021, pp. 35–41.

[5] J. Calvo-Zaragoza, J. Hajic Jr, and A. Pacha, "Understanding optical music recognition," *ACM Computing Surveys (CSUR)*, vol. 53, no. 4, pp. 1–35, 2020.

[6] M. Alfaro-Contreras, J. M. Iñesta, and J. Calvo-Zaragoza, "Optical music recognition for homophonic scores with neural networks and synthetic music generation," *Int. J. Multim. Inf. Retr.*, vol. 12, no. 1, p. 12, 2023.

[7] J. Mayer, M. Straka, J. H. Jr., and P. Pecina, "Practical end-to-end optical music recognition for pianoform music," *CoRR*, vol. abs/2403.13763, 2024.

[8] M. Villarreal and J. A. Sánchez, "Synchronous recognition of music images using coupled n-gram models," in *Proceedings of the ACM Symposium on Document Engineering 2023*, 2023, pp. 1–9.

[9] A. Ríos-Vila, J. Calvo-Zaragoza, D. Rizo, and T. Paquet, "Sheet music transformer ++: End-to-end full-page optical music recognition for pianoform sheet music," *CoRR*, vol. abs/2405.12105, 2024.

[10] Y. Bengio, Y. Lecun, and G. Hinton, "Deep learning for AI," *Communications of the ACM*, vol. 64, no. 7, pp. 58–65, 2021.

[11] J. C. Martinez-Sevilla, A. Rosello, D. Rizo, and J. Calvo-Zaragoza, "On the performance of optical music recognition in the absence of specific training data," in *Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR 2023, Milan, Italy, November 5-9, 2023*, 2023, pp. 319–326.

[12] L. Tuggener, R. Emberger, A. Ghosh, P. Sager, Y. P. Satyawan, J. Montoya, S. Goldschagg, F. Seibold, U. Gut, P. Ackermann *et al.*, "Real world music object recognition," *Transactions of the International Society for Music Information Retrieval*, vol. 7, no. 1, pp. 1–14, 2024.

[13] 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.

[14] P. Torras, A. Baró, L. Kang, and A. Fornés, "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*. Online: ISMIR, Nov. 2021.

[15] Y. Li, H. Liu, Q. Jin, M. Cai, and P. Li, "Tromr: Transformer-based polyphonic optical music recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[16] A. Ríos-Vila, D. Rizo, J. M. Iñesta, and J. Calvo-Zaragoza, "End-to-end optical music recognition for pianoform sheet music," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 26, no. 3, p. 347–362, 2023.

[17] J. C. Martinez-Sevilla, A. Rios-Vila, F. J. Castellanos, and J. Calvo-Zaragoza, "A holistic approach for aligned music and lyrics transcription," in *International Conference on Document Analysis and Recognition*. Springer, 2023, pp. 185–201.

[18] A. Pacha, "Incremental supervised staff detection," in *Proceedings of the 2nd international workshop on reading music systems*, 2019, pp. 16–20.

[19] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "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, June 25-29, 2006*, 2006, pp. 369–376.

[20] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind one embedding space to bind them all," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023, pp. 15 180–15 190.

[21] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *3rd Int. Conf. on Learning Representations*, Y. Bengio and Y. LeCun, Eds., San Diego, USA, 2015.

[22] 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 *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, 2019, pp. 557–564.

[23] M. E. Thomae, J. E. Cumming, and I. Fujinaga, "Digitization of choirbooks in guatemala," in *Proceedings of the 9th International Conference on Digital Libraries for Musicology*, ser. DLfM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 19–26.

[24] J. Calvo-Zaragoza, D. Rizo, and J. M. I. Quereda, "Two (note) heads are better than one: Pen-based multimodal interaction with music scores," in *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016, New York City, United States, August 7-11, 2016*, 2016, pp. 509–514.

[25] A. Madueño, A. Rios-Vila, and D. Rizo, "Automatized incipit encoding at the andalusian music documentation center," in *Proceedings of the 9th International Conference on Digital Libraries for Musicology*, ser. DLfM '21, 2021.

[26] L. Pugin, R. Zitellini, and P. Roland, "Verovio: A library for engraving MEI music notation into SVG," in *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014*, 2014, pp. 107–112.

[27] J. Calvo-Zaragoza and D. Rizo, "Camera-primus: Neural end-to-end optical music recognition on realistic monophonic scores," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, 2018, pp. 248–255.

[28] A. Ríos-Vila, M. Esplà-Gomis, D. Rizo, P. J. Ponce de León, and J. M. Iñesta, "Applying automatic translation for optical music recognition's encoding step," *Applied Sciences*, vol. 11, no. 9, 2021.

[29] A. Awasthi and S. Sarawagi, "Continual learning with neural networks: A review," in *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, 2019, pp. 362–365.