






Complete Optical Music Recognition via Agnostic Transcription and Machine Translation

Antonio Ríos-Vila¹ , David Rizo^{1,2} , and Jorge Calvo-Zaragoza¹ 

¹ U.I. for Computing Research, University of Alicante, Alicante, Spain
{arios,drizo,jcalvo}@dlsi.ua.es

² Instituto Superior de Enseñanzas Artísticas de la Comunidad Valenciana
(ISEA.CV), Valencia, Spain

Abstract. Optical Music Recognition workflows currently involve several steps to retrieve information from music documents, focusing on image analysis and symbol recognition. However, despite many efforts, there is little research on how to bring these recognition results to practice, as there is still one step that has not been properly discussed: the encoding into standard music formats and its integration within OMR workflows to produce practical results that end-users could benefit from. In this paper, we approach this topic and propose options for completing OMR, eventually exporting the score image into a standard digital format. Specifically, we discuss the possibility of attaching Machine Translation systems to the recognition pipeline to perform the encoding step. After discussing the most appropriate systems for the process and proposing two options for the translation, we evaluate its performance in contrast to a direct-encoding pipeline. Our results confirm that the proposed addition to the pipeline establishes itself as a feasible and interesting solution for complete OMR processes, especially when limited training data is available, which represents a common scenario in music heritage.

Keywords: Optical Music Recognition · Graphics recognition · Machine translation

1 Introduction

Music represents a valuable component of our cultural heritage. Most music has been preserved in printed or handwritten music notation documents. In addition to the efforts to correctly maintain the documents that inherently get damaged over time, huge efforts are being done to digitize them. The same way Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) are successfully being applied to extract the content from text images, Optical Music Recognition (OMR) systems are applied to encode the music content that appears in score sheets [6].

Specifically, OMR joins the knowledge from areas like computer vision, machine learning and digital musicology to perform the recognition and the digitising of music scores. Despite of being sometimes addressed as “OCR for music” [3], its two-dimensional nature, along with many ubiquitous contextual dependencies among symbols, differentiates OMR from other document recognition areas, such as OCR and HTR [5]. To illustrate this, we could use a simple example: A note is identified, graphically speaking, at a specific position in the staff. However, its interpretation could change depending on multiple factors, such as the clef position, the accidentals that may be present nearby, the key signature, or just a bar line cancelling previous alterations. Indeed, there is also a required temporal interpretation that does not depend on that specific position in the staff, but on the shape of the note (as each glyph represents a different duration of the note during interpretation).

Most of the existing OMR literature is framed within a multi-stage workflow, with steps involving image pre-processing—including staff-line detection and removal [10]—symbol classification [20] and notation assembly [19]. Advances in Deep Learning (DL) lead the image processing steps to be replaced with neural approaches such as Convolutional Neural Networks (CNN). But more importantly, DL brought alternatives to these traditional multi-stage workflows. On the one hand, we have the segmentation-based approach, where the complex multi-stage symbol isolation workflows have been replaced for region-based CNN that directly recognize symbols in a music staff [21, 29]. On the other hand, there are end-to-end approaches. Specifically, we find solutions based on Convolutional Recurrent Neural Networks (CRNN) that come in varying configurations: the so-called Sequence to Sequence (*Seq2Seq*) architecture [2] ones, and also those trained using the Connectionist Temporal Classification (CTC) loss function [8].

Typically, these DL-based approaches cover all the processes that involve the transcription of an input image, which is usually a music staff, into a sequence that represents the recognized glyphs and positions of the symbols in the given score. Even obtaining such descriptive sequences, these results cannot be used by an end-user or reproduced in a music tool or visualizer, as there exists the need to recover music semantics as well. This last step to achieve an actual digital score is the so-called encoding process, where the graphical recognition outputs, without specific musical meaning, are converted into a standard semantic encoding. Unfortunately, this step is hardly addressed in the DL-based OMR literature, due to the focus the community has given to the challenges of the previous steps require.

In this paper, we research how to complete the OMR process, starting from a music-staff image as an input and producing a semantic standard encoding sequence as output. As a novelty, we introduce the use of Machine Translation to perform this last step of parsing a purely visual representation extracted from a graphic recognition process and exporting it in a standard musical encoding document.

The rest of the paper is organized as follows: in Sect. 2 we discuss why we approach the OMR encoding step with Machine Translation techniques, instead of

hand-crafted rule-based heuristic approaches. In Sect. 3, we describe the specific implemented systems used to perform our experiments. In Sect. 4, we explain our experimentation environment for the sake of replicability; in Sect. 5 we present and discuss the obtained results regarding the comparison between different alternatives; and, we conclude our work in Sect. 6.

2 Machine Translation for Encoding in Optical Music Recognition

We discuss in this section how to approach the encoding step of an OMR system, as it is an issue that has not been fully solved in previous works. We remind the reader that the encoding step consists of the production of a symbolic music notation format from the symbol recognition phase in the previous OMR step, which typically works at the image level. This means eventually producing a score encoded in a standard digital format from a collection of musical glyphs located in a two-dimensional space. From now on, we will denote the output from the graphical process as *agnostic encoding*; while the music standard format is referred to as *semantic encoding*. These terms are becoming common in OMR literature [7, 22].

A usual approach in most commercial systems to convert from agnostic encoding to semantic encoding is laying the task on a rule-based translation system. This has been proved to be a challenging task in complex scores [5, 13]. This approach also has significant issues in both extrapolability for different notation types, and scalability terms, as it requires the redesign of the rules or systems when the source and/or the target encoding vary. This scalability issue also appears when moving into more complex music domains, such as polyphonic scores, where the task of designing rules which adapt to these documents may become unfeasible. As we can observe, this is hardly maintainable when complexity both on the document type and the notations scale. This situation leaves us to look for more sophisticated models in the Machine Translation community.

One simple approach could be to apply Statistical Machine Translation (SMT) techniques [16], which are data-driven approaches for Machine Translation where several independent models are combined to produce both a translation unit and a language model to convert a source language sequence into a target language one. These combinations allow SMT to provide balanced predictions in accuracy and fluency, as they implement mechanisms to deal with translation issues such as word reordering. Another benefit they bring is that, currently, there exist standard and well-known toolkits to perform SMT, such as Moses [17]. However, during preliminary experimentation with these techniques [26], we observed that, despite offering interesting results with few labeled data, they do not produce flexible models where the input can have structural errors. This is a significant drawback in our case, as we cannot expect the graphical recognition step of the OMR pipeline to be completely accurate. In addition to this issue, SMT techniques also require an additional feature engineering process for both the source and the target languages, as we are dealing with data-driven

models which might not get their best results by just inputting raw sequences. This preprocessing requirement implies an additional workload that may become unfeasible if the considered encodings got extended.

For all the above, we decided to implement Neural Machine Translation (NMT). As other knowledge areas, the Machine Translation community has also moved towards DL techniques to perform automatic translation between languages. These neural models typically need more training data than SMT. However, they produce models that have proven to be more robust in the musical context [26], one aspect that we discuss further below. Another benefit of integrating these systems into the OMR pipeline is that they share technological features with the previously performed steps, so it is possible to easily produce a complete system that includes both the recognition steps and the translation one, which can be packaged to be served in practical applications. Therefore, given the advantages that this approach offers, we propose to tackle the encoding step via Machine Translation techniques, specifically with NMT.

2.1 Target Encoding Format

One relevant goal of this work is to showcase a suitable music notation format to be used as the target semantic encoding of the NMT process. We have analyzed which format is more beneficial in terms of exportability and later compatibility with musicology tools, which are the target destinations of our pipeline outputs.

The first options that may be considered are the most extended semantic encodings in music information retrieval and musicology contexts: MEI [14] and MusicXML [11], which represent music score components and metadata in XML languages. These can be understood as analogous markup-based encoding languages such as TEI [4] in the text recognition context. Despite being comprehensive formats, these semantic representations have a significant drawback in a Machine Translation context: they are highly verbose. This means that the target language would require a huge number of tokens for even small music excerpts, thereby making the automatic encoding task unnecessarily complicated.

Previous works on this area have proposed alternatives to tackle this issue [25], such as using *Humdrum **kern* [15]. This is a robust and widely-used semantic encoding for many musicological projects. Its benefits for our purpose lie in a simple vocabulary, a sequential-based format, and in its compatibility with dedicated musicology software like Verovio Humdrum Viewer [28], which brings the possibility of automatically converting into other formats.

For all the above, we selected ***kern* as our target semantic encoding language. An example of the convenience of this format over other XML-based ones, like MEI, is shown in Fig. 1.

3 Methodology

We define a complete-pipeline OMR as a process that eventually exports the written notation in a standard digital format. Our methodology assumes an



(a) Example music excerpt

```

... <music> ↵ <body> ↵ <mdiv> ↵ <score> ↵
<scoreDef xml:id="scoredef-0000000430793170" key.sig="2s" meter.sym="common">
<scoreDef xml:id="scoredef-0000000430793170" key.sig="2s" meter.sym="common">
<staffGrp xml:id="staffgrp-0000000321535565">
<staffGrp xml:id="staffgrp-0000000321535565">
<staffDef xml:id="staffdef-0000000979385103" clef.shape="G"
  clef.line="2" n="1" lines="5" />
</staffGrp> ↵ </scoreDef> ↵ <section xml:id="section-0000002102168953"> ↵
<measure xml:id="measure-0000000817881159" right="single">
<staff xml:id="staff-0000000752632627" n="1">
<layer xml:id="layer-0000001525105800" n="1">
<note xml:id="note-0000000088370008" dur="2" oct="5" pname="d" tie="i" />
<beam xml:id="beam-0000000838622227">
<note xml:id="note-0000001323524379" dur="16" oct="5" pname="d" tie="t" />
<note xml:id="note-0000000788593928" dur="16" oct="5" pname="e" />
<note xml:id="note-0000001776562259" dur="16" oct="5" pname="d" />
<note xml:id="note-0000000069259125" dur="16" oct="5" pname="c" />
</beam> ...

```

(b) MEI representation of the music excerpt ('↵' represents the end-of-line character.)

```
**skern ↵ *clefG2 ↵ *k[f#c#] ↵ *met(C) ↵J2dd[ ↵ 16dd] ↵ 16ee ↵ 16dd ↵ 16cc#
```

(c) **kern representation of the music excerpt ('↵' represents the end-of-line character.)

Fig. 1. Example of MEI and **kern representations of a simple music excerpt, showing the different verbosity between formats.

initial pre-process to divide a full-page score into a sequence of staves, much in the same way as HTR typically assumes a previous text-line segmentation [27]. This is not a strong assumption as there exist specific layout analysis algorithms for OMR that decompose the image into staves [24].

Our OMR pipeline is divided here into a two-step procedure that first recovers the graphical information and then performs a proper encoding. Formally, let \mathcal{X} be the input image space of single-staff sections, and Σ_a and Σ_s be denoted as the alphabet of *agnostic* symbols and the alphabet of *semantic* symbols, respectively. Then, the OMR system becomes a *graphical recognition* $f_g : \mathcal{X} \rightarrow \Sigma_a$ followed by a *translation process* $f_t : \Sigma_a \rightarrow \Sigma_s$. An overview of the methodology is illustrated in Fig. 2.

Additionally, a *direct encoding* approach $f_d : \mathcal{X} \rightarrow \Sigma_s$ will be proposed as a baseline for our experimentation, in order to demonstrate the benefits of the two-step strategy.

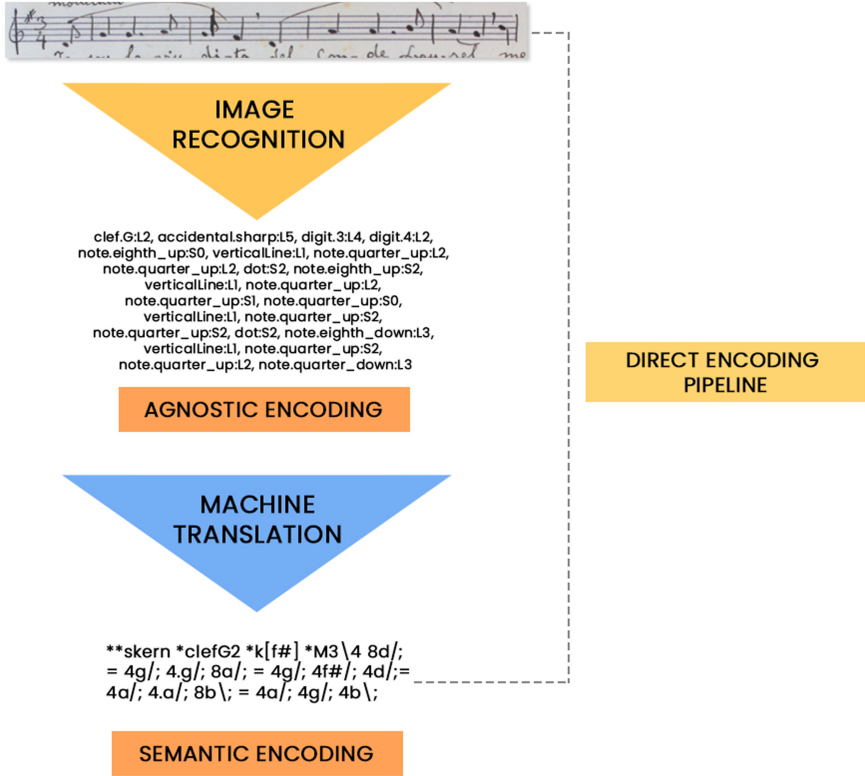


Fig. 2. Overview of the procedures proposed for complete OMR, receiving a staff-section image as input and predicting a semantic music encoding sequence as output.

3.1 Graphical Recognition

The graphical recognition step f_g takes an input image and produces a sequence of agnostic symbols. Given an input staff-section image $x \in \mathcal{X}$, f_g seeks for the sequence \hat{s} such that

$$\hat{s} = \arg \max_{s \in \Sigma_a^*} P(s | x). \tag{1}$$

To implement this step, we consider the state-of-the-art model OMR, which consists of a CRNN model trained with a CTC loss function. We follow the configuration specified in [6].

The convolutional block learns relevant features of the image and the recurrent block interprets them as a sequence of frames. The model eventually computes the probability of each symbol appearing in each input frame. To approximate \hat{s} of Eq. 1, we resort to a *greedy* decoding algorithm that takes the most likely symbol per frame, concatenates consecutive frames with the same symbol,

and then removes the ‘blank’ symbol introduced to train the model with the CTC loss function [12].

3.2 Translation Process

The graphical recognition produces a discrete sequence of agnostic symbols, where just the shape and the position within the staff are encoded (graphical features). As discussed above, this is insufficient to retrieve meaningful music information, so we need an additional step to obtain a semantic output.

Given a sequence of agnostic symbols, $\mathbf{s} \in \Sigma_a^*$, the translation step f_t can be expressed as seeking the sequence $\hat{\mathbf{t}}$, such that

$$\hat{\mathbf{t}} = \arg \max_{\mathbf{t} \in \Sigma_s^*} P(\mathbf{t} | \mathbf{s}). \quad (2)$$

We compute this probability by means of NMT. Given the novelty of this approach in the context of OMR, we consider two alternatives, whose effectiveness will be analyzed empirically.

The first approach is a Seq2Seq model with Attention mechanisms, hereafter referred to as *Seq2Seq-Attn*. This model is an encoder-decoder approach based on Recurrent Neural Networks (RNN), where the first part (the encoder) creates an embedding representation of the input sequence, usually known as the context vector, and the decoder produces, from this context vector and the previously predicted tokens, the next token of the translated sequence. Specifically, we resort to an advanced strategy which implements attention mechanisms: the ‘Global Attention’ approach, proposed by Luong et al. [18], where the previously mentioned context vector is regulated by an attention matrix, whose scoring regulators are given by the scalar product between the encoder and the decoder outputs.

The second considered model is the *Transformer* [30], that currently represents the backbone of state-of-the-art NMT. This model implements an encoder-decoder architecture, such as the previously described system, that replaces all the recurrent layers by attention-based ones, which are referred to in the literature as the multi-headed attention units. These units are not only able to compute faster the training process (as they are easily parallelizable) but have also proven to obtain better quality context vectors and translation decodings, which allows them to learn relevant grammatical and semantic information from the input sequences themselves.

In both cases, the specific configuration is set as done in our previous work [26], where good results for processing music encoding formats were attained.

3.3 Direct Encoding

A direct encoding performs a function $f_d : \mathcal{X} \rightarrow \Sigma_s$. Formally, given an input staff-section image $x \in \mathcal{X}$, it seeks for a sequence $\hat{\mathbf{t}}$ such that

$$\hat{\mathbf{t}} = \arg \max_{\mathbf{t} \in \Sigma_s^*} P(\mathbf{t} | x) \quad (3)$$

As far as we know, there is no single-step complete OMR system in the literature. In our case, we decided to implement the CRNN-CTC model used for image recognition (Sect. 3.1), but modifying the output alphabet to be that of the semantic output.

This implementation establishes a good comparison baseline, as it is the easiest and simplest model to implement and reduces the number of steps to one.

4 Experimental Setup

In this section, we present our experimental environment to evaluate the OMR pipelines. We detail the corpora used to perform and the evaluation process considered to obtain the results presented in Sect. 5.

4.1 Corpora

Two corpora of music score images, with varying features in printing style, have been used to assess and discuss the performance of the different pipelines.

The first considered corpus is the “Printed Images of Music Staves” (PrIMuS) dataset; specifically, the camera-based version [7]. It consists of 87,678 music incipits¹ from the RISM collection [1]. They consist of music scores in common western modern notation, rendered with Verovio [23] and extended with synthetic distortions to simulate the imperfections that may be introduced by taking pictures of sheet music in a real scenario, such as blurring, low-quality resolutions, and rotations.

The second considered corpus is a collection of four groups of handwritten score sheets of popular Spanish songs taken from the ‘Fondo de Música Tradicional IMF-CSIC’ (FMT),² that is a large set of popular songs manually transcribed by musicologists between 1944 and 1960.

The characterization of these corpora can be found in Table 1, while representative examples are shown in Fig. 3 and Fig. 4 for PrIMuS and FMT, respectively, along with agnostic and semantic annotations.

4.2 Evaluation Process

One issue that one may find when performing OMR experiments is to correctly evaluate the performance of a proposed model, as music notation has specific features to take into account. However, OMR does not have a standard evaluation protocol [6]. In our case, it seems convenient to use text-related metrics to approach the accuracy of the predictions. Despite not considering specific music features, in practical terms, we are dealing with text sequences.

¹ Short sequence of notes, typically the first ones, used for identifying a melody or musical work.

² <https://musicatradicional.eu>.

Table 1. Details of the considered corpora.

	PrIMuS	FMT
Engraving	Printed	Handwritten
Size of the corpus (staves)	87,678	872
Agnostic vocabulary size ($ \Sigma_a $)	862	266
Semantic vocabulary size ($ \Sigma_s $)	1,421	206
Running symbols (agnostic)	2,520,245	18,329
Running symbols (semantic)	2,425,355	18,616



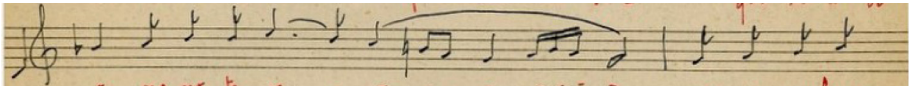
(a) Staff-section image.

clef.G-L2, accidental.sharp-L5, accidental.sharp-S3, accidental.sharp-S5,
accidental.sharp-L4, digit.2-L4, digit.4-L2, note.quarter-L3, note.eighth-S2,
dot-S2, note.sixteenth-L2, barline-L1

(b) Agnostic encoding of the first bar.

**skern ↵ *clefG2 ↵ *k[f#c#g#d#] *M2/4 ↵ 4b ↵ 8.a ↵ 16g#y ↵ = ↵

(c) Semantic encoding of the first bar ('↵' represents the end-of-line token).

Fig. 3. Selected example from PrIMuS: Incipit RISM ID no. 000102547, Incipit 1.1.1 *Peace troubled soul whose plaintive moan*. Anonymous.

(a) Staff-section image.

clef.G:L2, accidental.flat:L3, ..., note.beamedRight2_up-S2,
note.beamedBoth2_up-L3, note.beamedLeft2_up-S2, ...

(b) First tokens of the agnostic encoding.

**skern ↵ *clefG2 ↵ *k[] ↵ 4b- ↵ ... 24aL ↵ 24b ↵ 24aJ ...

(c) First tokens of the semantic encoding.

Fig. 4. Selected example from FMT (Canción de Trilla) [9]

For the above, we measured the performance of the proposed models with the Sequence Error Rate (SER). Let H be the predicted sequence and R the reference one, this metric computes the edit distance between H and R and divides it by the length (in tokens) of R . We chose this metric as it both represents accurately the performance of the model in recognition tasks and correlates with the effort a user would have to invest to manually correct the results.

To obtain a more robust approximation of the generalization error, we follow a 5-fold cross-validation process, where the resultant SER is the average of the produced test error within the five data partitions.

5 Results

The experimentation results are given in Table 2, comparing the proposed two-step approach with a direct encoding, that acts as a baseline. We also report the intermediate results of the former, to provide more insights. In the case of the translation process, the intermediate results show the SER obtained starting from a ground-truth agnostic sequence.

Concerning the intermediate results, it can be observed that the graphical recognition step performs well on the printed dataset and gets much worse results in the handwritten one, as might be expected in terms of their training set size and complexity. In the translation task, the tendency is similar but this time we observe a more reasonable SER in both cases. The Transformer is the best only-translation option when there is enough training data, while the Seq2Seq-Attn results better in the case of limited training data. As discussed next, this fact does not extrapolate to the complete pipeline.

If we analyse the complete pipeline, the results obtained using the combination of CRNN and NMT models outperform the direct encoding approach, both in the PRIMuS and the FMT dataset. The difference is especially significant in the handwritten small-sized corpus FMT, where the SER of the CRNN+Seq2Seq-Attn outperforms the direct encoding approach by a wide margin (around 20% of SER). One interesting fact from these results is that the NMT models can deal reasonably well with the inconsistencies introduced during the graphics recognition, as we observe that the final SER figures are much more correlated to the graphical recognition than to the translation process.

Furthermore, it is interesting to note that the Transformer is the most NMT accurate model when translating from ground-truth data. However, if we pay attention to the complete pipeline, it does not produce a model as robust to inconsistencies as the Seq2Seq-Attn model does. This scenario, in practical terms, is the most frequent in OMR, where the graphical recognition step tends to make mistakes. Therefore, the Seq2Seq-Attn approach is, as far as our results generalize, the most suitable alternative for the translation process in the two-step pipeline.

Table 2. Average SER (%) over the test set. The table shows the error amount produced in the recognition and encoding steps (as they have been trained separately) and the final error done by the complete pipeline, which receives an image as input and a semantic sequence as output. We highlighted in bold typeface the results that show better performance in the complete pipeline.

	PrIMuS	FMT
<i>Intermediate results</i>		
Graphical recognition (CRNN)	3.52	34.88
Translation process w/ Seq2Seq-Attn	2.04	9.53
Translation process w/ Transformer	0.53	15.43
<i>Complete pipeline</i>		
CRNN + Seq2Seq-Attn	4.28	36.76
CRNN + Transformer	6.35	38.88
CRNN Direct encoding (baseline)	4.66	52.24

Despite the aforementioned evidence, some doubts may appear referring to the error fluctuation between the presented pipelines, as we observe a drastic change in the performance between the two datasets. To further analyze the situation, we repeated the same experimentation in reduced versions of the PrIMUS dataset, where we try to find an intermediate point between FMT and PrIMuS complexities. This resulted in three new corpora with 10,000, 5,000 and 1,000 *samples*, (the FMT corpus has nearly 900 samples). The obtained results are graphically shown in Fig. 5. It can be observed that the tendency described from the original PrIMUS results, where the CRNN+Transformer performed the worst, is maintained until dropping to 5,000 samples, where the direct approach is then outperformed by it. In all cases, however, the CRNN+Seq2Seq-Attn is postulated as the best option by different margins, depending on the complexity of the dataset.

This new experiment summarized the behaviour of all alternatives. On the one hand, a direct encoding pipeline—which acted as baseline—depends highly on the amount of training data, attaining competitive results in such case. On the other hand, the two-step process, especially when using the Seq2Seq-Attn as translation mechanisms, clearly stands for the best option when training data is limited, also reaching the best performance when the training set is of sufficient size.

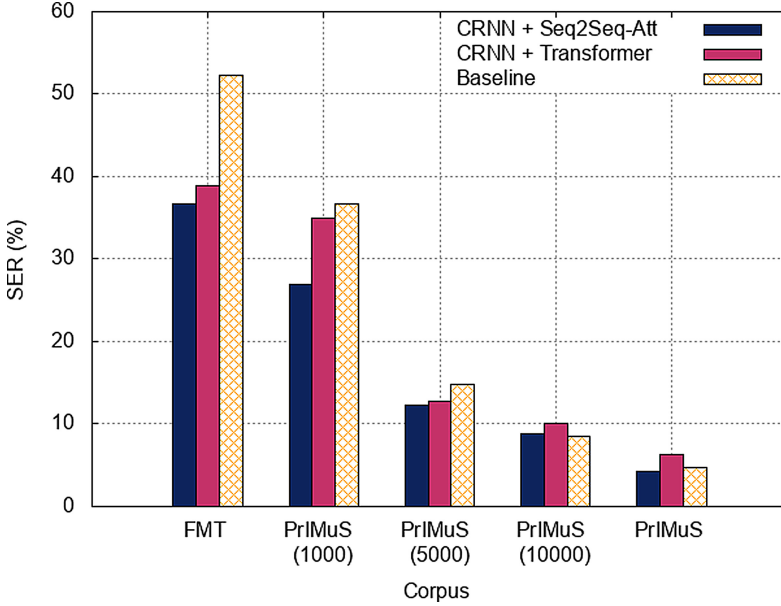


Fig. 5. Graphic bar plot comparison of the average SER produced by the proposed pipelines with the different corpora, which consists in the initially proposed datasets and two reductions on PrIMUS size in order to establish intermediate points between the handwritten and the printed corpus. The Baseline results refer to the Direct Encoding approach described in other sections.

6 Conclusions

We studied in this paper the development of complete OMR pipelines, which receive a music staff image as input and produce a standard music encoding sequence as output. We discussed how to approach this task by proposing a two-step pipeline based on a state-of-the-art image recognition model in OMR combined with NMT solutions for the encoding step. We also included a direct encoding pipeline that outputs directly the final encoding from the image. To evaluate these approaches, we experimented in two corpora of varying characteristics. After the experimentation, we observed different aspects about how these approaches perform over different corpora, where we obtained a relevant idea that outlines this work: the two-step pipeline with NMT is a good option when the target corpus to digitize does not have enough labeled data for learning the inherent complexity of the OMR, which is, in fact, an interest aim of this paper.

From a practical perspective, specifically in the case of early music heritage, it is common to find scenarios where manual data labeling is required in order to constitute a corpus before using OMR tools. As we saw in our experimentation, the OMR processes that include NMT models to perform the encoding step behave reasonably well in this case. This feature involves a great practical

advantage for these scenarios, as there is no need to label a vast amount of data to start using this tool. However, the two-step pipeline also has a considerable drawback: the corpus has to be labeled in two encoding languages (agnostic and semantic) in order to make it work. Despite this issue, there are possible ways of mitigation because the translation process does not depend on a specific manuscript; therefore, just one pretrained translation model, relieving the effort of manually creating the semantic annotation.

Despite the advances presented in this paper, we consider that further research is required to maximize the benefits this approach might bring, as this paper only proves that it is a feasible option for cases where the corpus does not provide enough data. This future research may focus on different topics such as improving the robustness to input inconsistencies of the NMT models (especially the Transformer) with data augmentation, the modelling of cohesive vocabularies to obtain more profit from the encoding models, or the study on how to integrate these systems to produce a single-step OMR pipeline with a dual training process.

Acknowledgments. This work was supported by the Generalitat Valenciana through project GV/2020/030.

References

1. Répertoire International des Sources Musicales (RISM) Series A/II: Music manuscripts after 1600 on CD-ROM. Technical report (2005)
2. 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)
3. Burgoyne, J.A., Devaney, J., Pugin, L., Fujinaga, I.: Enhanced bleedthrough correction for early music documents with recto-verso registration. In: Bello, J.P., Chew, E., Turnbull, D. (eds.) ISMIR 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA, 14–18 September 2008, pp. 407–412 (2008)
4. Burnard, L., Bauman, S. (eds.): A gentle introduction to XML. Text encoding initiative consortium. In: TEI P5: Guidelines for Electronic Text Encoding and Interchange (2007). <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SG.html>
5. Byrd, D., Simonsen, J.: Towards a standard testbed for optical music recognition: definitions, metrics, and page images. *J. New Music Res.* **44**, 169–195 (2015). <https://doi.org/10.1080/09298215.2015.1045424>
6. Calvo-Zaragoza, J., Hajič Jr., J., Pacha, A.: Understanding optical music recognition. *ACM Comput. Surv.* **53**(4) (2020)
7. Calvo-Zaragoza, J., Rizo, D.: 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, 23–27 September 2018, pp. 248–255 (2018)
8. Calvo-Zaragoza, J., Toselli, A.H., Vidal, E.: Handwritten music recognition for mensural notation with convolutional recurrent neural networks. *Pattern Recogn. Lett.* **128**, 115–121 (2019)

9. Clares Clares, E.: Canción de trilla. Fondo de música tradicional IMF-CSIC. <https://musicatradicional.eu/es/piece/12551>. Accessed 01 Feb 2021
10. Dalitz, C., Droettboom, M., Pranzas, B., Fujinaga, I.: A comparative study of staff removal algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**(5), 753–766 (2008)
11. Good, M., Actor, G.: Using MusicXML for file interchange. In: *International Conference on Web Delivering of Music 0*, 153 (2003)
12. 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)
13. Hajic, J., Pecina, P.: The MUSCIMA++ dataset for handwritten optical music recognition. In: *ICDAR (2017)*
14. Hankinson, A., Roland, P., Fujinaga, I.: The music encoding initiative as a document-encoding framework. In: *Proceedings of the 12th International Society for Music Information Retrieval Conference (2011)*
15. Huron, D.: *Humdrum and Kern: Selective Feature Encoding*, pp. 375–401. MIT Press, Cambridge (1997)
16. Koehn, P.: *Statistical Machine Translation*. Cambridge University Press, Cambridge (2009)
17. Koehn, P., et al.: Moses: open source toolkit for statistical machine translation. In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 177–180. Association for Computational Linguistics, Prague (2007)
18. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Márquez, L., Callison-Burch, C., Su, J., Pighin, D., Marton, Y. (eds.) *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, Lisbon, Portugal, 17–21 September 2015, pp. 1412–1421. The Association for Computational Linguistics (2015)
19. Pacha, A., Calvo-Zaragoza, J., Hajič jr., J.: Learning notation graph construction for full-pipeline optical music recognition. In: *20th International Society for Music Information Retrieval Conference*, pp. 75–82 (2019)
20. Pacha, A., Eidenberger, H.: Towards a universal music symbol classifier. In: *14th International Conference on Document Analysis and Recognition*, pp. 35–36. IAPR TC10 (Technical Committee on Graphics Recognition), IEEE Computer Society, Kyoto (2017)
21. Pacha, A., Hajič, J., Calvo-Zaragoza, J.: A baseline for general music object detection with deep learning. *Appl. Sci.* **8**(9), 1488 (2018)
22. Parada-Cabaleiro, E., Batliner, A., Schuller, B.W.: A diplomatic edition of il lauro secco: ground truth for OMR of white mensural notation. In: Flexer, A., Peeters, G., Urbano, J., Volk, A. (eds.) *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, Delft, The Netherlands, 4–8 November 2019, pp. 557–564 (2019)
23. Pugin, L., Zitellini, R., Roland, P.: Verovio: a library for engraving MEI music notation into SVG. In: *Proceedings of the 15th International Society for Music Information Retrieval Conference*, pp. 107–112. ISMIR, October 2014
24. Quirós, L., Toselli, A.H., Vidal, E.: Multi-task layout analysis of handwritten musical scores. In: Morales, A., Fierrez, J., Sánchez, J.S., Ribeiro, B. (eds.) *IbPRIA 2019*. LNCS, vol. 11868, pp. 123–134. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-31321-0_11

25. Ríos-Vila, A., Calvo-Zaragoza, J., Rizo, D.: Evaluating simultaneous recognition and encoding for optical music recognition. In: 7th International Conference on Digital Libraries for Musicology, DLfM 2020, pp. 10–17. Association for Computing Machinery, New York (2020)
26. 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) (2021)
27. Sánchez, J., Romero, V., Toselli, A.H., Villegas, M., Vidal, E.: A set of benchmarks for handwritten text recognition on historical documents. *Pattern Recognit.* **94**, 122–134 (2019)
28. Sapp, C.S.: Verovio humdrum viewer. In: Proceedings of Music Encoding Conference (MEC), Tours, France (2017)
29. Tuggener, L., Elezi, I., Schmidhuber, J., Stadelmann, T.: Deep watershed detector for music object recognition. In: 19th International Society for Music Information Retrieval Conference, Paris, 23–27 September 2018 (2018)
30. Vaswani, A., et al.: Attention is all you need (2017)