Sheet Music Transformer: End-To-End Optical Music Recognition Beyond Monophonic Transcription

Antonio Ríos-Vila[1[0000−0002−7770−9726]], Jorge Calvo-Zaragoza[1[0000−0003−3183−2232]], and Thierry Paquet[2[0000−0002−2044−7542]]

[1] Pattern Recognition and Artificial Intelligence Group, University of Alicante, Spain
{arios, jcalvo}@dlsi.ua.es
[2] LITIS Laboratory - EA 4108, Rouen University, France
thierry.paquet@litislab.eu

Abstract. State-of-the-art end-to-end Optical Music Recognition (OMR) has, to date, primarily been carried out using monophonic transcription techniques to handle complex score layouts, such as polyphony, often by resorting to simplifications or specific adaptations. Despite their efficacy, these approaches imply challenges related to scalability and limitations. This paper presents the Sheet Music Transformer (SMT), the first end-to-end OMR model designed to transcribe complex musical scores without relying solely on monophonic strategies. Our model employs a Transformer-based image-to-sequence framework that predicts score transcriptions in a standard digital music encoding format from input images. Our model has been tested on two polyphonic music datasets and has proven capable of handling these intricate music structures effectively. The experimental outcomes not only indicate the competence of the model, but also show that it is better than the state-of-the-art methods, thus contributing to advancements in end-to-end OMR transcription.

Keywords: Optical Music Recognition · SMT · Transformer · Polyphonic music transcription · GrandStaff · Quartets

1 Introduction

Music is a valuable component of our cultural heritage, as it is a resource that enables an understanding of the social, cultural and artistic trends of each period of history. Most existing documents have been transmitted in the form of printed and handwritten documents. In the same way that Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) are successfully applied in order to extract content from text images, OMR is the research area that produces models that automatically recognize and transcribe sheet music [6].

The progress made in OMR, which initially depended on multi-stage workflows [27], has diversified with the emergence of Deep Learning solutions. Two major approaches currently exist: object detection workflows [36,5],—in which

notes are individually detected and assembled into a digital document—and holistic workflows [32,8,4,34,2], also known as *end-to-end* methods, in which systems directly generate a symbolic representation of a region of a given document. The latter paradigm dominates the current state of the art of other applications, such as the recognition of text, speech or mathematical formulae [11,10,38], and has also proven successful in OMR.

Despite the successful results obtained, OMR has, to date, found solutions that are applicable only to monophonic scores[3]. However, many non-monophonic music documents have not been dealt with by the literature concerning OMR.

This need to cover all music documents signifies that end-to-end OMR is now shifting toward new projects that were not, until recently, considered feasible. Recent literature focuses on applying OMR systems to complex scenarios, such as full-page documents [9], the simultaneous recognition of music and lyrics [24] and polyphonic scores [15,29]. However, most of these advances are *ad-hoc* adaptations or pipeline-based solutions of the current state of the art for monophonic music transcription. There is no solution that currently goes beyond monophonic transcription, but rather adaptations that reduce the problem in order to make it close to a monophonic scenario and solve it using state-of-the-art methods. According to the recent works of HTR [33,13] and Document Understanding (DU) [12,22], OMR should seek to break this monophonic-dependent barrier. In this paper, we propose the Sheet Music Transformer (SMT), an image-to-sequence approach—based on autoregressive Transformers—that is able to transcribe music input images beyond monophony, without adaptations or specific preprocessing steps. In order to test this approach, we deal with the challenge of polyphonic music transcription, which is the most complex in OMR, and compare it to other current state-of-the–art approaches. We specifically test it in two common scenarios regarding sheet music: pianoform scores and those for string quartets. Our experiments empirically demonstrate that our solution provides robust results that clearly outperform the state of the art. The scientific contributions of this paper can be summarized as follows:

- We propose the SMT, the first image-to-sequence-based approach for music transcription that is able to deal with transcripts beyond the monophonic level. In our experiments, we demonstrate that this approach performs better than current state-of-the-art solutions.
- We explore and analyze different configurations for feature extraction in order to produce a model that is better suited to complex music layouts.
- We create an adaptation of a well-known music dataset for end-to-end OMR that goes beyond monophonic-level transcription.

The remainder of the paper is structured as follows: in Section 2, we present the polyphonic transcription challenge in OMR, explain the current state of the art and discuss its limitations. In Section 3, we describe the end-to-end SMT Neural Network proposed in this paper, while the experiments conducted are presented in Section 4, along with the results, Section 5, and analysis, Section

[3] Monophonic scores are pieces of music in which only one voice is present

6, that are drawn from them. Finally, we conclude the paper in Section 7, in addition to discussing other avenues for future work.

2 End-to-end polyphonic OMR

The state-of-the-art holistic OMR currently addresses music transcription as a sequence recognition problem, in which the most probable symbolic representation $\hat{\mathbf{s}}$—encoded in the Σ_a music notation vocabulary—for each staff-section image x is sought.

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s} \in \Sigma_a} P(\mathbf{s} \mid x) \tag{1}$$

Neural network approaches approximate this probability by training with the Connectionist Temporal Classification (CTC) loss [18]. The output of the network consists of a probability map, which contains the probabilities of all the tokens within the vocabulary Σ_a. In order to allow for the possibility of no prediction at a given timestep, CTC provides an extra blank token (ϵ), and the output vocabulary of the network therefore becomes $\Sigma'_a = \Sigma_a \cup \{\epsilon\}$.

At inference, OMR methods employ a *greedy* decoding, from which the most probable sequence is retrieved given an input image x. This formulation forces networks to define a reshape function based on the vertical collapse of the feature map, as a sequence must be retrieved. This vertical collapse assumes that the symbols can be read from left to right and that the frames[4] of the feature map always contain information about the same symbol in this case.

2.1 The challenge of transcribing polyphonic scores

This methodology, which is able to address single-staff music transcription, has proven successful for the transcription of both printed and handwritten scores [34,4,8,7] and has attained very accurate results, surpassing those of any other image-to-sequence approach [28].

Despite this, the approach is severely limited. As mentioned in Section 2, the approach performs a vertical collapse to group all the information regarding the image into a sequential structure, which simplifies the problem. However, this also prevents the network from dealing with multiple structures, i.e., music staves or symbols. This entails a difficulty when transcribing several music structures, even with the simplest graphical complexity.

If we consider scores other than those of a monophonic nature, such as polyphonic scores, the challenge increases. These engravings must be read in a particular order during interpretation, since there are staves that must be read simultaneously. Rather than performing a line-by-line reading from top to bottom and left to right, interpretation is tied to staff groups, namely systems, in which all elements are read simultaneously from left to right (see Figure 2).

[4] Column-wise elements of the image

This particularity poses several challenges, since the premise that a frame contains the information regarding a single music symbol—even at the staff level—no longer holds in this case. Some vocabulary-based shortcuts can be taken if complexity does not increase much, as occurs in the work of Alfaro-Contreras et al. [1]. These approaches could be extended to more complex scenarios, such as polyphonic scores, but at the cost of increasing the length of the ground truth sequence. This scalability issue becomes unfeasible with larger samples—e.g. scores for string quartets—since vertical collapse cannot produce sufficient frames with which to accurately transcribe the score. It would, therefore, appear that methodological advances must be made in order to produce more robust and generic systems.

2.2 Current approaches

The first avenue found in the literature concerns employing a *divide and conquer* approach, in which single-staff images are segmented by means of a Layout Analysis method and transcribed using state-of-the-art methods, as occurs in [15]. However, despite proposing a robust solution to transcription, the alignment retrieval—in which simultaneous music notes must be placed in the same timestep in the digital document—of the music score becomes a challenging process, since the system retrieves individual decontextualized single-staff transcriptions. Exhaustive post-processing is, therefore, required in order to retrieve a correct and usable music document.

A holistic approach dealing with the aforementioned challenge recently appeared in [29]. Its authors specifically suggest an extension to the current state-of-the-art-based musical score alignment by exploiting music encoding features and score rotations. This method consists of a neural network that unfolds the rotated input score image and sequentially transcribes all simultaneous events as text lines in the ground truth—with special tokens that indicate polyphonic timesteps and the conclusion of simultaneous events. This system was tested on excerpts from single-line pianoform music, which are the most complex in OMR[5]. Although the aforementioned paper presents very competitive results in the case of excerpts from single-system pianoform music, the approach is very limited. Indeed, since this model unfolds the score in order to detect simultaneous events, this system is still limited to input images in which all the events are simultaneous, and does not allow multiple systems in a single image. A complementary Layout Analysis should, therefore, be performed in order to address complex documents. Moreover, this method has not been stressed in larger images with simultaneous events, with scalability being another potential issue when using this system.

[5] The reason for the difficulty related to pianoform scores is that they contain multiple voices within a single staff and cross-staff interactions that are not easily seen in the image of the document.

3 Sheet Music Transformer

In this paper, we present the Sheet Music Transformer (SMT). This model is an autoregressive end-to-end neural network that generates the transcription of a given polyphonic music system input image. The model consists of two fundamental components: an encoder and a decoder. The encoder acts as a feature extractor of the image x, producing a feature map x'_e. The decoder consists of an autoregressive conditioned language model that predicts the probability of each symbol of the vocabulary in a timestep given the output feature vector from the encoder and the previously generated symbols. This is formalized as:

$$\hat{y} \underset{\hat{y} \in \Sigma}{=} P(\hat{y}_t \mid x'_e, (\hat{y}_0, \hat{y}_1, \hat{y}_2, ..., \hat{y}_{t-1})) \tag{2}$$

where Σ represents the symbol vocabulary used in order to encode the music content, x is the input feature map and t is the current timestep. A graphic scheme of the SMT is depicted in Figure 1.

3.1 Encoder

Let $x \in \mathbb{R}^{c \times h \times w}$ represent the input image of the system, where h and w respectively denote its height and width in pixels and c denotes the number of channels. This block is typically implemented through the use of Convolutional Neural Networks [12,33], owing to their capacity to process image signals. This module therefore results in an image processor that outputs c_e two-dimensional feature maps denoted by $x_e \in \mathbb{R}^{h_e \times w_e \times c_e}$. Note that, h_e and w_e respectively relate to the image dimensions $h_e = \frac{h}{r_h}$ and $w_e = \frac{w}{w_t}$, where r_h and r_w represent the corresponding downscaling produced by this network.

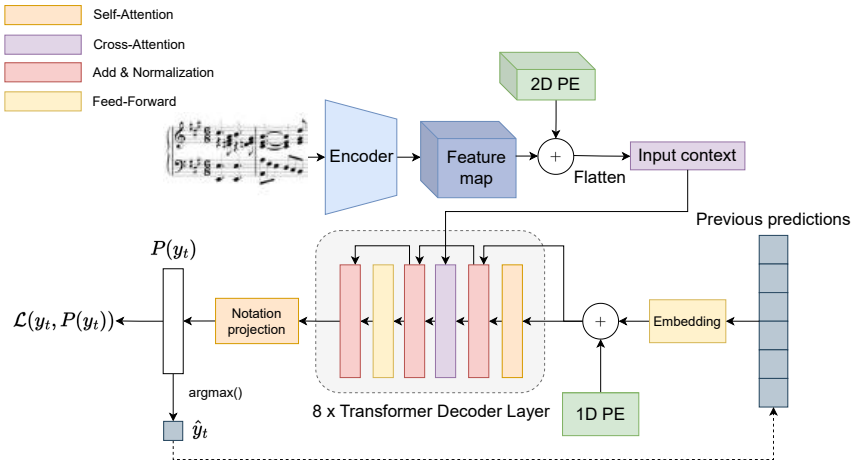


Fig. 1: Graphic scheme of the SMT architecture.

3.2 Decoder

In this work, we have opted for a Transformer [37] decoder, as it is currently the state-of-the-art method for tasks involving conditional sequence generation of varying lengths owing to its capacities to capture temporal relationships within the data through the use of multi-head attention mechanisms.

The decoder consists of a language model that, at each time step t (which is conditioned by the input feature vector produced by the encoder, x'_e along with the previously predicted tokens, $(\hat{y}_0, \ldots, \hat{y}_{t-1})$) outputs a probability distribution $p_t \in \mathbb{R}^{|\Sigma|}$ over the Σ vocabulary symbols, with the predicted token $\hat{y}_t$ being that which maximizes this probability score. The process starts with a special start-of-transcription symbol, $\hat{y}_0 = <sot>$, and continues until an end-of-transcription token is predicted, $\hat{y}_{|\hat{\mathbf{y}}|} = <eot>$.

We should stress that the encoder outputs a two-dimensional feature map to serve as a context in which to condition the decoder. However, the decoder works with one less dimension, since it is a sequence generator. In order to connect these two modules, the two-dimensional feature maps must be converted into a one-dimensional format that is suitable for the decoder, with a straightforward approach being that of flattening this structure across height and width, resulting in sequences of length $h_e \times w_e$.

Another aspect to take into account is that the Transformer implements a positional encoding (PE) mechanism with which to model its otherwise order-agnostic operation. This mechanism adds a position vector to each input element, determined by its location in the sequence. The addition of a one-dimensional PE to the unfolded feature maps could initially be considered possible, but while this might be sufficient for monophonic music sequences, it would result in a loss of spatial information when targeting polyphonic scores in which multiple voices are played simultaneously. In order to ensure that the model is aware of all the dimensions of the image, we, therefore, incorporate a two-dimensional PE within the feature maps before they are flattened into a one-dimensional sequence. The two-dimensional PE proposal is based on sine and cosine functions, akin to the original one-dimensional PE, in which the first half of the feature dimensions—i.e., $[0, c_e/2)$—is meant for horizontal positions, whereas the second half—i.e., $[c_e/2, c_e)$—is used for the vertical ones, the same that in [33,12].

$$\begin{aligned}
\mathrm{PE}_{\mathrm{2D}}\left(\mathrm{pos}_t, 2i\right) &= \sin\left(\mathrm{pos}_t / 10000^{2i/c_e}\right) \\
\mathrm{PE}_{\mathrm{2D}}\left(\mathrm{pos}_t, 2i+1\right) &= \cos\left(\mathrm{pos}_t / 10000^{2i/c_e}\right) \\
\mathrm{PE}_{\mathrm{2D}}\left(\mathrm{pos}_f, c_e/2+2i\right) &= \sin\left(\mathrm{pos}_f / 10000^{2i/c_e}\right) \\
\mathrm{PE}_{\mathrm{2D}}\left(\mathrm{pos}_f, c_e/2+2i+1\right) &= \cos\left(\mathrm{pos}_f / 10000^{2i/c_e}\right)
\end{aligned} \tag{3}$$

where pos_t and pos_f respectively specify the horizontal (width) and vertical (height) pixel positions and $i \in [0, c_e/4)$ denotes the feature dimension of the output map.

3.3 Output Encoding

The definition of an output encoding for our system is another key step to cover. Of the variety that OMR provides, the first choices to consider are the most widespread musical encodings in digital musicology: MEI [20] and MusicXML [17], which represent the components and metadata of a musical score in an XML-based markup encoding. Despite their extensive capabilities, these formats are overly verbose, which is not convenient for OMR systems, as their conversion to deep-learning-friendly formats tends to be lossy.

In this paper, we have employed the text-based Humdrum **kern encoding format, which is included in the Humdrum tool-set [21] and is hereafter referred to simply as KERN. This music notation is one of the representations found most frequently in computational music analysis. Its features include a simple vocabulary and an easy-to-parse file structure, which is highly suitable for end-to-end OMR applications. Moreover, KERN files are compatible with software dedicated to music [31,26] and can be automatically converted to other music encodings, such as those mentioned above, by means of straightforward operations. A simple example of a Humdrum **kern-encoded score[6] is shown in Figure 2.

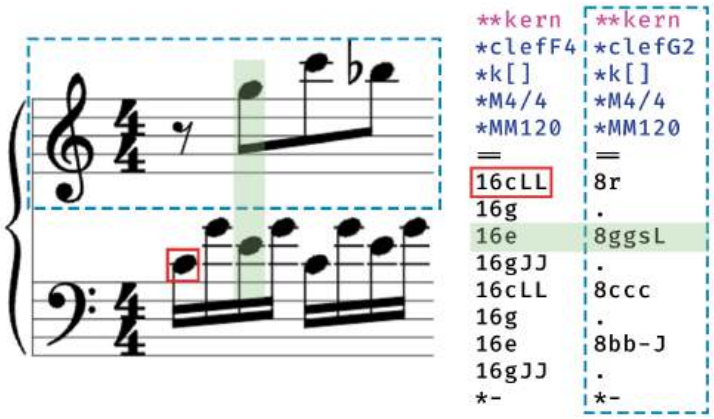


Fig. 2: Example of an excerpt from Humdrum **kern-encoded pianoform music. The blue dashed line represents a **kern spine, which is a voice in the musical score, in this case, a staff. The green box represents simultaneous notes during interpretation, which are represented as a text line in the ground truth. The red box is a single musical symbol (in this case, a note). The ground truth is read top to bottom and left to right, while the score is read from left to right and bottom to top.

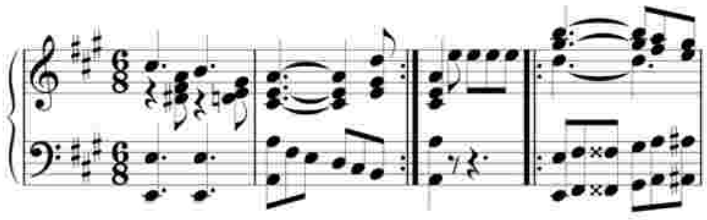
[6] Please refer to the official Humdrum **kern syntax (`https://www.humdrum.org/rep/kern/`) for a more detailed explanation.

4 Experimental setup

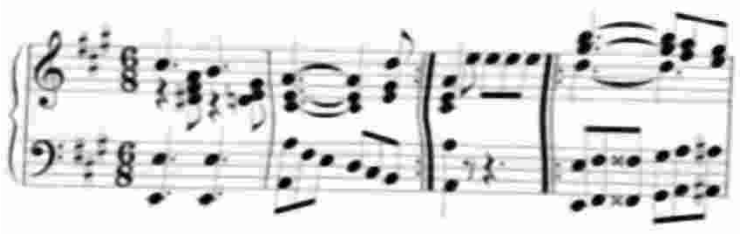
In this section, we present the experimental framework designed in order to evaluate the performance of our method in comparison to other state-of-the-art methods[7].

4.1 Corpora

Two different datasets of polyphonic music scores have been used in order to perform our experiments.



(a) GrandStaff music excerpt



(b) Distorted version of (a).

Fig. 3: Examples of the data contained in the GrandStaff and Camera GrandStaff dataset.

GrandStaff The GrandStaff dataset is a publicly available corpus[8] [29] that consists of 53,882 printed images of single-line (or system) pianoform scores, along with their digital score encoding. The dataset is composed of both original works from six authors—from the Humdrum[9] repository—and synthetic augmentations of the music encodings that make it possible to provide a greater

[7] The implementation of the model and the datasets are available in `https://grfia.dlsi.ua.es/sheet-music-transformer/`

[8] `https://sites.google.com/view/multiscore-project/datasets`

[9] `https://github.com/humdrum-tools/humdrum-data`

variety of musical sequences and patterns. The dataset comes with an official partition, in which the 7,000 original scores are used as a test set, and the 46,882 samples generated from augmentations comprise the training set. The dataset comes with an alternative version that introduces distortions into images to make them resemble low-quality photocopies. This version is, from here on, referred to as Camera GrandStaff. Figure 3 depicts two samples of this dataset.

Quartets Since the GrandStaff dataset is the only publicly available corpus designed for polyphonic end-to-end OMR, we produced an additional dataset with which to test both the state-of-the–art methods and our approach in a different polyphonic scenario. In this paper, we introduce the Quartets dataset. Quartets is a well-known collection employed in the Audio to Score field [30,3]. As the dataset provides the Humdrum **kern transcriptions from the excerpts of music, we produced a single-system transcription version of it. The dataset provides pieces that were randomly split from the original audios, namely pieces, into portions of approximately seven seconds, resulting in a total of 38 051 excerpts. These excerpts were rendered into printed music images using the Verovio Tool [26]. Once the music images had been generated, we distorted the image using the same operations as those employed with Camera GrandStaff[10] and included an additional distortion that simulated old printed ink, which contains bleeding and erasing errors. This distorted image was eventually fused with a random texture from a set of images on old paper. An example of this dataset is depicted in Figure 4.

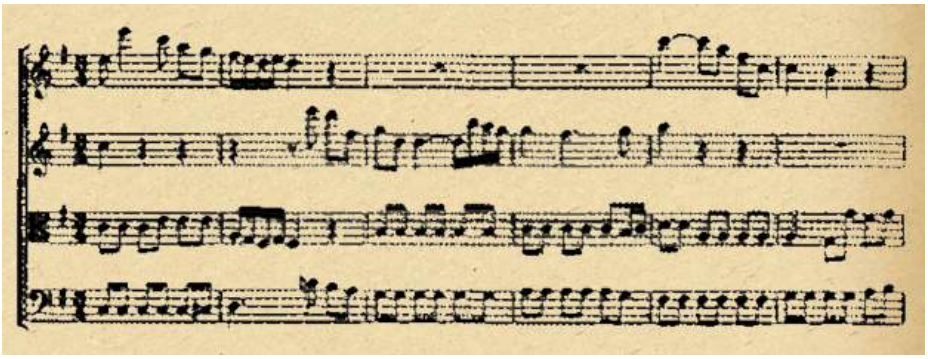


Fig. 4: Examples of an excerpt of music from the Quartets dataset.

We followed the same partitions as those provided in the original dataset. The training set specifically contains 18 162 samples for Haydn, 7 435 samples for Mozart, and 12 454 samples for Beethoven. Each corpus is divided into three

[10] All of the operations pertain to the ImageMagick image processing library.

splits at piece level: train (70%), validation (15%), and test (15%), and are combined in order to retrieve the partitions of the corpus. Please note that the partitions are made at the piece, level, meaning that excerpts originating from the same score will always be in the same partition, thus avoiding potential test biases.

4.2 Neural networks configuration

This paper introduces an image-to-sequence approach whose objective is to transcribe music scores that are more complex than monophony.

Baseline Section 2.2 mentions the work of [29], which is currently the only end-to-end approach that transcribes musical scores other than those of a monophonic nature, with some *ad-hoc* adaptations. We propose this system as our baseline, since its results were obtained using the GrandStaff dataset and can easily be applied to that of the Quartets. We first implemented the best performing version of the model the authors presented with the raw Humdrum **kern notation, which is a Fully Connected Convolutional Neural Network, as the encoder, with a Recurrent Neural Network as a decoder, namely CRNN. In order to provide a fair comparison—which has not been explored in state of the art for the Quartets dataset—we also implemented the Transformer decoder version that the authors presented in the paper, which is referred to as CNNT[11].

SMT We propose an image-to-sequence approach with three different feature extractors. The common aspect of all the SMT tested in this paper is the implementation of a decoder, and we specifically follow the implementation of [12]. The SMT contains a transformer decoder with eight layers, four attention heads and an embedding size of 256 features in both the attention and feed-forward modules. Given this common framework, the three variants of the feature extractor are the following:

- **CNN** As occurs in the recent works of HTR [33,12,13], the feature extractor is a fully-connected Convolutional Neural Network (CNN). We specifically implement the same backbone as that employed in [12], as it is the lightest network that can achieve good results. This convolutional backbone is updated in order to rescale the height of the image by 16, since this attained the best results in previous experimentation. This implementation is referred to as SMT_{CNN}.
- **Swin Transformer** The results of [22] inspired us to consider the Sliding Windows Transformer (Swin-T) model as our feature extractor. This model is an adaptation of the Vision Transformer [14] with sliding-window mechanisms and patch-merging layers that imitate the hierarchical behavior of

[11] The baseline experiments were performed with the public implementation of the model, which can be found at `https://github.com/multiscore/e2e-pianoform`

well-known CNN backbones. In order to attain comparable results to those of the SMT_{CNN}, our Swin-T implementation maintains only the first three layers from the network, with feature spaces of 64, 128 and 256, maintaining its original layer ratio. This implementation is referred to as SMT_{SWIN}.

– **ConvNexT** Given the performance results of the Swin-T in the Computer Vision field, the work by Liu et. al. [23] proposes an adaptation of traditional CNN backbones, specifically the ResNet50, so as to imitate the behavior of the Swin-T. For the sake of completion, we implement an adaptation of the ConvNexT as a backbone for our network. We specifically maintain only the first three layers—with the same ratios—of the network with depths of 64, 128 and 256.

4.3 Metrics

The assessment of the performance of OMR experiments is one of the main challenges of the field. Despite several efforts [35,19], OMR does not provide a standard evaluation framework. In the case of end-to-end approaches, it is convenient to employ text-related metrics in order to perform this evaluation. Since we are comparing our method to current state-of-the-art approaches in this paper, we employ the same metrics as those presented in [29]. All of these measures are based on the normalized mean edit distance between a hypothesis sequence and a reference sequence by the length of the reference. These measures are the Character Error Rate (CER), which evaluates among minimum semantic units in the output encoding, Symbol Error Rate (SER), a commonly-used metric that takes complete symbols as evaluation tokens, and Line Error Rate (LER), which approximates the accuracy of the model as regards both content and **kern document structure retrieval.[12]

5 Results

Table 1 shows the results attained for both the baseline and our approach for the GrandStaff, Camera GrandStaff and Quartets collections.

We first demonstrate that the state-of-the-art method struggles when the number of voices increases in the music score, with a CER of 21.9%, a SER of 22.5% and a LER of 67.0% in the Quartets Dataset, which are much worse than those ones for both GrandStaff and Camera GrandStaff. Given the results attained by our approach, we can clearly determine that the Convolution-based approaches, SMT_{CNN} and SMT_{NexT}, considerably outperform the state-of-the–art method, particularly as regards Camera GrandStaff and Quartets. The SMT_{SWIN}, however, attains the worst results. This underperformance can be explained by the fact that the Swin-T model, although light and powerful, is composed entirely of self-attention layers. Since the Transformer architecture has

[12] Evaluating the structure of the document is a key aspect in OMR when the transcription is more complex than the staff level, since the correct alignment between transcribed notes is relevant for a retrieval of the correct interpretation

Table 1: Average CER, SER, and LER (%) obtained by the models studied in the test set for both the perfectly printed and the distorted versions of the corpora presented in this work.

Method	**GrandStaff**						**Quartets**		
	Ideal			**Camera**					
	CER	**SER**	**LER**	**CER**	**SER**	**LER**	**CER**	**SER**	**LER**
Baseline									
CRNN	5.0	7.3	23.2	7.2	9.9	29.5	21.9	22.5	67.0
CNNT	7.9	11.1	32.4	9.4	12.3	33.3	30.8	32.6	82.4
Approach									
SMT_{CNN}	5.7	7.6	19.5	6.9	8.5	20.2	2.8	3.0	10.9
SMT_{SWIN}	53.2	70.1	98.2	60.3	82.3	100.0	40.5	50.2	75.7
SMT_{NexT}	**3.9**	**5.1**	**13.1**	**5.3**	**6.2**	**13.5**	**1.3**	**1.4**	**5.6**

a high inductive bias and no pretraining is performed, it is reasonable to assume that this model requires a greater amount of samples than the CNN and the ConvNexT in order to learn to extract relevant features from the image. Despite this, the SMT_{NexT} proposal attains outstanding results for all the datasets. There are improvements as regards CER, SER and LER of 22%, 32.9% and 43.5% in GrandStaff; 26.4%, 37.4% and 54.2% in Camera GrandStaff, and 91.8%, 91.6% and 89.1% in the Quartets dataset. These results demonstrate that our image-to-sequence approach is both a viable and scalable option with which to transcribe music more complex than monophony without *ad-hoc* adaptations.

6 Discussion

Although the results described in the previous section are positive, an analysis of the SMT performance is required in order to explain and understand these improvements. This study is carried using the model that performed best in our experiments: the SMT_{NexT}.

One of the most noteworthy results is the substantial improvement as regards the LER metric. For example, the SMT_{CNN}, despite producing slightly worse CER and SER than the CRNN baseline, notably improves the LER. Although this metric suggests an overall improvement to the output quality, it is difficult to assess what this implies. In order to clarify this, we evaluated what percentage of the test results had a correct **kern structure. That is, we assessed how many documents could be processed by a standard musicological tool, such as Verovio Humdrum Viewer [26] or MuseScore. The results of this method are depicted in Table 2.

These results make it possible to state that there is a clear correlation between LER and overall document quality, as an improvement to this metric significantly increases the number of directly usable documents, which can be directly edited and processed with musicological tools. It is consequently possible to state that

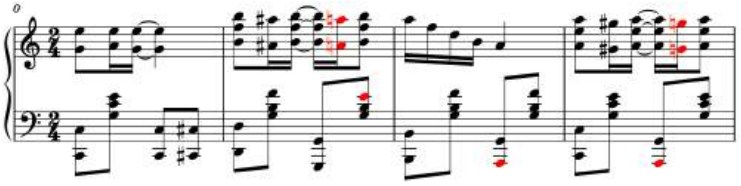
Table 2: Average Render percentage (%) values for the best-performing models for both the baseline and our approach in the test sets for the collections evaluated.

	Samples	CRNN		SMT_{NexT}	
		LER	Render %	LER	Render %
GrandStaff	7661	23.2	70.3	13.1	98.6
Camera GrandStaff	7661	29.5	51.1	13.5	96.4
Quartets	6107	67.0	22.5	5.6	90.7

the SMT produces outputs that are easier for end-users to edit and handle, thus making them more usable.

Two visualization errors from both the Camera GrandStaff and Quartets dataset are shown in Figures 5 and 7, respectively. The highlighted errors were computed using the musicdiff tool [16] and visualized with the MuseScore editor[13].

(a) Input test image



(b) Render of the prediction made by the SMT_{NexT} with highlighted transcription errors.

Fig. 5: Test example from the GrandStaff dataset with the errors highlighted. This specific sample attained a CER of 5.0%, a SER of 6.0% and a LER of 20.5%.

[13] https://musescore.org/es

Both examples show that, visually, the model is capable of generating a correct music sequence in terms of syntax, as no bar-completion or time-related errors are found. Indeed, most of the errors that are highlighted in these examples show that most of them are pitch misplacements. This means that, although the model is able to recognize the shape of the notes accurately, it generally fails to predict their position in their staff, or to place accidentals. Another interesting case is that shown in Figure 6.

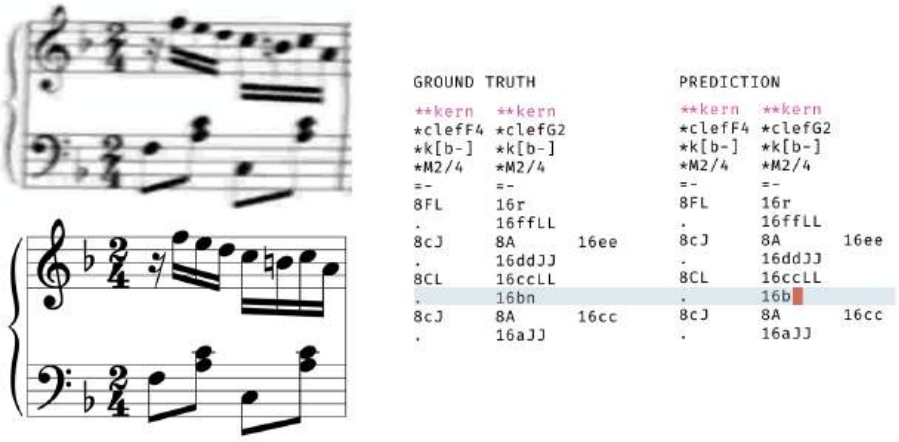
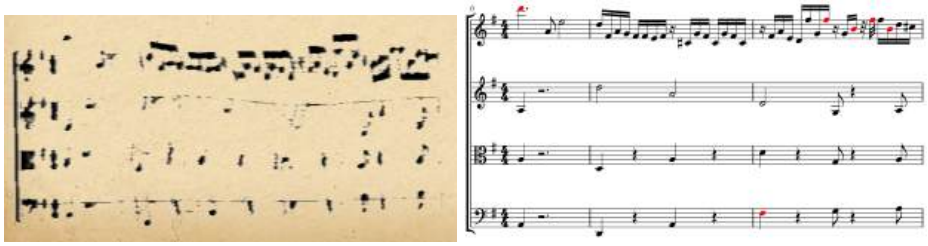


Fig. 6: Example of a redundant error in the Camera GrandStaff dataset, in which the ground truth contains redundant annotations that the language model avoids.

In this figure, note that the SMT predicted a visually identical excerpt of music. However, when reading the transcription produced in comparison to the ground truth, there is an insertion error—in line 11, which is highlighted in blue and red. Specifically, the model should have predicted the note 'B' with a natural accidental (♮). This is a specific difficulty that the Humdrum **kern annotation rules has with music renderers. The Humdrum **kern syntax specifies that if a note is affected by an accidental, the annotator must always indicate it, independently of how it is graphically seen. However, music renderers perform the simplifications of Common Western Notation rules when visualizing the music score. This results in cases in which some accidentals are rendered or the music rules are not provided, but these can be found in the ground-truth transcription.

This is a specific case in music notation in which text-based metrics penalize unfairly, as the piece produced is musically correct and equivalent to the ground truth. It is, therefore, possible that the metrics presented in Table 1, although necessary to assess document-quality information, are pessimistic figures of merit for the methodological evaluation. Even though no consensus has been reached as to the best way in which to evaluate OMR, and that this is still, therefore, an



(a) Input test image (b) Prediction from the SMT_{NexT}.

Fig. 7: Test example from the Quartets dataset with the errors highlighted. This specific sample attained a CER of 1.0%, a SER of 1.8% and a LER of 3.2%.

open research question, we corroborate that the results of this paper are robust, usable and musically-accurate, and this is also supported by the information provided in Table 2.

7 Conclusion

In this paper, we tackle the challenge of advancing Optical Music Recognition beyond monophonic transcription, which has traditionally involved simplifications or *ad-hoc* adaptations of existing methods. We introduce the Sheet Music Transformer (SMT) model, an autoregressive Transformer-based model that utilizes attention computation and language modeling to transcribe input scores into digital music encodings. We evaluate this approach in two polyphonic music scenarios: pianoform and string quartet scores. Our results demonstrate that the SMT model not only effectively transcribes complex musical layouts, but also outperforms current state-of-the-art methods, thus implying a significant advance in OMR.

This research also opens avenues for further improvements to OMR. Firstly, and as discussed in Section 6, standard OMR evaluation methods may overlook musicological interpretations, leading to pessimistic results. Addressing this issue could involve exploring graphic-based output music encodings [7] or employing additional musicological metrics in order to assess transcription accuracy [25]. The development of segmentation-free full-page transcription methods is another promising direction, as it is not limited by image features or layout constraints. Finally, we propose that the Universal OMR transcription challenge, a desired goal in OMR research, could be addressed by profiting from the language modeling capabilities of the SMT, thus allowing it to be trained to transcribe music engraved in various manners.

8 Acknowledgements

This paper is part of the project I+D+i PID2020-118447RA-I00 (MultiScore), funded by MCIN/AEI/10.13039/501100011033. The first author is supported by grants ACIF/2021/356 and CIBEFP/2022/19 from the "Programa I+D+i de la Generalitat Valenciana".

References

1. Alfaro-Contreras, M., Calvo-Zaragoza, J., Iñesta, J.M.: Approaching end-to-end optical music recognition for homophonic scores. In: 9th Iberian Conference Pattern Recognition and Image Analysis. Lecture Notes in Computer Science, vol. 11868, pp. 147–158. Springer, Madrid, Spain (2019)
2. Alfaro-Contreras, M., Ríos-Vila, A., Valero-Mas, J.J., Iñesta, J.M., Calvo-Zaragoza, J.: Decoupling music notation to improve end-to-end optical music recognition. Pattern Recognition Letters **158**, 157–163 (2022)
3. Arroyo, V., Valero-Mas, J.J., Calvo-Zaragoza, J., Pertusa, A.: Neural Audio-To-Score Music Transcription For Unconstrained Polyphony Using Compact Output Representations. In: Proceedings of the 47th IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 4603–4607. Singapore, Singapore (2022)
4. Baró, A., Riba, P., Calvo-Zaragoza, J., Fornés, A.: From optical music recognition to handwritten music recognition: A baseline. Pattern Recognition Letters **123**, 1–8 (2019)
5. Baró, A., Riba, P., Fornés, A.: Musigraph: Optical music recognition through object detection and graph neural network. In: International Conference on Frontiers in Handwriting Recognition. pp. 171–184. Springer (2022)
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.: End-to-end neural optical music recognition of monophonic scores. Applied Sciences **8**(4), 606 (2018)
8. Calvo-Zaragoza, J., Toselli, A.H., Vidal, E.: Handwritten music recognition for mensural notation with convolutional recurrent neural networks. Pattern Recognition Letters **128**, 115–121 (2019)
9. Castellanos, F.J., Calvo-Zaragoza, J., Iñesta, J.M.: A neural approach for full-page optical music recognition of mensural documents. In: Proc. of the 21th Int. Society for Music Information Retrieval Conference. pp. 12–16 (2020)
10. Chiu, C.C., Sainath, T.N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R.J., Rao, K., Gonina, E., Jaitly, N., Li, B., Chorowski, J., Bacchiani, M.: State-of-the-art speech recognition with sequence-to-sequence models. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4774–4778 (2018)
11. Chowdhury, A., Vig, L.: An efficient end-to-end neural model for handwritten text recognition. In: British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018. p. 202. BMVA Press (2018)
12. Coquenet, D., Chatelain, C., Paquet, T.: Dan: a segmentation-free document attention network for handwritten document recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(7), 8227–8243 (2023)

13. Dhiaf, M., Rouhou, A.C., Kessentini, Y., Salem, S.B.: Msdoctr-lite: A lite transformer for full page multi-script handwriting recognition. Pattern Recognition Letters **169**, 28–34 (2023)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021)
15. Edirisooriya, S., Dong, H.W., Mcauley, J., Berg-Kirkpatrick, T.: An Empirical Evaluation of End-to-End Polyphonic Optical Music Recognition. In: Proceedings of the 22nd International Society for Music Information Retrieval Conference. pp. 167–173. ISMIR (Oct 2021)
16. Foscarin, F., Jacquemard, F., Fournier-S'niehotta, R.: A diff procedure for music score files. In: 6th International Conference on Digital Libraries for Musicology. pp. 58–64 (2019)
17. Good, M., Actor, G.: Using MusicXML for file interchange. Web Delivering of Music, International Conference on **0**, 153 (2003)
18. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd International Conference on Machine Learning. pp. 369–376. ACM (2006)
19. Jan Hajič, j., Pecina, P.: The MUSCIMA++ Dataset for Handwritten Optical Music Recognition. In: 14th International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, November 13 - 15, 2017. pp. 39–46. Dept. of Computer Science and Intelligent Systems, Graduate School of Engineering, Osaka Prefecture University, IEEE Computer Society, New York, USA (2017)
20. Hankinson, A., Roland, P., Fujinaga, I.: The Music Encoding Initiative as a Document-Encoding Framework. Proceedings of the 12th International Society for Music Information Retrieval Conference (2011)
21. Huron, D.: Humdrum and kern: Selective feature encoding. Beyond MIDI (1997)
22. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Ocr-free document understanding transformer. In: European Conference on Computer Vision (ECCV) (2022)
23. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11976–11986 (June 2022)
24. Martinez-Sevilla, J.C., Ríos-Vila, A., Castellanos, F.J., Calvo-Zaragoza, J.: A holistic approach for aligned music and lyrics transcription. In: Document Analysis and Recognition - ICDAR 2023 - 17th International Conference, San José, CA, USA, August 21-26, 2023, Proceedings, Part I. Lecture Notes in Computer Science, vol. 14187, pp. 185–201. Springer (2023)
25. McLeod, A., Steedman, M.: Evaluating automatic polyphonic music transcription. In: International Society for Music Information Retrieval Conference (ISMIR). pp. 42–49 (2018)
26. Pugin, L., Zitellini, R., Roland, P.: Verovio - A library for Engraving MEI Music Notation into SVG. In: International Society for Music Information Retrieval (jan 2014)
27. Rebelo, A., Fujinaga, I., Paszkiewicz, F., Marcal, A.R., Guedes, C., Cardoso, J.S.: Optical music recognition: state-of-the-art and open issues. International Journal of Multimedia Information Retrieval **1**(3), 173–190 (2012)

28. Ríos-Vila, A., Iñesta, J.M., Calvo-Zaragoza, J.: On the use of transformers for end-to-end optical music recognition. In: Pattern Recognition and Image Analysis - 10th Iberian Conference, IbPRIA 2022, Aveiro, Portugal, May 4-6, 2022, Proceedings. Lecture Notes in Computer Science, vol. 13256, pp. 470–481. Springer (2022)
29. Ríos-Vila, A., Rizo, D., Iñesta, J.M., Calvo-Zaragoza, J.: End-to-end optical music recognition for pianoform sheet music. Int. J. Document Anal. Recognit. **26**(3), 347–362 (2023)
30. Román, M.A., Pertusa, A., Calvo-Zaragoza, J.: A Holistic Approach to Polyphonic Music Transcription with Neural Networks. In: Proceedings of the 20th International Society for Music Information Retrieval Conference. pp. 731–737. Delft, The Netherlands (2019)
31. Sapp, C.S.: Verovio humdrum viewer. Proceedings of Music Encoding Conference (MEC), Tours, France (2017)
32. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(11), 2298–2304 (2017)
33. Singh, S.S., Karayev, S.: Full Page Handwriting Recognition via Image to Sequence Extraction. In: Proceedings of the 16th International Conference on Document Analysis and Recognition. pp. 55–69. Lausanne, Switzerland (2021)
34. Torras, P., Baró, A., Kang, L., Fornés, A.: On the Integration of Language Models into Sequence to Sequence Architectures for Handwritten Music Recognition. In: Proceedings of the 22nd International Society for Music Information Retrieval Conference. pp. 690–696. ISMIR (2021)
35. Torras, P., Biswas, S., Fornés, A.: The common optical music recognition evaluation framework (2023)
36. Tuggener, L., Emberger, R., Ghosh, A., Sager, P., Satyawan, Y.P., Montoya, J., Goldschagg, S., Seibold, F., Gut, U., Ackermann, P., et al.: Real world music object recognition (2024)
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
38. Zhang, J., Du, J., Zhang, S., Liu, D., Hu, Y., Hu, J., Wei, S., Dai, L.: Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition. Pattern Recognition **71**, 196–206 (2017)