# Handwritten Historical Music Recognition by Sequence-to-Sequence with Attention Mechanism

Arnau Baró*, Carles Badal† and Alicia Fornés*
*Computer Vision Center - Computer Science Department
†Art and Musicology Department
Universitat Autònoma de Barcelona, Bellaterra, Catalonia, Spain
Email: abaro@cvc.uab.cat, carles.badal@uab.cat, afornes@cvc.uab.es

***Abstract*—Despite decades of research in Optical Music Recognition (OMR), the recognition of old handwritten music scores remains a challenge because of the variabilities in the handwriting styles, paper degradation, lack of standard notation, etc. Therefore, the research in OMR systems adapted to the particularities of old manuscripts is crucial to accelerate the conversion of music scores existing in archives into digital libraries, fostering the dissemination and preservation of our music heritage. In this paper we explore the adaptation of sequence-to-sequence models with attention mechanism (used in translation and handwritten text recognition) and the generation of specific synthetic data for recognizing old music scores. The experimental validation demonstrates that our approach is promising, especially when compared with long short-term memory neural networks.**

## I. Introduction

The recognition of music scores is a classical research field within the document image analysis and recognition community. Optical Music Recognition (OMR) [1], [2], [3] consists in converting images of music scores into a digital format, such as MEI, MusicXML, MIDI, etc. Thus, OMR helps to ease the edition of scores never edited, renewal of old scores, production of audio files, transposing a music score into other clef or key signature, etc.

OMR has lately reached a very good performance on scanned printed music scores, especially for monophonic scores with low density of symbols. However, the recognition of handwritten scores is still a challenge due to the variabilities in the handwriting styles. Moreover, in the case of old and historical scores, the difficulties increase due to paper degradation, the frequent appearance of touching elements (e.g. lyrics and music symbols often overlap), or even the lack of standard notation, where one can find music scores that do not follow current music notation rules. Besides, the availability of labelled datasets of old handwritten music scores is scarce, which hinder the training of deep-learning based architectures.

Nevertheless, a low OMR performance could be substituted by a manual transcription, but such task is very time consuming and requires a huge amount of human resources (for example, a music transcriber can devote 1-3 hours to transcribing one music page, depending on the density of music symbols). Although there are several projects [1],[2],[3],[4],[5] on ancient manuscript music cataloging, digitization and/or transcription, the immense amount and variability of existing music works in archives makes it impossible in practice to transcribe and disseminate the entire source and forces musicology to, most of the times, carry out strictly qualitative research. This problem, obvious at a global level, is equally self evident at the local one. In the case of European church music, for example, in a significant parish or cathedral, which will probably have more than 500 years of history of handwritten music behind, one can find more than 2000 records of different works by various composers. This can easily mean tens of thousands of handwritten score pages in one single church. To give an example of the magnitude of this problem, we should consider that in a city like Barcelona there are four such musical chapels. Although there are few exceptions, their documentation is almost completely preserved in their archives, so the amount of music to be transcribed and analyzed is overwhelming for a manual process.

Therefore, the research in OMR systems adapted to the particularities of old music scores is crucial to accelerate the process from its discovery to its digital transcription, enabling researchers to analyze, publicize and divulge unknown composers and compositions that traditional methods are forced to neglect. This paradigm shift from traditional musicological research -which is usually focused on the aesthetic assessment and compositional characteristics of a certain number of composers- far from opposing it, would become a fundamental tool to complement these studies. This would provide a much more accurate overview of the local characteristics of each music and its relation to other geographical contexts (transmission, influences, circuits, etc.).

For the above reasons, in this paper we propose an OMR system for old handwritten scores. Our method is based on sequence-to-sequence models with an attention mechanism, which have been successfully applied to translation and handwritten text recognition. Also, and since the lack of available transcribed scores for training deep learning systems pose a

[1] SIMSSA: https://simssa.ca/

[2] Hispamus: https://grfia.dlsi.ua.es/hispamus/

[3] IFMuC: http://pagines.uab.cat/ifmuc/es

[4] RISM: http://www.risme.info/home.html

[5] MDC: http://mdc.csuc.cat/cdm/search/collection/musicatedra!partiturBC

978-1-7281-9966-5/20/$31.00 ©2020 IEEE
DOI 10.1109/ICFHR2020.2020.00046

challenge, we also generate specific synthetic data that emulates the particularities of old scores (e.g. lyrics touch the stave or even the musical symbols.). The experiments demonstrate the suitability of our approach, especially when compared to Long Short-Term Memory Recurrent Neural Networks.

The contributions are: 1) The adaptation of a sequence-to-sequence model with attention for historical music recognition. 2) A novel synthetic data generation, emulating old handwritten scores. 3) The historical handwritten music dataset is made available [6].

## II. Related Work

In this section we describe the most relevant approaches in Optical Music Recognition related to our work. First, we overview the traditional approaches for OMR, and then, we review the recent deep-learning based OMR approaches.

For decades, computer vision and pattern recognition have tackled the problem of OMR through traditional techniques. For example, and since monophonic scores follow a sequence, Hidden Markov models have been applied [4], [5]. Other works are based on symbol segmentation and recognition [6], [7], [8]. Since errors or ambiguities are frequent, grammars or syntactic rules[9], [10] are used to minimize them.

In the last years the OMR performance has significantly improved thanks to deep learning architectures. We can mention the sequence to sequence model (without attention) of Van der Wel and Ullrich[11], the long short-term memory recurrent neural networks (BLSTMs) of Calvo-Zaragoza *et.al.*[12], or the segmentation and classification models proposed by Wen *et.al.*[13]. However all the above methods are applied to printed music scores, which are easier to recognize. In addition, labelled printed score datasets are available (or easy to generate) for training those models.

There is little research for handwritten scores, and the few existing approaches are focused on Western music notation, such as the well-known MUSCIMA++ dataset [14], [15]. Some methods have been applied to this particular dataset. For example, Pacha *et.al.*[16] detect music primitives through deep convolutional neural networks (CNNs), Baró *et.al.*[17] combine CNNs and BLSTMs, and Tuggener *et.al.* use ResNets[18].

Recognizing historical documents implies dealing with few labelled data. One solution is to train with synthetic data and refine with real handwritten one, as in [17]. Another solution is to explore unsupervised domain adaptation techniques, as proposed in [19] for handwritten text recognition. Although Adversarial Networks for domain adaptation have been explored in *et. al* for recognizing music symbols, their application to music score recognition is still an open problem.

Concerning old scores, the are some works for mensural notation. For example, Calvo-Zaragoza *et.al.* propose to use hidden Markov models and N-gram language models [20], whereas in [21] they opt for convolutional neural network with a recurrent neural network and language models. Pacha *et. al* [22] use a R-CNN with Inception ResNet V2 for music object detection i mensural scores. However, mensural notation is rather simple compared to western notation, so the research in OMR for historical documents dated from 17th-18th is still necessary. Moreover, the adaptation of the groundtruth to train an object detection method is very tedious. Methods like [22], not only need an expert to transcribe the music score, but also to annotate in the image score the position of each symbol. Note that such a detailed annotation is not needed for training recurrent neural networks, including sequence-to-sequence methods.

For these reasons, and inspired by the succeed of sequence to sequence models that incorporate attention mechanisms, we explore their adaptation to the recognition of such scores. As far as we know, this is the first OMR method based on sequence-to-sequence (seq2seq) model with attention mechanism adapted for historical music score recognition. Besides, we believe that historical handwritten scores (including synthetic datasets that emulate the real ones), are required for further research in the recognition of this kind of documents.

## III. Baseline: Long Short-Term Memory Model

Before describing our Seq2Seq OMR architecture, we first describe our baseline, based on Long Short-Term Memory Recurrent Neural Networks [12], [17]. Note that our baseline is based on recurrent models because of the sequentiality of monopohonic music staves.

Although long short-term memory networks are capable of directly treating the raw image, the performance improves when adding a Convolutional Neural network (CNN) as a feature extractor. Thus, our baseline is composed of a Convolutional Neural Network and a bidirectional Long Short-Term Memory neural network (BLSTM) with Connectionist Temporal Classification (CTC) loss. Figure 1 shows the model architecture. The modules are described next.

- **Convolutional Network:** In this step we extract the features that will be used in the next steps. The convolutional network is composed of the first three layers of the ResNet18 [23], consisting of convolution, batch normalization and rectified linear unit activation.
- **Bidirectional LSTM:** The BLSTM gets as input the features from the CNN. We use a LSTM to reduce the vanish gradient problem since LSTMs can remember information for longer time. We use bi-directonal LSTMs to increase context information (from left and right sides in the image) and reduce the number of ambiguities.
- **Fully connected layers:** The results obtained by the BLSTM network are passed to a fully connected layer to return the final result.
- **Connectionist Temporal Classification:** This step helps to evaluate the output and check that the predictions are correct. As a loss function we use the Connectionist Temporal Classification (CTC) [24], which is trained using Stochastic Gradient Descent (SGD) optimizer with Momentum.

[6] http://www.cvc.uab.es/people/abaro/datasets.html

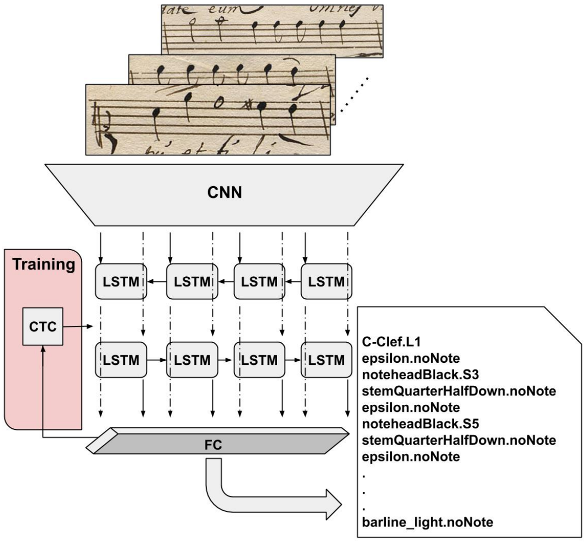


Fig. 1. Convolutional Neural Network and Bidirectional Long Short-Term Memory model.

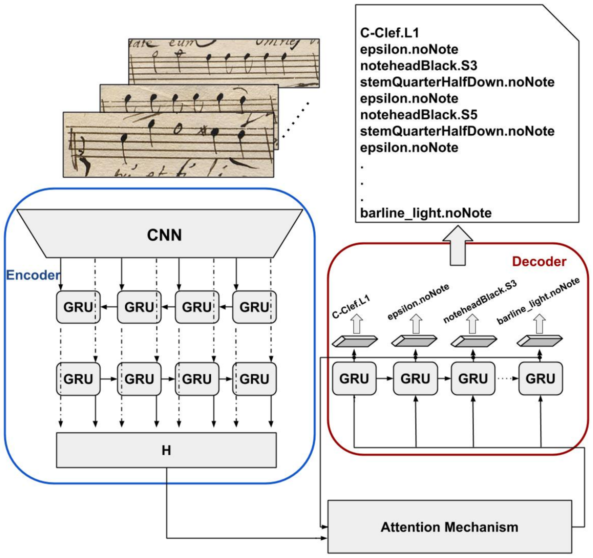


Fig. 2. Sequence-to-sequence model with attention mechanism.

## IV. Architecture: Sequence to Sequence model

As explained before, music scores are written on staves following a sequence, so our approach is also based on recurrent models. Concretely, our method is based on the sequence-to-sequence (seq2seq) text recognition method [25], and adapted to music scores.

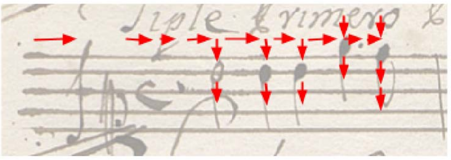
### A. Sequence-to-Sequence model with attention mechanism

This methodology makes use of an attention-based encoder-decoder framework. Thus, our model consists of 3 components, the encoder, the attention mechanism and the decoder. Figure 2 depicts our proposed architecture for optical music recognition.

- **Encoder.** Given an input image, the encoder extracts high-level features encoding the contents of the image. These features will be later used to obtain the contents of the image in a machine readable format. In this work, the proposed encoder is implemented with a VGG-19-BN network [26] with pre-trained weights from ImageNet. Moreover, the last max-pooling is removed. Finally, the VGG features are reshaped into a two-dimensional feature map that will be further used as the input to a multi-layered Bidirectional Gated Recurrent Unit (BGRU) which provides extra positional information.
- **Attention Mechanism.** As an attention module, we use a location-based attention as proposed by Chorowski *et.al.* [27]. This takes into account the location information explicitly for a better alignment. Otherwise, content-based attention expects the location information to be coded in the extracted features in order to differentiate the different representations of the features of the same symbol in different positions. The attention mechanism is in charge of aligning our feature representations with our decoding steps.
- **Decoder.** Finally, the decoder module is formed by a one-directional multi-layered GRU. The decoder provides the recognized symbols in several steps following a sequential order. At each time step the decoder GRU receives the concatenation of its previous embedding vector (in step $i-1$) and the current context vector (defined by the encoded features and our attention mechanism) in order to predict a new symbol. Moreover, to enhance the decoder we have used, on the one hand a multinominal which takes into account several decoding paths to obtain the final prediction and, on the other hand, label smoothing that allows a better generalization preventing over-confident predictions.

### B. Adaptation to music scores

Recurrent methods, including sequence to sequence (seq2seq) models, have shown good performance when applied to handwritten text recognition (HTR). But the recognition of music is much more complex than text. The main reason is that the nature of text is one-dimensional: a sequence of characters. In music, however, we have to deal with two-dimensional sequences. First, music notes are composed of rhythm and melody. Second, some elements, such as ornaments or articulations, usually appear above or below notes. In addition, groups of notes (e.g. chords) or even other symbols (e.g. slurs or ties group notes, providing musicality to the work) appear at the same instant of time. Furthermore, music notation allows notes (e.g. 8th, 16th, etc.) to be written isolated or together (grouped using beams). Besides, an 8th note can have a stem looking up or down. Although isolated or compound symbols could



C-Clef.L1~epsilon~timeSig_common~epsiln~halfRest~epsilon ~noteheadHalf.L4~stemDown~epsilon~noteheadBlack.L4~ stemDown~epsilon~noteheadBlack.L4~stemDown~epsilon~ noteheadBlack.S5~stemDown~epsilon~dot~epsilon~ noteheadBlack.L5~flag8thDown~epsilon~barline_light

Fig. 3. Example of the labelling of the groundtruth, creating a 1D sequence. The transcription is written reading each measure from left to right, and from top to bottom if the symbol is divisible into primitives.

seem the same to a non-expert user, a musician or musicologist makes differences (especially during the interpretation).

For the above reasons, we need to adapt the seq2seq model designed for text recognition to the particularities of music scores. It is true that, since music elements are located in a 2D space on the staff, these elements could be represented using a graph, such as in [15]. Thus, one possible solution is to treat the problem as a graph serialization task, which can be defined as the conversion of a 2D graph into a 1D string. In our case, music scores have been annotated at primitive level (i.e., note heads, stems, beams, flags, rests, etc.), so the output of our architecture will be a sequence of 1D music primitives. Therefore, we can solve the problem by defining a reading order, from left to right and from top to bottom, as illustrated in Figure 3. In a horizontal lecture, when we move one step in the staff (the position of the horizontal arrow), we use the symbol epsilon ($\epsilon$) as a separator. Contrary, if the vertical primitives belong together (e.g. same symbol), they appear at the same time step, as denoted using vertical arrows.

## V. Dealing with the lack of data

Deep learning methods are data hungry, i.e. they need a lot of labelled data to train. Since the amount of historical labelled data is scarce, we must look for alternatives. Therefore, we have generated two synthetic datasets using Lilypond [7]. Each one contains about 30,000 bar images, and are divided into 60% train, 20% validation and 20% test. These two datasets are complementary: one simulates the particularities of historical scores, whereas the other provides examples of a large diversity of symbols, including polyphony. These datasets are described next:

- Old synthetic (monophonic): This dataset tries to imitate the texture and degradation of the paper of historical scores adding a background. Also, the type and diversity of symbols is limited, similar to the historical scores used in the experiments. Figure 4 shows a measure from the old synthetic dataset.
- Modern synthetic (polyphonic): This dataset contains polyphonic symbols written in one staff, i.e. stacks of notes meant to be played all at once, such as chords. Figure 5 shows a measure from this dataset. This data will allow our model to generalize to any kind of historical music score, either monophonic or poliphonic.

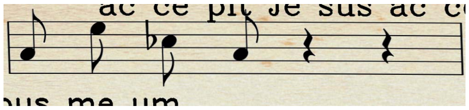


Fig. 4. Example image from the old synthetic dataset.

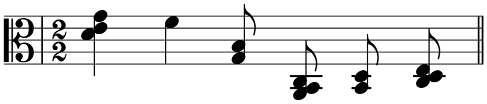


Fig. 5. Example image from the modern (polyphonic) dataset.

These synthetic datasets are used to pretrain our system. We train our model using curriculum learning [28] for improving the performance. At the beginning, we train with few real historical measures and lots of synthetic ones. After $n$ epochs, we increment the number of historical measures and decrease the synthetic ones. At the end, the training data is 100% historical.

## VI. Experimental Validation

This section experimentally validates our approach.

### A. Historical Dataset

The historical data used in the experimental validation is a motet composed by Pau Llinàs, a catalan musician who worked as chapel master in Santa Maria del Pi of Barcelona between 1709 and 1749. Most probably, the work was written around that time since Llinàs spent most of his life as a chapel master and, thus, composing professionally in this chapel [29]. This religious motet (psalm number 148: Laudate Domine - Praise the Lord) is preserved in 12 separated parts, instead of a full score (most common in this time period). This motet actually belongs to the *Fons Musical de la Catedral de Barcelona* and has been incorporated at the *Biblioteca Nacional de Catalunya* (BNC) catalogue [8].

For our experimental validation, we have manually labeled 40 music staves, containing 245 measure images. These are divided into 147 measures for training, 49 for validation and 49 for test. Figure 6 shows a page from this historical dataset, illustrating their main difficulties. On the first staff we can observe that the lyrics are touching the staff and in some cases even the symbols. Also, at the end of the third staff or at the beginning of the fifth staff the are ink stains.

### B. Results on historical music scores

We have used the Symbol Error rate (SER) metric to evaluate our approach. This has already been used in several

[7] https://lilypond.org/

[8] http://mdc.csuc.cat/cdm/landingpage/collection/musicatedra

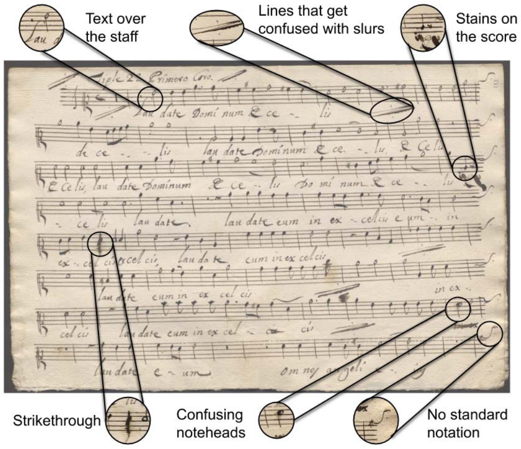


Fig. 6. Page example from the historical dataset.

music recognition publications as a substitute for the well-known Character/Word Error Rate in text recognition. The SER is defined by

$$SER = \frac{S + D + I}{N}$$

where S denotes the substitutions, D the deletions and I the insertions and N the number of symbols in the groundtruth. As it is a metric that evaluates the error, the lower the better. Next, we evaluate our sequence-to-sequence architecture, and compare with the baseline described in Section III.

*1) Quantitative Results:* Table I shows the comparison between our Seq2Seq model and the baseline model using Convolutional Neural Network and Bidirectional Long Short Term Memory Neural Networks with Connectionist time classification (CNN+BLSTM) [17]. The first column indicates the method used, the second column indicates which dataset has been used for training and the third column indicates the percentage of Symbol Error Rate (SER). From the Table I, we can observe that, in all setups, the Seq2Seq outperforms the BLSTMs by a large margin. As expected, the best result is obtained when training with real historical data, even though the amount of real labelled data is very low. We also observe that training with the modern synthetic dataset leads to a very low performance. However, if we train with the old synthetic dataset, we can reduce the SER by 20 points. Finally, if we combine both synthetic datasets (50% modern and 50% old), there is more varied data during training, so the methods obtain a slightly better SER.

Given that the best results are obtained using our proposed Seq2Seq approach and combining all synthetic data (modern+old), we have performed a second experiment considering the scenario where both real and synthetic data are available for training. As explained in Section V, we use curriculum learning to train with easy examples first, and gradually incorporate more difficult ones. Table II shows how we have modified the percentage of historical and synthetic data at training time. The first four columns of the table shows the percentage of measures used for training and validation for each dataset, whereas the last column shows the SER on the real historical test set. We start the first epochs (see the first row) with few historical data and a high percentage of synthetic data. Every 10 epochs we augment the percentage of real data, while decreasing the amount of synthetic one. To minimize the overfitting problem, and given that the amount of synthetic scores are much higher than the historical ones, in the validation set, we do exactly the opposite: we have started with a high percentage of historical data, which is progressively decreased during training. At the end of the training phase, the training set has mainly historical data whereas the validation set has mainly synthetic one.

TABLE I
QUANTITATIVE RESULTS COMPARING THE CNN+BLSTM MODEL AND OUR SEQUENCE-TO-SEQUENCE (SEQ2SEQ) MODEL.

| Architecture | Dataset Train | Test SER (%) |
|---|---|---|
| CNN+ BLSTM | Historical | 56.20 |
| | Modern Synthetic | 96.20 |
| | Old Synthetic | 75.20 |
| | Modern + Old Synthetic | 74.40 |
| Seq2Seq | Historical | 40.39 |
| | Modern Synthetic | 83.80 |
| | Old Synthetic | 61.89 |
| | Modern + Old Synthetic | 60.69 |

TABLE II
RESULTS USING OUR SEQ2SEQ MODEL WITH CURRICULUM LEARNING. WE SHOW THE AMOUNT OF DATA OF EACH KIND USED DURING TRAINING.

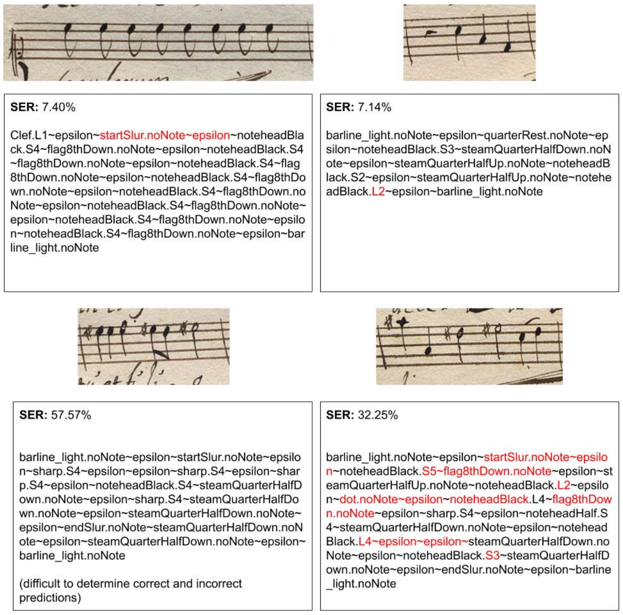
| Percentage in Training (%) | | Percentage in Validation (%) | | Test SER (%) |
|---|---|---|---|---|
| Historical | Modern+Old Syn. | Historical | Modern+Old Syn. | |
| 10 | 90 | 100 | 0 | 60.03 |
| 40 | 60 | 70 | 30 | 66.20 |
| 60 | 40 | 50 | 50 | 43.38 |
| 80 | 20 | 30 | 70 | 37.86 |
| 90 | 10 | 20 | 80 | 34.56 |
| 100 | 0 | 10 | 90 | **31.79** |

From the results reported in Table II, we can conclude that training with real and synthetic data highly benefits the overall system performance. Indeed, the obtained SER of 31.79% is significantly lower than the SER of 40.39% that was obtained when training with historical data only, as shown in Table I.

*2) Qualitative Results:* Figure 7 shows some qualitative results from the sequence-to-sequence model. We have highlighted in red some common mistakes. In the first example, we see that the lyric is often confused by slurs. Some times the shape between the stem and the flag is also confused. The position of a notehead can be frequently displaced *i.e.* a note in the space 3(S3) could be wrongly predicted as to be in line 3(L3) or line 4(L4).

## VII. CONCLUSIONS AND FUTURE WORK

In this work we have proposed a sequence-to-sequence architecture with attention mechanism for recognizing historical handwritten music scores. We have experimentally demonstrated that our model obtains promising results, especially compared to Bidirectional Long Short-Term Memory



```
SER: 7.40%

Clef.L1~epsilon~startSlur.noNote~epsilon~noteheadBla
ck.S4~flag8thDown.noNote~epsilon~noteheadBlack.S4
~flag8thDown.noNote~epsilon~noteheadBlack.S4~flag
8thDown.noNote~epsilon~noteheadBlack.S4~flag8thDo
wn.noNote~epsilon~noteheadBlack.S4~flag8thDown.no
Note~epsilon~noteheadBlack.S4~flag8thDown.noNote~
epsilon~noteheadBlack.S4~flag8thDown.noNote~epsilo
n~noteheadBlack.S4~flag8thDown.noNote~epsilon~bar
line_light.noNote
```

```
SER: 7.14%

barline_light.noNote~epsilon~quarterRest.noNote~ep
silon~noteheadBlack.S3~steamQuarterHalfDown.noN
ote~epsilon~steamQuarterHalfUp.noNote~noteheadB
lack.S2~epsilon~steamQuarterHalfUp.noNote~notehe
adBlack.L2~epsilon~barline_light.noNote
```

```
SER: 57.57%

barline_light.noNote~epsilon~startSlur.noNote~epsilo
n~sharp.S4~epsilon~epsilon~sharp.S4~epsilon~shar
p.S4~epsilon~noteheadBlack.S4~steamQuarterHalfD
own.noNote~epsilon~sharp.S4~steamQuarterHalfDo
wn.noNote~epsilon~steamQuarterHalfDown.noNote~
epsilon~endSlur.noNote~steamQuarterHalfDown.noN
ote~epsilon~steamQuarterHalfDown.noNote~epsilon~
barline_light.noNote

(difficult to determine correct and incorrect
predictions)
```

```
SER: 32.25%

barline_light.noNote~epsilon~startSlur.noNote~epsilo
n~noteheadBlack.S5~flag8thDown.noNote~epsilon~st
eamQuarterHalfUp.noNote~noteheadBlack.L2~epsilo
n~dot.noNote~epsilon~noteheadBlack.L4~flag8thDow
n.noNote~epsilon~sharp.S4~epsilon~noteheadHalf.S
4~steamQuarterHalfDown.noNote~epsilon~notehead
Black.L4~epsilon~epsilon~steamQuarterHalfDown.no
Note~epsilon~noteheadBlack.S3~steamQuarterHalfD
own.noNote~epsilon~endSlur.noNote~epsilon~barline
_light.noNote
```

Fig. 7. Qualitative results. Mistakes shown in red color. From left to right and from up to down. First image: a slur is predicted instead of lyrics. Second image: the pitch of one notehead is confused. Third and fourth images: multiple mistakes because lyrics are too close to music symbols.

networks. We have also shown that the generation of specific synthetic data that simulates old scores is beneficial. In this sense, we have demonstrated that curriculum learning can gain leverage from the combination of real and synthetic data, improving the overall performance.

Nevertheless, the difficulties of historical scores in terms of paper degradation, touching lyrics and music symbols as well as the lack of annotated data still pose a challenge for optical music recognition. Concerning this last issue, we believe that the research community can benefit from our three labelled datasets, which will be publicly available.

As future work we plan to tackle polyphonic scores and improve our Seq2Seq architecture by exploring the incorporation of language models and domain adaptation techniques.

## Acknowledgment

This work has been partially supported by the Spanish projects RTI2018-095645-B-C21 and PGC2018-098986-B-C33, the Ramon y Cajal Fellowship RYC-2014-16831, the CERCA Program/ Generalitat de Catalunya, and the FI fellowships AGAUR 2020 FI_B2 00021 and FI_B2 00149 (with the support of the Secretaria d'Universitats i Recerca of the Generalitat de Catalunya and the Fons Social Europeu). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

## References

[1] A. Rebelo, I. Fujinaga, F. Paszkiewicz, A. R. S. Marçal, C. Guedes, and J. S. Cardoso, "Optical music recognition: state-of-the-art and open issues." *IJMIR*, vol. 1, no. 3, pp. 173–190, 2012.

[2] A. Fornés and G. Sánchez, "Analysis and recognition of music scores," in *Handbook of Document Image Processing and Recognition*. Springer-Verlag London, 2014, pp. 749–774.

[3] J. Calvo-Zaragoza, J. Hajic Jr, and A. Pacha, "Understanding optical music recognition," *arXiv preprint arXiv:1908.03608*, 2019.

[4] L. Pugin, "Optical music recognition of early typographic prints using hidden markov models," in *ISMIR*, 2006, pp. 53–56.

[5] L. Pugin, J. A. Burgoyne, and I. Fujinaga, "Map adaptation to improve optical music recognition of early music documents using hidden markov models," in *ISMIR*, 2007, pp. 513–516.

[6] A. Fornés, J. Lladós, G. Sánchez, and D. Karatzas, "Rotation invariant hand drawn symbol recognition based on a dynamic time warping model," *IJDAR*, vol. 13, no. 3, pp. 229–241, 2010.

[7] A. Rebelo, G. Capela, and J. S. Cardoso, "Optical recognition of music symbols: A comparative study," *IJDAR*, vol. 13, no. 1, pp. 19–31, 2010.

[8] S. Escalera, A. Fornés, O. Pujol, P. Radeva, G. Sánchez, and J. Lladós, "Blurred Shape Model for binary and grey-level symbol recognition," *Pattern Recognition Letters*, vol. 30, no. 15, pp. 1424–1433, 2009.

[9] B. Coüasnon and B. Rétif, "Using a grammar for a reliable full score recognition system," 1995.

[10] A. Baró, P. Riba, and A. Fornés, "Towards the recognition of compound music notes in handwritten music scores," in *ICFHR*, 2016, pp. 465–470.

[11] E. van der Wel and K. Ullrich, "Optical music recognition with convolutional sequence-to-sequence models," in *ISMIR*, 2017, pp. 731–737.

[12] J. Calvo-Zaragoza and D. Rizo, "End-to-end neural optical music recognition of monophonic scores," *Applied Sciences*, vol. 8, pp. 1–23, 2018.

[13] C. Wen, A. Rebelo, J. Zhang, and J. Cardoso, "A new optical music recognition system based on combined neural network," *Pattern Recognition Letters*, vol. 58, pp. 1 – 7, 2015.

[14] A. Fornés, A. Dutta, A. Gordo, and J. Lladós, "CVC-MUSCIMA: a ground truth of handwritten music score images for writer identification and staff removal," *IJDAR*, vol. 15, no. 3, pp. 243–251, 2012.

[15] J. Hajič jr. and P. Pecina, "The MUSCIMA++ Dataset for Handwritten Optical Music Recognition," in *ICDAR*, 2017, pp. 39–46.

[16] A. Pacha, K.-Y. Choi, B. Coüasnon, Y. Ricquebourg, R. Zanibbi, and H. M. Eidenberger, "Handwritten music object detection: Open issues and baseline results," in *DAS*, 2018, pp. 163–168.

[17] A. Baró, P. Riba, J. Calvo-Zaragoza, and A. Fornés, "From optical music recognition to handwritten music recognition: A baseline," *Pattern Recognition Letters*, vol. 123, pp. 1 – 8, 2019.

[18] L. Tuggener, I. Elezi, J. Schmidhuber, and T. Stadelmann, "Deep watershed detector for music object recognition," in *ISMIR*, 2018, pp. 271–278.

[19] L. Kang, M. Rusiñol, A. Fornés, P. Riba, and M. Villegas, "Unsupervised adaptation for synthetic-to-real handwritten word recognition," in *WACV*, 2020.

[20] J. Calvo-Zaragoza, A. H. Toselli, and E. Vidal, "Handwritten music recognition for mensural notation: Formulation, data and baseline results," in *ICDAR*, 2017, pp. 1081–1086.

[21] 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.

[22] A. Pacha and J. Calvo-Zaragoza, "Optical music recognition in mensural notation with region-based convolutional neural networks," in *ISMIR*, 2018, pp. 240–247.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[24] A. Graves, S. Fernández, and F. Gomez, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006, pp. 369–376.

[25] L. Kang, J. Toledo, P. Riba, M. Villegas, A. Fornés, and M. Rusiñol, "Convolve, attend and spell: An attention-based sequence-to-sequence model for handwritten word recognition," in *GCPR*, 2019, pp. 459–472.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv 1409.1556*, 09 2014.

[27] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *NIPS*, p. 577–585, 2015.

[28] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *ICML*, 2009, pp. 41–48.

[29] C. Badal, "Pau llinàs i la vida a santa maria del pi (1711-1749): biografia i inventari," *Revista Catalana de Musicologia*, vol. IX, pp. 153–173, 2017.