ON THE PERFORMANCE OF OPTICAL MUSIC RECOGNITION IN THE ABSENCE OF SPECIFIC TRAINING DATA

Juan C. Martinez-Sevilla[1] **Adrian Rosello**[1]
David Rizo[1,2] **Jorge Calvo-Zaragoza**[1]
[1] U. I. for Computing Research, University of Alicante, Spain
[2] Instituto Superior de Enseñanzas Artísticas de la Comunidad Valenciana, Spain
adrian.rosello@ua.es, {jcmartinez, drizo, jcalvo}@dlsi.ua.es

ABSTRACT

Optical Music Recognition (OMR) has become a popular technology to retrieve information present in musical scores in conjunction with the increasing improvement of Deep Learning techniques, which represent the state-of-the-art in the field. However, its effectiveness is limited to cases where the target collection is similar in musical context and graphical appearance to the available training examples. To address this limitation, researchers have resorted to labeling examples for specific neural models, which is time-consuming and raises questions about usability. In this study, we propose a holistic and comprehensive study for dealing with new music collections in OMR, including extensive experiments to identify key aspects to have in mind that lead to better performance ratios. We resort to collections written in Mensural notation as a specific use case, comprising 5 different corpora of training domains and up to 15 test collections. Our experiments report many interesting insights that will be important to create a manual of best practices when dealing with new collections in OMR systems.

1. INTRODUCTION

Manual sheet music transcription is a tedious process, prone to errors, and generally requires professionals with precise knowledge of the type of notation and/or music at issue. The alternative to this manual digitization of content is to resort to cutting-edge technology based on artificial intelligence, which performs an automated reading of documents. This technology is known as Optical Music Recognition (OMR).

OMR has been an active research area for decades [1], although the field progressed slowly [2]. Recently, the use of modern machine learning techniques, namely Deep Learning, has led to a paradigm shift that has partially unlocked this situation [3, 4]. Indeed, it has been shown that current OMR technologies, despite the fact that they are not yet fully mature, are usually a better alternative than performing the entire transcription by hand [5].

Concerning the machine learning methods, the related literature reports that the models provide sufficient precision when the collections to be transcribed are from the same graphic and content domain as the corpus used to train them. This, however, makes it difficult to transfer technology to new collections, since it is not always possible, desirable, or efficient to invest resources in annotating a small portion of the target collection. Although it is naive to assume the availability of training sets from the same domain as a given target collection, in the current data era we can assume to have at least a series of labeled collections, even with different graphic and musical characteristics. This, of course, can and should be used to improve the efficiency of fitting OMR models to new collections for which we do not have specific training sets.

In this paper, we report on a case study focused on Mensural notation to answer questions about the transferability of OMR models to new music collections. To our best knowledge, this work constitutes the first to analyze this issue in the field. We consider Mensural notation as the structuring experimental body because the OMR technology can be considered mature for this notation. Also, we have a significant number of labeled and unlabeled collections in this notation, which allows us to carry out an exhaustive study that is expected to lead to more generalizable conclusions. Specifically, we consider 5 labeled collections that will be used as training sets, along with their possible combinations, and up to 15 unlabeled collections as target.

The rest of the paper is structured as it follows: in Section 2, we provide some background to the topic; in Section 3, we present our methodology to analyze the question at issue; the experimental setup is described in Section 4, while the results and analysis are given in Section 5; finally, we conclude the paper in Section 6, while pointing out some interesting avenues for future work.

2. BACKGROUND

Recent advances in artificial intelligence, with extensive use of Deep Learning (DL) technologies, resulted in about successful approaches to OMR. Specifically, a holistic ap-

proach, also known as end-to-end formulation, which has been dominating the state of the art in other applications such as text or speech recognition [6, 7], is currently considered the reference model in OMR. The related literature includes many successful solutions of this type [8–10]. In this work, we resort to this approach as representative of the state of the art based on DL.

However, as introduced above, there is still no computational approach for creating a universal OMR system; *i.e*, one that is capable of dealing with any kind of collection. The underlying issue is an overly unsolved challenge in artificial intelligence [11]: DL works well if the problem is statistically regular and there is abundant training data to adequately and representatively learn such regularity. This is, unfortunately, quite difficult to expect when dealing with ancient documents. Instead of trying to solve the underlying problem of machine learning, we take a more practical path to provide a series of best practices to tackle the situation of target collections in the absence of specific training data successfully.

It is important to highlight that, in the OMR literature, there are very few works dedicated to studying the practical aspects of the technology. Pugin and Crawford [12] estimated through a quantitative evaluation the suitability of using the Aruspix machine-learning-based OMR system on a real collection. Furthermore, Alfaro-Contreras et al. [5] analyzed the benefits of using OMR in cases where the accuracy of the system was not perfect. Our work further contributes to this barely explored line of practical aspects for the application of OMR to real-world scenarios from the perspective of the available training data.

3. METHODOLOGY

The focus of the work is essentially experimental. We want to be able to answer specific questions about how to approach the generation of generalizable OMR models. Our objective is to reduce the uncertainty when facing the recognition of collections for which there is no specific training set.

To answer these questions, we will consider as a starting point the availability of N training sets that, even depicting the same musical notation (Mensural notation), differ in graphic characteristics. This will allow drawing more interesting conclusions about the synergy of using a heterogeneous set of training collections. To cover all possibilities, we create models from all possible combinations of these sets ($2^N - 1$ possibilities). Each of these possibilities will be directly evaluated on M test sets (not seen in any training case), also showing heterogeneous characteristics.

As previously mentioned, we will consider a deep end-to-end model as representative of the state of the art in OMR. Below we explain in more detail how this model works.

3.1 Learning framework

For the task, a Convolutional Recurrent Neural Network (CRNN) scheme is proposed for the end-to-end optical music transcription pipeline. The CRNN architecture consists of a block of convolutional layers that learns the relevant features from the input image (single staff), followed by a group of recurrent stages that model the temporal dependencies of the feature-learning block. Finally, a fully-connected network with a softmax activation is used to retrieve the posteriogram, which is decoded to obtain the predicted musical symbols. [1]

The Connectionist Temporal Classification (CTC) [13] training procedure is used to train the CRNN model using unsegmented sequential data. The training set $\mathcal{T}$ consists of pairs of single musical staff images x_i and their corresponding symbol sequence $\mathbf{z}_i$ in a symbol vocabulary Σ, with 261 units corresponding to the number of different symbols among the training sets. To use CTC as an end-to-end sequence labeling framework, an additional "blank" symbol is included in the vocabulary Σ'.

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

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

4. EXPERIMENTAL SETUP

In this section, we present our choices for the experimental design. First, we describe the considered evaluation metric. Then, we give more implementation details of the deep learning model. Finally, we present and describe the collections selected as train and target sets.

4.1 Evaluation

To evaluate the performance of the OMR model, we resort to the *Symbol Error Rate* (SER). This is computed as the average number of elementary editing operations (insertions, deletions, or substitutions) required to convert prediction $\hat{\mathbf{z}}_i$ into reference $\mathbf{z}_i$, normalized by the length of the latter.

In general, we are interested in computing the amount of effort it would take for a person to correct the remaining errors in the system. Since computing this human effort

[1] Understanding *musical symbol* as the conjunction of `glyph:position`, *i.e.*, `note_half:L2` (a glyph `note_half` present in the second staff line).

does not scale well in practice (it consumes huge amounts of resources), we believe that this metric is suitable to measure the transcription correctness. In addition, it is a metric that has been commonly applied in previous works on this subject (cf. Section 2).

4.2 Neural model configuration

The CRNN topology is based on the one used in the research [14], where the authors adopt a 4 convolutional layer block with batch normalization, Leaky ReLu activation, and max-pooling down-sampling. The feature maps extracted from the convolutional block are fed into two Bidirectional Long Short-Time Memory layers with 256 hidden units each and a dropout value of $d = 50\%$ followed by a fully-connected network with $|\Sigma'|$ units.

The models were trained with a batch size of 16 elements—note that in experiments where multiple training sets were used all the generated batches in the training process were balanced so the net didn't adjust to a certain corpus. The ADAM optimizer [15] was considered and a fixed learning rate of 10^{-3}. We iterate for 300 epochs, keeping the weights that minimize the SER metric in the validation partition with an early stopping policy of 30 epochs. Finally, all experiments were run using the Python language (v. 3.8.13) with the PyTorch framework (v. 1.13.0) on a single NVIDIA GeForce RTX 4090 card with 24GB of GPU memory.

4.3 Datasets

A set of 20 different white Mensural notation works has been collected for this work, consisting of pairs of staff images and their transcription into sequences of musical symbols. The pieces have been selected looking for diverse cases concerning printers or copyists, layouts, authors, the period in history, and extension. [2]

4.3.1 Training Datasets

For training, 4 different datasets were chosen from real collections, trying to cover as much variability as possible. When facing a new transcription project, it is usual that no training collection is similar or big enough for building a model to obtain reliable results from the automatic recognition process. In this scenario, the creation of synthetic training data from scratch is a valid alternative that will be evaluated in the work with the PRIMENS dataset. Therefore, we will add this synthetic collection to the set of training sets, resulting in 5 different collections. These training collections are described below.

- **CAPITAN**. The Capitan dataset contains 100 handwritten pages of ca. 17th-century manuscripts in late white Mensural notation extracted from the work with signature B59.850 in the Catedral del Pilar in Zaragoza [16].
- **SEILS**. The SEILS dataset contains 151 printed pages of the "Il Lauro Secco" collection corresponding to an anthology of 16th-century Italian madrigals in white Mensural notation [17].
- **GUATEMALA**. The Guatemala dataset presents 383 handwritten pages from a polyphonic choir book, part of a larger collection held at the "Archivo Histórico Arquidiocesano de Guatemala" [18].
- **MOTTECTA**. This dataset corresponds to the work "Mottecta (Mottecta Francisci Guerreri, que partim quaternis partim quinis alia senis alia octonis concinuntur vocibus, liber secundus dataset)", authored by Francisco Guerrero in the 16th-century and edited by Giacomo Vincenti in the 17th-century. This 297-printed mensural pages corpus has been obtained from the collection of mensural books of the Biblioteca Digital Hispánica. [3]
- **PRIMENS**. The Printed Images of Mensural Staves (PrIMenS) dataset is a synthetic corpus that tries to resemble low-quality real scans of printed mensural sources. It has been built from works composed by Agricola, Frye, and Ockeghem available in the Josquin Research Project [4]. Given polyphonic scores encoded in **kern [19] format, each voice is separated into a single file. In order to increase the variability, the original clefs are modified according to the instrument annotation in the voice. To obtain single staves, the whole work has been divided into a random number of measures from 3 to 18, and the resulting files have been converted into **mens [20] format. The corresponding agnostic encoding has been generated following the method described in [17]. The images have been obtained using the digital engraver Verovio [21] by applying random values to all the options in the allowed ranges. Finally, those images have been distorted to simulate real printed image scans by using a random sequence of graphical filters with the GraphicsMagick Image Processing. Additionally, this real-image simulation process has been complemented by composing randomly damaged old paper textures with distorted images.

To better understand the differences that might appear among these corpora, we provide a staff example from each corpus in Fig. 1.

4.3.2 Target Datasets

For the task of testing the suitability of each model, 15 datasets have been chosen. These corpora have been carefully and specifically labeled for this work, and are summarized in Table 1 and Fig. 2.

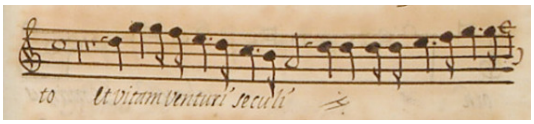
The printed sets have been extracted from the publicly available collection of Mensural books in the Biblioteca Digital Hispánica. [5] The handwritten collections are obtained from archive of Catedral del Pilar in Zaragoza [16].

[2] The whole set, along with a comprehensive description of the contents, can be found at `https://grfia.dlsi.ua.es/polifonia/ismir2023.html`.

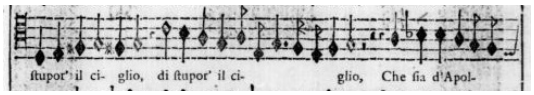
[3] `bdh.bne.es/bnesearch/detalle/bdh0000008932`

[4] `https://josquin.stanford.edu/` (accessed September 1st, 2022).

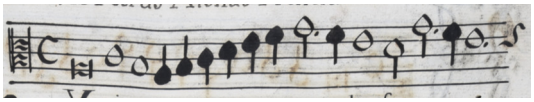
[5] `https://www.bne.es/es/catalogos/biblioteca-digital-hispanica` (accessed March 7th, 2023)



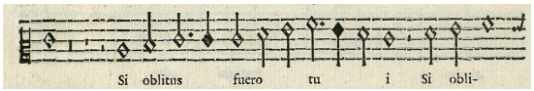
(a) CAPITAN



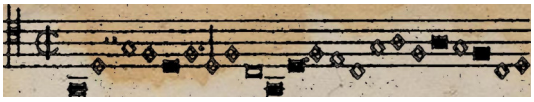
(b) SEILS



(c) GUATEMALA



(d) MOTTECTA



(e) PRIMENS

Figure 1: Samples of staves of the different training datasets employed.

Name (ID)	Number of staves	Printer
Amorosa (Amo)	224	H. of G. Scoto
Chansons (Cha)	173	A. Le Roy, R. Ballard
Dolci (Dol)	170	H. of G. Scoto
Lamentationes (Lam)	528	G. G. Carlino
Madrigali (Mad)	201	G. Scotto
Magnificat (Mag)	1361	Antonio Gardano
Missarum (Mis)	489	H. of G. Scoto
MusicaNova (Mus)	874	Antonio Gardano
Orlande (Orl)	259	A. Le Roy, R. Ballard
Responsoria (Res)	666	G. G. Carlino
Sacrarum (Sac)	460	Antonio Gardano
Villanelle (Vil)	59	G. G. Carlino
B3.28 (B3)	60	*Handwritten*
B50.747 (B50)	80	*Handwritten*
B53.781 (B53)	32	*Handwritten*

Table 1: Features of the different target collections considered in this work.

5. RESULTS

Given the number of training corpora (5), the test datasets (15), and the number of experiments (31), we are able to report up to 465 different SER results. This enables us to properly summarize the experimentation, extracting meaningful learnings that will be used to state the best practices to deal with training data on new projects. The analysis of the results follows. The extended raw results of each experiment are attached to this document in the supplementary material.

5.1 Importance of size and variability

In order to understand which is the best training set selection strategy when facing a new unseen collection, all the possible combinations of the datasets available for training have been evaluated against the different target sets.

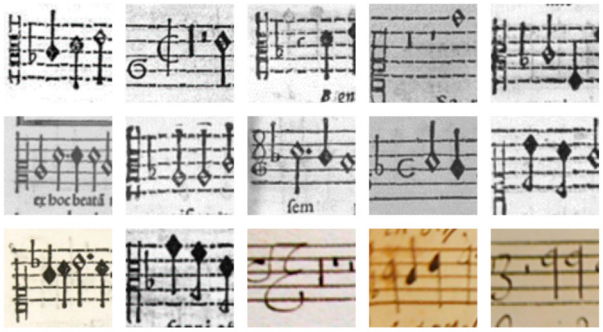


Figure 2: Image examples from the selected corpora as test partition. The images follow a left-right-top-bottom order concerning the list order from Table 1.

The more training sets we include in the combination the greater the number of staves of that combined training set will be. To evaluate which factor is more important, either the variability, given by the number of different training sets included in each combination, or the size as the total number of staves to train, we have plotted in Fig. 3 the summary statistics of the SER obtained by each trained model over all the target collections.

In general, the best behavior has been obtained when merging all the available training corpora. This first outcome may seem obvious, but due to the variability of the training datasets and some of the test works, it was not illogical to expect otherwise. From this result, the fact to be explained is why it performs the best, either due to the size of the training set in terms of the number of staves or the generality the model encompasses due to the training corpora of different natures included.

The plot shows that, although adding more training corpus does not worsen the results, it is not a determining factor. In general, good results are generally obtained with combinations of at least 3 training sets. However, a combination of just two corpora (*i.e.* CS) yields a good performance both in mean and dispersion that denotes its robustness. These two corpora are complementary from the graphical point of view and seem to be representative of both printed sources (SEILS) and handwritten manuscripts (Capitan). When applying 3-corpora training set combinations, the results are equivalent: CGS experiment compared with the GMP, wherein the combination of the first two handwritten corpora and one printed appear compared to the collection of one handwritten and two printed training sets. From these evidences, it can be deduced that the variability of training sets is relevant for better overall performance.

If we focus on the size of the training collection, *i.e.*, the total number of staves used for training, the plot shows that it is not as important as the variability for the final performance. For example, experiment CMS, having less than 4 000 staves, brings better results than experiment GP with over 8 000 samples for training.

To confirm the size is not all that matters, Fig. 4 illustrates the results reported by calculating the number of experiments where the SER is minimized in any of the target datasets, taking into account the number of datasets used

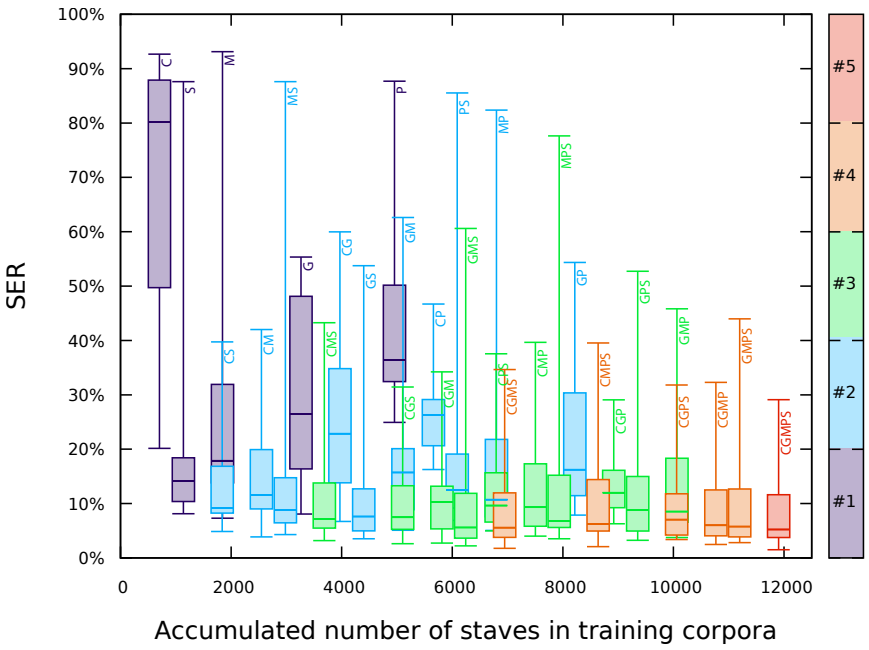


Figure 3: The boxplot shows different statistical SER figures (min, Q1, mean, Q3, max) over the test corpora using a different combination of training corpora. The colors shown in the right bar represent the number of training corpora used in each experiment. The labels are the initials of the corpora included in each training set: **C**: Capitan, **S**: SEILS, **M**: Mottecta, **G**: Guatemala, **P**: PrIMenS.

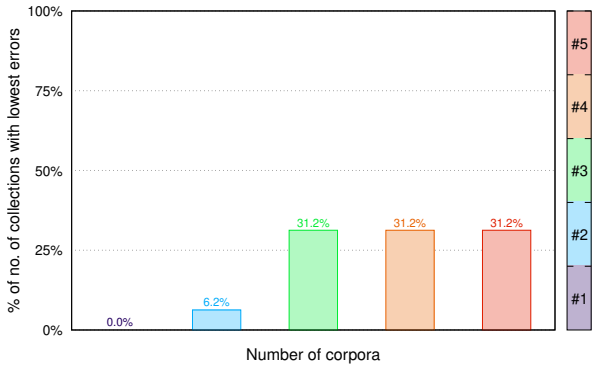


Figure 4: Percentage of experiments that minimize the SER value for any of the available test corpora.

to train. It can be noticed that trained experiments with sizes 3, 4, and 5 report a value of 31.2%. Aside from the value itself, what this aspect exposes is that the size of your dataset at a given point is no longer a critical factor for the transcription quality.

5.2 The complexity of a corpus

The average SER values for all experiments on each target dataset are plotted in Fig. 5. The main noticeable aspect is the difference between Q1 and Q3 (the colored box ends) in the diverse corpora. This substantial contrast in dispersion is what we named "The complexity of a corpus". The plot shows that, as expected, the performance depends on the precise selection of the combination of training corpora to use. The maximum SER values are obtained when the training data is built from just one dataset.

In general, the worst results in the graph are obtained for handwritten target works (those named with the prefix "B") because, intrinsically, they are more difficult to deal with and need a higher variability in the number of training corpora of handwritten works.

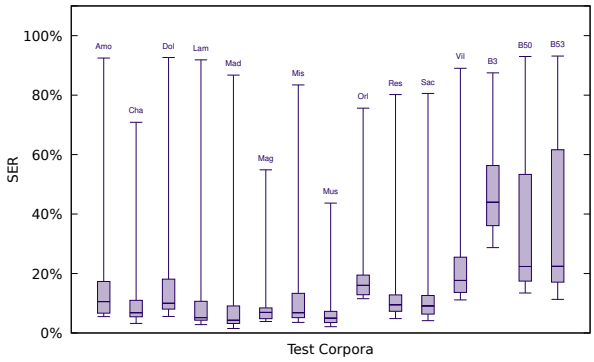


Figure 5: The boxplot shows different statistical SER figures over all experiments made in each one of the testing corpora.

5.3 The importance of leveraging the availability of training corpora

Figure 6 shows the results of the experiments that use each specific training corpus compared to the experiments that do not use it. The image presents the casuistry when having to choose either adding new samples from a different dataset or continue increasing the size of existing labeled samples. As the image reveals, every dataset available for the train, no matter the type—printed or handwritten, real or synthetic—should be included. It is worth mentioning, that the relevance of adding a new corpus is more noticeable than others. For example, referring to the Capitan corpus, if we compare the experiments CMP – MP, CPS – PS,

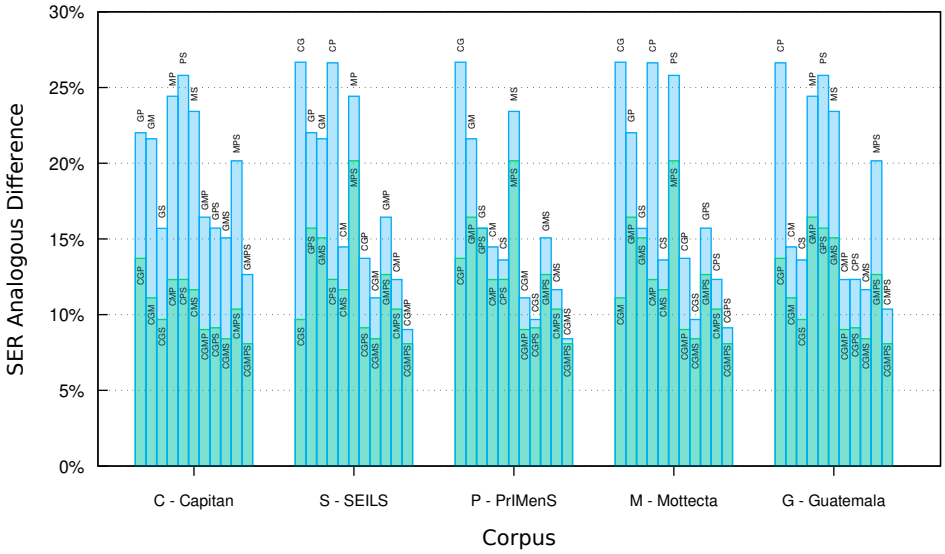


Figure 6: Comparison between all experiments containing (green) and not containing (blue) each training set.

and CMS – MS, we can observe this phenomenon: because of the variability that Capitan adds to the training set, the improvement is noticeable. Therefore, a new corpus seems to generally improve the model performance, as outlined in Fig. 5.

But not only adding a different corpus helps to improve, as the key is to be aware of what is missing in terms of graphical variability in the available training data to build a more robust model. An interesting piece of evidence in the plot that shows how to proceed when this happens is to notice that even a synthetic corpus helps in improving the overall results when it complements the available original training data. Note the reduction in SER when adding PrIMenS, that synthetically simulates printed sources, to complement two other handwritten datasets (Capitan and Guatemala).

5.4 Lessons learned

In order to summarize and establish a set of best practices to improve the generalization performance of OMR systems in the absence of specific training data, we will introduce some questions and answers related to the knowledge acquired from the experimental outcomes.

- **Which is the best choice to transcribe a new collection?** In general, one must use all the available training corpora even if some of them are quite different from the target collection.
- **Is it better to have fewer collections with a high number of samples or more collections with fewer samples each?** It is preferable to have more variability even at the cost of a smaller sample set.
- **How important is it to be aware of the collection to transcribe for selecting the right corpora to train the model?** It is indeed relevant, and depending on the difficulty (for example, whether or not it is handwritten) the differences in performance can be very varied.
- **Does the introduction of a synthetic corpus improve the performance?** Yes, the introduction of a reliable synthetic collection adds size and variability to the training data, enabling better performance rates.

We consider that these answers can be used as general *rules of thumbs*, although of course in certain cases they may not hold.

6. CONCLUSIONS

OMR promises to make written music collections more accessible and browsable by automatically recognizing the symbolic content from their images. However, modern technologies are based on machine learning with deep neural networks, which typically causes unpredictable performance when processing a collection for which no specific training data is available. In this work, we have studied this issue using a large number of training and test collections depicting Mensural notation. This extensive study has been developed considering a state-of-the-art model as representative of the ability to transfer knowledge between collections with dissimilar characteristics.

Our experiments allowed us to analyze various phenomena related to the synergies created between different training collections, the importance of choosing a good recognition trained model to alleviate the uncertainty about performance in a new collection, as well as a series of general good practices on how to proceed for training general OMR models.

As future work, we want to keep on in this line of investigating practical aspects of OMR systems that have a direct impact on particular use cases. For example, we want to extend the case study to the scenario of transfer learning and fine-tuning, where a (limited) amount of training data from a new collection can be assumed. Also, it is interesting to analyze the nature of the errors made by the different OMR models, as well as to have a more precise estimate of the impact of the different errors on the amount of effort required during the post-editing correction process.

7. ACKNOWLEDGMENT

This paper is part of the I+D+i TED2021-130776A-I00 (PolifonIA) project, funded by MCIN/AEI /10.13039/501100011033 and European Union NextGenerationEU/PRTR.

8. REFERENCES

[1] D. Bainbridge and T. Bell, "The challenge of optical music recognition," *Computers and the Humanities*, vol. 35, pp. 95–121, 2001.

[2] A. Rebelo, I. Fujinaga, F. Paszkiewicz, A. R. Marcal, C. Guedes, and J. S. Cardoso, "Optical music recognition: state-of-the-art and open issues," *International Journal of Multimedia Information Retrieval*, vol. 1, pp. 173–190, 2012.

[3] A. Pacha, K.-Y. Choi, B. Coüasnon, Y. Ricquebourg, R. Zanibbi, and H. Eidenberger, "Handwritten music object detection: Open issues and baseline results," in *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*. IEEE, 2018, pp. 163–168.

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

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

[6] A. Chowdhury and L. Vig, "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*. BMVA Press, 2018, p. 202.

[7] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4774–4778.

[8] 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.

[9] P. Torras, A. Baró, L. Kang, and A. Fornés, "On the integration of language models into sequence to sequence architectures for handwritten music recognition," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12*, 2021.

[10] M. Alfaro-Contreras, A. Ríos-Vila, J. J. Valero-Mas, J. M. Iñesta, and J. Calvo-Zaragoza, "Decoupling music notation to improve end-to-end optical music recognition," *Pattern Recognition Letters*, vol. 158, pp. 157–163, 2022.

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

[12] L. Pugin and T. Crawford, "Evaluating OMR on the early music online collection," in *Proceedings of the 14th International Society for Music Information Retrieval Conference, ISMIR 2013, Curitiba, Brazil, November 4-8, 2013*, A. de Souza Britto Jr., F. Gouyon, and S. Dixon, Eds., 2013, pp. 439–444.

[13] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the Twenty-Third International Conference on Machine Learning, (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, 2006, pp. 369–376.

[14] J. Calvo-Zaragoza, A. Toselli, and E. Vidal, "Handwritten music recognition for mensural notation with convolutional recurrent neural networks," *Pattern Recognition Letters*, vol. 128, 08 2019.

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

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

[17] E. Parada-Cabaleiro, A. Batliner, and B. Schuller, "A diplomatic edition of il lauro secco: Ground truth for omr of white mensural notation," 10 2019.

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

[19] D. Huron, "Humdrum and Kern: Selective Feature Encoding BT - Beyond MIDI: The handbook of musical codes," in *Beyond MIDI: The handbook of musical codes*. Cambridge, MA, USA: MIT Press, jan 1997, pp. 375–401.

[20] D. Rizo, N. Pascual-León, and C. S. Sapp, "White mensural manual encoding: from humdrum to mei,"

Cuadernos de Investigación Musical, no. 6, pp. 373–393, 2018.

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