UNSUPERVISED SYNTHETIC-TO-REAL ADAPTATION FOR OPTICAL MUSIC RECOGNITION

Noelia Luna-Barahona[1] **Adrián Roselló**[1] **María Alfaro-Contreras**[1]
David Rizo[1,2] **Jorge Calvo-Zaragoza**[1]
[1] Pattern Recognition and Artificial Intelligence Group, University of Alicante, Spain
[2] Instituto Superior de Enseñanzas Artísticas de la Comunidad Valenciana, Spain
{noelia.luna, adrian.rosello}@ua.es, malfaro, drizo, jcalvo}@dlsi.ua.es

ABSTRACT

The field of Optical Music Recognition (OMR) focuses on models capable of reading music scores from document images. Despite its growing popularity, OMR is still confined to settings where the target scores are similar in both musical context and visual presentation to the data used for training the model. The common scenario, therefore, involves manually annotating data for each specific case, a process that is not only labor-intensive but also raises concerns regarding practicality. We present a methodology based on training a neural model with synthetic images, thus reducing the difficulty of obtaining labeled data. As sheet music renderings depict regular visual characteristics compared to scores from real collections, we propose an unsupervised neural adaptation approach consisting of loss functions that promote alignment between the features learned by the model and those of the target collection while preventing the model from converging to undesirable solutions. This unsupervised adaptation bypasses the need for extensive retraining, requiring only the unlabeled target images. Our experiments, focused on music written in Mensural notation, demonstrate that the methodology is successful and that synthetic-to-real adaptation is indeed a promising way to create practical OMR systems with little human effort.

1. INTRODUCTION

Encoding and transcribing sheet music by hand is a complex and error-prone task that often requires individuals with specialized knowledge of the music notation at hand. An alternative to this manual digitization is the utilization of advanced artificial intelligence technologies, which enable the automated interpretation of musical documents. This technology is known as Optical Music Recognition (OMR) [1].

OMR has been a subject of research for several years [2], experiencing slow progress initially [3]. However, the recent adoption of advanced machine learning techniques, notably Deep Learning, has catalyzed significant improvements in the field [4]. Current OMR systems, albeit not fully perfected, present a more efficient and accurate alternative to manual transcription efforts [5].

In the context of machine learning, existing literature reports models that achieve satisfactory levels of accuracy when processing collections that share graphic characteristics with the training corpus [6–9]. This situation poses challenges for applying OMR technology to new collections, as it is not always feasible, practical, or resource-efficient to dedicate efforts towards annotating a segment of the target collection for training purposes.

This work explores the potential of creating OMR models to address diverse music collections by leveraging synthetic data for training. Given the vast availability of symbolic music data and score engraving tools, generating synthetic data for training presents itself as a viable and promising approach. However, the significant graphical disparities between renderings and real music collections suggest that a straightforward application of such synthetic data might not suffice. To address this issue, we consider the strategy proposed by Alfaro-Contreras & Calvo-Zaragoza [10], aimed at adapting pre-trained transcription models—in our case, initially trained on synthetic data to accommodate real-world music collections. Some previous works on OMR also implement domain adaptation but in other related tasks such as layout analysis [11] or music-object detection [12].

Our experimentation focuses on early monophonic music written in Mensural notation, as there exists a significant number of collections in this notation, each with specific characteristics. This abundance enables us to conduct a thorough examination, aiming to derive conclusions that are broadly applicable and representative. We will use the same synthetic data (and model) to independently adapt to five different Mensural collections. Our experiments indicate that our approach enables consistent synthetic-to-real adaptation, leading to notable improvements in many settings compared to the baseline. While there is still potential for better adaptation, our method represents a significant step towards developing practical OMR models that do not rely on corpus-specific labeled data.

Attribution: N. Luna-Barahona, A. Roselló, M. Alfaro-Contreras, D. Rizo, J. Calvo-Zaragoza, “Unsupervised Synthetic-to-Real Adaptation for Optical Music Recognition”, in *Proc. of the 25th Int. Society for Music Information Retrieval Conf.,* San Francisco, United States, 2024.

2. BACKGROUND

The traditional OMR pipeline comprises four stages [13]: (i) *image pre-processing*, which includes tasks such as binarization, distortion correction, or staff separation; (ii) *music symbol detection*, which involves steps such as staff-line removal, connected-component search, and classification; (iii) *notation assembly*, which relates the individual identified components to reconstruct the musical notation; and (iv) *encoding*, which exports the recognized notation to a specific language for storage and further computational processing.

With the rise of Deep Learning, the so-called *end-to-end* formulation has emerged as an alternative to OMR. This approach, which has been dominating the state of the art in other applications such as text or speech recognition [14, 15], is currently considered the reference model in OMR. The related literature includes many successful solutions of this type [16–18], often with some prior pre-processing such as staff segmentation [19, 20].

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. Instead, in this work, we take a more practical strategy that leverages synthetic data and domain adaptation. Synthetic data, generated through score engraving tools, provides a seemingly infinite resource for training machine learning models without the necessity for laborious manual annotation.

Nevertheless, the utilization of synthetic data presents a critical challenge: while synthetic scores are generated under precise, controlled conditions, real-world music scores exhibit a wide variety of visual characteristics. This variance results in a significant domain gap, where models trained exclusively on synthetic data struggle to generalize. Domain Adaptation (DA) becomes essential to reduce performance degradation by fine-tuning a pre-trained model with unlabeled data from the target domain [21]. While DA has been applied to some stages of the legacy OMR workflow [12, 22], its application to end-to-end approaches remains unexplored. Our contribution is the introduction of an unsupervised synthetic-to-real DA method that employs a specific set of loss functions to adapt pre-trained models using only target staff images.

3. METHODOLOGY

The methodology followed in this work is illustrated in Figure 1. First, a general OMR model is trained in a supervised way using synthetic data. Then, before processing a real collection, for which images but no annotations are available, we apply an unsupervised adaptation approach that modifies the pre-trained model. Then, the adapted model is used to perform OMR on the targeted collection.

The following sections describe the operation of the OMR model and the unsupervised adaptation approach.

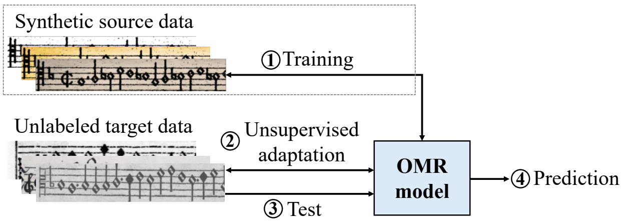


Figure 1: Overview of the unsupervised synthetic-to-real Optical Music Recognition methodology followed in this work.

3.1 Optical Music Recognition model

Our OMR model works at the staff level, assuming that a certain layout analysis has already detected the different staves of the score, as in recent literature [6, 7, 9, 23]. Then, the goal of the model is to retrieve the sequence of music-notation symbols that appear in a given staff.

The state of the art for the aforementioned formulation is to train a Convolutional Recurrent Neural Network (CRNN), using the so-called Connectionist Temporal Classification (CTC) [9, 23]. The convolutional part learns discriminative features from images, while the recurrent block models these features in terms of music-symbol sequences. CTC allows training without explicit information about the location of the symbols in the image [24], which enables an end-to-end learning framework from just pairs of staff images and corresponding transcripts.

Given a staff image $\mathbf{x}$, the output of the CRNN is a stochastic sequence $\pi_{\mathbf{x}} = (\pi_{x_1}, \ldots, \pi_{x_K})$, $\pi_{x_i} \in [0,1]^{\Sigma}$, where K is the number of frames (columns) processed by the recurrent block and Σ represents the vocabulary of music-notation symbols.[1] $\pi^{\sigma}_{x_i}$ represents the probability of observing music-notation symbol σ in the i-th frame of the input ($\sum_{\sigma \in \Sigma} \pi^{\sigma}_{x_i} = 1$). The whole sequence $\pi_{\mathbf{x}}$ is often referred to as the *posteriorgram* of $\mathbf{x}$.

For performing OMR, the posteriorgram is converted into an actual sequence of music-notation symbols by following a *greedy policy* based on retrieving the most probable symbol per frame and applying some direct operations to remove repeated symbols and “blank” tokens.

3.2 Unsupervised adaptation

The model explained in the previous section has demonstrated its goodness in scenarios where the training data belongs to the same collection to be processed. However, this is not interesting in most practical cases, especially when the model is trained with synthetic data, as it barely generalizes to real collections. In this section, we explain the considered approach to adapt a pre-trained model to a (real) target collection using only its images.

Specifically, given a mini-batch $b = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ of target staff images, we fine-tune the pre-trained model with the following loss:

[1] The number of frames is usually less than the number of columns of the original image because the convolutional block typically includes pooling operations.

$$\mathcal{L} = \alpha \cdot \mathcal{L}_a(b) + \beta \cdot \mathcal{L}_r(b) \tag{1}$$

The loss involves two terms, weighted by parameters α and β (to be tuned empirically), respectively: i) one that modifies the model's weights to perform adaptation to a real collection ($\mathcal{L}_a$), and ii) a regularization term that prevents meaningless convergence for OMR ($\mathcal{L}_r$). These are formally introduced in the following sections.

For this adaptation stage, we are not allowed to use the synthetic training corpus, despite being available in our particular case. This is because we are interested in the case for which the original training set is not accessible.

3.2.1 Adaptation term

The first mechanism aims to reduce the discrepancy between the pre-trained model and the target collection by aligning extracted features. Specifically, we approximate the distribution of the pre-trained model as a Gaussian distribution $\mathcal{N}_S(\mu_S, \sigma_S^2)$, using the Batch Normalization (BN) statistics stored in the corresponding layers. [2] During training, BN normalizes layer outputs within each mini-batch, ensuring zero-mean and unit-variance. Exponentially weighted averages of these mean and variance vectors, represented as μ_S and σ_S^2, respectively, are stored during training so that they can be used in the prediction phase to perform standardization.

To reduce distribution discrepancies between the pre-trained and the target collection, we fine-tune the layers preceding BN by forcing their extracted features to have mean and variance vectors similar to those of the source data. Specifically, when given batch b, we compute the mean μ_b and variance σ_b^2. The target batch feature distribution is subsequently approximated as $\mathcal{N}_b(\mu_b, \sigma_b^2)$. We then employ the feature-averaged Kullback-Leibler (KL) divergence to align the target batch feature distribution with the pre-trained feature distribution:

$$\mathcal{L}_a(b) = \mathcal{D}_{\text{KL}}\left(\mathcal{N}_b || \mathcal{N}_S\right) \tag{2}$$

This is described in the context of a single BN layer, but it can be applied to many of them by calculating the loss for each and then adding them up.

3.2.2 Regularization

The previous mechanism can lead to an informational collapse, where the model consistently extracts the same features, regardless of the input image, to match the expected distribution. This would lead to eventually predicting the same music-notation symbol in all frames, which is useless for OMR.

Furthermore, we want to encourage predictions that exhibit music-symbol diversity. This can be induced by maximizing entropy within each frame's predictions across the batch with the following loss: [3]

$$-\sum_{k=1}^{K}\sum_{\sigma\in\Sigma}\mathcal{H}(\pi^{\sigma}_{\mathbf{b}_k}) = \sum_{k=1}^{K}\sum_{\pi\in\Sigma'_S}\sum_{i=1}^{|b|}\left(\pi^{\sigma}_{x_{i_k}}\log\pi^{\sigma}_{x_{i_k}}\right) \tag{3}$$

Specifically, this term penalizes that the same frame in different samples of the batch provides an identical probability distribution over the vocabulary Σ.

Unfortunately, minimizing Eq. 3 might lead to probabilities for a specific frame to be uniformly distributed. In other words, this encourages the model to predict that all music-notation symbols are equiprobable in each frame. However, these distributions should ideally resemble a one-hot distribution, linking each image frame to a single symbol from Σ. To mitigate this, we must further regularize the model to encourage the predictions to behave as one-hot vectors by minimizing the entropy of each frame's output:

$$\sum_{i=1}^{|b|}\sum_{k=1}^{K}\mathcal{H}(\pi_{x_{i_k}}) = -\sum_{i=1}^{|b|}\sum_{k=1}^{K}\sum_{\sigma\in\Sigma}\left(\pi^{\sigma}_{x_{i_k}}\log\pi^{\sigma}_{x_{i_k}}\right) \tag{4}$$

Therefore, the regularization term of our unsupervised adaptation process becomes:

$$\mathcal{L}_r(b) = \sum_{i=1}^{|b|}\sum_{k=1}^{K}\mathcal{H}(\pi_{x_{i_k}}) - \sum_{k=1}^{K}\sum_{\sigma\in\Sigma}\mathcal{H}(\pi^{\sigma}_{\mathbf{b}_k}) \tag{5}$$

where predictions are encouraged to behave like the output of an OMR process, while preventing all predictions from providing the same symbol.

4. DATA

This section covers data handling and preparation, encompassing synthetic data generation to pre-train the model, the considered real datasets for the adaptation experiments, and the encoding of the output vocabulary of the OMR.

4.1 Synthetic data generation

We have considered a modified version of the Printed Images of Mensural Staves (PRIMENS) dataset [9]. The PRIMENS dataset is a synthetic corpus designed to emulate low-quality scans of printed mensural sources. It was obtained by transforming compositions by composers such as Agricola, Frye, and Ockeghem, which are accessible through the Josquin Research Project (JRP) [4]. The original JRP files consist of transcriptions in Common Western Modern notation encoded using `**kern` format. To obtain a Mensural notation dataset, Martínez-Sevilla et al. converted the original files to `**mens` format [25]. Given the polyphonic nature of these compositions, they isolated individual monophonic excerpts by segmenting them into randomly chosen measures spanning from 3 to 18. The

[2] Assuming BN layers for this purpose is a soft constraint since most of the considered CRNN architectures for OMR include these.

[3] Note that the equation is negating the entropy so that the loss is performing *maximization* during gradient descent.

[4] https://josquin.stanford.edu/. Last accessed April 12th, 2024.

authors also modified the original clefs accordingly to increase variability and thus expand the dataset size.

The images were generated using the digital engraver Verovio [26], with random values applied to all available options within permitted ranges. Subsequently, these images were also distorted to mimic genuine printed image scans by employing a random sequence of graphical filters through GraphicsMagick Image Processing System. Furthermore, this simulation of real images was further enhanced by blending randomly damaged old paper textures with distorted images. Figure 2a shows a staff example of the PrIMEnS dataset.

When analyzing the music-symbol distribution of the original PrIMEnS dataset, we found that it lacked bar lines and custodes, two common elements in Mensural corpora. To standardize the vocabulary, we randomly introduced bar lines with a probability of 10% per monophonic excerpt. Custodes were added at the end of each staff, positioned at the most repeated pitch of that region to ensure meaningful vertical staff alignment.

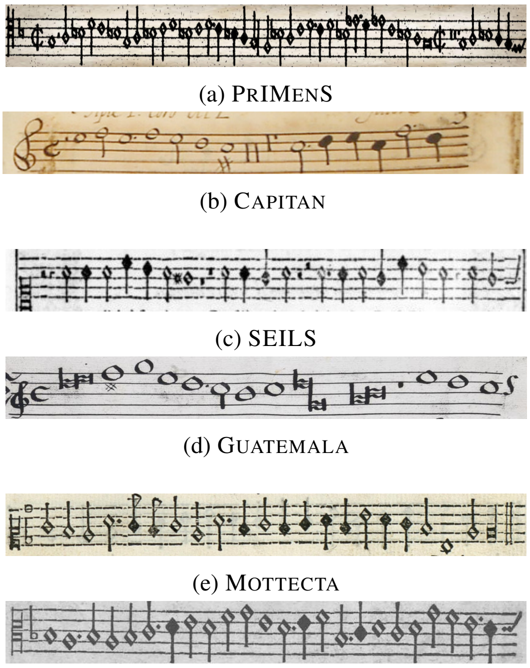
4.2 Real datasets

We have considered five corpora of Mensural music, both handwritten and typeset:

- Capitan corpus [27]: a set of 97 manuscript pages dated from the 17th century of liturgical music. An example of a particular staff from this corpus is depicted in Figure 2b.
- Il Lauro Secco (SEILS) corpus [28]: a collection of 151 typeset pages corresponding to an anthology of Italian madrigals of the 16th century. Figure 2c shows a staff example of this set.
- Guatemala corpus [29]: a collection of 385 handwritten pages from a polyphonic choir book, part of a larger collection held at the "Archivo Histórico Arquidiocesano de Guatemala". An example of a particular staff from this corpus is depicted in Figure 2d.
- Mottecta corpus [9]: a set of 297 printed pages from a collection of the "Biblioteca Digital Hispánica" dated from the 17th century. Figure 2e shows a staff example of this set.
- Magnificat corpus [5]: a set of 127 typeset pages corresponding to a Spanish choir book of the 16th century. See Figure 2f for a sample of this corpus.

4.3 Output encoding

OMR primarily deals with image signals, leading OMR systems to prioritize learning graphic concepts over musical ones. This explains why, when training end-to-end OMR models, an internal representation referred to as "agnostic" is used instead of a semantic representation where music symbols are encoded based on their musical significance [28, 30]. This agnostic representation categorizes elements within a collection of musical symbols according to their form, representing event duration, and their height or vertical position on the staff, denoting pitch. In essence, each symbol is denoted as the 2-tuple $z_i = \langle f_i, h_i \rangle : f_i \in \Sigma_F,\ h_i \in \Sigma_H$, where Σ_F and Σ_H represent the spaces for the different form and height labels, respectively. This approach effectively describes all symbols, including rests that symbolize silence and can be positioned at various vertical locations.


(a) PrIMEnS

(b) Capitan

(c) SEILS

(d) Guatemala

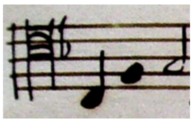
(e) Mottecta

(f) Magnificat

Figure 2: Staff samples of the synthetic data (a) used to train the initial OMR model, which is then adapted to the five Mensural corpora (b-f).

The concise structure of the agnostic representation not only facilitates faster convergence of OMR models but also enables non-experts to annotate music data, making the subsequent conversion to a semantic representation automatable [31]. However, holistic OMR models do not leverage this dual dimensionality. Instead, they treat each combination of form and height as a single category—$|\Sigma| = |\Sigma_F| \times |\Sigma_H|$. Recent works [8, 23] have shown that splitting the symbols in $\mathbf{z}_i$ into their two components and retrieving them sequentially—first, the form and then, the height—leads better recognition rates. Note that the cardinality of the set of symbols in this *split-sequence encoding* is $|\Sigma| = |\Sigma_F| + |\Sigma_H|$, much lower than that of the *standard encoding*, at the expense of doubling the length of the sequence to be predicted. Figure 3 shows a staff sample and its encoding representations in standard and split-sequence encoding.

In this work, we consider both the standard encoding and the split-sequence encoding representations. When using the latter encoding, we adhere to the 2D-greedy decoding method proposed in [8]. This method adjusts the standard CTC greedy decoding to ensure that the output predictions conform to the form-height pattern of the split-



```
barline:L1, clef.C:L4, note.quarter_up:S0,
note.wholeBlack:L2, note.half_up:S2
```

```
barline, L1, clef.C, L4, note.quarter_up, S0,
note.wholeBlack, L2, note.half_up, S2
```

Figure 3: Staff sample and its encoding representations in standard (above) and split-sequence (below) encoding. Note that Ln and Sn respectively denote the line or space of the staff on which the symbol may be placed, which refers to its height property.

sequence encoding representation.

Note that when using the split-sequence encoding representation, we transition from a cardinality of $\Sigma_F \times \Sigma_H$ to one of $\Sigma_F + \Sigma_H$. This implies fewer different symbols and subsequently enables a greater overlap of vocabularies between the source and target collections. This feature makes it particularly suitable for our synthetic-to-real scenario.

Table 1 provides a summary of the characteristics of each label space for the considered corpora. As for data partitioning, we adhere to the same training, validation, and test splits as outlined in the referenced works.

Table 1: Overview of the corpora used in this work: number of staves, vocabulary size for each label space considered (form and height separately for the split-sequence encoding, and a single token combining these two pieces of information for the standard encoding), and engraving style.

	Staves	**Vocabulary**			**Engraving style**
		Form	Height	Combined	
PRIMENS	42 136	37	34	386	Synthetic
CAPITAN	828	62	16	372	Handwritten
SEILS	1 136	37	17	205	Typeset
GUATEMALA	3 263	52	17	315	Handwritten
MOTTECTA	1 847	38	15	228	Typeset
MAGNIFICAT	1 340	42	19	220	Typeset

5. EXPERIMENTAL SET UP

This section describes the evaluation protocol and the implementation details.

5.1 Evaluation metric

We consider the Symbol Error Rate (SER) for assessing the performance of the presented recognition scheme, as in previous works [6–9]. This metric is computed as the average number of elementary editing operations (insertions, deletions, or substitutions) required to match the sequence predicted by the model with that in the ground truth, normalized by the length of the latter. In mathematical terms, this is expressed as:

$$\text{SER}\ (\%) = \frac{\sum_{i=1}^{|\mathcal{S}|} \text{ED}\left(\hat{\mathbf{z}}_i,\ \mathbf{z}_i\right)}{\sum_{i=1}^{|\mathcal{S}|} |\mathbf{z}_i|} \qquad (6)$$

where $\mathcal{S} \subset \mathcal{X} \times \mathcal{Z}$ is a set of test data, ED : $\mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{N}_0$ denotes the string edit distance [32], and $\hat{\mathbf{z}}_i$ and $\mathbf{z}_i$ respectively represent the estimated and target sequences. For comparative purposes, we convert all predicted and ground-truth sequences to split-sequence before computing the metric.

5.2 Implementation details

The CRNN scheme is based on that used typically for OMR [7, 9, 27]. Specifically, we used four convolutional layers that applied 64 filters of size 5×5, 64 filters of size 5×5, 128 filters of size 3×3, and 128 filters of size 3×3, respectively. We considered a Leaky ReLU activation with a negative slope of $\alpha = 0.2$ and max-pooling stages of size and striding factors of 2×1 (except the first convolutional layer, which is 2×2). The produced feature maps were fed into two Bidirectional Long Short-Time Memory layers with 256 hidden units each and a dropout value of 50%, followed by a fully-connected network with $|\Sigma'|$ units that provide a probability for each possible music-notation token.

The evaluation pipeline consisted of two stages: (i) training the source model, and (ii) adapting it to the target dataset using the AMD method. For (i), we used the ADAM optimizer with a batch size of 16 elements and a fixed learning rate of 10^{-3}. We stopped the training using an early stopping strategy with a patience of 20 epochs, retaining the weights that minimize the SER metric in the validation partition. In (ii) we maintained the batch size of 16, the learning rate was selected through a random search ranging from 10^{-3} to 3×10^{-4}, and a maximum of 50 training-adaptation epochs was considered as we finetuned an already trained model.

Regarding data pre-processing, we replicated the exact experimental conditions outlined in the aforementioned reference works. Specifically, we resized each staff image to a height of 64 pixels, preserving the aspect ratio (individual samples may vary in width), and converted them to grayscale without any additional pre-processing steps. Additionally, following the approach outlined in the aforementioned works, we incorporated a data augmentation step during the training of the source models.

6. RESULTS

This section presents the results obtained from applying the experimental scheme to the different presented corpora. Specifically, Table 2 depicts the performance of the PRIMENS model before and after adaptation for each real target Mensural corpus in terms of the SER metric. [5]

The most important remark is that the considered synthetic-to-real adaptation framework improves the performance of the synthetic-only scenario across all datasets. The approach does not solve the adaptation challenge completely (the reference value is still far in most cases), but it allows taking the model to more usable levels without

[5] Code at: https://github.com/OMR-PRAIG-UA-ES/ISMIR-2024-SYNTHETIC2REAL-OMR.

Table 2: Results in terms of the SER (%) metric for each real Mensural corpus before and after the unsupervised adaptation of the OMR model trained with the synthetic PrIMenS dataset. For completeness, we also include the in-collection performance, where the real corpus is used for both training and testing. These performances are shaded in gray, serving as an upper-bound reference. Final row reports the relative improvement (↓∆ %) when adaptation is performed.

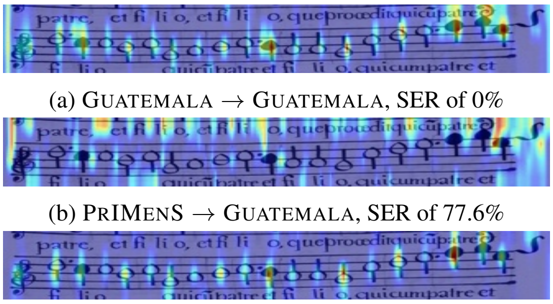
	Target corpus									
	Capitan		SEILS		Guatemala		Mottecta		Magnificat	
	Standard	Split-sequence	Standard	Split-sequence	Standard	Split-sequence	Standard	Split-sequence	Standard	Split-sequence
In-collection (reference)	6.7	6.2	1.8	1.7	1.6	1.6	3.3	2.9	1.5	1.5
Before adaptation	45.9	43.1	23.9	21.3	46.9	53.4	25.8	29.2	12.8	12.8
After adaptation	32.9	32.9	19.3	18.5	21.4	18.1	17.1	16.4	9.1	10.6
Relative improvement	↓28 %	↓23 %	↓19 %	↓14 %	↓54 %	↓66 %	↓34 %	↓44 %	↓29 %	↓17 %

the need to initially annotate data. This is quite useful, for example, in the context of OMR plus post-correction.

The degree of relative improvement varies depending on the specific dataset, ranging from 66% to 14%. In this sense, it is difficult to draw a correlation between the different factors and the degree of improvement. However, it is worth highlighting that the scenarios with a greater margin (for example, Guatemala and Capitan) lead to a greater absolute improvement. This may indicate that there is a glass ceiling to the performance that can be obtained by training with a synthetic corpus, since in the cases where the result is already relatively successful (e.g. Magnificat) the improvement is rather limited.

Concerning the output encoding, the split-sequence encoding generally yields better SER figures in the in-collection scenario. However, the differences are marginal in the other two scenarios. Therefore, this does not represent a relevant factor for adaptation.

To provide more insights into the adaptation process, we explored the “relevant” parts of the image that the different OMR models consider to predict the symbols. Gradient-weighted Class Activation Mapping (Grad-CAM) [33] is an interpretability method that uses the gradients of any target prediction to produce a coarse localization map highlighting the important regions in the image for such predictions. Figure 4 shows the activation map over the same test image for the three different scenarios considered: (a) in-collection, (b) before adaptation, and (c) after adaptation. Specifically, we display here the case of processing the real collection Guatemala. We can observe how the initially misplaced pixel activations in scenario (b) are corrected to the actual music symbols after adaptation in scenario (c), showing a high degree of similarity to the activation map of the in-collection model of scenario (a).



(a) Guatemala → Guatemala, SER of 0%

(b) PrIMenS → Guatemala, SER of 77.6%

(c) PrIMenS → Guatemala (adapted), SER of 0.9%

Figure 4: Activation maps over the same Guatemala test image using an OMR model trained with (a) Guatemala scores, (b) PrIMenS scores, and (c) PrIMenS scores but adapted to Guatemala images.

7. CONCLUSIONS

Existing end-to-end OMR approaches have exhibited remarkable performance in transcribing collections that share graphic characteristics with the training corpus. However, when this condition is not met, allocating resources to manually annotate training data to maintain performance levels becomes impractical and resource-intensive. Our work proposes a possible solution to this challenge. Firstly, we train an initial OMR model with synthetic scores. By doing so, we eliminate the need for human manual annotation of training data. Subsequently, we tailor this model to the specific characteristics of the target corpus through unsupervised adaptation, using only unlabeled images from the target corpus. This adaptation process employs a loss function to align the learned features of the model with those of the target collection while ensuring the model does not converge to undesirable solutions. Our experiments across five distinct Mensural datasets validate the effectiveness of our synthetic-to-real adaptation as a viable approach to developing universal OMR systems with little human effort. However, there remains room for improvement. Future research avenues may explore leveraging self-labeled samples obtained through the adapted model to further enhance its performance and robustness or exploring few-shot scenarios.

8. ACKNOWLEDGEMENTS

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

9. REFERENCES

[1] J. Calvo-Zaragoza, J. H. Jr, and A. Pacha, “Understanding Optical Music Recognition,” *ACM Computing Surveys*, vol. 53, no. 4, pp. 1–35, 2020.

[2] D. Bainbridge and T. Bell, “The challenge of optical

music recognition," *Computers and the Humanities*, vol. 35, pp. 95–121, 2001.

[3] 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.

[4] J. Calvo-Zaragoza, J. C. Martinez-Sevilla, C. Peñarrubia, and A. Ríos-Vila, "Optical music recognition: Recent advances, current challenges, and future directions," in *Proceedings in International Conference on Document Analysis and Recognition*, ser. Lecture Notes in Computer Science, M. Coustaty and A. Fornés, Eds., vol. 14193. San José, CA, USA: Springer, 2023, pp. 94–104.

[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. Baró, C. Badal, and A. Fornés, "Handwritten Historical Music Recognition by Sequence-to-Sequence with Attention Mechanism," in *Proceedings of the 17th International Conference on Frontiers in Handwriting Recognition*. Dortmund, Germany: IEEE, Sep. 2020, pp. 205–210.

[7] M. Villarreal and J. A. Sánchez, "Handwritten Music Recognition Improvement through Language Model Re-interpretation for Mensural Notation," in *Proceedings of the 17th International Conference on Frontiers in Handwriting Recognition*. Dortmund, Germany: IEEE, Sep. 2020, pp. 199–204.

[8] 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.

[9] J. C. Martínez-Sevilla, A. Roselló, D. Rizo, and J. Calvo-Zaragoza, "On the Performance of Optical Music Recognition in the Absence of Specific Training Data," in *Proceedings of the 24th International Society for Music Information Retrieval Conference*. Milan, Italy: ISMIR, Nov. 2023, pp. 319–326.

[10] M. Alfaro-Contreras and J. Calvo-Zaragoza, "Align, minimize and diversify: A source-free unsupervised domain adaptation method for handwritten text recognition," *arXiv preprint arXiv:2404.18260*, 2024.

[11] F. J. Castellanos, A. J. Gallego, J. Calvo-Zaragoza, and I. Fujinaga, "Domain adaptation for staff-region retrieval of music score images," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 25, no. 4, pp. 281–292, 2022.

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

[13] I. Fujinaga and G. Vigliensoni, "The art of teaching computers: the SIMSSA optical music recognition workflow system," in *Proceedings of the 327th European Signal Processing Conference*. IEEE, 2019, pp. 1–5.

[14] A. Chowdhury and L. Vig, "An Efficient End-to-End Neural Model for Handwritten Text Recognition," in *British Machine Vision Conference*. Newcastle, UK: BMVA Press, Sep. 2018, p. 2018.

[15] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 4774–4778.

[16] Pau Torras and Arnau Baró and Lei Kang and Alicia Fornés, "On the Integration of Language Models into Sequence to Sequence Architectures for Handwritten Music Recognition," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference*. Online: ISMIR, Nov. 2021.

[17] M. Alfaro-Contreras, J. M. Iñesta, and J. Calvo-Zaragoza, "Optical music recognition for homophonic scores with neural networks and synthetic music generation," *International Journal of Multimedia Information Retrieval*, vol. 12, no. 1, p. 12, 2023.

[18] A. Ríos-Vila, J. Calvo-Zaragoza, and T. Paquet, "Sheet Music Transformer: End-To-End Optical Music Recognition Beyond Monophonic Transcription," *arXiv preprint arXiv:2402.07596*, 2024.

[19] M. Kletz and A. Pacha, "Detecting Staves and Measures in Music Scores with Deep Learning," in *Proceedings of the 3rd International Workshop on Reading Music Systems*, Alicante, Spain, 2021, pp. 8–12.

[20] F. J. Castellanos, C. Garrido-Munoz, A. Ríos-Vila, and J. Calvo-Zaragoza, "Region-based layout analysis of music score images," *Expert Systems with Applications*, vol. 209, p. 118211, 2022.

[21] W. M. Kouw and M. Loog, "A Review of Domain Adaptation without Target Labels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 766–785, 2021.

[22] F. J. Castellanos, A. J. Gallego, J. Calvo-Zaragoza, and I. Fujinaga, "Domain adaptation for staff-region retrieval of music score images," *International Journal*

on Document Analysis and Recognition, vol. 25, no. 4, pp. 281–292, 2022.

[23] A. Ríos-Vila, J. Calvo-Zaragoza, and J. M. Iñesta, "Exploring the two-dimensional nature of music notation for score recognition with end-to-end approaches," in *Proceedings of the 17th International Conference on Frontiers in Handwriting Recognition*. Dortmund, Germany: IEEE, Sep. 2020, pp. 193–198.

[24] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh, USA: Association for Computing Machinery, Jun. 2006, pp. 369–376.

[25] D. Rizo, N. Pascual-León, and C. Sapp, "White mensural manual encoding: from humdrum to mei," *Cuadernos de Investigación Musical*, 2019.

[26] L. Pugin, R. Zitellini, and P. Roland, "Verovio - A library for Engraving MEI Music Notation into SVG," in *Proceedings of the 15th International Society for Music Information Retrieval Conference*. Taipei, Taiwan: ISMIR, Jan. 2014.

[27] 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.

[28] E. Parada-Cabaleiro, A. Batliner, and B. W. Schuller, "A Diplomatic Edition of Il Lauro Secco: Ground Truth for OMR of White Mensural Notation," in *Proceedings of the 20th International Society for Music Information Retrieval Conference*. Delft, The Netherlands: ISMIR, Nov. 2019, pp. 557–564.

[29] 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*. Prague, Czech Republic: Association for Computing Machinery, Jul. 2022, pp. 19–26.

[30] J. Calvo-Zaragoza and D. Rizo, "Camera-PrIMuS: Neural End-to-End Optical Music Recognition on Realistic Monophonic Scores," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*. Paris, France: ISMIR, Sep. 2018, pp. 248–255.

[31] A. Ríos-Vila, M. Esplà-Gomis, D. Rizo, P. J. Ponce de León, and J. M. Iñesta, "Applying Automatic Translation for Optical Music Recognition's Encoding Step," *Applied Sciences*, vol. 11, no. 9, pp. 3890–3912, 2021.

[32] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals," *Soviet Physics-Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.