A NEURAL APPROACH FOR FULL-PAGE OPTICAL MUSIC RECOGNITION OF MENSURAL DOCUMENTS

Francisco J. Castellanos **Jorge Calvo-Zaragoza** **Jose M. Inesta**

Department of Software and Computing Systems, University of Alicante, Spain

fcastellanos@dlsi.ua.es, jcalvo@dlsi.ua.es, inesta@dlsi.ua.es

ABSTRACT

The digitization of the content within musical manuscripts allows the possibility of preserving, disseminating, and exploiting that cultural heritage. The automation of this process has been object of study for a long time in the field of Optical Music Recognition (OMR), with a wide variety of proposed solutions. Currently, there is a tendency to use machine learning strategies based on neural networks because of their high performance and flexibility to adapt to different scenarios by changing only the training data. However, most of the recent literature addresses only specific parts of the traditional OMR workflow such as music object detection or symbol classification. In this paper, we progress one step further by proposing a full-page OMR system for Mensural notation scores that consists of simply two processes, which are enough to extract the symbolic music information from a full page. More precisely, our pipeline uses Selectional Auto-Encoders to extract single staff regions, combined with end-to-end staff-level recognition based on Convolutional Recurrent Neural Networks for retrieving the music notation. The results confirm the adequacy of our method, reporting a successful behavior on two Mensural collections (CAPITAN and SEILS datasets) with a straightforward implementation.

1. INTRODUCTION

The digitization of the content within documents [1] is a process that helps to preserve cultural heritage and enables easier dissemination and knowledge creation. Traditionally, this content digitization is done manually, with an undeniably high cost that is very prone to introduce mistakes as well. In the music context, the development of Optical Music Recognition (OMR) systems promises to perform this task automatically with minimum human involvement. Research efforts have promoted the progress in this field achieving excellent, yet partial results [2–4]; therefore, full digitization of music documents is still to be studied in practical contexts.

Recent advances in machine learning enable new approaches in the OMR field [5]. The use of deep neural networks provides novel ways of avoiding complex multi-step workflows that are considered in legacy OMR research [6]. A successful example of this new trend is the so-called end-to-end approach, that operates at the staff level; in other words, a single step that completely processes the image of a single staff and retrieves the series of symbols that appear therein [7].

While end-to-end strategies can be used to read a sequence of symbols at the staff level, it is still necessary to previously detect all the staves contained in the documents as region blocks, for then transcribing the music content. This staff detection task has been addressed in recent literature [8, 9]. However, these works only assess staff detection as a computer vision problem—i.e., how accurate is, in geometric terms, the region extracted, without considering how useful it is for the subsequent steps. These partial results are not sufficient to determine with certainty the goodness of the approaches within a complete OMR pipeline.

In this work, we carry out a study to determine how the recent advances in OMR interact with each other. Also, we eventually offer, for the first time, results that validate that only two steps—the staff-region detection combined with an end-to-end method—are sufficient to develop a complete page-level OMR system with excellent recognition rates through neural networks.

As we will explain later, this approach is successful if the graphical complexity of the scores follows certain criteria: single-staff systems with a single voice in each. That is why our experiments are restricted to Mensural manuscripts, of great historical interest, where these requirements are common.

2. BACKGROUND

Although the term OMR covers a wide range of scenarios—different research might be carried out according to the notational type or the engraving mechanism of the manuscripts—there has been a general pipeline that addresses the challenge through a series of independent stages that work on different parts of the problem [5].

Traditionally, individual challenges were very complex, so procedures were developed to work on specific manuscripts [10–13]. However, the systems ended up be-

Attribution: F.J. Castellanos, J. Calvo-Zaragoza and J.M. Iñesta, “A Neural Approach for Full-Page Optical Music Recognition of Mensural Documents”, in *Proc. of the 21st Int. Society for Music Information Retrieval Conf.,* Montréal, Canada, 2020.

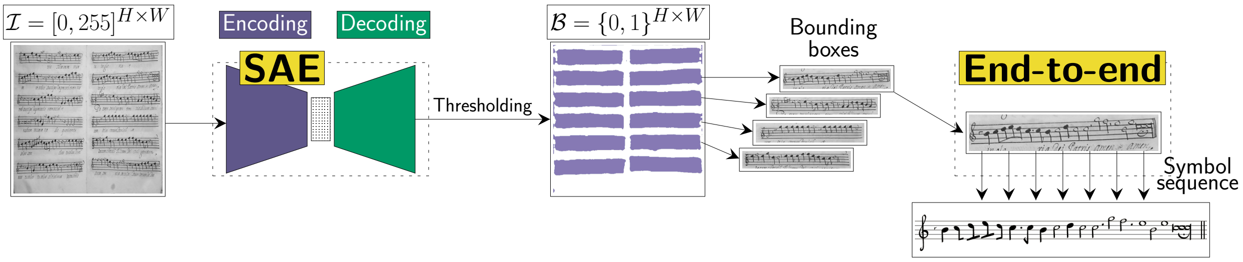

$\mathcal{I} = [0, 255]^{H\times W}$
Encoding
Decoding
SAE
Thresholding
$\mathcal{B} = \{0, 1\}^{H\times W}$
Bounding boxes
End-to-end
Symbol sequence

Figure 1: Scheme of the considered methodology.

ing very specific, so previous efforts were hardly reused. In other words, there was little scientific progress in the field.

Recently, the early stages of the process have been reformulated as object detection tasks [14], thus bypassing some of the stages of the traditional workflow. Pacha et al. [15] provided a baseline for direct music-object detection in music score images, experimenting with several models and corpora of different typology.

At the same time, there have been approaches for end-to-end OMR. Such models perform the complete recognition of musical notation from an image, directly providing the sequence of music symbols present therein as output. In addition to high performance, this prevents the need for the training set to be annotated at the symbol-level position and post-process strategies that convert the individually detected elements to the actual music notation. Concerning this formulation, Pugin [16] pioneered the end-to-end approach for printed Mensural notation using Hidden Markov Models (HMM) with the Aruspix system. However, although HMMs represent models that fit perfectly well with the task at issue, other tasks of a similar nature, like handwritten text recognition, experienced a leap in performance using deep neural networks [17]. It has been demonstrated that neural approaches outperform those based on HMM for end-to-end OMR as well [7].

To date, however, there is no existing end-to-end approach that works at the full-page level, but only at the single-staff level. It is not only a challenge to be solved in the field of OMR but also in text recognition—a task that we could consider even simpler. In text recognition, the end-to-end approaches face the recognition process at the line level [18]. For this, there exist line extraction algorithms [19], which enable working at the page level in combination with the line-level end-to-end neural networks. In the case of music, a similar idea is to use staff extraction algorithms combined with the end-to-end staff-level recognition.

Recently, several methods have been proposed to solve the staff detection task [8, 9, 20]. The problem with these works is that they only studied the extraction of staves as a computer vision challenge. Similarly, the end-to-end staff-level neural networks for OMR only experimented with staves detected manually. Therefore, it is not known how well the combination of staff retrieval with staff-level end-to-end neural networks performs in real scenarios.

For all the above, this paper fills a gap in the existing literature and presents, for the first time, a neural full-page OMR system that takes advantage of recent advances in deep learning to solve the task in just two steps: staff retrieval and end-to-end staff-level recognition. As we will see later, this allows us to provide a general approach that works successfully in different manuscripts by simply providing training data.

3. METHODOLOGY

The proposed methodology outlines an approach by which to evaluate a full page-level OMR system using only two procedures: staff retrieval and end-to-end staff-level recognition. Both of them are solved in a single step each by using deep neural networks. A graphical overview of the complete methodology is shown in Figure 1.

One of the main advantages of our methodology is that it is completely based on machine learning: it is enough to provide annotated examples (of each task) to build new and accurate models—which is usually easier and cheaper than developing a pipeline anew.

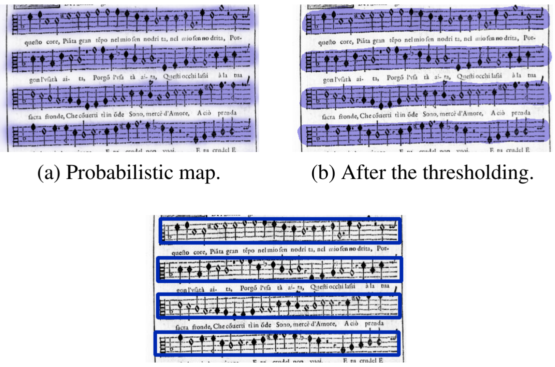
Although this approach might not work for arbitrary types of music scores—e.g., recognizing each staff separately does not make that sense for scores that include multi-staff systems—we believe it is worth studying and providing simple, generalizable, and effective solutions in those cases where the structural complexity of the scores makes it possible. Furthermore, our approach is not necessarily restricted to the case of monophony but can be applied in the case where only one voice appears on each staff. In our case study, whose details are available in Section 5, we will apply our methodology to vocal polyphony scores in Mensural notation—where different voices appear independently.

3.1 Staff retrieval

The first step in the considered methodology needs to detect and extract the individual staves. With the premise that all individual staves are compact blocks within the image, we can apply a layout analysis to estimate the probability of each pixel to belong to one of the staves. Previous work [21] presented a Selectional Auto-Encoder (SAE)-based framework focused on performing layout analysis by patches to split the image into different information layers: staff lines, symbols, lyrics and background. Here, we adapt that method to directly detect staff regions. Since the staff blocks are extensive and compact, a patch-wise model may introduce additional errors in their detection. For avoiding that, we propose to adjust the original image to the input size requirements of the model, being this new resolution enough to discern the different staves.

Let $\mathcal{I} = [0, 255]^{H \times W}$ be a grayscale image[1] with height H and width W in terms of pixels. The SAE model processes $\mathcal{I}$ to return another image $\mathcal{S} = [0, 1]^{H \times W}$ such that $\mathcal{S}_{i,j} \equiv P(\mathcal{I}_{i,j} =$ 'staff')—i.e., $\mathcal{S}$ stands for an image with the same size of $\mathcal{I}$, and whose values represent the probability of each pixel of belonging to any staff region. Note that the SAE model is trained through supervised learning mechanisms, hence a set of documents annotated with the location of each staff is required.

After obtaining $\mathcal{S}$, a threshold $\tau \in [0, 1]$ is applied to obtain a binary map $\mathcal{B} = \{0, 1\}^{H \times W}$. Then, the staff regions can be retrieved by performing a connected-component analysis over the map $\mathcal{B}$. Afterwards, we compute the rectangular coordinates of each component for retrieving the bounding boxes. An example of this process can be found in Figure 2.



(a) Probabilistic map.

(b) After the thresholding.

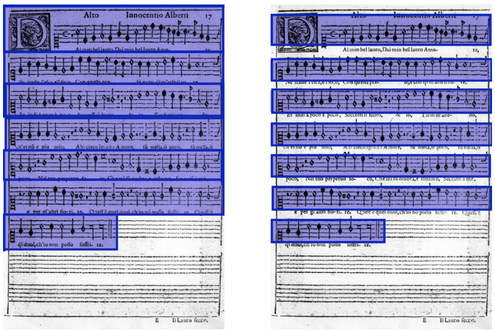
(c) Obtaining bounding boxes.

Figure 2: Example of staff-region prediction, with the probabilistic map obtained by the SAE model, the result of applying a threshold to determinate the areas most likely being staves, and finally the bounding box retrieval.

A drawback to this approach is that it requires that the different staves do not overlap with each other, as shown in Figure 3, as that would prevent distinguishing them once $\mathcal{B}$ has been computed. To reduce this possibility, we propose to apply a vertical reduction factor δ, with which the bounding boxes in the ground truth will be trimmed vertically, largely avoiding the overlapping in the annotated documents. Note that, since clipping is necessary to make the subsequent prediction easier, the bounding boxes obtained by the SAE model should be expanded by the same factor after being retrieved. In this way, ideally, the detected bounding boxes will cover the staves completely.

3.2 End-to-end staff-level recognition

Once individual staves are extracted, the symbol recognition at this level can be performed by an end-to-end methodology based on deep neural networks. Within the many options for this, we consider the approach initially proposed by Shi et al. [17], given that it outperformed competing methods for OMR [7].



(a) Original page.

(b) With $\delta = 20\%$.

Figure 3: Example of ground truth with overlapping that can be solved by means of applying a reduction factor ($\delta = 20\%$, i.e. 20% top and bottom trims).

Given an image $\mathbf{x}$, corresponding to a single-staff region, we want to retrieve the most probable sequence from a fixed alphabet Σ of music symbols. $\mathbf{x}$ can be interpreted as a sequence of frames (single image columns), so the aforementioned problem can be solved by using a recurrent neural network [22]. These networks can provide a probability per frame $P(\sigma \mid x_i), 1 \leq i \leq |\mathbf{x}|, \sigma \in \Sigma \cup \{\epsilon\}$, where ϵ is a special token required to separate consecutive predictions of the same symbol [23].

This stochastic representation of $\mathbf{x}$ can be decoded into an actual sequence of music symbols by first retrieving the most probable sequence of symbols per frame

$$\sigma_i = \arg \max_{\sigma \in \Sigma \cup \{\epsilon\}} P(\sigma \mid \mathbf{x}_i)$$

and then following a greedy approach which merges consecutive frames with the same symbol and removes the frames whose predicted symbol is ϵ [7].

In our case, we add a convolutional neural network on top of the recurrent neural network to automatically learn features that are appropriate for the specific music manuscript at issue [24].

The joint Convolutional Recurrent Neural Network (CRNN) can be trained in an end-to-end fashion by using the so-called Connectionist Temporal Classification (CTC) loss function [23]. Given a ground-truth sample consisting of a single-staff region $\mathbf{x}$ and its corresponding sequence of music symbols σ, CTC is used to modify the network's weights to maximize the probability of retrieving σ from $\mathbf{x}$ without the need of providing a framewise localization of the symbols.

4. EXPERIMENTAL SETUP

4.1 Parameterization of the Neural Networks

In this section, we present the setup of the neural models for both staff retrieval and staff-level symbol recognition. For the first one, we considered the use of SAE due to its high performance and efficiency in the document analysis

[1] This is with no loss of generality, as the approach can be easily extended to deal with color images as well.

Input	Encoding	Decoding	Output
$[0, 255]^{512\times512}$	Conv2D$(128, 5 \times 5,$ ‘ReLU’$)$	Conv2D$(128, 5 \times 5,$ ‘ReLU’$)$	$[0, 1]^{512\times512}$
	MaxPool(2×2)	UpSamp(2×2)	
	Conv2D$(128, 5 \times 5,$ ‘ReLU’$)$	Conv2D$(128, 5 \times 5,$ ‘ReLU’$)$	
	MaxPool(2×2)	UpSamp(2×2)	
	Conv2D$(128, 5 \times 5,$ ‘ReLU’$)$	Conv2D$(128, 5 \times 5,$ ‘ReLU’$)$	
	MaxPool(2×2)	UpSamp(2×2)	
		Conv2D$(1, 5 \times 5,$ ‘sigmoid’$)$	

Table 1: Detailed description of the selected SAE architecture, implemented as a Fully-Convolutional Network (FCN). ‘ReLU’ and ‘sigmoid’ denote the Rectifier Linear Unit and Sigmoid activations, respectively.

task [21]. As of the second process, the symbolic music sequence is obtained by means of a CRNN.

The following notation will be used for the specifications given below: Conv2D$(n, h \times w,$ ‘act’$)$ indicates a two-dimensional convolution operator of n filters and kernel size of $h \times w$ with ‘act’ denoting the actual activation function; MaxPool$(h \times w)$ represents a down-sampling max-pooling operation with a $h \times w$ window; UpSamp$(h \times w)$ denotes an up-sampling operator of h rows and w columns; BLSTM(n) stands for a bidirectional Long Short-Term Memory unit of n neurons; Dropout(p) represents a dropout operation with a ratio of p neurons; Dense$(n,$ ‘act’$)$ indicates a dense layer of n neurons with ‘act’ denoting the actual activation function.

4.1.1 Selectional Auto-Encoder

The SAE configuration used in this work is set according to previous works for layout analysis, whose details are given in Table 1. In the staff analysis, the model does not need to predict small details since staves are extensive and compact within the document. For this, we can rescale the original image to the size of the input model, being of enough resolution to differentiate the staves of the ground truth. After some informal testing, we configured the input as an image of 512×512 px. The image rescaling was performed through the OpenCV library.

In addition, as discussed in Section 3.1, a vertical reduction factor δ and a threshold τ to determine the pixels belonging to a staff are necessary. We set $\delta = 20\%$, so the ground-truth staves are top and bottom trimmed by that factor, and $\tau = 0.5$ to indicate that a probability higher or equal to 50% is assumed to represent a pixel from a staff. In our preliminary experiment experiments, δ played an important role for avoiding overlapping, whereas the model was quite robust against different values of τ.

4.1.2 Convolutional Recurrent Neural Network

The CRNN follows the best architecture from the work by Calvo-Zaragoza et al. [7]. It consists of four convolutional layers and max-pooling down-sampling, connected with a recurrent block of two Bidirectional Long Short-Term Memory (LSTM) layers. The specifications of the model are given in Table 2.

Input: $[0, 255]^{64\times\mathcal{W}}$
Conv2D(64, 5×5,‘ReLU’), MaxPool(2×2)
Conv2D(64, 5×5,‘ReLU’), MaxPool(2×2)
Conv2D(128, 3×3,‘ReLU’), MaxPool(2×1)
Conv2D(128, 3×3,‘ReLU’), MaxPool(2×1)
BLSTM(256), Dropout(0.5)
BLSTM(256), Dropout(0.5)
Dense($\lvert\Sigma \cup \{\epsilon\}\rvert$, ‘softmax’)

Table 2: Architecture of the CRNN considered for staff-level recognition. ‘softmax’ indicates the Softmax activation, that normalized the output to a probability over the set of symbols (plus the ‘blank’ symbol denoted by ϵ). Given that the images are of variable width, this dimension of the input is not specified (indicated as $\mathcal{W}$).

4.2 Corpora

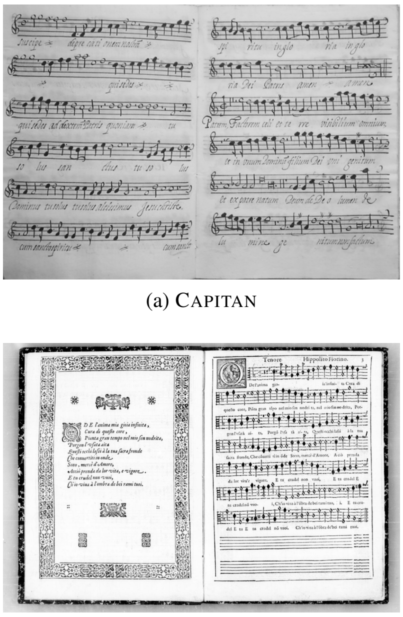
To evaluate our method, we consider the following corpora of Mensural manuscripts:

- The CAPITAN dataset, which encodes a complete *Missa* composed during the second half of the 17th century. Annotations are specifically provided for OMR [25].
- The Symbolically Encoded Il Laurro Secco (SEILS) dataset, which consists of scores from the 16th-century anthology of Italian madrigals *Il Lauro Secco*. Among many formats, the dataset includes the required format to perform OMR [26].

	CAPITAN	SEILS
Engraving	Handwritten	Printed
Pages	97	150
Staves	737	1 278
Running symbols	17 112	31 589
Symbol categories	53	33

Table 3: Corpora statistics.

Page samples from these corpora can be seen in Figure 4. As observed, CAPITAN is handwritten and SEILS is printed. This heterogeneity benefits the verification that the proposed methodology is generalizable to a variety of manuscript types. In addition, some descriptive statistics about the corpora are provided in Table 3.



(a) CAPITAN

(b) SEILS

Figure 4: An example page image from of each dataset considered in the experimentation.

4.3 Evaluation protocol

As of the experimentation, the results have been computed using a 5-fold cross-validation technique (5-CV), each of which takes three data partitions—training, validation and testing—with 60%, 20% and 20% of the whole set of documents, respectively. The training process was performed during 100 epochs, monitoring with the validation partition and reporting the results on the test partition. The experiments have been performed using the Keras v.2.3.1 [27] library with TensorFlow v.1.14 as backend.

In the literature, the experiments are commonly evaluated partially, focusing only on individual processes, regardless of the impact they may have on the transcription into a digital format, which is precisely the ultimate purpose of the OMR field. The main goal of this paper is to evaluate a full-page OMR system combining a staff-retrieval method based on SAE and an end-to-end staff-level recognition. Therefore, we will be able to analyze the effect of staff retrieval in the symbol recognition step.

Nevertheless, experiments have been divided into two parts: an assessment of the staff bounding-box recognition, in which we will present the computation of the average of Intersection over Union (IoU), which provides a measure of the overlapping between the set of retrieved staves and the ground-truth ones (the higher, the better). With regard to the second step of the proposal, the objective is to check the recognition of the sequence of symbols within each staff obtained in the first step. In practical OMR systems, a critical factor to be considered is the number of corrections the user has to perform. Hence, we decided to report the final results in terms of Symbol Error Rate (SER), which is computed as the ratio of editing operations needed to correct the transcription of the symbol sequence (the lower, the better).

5. RESULTS

Concerning the staff retrieval itself, this step detected all of the 737 staves that CAPITAN contains (considering all the test partitions within the 5-CV) but also retrieved an additional box that did not correspond to a staff. Similarly, the model for SEILS retrieved correctly all the 1 278 staves, while 152 regions were detected where there were none. Therefore, whereas all real staves are retrieved, the process also yields some *false positives*, that are supposed to be easily removed in an interactive environment. As a reference in geometric terms, the model for CAPITAN obtained 86.3% of IoU, while SEILS achieved 79.7%. This indicates that the retrieved boxes generally fit well the ground-truth location of the staves. We will see below that this is accurate enough for the task of retrieving the inner symbol sequences.

To complement these numerical results, Figure 5 contains an example of a comparison between a ground-truth staff with the predicted one. Note that, although the IoU obtained in that example is 80.5%, the retrieved staff properly covers the music information. As evidenced in Figure 6, most of the predicted bounding boxes have an IoU between 70% and 95%.

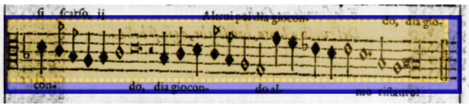


Figure 5: Example of a retrieved staff from SEILS, colored in yellow, compared with the ground-truth box, colored in blue. For this example, the IoU reaches 80.5%.

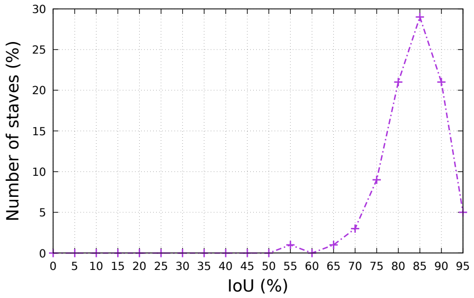


Figure 6: Average histogram of staves predicted by the SAE model and ordered by IoU (with a granularity of 5%).

We now proceed to present the results of the symbol recognition step. To fully evaluate the two-step proposed method, we should first take into account two main points: first, the staff retrieval SAE can only be trained with ground-truth data, since it constitutes the earlier step in the approach; and second, the end-to-end model would be ideally trained with the bounding boxes predicted in the last step—closer to the real scenario—but it is also possible to directly use the ground-truth staves. While in all cases the sequence of music symbols to be predicted is the same, we can compare the results with different configurations of the input images provided for training the CRNN: ground-truth staff regions (GT); staff regions predicted by the SAE, once trained (Pred.); both ground-truth staff regions and staff regions predicted by the SAE, once trained (GT+Pred.).

Likewise, for the test partition, a real case only contemplates evaluating with the automatically detected staves (real scenario); however, we have also evaluated on ground-truth staves to provide a reference of the loss caused by the automatic staff retrieval.

We report in Table 4 the results of the end-to-end staff-level recognition of symbols, which has been tested with both CAPITAN and SEILS datasets, and each one in turn evaluated with the cases mentioned above.

Data		CAPITAN	SEILS
Training staves	Test staves		
Real scenario			
GT	Pred.	16.8 ± 3.7	5.2 ± 1.4
Pred.	Pred.	14.8 ± 3.6	4.4 ± 0.5
GT+Pred.	Pred.	**11.5 ± 2.2**	**3.7 ± 0.8**
Reference			
GT	GT	13.2 ± 1.1	4.4 ± 1.2
GT+Pred.	GT	**10.8 ± 1.1**	**3.6 ± 0.9**

Table 4: Average ± std. deviation results in terms of SER (%) of a 5-CV experiment for the staff-level end-to-end recognition with different combinations of training and test data during the staff retrieval stage. GT stands for ground-truth staves, while Pred. represents the predicted ones.

First, we focus on the results obtained in the real scenario, i.e. those in which the test staves have been provided by the staff-retrieval step. It can be observed that training with GT achieves successful outcomes, being the SER metric 16.8% and 5.2% for CAPITAN and SEILS, respectively. Results are improved if the training data contains predicted staves instead of GT, with figures that reach 14.8% for CAPITAN and 4.4% for SEILS. The reason behind this phenomenon may come from what is seen in Figure 5: the staff is correctly detected but the box is actually different from the ground-truth one. Therefore, if this difference is also introduced during training, the model is better prepared for what occurs in the real case. Despite this, the experiments reveal that the robustness of the end-to-end model is optimized if a combination of GT and Pred. is performed in the training process, allowing to reduce the symbol error rate until 11.5% and 3.7% for CAPITAN and SEILS datasets, respectively. This combination in the training data seems similar to the typical machine learning strategy called data augmentation [28, 29], given that the Pred. boxes depict variations with respect to the GT ones.

If we analyze the reference results, i.e., those obtained by the end-to-end step tested with GT boxes, we observe that the end-to-end model outperforms the real case, as both training and test data are part of the annotated bounding boxes by the user. Similarly, the data augmentation strategy allows to even improve the results. When comparing to the real scenario, the reference case reports slightly better results, with negligible differences (from 11.5% to 10.8% for CAPITAN and from 3.7% to 3.6% for SEILS, at best).

What is important about the figures above is that they demonstrate that introducing an automatic staff detection step barely affects the overall performance of the system—according to the best values obtained with predictions compared to the best values obtained with ground-truth boxes. Therefore, we can validate our methodology as suitable to deal with the complete recognition of Mensural manuscripts or even any type of musical document that depicts a comparable structure.

Finally, although this is not of special relevance within the scope of this paper, we see that the machine learning models find it easier to deal with printed manuscripts, as the error figures from SEILS are clearly below those from the CAPITAN one. Probably, the regularity of the printed symbols makes the task easier than in the handwritten case.

6. CONCLUSIONS

OMR is an interesting field of study, but most of its research focus on individual steps that avoid evaluating the impact within the full system. In this paper, a full-page OMR system with neural networks has been presented. It is based on the combination of staff-retrieval and symbol sequence recognition steps.

The first step—staff retrieval—has been implemented as a SAE model based on a successful architecture used in previous work for layout analysis. This neural network predicts staff regions as compact blocks, processing the whole image in only one step, and then bounding boxes of predicted staves are extracted. The second step—end-to-end staff-level recognition—transcribes the content of the predicted staves into a digital format, which is the main goal of OMR.

The paper includes a study of the impact of the first step in the final performance in the digitization for two Mensural manuscripts. The methodology has been assessed in terms of SER, which determines the number of corrections that a user should make to have the correct sequence transcribed. The results reveal that ground-truth staves are not the best option for training the end-to-end model in a real case, in which the transcription will be performed from predicted staves. The assessment of the model trained with predicted staves shows performances as good as in the ideal case, in which the training and the test datasets consist of ground truth regions. This means that the precision in staff retrieval is not the most important issue in the symbol recognition task. Furthermore, we observed that training the end-to-end symbol recognizer with a combination of predicted and ground-truth staves provides the best results for both, real and ideal situations, with non-significant differences between them. Therefore, between a fully automatic OMR system and other where the bounding boxes are annotated by the user, the performance hardly varies. We can then conclude that our approach allows transcribing reliably the music content with minimum human effort.

In future works, we plan to keep on researching simple and generalizable strategies for more complex manuscripts, such as those corresponding to polyphonic scores in Western modern notation.

7. ACKNOWLEDGMENT

This work was supported by the Spanish Ministry HISPAMUS project TIN2017-86576-R, partially funded by the EU. First author also acknowledges the support from "Programa I+D+i de la Generalitat Valenciana" through grant ACIF/2019/ 042.

8. REFERENCES

[1] D. Doermann, K. Tombre *et al.*, *Handbook of Document Image Processing and Recognition*. Springer, 2014.

[2] T. Pinto, A. Rebelo, G. Giraldi, and J. S. Cardoso, "Music score binarization based on domain knowledge," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2011, pp. 700–708.

[3] A.-J. Gallego and J. Calvo-Zaragoza, "Staff-line removal with selectional auto-encoders," *Expert Systems with Applications*, vol. 89, pp. 138–148, 2017.

[4] Z. Huang, X. Jia, and Y. Guo, "State-of-the-art model for music object recognition with deep learning," *Applied Sciences*, vol. 9, no. 13, p. 2645, 2019.

[5] J. Calvo-Zaragoza, J. H. Jr., and A. Pacha, "Understanding optical music recognition," *ACM Comput. Surv.*, vol. 53, no. 4, Jul. 2020.

[6] D. Bainbridge and T. Bell, "The challenge of optical music recognition," *Computers and the Humanities*, vol. 35, no. 2, pp. 95–121, 2001.

[7] 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.

[8] L. Quirós, A. H. Toselli, and E. Vidal, "Multi-task layout analysis of handwritten musical scores," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2019, pp. 123–134.

[9] A. Pacha, "Incremental supervised staff detection," in *Proceedings of the 2nd International Workshop on Reading Music Systems*, Delft, The Netherlands, 2019, pp. 16–20.

[10] C. Dalitz, G. K. Michalakis, and C. Pranzas, "Optical recognition of psaltic byzantine chant notation," *International Journal of Document Analysis and Recognition*, vol. 11, no. 3, pp. 143–158, 2008.

[11] C. Dalitz and C. Pranzas, "German lute tablature recognition," in *10th International Conference on Document Analysis and Recognition*, 2009, pp. 371–375.

[12] L. J. Tardón, S. Sammartino, I. Barbancho, V. Gómez, and A. Oliver, "Optical music recognition for scores written in white mensural notation," *EURASIP Journal on Image and Video Processing*, vol. 2009, no. 1, p. 843401, 2009.

[13] Y.-H. Huang, X. Chen, S. Beck, D. Burn, and L. Van Gool, "Automatic handwritten mensural notation interpreter: From manuscript to MIDI performance," in *16th International Society for Music Information Retrieval Conference*, M. Müller and F. Wiering, Eds., Málaga, Spain, 2015, pp. 79–85.

[14] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.

[15] A. Pacha, J. Hajič, and J. Calvo-Zaragoza, "A baseline for general music object detection with deep learning," *Applied Sciences*, vol. 8, no. 9, p. 1488, 2018.

[16] L. Pugin, "Optical music recognitoin of early typographic prints using hidden markov models," in *Proceedings of the 7th International Conference on Music Information Retrieval, Victoria, Canada, 8-12 October*, 2006, pp. 53–56.

[17] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298–2304, 2017.

[18] A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks," in *Advances in neural information processing systems*, 2009, pp. 545–552.

[19] M. Pastor, "Text baseline detection, a single page trained system," *Pattern Recognit.*, vol. 94, pp. 149–161, 2019.

[20] V. Bosch, J. Calvo-Zaragoza, A. H. Toselli, and E. Vidal-Ruiz, "Sheet music statistical layout analysis," in *15th International Conference on Frontiers in Handwriting Recognition, ICFHR 2016, Shenzhen, China, October 23-26*, 2016, pp. 313–8.

[21] F. J. Castellanos, J. Calvo-Zaragoza, G. Vigliensoni, and I. Fujinaga, "Document analysis of music score images with selectional auto-encoders," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, 2018, pp. 256–263.

[22] A. Graves, "Supervised sequence labelling with recurrent neural networks," Ph.D. dissertation, Technical University Munich, 2008.

[23] 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*, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 369–376.

[24] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proceedings of the 13th European Conference on Computer Vision, Zurich, Switzerland, September 6-12, Part I*, 2014, pp. 818–833.

[25] J. Calvo-Zaragoza, A. H. Toselli, and E. Vidal, "Handwritten music recognition for mensural notation: Formulation, data and baseline results," in *14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, November 9-15*, 2017, pp. 1081–1086.

[26] 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, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, 2019, pp. 557–564.

[27] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[28] Y. Xu, R. Jia, L. Mou, G. Li, Y. Chen, Y. Lu, and Z. Jin, "Improved relation classification by deep recurrent neural networks with data augmentation," in *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, 2016, pp. 1461–1470.

[29] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: When to warp?" in *2016 International Conference on Digital Image Computing: Techniques and Applications, DICTA 2016, Gold Coast, Australia, November 30 - December 2, 2016*, 2016, pp. 1–6.