A Starting Point for Handwritten Music Recognition

Arnau Baró, Pau Riba and Alicia Fornés
Computer Vision Center - Computer Science Department
Universitat Autònoma de Barcelona, Bellaterra, Catalonia, Spain
Email: {abaro,priba,afornes}@cvc.uab.cat

***Abstract*—In the last years, the interest in Optical Music Recognition (OMR) has reawakened, especially since the appearance of deep learning. However, there are very few works addressing handwritten scores. In this work we describe a full OMR pipeline for handwritten music scores by using Convolutional and Recurrent Neural Networks that could serve as a baseline for the research community.**

***Index Terms*—Optical Music Recognition, Long Short-Term Memory, Convolutional Neural Networks, MUSCIMA++, CVC-MUSCIMA.**

I. Introduction

For many decades, music scores have been manually written in a sheet format. Nowadays, there are archives with thousands of handwritten music scores waiting to be transcribed. Since a manual transcription becomes unfeasible, it is necessary to develop an automatic method to transcribe music scores.

Optical Music Recognition (OMR) is the process to convert a music score image into a machine-readable format. There are some OMR software such as PhotoScore [1] or SharpEye[2] that work very well on printed scores. However, when they have to deal with handwritten music scores, the accuracy decreases significantly. As far as we know, the few existing OMR handwritten methods only focus on a specific stage of the full OMR pipeline, such as layout analysis [1], detection and classification of graphic primitives [2], [3]. Thus, we believe that it is time to design a full OMR pipeline for handwritten scores.

In this paper we propose a full staff-wise OMR system for handwritten music scores which can serve as a baseline for the research community. Our method is composed by a Convolutional Neural Network followed by a Bidirectional Long Short-Term Recurrent Neural Network. This architecture is based on our previous publication [4], where we proposed to recognize printed music scores as a sequential recognition task by using Bidirectional Long Short-Term Memory networks. In the current work we improve and adapt this architecture for dealing with handwritten music scores. First, we have added a Convolutional Neural Network in order to extract meaningful features. Since the amount of annotated handwritten data is very limited, we propose a specific data augmentation technique to increase the amount of training music scores. Finally, we have studied transfer learning techniques to benefit from printed and synthetic music scores.

[1]http://www.neuratron.com/photoscore.htm
[2]http://www.visiv.co.uk/

II. Proposed architecture

Single staff sheet music can be seen as a sequence. In this way, first of all we have cropped the different staves of each page to read them from left to right. Then, our architecture is composed of the following steps:

Input: Each music score is first resized to a height of 100 pixels and binarized to feed each column in the proposed architecture. Since the aspect ratio is kept, each image will have different width. Afterwards, all images belonging to the same batch are padded to the maximum width.

Convolutional Block: This block uses three convolutional layers with a kernel of 3x3 which increases the depth. Batch Normalization and Rectified Linear Unit activations are located after each convolutional layer. Then a max-pool 2x1 operator is used to maintain the image width in order to feed our model column-by-column.

Recurrent Block: This block is constructed by 4 Bidirectional Long Short-Term Memory (BLSTM) layers of 512 neurons each. The bidirectionallity provides more context than a single LSTM, because ambiguities can be reduced when taking into account the forward and backward directions in the image. For example, if one direction is reading a vertical line and the other direction is seeing a notehead, the network can correctly predict a quarter note.

Dense Layers: After the recurrent block, two fully connected (FC) layers are located. We use two separated fully connected layers in order to reduce the large combination between rhythm and melody. Beside this, by separating rhythm and melody, the system is able to learn the shape of a symbol independently of its position on the staff and vice versa. In other words, the system can learn the shape of a quarter note no matter if it is located in the first or fourth line in the staff.

Output: Lastly, each fully connected layer returns a binary matrix. Each matrix has the same width as the original image. The rhythm is defined using a matrix with a height of 80 classes and the pitch with 28 classes. Finally, a threshold is applied to decide which symbols appear in the music score.

Both matrices are finally converted into one array. So, we obtain three arrays, one for the rhythm, other for the pitch and another one with the combination of both. These arrays will be used to evaluate the method.

III. Training strategies

Since there are very few labelled handwritten music scores to train the systems, this leads to overfitting problems. For this reason, we opt to use printed data as pre-training.

TABLE I
RESULTS. SYMBOL ERROR RATES ARE BETWEEN [0-1], AND GIVEN BY THE MEAN OF FIVE EXECUTIONS.

Pre-train Printed	D. Augm. Printed	BLSTM	CNN	D. Augm. Handwritten		Rhythm SER	Pitch SER	Rhythm+Pitch SER
				Shuffle	Morph.			
-	-	-	-	-	-	0.826	0.709	0.899
✓	-	-	-	-	-	0.771	0.668	0.872
✓	✓	-	-	-	-	0.762	0.690	0.854
✓	✓	✓	-	-	-	0.523	0.464	0.610
✓	✓	✓	✓	-	-	0.493	0.396	0.559
✓	✓	✓	✓	✓	-	**0.476**	**0.387**	**0.545**
✓	✓	✓	✓	✓	✓	0.490	0.393	0.554

Concretely, we propose to train our model with printed data and afterwards retrain it with the few available handwritten data. In other words, we propose transfer learning by fine-tuning the pre-trained system with the handwritten data.

In addition, some distortions have been applied to the training data so that the system can learn a more robust and general shape of every symbol. First, we have applied three distortion methods -dilating, eroding or blurring- into both the printed and handwritten training sets. Secondly, and in order to increase the amount of possible melodies, the number of handwritten training music scores has been increased by cropping each measure (bar unit) and shuffling them in the music score.

IV. EXPERIMENTATION

To evaluate our OMR system on handwritten music scores, we have labelled at symbol level a subset of 20 pages of the MUSCIMA++ dataset [5], which is a subset of the CVC-MUSCIMA dataset [6]. For measuring the performance, we have used the Symbol Error Rate (SER) as evaluation metric, defined as the sum of edit operations that are needed to convert the output of our architecture into the label in terms of symbol insertions, substitutions and deletions.

Table I shows the results of our method. In the table, the lower SER, the better. Note that each line introduces an improvement to the previous system. In the first row we are testing our method without any of the proposed improvements nor pre-training on printed data.

It can be observed that the incorporation of the data augmentation on the printed dataset, and shuffling the measures indeed improves the overall system performance. Also, instead of using the raw image as input, the convolutional block shows that it indeed helps to extract the discriminative features to recognize the different music symbols. Nevertheless, the BLSTM is the key modification to reduce the error rates by a large margin. This is because the bidirectionality is able to reduce most of the ambiguities. Finally, note that using morphological operations for data augmentation only introduces noise and increases the error rates.

V. CONCLUSIONS AND FUTURE WORK

In this work, we have proposed a full Optical Music Recognition system to recognize handwritten music scores. This method uses a convolution block to extract features followed by a recurrent block based on the Bidirectional Long Short-Term Memory Recurrent Neural Networks. It also includes a specific data augmentation and uses transfer learning from printed data.

We believe that the provided results can be used as a baseline for future improvements in the OMR research field. Bearing in mind that we have only used 20 pages of the MUSCIMA++, these results are encouraging. We think that increasing the amount of annotated handwritten music scores at symbol level, the system could obtain much better results.

Future work will be focused on investigating more suitable data augmentation methods for music scores. Moreover, we would like to study how a postprocessing step based on grammars or semantics could solve ambiguities at higher level, and thus improve the performance.

ACKNOWLEDGMENTS

This work has been partially supported by the project TIN2015-70924-C2-2-R, the CERCA Program/Generalitat de Catalunya, and the Fellowships AGAUR 2018 FI_B 00546, FPU15/06264 and RYC-2014-16831.

REFERENCES

[1] J. Calvo-Zaragoza, F. J. Castellanos, G. Vigliensoni, and I. Fujinaga, "Deep neural networks for document processing of music score images," *Applied Sciences*, vol. 8, no. 5, pp. 654–674, 2018.

[2] J. Hajič jr. and P. Pecina, "Detecting noteheads in handwritten scores with convnets and bounding box regression," *CoRR*, vol. abs/1708.01806, 2017.

[3] L. Tuggener, I. Elezi, J. Schmidhuber, and T. Stadelmann, "Deep watershed detector for music object recognition (accepted)," in *ISMIR*, 2018.

[4] A. Baró, P. Riba, J. Calvo-Zaragoza, and A. Fornés, "Optical music recognition by long short-term memory networks," in *Graphic Recognition. Current Trends and Challenges*, 2018.

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

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