Optical Music Recognition by Recurrent Neural Networks

Arnau Baró*, Pau Riba*, Jorge Calvo-Zaragoza† and Alicia Fornés*

*Computer Vision Center - Computer Science Department
Universitat Autònoma de Barcelona, Bellaterra, Catalonia, Spain
Email: {abaro,priba,afornes}@cvc.uab.cat

†Schulich School of Music
McGill University, Montreal, Canada
Email: jorge.calvozaragoza@mail.mcgill.ca

***Abstract*—Optical Music Recognition is the task of transcribing a music score into a machine readable format. Many music scores are written in a single staff, and therefore, they could be treated as a sequence. Therefore, this work explores the use of Long Short-Term Memory (LSTM) Recurrent Neural Networks for reading the music score sequentially, where the LSTM helps in keeping the context. For training, we have used a synthetic dataset of more than 40000 images, labeled at primitive level.**

***Index Terms*—Optical Music Recognition; Recurrent Neural Network; Long Short-Term Memory.**

I. Introduction

The transcription of sheet music into some machine-readable format can be carried out manually. However, the complexity of music notation inevitably leads to burdensome software for music score editing, which makes the whole process very time-consuming and prone to errors. Consequently, automatic transcription systems for musical documents represent interesting tools. The field devoted to address this task is known as Optical Music Recognition (OMR) [1]. Typically, an OMR system takes an image of a music score and automatically export its content into some symbolic structure such as MEI or MusicXML.

The process of recognizing the content of a music score is complex, and therefore the workflow of an OMR system is very extensive [2]. However, our work focuses on recognizing the content appearing on a single staff section (e.g. scores for violin, flute, etc.), much in the same way as most text recognition research focuses on recognizing words appearing in a given line image [3]. This should not be an issue as there exist successful algorithms to both isolate staff sections and separate music and lyrics (accompanying text) [4].

To address this specific task, we propose the use of Recurrent Neural Networks (RNN), which have been applied with great success to many sequential recognition tasks such as speech [5] or handwriting [3] recognition.

The rest of the paper is organized as follows. Section II describes the method that we have used to solve the problem. Section III discusses the preliminar qualitative results. Finally, conslusions and future work are drawn in Section IV

II. Method

Music scores that are written in single staff can be represented as a sequence. In this way, we can read a music score from left to right. In order to make a similar lecture by a computer a RNN is appropriate to perform this task. In this work a Long Short-Term Memory (LSTM) [6] RNN has been used. The architecture is very similar but LSTM decides which information has to keep as context an which one has to remove (i.e. forget).

To evaluate our approach, we used a synthetic dataset that is composed of more than 40000 music scores with 3 different typographies. The dataset corresponds to incipts from the RISM dataset [1].

These images are resized to a height of 50 pixels in order to feed pixel columns into the proposed model. The maximum width can be variable depending on the widest image in the batch. Images with a width shorter than the widest one will have padding.

Beside this, the ground truth is represented by two binary matrices, one for the rhythm and other for the melody (the pitch). Each part in the horizontal axis is as long as the input image. And the vertical axis is 54 for the melody and 26 for the rhythm. 54 and 26 indicate the number of different possible symbols in the dataset, and the number of different possible pitches (i.e. locations in the staff) of the notes in the music score. Some symbols have been manually added to make easier the recognition task. One of these symbols is the epsilon(ε), which is used to know where each symbol starts and ends. This symbol can be seen as a separator. Where this symbol is activated, it means that it is not possible to have any other symbol activated as well. The main reason to divide the ground truth in two is that the combination between melody and rhythm is almost infinite.

he LSTM has 3 layers and the hidden size is 128. At the end of the network, after the LSTM's output, two fully connected (FC) layers allow us to separate the rhythm and the melody in two different outputs. After the FC layers, the next step is to calculate the loss. In music, in one instance of time we can find one or more symbols, for example, the case of chords. For this reason we need a loss function that allows us to have more than one class per time step. We tried two different loss functions: $SmoothL1Loss$ and $MultiLabelSoftMarginLoss$. The loss is calculated independently. One loss for the rhythm and other for the melody

[1] http://www.rism.info/

(pitch). Once both losses are calculated, they are summed and backpropagated.

III. Results

In Table I we show the average and standard deviation applying LSTM and BLSTM. The first column shows the loss function and the network that has been used, the second column the results in term of error rate of the rhythm, the third column the results in term of error rate of the pitch and the last column the results in term of error rate of the rhythm and pitch jointly. Note that BLSTM produces better results. It process the input in both directions getting information of the whole symbol, and therefore, it is more accurate. For example, if one direction recognizes a note-head, the other direction can discard that the vertical line that it is reading is not a bar line, but instead a note stem (both stems and bar lines are straight vertical lines).

Table I: Results using LSTM and BLSTM. All results are between [0-1]. The first number is the mean of the five executions and the number between parenthesis is the standard deviation

	Rhythm (R) Symbol ER	Pitch (P) Symbol ER	R + P Symbol ER
LSTM Smooth L1	0.326 ($\pm$ 0.007)	0.293 ($\pm$ 0.008)	0.426 ($\pm$ 0.009)
BLSTM Smooth L1	**0.020** ($\pm$ 0.001)	**0.015** ($\pm$ 0.001)	**0.028** ($\pm$ 0.002)
LSTM Multi Label Soft Margin	0.431 ($\pm$ 0.017)	0.567 ($\pm$ 0.051)	0.747 ($\pm$ 0.063)
BLSTM Multi Label Soft Margin	0.027 ($\pm$ 0.002)	0.023 ($\pm$ 0.002)	0.036 ($\pm$ 0.003

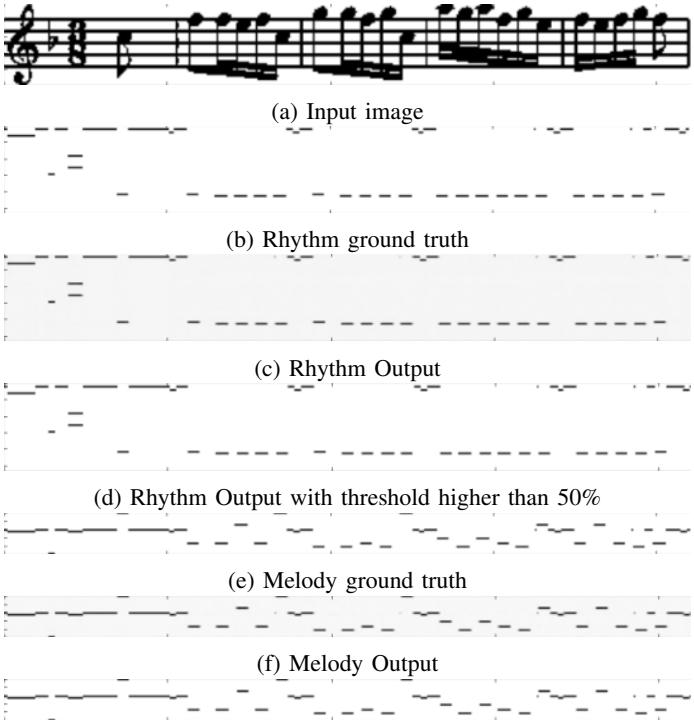
In Figure 1 we can see some qualitative results. In Figure 1a we can see the input of the BLSTM. Figure 1b is the Rhythm Ground Truth and Figure 1c is the BLSTM's output. Figure 1d is the same output but only activating the positions with confidence higher than 50%. In Figure 1e we can see the Melody Ground Truth. Figure 1f is the BLSTM's output and Figure 1g the same output but only activating the positions with confidence higher than 50%.

As we can see in the images, the results are promising, since the output of the network is very close to the ground truth.

IV. Conclusion and Future Work

In this work, we have proposed an optical music recognition method. This method is based on following a sequence path with LSTM and BLSTM Recurrent Neural Networks.

The results are encouraging. Thus, we could conclude that single staff music scores could be recognized as a sequence. However, more difficult music scores, for example scores with multiple voices, will need other kind of methods. Future work will be focused on investigate transfer learning methods to recognize handwritten music scores. Finally, we would also like to apply music rules or semantics as in our previous work [7] in order to solve ambiguities, and also, investigate more suitable techniques for recognizing complex polyphonic music scores.



(a) Input image

(b) Rhythm ground truth

(c) Rhythm Output

(d) Rhythm Output with threshold higher than 50%

(e) Melody ground truth

(f) Melody Output

(g) Melody Output with threshold higher than 50%

Figure 1: Qualitative Results Example

Acknowledgment

This work has been partially supported by the Spanish project TIN2015-70924-C2-2-R, the Ramon y Cajal Fellowship RYC-2014-16831, the CERCA Program/Generalitat de Catalunya, FPU fellowship FPU15/06264 from the Spanish Ministerio de Educación, Cultura y Deporte and the social Sciences and Humanities Research Council of Canada. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

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

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

[3] A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks," in *NIPS*, 2009, pp. 545–552.

[4] J. A. Burgoyne, Y. Ouyang, T. Himmelman, J. Devaney, L. Pugin, and I. Fujinaga, "Lyric extraction and recognition on digital images of early music sources," in *ISMIR*, 2009, pp. 723–727.

[5] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 6645–6649.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 9, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

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