How can Machine Learning make Optical Music Recognition more relevant for practicing musicians?

Heinz Roggenkemper
Los Gatos, Ca, USA
heinz@roggenkemper.net

Ryan Roggenkemper
Berkeley, CA, USA
rrogenkemper@berkeley.edu

***Abstract*—Optical music recognition has not become relevant yet for musicians. Based on a proof of concept we attempt to outline user expectations, what is missing, and that community involvement is key to make machine learning more relevant.**

Disclaimer: This has been written from the point of view of practicing amateur musicians with limited experience with optical music recognition.

Keywords—optical music recognition, machine learning, user expectations, community involvement

I. Introduction

In recent years optical character recognition (OCR) has moved to be deeply embedded in products and services: scanners often perform OCR directly after scanning so that the PDFs become searchable. Dropbox has automatic OCR as a feature to their business users, and Google offers it in Google Drive (when you open a PDF with Google Docs, OCR is performed in the background). Machine learning and especially the advances of deep learning in areas like image recognition have made this possible. It seems fair to say that OCR has become relevant for many users (sometimes without them being aware that they are using it).

On the first glance OCR and optical music recognition (OMR) seem similar. It would appear that for instance OpenScore (https://openscore.cc/) would be a natural candidate for the application of OMR. However, it focuses on crowdsourced human effort to reach its goal.

Obviously the number of potential users, and the resulting interest and investment differ for OCR and OMR, and there is crucial difference between the two from a user's perspective: a page of text that an OCR system processes with 99% accuracy is likely very useful – important services like search documents work, and a user reads the document, the human brain will recognize the meaning of the words and ignore the errors. However, if a violinist is given a one-page score with the same accuracy, it is most likely useless for her/him – the human ear will neither ignore nor forgive the errors. In addition, a string player would expect at least slurs and staccatos from a score.

II. Background

In the middle of 2017, and after experiencing the amount of work necessary to produce usable scores through manual effort, we got excited by the idea of combining computer vision and machine learning to the OMR after reading how Dropbox had approached OCR [4], and decided to do a proof of concept with the following goals:

> Build an OMR system that combines computer vision and machine learning, and achieves an accuracy that is higher than any of the commercial OMR system that was analyzed in [5] for string quartets. Accuracy in the proof of concept is defined as getting pitch and duration right (slurs, accents, dynamics etc. are ignored.) That translates an accuracy of over 90%. Stretch goal is to achieve an accuracy that is higher than the combined output of multiple sources (95%).

Since the goal of the proof of concept was to compare the results to ones that were done using a Mozart string quartet [4], that is what we used as well as a starting point (K.458, IMSLP482550, MusicXML from Project Gutenberg - https://www.gutenberg.org/ebooks/4951). We wrote a set of tools to create individual images from the IMSLP PDF, extracted the labels from the MusicXML file, matched the labels to the images, and checked them carefully (this turned out to be necessary). In addition, we created synthetic files, rendered through both MuseScore and Finale. The main reason was completeness - we wanted as many combinations of pitch and duration represented as possible.

In total we used about 10,000 labeled image files of musical symbols (notes, rests, accidentals) in the proof of concept, with about 90% synthetic files. The results from this were:

(1) We were able to create models for separating the types of symbols, and recognizing both pitch and duration of the notes, and duration of rests, and reached the stretch goal of the proof of concept. For separating the symbol types, an SVM model worked well, for note pitch and duration we use a 3-layer CNN.

(2) Working with scanned images was a lot harder than working with synthetic notes. That of course is not surprising, but the magnitude of the problems that we encountered was startling. For example, when we compared the MusicXML symbols of the first movement of the Mozart string quartet to the IMSLP score, we found 119 differences.

(3) When we applied the trained model to images from a different score (IMSLP 10870) with the same dpi, the accuracy dropped to 82% (the pages did look visibly grainier). It seemed obvious to us that we would need to increase the amount of training material from scanned images very substantially to achieve better results.

III. User Hopes

We took a step back and examined the hopes of a practicing musician whose main goal it is to quickly get to a usable score. This means:

- The musician will most likely search for a score in IMSLP. If he/she wants to additional services for the score (like transposing it, play the whole score or sections of the score at a desired speed, allow basic editing), there should be an option to open the score in a tool that supports these basic services (similar to opening a PDF with Google Docs).
- It should estimate the accuracy of the score.
- It should highlight obvious problems (e.g. the note and rest values not adding up to the time signature).

Is this a pipe dream? From a technical perspective we do not see anything that is impossible.

IV. WHAT IS MISSING?

While we were working on our proof of concept, we became aware of other projects pursuing the same goal on slightly different paths (see [1], [7], and especially [2]). These projects use much more data (at this point in time synthetic data), with incredibly sophisticated ML models which achieved very high accuracy.

What is missing (based on how little we know and understand):

(1) Currently only monophonic music is covered.

(2) Additional models (for instance for predicting the accuracy of OMR for a score, or identifying possible inconsistencies) are not available.

(3) It seems unclear how the models will perform on scanned images (like IMSLP scores).

(4) Machine learning systems that effectively and efficiently collect feedback can almost automatically improve over time. (An obvious example is Google Maps for driving instructions or identifying areas of interest.)

We are certain that (1) is already being addressed.

In our view (2) is something that can be addressed relatively easily once results for a larger number of scores are available. It may offer the opportunity to involve the machine learning community by conducting a Kaggle competition, for example for the prediction of page-level accuracy. (Training input would be the page image, the accuracy of the best available model, and set of page level attributes, with required deliverable an explainable model – one predicts accuracy, and explains the result.)

Another area for involving the machine learning community could be the use of distributed machine learning in the form of dynamic task graphs [3].

As for (3), based on our experience we are concerned about the performance of models on scanned images, especially since the quality of scans can vary substantially. There seems to be two ways to deal with the problems:

- One uses image augmentation to make the model more robust so it can handle lower quality inputs (as described in [7]).
- Creating substantial amounts of correctly labeled scanned images. This approach requires a lot of work. (We currently see no way to avoid this. If this is correct, the question is about the best way to achieve this. It may well require the involvement of the musician community.)

As [2] as shown, a machine-learning OMR system can be competitive against a leading commercial OMR tool. We believe that a machine learning system that manages to collect feedback, manages it to add to the training set, and trigger retraining of the model can potentially and measurably improve over time. This will require human involvement by the musician community to ensure that the quality of the training set remains high.

V. FINAL REMARKS

We feel that the hopes of practicing musician will not be fulfilled in the near future, but we are confident what we described is not a pipe dream. To get there will require time to make progress in the areas that we identified, which may require substantial involvement of the musician community to improve OMR. (In that sense it seems that crowd sourcing approach of OpenScore is right, and should be extended.) OMR would benefit from a higher awareness in, and possible involvement of the machine learning community. We think that this is almost necessary: compared to OCR, OMR is a much smaller market, with less attention by big players.

REFERENCES

[1] Jorge Calvo-Zaragoza, Jose J. Valero-Mas, Antonio Pertusa, “End-to-end Optical Music Recognition using Neural Networks”, 18th International Society for Music Information Retrieval Conference, Suzhou, China, 2017.

[2] Jorge Calvo-Zaragoza, David Rizo, “End-to-end Neural Optical Music Recognition of Monophonic Scores”, Applied Sciences, 2018, 8, 606K.

[3] Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, William Paul, Michael I. Jordan, Ion Stoica, “Ray: A Distributed Framework for Emerging AI Applications”, arXiv: 1712.05889V1, 16 Dec 2017

[4] Brad Neuberg, “Creating a Modern OCR Pipeline Using Computer Vision and Deep Learning”, Dropbox Tech Blog, https://blogs.dropbox.com/tech/2017/04/creating-a-modern-ocr-pipeline-using-computer-vision-and-deep-learning/, April 12, 2017

[5] Victor Padilla, Alex McLean, Alan Marsden & Kia Ng. “Improving Optical Music Recognition by Combining Outputs from Multiple Sources”. 16th International Society for Music Information Retrieval Conference, 2015

[6] Ana Rebelo, Ichiro Fujinaga, Filipe Paszkeiewicz, Andre R.S. Marcal, Carlos Guedes, Jaime S. Cardoso. “Optical Music Recognition - State-of-the-Art and Open Issues”, International Journal of Multimedia Information Retrieval, Vol. 1, No. 3, pp. 173-190, 2012.

[7] Eelen van der Wel, Karen Ulrich, “Optical Music-Recognition with Convolutional Sequence-to-Sequence Models”, 18th International Society for Music Information Retrieval Conference, Suzhou, China, 2017