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*—We describe our experience with building a simple optical music recognition system using machine learning, continue with what we believe the user wants, followed by how machine learning can contribute with better models and through community involvement, and final thoughts.**

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 as well: 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 a 99% pitch accuracy, it is quite possibly useless for her/him – the human ear will neither ignore nor forgive the errors.

II. Proof of Concept

As a family of musicians with different skill levels we wanted better access to symbolic music to get scores that fitted our needs. Since not much is available, we tried to use embedded OMR systems with poor results, and found the effort necessary to produce usable symbolic scores through manual work much too high. After learning about how Dropbox had combined computer vision and machine learning to approach OCR [4], we got excited enough to start working on a four-months proof of concept in August 2017, 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 systems that was analyzed in [5] for string quartets. Accuracy in the proof of concept is defined as getting positional pitch and duration right (slurs, accents, dynamics etc. are ignored.) This requires an accuracy of over 90% for pitch and duration. Stretch goal is to achieve an accuracy that is higher than the combined output of multiple sources (95%).

In [5] the comparison was based on Mozart string quartets. We found for string quartet K458 both an IMSLP file and an unrelated MusicXML encoding (IMSLP482550, MusicXML: 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: images were too small, included more than one symbol, or labels were matched to the wrong image. In some cases we found error rates of up to 8%.

In addition, we created synthetic images, mainly for completeness - we wanted all combinations of pitch (G3 to D6) and note length (whole to 1/128) represented. We rendered them both through MuseScore and Finale for difference in appearance.

The individual image files had different sizes, which were normalized to 48x144 pixel, and converted to 8-bit grayscale. The target vector that we extracted from MusicXML contained symbol type (note, rest), pitch, and duration. (The training data for accidentals was derived from synthetic music only.)

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 classifying 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. SVM models worked well for classification, and for pitch and duration we used 3-layer CNNs, with scikit-learn and TensorFlow as frameworks.

(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 unexpectedly high. When we found it difficult to match the images from IMSLP to the targets we had extracted from MusicXML in the first movement of the Mozart string quartet, we rendered the MusicXML and compared it measure by measure to the IMSLP score. We found 119 differences (none for pitch, all for duration – we ignored slurs/ties and other differences). Based on 4074 notes and 1025 rests in the first movement, the 2.3% difference may not look high, and the differences had little or no musical meaning: almost nobody would notice whether viola and cello play a dotted quarter note, or a quarter note

followed by a 1/8 rest (measure 4), or whether there is one ¼ rest or two 1/8 rests (measure 137). However, this complicates preparing training data substantially, since automatic matching between image and label is either not possible or can even be wrong.

(3) When we applied the trained model to additional 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.

(4) Looking back, there is one decision that we now think we got wrong: to focus on positional pitch and to treat accidentals as its own symbol type. It would have been better to treat the accidental as part of the note.

III. What do Musicians want?

We took a step back and, based on our own experience and talks others (members of Ryan's youth orchestra, the conductor of the youth orchestra, amateur musicians and music teachers in the US and Germany, and a composer) to see how musicians currently interact with scores:

- Musicians still buy scores, but everyone uses IMSLP, increasingly on iPads with products like forScore.
- A (small) fraction interacts with symbol music through notation software.
- A fraction of those use OMR software. (The people we talked to use what is bundled with the notation software, and most were with it.)

We came to believe that musicians want services enabled by symbolic music. Imagine the following:

- A musician searches for a score in IMSLP. If she/he wants 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 access the result of a high-quality OMR process (like opening a PDF with Google Docs), if the musician is satisfied with the predicted recognition accuracy of the score with an emphasizes pitch accuracy.
- The tool highlights obvious problems that the OMR process could not resolve (e.g. the note and rest values not adding up to the time signature).

We think that a lot of musicians would use this. Is this a pipe dream? From a technical perspective: no.

IV. What can machine Learning contribute?

Machine learning can support this in the following ways:

(1) Delivering high-accuracy models: [2], [7], and especially [3] have shown that with very sophisticated machine learning models high accuracy can be achieved, competitive with a leading commercial OMR tool.

(2) As soon good models and data exist, additional models (e.g. predicting pitch accuracy for a new page) become easier.

(3) Machine learning systems can improve over time once they effectively and efficiently collect feedback, and learn from it. (An obvious example is Google Maps for driving instructions or identifying areas of interest.)

This would require:

(a) Easily accessible data and pipelines: among other, both [3] and [7] point the way by making their data accessible.

(b) If image augmentation as described in [7] is not sufficient to handle lower quality inputs like scanned images from IMSLP PDFs, then a good way to get to enough training data is needed. In our view, this will require better tools and the involvement of the musician community.

(c) As for (3) we believe that involving the musician community will be key here as well, since feedback needs to be evaluated, for instance to decide whether and how it should be added to the training set, and how retraining is triggered and measured.

Finally, OMR topics seem to not be well-known in the machine learning community. We wonder whether exposing OMR problems in a Kaggle competition could help to change this. An example could be 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 being a model that not only predicts accuracy, but explains the results as well.)

V. Final remarks

We believe that progress with OMR will require the involvement of the musician and machine learning community. (In that sense, the approach of OpenScore is correct, but in our opinion too limited.) As for what we outlined in III. it is desirable and feasible [1]. Viability is another matter - in OCR there were Google, Dropbox and others, and we currently do not see their equivalent in this space.

References

[1] Tim Brown: "Change by Design: How Design Thinking Transforms Organizations and Inspires Innovation", 2009.

[2] 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.

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

[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