SELF-LEARNING OPTICAL MUSIC RECOGNITION

Alexander Pacha

E193-06 - Institute of Visual Computing and Human-Centered Technology

INTRODUCTION

Music is an essential part of our culture and our cultural heritage. It has been passed on through the centuries primarily in two forms: as aural transmission and in written documents, called music scores (see Figure 1). Reading music scores and playing or singing them requires years of studies, which many people cannot. Optical Music Recognition is the research field, that can help to overcome this problem by automating the decoding of written music into a machine-readable format, which allows for further processing, such as generating an audio file that a user can listen to.

PROBLEM STATEMENT

The idea of Optical Music Recognition (OMR) dates back to the 1960s and 1970s, when the first flat-bed scanners became available [1]. The classical approach is a pipeline of steps, that tries to find the structural elements of music scores, called staff lines, followed by several steps that try to isolate and extract the smaller symbols, called notes and rests, that convey the intended temporal sequence of sounds and rests. After finding those primitives, a semantic reconstruction attempts to recover the relationships between symbols as well as their notational semantics. Finally, the internal representation is exported into formats such as MIDI, which allows to play the music back to the user.

The whole process has often been referred to as Optical Character Recognition (OCR) for music. And although the two share many similarities, OMR presents a significantly harder challenge and has notoriously been underestimated, which resulted in the unsatisfactory result, that there exists no system which is capable of robustly recognizing music scores for any except the most simple cases.

The aim of this research is to improve Optical Music Recognition by breaking with the traditional pipeline and replacing it with a machine-learning approach, that makes use of the recent advances from the field of computer vision by the means of deep learning. Instead of using hand-crafted algorithms, that were tailored to work well on a particular dataset, a machine is trained to recognize music scores by itself from a large dataset, given initial human supervision.

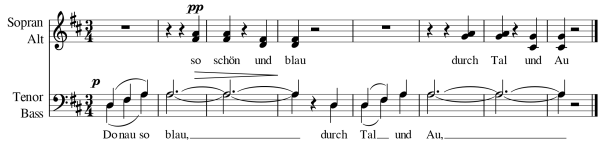

Figure 1: Beginning of Johann Strauss Junior's waltz "An der schönen blauen Donau", arranged for four voices, written in modern music notation.

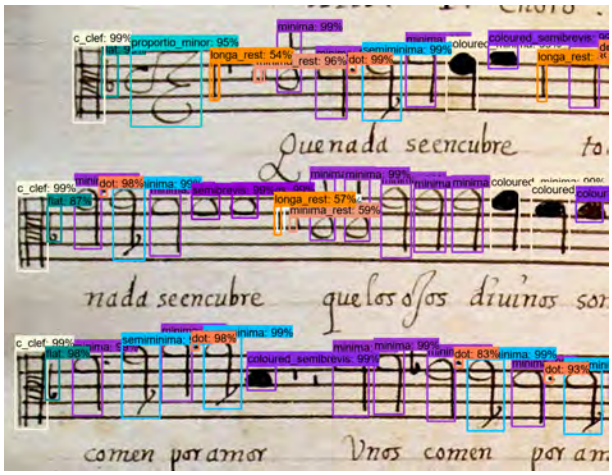

Figure 2: Hand-written music scores, written in mensural notation, with detected symbols, highlighted in boxes with their associated classes on top.

EXPERIMENTS

So far, we conducted several experiments, that tried to answer the following questions: Can a machine learn the concept of "what scores look like" and distinguish music scores from something else? [2] Can a machine learn to distinguish between isolated symbols just by providing enough samples of each class? [3] Can a machine learn to detect all symbols in hand-written music scores? [4]

For each of these questions, a deep convolutional neural network was trained on a large dataset of thousands of examples that were manually annotated by humans. For the last question, various state-of-the-art object detectors such as Faster R-CNN [5] were evaluated and adapted to work well on previously unseen data (see Figure 2). This breakthrough allows the research community to move on to other remaining challenges, such as the semantic reconstruction, which has to deal with a substantial amount of incomplete information and notational subtleties that previously found little attention, because researcher were struggling to solve the preceding steps.

RESULTS AND DISCUSSION

The conducted experiments showed very promising results that were comparable or even better than the performance of humans on the same task. A single neural network for example for capable of distinguishing 79 different classes of symbols with a precision of over 98% and the work on detecting music objects in the scores represents a milestone with detection results of over 80% mean average precision (mAP). For the first time, it is possible to accurately detect the full vocabulary of symbols in hand-written music scores, by just training a computer on a suitable dataset. Nevertheless, there is still plenty of room for improvement, before the machine is capable of reading music scores as good as humans.

CONCLUSION

This work has shown that with recent advances in the field of computer vision and deep learning, it is possible to replace a hand-crafted and often very limited process with an end-to-end trainable neural network, that is capable of learning abstract concepts and solving very specific problems with high accuracy, given the right approach and a sufficient amount of data. We will continue this way and aim towards a system, where the entire process of OMR is end-to-end trainable, allowing the computer to learn and improve by simply providing more data.

REFERENCES

[1] Rebelo A., Fujinaga I., Paszkiewicz F., Marcal A.R.S., Guedes, C., Caroso J.S, Optical music recognition: state-of-the-art and open issues, International Journal of Multimedia Information Retrieval, 2012

[2] Pacha A., Eidenberger H., Towards a Universal Music Symbol Classifier, Proceedings of the 12th IAPR International Workshop on Graphics Recognition, 2017

[3] Pacha A., Eidenberger H., Towards Self-Learning Optical Music Recognition, Proceedings of the 16th IEEE International Conference On Machine Learning and Applications, 2017

[4] Pacha A., Choi K.-Y., Coasnon B., Ricquebourg Y., Zanibbi R., Eidenberger H., Handwritten Music Object Detection: Open Issues and Baseline Results, 2018 13th IAPR Workshop on Document Analysis Systems (DAS), 2018 (in press)

[5] Ren S., He K., Girshick R., Sun J., Faster R-CNN: Towards real-time object detection with region proposal networks, Advances in neural information processing systems, 2016