MultiScore Project: Multimodal Transcription of Music Scores

Jorge Calvo-Zaragoza, Antonio Pertusa, Antonio-Javier Gallego, José M. Iñesta, Luisa Micó,
Jose Oncina, Carlos Pérez-Sancho, Pedro J. Ponce de León, David Rizo

U.I. for Computer Research, University of Alicante, Spain
{jcalvo,pertusa}@dlsi.ua.es

Optical Music Recognition (OMR) and Automatic Music Transcription (AMT) are the research fields that investigate how to computationally transcribe music score images and audio recordings, respectively, into machine-readable formats encoding symbolic data, namely digital scores. A digital score contains a series of symbols that can be subsequently rendered to be read by musicians, or analyzed by music algorithms to extract meaningful information. However, while musicians can read very complex music scores or transcribe them by ear, there is still no automatic system capable of doing so with comparable performance. These tasks remain a research challenge nowadays.

Despite the extensive literature around AMT, there are no reliable methods that can retrieve symbolic scores from audio recordings (Benetos et al., 2019). In general, most existing AMT approaches work in a pipeline-based fashion consisting of two basic steps: Multi-pitch estimation and Note tracking. This pipeline allows representing an audio file in a so-called piano-roll music notation format, which is a 2D visualization of music where the y-axis contains all the possible notes ordered by frequency and the x-axis represents their evolution in time. Notwithstanding, while piano-rolls retrieve some music insights from the audio signal, they are insufficient for further analysis or human understanding. A complete digital score should also contain meter, clef, voices, key, note lengths (half, quarter,...), barlines, ties and accidentals, among others.

Furthermore, the OMR process typically follows multi-stage approaches (Calvo-Zaragoza et al., 2020), as well. In this case, most existing systems first aim to detect all the primitives (graphic units that make up the notation), and then try to relate them to reconstruct the music scores. These two stages, in turn, consist of many other steps (image pre-processing, segmentation, classification, ...) that end up propagating errors that lead to several failures, since contextual information, which is key to decoding the meaning of a score, does not transfer from one stage to another.

The aforementioned state of the art is the starting point of the **Multimodal Transcription of Music Scores (MultiScore)** project. With the aim of making a relevant contribution to the current situation and building upon previous efforts, MultiScore proposes the development of neural models that leverage large data sets to learn both OMR and AMT holistically (end-to-end) under a common framework for transcribing music. This common formulation brings additional benefits such as unifying AMT and OMR, so far addressed with completely different approaches. This enables promoting synergies between both fields and pursuing new scientific tasks, such as developing common language models, multimodal audio and image transcription—that has preliminary obtained successful results (de la Fuente et al., 2021)—or learning unique models capable of dealing with both tasks independently (in a multi-task learning fashion).

Acknowledgements

This report is part of the project I+D+i PID2020-118447RA-I00, funded by MCIN/AEI/10.13039/50110001103.

References

Benetos, E., Dixon, S., Duan, Z., and Ewert, S. (2019). Automatic music transcription: An overview. *IEEE Signal Process. Mag.*, 36(1):20–30.

Calvo-Zaragoza, J., Jr., J. H., and Pacha, A. (2020). Understanding optical music recognition. *ACM Comput. Surv.*, 53(4):77:1–77:35.

de la Fuente, C., Valero-Mas, J. J., Castellanos, F. J., and Calvo-Zaragoza, J. (2021). Multimodal image and audio music transcription. *International Journal of Multimedia Information Retrieval.*