# MULTIMODAL AUDIO AND IMAGE MUSIC TRANSCRIPTION

**Carlos de la Fuente** **Jose J. Valero-Mas** **Francisco J. Castellanos**
**Jorge Calvo-Zaragoza** **María Alfaro-Contreras** **José M. Iñesta**
Department of Software and Computing Systems, University of Alicante, Alicante, Spain
cdlf4@alu.ua.es, {jjvalero, fcastellanos, jcalvo, malfaro, inesta}@dlsi.ua.es

## ABSTRACT

Optical Music Recognition (OMR) and Automatic Music Transcription (AMT) stand for the research fields which aim at obtaining a structured digital representation of the music content present in either a sheet music image or an acoustic recording, respectively. While these fields have historically evolved separately, the fact that both tasks share the same output representation poses the question of whether they could be combined in a multimodal framework that exploits the individual transcription advantages depicted by each modality in a synergistic manner. To assess this hypothesis, this work presents a proof-of-concept research piece that combines the predictions given by end-to-end AMT and OMR systems over a corpus of monophonic music pieces considering a local alignment approach. The results obtained, while showing a narrow improvement with respect to the best individual modality, validate our initial premise.

## 1. INTRODUCTION

The attainment of structured digital representations of music sources, typically known as *transcription*, remains as one of the key, yet challenging, tasks in the Music Information Retrieval (MIR) field [1]. Under this transcription framework, two particular research lines stand out within the MIR community: on the one hand, when tackling music scores, Optical Music Recognition (OMR) is the field that investigates how to computationally read music notation from these documents and to store them in a digital structured format [2]; on the other hand, when considering acoustic music signals, Automatic Music Transcription (AMT) represents the field that researches on the design of computational algorithms to transcribe them into some form of structured digital music notation [3].

Nevertheless, despite pursuing the same goal, these two fields have historically worked in a disjoint manner due to the different nature of the source data, either scores or acoustic pieces. However, what if, assuming that we have both a score and a recording of a performance of a music composition, the individual OMR and AMT systems be combined to obtain a digital transcription of the piece which benefits from the advantages of each method?

In this work, we aim at exploring, as a proof of concept, whether the transcription results of a multimodal combination of sheet scores and acoustic performances of music pieces improves those of the stand-alone modalities. For that, we consider a fusion policy based on the combination of the most probable hypotheses depicted by each source of data (prediction-level fusion) for monophonic compositions considering end-to-end OMR and AMT systems.



## 2. METHODOLOGY

We shall now describe the end-to-end neural architecture considered for both the OMR and AMT processes as well as the prediction-level fusion policy proposed.

### 2.1 End-to-end base recognition systems

Concerning the end-to-end neural architectures, we have considered a Convolutional Recurrent Neural Network (CRNN) scheme [4] together with the Connectionist Temporal Classification (CTC) training algorithm [5]. This network is formed by an initial block of *convolutional* layers devised to learn the adequate features for the particular recognition task followed by another group of *recurrent* stages which model the temporal/spatial dependencies of those features.

As commented, the network is trained using the CTC training function as it allows training the CRNN scheme using unsegmented sequential data. In a practical sense, this method only requires the different input signals to the scheme and their associated sequences of characters drawn from vocabulary $\Sigma$ as its expected output, without any specific input-output alignment. It must be mentioned that CTC requires the inclusion of an additional "*blank*" symbol within the set of considered symbols, i.e., $\Sigma' = \Sigma \cup \{blank\}$ due to its particular training procedure. This symbol is used for enabling the detection of consecutive repeated elements.

Since CTC assumes that the architecture contains a fully-connected network of $|\Sigma'|$ outputs with a *softmax* activation, the actual output is a posteriogram with a number of frames given by the recurrent stage with $|\Sigma'|$ tokens each. Most commonly the final prediction is obtained

out of this posteriogram using a *greedy* approach which retrieves the most probable symbol per step and a posterior squash function which merges consecutive repeated symbols and removes the *blank* label. In our case, we slightly modify this decoding approach for allowing the multimodal fusion of both sources of information.

### 2.2 Fusion policy

The proposed policy takes as starting point the posteriograms of the two recognition modalities, OMR and AMT. For each posteriogram, a greedy decoding policy is applied to each of them for obtaining their most probable symbols per frame together with their per-symbol probabilities.

After that, the CTC squash function merges consecutive symbols for each modality with the particularity of deriving the per-symbol probability by averaging the individual probability values of the merged symbols. For example, when any of the models obtains a sequence in which it predicts the same symbol for 4 consecutive frames, the algorithm combines them and computes the average probabilities of these involved frames. Note that the *blank* symbols estimated by CTC are also removed.

Given that the resulting sequences for each modality may not match in terms of length, it is necessary to align both estimations for properly merging them. In this regard, we make use of the Smith-Waterman (SW) local alignment algorithm [6] which performs a search for the most similar regions between pairs of sequences.

Eventually, the final estimation is obtained from these two aligned sequences following these premises: (i) if both sequences match on a token, it is included in the resulting estimation; (ii) if the sequences disagree on a token, the one with the highest probability is included in the estimation; (iii) if one of the sequences poses a *blank* symbol, that of the other sequence is included in the estimation.

## 3. EXPERIMENTATION

For the evaluation of our approach, we considered the Camera-based Printed Images of Music Staves (Camera-PrIMuS) database [7]. This corpus contains 87,678 real music staves of monophonic incipits [1] extracted from the *Répertoire International des Sources Musicales* (RISM). For each incipit, different representations are provided: an image with the rendered score (both plain and with artificial distortions), several encoding formats for the symbol information, and a MIDI file of the content.

Regarding the particular type of data used by each recognition model, the OMR system takes as input the artificially distorted staff image of the incipit scaled to a height of 64 pixels, maintaining the aspect ratio. Regarding the AMT model, an audio file is synthesized from the MIDI file for each incipit with the FluidSynth software [2] and a piano timbre considering a sampling rate of 22,050 Hz; then a time-frequency representation is obtained by means of the Constant-Q Transform with a hop length of 512 samples, 120 bins, and 24 bins per octave. This result is embedded as an image whose height is scaled to 256 pixels, maintaining the aspect ratio.

Table 1 summarizes the details of the data considered for each modality and partition after data curation and balancing processes.

**Table 1**. Number of incipits considered for each modality and partition. In preliminary experimentation, the OMR system remarkably outperformed the AMT one, thus we reduced the training set of the OMR system not to eclipse the possible contribution of AMT to the combined result.

| Modality | Train | Validation | Test |
|---|---|---|---|
| OMR (Image) | 802 | 4,457 | 4,457 |
| AMT (Audio) | 13,371 | 4,457 | 4,457 |

Regarding the performance evaluation, we considered the Symbol Error Rate (Sym-ER) as in other neural-based transcription systems. This measure is defined as the average number of elementary editing operations (insertions, deletions, or substitutions) necessary to match the predicted sequence with the ground truth one, normalized by the length of the latter.

The results obtained with the experimental set-up considered for the AMT and OMR systems as well as the presented fusion policy are depicted in Table 2. It must be pointed out that these results constitute the ones achieved after optimizing the alignment parameters of the SW algorithm on the validation partition.

**Table 2**. Symbol Error Rate result for the OMR, AMT, and fusion policy for the test partition considered.

| Metric | OMR | AMT | Fusion |
|---|---|---|---|
| Sym-ER (%) | 14.29 | 27.53 | 12.95 |

As it can be observed, the stand-alone OMR method consistently outperforms the AMT one, achieving the former system a Sym-ER figure approximately $13\%$ lower than that of the latter. In this context, one could argue that combining the outputs of these two systems may report an improvement due to being the OMR system considerably more robust than the AMT one. Nevertheless, the fusion method is able to decrease the error rate obtained by the best transcription model when the alignment method is properly adjusted. More precisely, the fusion method achieves a Sym-ER $1.4\%$ lower than that of the OMR model, i.e. the error is reduced over $9.4\%$.

Finally, it must be highlighted that the improvement over a $9.4\%$ of the Sym-ER metric supports the initial hypothesis that the multimodal combination of OMR and AMT technologies may enhance that of stand-alone systems, being, hence, worthwhile studying this new paradigm for transcription tasks. Note that this work constitutes a proof-of-concept research piece meant to validate the commented hypothesis.

[1] Short sequence of notes, typically the first measures of the piece, used for indexing and identifying a melody or musical work.

[2] https://www.fluidsynth.org/

## 4. ACKNOWLEDGMENTS

This work is supported by the Spanish "Ministerio de Educación y Formación Profesional" through grant 20CO1/000966, the "Programa I+D+i de la Generalitat Valenciana" through grants ACIF/2019/042 and APOSTD/2020/256 and the Spanish "Ministerio de Universidades" through grant FPU19/04957.

## 5. REFERENCES

[1] X. Serra, M. Magas, E. Benetos, M. Chudy, S. Dixon, A. Flexer, E. Gómez Gutiérrez, F. Gouyon, P. Herrera, S. Jordà *et al.*, *Roadmap for music information research*. The MIReS Consortium, 2013.

[2] J. Calvo-Zaragoza, J. Hajič Jr, and A. Pacha, "Understanding optical music recognition," *ACM Computing Surveys (CSUR)*, vol. 53, no. 4, pp. 1–35, 2020.

[3] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2018.

[4] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298–2304, 2017.

[5] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 369–376.

[6] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0022283681900875

[7] J. Calvo-Zaragoza and D. Rizo, "Camera-PrIMuS: Neural End-to-End Optical Music Recognition on Realistic Monophonic Scores," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*, Paris, France, Sep. 2018, pp. 248–255. [Online]. Available: https://doi.org/10.5281/zenodo.1492395