International Journal of Multimedia Information Retrieval (2022) 11:77–84
https://doi.org/10.1007/s13735-021-00221-6

REGULAR PAPER

Multimodal image and audio music transcription

Carlos de la Fuente[1] · Jose J. Valero-Mas[1] · Francisco J. Castellanos[1] · Jorge Calvo-Zaragoza[1]

Received: 13 July 2021 / Revised: 26 September 2021 / Accepted: 6 October 2021 / Published online: 11 November 2021

Abstract
Optical Music Recognition (OMR) and Automatic Music Transcription (AMT) stand for the research fields that aim at obtaining a structured digital representation from sheet music images and acoustic recordings, respectively. While these fields have traditionally evolved independently, the fact that both tasks may share the same output representation poses the question of whether they could be combined in a synergistic manner to exploit the individual transcription advantages depicted by each modality. To evaluate this hypothesis, this paper presents a multimodal framework that combines the predictions from two neural end-to-end OMR and AMT systems by considering a local alignment approach. We assess several experimental scenarios with monophonic music pieces to evaluate our approach under different conditions of the individual transcription systems. In general, the multimodal framework clearly outperforms the single recognition modalities, attaining a relative improvement close to 40% in the best case. Our initial premise is, therefore, validated, thus opening avenues for further research in multimodal OMR-AMT transcription.

Keywords Multimodal recognition · Automatic music transcription · Optical music recognition and deep learning

1 Introduction

Bringing music sources into a structured digital representation, typically known as *transcription*, remains as one of the key, yet challenging, tasks in the Music Information Retrieval (MIR) field [17,21]. Such digitization not only improves music heritage preservation and dissemination [11], but it also enables the use of computer-based tools which allow indexing, analysis, and retrieval, among many other tasks [20].

In this context, two particular research lines stand out within the MIR community: on the one hand, when tackling music scores images, the field of Optical Music Recognition (OMR) investigates how to computationally read these documents and store their music information in a symbolic format [3]; on the other hand, when considering acoustic music signals, Automatic Music Transcription (AMT) represents the field devoted to the research on computational methods for transcribing them into some form of structured digital music notation [1]. It must be remarked that, despite pursuing the same goal, these two fields have been developed separately due to the different nature of the source data.

Multimodal recognition frameworks, understood as those which take as input multiple representations or modalities of the same piece of data, have proved to generally achieve better results than their respective single-modality systems [25]. In such schemes, it is assumed that the different modalities provide complementary information to the system, which eventually results in an enhancement of the overall recognition performance. Such approaches are generally classified in one of these fashions [7]: (i) those in which the individual features of the modalities are directly merged with the constrain of requiring the input elements to be synchronized to some extent (*feature* or *early-fusion* level); or those in which the merging process is done with the hypotheses obtained by each individual modality, thus not requiring both systems to be synchronized (*decision* or *late-fusion* level).

Regarding the MIR field, this premise has also been explored in particular cases as music recommendation, artist identification or instrument classification, among others [22].

✉ Carlos de la Fuente
cdlf4@alu.ua.es

Jose J. Valero-Mas
jjvalero@dlsi.ua.es

Francisco J. Castellanos
fcastellanos@dlsi.ua.es

Jorge Calvo-Zaragoza
jcalvo@dlsi.ua.es

[1] Department of Software and Computing Systems, University of Alicante, Alicante, Spain

Music transcription is no strange and has also contemplated the use of multimodality as a means of solving certain glass ceiling reached in single-modality approaches. For instance, research on AMT has considered the use of additional sources of information as, for instance, onset events, harmonic information, or timbre [2]. Nevertheless, to our best knowledge, no existing work has considered that a given score image and its acoustic performance may be considered two different modalities of the same piece to be transcribed. Under this premise, transcription results may be enhanced if the individual, and somehow complementary, descriptions by the OMR and AMT systems are adequately combined.

While this idea might have been discussed in the past, we consider that classical formulations of both OMR and AMT frameworks did not allow exploring a multimodal approach. However, recent developments in these fields define both tasks in terms of a *sequence labeling* problem [10], thus enabling research on the combined paradigm. Note that when addressing transcription tasks within this formulation, the input data (either image or audio) is directly decoded into a sequence of music-notation symbols, having this typically been carried out considering neural end-to-end systems [4,19].

One could argue whether it may be practical, or even realistic, having both the acoustic and image representations of the piece to be transcribed. We assume, however, that for a music practitioner it would be, at least, more appealing to play a composition reading a music sheet rather than manually transcribing it. Note that we find the same scenario in the field of Handwritten Text Recognition, where producing a uttering out of a written text and using a speech recognition system for then fusing the decisions required less effort than manually transcribing the text or correcting the errors produced by the text recognition system [8].

This work explores and studies 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 propose a decision-level fusion policy based on the combination of the most probable symbol sequences depicted by two end-to-end OMR and AMT systems. The experiments have been performed with a corpus of monophonic music considering multiple scenarios which differ in the manner the individual transcription systems are trained, hence allowing a thorough analysis of the proposal. The results obtained prove that the combined approach improves the transcription capabilities with respect to single-modality systems in cases in which their individual performances do not remarkably differ. This fact validates our initial premise and poses new research questions to be addressed and explored.

The rest of the paper is structured as follows: Sect. 2 contextualizes the work within the related literature; Sect. 3 describes our multimodal framework; Sect. 4 presents the experimental set-up considered as well as results and discussion; finally, Sect. 5 concludes the work and poses future research.

2 Related work

While multimodal transcription approaches based on the combination of OMR and AMT have not been yet explored in the MIR field, we may find some research examples in the related areas of Text Recognition (TR) and Automatic Speech Recognition (ASR). It must be noted that the multimodal fusion in these cases is also carried out at the decision level, keeping the commented advantage of not requiring multimodal training data for the underlying models.

One of the first examples in this regard is the proposal by Singh et al. [23], in which TR and ASR where fused in the context of postal code recognition using a heuristic approach based on the Edit distance [14]. More recent approaches related to handwritten manuscripts have resorted to probabilistic frameworks for merging the individual hypotheses by the systems as those of using *confusion networks* [8] or the *word-graph* hypothesis spaces [9].

It is worth noting that this type of multimodality may be also found in other fields as now the Gesture Recognition (GR) one. For instance, the work by Pitsikalis et al. [16] improves the recognition rate by re-scoring the different hypotheses of the GR model with information from an ASR system. Within this same context other works have explored the alignment of different hypotheses using Dynamic Programming approaches [15] or, again, a *confusion networks* framework [13].

In this work, we tackle this multimodal music transcription problem considering the alignment, at a sequence level, of the individual hypotheses depicted by stand-alone end-to-end OMR and AMT systems. As it will be shown, when adequately configured, this approach is capable of successfully improving the recognition rate of the single-modality transcription systems.

3 Methodology

We consider two neural end-to-end transcription systems as the base OMR and AMT methods for validating our fusion proposal. As commented, the choice of these particular approaches is that they allow a common formulation of the individual modalities, thus facilitating the definition of a fusion policy. Note that, in this case, the combination policy works at a decision, or sequence, level, as it can be observed in Fig. 1. To properly describe these design principles, we shall introduce some notation.

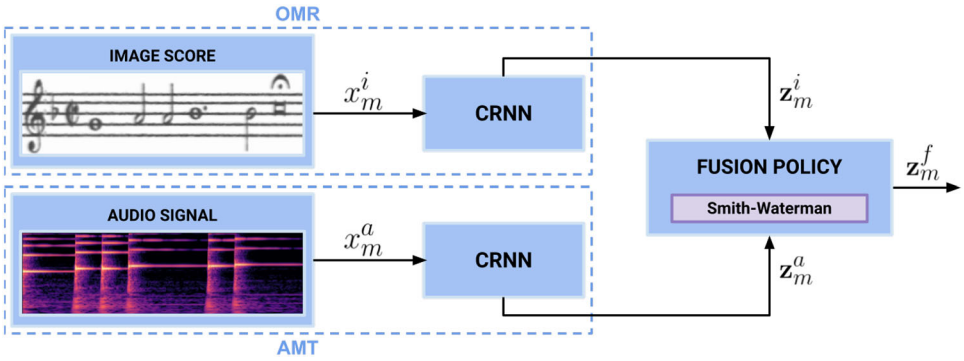

Fig. 1 Graphical description of the scheme proposed. For a given music piece, a score image x_m^i and an audio signal (as a CQT spectrogram) x_m^a are provided to the OMR and AMT systems, retrieving sequences $\mathbf{z}_m^a$ and $\mathbf{z}_m^i$, respectively. The multimodal fusion policy eventually produces the sequence $\mathbf{z}_m^f$

Let $\mathcal{T} = \{(x_m, \mathbf{z}_m) : x_m \in \mathcal{X},\ \mathbf{z}_m \in \mathcal{Z}\}_{m=1}^{|\mathcal{T}|}$ represent a set of data where sample x_m drawn from space $\mathcal{X}$ corresponds to symbol sequence $\mathbf{z}_m = (z_{m1}, \ldots, z_{mN_m})$ from space $\mathcal{Z}$ considering the underlying function $g : \mathcal{X} \rightarrow \mathcal{Z}$. Note that the latter space is defined as $\mathcal{Z} = \Sigma^*$ where Σ represents the score-level symbol vocabulary.

Since we are dealing with two sources of information, we have different representation spaces $\mathcal{X}^i$ and $\mathcal{X}^a$ with vocabularies Σ^i and Σ^a related to the image scores and audio signals, respectively. While not strictly necessary, for simplicity we are constraining both systems to consider the same vocabulary, i.e., $\Sigma^i = \Sigma^a$. Also note that, for a given m-th element, while staff $x_m^i \in \mathcal{X}^i$ and audio $x_m^a \in \mathcal{X}^a$ signals depict a different origin, the target sequence $\mathbf{z}_m \in \mathcal{Z}$ is deemed to be the same.

3.1 Neural end-to-end base recognition systems

Concerning the recognition architectures, we consider a Convolutional Recurrent Neural Network (CRNN) scheme to approximate $g(\cdot)$. Recent works have applied this approach to both OMR [5,6] and AMT [18,19] transcription systems with remarkably successful results. Hence, we shall resort to these works to define our baseline single-modality transcription architectures within the multimodal framework.

More in depth, a CRNN architecture is formed by an initial block of *convolutional* layers devised to learn the adequate features for the task at issue followed by another group of *recurrent* layers that model their temporal dependencies. To achieve an end-to-end system with such architecture, CRNN models are trained using the Connectionist Temporal Classification (CTC) algorithm [10]. In a practical sense, this algorithm only requires the different input signals and their associated transcripts as sequences of symbols, without any specific input-output alignment at a finer level. Note that CTC requires the inclusion of an additional "*blank*" symbol within the Σ vocabulary, i.e., $\Sigma' = \Sigma \cup \{blank\}$ due to its training procedure.

Since CTC assumes that the architecture contains a fully-connected layer of $|\Sigma'|$ outputs with a *softmax* activation, the actual output is a posteriogram with a number of frames given by the recurrent stage and $|\Sigma'|$ activations 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.

3.2 Multimodal 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 the same symbol is predicted for 4 consecutive frames, the algorithm combines them and computes the average probabilities of these involved frames. After that, the *blank* symbols estimated by CTC are also removed, retrieving predictions $\mathbf{z}_m^i$ and $\mathbf{z}_m^a$, which correspond to the image and audio recognition models, respectively.

Since sequences $\mathbf{z}_m^i$ and $\mathbf{z}_m^a$ may not match in terms of length, it is necessary to align both estimations for merging them. Hence, we consider the Smith-Waterman (SW) local alignment algorithm [24], which performs a search for the most similar regions between pairs of sequences.

Eventually, the final estimation $\mathbf{z}_m^f$ 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 misses a symbol, that of the other sequence is included in the estimation.

4 Experiments

Having defined the individual recognition systems as well as the multimodal fusion proposal, this section presents the experimental part of the work. For that, we introduce the CRNN schemes considered for OMR and AMT, we describe the corpus and metrics for the evaluation, and finally we present and discuss the results obtained. As previously stated, the combination of OMR and AMT has not been previously addressed in the MIR field. Hence, the experimental section of the work focuses on comparing the performance of the multimodal approach against that of the individual transcription models, given that no other results can be reported from the literature.

4.1 CRNN models

The different CRNN topologies considered for both the OMR and the AMT systems are described in Table 1. These configurations are based on those used by recent works addressing the individual OMR and AMT tasks as a sequence labeling problem with deep neural networks [4,19]. It is important to highlight that these architectures can be considered as the state of the art in the aforementioned transcription tasks, thus being good representatives of the attainable performance in each of the baseline cases. Note that, as aforementioned, the last recurrent layer of the schemes is connected to a dense unit with $|\Sigma^i| + 1 = |\Sigma^a| + 1$ output neurons and a softmax activation.

These architectures were trained using the backpropagation method driven by CTC for 115 epochs using the ADAM optimizer [12]. Batch size was fixed to 16 for the OMR system, while for the AMT it was set 1 because of being more memory-intensive.

4.2 Materials

For the evaluation of our approach, we considered the Camera-based Printed Images of Music Staves (Camera-PrIMuS) database [4]. 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. Although this dataset does not represent the hardest challenge for OMR or AMT, it provides both audio and images of the same pieces while allowing an artificial control of the performances for studying different scenarios.

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

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, while maintaining the aspect ratio. Concerning 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.

An initial data curation process was applied to the corpus for discarding samples which may cause a conflict in the combination, resulting in 67,000 incipits.[3] Since this reduced set still contains a considerably large amount of elements, we randomly selected approximately a third of this curated set for our experiments to take a considerable amount of memory and time, resulting in 22,285 incipits with a label space of $|\Sigma^i| = |\Sigma^a| = 1,180$ tokens. Eventually, we derive three partitions—train, validation, and test—which correspond to the 60%, 20%, and 20% of the latter amount of data, respectively.

With regard to the performance evaluation, we considered the Symbol Error Rate (SER) as in other neural end-to-end transcription systems [4,19]. This measure is defined as:

$$\text{SER}\ (\%) = \frac{\sum_{m=1}^{|\mathcal{S}|} \text{ED}\left(\mathbf{z}_m,\ \mathbf{z}'_m\right)}{\sum_{m=1}^{|\mathcal{S}|} |\mathbf{z}_m|} \tag{1}$$

where $\text{ED}\,(\cdot,\cdot)$ stands for the string Edit distance, $\mathcal{S}$ a set of test data, and $\mathbf{z}_m$ and $\mathbf{z}'_m$ the target and estimated sequences, respectively.

4.3 Results

In preliminary experimentation, when training both the OMR and AMT systems with the same amount of data, the former one depicted a remarkably better performance. This fact hindered the possible improvement of the multimodal proposal as the AMT recognition model rarely corrected any flaw of the (almost perfect) OMR one. Thus, we propose four controlled scenarios with the goal of thoroughly analyzing the multimodal transcription proposal.

For the sake of compactness, all the results are depicted in Table 2 while the following sections provide an individual analysis for each case. A last additional section further explores the results to analyze the error typology by each

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

[3] This is the case of samples containing long multi-rests, which barely extend the length of the score image but take many frames in the audio signal.

Table 1 CRNN configurations considered

Model	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6
OMR	Conv(64, 5 × 5)	Conv(64, 5 × 5)	Conv(128, 3 × 3)	Conv(128, 3 × 3)		
	BatchNorm	BatchNorm	BatchNorm	BatchNorm	BLSTM(256)	BLSTM(256)
	LeakyReLU(0.20)	LeakyReLU(0.20)	LeakyReLU(0.20)	LeakyReLU(0.20)	Dropout(0.50)	Dropout(0.50)
	MaxPool(2 × 2)	MaxPool(1 × 2)	MaxPool(1 × 2)	MaxPool(1 × 2)		
AMT	Conv(8, 2 × 10)	Conv(8, 5 × 8)				
	BatchNorm	BatchNorm	BLSTM(256)	BLSTM(256)		
	LeakyReLU(0.20)	LeakyReLU(0.20)	Dropout(0.50)	Dropout(0.50)		
	MaxPool(1 × 2)	MaxPool(1 × 2)				

Notation: Conv(f, $w \times h$) stands for a convolution layer of f filters of size $w \times h$ pixels, BatchNorm performs the normalization of the batch, LeakyReLU(α) represents a leaky rectified linear unit activation with negative slope value of α, MaxPool2D($w_p \times h_p$) stands for the max-pooling operator of dimensions $w_p \times h_p$ pixels, BLSTM(n) denotes a bidirectional long short-term memory unit with n neurons, and Dropout(d) performs the dropout operation with d probability

Table 2 Symbol error rate (%) results for the OMR, AMT, and fusion policy for the scenarios considered

Scenario	OMR (%)	AMT (%)	Fusion (%)
A	26.09	27.53	18.56
B	18.57	27.53	15.14
C	10.82	11.64	6.64
D	2.38	27.53	5.70

transcription method as well as the incorrect hypotheses the fusion policy is able to correct.

4.3.1 Scenario A: $\mathbf{SER_{OMR} \sim SER_{AMT}}$

This first scenario poses the case in which the OMR and AMT systems depict a similar performance. For obtaining such situation, we reduced the training data of the OMR to, approximately, a 2% of the initial partition considered while that of the AMT system remained unaltered. Under these conditions, the individual OMR and AMT frameworks achieve error rates of 26.09% and 27.53%, respectively.

As it may be checked, the proposed fusion policy reduces the error rate to a figure of 18.56%, which supposes a relative error decrease of approximately 28.86% with respect to that of the OMR system. This fact suggests that the fusion policy somehow exhibits a synergistic behavior in which the resulting sequence takes the most accurate estimations of the OMR and AMT transcription methods.

4.3.2 Scenario B: $\mathbf{SER_{OMR} < SER_{AMT}}$

The second scenario shows the case in which the individual performance of one of the transcription systems is considerably superior than that of the other one. For that, we reduced the training data devoted to the OMR system to, approximately, a 3% of the initial partition considered, remaining AMT unaltered.

With this particular configuration the starting point is that OMR improves the error rate of AMT in, approximately, a 9%. While such difference may, in principle, suggest that no improvement would be expected, it is eventually observed that the fusion decreases the error rate to 15.14%, which supposes a relative improvement of almost 19% with respect to the OMR system.

This experiment shows that, even in cases where a modality depicts a better performance than the other one, there is still a margin for improvement.

4.3.3 Scenario C: $\mathbf{SER_{OMR} \sim SER_{AMT} \Downarrow\Downarrow}$

The third posed scenario considers the case in which both transcription systems also achieve similar recognition rates but with a remarkably better performance than those shown in *Scenario A*. To artificially increase the performance of the AMT process, we removed the music incipits from the test set whose error was superior to 30% according to this model. After the process, the number of elements in this test partition is reduced to a 60% of the initial size while the others remain as in *Scenario B*.

In this case, the error rates depicted by the individual systems range between 10% and 11%, which already represent competitive transcription figures, at least in this type of architectures. However, when combining both modalities, the error rate decreases to 6.64%, which represents a relative improvement of, roughly, a 40%.

This particular experiment proves that, even in cases where both stand-alone transcription methods report competitive performances, the multimodal framework may report a noticeable benefit in the recognition process.

Table 3 Example of the multimodal fusion on a music incipit

OMR	AMT	Fusion	Ground truth
Clef-G2	**Clef-C1**	Clef-G2	Clef-G2
KeySignature-FM	–	KeySignature-FM	KeySignature-FM
TimeSignature-C	TimeSignature-C	TimeSignature-C	TimeSignature-C
Rest-half	Rest-half	Rest-half	Rest-half
Note-A4_eighth	Note-A4_eighth	Note-A4_eighth	Note-A4_eighth
Note-D5_eighth	Note-D5_eighth	Note-D5_eighth	Note-D5_eighth
Note-D5_sixteenth	Note-D5_sixteenth	Note-D5_sixteenth	Note-D5_sixteenth
Note-C5_sixteenth	**Note-C#5_sixteenth**	Note-C#5_sixteenth	Note-C#5_sixteenth
Note-D5_sixteenth	Note-D5_sixteenth	Note-D5_sixteenth	Note-D5_sixteenth
Note-E5_sixteenth	Note-E5_sixteenth	Note-E5_sixteenth	Note-E5_sixteenth
Barline	Barline	Barline	Barline
Note-F5_eighth	Note-F5_eighth	Note-F5_eighth	Note-F5_eighth
Note-D5_eighth	Note-D5_eighth	Note-D5_eighth	Note-D5_eighth
Rest-eighth	Rest-eighth	Rest-eighth	Rest-eighth
Note-C5_eighth	**Note-C#5_eighth**	Note-C#5_eighth	Note-C#5_eighth
Note-D5_eighth	Note-D5_eighth	Note-D5_eighth	Note-D5_eighth

The OMR and AMT columns depict the estimated sequences by the stand-alone systems while the Fusion one shows the combined estimation. The ground-truth transcription is also provided. Disagreements between modalities are highlighted in bold

4.3.4 Scenario D: $\mathbf{SER_{OMR} \ll SER_{AMT}}$

In this last scenario, we pose the case where one of the systems greatly outperforms the other one. For that, we have considered the original data partitions introduced in Sect. 4.2 for both OMR and AMT transcription systems.

In this particular case, it may be observed that the OMR model achieves an individual SER of 2.38%, while the AMT one remains at 27.53%. As expected, when fusing the two sources of information, the error increases to 5.70%, which supposes a remarkable performance decrease compared to the system achieving the best results, i.e., the OMR one.

Not surprisingly, when one of the modalities has a very limited room for improvement, these results show that the multimodal framework is not expected to bring any benefit.

4.3.5 Multimodal fusion example

The previously posed scenarios show the performance of the multimodal music transcription framework proposed, on a macroscopic level. Hence, we shall now analyze in detail the actual behavior of the method. For that Table 3 shows an example of the results obtained for a given incipit with the OMR and AMT systems, as well with the multimodal fusion proposed. The reference transcription is also provided.

A first point which can be observed is that, for this particular case, there is a strong agreement between the OMR and AMT modalities, being only four cases in which the two sequences estimate different labels: one related to the clef, another one for the key signature, and the remaining related to actual music notes. We shall now examine how these conflicts are solved by the merging policy.

Focusing on the clef and key errors, note that the devised fusion policy estimates the correct labels to be the ones by the OMR recognition system. Given that this disagreement is solved, on a broad sense, by taking the token with a superior probability among the different modalities, it is possible to affirm that the OMR performs better on this particular information than the AMT system. This conclusion is no strange since these two data (clef and key) are explicitly drawn in the score image while, for the case of audio data, this information must be inferred.

Furthermore, the errors present in the notes of the piece are better estimated by the AMT system rather than the OMR one. Again, this behavior is very intuitive since, while the note information is explicitly present in the audio data, in a score some information is elided due to the graphical representation rules. As an example, if the music piece depicts pitch alterations (sharp and/or flat notes), this information is explicitly engraved in the key signature of the piece and not represented with the notes to be recognized; oppositely, acoustic data directly contains the note with its possible alteration in the audio stream.

Finally, it must be remarked that the relative improvement in terms of error rate of almost a 40% achieved in *Scenerio C* supports the initial hypothesis that the multimodal combination of OMR and AMT technologies may enhance that of stand-alone systems, at least in some particular scenarios where there is margin for improvement. This facts endorses

the idea of further studying this new multimodal image and audio paradigm for music transcription tasks.

5 Conclusions

Music transcription, understood as obtaining a structured digital representation of the content of a given music source, is deemed as a key challenge in the Music Information Retrieval (MIR) field for its applicability in a wide range of tasks including music heritage preservation, dissemination, and analysis, among others.

Within this MIR field, depending on the nature of the data at issue, transcription is approached from either the Optical Music Recognition (OMR) perspective if dealing with image scores or the so-called Automatic Music Transcription (AMT) when tackling acoustic recordings. While these fields have historically evolved separately, the fact that both tasks may represent their expected outputs in the same way allows developing a synergistic framework with which achieving a more accurate transcription.

This work presents a first proposal that combines the predictions depicted by a couple of neural end-to-end OMR and AMT systems considering a local alignment approach over different scenarios dealing with monophonic music data. The results obtained validate our initial hypothesis that the multimodal combination of these two sources of information is capable of retrieving an improved transcription result. While the actual improvement depends on the scenario considered, our results attain up to around 40% of relative error improvement with respect to the single-modality transcription systems. It must be also pointed out that, out of the different scenarios posed, the only case in which the multimodal fusion proposed does not imply any benefit is when one of the modalities remarkably outperforms the other one and reaches an almost perfect performance.

In light of these results, different research avenues may be explored to further improve the results obtained. The first one is the actual combination of the hypotheses depicted by the individual systems on a probabilistic framework, such as that of *word graphs* or *confusion networks*. In addition, while these proposals work on a prediction-level combination, it may be also explored the case in which this fusion is done in previous stages of the pipeline as, for instance, the feature extraction one. Finally, experimentation may be also extended to more challenging data as handwritten scores, different instrumentation, or polyphonic music.

Author Contributions C.F., J.J.V.-M., F.J.C. and J.C.-Z. made equally contributions as regards the conception of the work, the experimental work, the data analysis, and writing the paper.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This research was partially funded by the Spanish "Ministerio de Ciencia e Innovación" through project MultiScore (PID2020-118447RA-I00). The first author acknowledges the support from the Spanish "Ministerio de Educación y Formación Profesional" through grant 20CO1/000966. The second and third authors acknowledge support from the "Programa I+D+i de la Generalitat Valenciana" through grants ACIF/2019/042 and APOSTD/2020/256, respectively.

Data availability Data are available from the authors upon request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This paper contains no cases of studies with human participants performed by any of the authors.

Code availability Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Benetos E, Dixon S, Duan Z, Ewert S (2018) Automatic music transcription: an overview. IEEE Signal Process Mag 36(1):20–30
2. Benetos E, Dixon S, Giannoulis D, Kirchhoff H, Klapuri A (2013) Automatic music transcription: challenges and future directions. J Intell Inf Syst 41(3):407–434
3. Calvo-Zaragoza J, Hajič J Jr, Pacha A (2020) Understanding optical music recognition. ACM Comput Surv (CSUR) 53(4):1–35
4. Calvo-Zaragoza J, Rizo D (2018) 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, pp. 248–255. Paris, France
5. Calvo-Zaragoza J, Toselli AH, Vidal E (2017) Handwritten music recognition for mensural notation: formulation, data and baseline results. In: 14th IAPR International conference on document analysis and recognition, vol. 1, pp. 1081–1086
6. Calvo-Zaragoza J, Valero-Mas JJ, Pertusa A (2017) End-to-end optical music recognition using neural networks. In: Proceedings of the 18th international society for music information retrieval conference, pp. 472–477. Suzhou, China

7. Dumas B, Signer B, Lalanne D (2012) Fusion in multimodal interactive systems: an hmm-based algorithm for user-induced adaptation. In: Proceedings of the 4th ACM SIGCHI symposium on Engineering interactive computing systems, pp. 15–24
8. Granell E, Martínez-Hinarejos CD (2015) Multimodal output combination for transcribing historical handwritten documents. In: International conference on computer analysis of images and patterns, pp. 246–260. Springer
9. Granell E, Martínez-Hinarejos CD, Romero V (2018) Improving transcription of manuscripts with multimodality and interaction. In: Proceedings of IberSPEECH, pp. 92–96
10. Graves A, Fernández S, Gomez F, Schmidhuber J (2006) Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on machine learning, pp. 369–376. New York, USA
11. Iñesta JM, Ponce de León PJ, Rizo D, Oncina J, Micó L, Rico-Juan JR, Pérez-Sancho C, Pertusa A (2018) Hispamus: Handwritten spanish music heritage preservation by automatic transcription. In: 1st International workshop on reading music systems, pp. 17–18
12. Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In: 3rd International conference on learning representations. San Diego, USA
13. Kristensson PO, Vertanen K (2011) Asynchronous multimodal text entry using speech and gesture keyboards. In: Twelfth annual conference of the international speech communication association
14. Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions, and reversals. Sov Phys Dokl 10(8):707–710
15. Miki M, Kitaoka N, Miyajima C, Nishino T, Takeda K (2014) Improvement of multimodal gesture and speech recognition performance using time intervals between gestures and accompanying speech. EURASIP J Audio, Speech, Music Process 2014(1):1–7
16. Pitsikalis V, Katsamanis A, Theodorakis S, Maragos P (2017) Multimodal gesture recognition via multiple hypotheses rescoring. In: Escalera S, Guyon I, Athitsos V (eds) Gesture recognition. Springer, Cham, pp 467-496
17. Rebelo A, Fujinaga I, Paszkiewicz F, Marcal AR, Guedes C, Cardoso JS (2012) Optical music recognition: state-of-the-art and open issues. Int J Multimed Inf Retr 1(3):173–190
18. Román MA, Pertusa A, Calvo-Zaragoza J (2020) Data representations for audio-to-score monophonic music transcription. Exp Syst Appl 162:113769
19. Román M, Pertusa A, Calvo-Zaragoza J (2019) A holistic approach to polyphonic music transcription with neural networks. In: Proceedings of the 20th international society for music information retrieval conference, pp. 731–737. Delft, The Netherlands
20. Schedl M, Gómez E, Urbano J (2014) Music information retrieval: recent developments and applications. Found Trends Inf Retr 8:127–261. https://doi.org/10.1561/1500000042
21. Serra X, Magas M, Benetos E, Chudy M, Dixon S, Flexer A, Gómez E, Gouyon F, Herrera P, Jordà S, et al (2013) Roadmap for music information research. The MIReS Consortium. Creative Commons BY-NC-ND 3.0 license
22. Simonetta F, Ntalampiras S, Avanzini F (2019) Multimodal music information processing and retrieval: survey and future challenges. In: International workshop on multilayer music representation and processing, pp. 10–18
23. Singh A, Sangwan A, Hansen JHL (2012) Improved parcel sorting by combining automatic speech and character recognition. In: 2012 IEEE International conference on emerging signal processing applications, pp. 52–55. https://doi.org/10.1109/ESPA.2012.6152444
24. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147(1):195–197
25. Toselli AH, Vidal E, Casacuberta F (2011) Multimodal interactive pattern recognition and applications. Springer Science & Business Media, Berlin