INSIGHTS INTO TRANSFER LEARNING BETWEEN IMAGE AND AUDIO MUSIC TRANSCRIPTION

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

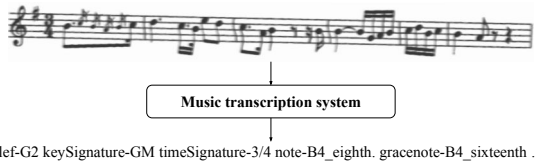
ABSTRACT

Optical Music Recognition (OMR) and Automatic Music Transcription (AMT) stand for the research fields that devise methods to transcribe music sources—documents or audio signals, respectively—into a structured digital format. Historically, they have followed different approaches to achieve the same goal. However, their recent definition in terms of sequence labeling tasks gathers them under a common formulation framework. Under this premise, one may wonder if there exist any synergies between the two fields that could be exploited to improve the individual recognition rates in their respective domains. In this work, we aim to further explore this question from a Transfer Learning (TL) point of view in the context of neural end-to-end recognition models. More precisely, we consider a music transcription system, trained on either image or audio data, and adapt its performance to the unseen domain during the training phase using different TL schemes. Results show that knowledge transfer slightly boosts model performance with sufficient available data, but it is not properly leveraged when the latter condition is not met. This opens up a new promising, yet challenging, research path towards building an effective bridge between two solutions of the same problem.

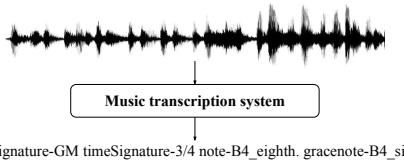
1. INTRODUCTION

The attainment of structured digital representations from music sources, typically known as *transcription*, stands as one of the major challenges in Music Information Retrieval (MIR) [1]. Within this community, there exist two main research lines that study how to computationally solve this problem when targeting either music documents—known as Optical Music Recognition (OMR) [2]—or acoustic music signals—namely, Automatic Music Transcription (AMT) [3]. Despite pursuing the same goal, these two fields have historically evolved in a disjoint manner since the differences in the nature of the data yielded specific task-oriented recognition frameworks, most commonly based on multi-stage processes [4].

However, some recent proposals in the MIR literature frame transcription problems under a sequence labeling formulation which approaches the task in a holistic or end-to-end manner [5]: the input data—either scores or acoustic pieces—are directly decoded into a sequence of music-notation symbols. Figure 1 graphically illustrates these music transcription approaches.



(a) Optical Music Recognition: music transcription using scores as inputs and a music-representation language as output.



(b) Automatic Music Transcription: music transcription using audio pieces as inputs and the same language as output.

Figure 1: End-to-end music transcription framework. OMR techniques deal with images and AMT techniques with audio signals; however, both tasks have to provide a result in a symbolic format that represents a score.

Currently, most end-to-end transcription approaches in the MIR literature resort to neural architectures [6, 7], which allows addressing OMR and AMT tasks with equivalent recognition models that only differ in the input data used for training the system. Furthermore, this common formulation enables exploring possible synergies that may exist between image and audio sources. Promoting such commonalities between both fields would open up a vast range of research avenues not previously explored in the related literature, such as: developing common language models, multimodal image and audio transcription, devising multi-task neural architectures capable of dealing with both tasks independently in a single model, or using pre-trained models with one modality and fine-tuning with the other. This latter case, which is commonly known as Transfer Learning, represents the focus of the work.

In a general sense, Transfer Learning (TL) addresses the case in which a model, initially trained with a particular source data distribution, is iteratively adapted to a differ-

ent, but related, target data domain [8]. Owing to a certain knowledge transference between the distributions, this process is expected to improve the recognition performance with respect to the case of exclusively training with the target domain [9]. While TL has been largely contemplated with successful results in other MIR tasks such as music classification [10] or vocal melody extraction [11], to our best knowledge no existing work has examined its use in transcription tasks due to the aforementioned limitations of legacy approaches.

This work studies the use of TL to exploit potential synergies between neural models of end-to-end image and audio transcription. More precisely, we examine the possible improvement achieved when training a recognition framework on a source domain—either image or audio—and transferring its knowledge to that not considered during the training stage. Such analysis may not only depict insights for improving the performance and robustness of the models, but it may also tackle the issue of data scarcity inherent to these fields by concurrently exploiting both domains [12, 13]. The results obtained show that, when a relatively large amount of training data is available, the proposed TL scheme achieves higher recognition rates than the case in which it is neglected. Moreover, this improvement is remarkably higher when adapting from AMT to OMR than the opposite case.

The rest of the paper is organized as follows: Section 2 develops the transfer learning framework considered; Section 3 describes the experimental setup; Section 4 presents and analyses the results; finally, Section 5 concludes the work and discusses possible ideas for future research.

2. METHODOLOGY

This section formally presents the neural end-to-end recognition and Transfer Learning frameworks contemplated in the work. To properly describe these design principles, we shall introduce some notation.

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 $\mathbf{z}_m = (z_{m1}, z_{m2}, \ldots, z_{mN})$ 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. In this regard, for the sake of clarity in the rest of the work, let $\mathcal{T}^i \subset \mathcal{X}^i \times \mathcal{Z}^i$ and $\mathcal{T}^a \subset \mathcal{X}^a \times \mathcal{Z}^a$ respectively represent the labeled sets of image scores and audio signals for training the transcription models.

2.1 Neural End-To-End Music Transcription

Due to its reported competitive performance in the related literature, we have considered a Convolutional Recurrent Neural Network (CRNN) scheme [14] with the Connectionist Temporal Classification (CTC) training algorithm [5] to approximate function g. This architecture comprises 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 method as it allows training the CRNN scheme using unsegmented sequential data. In a practical sense, this mechanism only requires the different input signals to the scheme and their associated sequences of characters drawn from vocabulary Σ 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., $\Sigma' = \Sigma \cup \{blank\}$ for enabling the detection of consecutive repeated elements.

Since CTC assumes that the architecture contains a fully-connected network of $|\Sigma'|$ neurons 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 that merges consecutive repeated symbols and removes the *blank* label.

2.2 Transfer Learning Framework

As commented, Transfer Learning (TL) has been largely considered in the context of neural learning-based systems due to its reported benefits in terms of performance improvement. In a practical sense, such approaches typically initialize a recognition model using a source domain of data which is then fine-tuned with the actual target corpus, attending to a particular adaptation policy.

In this work, we explore the use of TL in neural end-to-end music transcription involving image scores and audio recordings given that, in both domains, we aim at retrieving a symbolic representation of the data at issue. More precisely, we assess the knowledge transference by posing these two scenarios: (i) pre-training a transcription model on one domain (either image or audio) and fine-tuning with the other one; and (ii) evaluating the amount of data necessary in the target domain for an effective transfer which outperforms the base case of neglecting TL. We shall now introduce some notation for then formally defining these scenarios.

Let CRNN^i and CRNN^a respectively denote two CRNN transcription models trained with a source domain of image, $\mathcal{T}^i$, and audio, $\mathcal{T}^a$, data, respectively. These initial models are adapted to a novel domain by being iteratively re-trained with the target data, hence obtaining CRNN$^{i \rightarrow a}$ for the image-to-audio scenario or CRNN$^{a \rightarrow i}$ for the audio-to-image one. Additionally, given that certain layers may be prevented from being updated during the domain transference stage—process typically known as *freezing*—we include subscript (L) to denote the range of L layers which are not affected by the re-training process, i.e., CRNN$^{i \rightarrow a}_{(L)}$ and CRNN$^{a \rightarrow i}_{(L)}$.

The first proposed scenario aims to uncover possible synergies between image and audio domains with a direct

knowledge transfer. In this regard, we assess the performance of the recognition models both when considering and disregarding pre-training. Two situations are posed attending to the target data: on the one hand, we compare model CRNN^{i} against CRNN$^{a\rightarrow i}_{(L)}$ for the OMR case; on the other hand, we confront CRNN^{a} against CRNN$^{i\rightarrow a}_{(L)}$ for the AMT situation. Moreover, for each of these cases we also assess the influence of freezing different layers when performing the TL process. Figure 2 graphically illustrates this scenario, particularly the case of training with image data, CRNN^{i}, and then performing TL to the audio domain while preventing the fifth layer of the model from being updated, i.e., CRNN$^{i\rightarrow a}_{(5:5)}$.

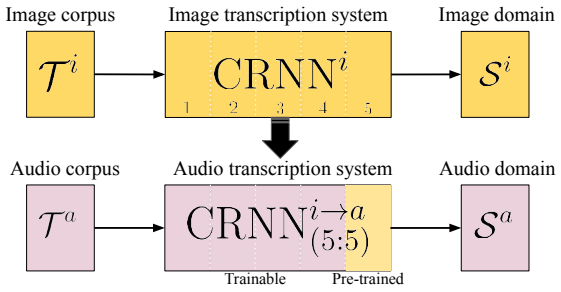


Figure 2: Graphical example of the Transfer Learning proposal: the CRNN^{i} image transcription model is adapted to the audio domain while preventing the parameters of the fifth layer from being updated: CRNN$^{i\rightarrow a}_{(5:5)}$.

It must be noted that, since the considered transcription models are based on deep neural networks, they generally require large amounts of labeled data to be trained [15]. This strong dependence on the size of the training set usually becomes an issue, especially when transcribing data domains with limited availability of transcriptions. Hence, as previously stated, TL poses itself as a promising alternative to solve the small-data problem by training the model on a certain domain for then transferring and adapting the knowledge acquired to the target one.

In this context, the second posed scenario simulates and studies this particular challenge attending to the target data domain: on one side, for the OMR case, CRNN^{i} is compared to CRNN$^{a\rightarrow i}_{(L)}$ when contemplating different subsets of the image corpus ($\mathcal{T}^{i-} \subset \mathcal{T}^{i}$); on the other side, when considering the AMT case, we compare CRNN^{a} against CRNN$^{i\rightarrow a}_{(L)}$ when considering subsets of the entire audio corpus ($\mathcal{T}^{a-} \subset \mathcal{T}^{a}$).

3. EXPERIMENTAL SETUP

This section presents the corpus considered, the definition of the different layers of the neural model, and the evaluation protocol used.

3.1 Corpus

We have considered the Camera-based Printed Images of Music Staves (Camera-PrIMuS) database [6]. This corpus contains 87,678 real music staves of monophonic incipits extracted from the *Répertoire International des Sources Musicales* (RISM). [1] 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.

In this regard, each transcription architecture considers a particular type of data: on the one hand, the OMR model takes as input the artificially distorted staff image of the incipit; on the other hand, for the AMT case, each MIDI file is synthesized with the FluidSynth software [2] and a piano timbre considering a sampling rate of 22,050 Hz, for then obtaining a time-frequency representation based on the Constant-Q Transform with a hop length of 512 samples, 120 bins, and 24 bins per octave, which is eventually embedded as an image that serves as the input. In both tasks, the height of the input figure considered is scaled to 256 pixels, maintaining the aspect ratio (thus, each sample might differ in width) and converted to grayscale, with no further preprocessing.

An initial data curation process was applied to the corpus to discard samples which may remarkably hinder the transcription, resulting in 67,000 incipits. [3] Since this reduced set still contains a considerably large amount of elements, we randomly selected a third of this curated set for our experiments, approximately, resulting in 22,285 incipits. Eventually, we derive three non-overlapping partitions—train, validation, and test—which correspond to the 60%, 20%, and 20% of the latter amount of data, respectively. Note that, since we are considering the same corpus for both image and audio data, both recognition tasks depict the same label space of $\Sigma^{i} = \Sigma^{a} =$ 1,166 tokens.

3.2 Neural Network Configuration

While the presented sequence labeling paradigm allows considering a common formulation for both OMR and AMT tasks, in practice there is no universal neural architecture capable of achieving state-of-the-art performances in both cases. Generally, these configurations depend on a wide range of parameters which comprises the particular corpus considered, the amount of accessible data, or the available computational resources, among others.

Nevertheless, this work considers a common architecture for both transcription modalities for simplicity. In this regard, since neural end-to-end OMR models generally outperform AMT approaches [4], we consider a state-of-the-art architecture originally devised for the latter field for addressing both domains. Note that this is not a strong assumption since the performance decrease in the OMR domain with respect to the best achievable result is not remarkable and does not bias the conclusions obtained in the work. Hence, the actual composition of each layer is depicted in Table 1.

[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/

[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: Layer-wise description of the CRNN model considered. 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 a negative slope of value α, MaxPool$(w \times h, a \times b)$ stands for the max-pooling operator of dimensions $w \times h$ pixels with $a \times b$ striding factor, BLSTM(n, d) denotes a bidirectional Long Short-Term Memory unit with n neurons and d dropout value parameters, Dense(n) means a fully-connected layer of n neurons, and Softmax$(\cdot)$ represents the softmax activation. Σ' denotes the considered alphabet, including the CTC-blank symbol.

Convolutional Block		**Recurrent Block**		**Classification Block**
Layer 1	**Layer 2**	**Layer 3**	**Layer 4**	**Layer 5**
Conv$(8, 2 \times 10)$	Conv$(8, 5 \times 8)$			
BatchNorm	BatchNorm	BLSTM(256)	BLSTM(256)	Dense$(\lvert\Sigma'\rvert)$
LeakyReLU(0.20)	LeakyReLU(0.20)	Dropout(0.50)	Dropout(0.50)	Softmax$(\cdot)$
MaxPool(2×2)	MaxPool(1×2)			

All models in the work are trained using the backpropagation method provided by CTC for 100 epochs using the ADAM optimizer [16] with a fixed learning rate of 0.001 and a batch size of 4 elements.

3.3 Evaluation Protocol

We consider the Symbol Error Rate (SER) for assessing the performance of the presented recognition schemes as in previous works addressing end-to-end transcription tasks [6, 7]. This figure of merit is computed as the average number of elementary editing operations (insertions, deletions, or substitutions) necessary to match the sequence predicted by the model with that in the ground truth, normalized by the length of the latter. Mathematically, this is expressed as:

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

where $\mathcal{S} \subset \mathcal{X} \times \mathcal{Z}$ is a set of test data—from either the image or the audio domains—, ED : $\mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{N}_0$ represents the string Edit Distance [17], and $\hat{\mathbf{z}}_m$ and $\mathbf{z}_m$ denote the estimated and target sequences, respectively.

4. RESULTS

This section presents the results obtained for the different TL scenarios posed with the experimental scheme considered. For the sake of clarity, we analyze each TL case in a different section: a first one, denoted as *Scenario A*, that assesses the performance of the recognition models when TL is both considered and ignored, and a second one, namely *Scenario B*, that studies the amount of data required in the target domain for an efficient transfer process that outperforms the base case of ignoring TL. In all cases, the figures provided represent those obtained with the test partition when the validation data achieved its best performance. [4]

4.1 Scenario A: Fine Tuning

The first scenario posed studies the impact of TL on the recognition performance of the transcription models. For that, we consider base OMR and AMT recognition models—CRNN^i and CRNN^a, respectively—and perform a TL process for adapting them to their respective opposite domain, i.e., CRNN$^{i \rightarrow a}$ and CRNN$^{a \rightarrow i}$. Since this adaptation process may prevent different layers from being updated, we also analyze the influence of this parameter in the success of the task.

Table 2 reports transcription results in terms of the Symbol Error Rate (SER) when both base CRNN^i and CRNN^a models as well as CRNN$^{i \rightarrow a}$ and CRNN$^{a \rightarrow i}$ architectures are used for transcribing an evaluation set of image and audio data, respectively denoted as $\mathcal{S}^i$ and $\mathcal{S}^a$. Note that, as aforementioned, subscript (L) in the TL-based schemes indicates the range of layers in the recognition model (cf. Table 1) that are unaffected by the re-training process, being N/A the case in which all parameters are updated.

Inspecting the reported results, a first point to remark is the improvement achieved when training a recognition framework on a source domain—either image or audio—and transferring its knowledge to that not considered during the training stage when following the best re-training policy. On the image domain, CRNN$^{a \rightarrow i}_{(N/A)}$ reduces the error rate of CRNN^i over approximately 44% (from 9.58% to 5.31%); whereas, on the audio domain, CRNN$^{i \rightarrow a}_{(5:5)}$ obtains around a 3% of relative error improvement with respect to its baseline CRNN^a (from 29.24% to 28.32%). Such results support the initial hypothesis that TL may serve itself as a feasible strategy for obtaining more reliable and robust models.

The chosen re-training strategy plays an important role in the model performance. In this sense, when initially trained for addressing OMR tasks and adapted to AMT scenarios (CRNN$^{i \rightarrow a}$), the classification block may remain unaltered (CRNN$^{i \rightarrow a}_{(5:5)}$). However, learning first to solve the music transcription task with audio data (AMT) accounts for no model re-usability when transferring the knowledge to the image domain (OMR). Nonetheless, results indicate that knowledge transfer is better leveraged when adapting from AMT to OMR—CRNN$^{a \rightarrow i}$—than from OMR to AMT—CRNN$^{i \rightarrow a}$—since, on average, the former achieves better performance rates than the latter.

The last point to remark is that, as expected, when models are evaluated on a domain different than the one they were

[4] The code developed in the work is publicly available for reproducible research at: https://github.com/mariaalfaroc/smc-2022.git

Table 2: Results obtained in terms of the Symbol Error Rate (SER) for each transcription model considered when confronted to the image and audio domains, denoted as $\mathcal{S}^i$ and $\mathcal{S}^a$, respectively. Each column in the TL models—CRNN$^{i\to a}_{(L)}$ and CRNN$^{a\to i}_{(L)}$—denotes the range of L = [1:5] layers in the recognition model (cf. Table 1) which are not affected by the re-training process, being N/A the case in which all parameters are updated. Non-adaptive schemes—CRNN^i and CRNN^a—are provided for reference purposes. Best results for each evaluation domain are highlighted in bold type.

	CRNN^i	CRNN^a	Transfer learning: CRNN$^{i\to a}_{(L)}$					Transfer learning: CRNN$^{a\to i}_{(L)}$				
			N/A	5 : 5	4 : 5	3 : 5	2 : 5	N/A	5 : 5	4 : 5	3 : 5	2 : 5
$\mathcal{S}^i$	9.6	331.2	392.4	198.2	104.0	87.2	97.3	**5.3**	17.6	41.1	96.0	96.0
$\mathcal{S}^a$	98.9	29.2	28.6	**28.3**	40.3	86.7	93.9	98.4	97.6	95.6	96.0	96.0

trained on, their performance suffers a drastic deterioration. This downgrade is even more accused for audio models evaluated on image data. For TL-based approaches, this issue is referred to as catastrophic forgetting [18], as it implies the loss of previously learned information. We believe the multi-task learning paradigm could mitigate this problem by devising single architectures capable of solving several tasks, OMR and AMT, in this case.

4.2 Scenario B: Influence of the Target Set Size

As stated in Section 2.2, TL stands as a suitable solution for tackling the data scarcity in learning-based systems. This framework guarantees the convergence of the model on a sufficiently large data domain whose performance may be then adapted to a target one. We now explore this premise in the context of neural end-to-end music transcription systems.

To simulate this scenario, we consider different subset sizes of the target corpus. These divisions are used for adapting the base OMR and AMT recognition models trained on the entire source corpus—namely, CRNN^i and CRNN^a, respectively—to the target domain, hence respectively obtaining CRNN$^{i\to a}(5:5)$ and CRNN$^{a\to i}(N/A)$. Note that, for this evaluation scheme, only the best transfer configurations, obtained in Scenario A, are considered.

Table 3 reports the transcription results obtained for both the base and TL-oriented recognition models when evaluating on the same domain as the target space considered during the training stage. Note that train subset sizes reported range within $[50, 1000]$ since figures obtained above these values did not report any difference in performance.

Focusing on the image recognition task ($\mathcal{S}^i$), the use of TL denotes a better initialization of the model parameters as the CRNN$^{a\to i}_{(N/A)}$ model overcomes the base CRNN^i case when considering train sizes up to 100 samples. However, for a larger amount of training samples, the results report that the TL-based scheme severely degrades with respect to the base case. This impedes the use of previously acquired information in situations where it is most needed, such as those with a limited quantity of data. Nonetheless, since Table 2 shows that the inherent bias when training with audio data ($\mathcal{T}^a$) is overcome by fine-tuning with the entire image corpus ($\mathcal{T}^i$), we may assume the existence of a certain threshold in terms of train data elements in which the achievable error rate is lower than that when TL is disregarded (CRNN^i case).

In the case of audio transcription, TL leads to an improvement, of over 30%, only if we have a relatively small train set of 50 samples. When the training size is increased, the differences between considering or disregarding TL, are marginal, as observed in Tables 2 and 3. This suggests new TL methods should be devised to enhance the knowledge transfer when adapting from OMR to AMT.

5. CONCLUSIONS

The transcription of music sources into a structured digital format is one of the main challenges of the Music Information Retrieval (MIR) field. To tackle this problem, two research lines are considered: Optical Music Recognition (OMR), when considering visual input data such as scores, and Automatic Music Transcription (AMT), when considering acoustic input data such as audio signals. While these fields have historically evolved separately, their recent definition within a sequence labeling formulation results in a common representation for their expected outputs. This enables OMR and AMT tasks to be addressed with equivalent recognition models that only differ in their input modality.

Under this context, the following question naturally arises: is there any relationship between the learning features that each data-specific model acquires for solving the music transcription task? This paper provides insights into the posed topic by considering both modalities under a Transfer Learning (TL) scenario, which means adapting an already trained model for a new task. Specifically, we start from image transcription models and adapt them to audio transcription, and vice versa. To uncover the synergies and commonalities that can be established between the image and audio modalities, we analyze: (i) whether such knowledge transfer influences the overall recognition performance of the model; and (ii) how scenarios depicting data scarcity leverage this kind of relationship.

The obtained results reveal two relevant conclusions: on one end, TL boosts the model performance, being this improvement more pronounced when adapting from AMT to OMR; on the other end, the use of TL is not enough to overcome the scarcity of data as the aforementioned enhancement of the performance is only observable with sufficient target training data.

In light of these results, this work opens up new promis-

Table 3: Transcription results obtained in terms of the Symbol Error Rate (SER) for different subsets $\mathcal{T}^-$ of the train partition when considering domain-equivalent target and evaluation sets. The reported TL-based models—$\text{CRNN}^{a\to i}_{(N/A)}$ and $\text{CRNN}^{i\to a}_{(5:5)}$—for the image and audio target domains represent the best performing ones from the previous scenario whereas the non-adaptive schemes—CRNN^{i} and CRNN^{a}—are provided for reference purposes. Best performing figures for each evaluation domain and train corpus size are highlighted in bold type.

	Size of train corpus $\mathcal{T}^-$					
	50	75	100	250	500	1000
Evaluating on $\mathcal{S}^i$						
CRNN^{i}	95.3	95.2	94.9	**63.6**	**37.0**	**23.7**
$\text{CRNN}^{a\to i}_{(N/A)}$	**92.1**	**89.7**	**90.8**	79.5	77.6	49.9
Evaluating on $\mathcal{S}^a$						
CRNN^{a}	93.5	**56.7**	57.0	**45.9**	**40.4**	**37.5**
$\text{CRNN}^{i\to a}_{(5:5)}$	**64.0**	57.7	**53.6**	47.3	41.9	38.4

ing, yet challenging, avenues for research. For instance, the TL approach may be further explored by considering domain adaptation techniques. Besides, few-shot learning could be considered for tackling the data scarcity issue in this transcription paradigm. Finally, multi-task techniques may be explored to strengthen the synergies between image and audio music transcription by devising neural architectures capable of tackling both data domains in a single model.

Acknowledgments

This paper is part of the project I+D+i PID2020-118447RA-I00 (MultiScore), funded by MCIN/AEI/10.13039/501100011033. The first and second authors are supported by grants FPU19/04957 from the Spanish Ministerio de Universidades and APOSTD/2020/256 from "Programa I+D+i de la Generalitat Valenciana", respectively.

6. 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] C. de la Fuente, J. J. Valero-Mas, F. J. Castellanos, and J. Calvo-Zaragoza, "Multimodal image and audio music transcription," *International Journal of Multimedia Information Retrieval*, pp. 1–8, 2021.

[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] 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.

[7] M. A. Román, A. Pertusa, and J. Calvo-Zaragoza, "A holistic approach to polyphonic music transcription with neural networks," in *Proceedings of the 20th International Society for Music Information Retrieval Conference*, Delft, The Netherlands, Nov. 2019, pp. 731–737.

[8] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 3320–3328.

[9] Q. Yang, Y. Zhang, W. Dai, and S. J. Pan, *Transfer learning*. Cambridge University Press, 2020.

[10] L. Wang, H. Zhu, X. Zhang, S. Li, and W. Li, "Transfer learning for music classification and regression tasks using artist tags," in *Proceedings of the 7th Conference on Sound and Music Technology (CSMT)*. Springer, 2020, pp. 81–89.

[11] W.-T. Lu and L. Su, "Vocal melody extraction with semantic segmentation and audio-symbolic domain transfer learning," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*. Paris, France: ISMIR, Sep. 2018, pp. 521–528.

[12] M. Alfaro-Contreras, D. Rizo, J. M. Iñesta, and J. Calvo-Zaragoza, "OMR-assisted transcription: a case study with early prints," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference*. Online: ISMIR, Nov. 2021, pp. 35–41.

[13] L. Liu and E. Benetos, "From audio to music notation," in *Handbook of Artificial Intelligence for Music*. Springer, 2021, pp. 693–714.

[14] 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.

[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[16] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[17] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet physics doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[18] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of learning and motivation*. Elsevier, 1989, vol. 24, pp. 109–165.