STRUCTURE-AWARE AUDIO-TO-SCORE ALIGNMENT USING PROGRESSIVELY DILATED CONVOLUTIONAL NEURAL NETWORKS

Ruchit Agrawal [†]   *Daniel Wolff* [⋆]   *Simon Dixon* [†]

[†] Centre for Digital Music, Queen Mary University of London, UK
[⋆] Institute for Research and Coordination in Acoustics/Music, Paris, France

ABSTRACT

The identification of structural differences between a music performance and the score is a challenging yet integral step of audio-to-score alignment, an important subtask of music information retrieval. We present a novel method to detect such differences between the score and performance for a given piece of music using progressively dilated convolutional neural networks. Our method incorporates varying dilation rates at different layers to capture both short-term and long-term context, and can be employed successfully in the presence of limited annotated data. We conduct experiments on audio recordings of real performances that differ structurally from the score, and our results demonstrate that our models outperform standard methods for structure-aware audio-to-score alignment.

Index Terms— Music Information Retrieval, Audio-to-Score Alignment, Music Alignment, Music Structure Analysis

1. INTRODUCTION

The analysis of music performance is a challenging area of Music Information Retrieval (MIR), owing to multiple factors such as suboptimal recording conditions, structural differences from the score, and subjective interpretations. Deviations from the structure and/or tempo prescribed by the score are common in several genres of music, particularly classical music and jazz [1]. Identification of such structural differences from the score is a challenging yet important step of performance analysis, especially for the audio-to-score alignment task. The alignment task aims at generating an accurate mapping between a performance audio and the score for a given piece of music. The particular path through the score that the musician is going to take is difficult to predict, and although there have been several methods proposed in the recent past to handle these impromptu changes; it still remains a problem that is not fully solved [2].

The majority of approaches for audio-to-score alignment are based on Dynamic Time Warping (DTW) [3] or Hidden Markov Models (HMM) [4]. These methods typically assume that the musician follows the score from the beginning to the end without inducing any structural modifications, which is often not the case in real world scenarios. The alignments computed using DTW-based methods are constrained to progress monotonically through the piece, and are thereby unable to model structural deviations such as repeats and jumps. Previous methods for handling structural changes during alignment are either reliant on robust Optical Music Recognition (OMR) to detect repeat and jump directives [5]; or on frameworks fundamentally different from DTW, such as the Needleman Wunsch Time Warping method [6]. The former method,

called *JumpDTW*, requires manually annotated frame positions for block boundaries, initially provided by an OMR system, and is unable to model deviations that are not foreseeable from the score. The latter is unable to align repeated segments, since it does not introduce backward jumps, and is based on a waiting mechanism upon mismatch of the two streams. Since structural changes like repeats and jumps can be arbitrarily added during a performance, especially during rehearsals; these are challenging to model using rule-based approaches, and machine learning methods offer promise at effectively addressing these challenges.

This paper is aimed at handling structural changes between the performance and the score for the offline audio-to-score alignment task. We propose a custom Convolutional Neural Network (CNN) based architecture for modeling these differences, coupled with a flexible DTW framework to generate the fine alignments. Our method does not require a large corpus of hand-annotated data and can also be trained exclusively using synthetic data if hand-annotated data is unavailable. Our architecture employs progressively dilated convolutions in addition to standard convolutions, with varying amounts of dilation applied at different layers of the network. The primary motivation behind our architecture is that it allows us to effectively capture both short-term and long-term context, using much fewer parameters than sequential models such as recurrent neural networks, and without facing the vanishing gradient problem. We conduct experiments on piano music and compare our method with three major alignment approaches; namely *JumpDTW* [5], *NWTW* [6], and *MATCH* [3]. We demonstrate results on two different test sets containing real performances. Our method yields higher performance than previous methods proposed for handling structural deviations, without requiring manually annotated data; and noticeably outperforms these methods given a limited amount of annotated data. To the authors' knowledge, this is the first method to employ dilated convolutional neural networks for structure-aware audio-to-score alignment.

2. RELATED WORK

Early work on structure-aware MIR focuses on structural segmentation of musical audio by constrained clustering [7] and music repetition detection using histogram matching [8]. Arzt and Widmer [9] propose a multilevel matching and tracking algorithm to deal with issues in score following due to deviations from the score in live performance. A challenge faced by this approach appears when complex piano music is played with a lot of expressive freedom in terms of tempo changes. Hence, they propose methods to estimate the current tempo of a performance, which could then be used to improve online alignment [10]. This is similar to the work proposed by Müller et al. [11], wherein they develop a method for automatic extraction of tempo curves from music recordings by comparing performances with neutral reference representations.
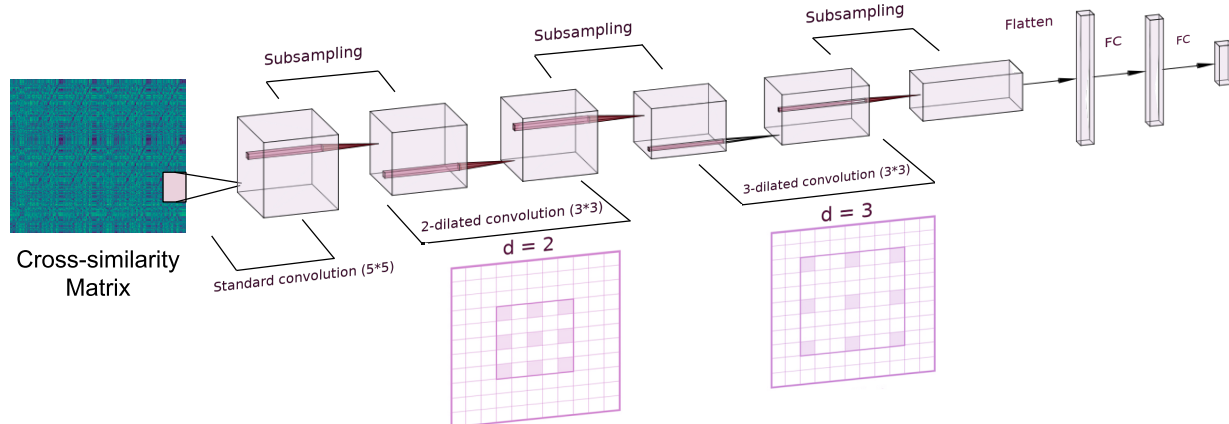
Fig. 1: Schematic diagram illustrating the general architecture of our models.
d: Dilation rate, FC: Fully connected layer

Work specifically on structure-aware audio-to-score alignment includes *JumpDTW* [5] and Needleman-Wunsch Time Warping (*NWTW*) [6] among others. Fremerey et al. [5] focus on tackling structural differences induced specifically by repeats and jumps, using a novel DTW variation called *JumpDTW*. This method identifies the "block sequence" taken by a performer along the score, however it requires manually annotated block boundaries to yield robust performance, which are generally not readily available at test time for real world applications. Additionally, it cannot align deviations that are not foreseeable from the score. *NWTW* [6], on the other hand, is a pure dynamic programming method to align music recordings that contain structural differences. This method is an extension of the classic Needleman-Wunsch sequence alignment algorithm [12], with added capabilities to deal with the time warping aspects of aligning music performances. A caveat of this method is that it does not successfully align repeated segments owing to its waiting mechanism, which skips unmatchable parts of either sequence, and makes a clean jump when the two streams match again.

Apart from *JumpDTW* and *NWTW*, which focus on offline alignment, work on online score following [13] has demonstrated the effectiveness of HMMs for modeling variations from the score. While this method focuses on real-time score following for monophonic music, our work deals with offline audio-to-score alignment for polyphonic music performance. Another work related to ours is that proposed by Jiang et al. [14], which focuses on offline score alignment for the practice scenario. Similar to Nakamura et al. [13], their approach is also based on HMMs; but they propose using pitch trees and beam search to model skips. However, their method struggles with pieces containing both backward and forward jumps, which is an important challenge we tackle using our progressively dilated convolutional models. Recent work on audio-to-score alignment has demonstrated the efficacy of multimodal embeddings [15], reinforcement learning [16], [17] and learnt frame similarities [18], albeit these are not structure-aware methods. Very recently, Shan et al. propose Hierarchical-DTW [19] to automatically generate piano score following videos given an audio and a raw image of sheet music. Their method is reliant on an automatic music transcription system [20] and a pre-trained model to extract bootleg score representations [21]. It struggles when this representation is not accurate, and also on short pieces containing jumps. While they work with raw images of sheet music and generate score following videos; our method works with symbolic scores, is not reliant on other pre-trained models, and performs well on both short and long pieces.

3. PROPOSED METHOD

We present a novel method to detect structural differences for audio-to-score alignment using a custom convolutional architecture. The closest to our method is the work proposed by Ullrich et al. [22], an adaptation of the onset detection method proposed by Schlüter et al. [23]. While Ullrich et al. [22] train a CNN as a binary classifier on spectrogram excerpts for popular music (using hand annotated data) to detect boundaries in a song, we focus on structure-aware alignment of a performance to the score, with the cross-similarity matrices as inputs. Additionally, we incorporate dilation in our CNN models to incorporate multi-scale context.

Our architecture combines standard convolution and dilated convolution [24], with varying amounts of dilation applied at different layers of the network. We approach the sturctural deviation identification problem as a multi-label prediction task, and train our models to detect synchronous subpaths between the score and the performance. These subpaths are detected by means of inflection points, which encode the positions of structural differences between the two streams. Figure 1 illustrates the general architecture of our models. Our networks operate on the cross-similarity matrix between the score and performance and predict the (x, y) co-ordinates of the inflection points as the output. We employ dilated convolution at the second and the third layer, and standard convolution at the first layer. The dilated convolution operation [24] of a discrete function F with a discrete filter f on an element $\mathbf{p}$ is defined as follows:

$$(F *_d f)(\mathbf{p}) = \sum_{\mathbf{t}} F(\mathbf{p} - d\mathbf{t}) f(\mathbf{t}) \qquad (1)$$

where d is the factor by which the kernel is inflated, referred to as the dilation rate. Inflating the kernel using dilation allows us to incorporate larger context without increasing the number of parameters, which is essential for modeling structure. The receptive field is exponentially increased, and for a dilation rate d, a kernel of size m effectively works as a kernel of size m' as follows:

$$m' = m + (d - 1) * (m - 1) \qquad (2)$$

This facilitates the incorporation of context better than standard convolutions, which can only offer linear growth of the effective receptive field as we move deeper into the network. We conduct experiments using varying dilation rates at different layers of the network to determine the optimal amount of dilation at each layer. The motivation behind using varying amounts of dilation is to incorporate both short-term and long-term context to model structure, which has

Model	On Mazurka dataset				With structural differences (Tido)				Without structural differences (Tido)			
	<25ms	<50ms	<100ms	<200ms	<25ms	<50ms	<100ms	<200ms	<25ms	<50ms	<100ms	<200ms
MATCH [3]	64.8	72.1	77.6	83.7	61.5	70.4	74.6	80.7	70.2	78.4	84.7	90.3
JumpDTW [5]	65.8	75.2	79.8	85.7	69.1	77.2	82.0	88.4	68.7	77.5	82.1	88.9
NWTW [6]	67.6	75.5	80.1	86.2	68.6	75.8	80.7	87.5	68.4	77.1	82.8	89.4
CNN_{1+1}	68.2	75.7	80.5	87.1	70.4	78.3	83.4	90.1	69.3	78.0	84.1	89.3
$DCNN_{2+2}$	**69.9**	76.4	81.6	88.9	72.7	80.1	84.5	91.4	**71.4**	79.5	85.3	90.5
$DCNN_{2+3}$	69.7	**77.2**	**82.4**	**89.8**	**73.9**	**81.3**	**85.6**	**92.8**	71.0	**80.3**	**85.8**	**91.8**
$DCNN_{3+3}$	69.2	76.1	81.2	88.7	72.3	79.5	84.2	90.4	70.6	78.8	84.9	91.2
$DCNNsyn_{2+3}$	68.1	75.9	80.7	87.5	70.5	78.6	83.8	90.5	69.2	78.3	84.6	89.8

Table 1: Alignment accuracy in % on the Mazurka and Tido test sets
$DCNN_{m+n}$: Dilated CNN model with dilation rates of m and n at the second and third layer respectively

proven to be useful in computer vision as well as natural language processing tasks [25, 26]. We observe that progressively increasing dilation as we move deeper into the network produces the best results for detecting the structural differences. We compare the results of our networks with previous methods proposed for handling structural changes during alignment ([5], [6]) as well as a baseline CNN model trained without any dilation.

In order to generate the fine alignments, the inflection points predicted by our dilated convolutional models are employed as potential jump positions to assist a DTW-based alignment algorithm. We implement such an extended DTW framework, inspired by *JumpDTW* [5], to allow for jumps between the synchronous subpaths. We assume $X = (x_1, x_2, ..., x_p)$ to be the feature sequence corresponding to the performance and $Y = (y_1, y_2, ..., y_q)$ to be the feature sequence corresponding to the score. Furthermore, let (a_i, b_i) denote the (x, y) co-ordinates of the i_{th} inflection point, and N denote the total number of inflection points. The odd numbered inflection points correspond to the end of the synchronous subpaths and the even numbered points correspond to the beginning of the subpaths. We modify the classical DTW framework to extend the set of possible predecessor cells for the cell (a_i, b_i) for all $i \in \{2, 4, 6, .., N\}$, as follows:

$$D(m,n) = e(m,n) + min \begin{cases} D(m, n-1) \\ D(m-1, n) \\ D(m-1, n-1) \\ D(a_{i-1}, b_{i-1}) \ \forall (m,n) = (a_i, b_i), \\ \qquad i \in \{2, 4, ..., N\} \end{cases}$$
(3)

where $e(m, n)$ is the Euclidean distance between points x_m and y_n, and $D(m, n)$ is the total cost to be minimized for the path until the cell (m, n). The path which yields the minimum value for $D(p, q)$ is taken to be the optimal alignment path between the performance and score sequences.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

We model the task of detecting the synchronous subpaths as a multi-label prediction task using progressively dilated CNNs, with each output label encoding a deviation in the performance from the score. A key challenge in modelling structural changes for alignment is the lack of hand annotated data, marked for repeats and jumps accurately at the frame level. This is also one of the caveats of *JumpDTW*, which is reliant on the accuracy of the OMR system to detect jump and repeat directives in the absense of manually annotated boundaries. To overcome the lack of annotated training data, we generated synthetic samples containing jumps and repeats using the audio

from the MSMD [15] dataset. Our training dataset contains 2625 pairs of audio recordings, corresponding to the MIDI score and performance respectively. 2475 of these are obtained from the MSMD dataset, with each piece utilized 4 times for varying number of repetitions (generated synthetically), and once without any repetition. In addition to synthetic data, we employed a small amount of hand annotated data from a private dataset procured from Tido UK Ltd., referred to as the Tido dataset further in the paper. The training set taken from the Tido dataset comprises audio pairs for 150 pieces, 80 of which contain structural differences. In addition to our models trained with different dilation rates on the entire training set, we also demonstrate the results obtained by our progressively dilated CNN model trained exclusively on synthetic data ($DCNNsyn_{2+3}$).

We compute the cross-similarity matrix for each performance-score pair using the Euclidean distance between the chromagrams corresponding to the score and performance respectively. We employ librosa [27] to compute the chromagrams as well as the cross-similarity. Our models consist of three convolutional and subsampling layers, with standard convolutions at the first layer and dilated convolutions with varying dilation rates at the second and third layer respectively. The output of the third convolutional and subsampling layer is sent through a flatten layer, following which it is passed through two fully connected layers of size 4096 and 1024 respectively to predict the (x, y) co-ordinates of the inflection points. The output of the final layer is a one-dimensional tensor of size 64, signifying that the model can predict up to 32 inflection points, with their (x, y) co-ordinates in chronological order. The output is compared with the ground truth using the L2 regression loss, since we want to capture the distance of the predicted inflection points from the ground truth inflection points in time. The outputs of each layer are passed through rectified linear units to add non-linearity, followed by batch normalization before being passed as inputs to the next layer. We employ a dropout of 0.5 for the fully connected layers to avoid overfitting. Our batch size is 64 and the models are trained for 40 epochs, with early stopping. Our dilated CNN models are denoted as $DCNN_{m+n}$, where m and n correspond to the dilation rates at the second and third layer respectively.

We test the performance of our models on two different datasets, both containing recordings of real performances. We demonstrate the results of our models on the publicly available Mazurka dataset [28]. In order to analyze specific improvements for structurally different pieces, we also demonstrate results on subsets of the Tido dataset *with* and *without* structural differences. Both the subsets contain 75 pieces each. We compare the results obtained by our models with *JumpDTW* [5], *NWTW* [6], *MATCH* [3] and a vanilla CNN model without dilation CNN_{1+1}. The number of parameters

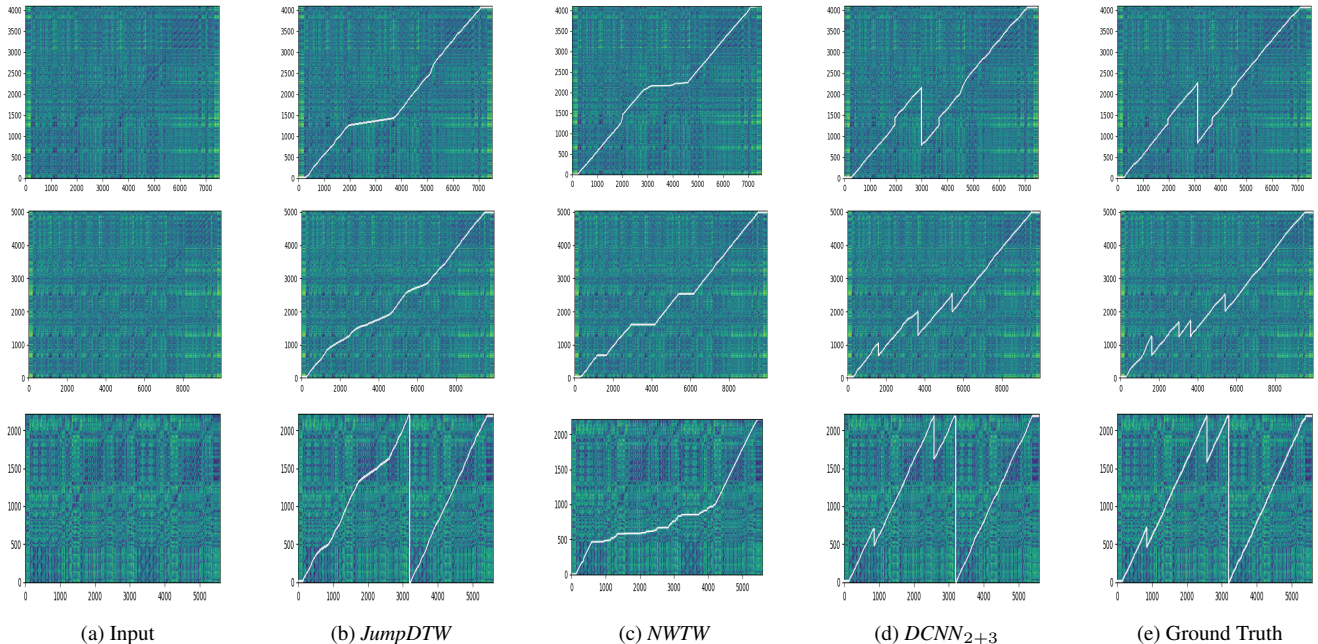|				
(a) Input	(b) *JumpDTW*	(c) *NWTW*	(d) $DCNN_{2+3}$	(e) Ground Truth

Fig. 2: Comparison of our alignment path with standard methods.
Input: Cross-similarity matrix between score and performance, X-axis: Frame index (performance), Y-axis: Frame index (score)

of our $DCNN_{m+n}$ networks is comparable with that of the baseline CNN_{1+1} network. For comparison with *JumpDTW*, we employ the SharpEye OMR engine to extract frame predictions for block boundaries from the sheet images [5]. These are then passed on to an implementation of *JumpDTW* to generate the alignment path. Similarly, we compare our models with an implementation of the *NWTW* method, and estimate the optimal gap penalty parameter γ [6] on our data.

4.2. Results and Discussion

The results obtained by our models are given in Table 1. We report alignment accuracy in %, where each value denotes the percentage of beats aligned correctly within the corresponding time durations of 25, 50, 100 and 200 ms respectively. Our models show an increase of 2-5% in alignment accuracy over *JumpDTW* and *NWTW* on the test subset containing structural differences and an increase of 1-3% on the test subset not containing structural differences. Compared with *MATCH*, our models show an increase of 9-10% on the subset with structural differences, and an increase of 1-2% on the subset without structural differences. Overall accuracy on the Mazurka dataset suggests that our models perform better than *MATCH* by 4-6% as well as the *JumpDTW* and *NWTW* frameworks by 1-4% (Table 1, columns 1-4). Our model trained exclusively on synthetic data (*DCNNsyn*$_{2+3}$) yields better alignment accuracy than *JumpDTW*, which requires manually labelled block boundaries to handle repeats and jumps [5]. This emphasizes the applicability of our method in real-world scenarios with scarce availability of hand-annotated data. Our models noticeably outperform all methods when a limited amount of real data is added to the synthetic data during training (Table 1, rows 5-7).

The experimentation with different dilation rates reveals that progressively increasing dilation as we move deeper ($DCNN_{2+3}$) yields better results than models trained using equal amounts of dilation ($DCNN_{2+2}$, $DCNN_{3+3}$). Models trained with dilation at the first layer and those trained using dilation rates of 4 and higher did not yield improvement over the vanilla CNN model CNN_{1+1} and hence are not reported. We speculate that progressively increasing dilation helps the model learn higher level features better further

down the network. Manual inspection of the alignments confirmed that long-term context was better captured by the progressively dilated CNNs than other models, and they could detect larger deviations in addition to short ones. We demonstrate the alignment paths generated by the comparative methods for three performance-score pairs in Figure 2 to facilitate qualitative understanding of our results. While *JumpDTW* struggles with deviations from the score, *NWTW* struggles with cases containing both forward and backward jumps (Fig. 2, example 1). This could be attributed to the waiting mechanism of $NWTW$, which makes backward jumps especially challenging. Our model $DCNN_{2+3}$ struggles with cases where there are multiple deviations within a short time span (Example 2). We speculate that this is due to the larger receptive fields of the dilated convolutions, which, while capturing greater context, are sometimes unable to capture multiple inflection points within a small context. Our model is specifically able to model forward jumps better than both the methods (Example 1), while also handling deviations not foreseeable from the score (Example 3).

We demonstrate that progressively dilated convolutional neural networks are effective at detecting structural differences between the score and the performance for structure aware audio-to-score alignment. While we used chroma-based features for score-performance audio pairs, our method can also be used with raw or scanned images of sheet music using learnt features, for instance, using multimodal embeddings trained on audio and sheet image snippets [15], [29]. Additionally, our method can also be utilized by frameworks other than DTW to generate the alignments. For instance, an unrolled score representation could be achieved via the inflection points predicted by our model, which could further be employed by score following approaches based on reinforcement learning [16], [17] for structure-aware tracking. The advantage of our method is that it does not require manually labelled block boundaries, and can effectively deal with deviations from the structure given in the score, in both the forward and backward directions. In the future, we would like to experiment with parallel dilation using different dilation rates and merging the learnt features. A current limitation of our method is the handling of trills and cadenzas, and we would also like to address these issues in future research.

5. REFERENCES

[1] Gerhard Widmer, "Getting closer to the essence of music: The con espressione manifesto," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 2, pp. 1–13, 2016.

[2] Andreas Arzt, *Flexible and Robust Music Tracking*, Ph.D. thesis, Universität Linz, Linz, 2016.

[3] Simon Dixon, "An on-line time warping algorithm for tracking musical performances," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2005, pp. 1727–1728.

[4] Meinard Müller, *Fundamentals of music processing: Audio, analysis, algorithms, applications*, Springer, 2015.

[5] Christian Fremerey, Meinard Müller, and Michael Clausen, "Handling repeats and jumps in score-performance synchronization," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2010, pp. 243–248.

[6] Maarten Grachten, Martin Gasser, Andreas Arzt, and Gerhard Widmer, "Automatic alignment of music performances with structural differences," in *Proceedings of 14th International Society for Music Information Retrieval Conference (ISMIR)*, 2013.

[7] Mark Levy and Mark Sandler, "Structural segmentation of musical audio by constrained clustering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 318–326, 2008.

[8] Aibo Tian, Wen Li, Linxing Xiao, Dong Wang, Jie Zhou, and Tong Zhang, "Histogram matching for music repetition detection," in *IEEE International Conference on Multimedia and Expo*, 2009, pp. 662–665.

[9] Andreas Arzt and Gerhard Widmer, "Towards effective 'anytime' music tracking," in *Proceedings of the Fifth Starting AI Researchers' Symposium*. IOS Press, 2010, pp. 24–36.

[10] Andreas Arzt and Gerhard Widmer, "Simple tempo models for real-time music tracking," in *Proceedings of the Sound and Music Computing Conference (SMC)*, 2010.

[11] Meinard Müller, Verena Konz, Andi Scharfstein, Sebastian Ewert, and Michael Clausen, "Towards automated extraction of tempo parameters from expressive music recordings," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2009, pp. 69–74.

[12] Vladimir Likic, "The Needleman-Wunsch algorithm for sequence alignment," *Lecture given at the 7th Melbourne Bioinformatics Course, Bi021 Molecular Science and Biotechnology Institute, University of Melbourne*, pp. 1–46, 2008.

[13] Tomohiko Nakamura, Eita Nakamura, and Shigeki Sagayama, "Real-time audio-to-score alignment of music performances containing errors and arbitrary repeats and skips," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 329–339, 2015.

[14] Yucong Jiang, Fiona Ryan, David Cartledge, and Christopher Raphael, "Offline score alignment for realistic music practice," *Proceedings of the Sound and Music Computing Conference (SMC)*, 2019.

[15] Matthias Dorfer, Jan Hajič Jr, Andreas Arzt, Harald Frostel, and Gerhard Widmer, "Learning audio–sheet music correspondences for cross-modal retrieval and piece identification," *Transactions of the International Society for Music Information Retrieval*, vol. 1, no. 1, 2018.

[16] Matthias Dorfer, Florian Henkel, and Gerhard Widmer, "Learning to listen, read, and follow: Score following as a reinforcement learning game," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2018.

[17] Florian Henkel, Stefan Balke, Matthias Dorfer, and Gerhard Widmer, "Score following as a multi-modal reinforcement learning problem," *Transactions of the International Society for Music Information Retrieval*, vol. 2, no. 1, 2019.

[18] Ruchit Agrawal and Simon Dixon, "Learning frame similarity using Siamese networks for audio-to-score alignment," in *2020 28th European Signal Processing Conference (EU-SIPCO)*. IEEE, 2021, pp. 141–145.

[19] Mengyi Shan and TJ Tsai, "Improved handling of repeats and jumps in audio-sheet image synchronization," *21st International Society for Music Information Retrieval Conference (ISMIR)*, 11-16 October 2020.

[20] Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse Engel, Sageev Oore, and Douglas Eck, "Onsets and frames: Dual-objective piano transcription," *International Society for Music Information Retrieval Conference (ISMIR)*, 2018.

[21] Thitaree Tanprasert, Teerapat Jenrungrot, Meinard Müller, and TJ Tsai, "Midi-sheet music alignment using bootleg score synthesis," *20th International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[22] Karen Ullrich, Jan Schlüter, and Thomas Grill, "Boundary detection in music structure analysis using convolutional neural networks.," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 417–422.

[23] Jan Schlüter and Sebastian Böck, "Improved musical onset detection with convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 6979–6983.

[24] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," *International Conference on Learning Representations (ICLR)*, 2016.

[25] Hyungtae Lee and Heesung Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4843–4855, 2017.

[26] Ruchit Rajeshkumar Agrawal, Marco Turchi, and Matteo Negri, "Contextual handling in neural machine translation: Look behind, ahead and on both sides," in *21st Annual Conference of the European Association for Machine Translation*, 2018, pp. 11–20.

[27] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science conference*, 2015, pp. 18–25.

[28] Craig Stuart Sapp, "Comparative analysis of multiple musical performances," in *8th International Conference on Music Information Retrieval (ISMIR)*, 2007, pp. 497–500.

[29] Stefan Balke, Matthias Dorfer, Luis Carvalho, Andreas Arzt, and Gerhard Widmer, "Learning soft-attention models for tempo-invariant audio-sheet music retrieval," *20th International Society for Music Information Retrieval Conference (ISMIR)*, 2019.