Improved Handling of Repeats and Jumps in Audio-Sheet Image Synchronization

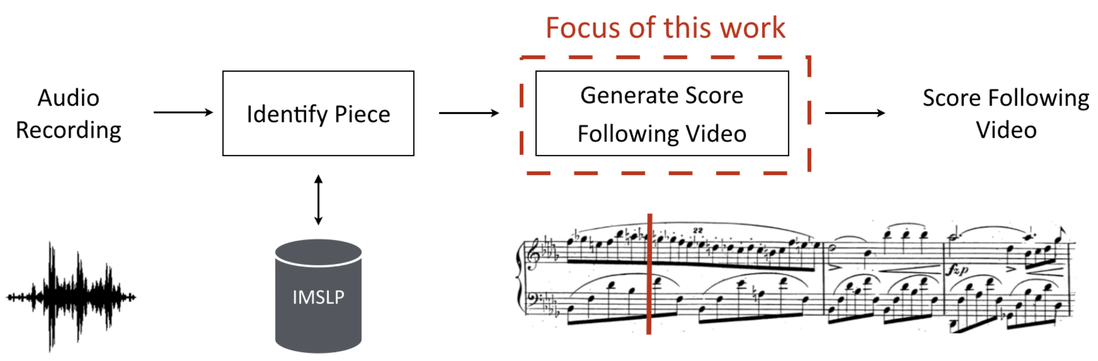
Mengyi Shan, TJ Tsai
Harvey Mudd College

Abstract

We propose an audio-sheet image synchronization system with a novel alignment algorithm called Hierarchical DTW that can handle repeats and jumps when jump locations are unknown a priori. On experiments with raw IMSLP data, it consistently outperforms a previously proposed Jump DTW algorithm.

Problem Statement

Problem:



Challenges:

- Raw sheet music PDFs from IMSLP
- Other pieces or filler pages in PDF
- Repeats & jumps

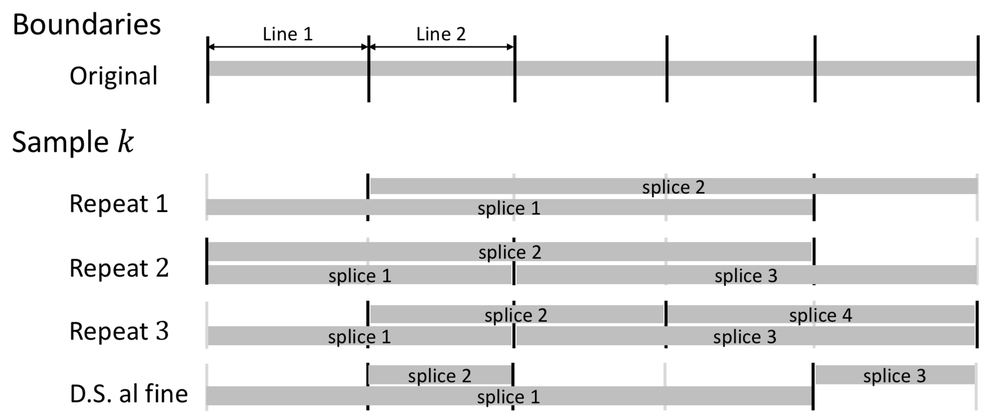
Experimental Setup

Data:

- 200 IMSLP piano sheet music PDFs. Completely unprocessed – may include other pieces or filler pages.
- 200 MIDI-synthesized audio files.
- Annotations of each sheet music line's start & end timestamp.

Benchmarks:

- Modify data to simulate repeats/jumps.
- 5 scenarios: No Repeat, Repeat x1, Repeat x2, Repeat x3, D.S. al fine.
- Jump locations randomly chosen among line breaks.

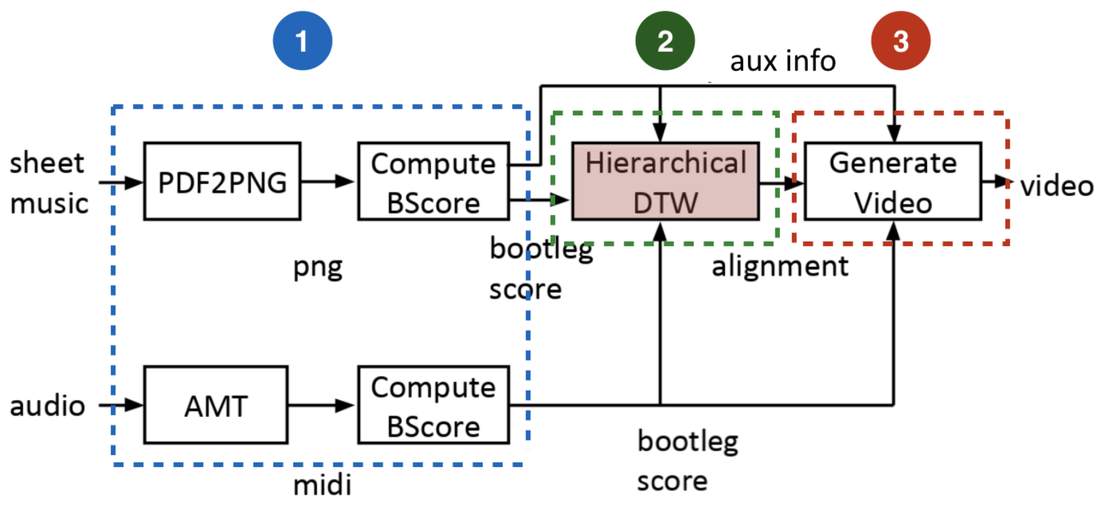


Evaluation:

- Accuracy: percentage of time correct line of music is shown.
- Uses scoring collar.

System Description

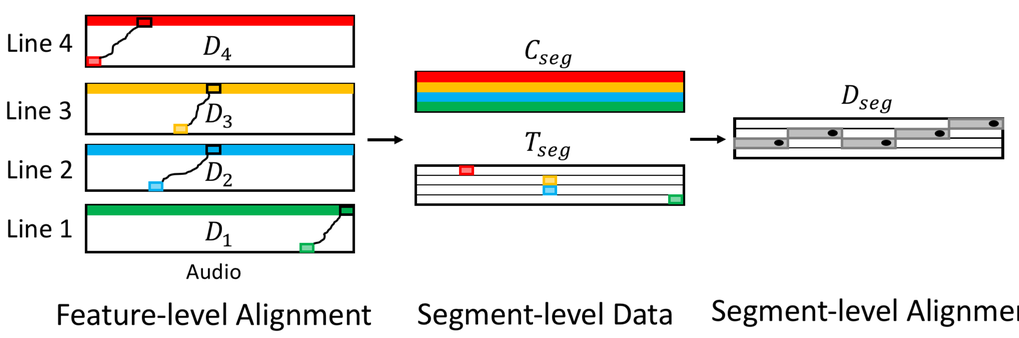
System architecture:



1 Feature Extraction:
The bootleg score is a feature representation for aligning piano sheet music images and MIDI [1].

2 Hierarchical DTW:
Aligns a sequence of sheet music lines against the audio. Handles jumps at unknown locations by allowing jumps at every line break. Three steps.

- **Feature-level alignment:** Align each sheet music line against audio using subsequence DTW. Cumulative cost matrices shown as D_i.
- **Segment-level data matrices:** (i) Matrix C_{seg} is segment-level cost matrix containing subsequence path scores. (ii) Matrix T_{seg} specifies starting location of every subsequence path.
- **Segment-level alignment:** Find the optimal path through C_{seg} using info in T_{seg} and set of allowable transitions. Can customize allowable transitions to impose domain knowledge.

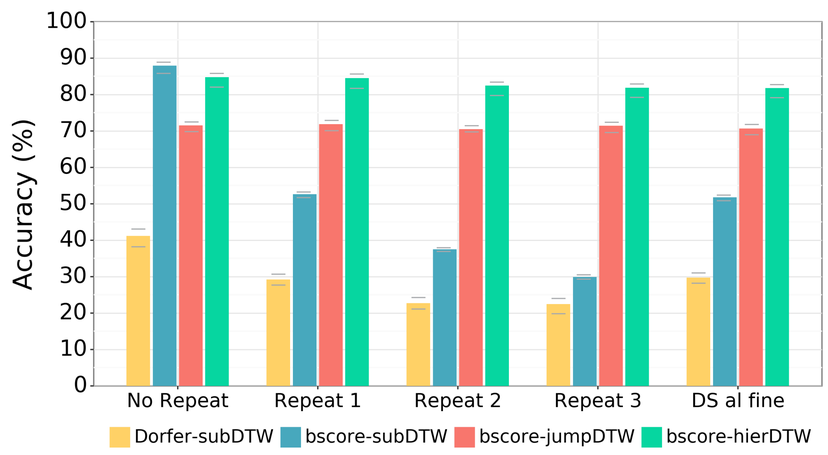


3 Video Generation:
Retrieve time information from alignment and pixel information from bootleg score. Show the predicted line of sheet music at every time instant in the audio recording.

Results and Analysis

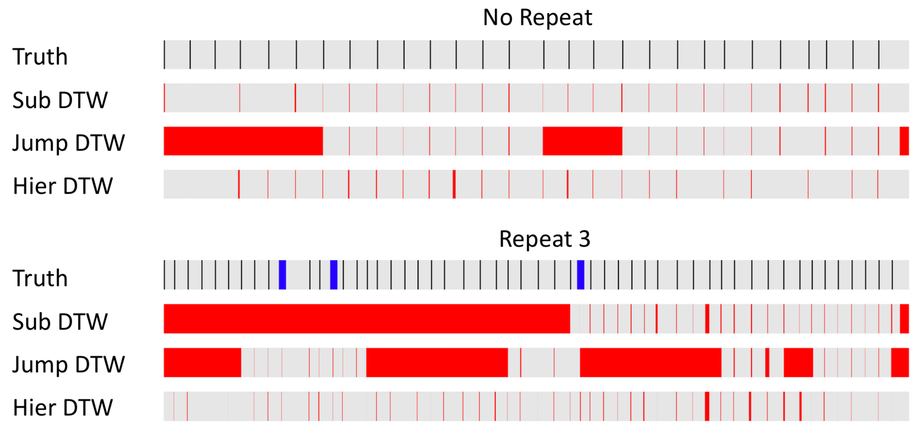
Results:
Compare with subsequence DTW, jump DTW [2] and Dorfer [3].

- Performs much better than Jump DTW across all benchmarks.
- Performs comparably with subsequence DTW with no repeats.



Analysis:

- Visualize predictions on specific examples. Errors in red, ground truth in black, repeats in blue.
- All systems have errors near line breaks. Subseq DTW cannot handle jumps. Jump DTW can handle jumps but also inserts spurious jumps.
- Hierarchical DTW's errors mostly come from bootleg score problems and repetitive music.



References

[1] Daniel Yang et al. "Midi passage retrieval using cell phone pictures of sheet music, " in ISMIR 2019.
[2] Christian Fremerey et al. "Handling repeats and jumps in score performance synchronization," in ISMIR 2010.
[3] Matthias Dorfer et al. "Learning audio-sheet music correspondences for cross-modal retrieval and piece identification," in TISMIR 2018.