BEAT ESTIMATION FROM MUSICIAN VISUAL CUES

Sutirtha Chakraborty [1], Senem Aktaş [2] , William Clifford [1] , Joseph Timoney [1]

[1] Maynooth University, Maynooth, Ireland

[2] Bolu Abant Izzet Baysal University, Bolu, Turkey

Introduction

Can robots play music ***WITH*** and ***LIKE*** humans?

- Tapping their feet, clapping their hands
- A fully connected network
- Expression and emotions

Literature

Understanding beats from dance moves

- Movements are prominent
- Performers are synchronized

Musical Conductor

- Controls the pace of ensemble
- Maintain the expression in of the performance
- Time varying leader – follower relationship

URMP : Multi-Modal Music Dataset

A dataset for facilitating audio-visual analysis of musical performances. The dataset comprises a few simple multi-instrument musical pieces assembled from coordinated but separately recorded performances of individual tracks.

Type of Performance	Name	Duration	Instrument*	Total Beats
Duet	Jupiter	01:03	Vn,Vc	86
	Sonata	00:46	Vn,Vn	44
	The Entertainer	01:27	Tpt, Tpt	216
Trio	Spring for the Four Seasons	00:35	Vn,Vn,Vc	65
	Hark the Herald Angels	00:47	Vn,Vn,Va	88
	Waltz from Sleeping Beauty	01:33	Fl,Fl,Cl	304
Quartets	Pirates of the Aegean	00:50	Vn,Vn,Va,Vc	163
	Pirates of the Aegean	00:50	Vn,Vn,Va,Sax	163
	In the Hall of Mountain King	01:25	Vn,Vn,Va,Vc	160
Quintets	Miserere Mei Deus	00:40	Fl,Fl,Ob,Cl,Bn	87
	Miserere Mei Deus	00:40	Fl,Fl,Ob,Cl,Bn	86
	Chorale	00:53	Tpt,Tpt,Hn,Tba,Tba	144
*Instruments : Vn=Violin, Vc=Cello, Tpt= Trumpet, Va= Viola, Fl=Flute, Ob=Oboe, Cl=Clarinet, Bn=Basson, Tba= Tuba				

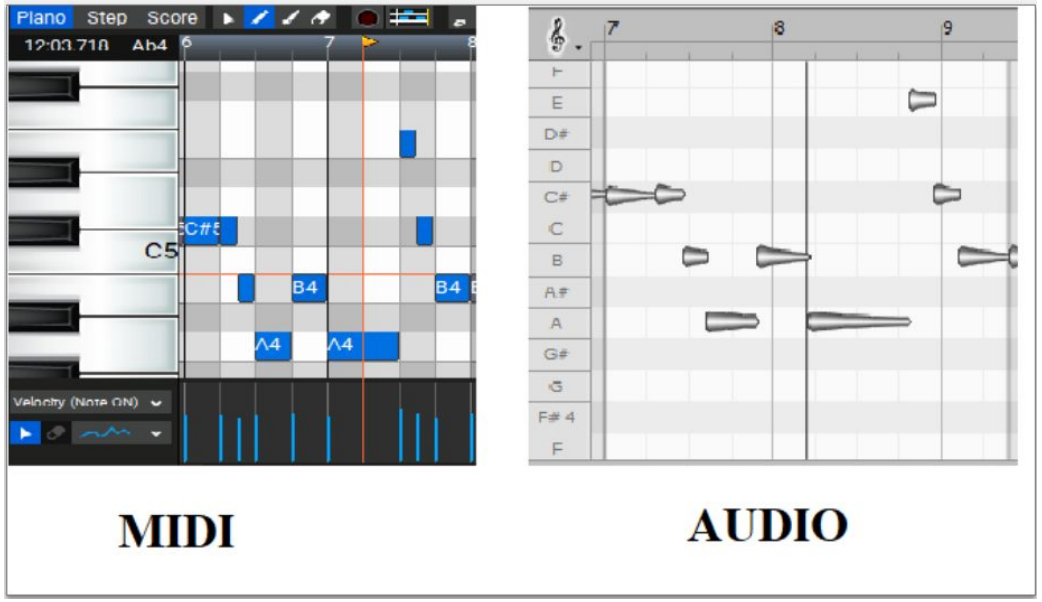
Understanding From Musicians

- Analyze the connection between musicians' body movements and the associated audio rhythms.

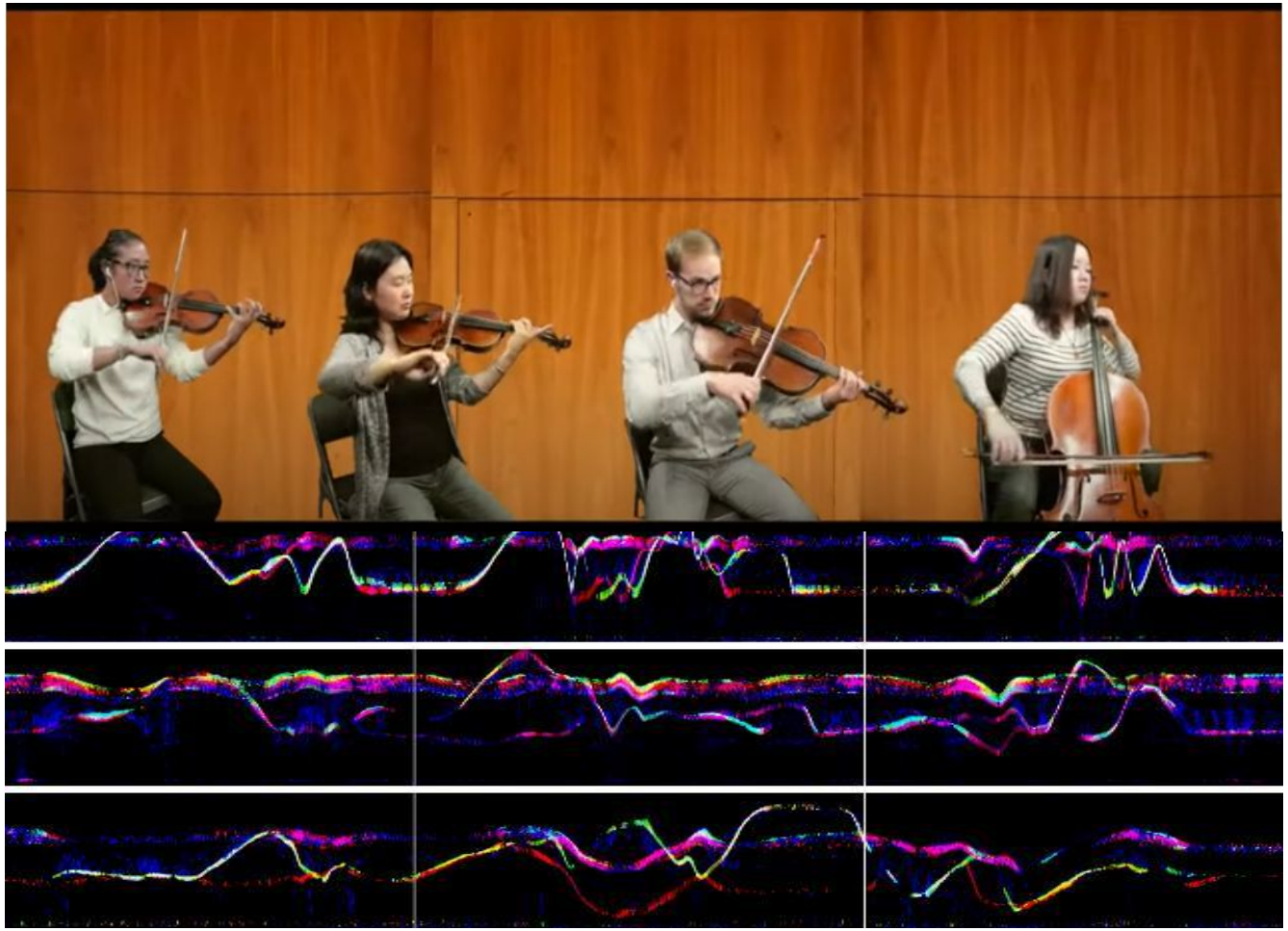
- Evaluating performance using two different approaches to predict rhythmic phases from multiple musicians' body sway.

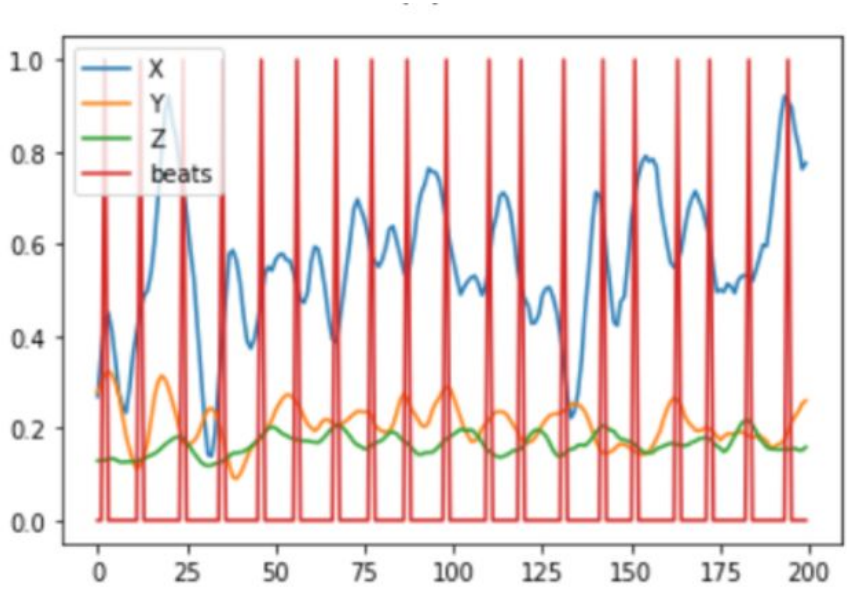
Getting Ground Truth

- Align Midi and Audio data
- Apply Dynamic Time Wrapping (DTW)
- Identify and get beats



Motiongram



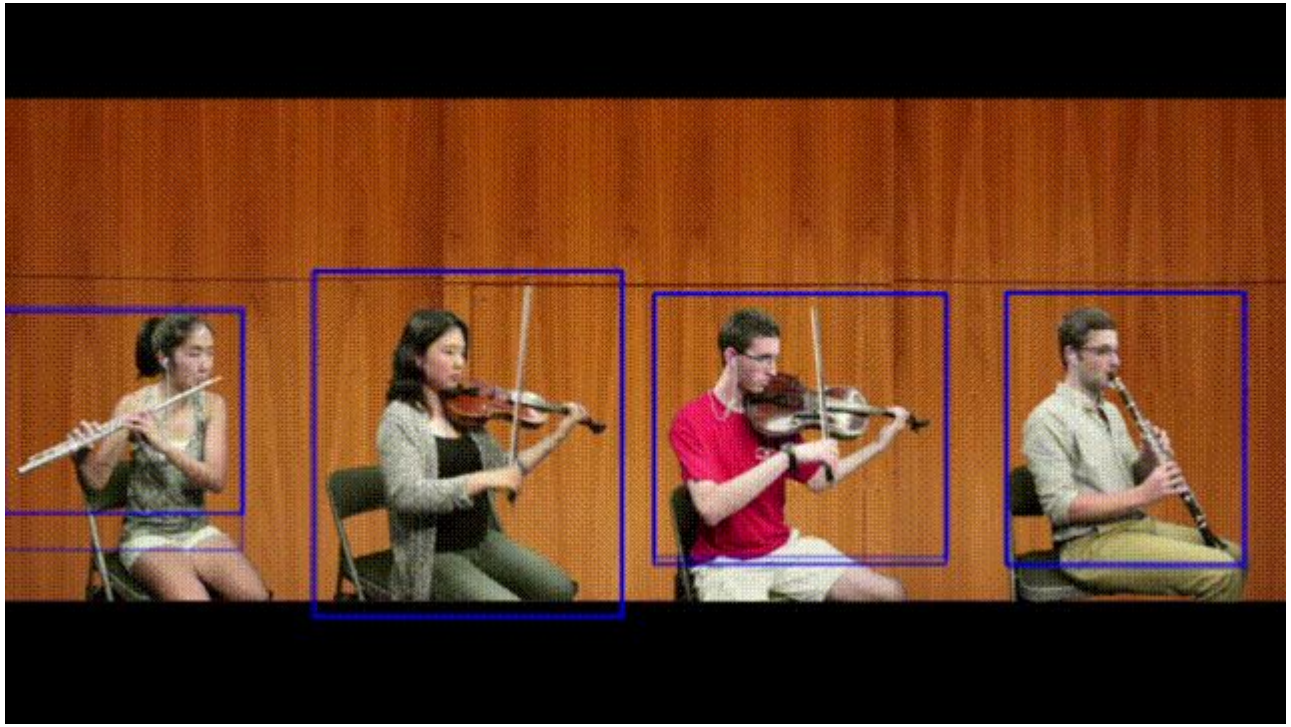


X= Horizontal Motion
Y= Vertical Motion
Z = Quality of Motion

Pose Estimation

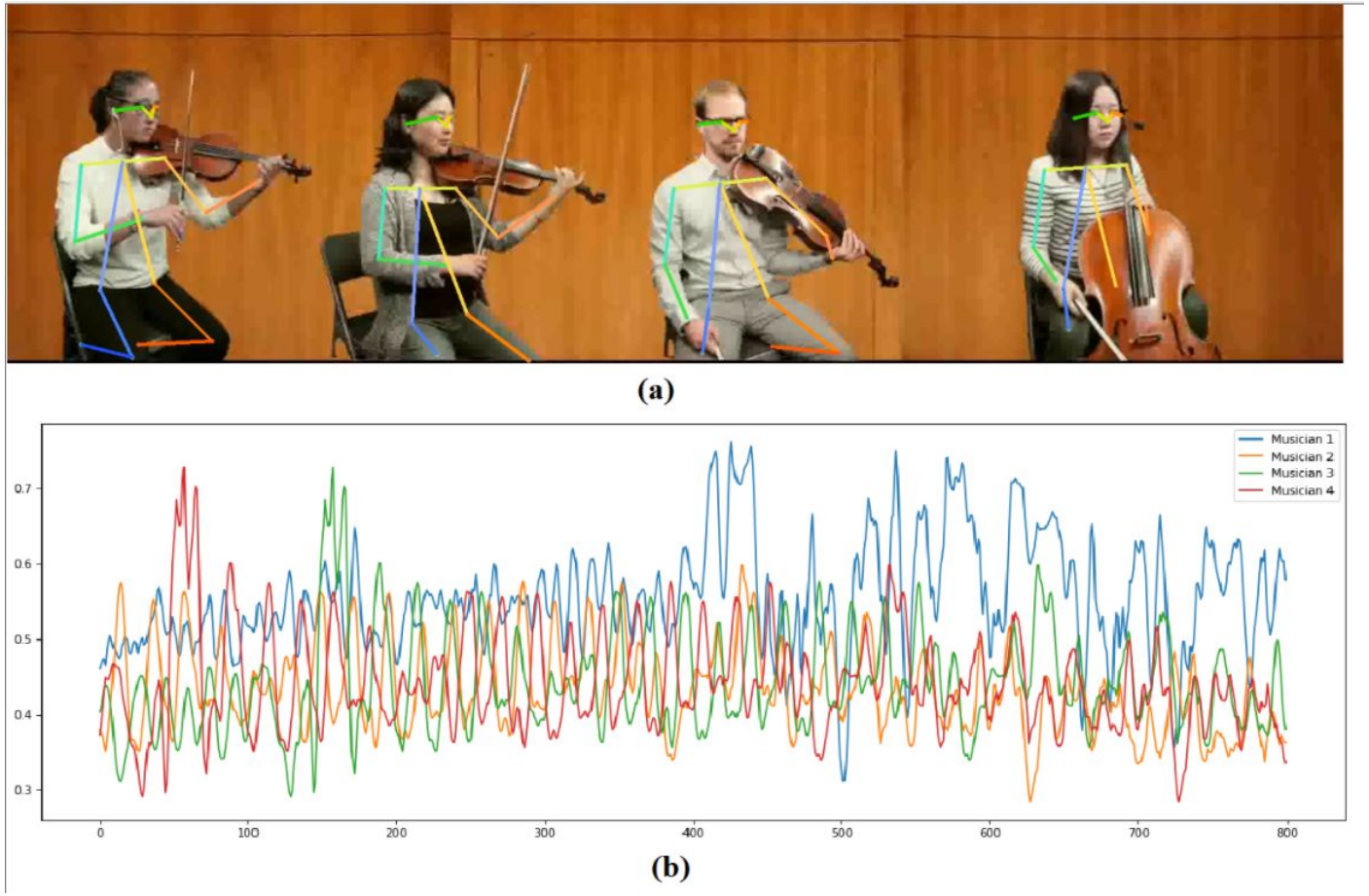
Two different approaches used :

1. First Frame as Reference
2. Spatial Derivative of keypoints

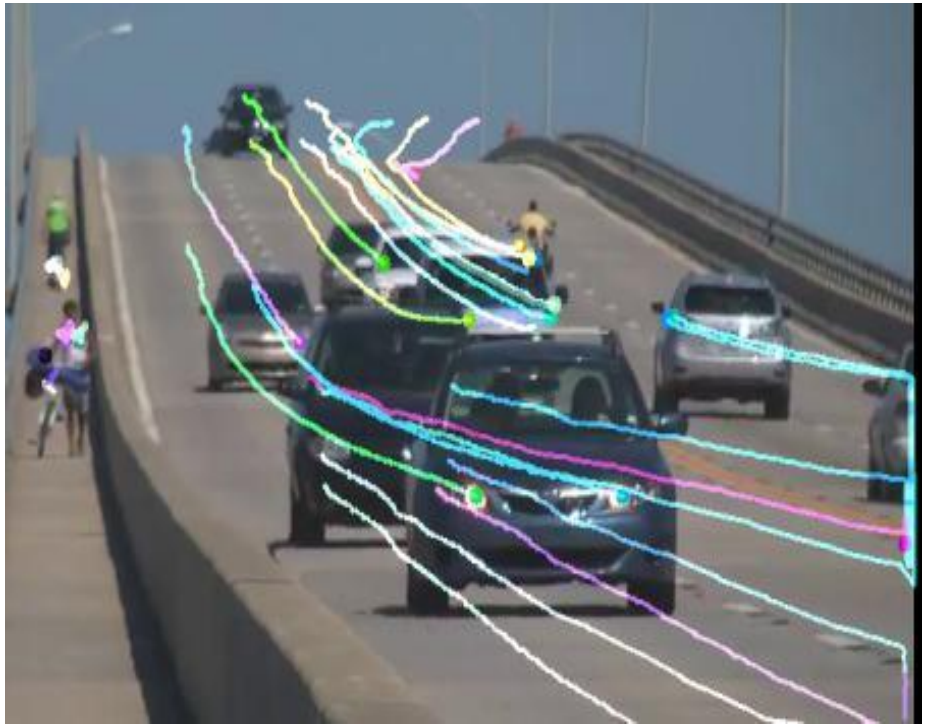


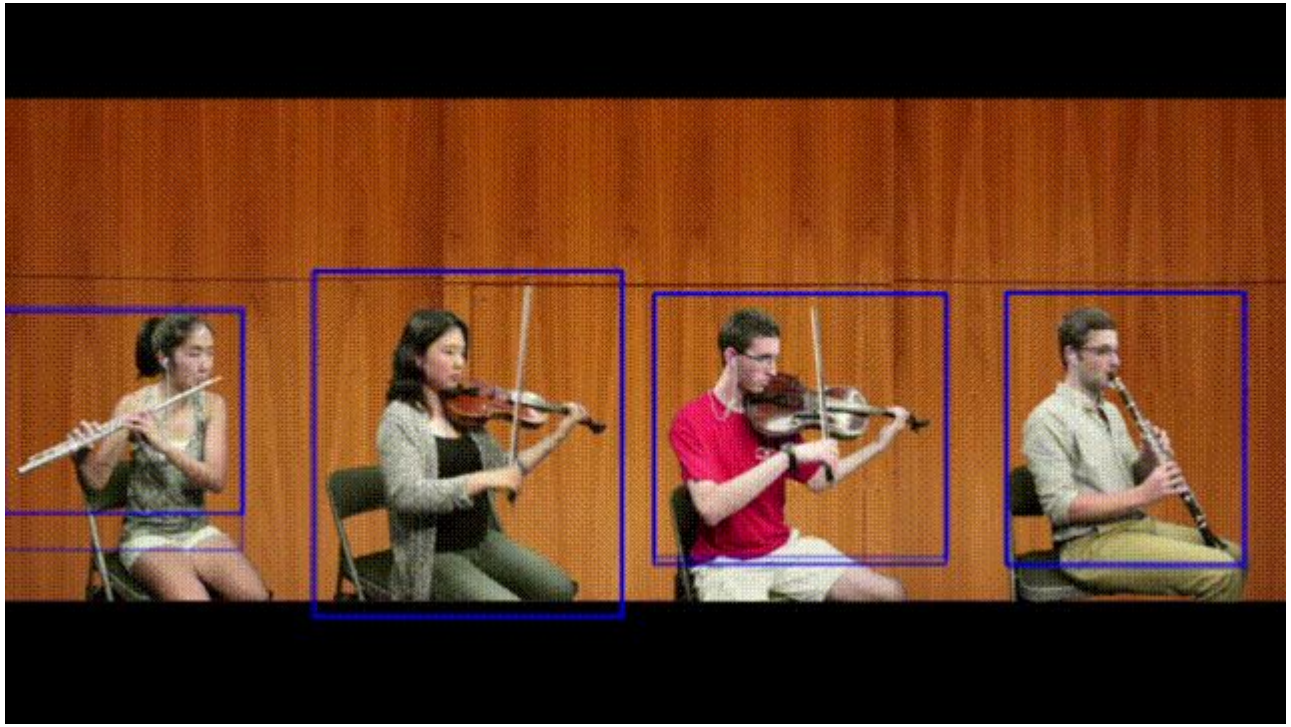
Pose Estimation - First Frame as Reference

- 17 keypoints extracted from each frame
- Estimating the average motion of each individual



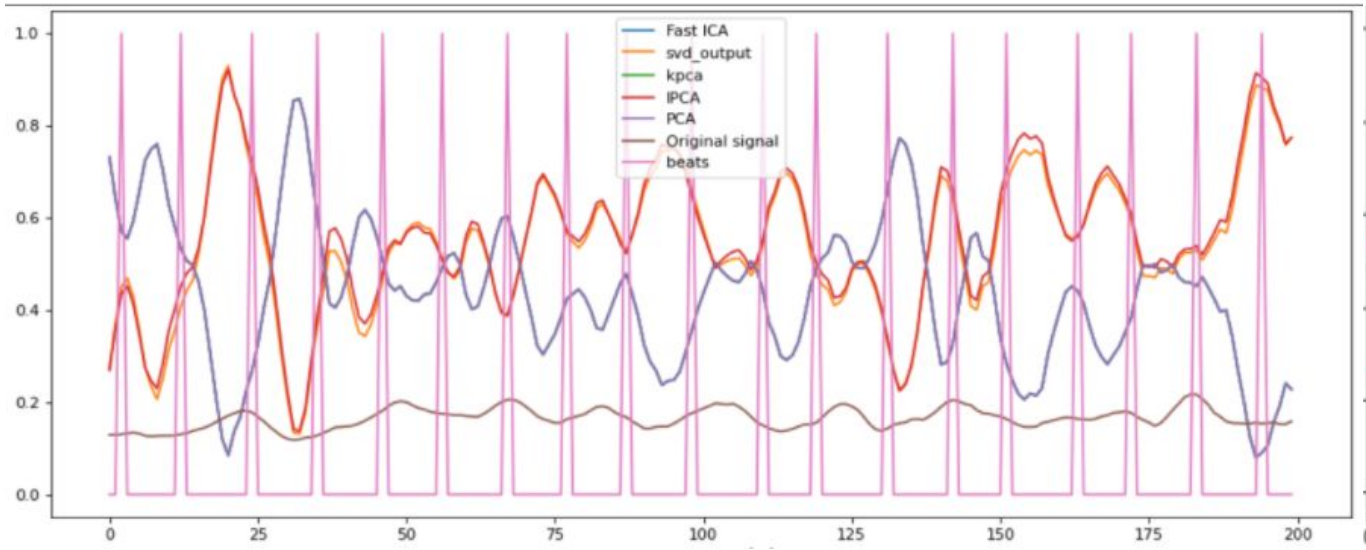
Pose estimation- Spatial Derivative of Keypoints

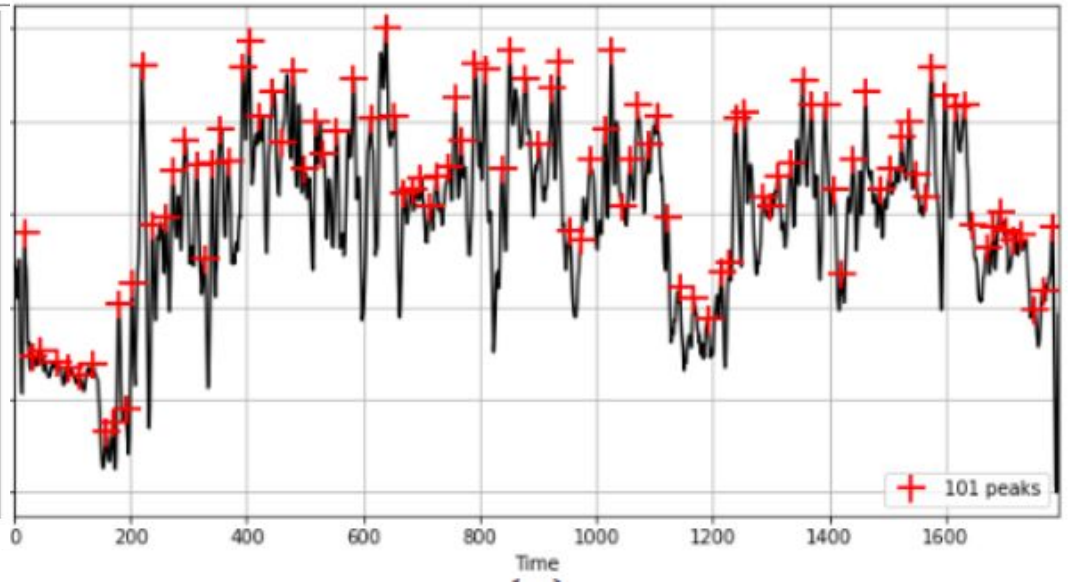




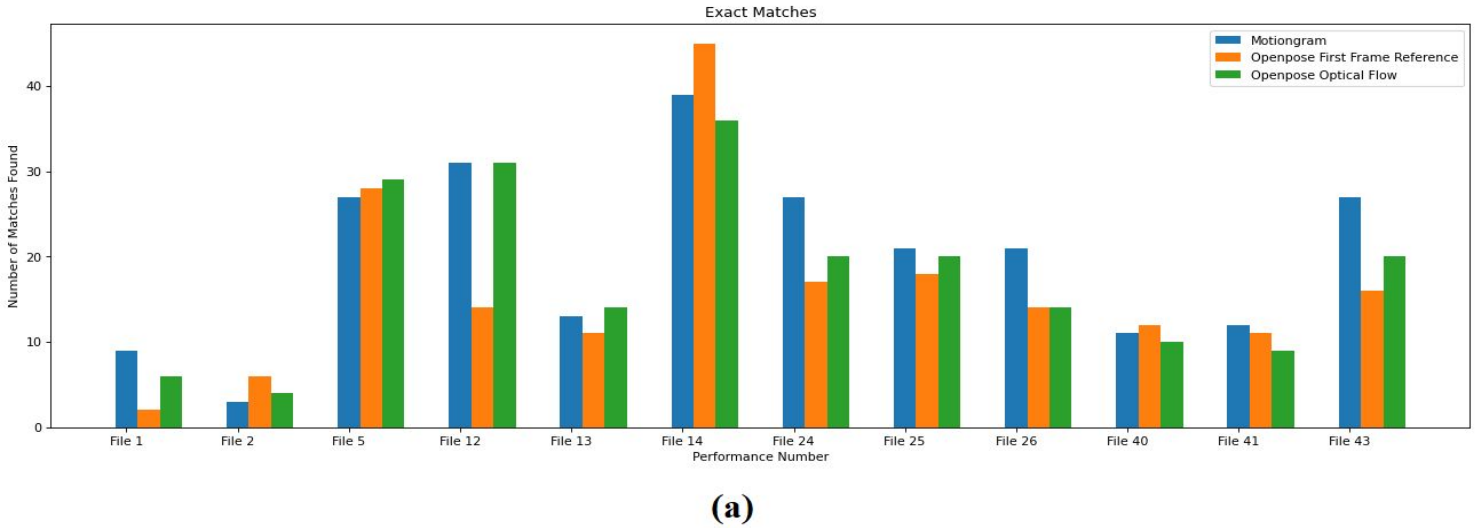
Peak Detection and Generating '*Trust*'

- Smoothing with Okada filter
- Used decomposition algorithm for signal multiplexing
- Used 'peak' and 'valley' detection on each algorithm
- Created a trust value based on timestamp

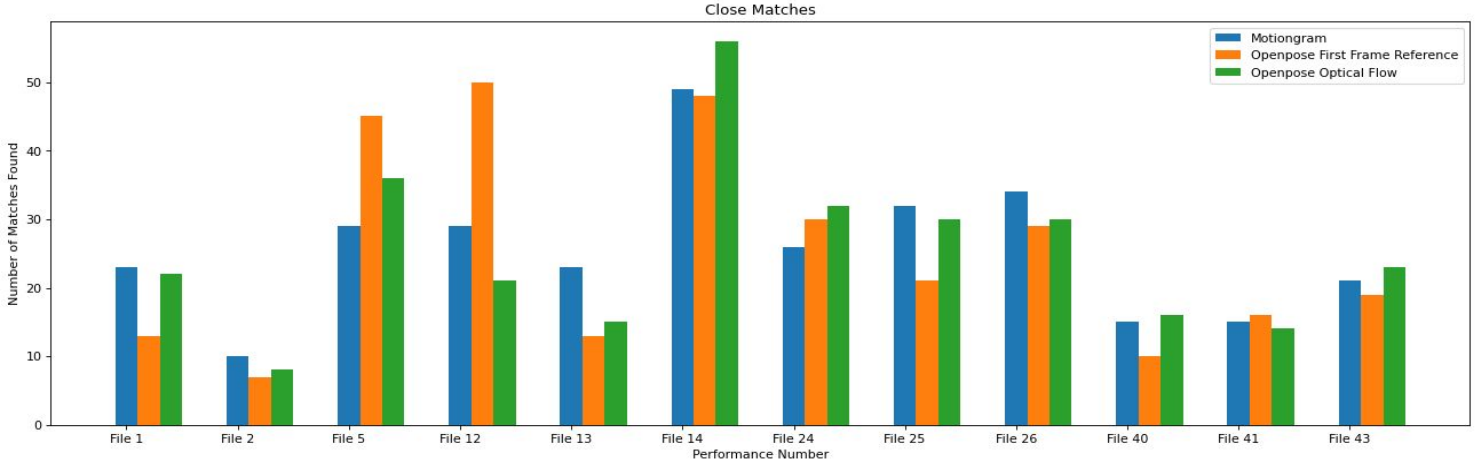




Results



(a)



(b)

Original Beats

(a) Exact Match
(b) Close Match with 100 ms difference

Predicted Beats

Conclusion

- Phases can be estimated from musician's body sway.
- Estimation techniques depends on musical instrument.
- Considering musicians' as independent oscillator.

Future Work

- A post processing for better result.
- Include both audio and video for beat prediction.
- Machine learning techniques for predictions.