See and Listen: Score-informed Association of Sound Tracks to Players in Chamber Music Performance Videos

Bochen Li, Karthik Dinesh, Zhiyao Duan, Gaurav Sharma

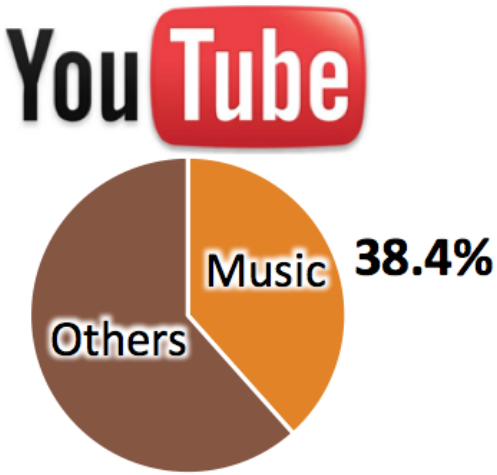
Department of Electrical and Computer Engineering, University of Rochester

March 7, 2017

UNIVERSITY of ROCHESTER

Background

- Music → Multi-modal art form
- See and listen → More enjoyment
- Popular music video streaming service

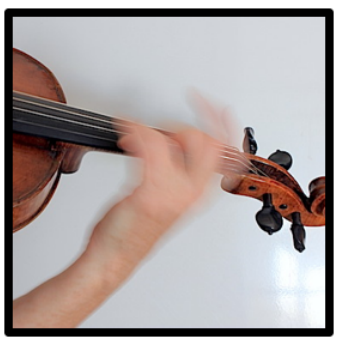


Background

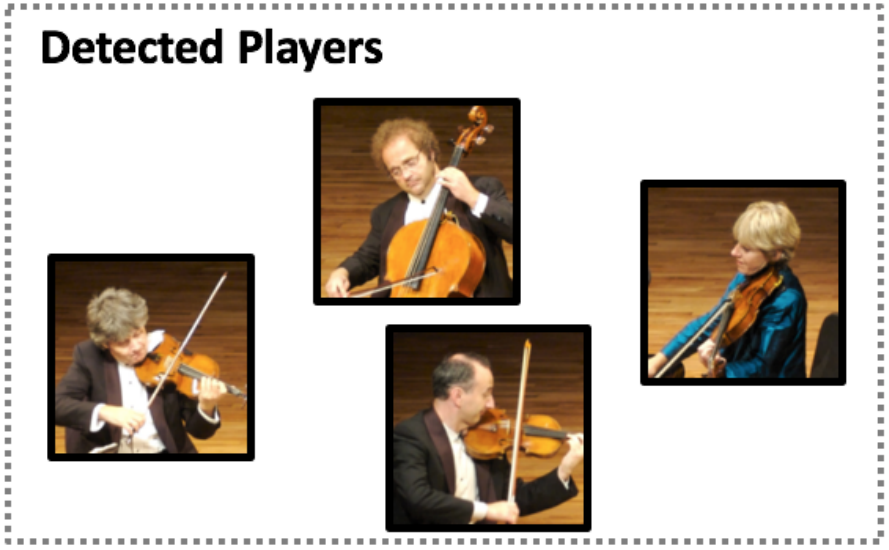
Multi-modal Music Information Retrieval

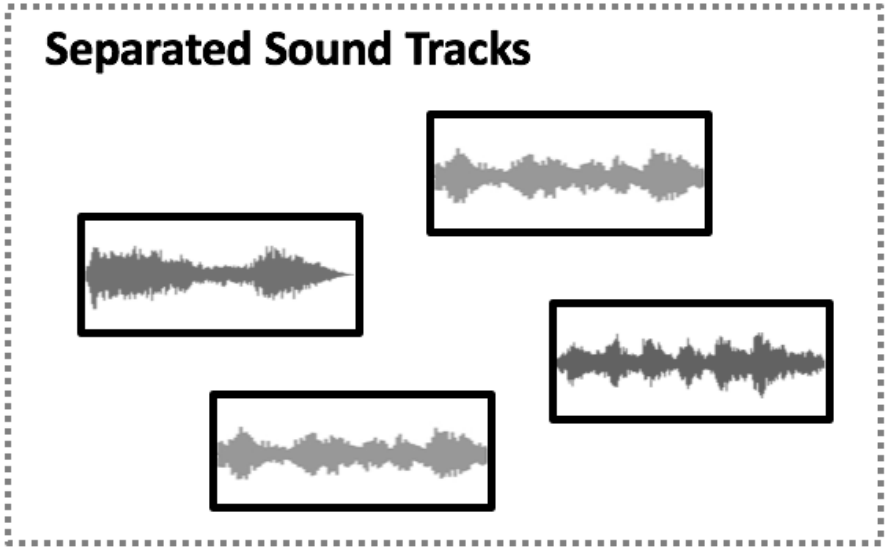
- Instrument Recognition
- Playing Activity Detection
- Polyphonic Music Analysis
- Fingering Investigation
- Conductor Following



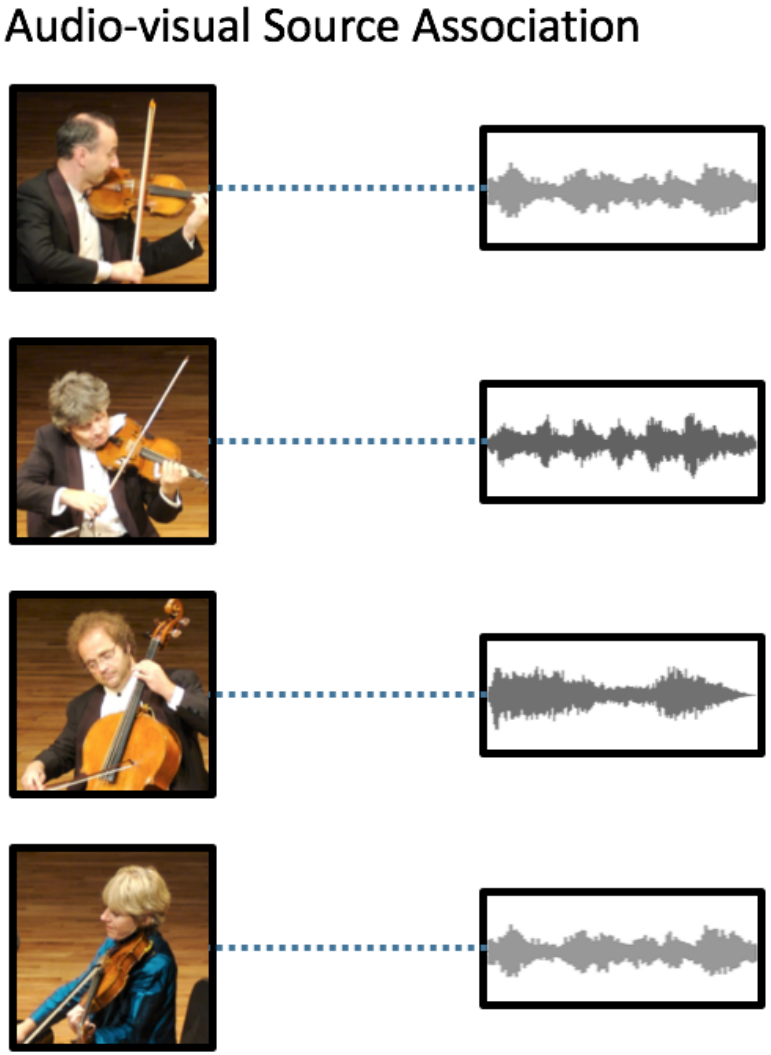



Source Association





Source Association



Background

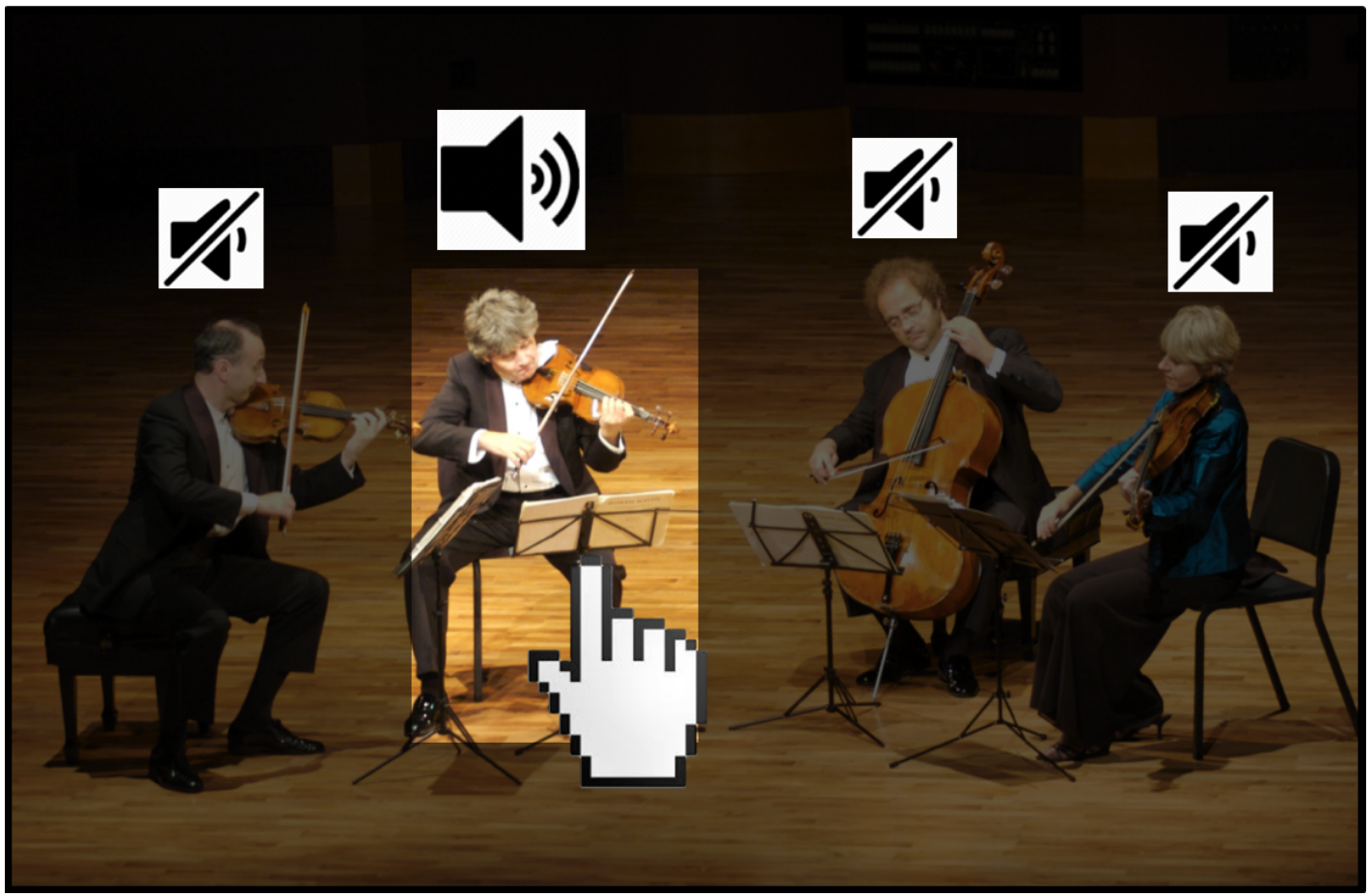
Applications

- Intuitive and user-friendly interaction with music performance videos
- Smart Music Editor

Background

Applications

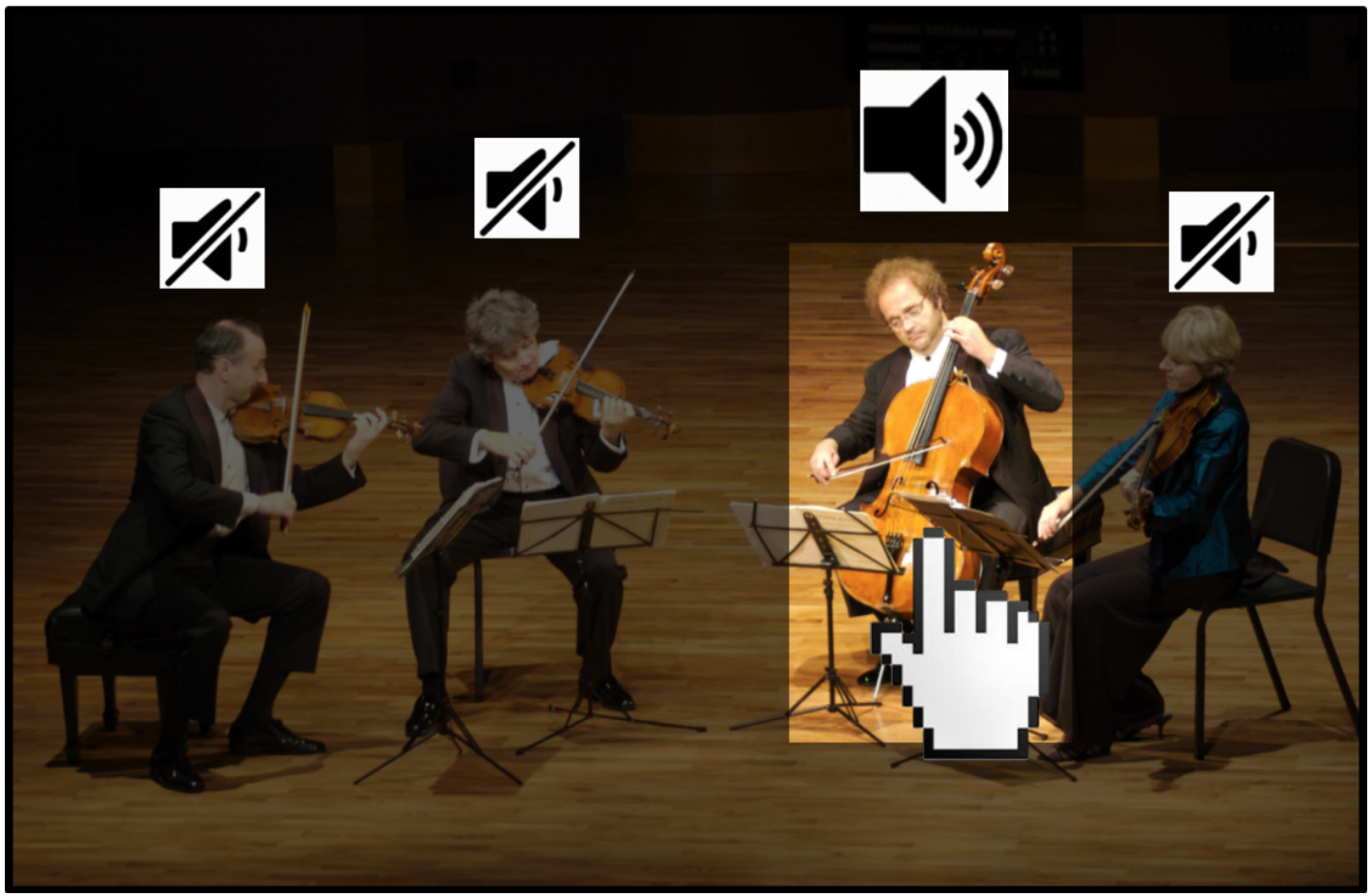
- Intuitive and user-friendly interaction with music performance videos
- Smart Music Editor



Background

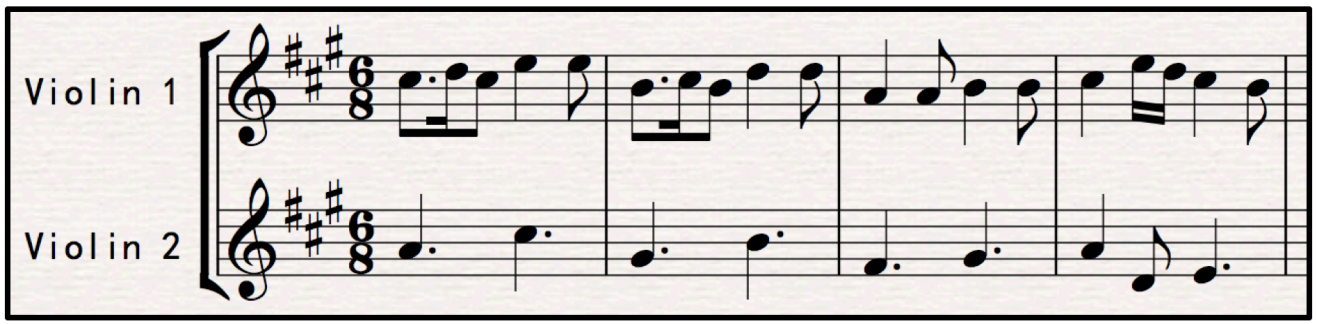
Applications

- Intuitive and user-friendly interaction with music performance videos
- Smart Music Editor

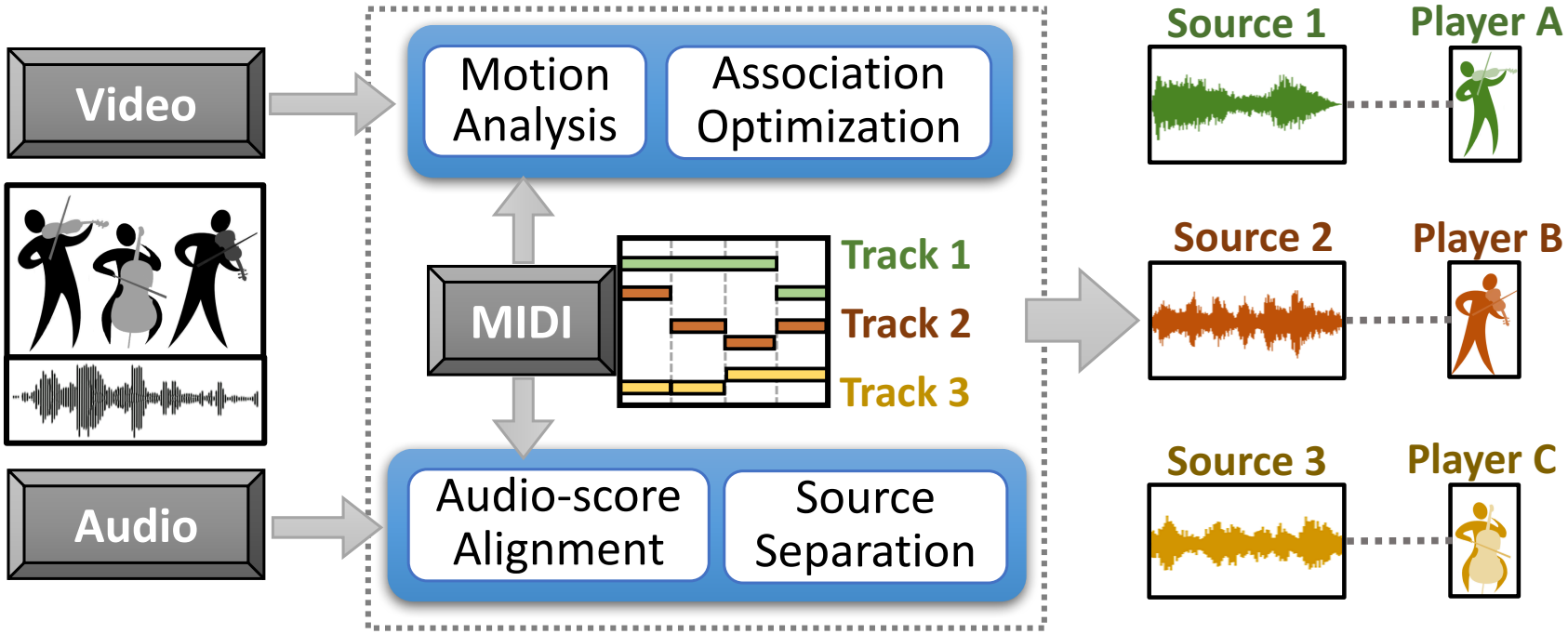


System Overview

Example:



System Overview

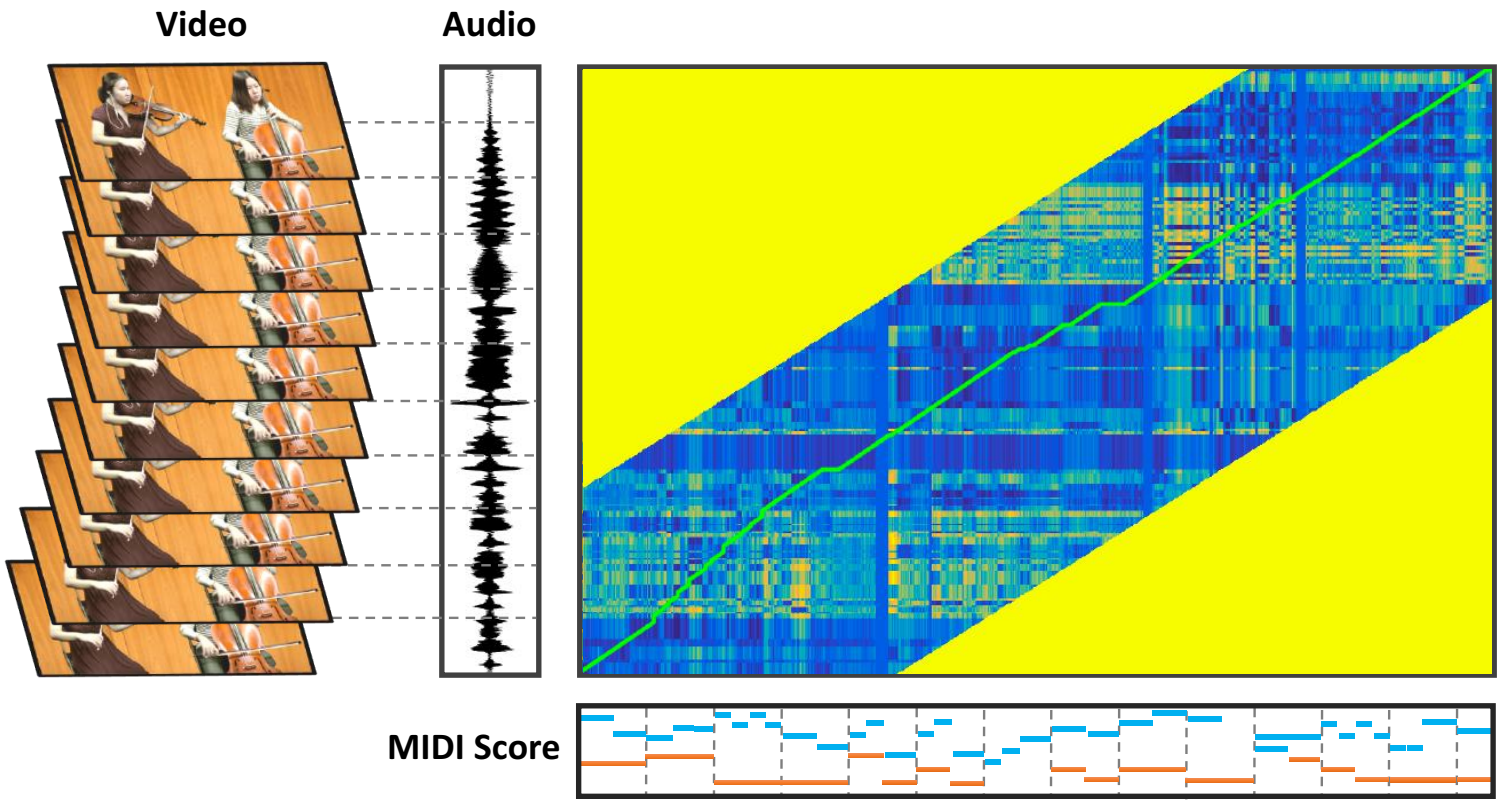


- Score-informed
- String-instruments
- Bow stroke $\Rightarrow$ audio event
- Correlate bow strokes with audio onsets

- Method: Audio Analysis
  - Audio-score Alignment
- Method: Video Analysis
  - Optical Flow Estimation
  - Player Detection
  - Bowing Motion Capturing
- Method: Association Optimization
- Experiments
  - Dataset
  - Evaluation Measure
  - Results

Method: Audio Analysis

Audio-score Alignment



- Chroma Feature & Dynamic Time Warping
- Video-score Alignment

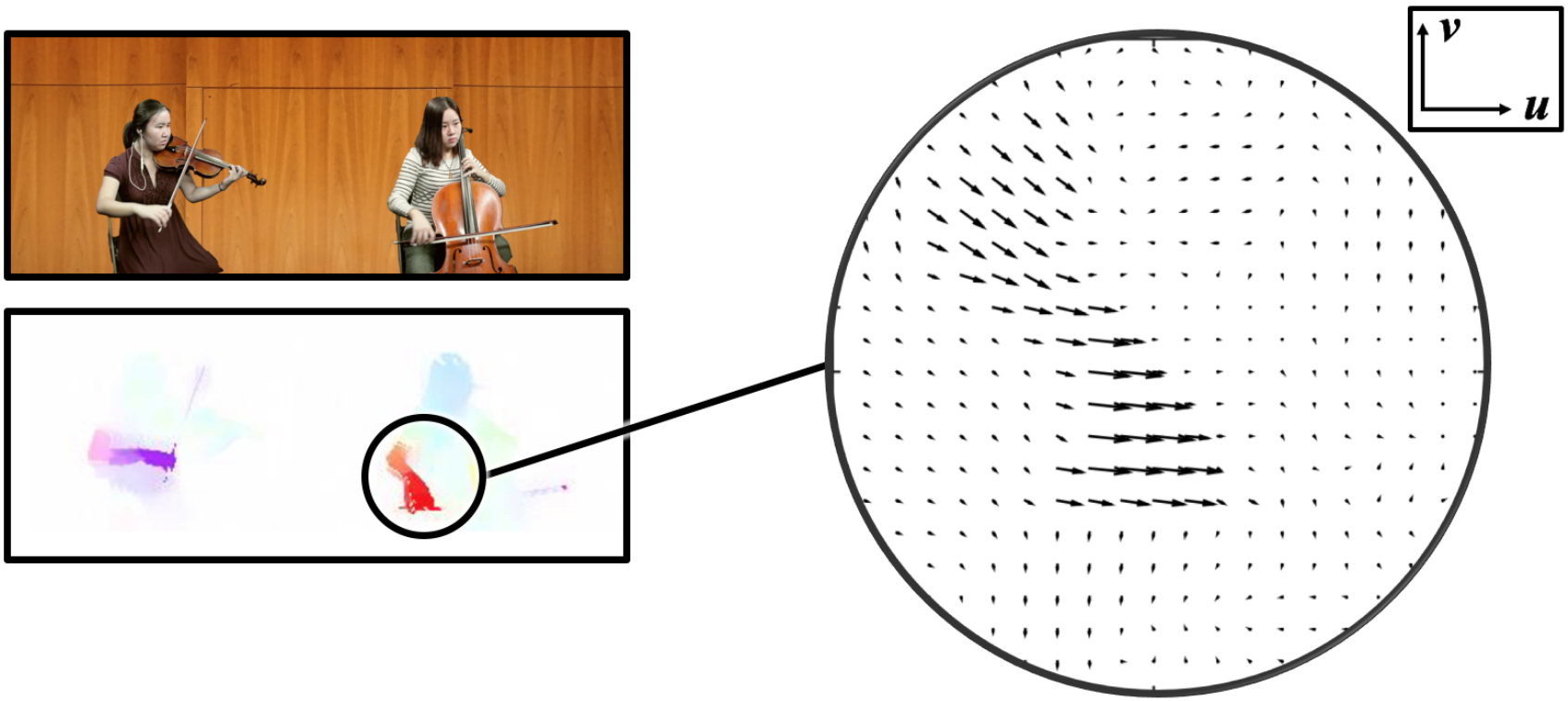
- Method: Audio Analysis
  - Audio-score Alignment
- Method: Video Analysis
  - Optical Flow Estimation
  - Player Detection
  - Bowing Motion Capturing
- Method: Association Optimization
- Experiments
  - Dataset
  - Evaluation Measure
  - Results

Method: Video Analysis

Optical Flow Estimation

The motion velocity of each pixel between two adjacent frames

Method: Sun et al. [2]

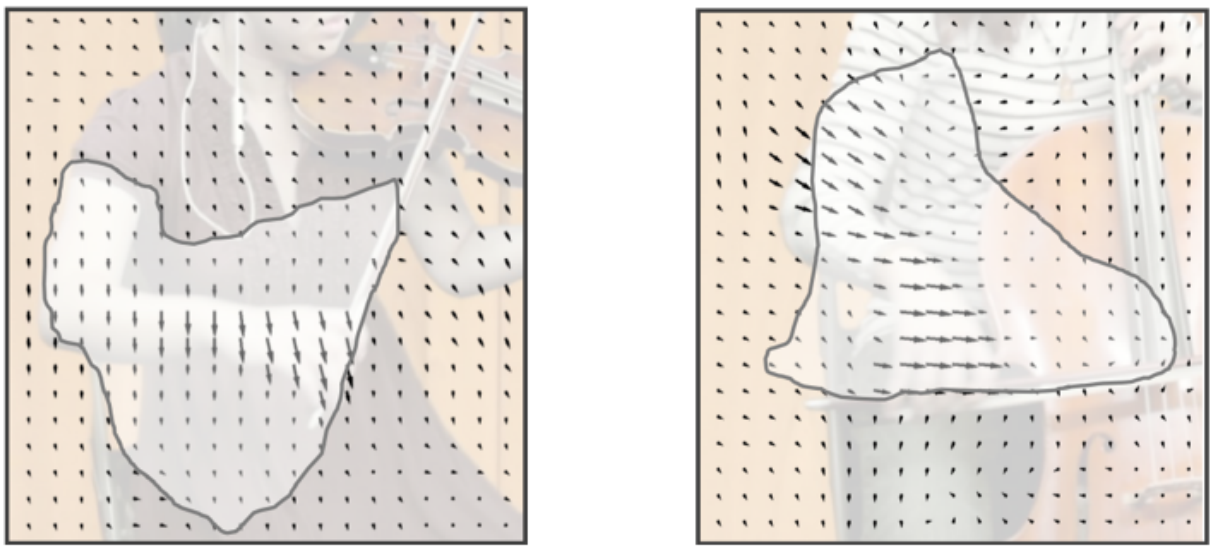


[2] D. Sun, S. Roth, and M. J. Black, Secrets of optical flow estimation and their principles, in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2010.

Method: Video Analysis

Player Detection

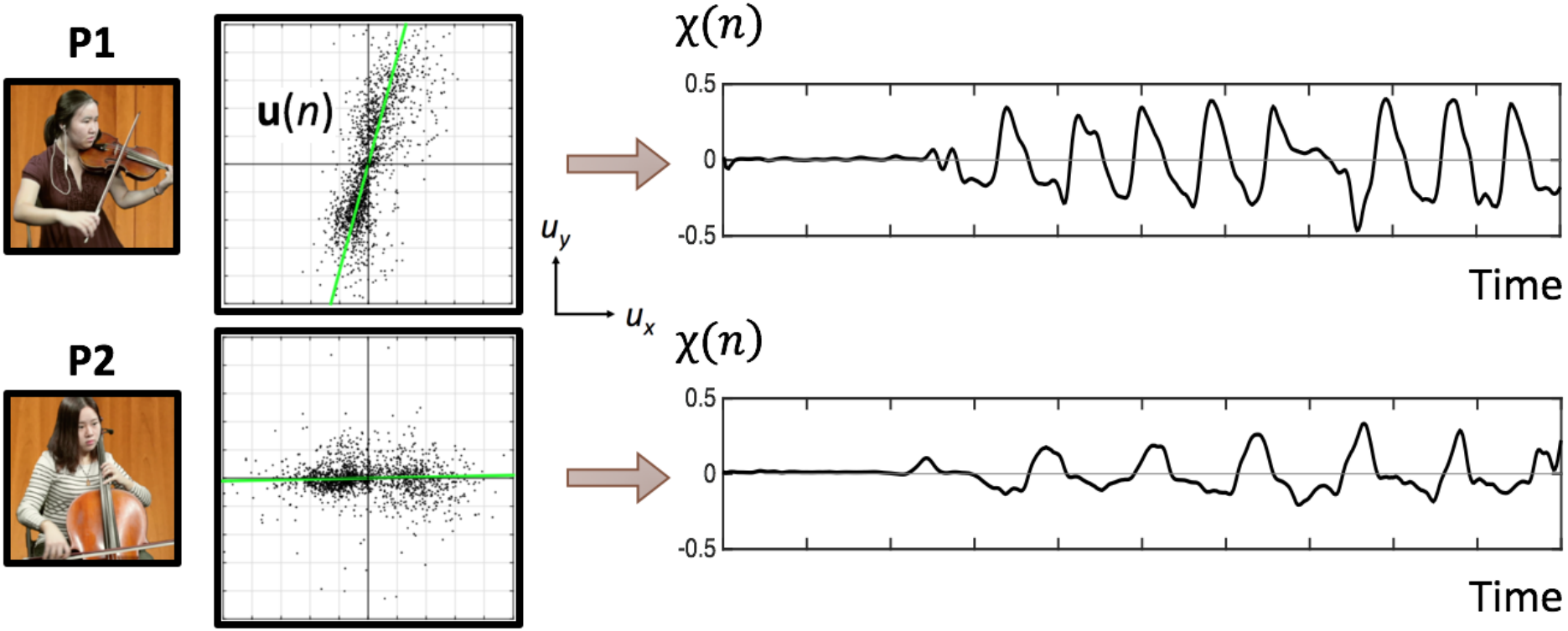
- Flow Magnitude ⇒ GMM Clustering ⇒ Player Region
- Thresholding ⇒ High Motion Region



Method: Video Analysis

Bowing Motion Capturing

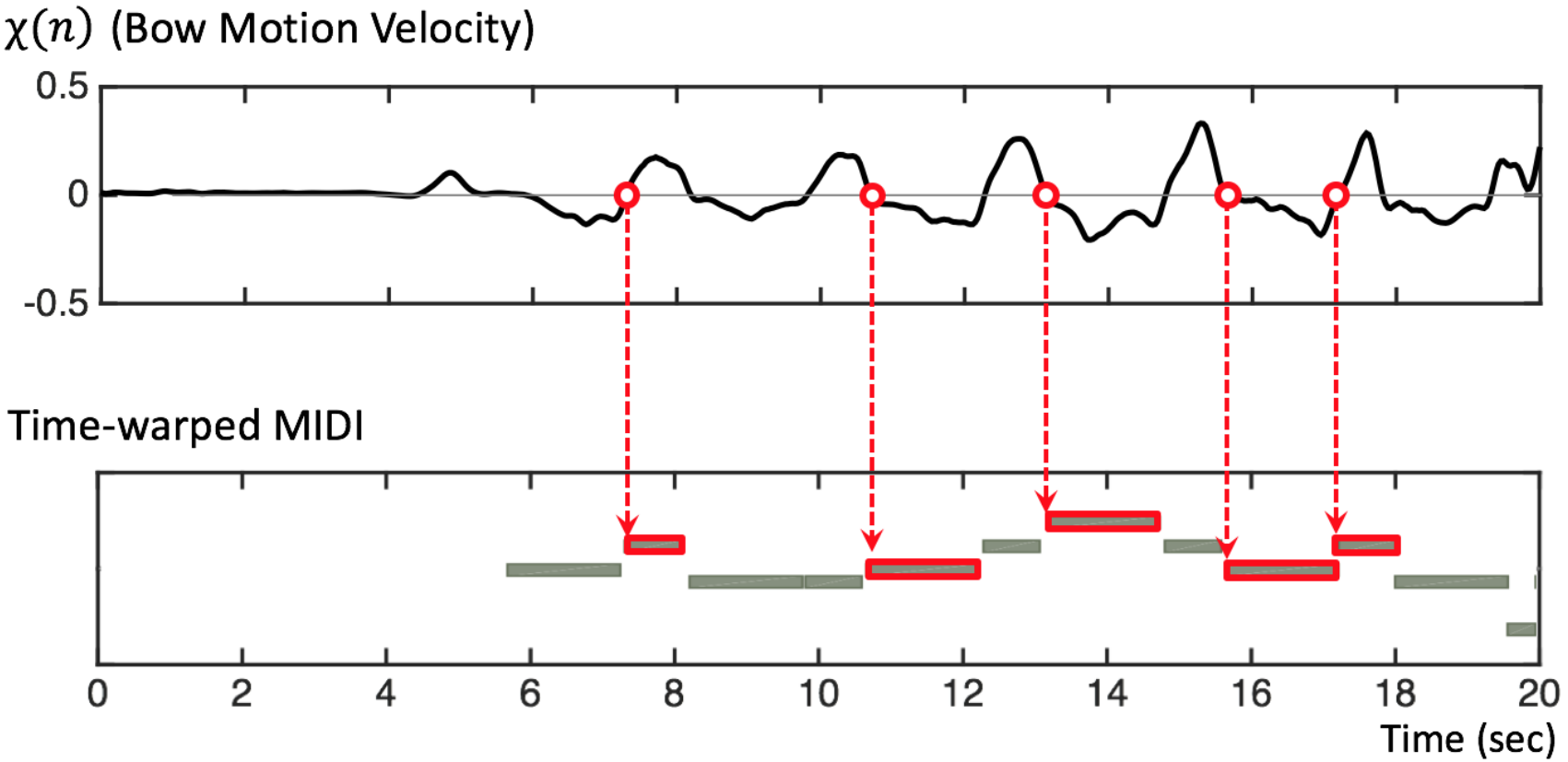
- Frame-wise global motion vector $\Rightarrow \mathbf{u}(n) = [u_x(n), u_y(n)]^T$.
- Principal component analysis (PCA) $\Rightarrow \tilde{\mathbf{u}} = (\tilde{u}_x, \tilde{u}_y)^T$
- Dimension reduction $\Rightarrow \chi(n) = \frac{\mathbf{u}(n)^T \tilde{\mathbf{u}}}{\|\tilde{\mathbf{u}}\|}$



Method: Video Analysis

Bowing Motion Capturing

Correlation:



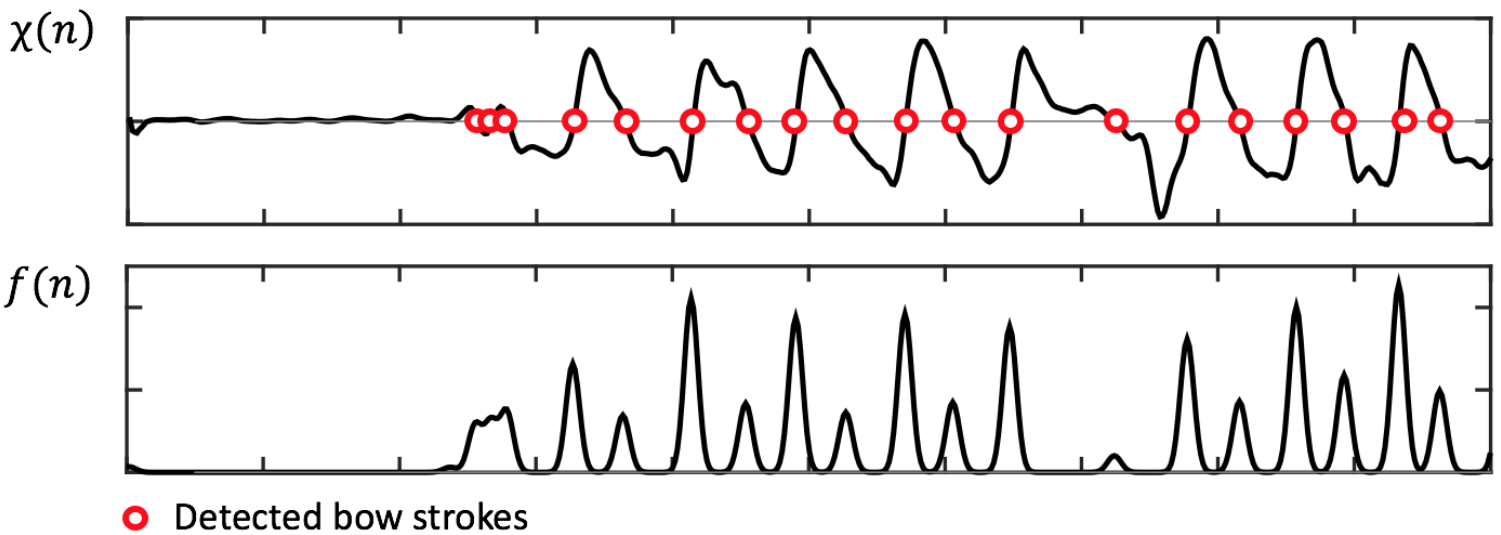
- Method: Audio Analysis
  - Audio-score Alignment
- Method: Video Analysis
  - Optical Flow Estimation
  - Player Detection
  - Bowing Motion Capturing
- Method: Association Optimization
- Experiments
  - Dataset
  - Evaluation Measure
  - Results

Method: Association Optimization

Pair-wise Matching

Bow Onset Likelihood:

$$f(n) = \left(\sum_{m \in \mathcal{Z}} \bar{\chi}(m) \cdot \delta(n, m) \right) * \mathcal{N}(0, \sigma^2) \tag{1}$$

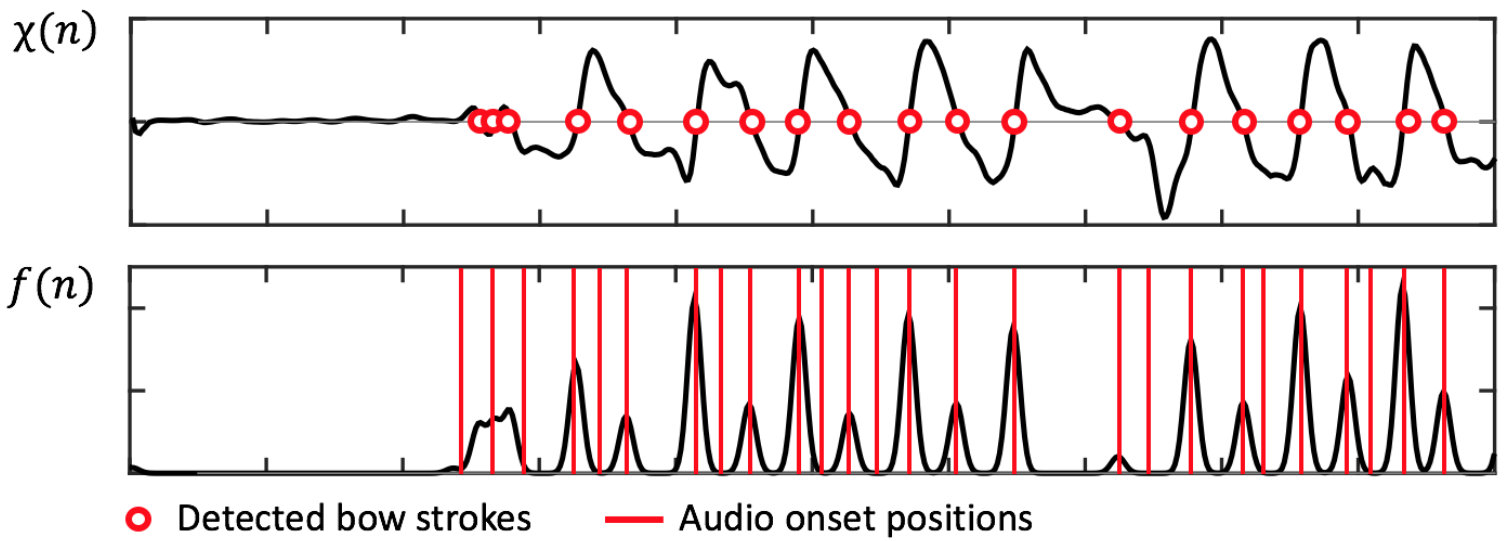


Method: Association Optimization

Pair-wise Matching

Bow Onset Likelihood:

$$f(n) = \left(\sum_{m \in \mathcal{Z}} \bar{\chi}(m) \cdot \delta(n, m) \right) * \mathcal{N}(0, \sigma^2) \tag{2}$$

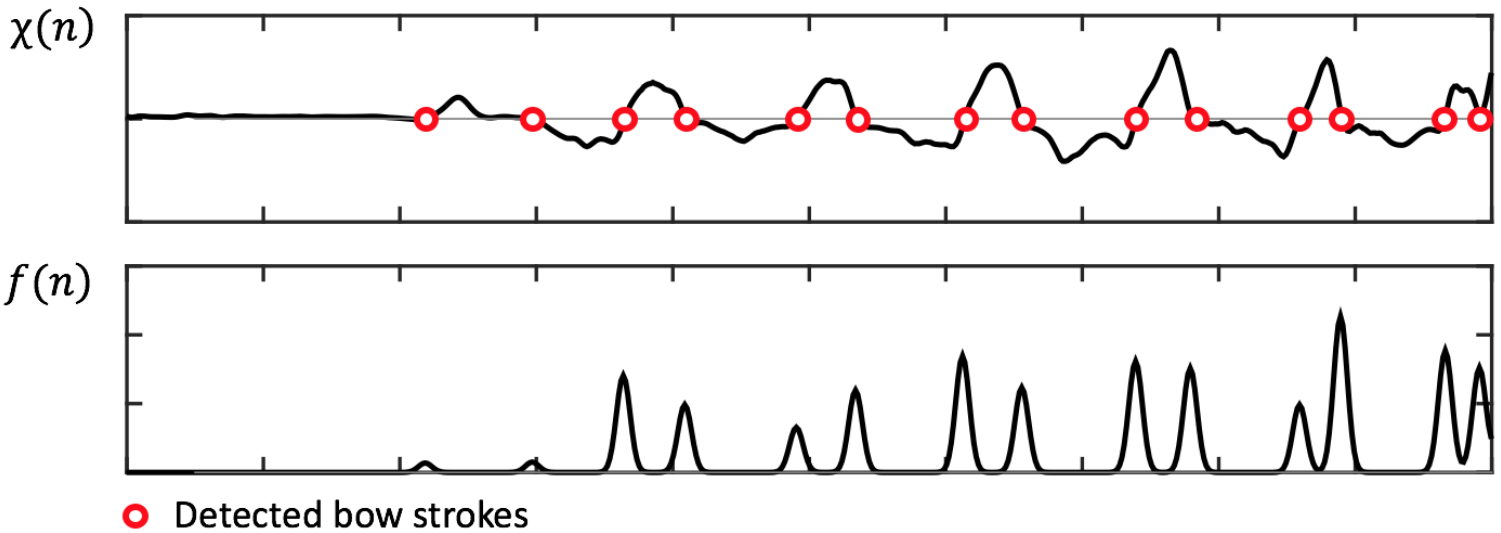


Method: Association Optimization

Pair-wise Matching

Bow Onset Likelihood:

$$f(n) = \left(\sum_{m \in \mathcal{Z}} \bar{\chi}(m) \cdot \delta(n, m) \right) * \mathcal{N}(0, \sigma^2) \tag{3}$$

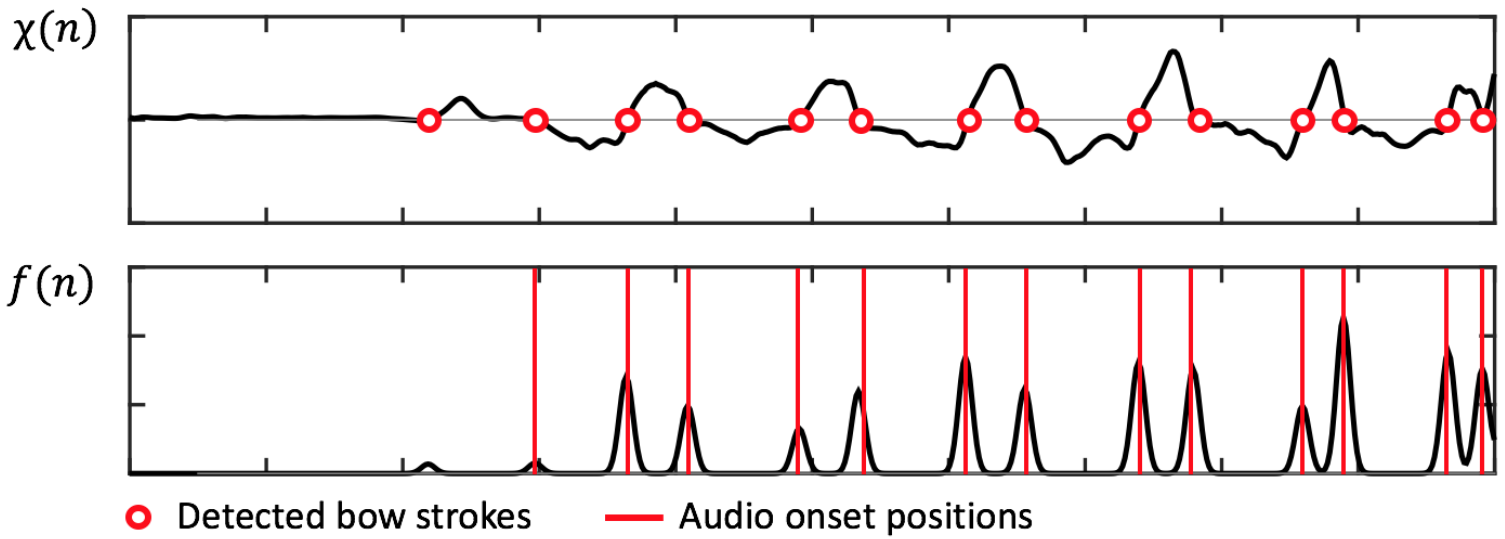


Method: Association Optimization

Pair-wise Matching

Bow Onset Likelihood:

$$f(n) = \left(\sum_{m \in \mathcal{Z}} \bar{\chi}(m) \cdot \delta(n, m) \right) * \mathcal{N}(0, \sigma^2) \tag{4}$$



Method: Association Optimization

Matching Function:

$$\begin{cases} M^{-}_{p,q} = f_p(n)^T g_q(n) / \sum_m g_q(m) \\ M^{+}_{p,q} = f_p(n)^T g_q(n) / \sum_m f_p(m) \\ M_{p,q} = \sqrt{M^{-}_{p,q} \cdot M^{+}_{p,q}}, \end{cases} . \tag{5}$$

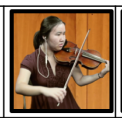
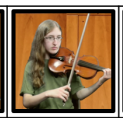
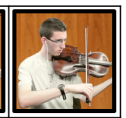
- $f_p(n) \rightarrow$ Bow onset likelihood for the p-th player.
- $g_q(n) \rightarrow$ Onset sequence for q-th track.
- $M^{-}_{p,q}$: This is low for legato bowing
- $M^{+}_{p,q}$: This is low for non-related body motion.

Method: Association Optimization

Association Score:

$$S_\sigma = \prod_{p=1}^{N} M_{p,\sigma(p)} \tag{6}$$

- For N players/tracks $\rightarrow$ N! bijections.
- $\sigma(\cdot) \rightarrow$ Permutation function.
- Select σ that maximizes S_σ.

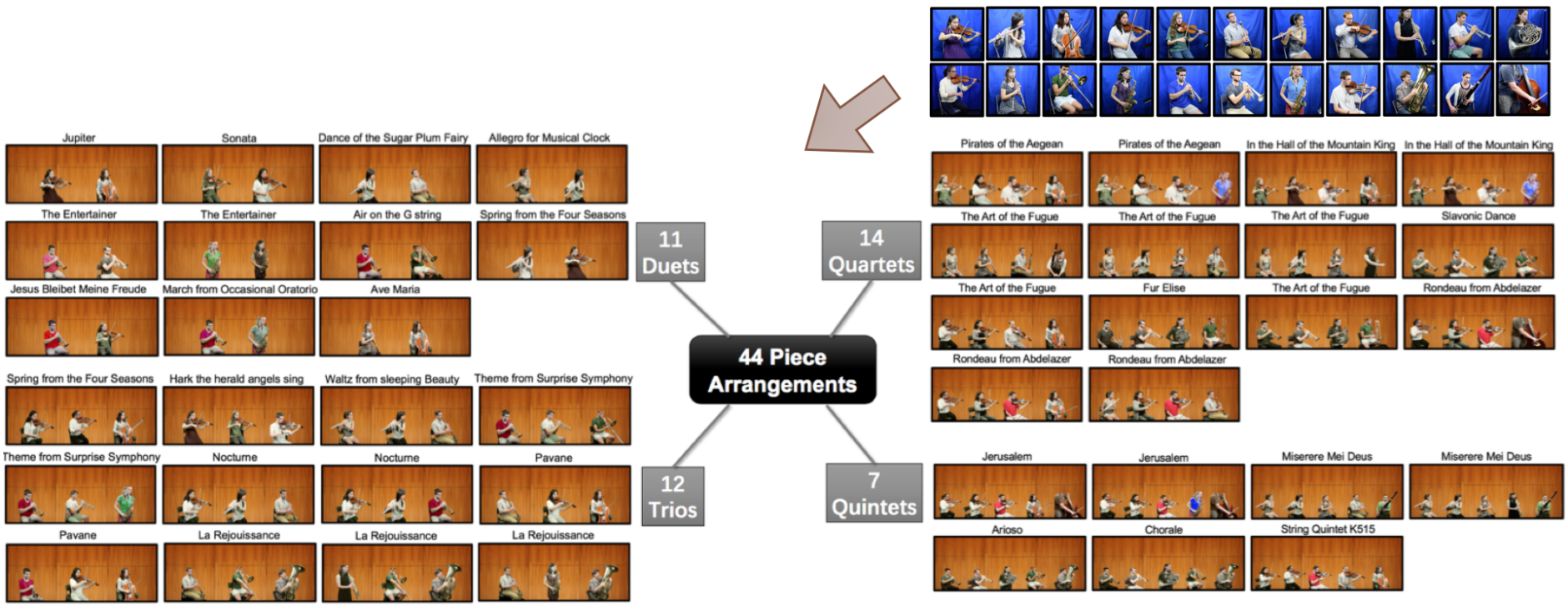
				
♫	$M_{1,1}$	$M_{2,1}$	$M_{3,1}$	$M_{4,1}$
♫	$M_{1,2}$	$M_{2,2}$	$M_{3,2}$	$M_{4,2}$
♫	$M_{1,3}$	$M_{2,3}$	$M_{3,3}$	$M_{4,3}$
♫	$M_{1,4}$	$M_{2,4}$	$M_{3,4}$	$M_{4,4}$

- Method: Audio Analysis
  - Audio-score Alignment
- Method: Video Analysis
  - Optical Flow Estimation
  - Player Detection
  - Bowing Motion Capturing
- Method: Association Optimization
- Experiments
  - Dataset
  - Evaluation Measure
  - Results

Experiments

Dataset: URMP Dataset [3]

- 14 instruments, 44 piece arrangements
- Individually recorded and assembled together



[3] B. Li *, X. Liu *, K. Dinesh, Z. Duan, and G. Sharma, Creating a musical performance dataset for multimodal music analysis: Challenges, insights, and applications, IEEE Trans. Multimedia, under review. (* Equal Contribution)

Experiments

Piece Selection

- We select 19 pieces $\rightarrow$ 5 duets, 4 trios, 7 quartets, 3 quintets
- Selection criteria: contains at most 1 non-string instrument

Overall Results

- Piece-level success rate: 89.5% (17 of 19 pieces)

** All sources within one piece should be correctly associated*

- Source-level success rate: 89.2% (58 of 65 sources)

Piece-wise Evaluation Measure

- Association Rank: the association score rank of the ground-truth association
- Metric Ratio: the ratio between the association score of the ground-truth association and the highest competitive one

Experiments

Piece-wise Results

Metadata				Association Measures		
No.	Instrument Type	Piece Length (mm:ss)	Polyphony - (No. permutations)	No. of Correctly Associated Sources	Rank of Correct Association	Metric Ratio
1	Vn. Vc.	01:03	2 - (2)	2	1	1.454
2	Vn1. Vn2.	00:46	2 - (2)	2	1	1.689
3	Fl. Vn.	00:35	2 - (2)	2	1	1.036
4	Tp. Vn.	03:19	2 - (2)	2	1	3.203
5	Ob. Vc.	01:44	2 - (2)	2	1	2.519
6	Vn1. Vn2. Vc.	02:12	3 - (6)	3	1	1.821
7	Vn1. Vn2. Va.	00:47	3 - (6)	3	1	1.048
8	Cl. Vn. Vc.	02:13	3 - (6)	3	1	1.247
9	Tp. Vn. Vc.	02:13	3 - (6)	3	1	1.289
10	Vn1. Vn2. Va. Vc.	00:50	4 - (24)	4	1	1.470
11	Vn1. Vn2. Va. Sax.	00:50	4 - (24)	4	1	1.142
12	Vn1. Vn2. Va. Vc.	01:25	4 - (24)	4	1	1.138
13	Vn1. Vn2. Va. Sax.	01:25	4 - (24)	2	5	0.769
14	Vn1. Vn2. Va. Vc.	02:54	4 - (24)	4	1	9.106
15	Vn1. Vn2. Va. D.B.	02:08	4 - (24)	4	1	1.330
16	Vn1. Vn2. Va. Vc.	02:08	4 - (24)	4	1	1.281
17	Vn1. Vn2. Va. Vc. D.B.	01:59	5 - (120)	5	1	1.438
18	Vn2. Vn2. Va. Sax. D.B.	01:59	5 - (120)	5	1	1.135
19	Vn1. Vn2. Va1. Va2. Vc.	03:45	5 - (120)	0	19	0.564

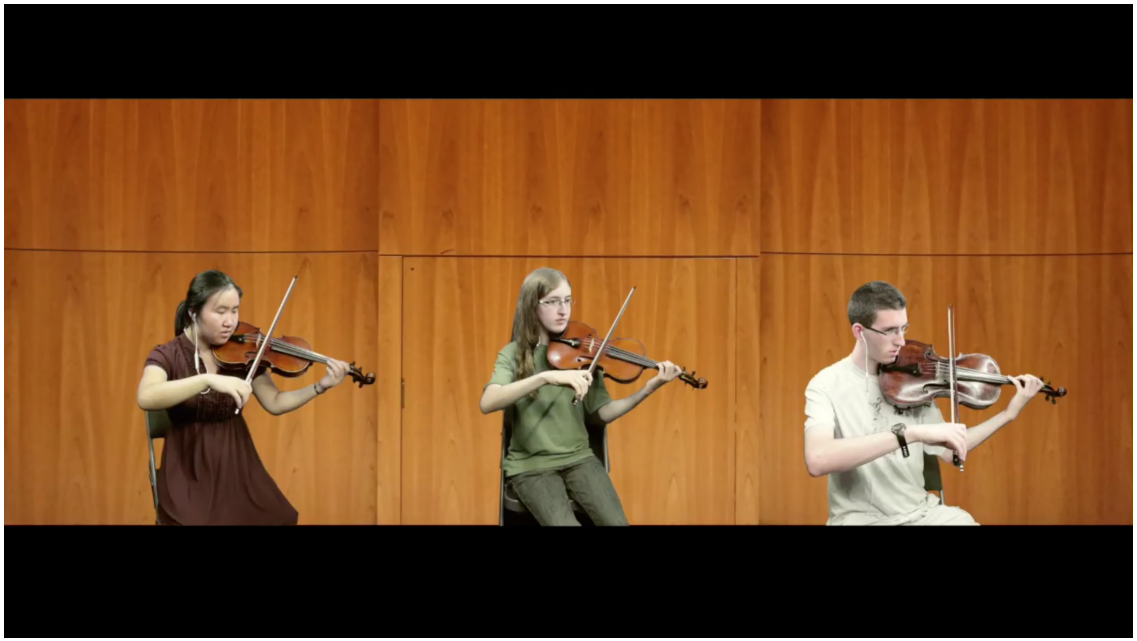
Failure case investigations:

- Non-string instrument $\rightarrow$ Motions not correlated
- Legato bowing $\rightarrow$ Audio onsets not correlated
- Same rhythmic patterns $\rightarrow$ Difficult to identify

Video

- Legato bowing

- Same rhythmic patterns



Conclusion

- Methodology for audio-visual source association
- High success rate
- Richer music enjoyment experiences