Transactions of the International Society for Music Information Retrieval

Li, B., Dinesh, K., Xu, C., Sharma, G., and Duan, Z. (2019). Online Audio-Visual Source Association for Chamber Music Performances, *Transactions of the International Society for Music Information Retrieval*, V(N), pp. xx–xx, DOI: https://doi.org/xx.xxxx/xxxx.xx

ARTICLE TYPE

Online Audio-Visual Source Association for Chamber Music Performances

Bochen Li[*], Karthik Dinesh[*], Chenliang Xu[†], Gaurav Sharma[*†], and Zhiyan Duan[*†]

Abstract

In audio-visual recordings of music performances, visual cues from instrument players exhibit good temporal correspondence with the audio signals and the music content. These correspondences provide useful information for estimating source associations, i.e., for identifying the affiliation between players and sound sources or score tracks. In this paper, we propose a computational system that models audio-visual correspondences to achieve source association for Western chamber music ensembles including strings, woodwind, and brass instruments. Through its three modules, the system models three typical types of correspondences between 1) body motions (e.g., bowing for string instruments and sliding for trombone) and note onsets, 2) finger motions (e.g., fingering for most woodwind/brass instruments) and note onsets, and 3) vibrato hand motions (e.g., fingering hand rolling for string instruments) with pitch fluctuations. Although the three modules are designed for estimating associations for different instruments, the overall system provides a universal framework for all common melodic instruments in Western chamber ensembles. The framework automatically and adaptively integrates the three modules, without requiring prior knowledge of the instrument types. The system operates in an online fashion, i.e., associations are updated as the audio-visual stream progresses. We evaluate the system on ensembles with different instruments and polyphony, ranging from duets to quintets. Results demonstrate that association accuracy increases as the duration of video excerpts increases. For string quintets, the accuracy is over 90% from just a 5-second video excerpt, while for woodwind, brass, and mixed-instrument quintets, a similar accuracy can be reached after processing 30 seconds of video. The result of the proposed framework is promising and enables novel applications such as interactive audio-visual music editing and auto-whirling camera in concerts.

Keywords: Source association, audio-visual analysis, music performance, motion analysis

1. Introduction

Visual aspects of music performances are often important. In live concerts, performers use various kinds of body movements to express their emotions and to impress audiences (Parncutt and McPherson, 2002; Sörgjerd, 2000). In music ensembles, visual interactions among musicians are important for coordination of timing and dynamics. In pop music, creative visual performances give artists a substantial competitive advantage. The inclusion of videos in music albums is shown to provide an eight-percent boost, on average, in purchase intent and improved perception (measured by Nielsen Holdings[1]). Even in prestigious classical music performances, research has shown that body movements and facial expressions of performers exert strong influences on the judgment of performance quality, for expert or novice audiences alike (Tsay, 2014).

On the technical side, the rapid expansion of digital storage and Internet bandwidth in the past decades has not only popularized video streaming services like YouTube but also significantly changed the way people

[*]Department of Electrical and Computer Engineering, University of Rochester, NY, USA.

[†]Department of Computer Science, University of Rochester, NY, USA.

[1] www.nielsen.com

Transactions of the International Society for Music Information Retrieval

Li, B., Dinesh, K., Xu, C., Sharma, G., and Duan, Z. (2019). Online Audio-Visual Source Association for Chamber Music Performances, *Transactions of the International Society for Music Information Retrieval*, V(N), pp. xx–xx, DOI: https://doi.org/xx.xxxx/xxxx.xx

ARTICLE TYPE

Online Audio-Visual Source Association for Chamber Music Performances

Bochen Li[*], Karthik Dinesh[*], Chenliang Xu[†], Gaurav Sharma[*†], and Zhiyan Duan[*†]

Abstract

In audio-visual recordings of music performances, visual cues from instrument players exhibit good temporal correspondence with the audio signals and the music content. These correspondences provide useful information for estimating source associations, i.e., for identifying the affiliation between players and sound sources or score tracks. In this paper, we propose a computational system that models audio-visual correspondences to achieve source association for Western chamber music ensembles including strings, woodwind, and brass instruments. Through its three modules, the system models three typical types of correspondences between 1) body motions (e.g., bowing for string instruments and sliding for trombone) and note onsets, 2) finger motions (e.g., fingering for most woodwind/brass instruments) and note onsets, and 3) vibrato hand motions (e.g., fingering hand rolling for string instruments) with pitch fluctuations. Although the three modules are designed for estimating associations for different instruments, the overall system provides a universal framework for all common melodic instruments in Western chamber ensembles. The framework automatically and adaptively integrates the three modules, without requiring prior knowledge of the instrument types. The system operates in an online fashion, i.e., associations are updated as the audio-visual stream progresses. We evaluate the system on ensembles with different instruments and polyphony, ranging from duets to quintets. Results demonstrate that association accuracy increases as the duration of video excerpts increases. For string quintets, the accuracy is over 90% from just a 5-second video excerpt, while for woodwind, brass, and mixed-instrument quintets, a similar accuracy can be reached after processing 30 seconds of video. The result of the proposed framework is promising and enables novel applications such as interactive audio-visual music editing and auto-whirling camera in concerts.

Keywords: Source association, audio-visual analysis, music performance, motion analysis

1. Introduction

Visual aspects of music performances are often important. In live concerts, performers use various kinds of body movements to express their emotions and to impress audiences (Parncutt and McPherson, 2002; Sörgjerd, 2000). In music ensembles, visual interactions among musicians are important for coordination of timing and dynamics. In pop music, creative visual performances give artists a substantial competitive advantage. The inclusion of videos in music albums is shown to provide an eight-percent boost, on average, in purchase intent and improved perception (measured by Nielsen Holdings[1]). Even in prestigious classical music performances, research has shown that body movements and facial expressions of performers exert strong influences on the judgment of performance quality, for expert or novice audiences alike (Tsay, 2014).

On the technical side, the rapid expansion of digital storage and Internet bandwidth in the past decades has not only popularized video streaming services like YouTube but also significantly changed the way people

[*]Department of Electrical and Computer Engineering, University of Rochester, NY, USA.

[†]Department of Computer Science, University of Rochester, NY, USA.

[1] www.nielsen.com

enjoy music. With the surge of Virtual Reality (VR) and Augmented Reality (AR) technologies and their adoption in music entertainment, visual aspects of music performances will further gain importance in innovative music enjoyment experiences.

While Music Information Retrieval (MIR) based on the audio signal and symbolic score (e.g., MIDI) has been widely studied, only limited explorations have been conducted on the interplay of visual and acoustic aspects of music performances. The auditory and visual modalities are intimately related in music performances. Sounds from acoustic instruments are invariably mediated by the instrument players' movements and characteristics of the movements are reflected in the resulting sounds. For example, the amplitude envelope and spectral evolution of a violin note are directly related to the velocity and pressure of a bowing motion (Askenfelt, 1989) and fingering force (Obata et al., 2009); the timing of a clarinet note is often correlated to the fingering movements; the loudness of a drum hit is strongly related to the drumstick's preparatory height and striking velocity (Dahl, 2004). These characteristics have been utilized to solve traditional MIR problems such as multi-pitch analysis (Dinesh et al., 2017), music transcription (Paleari et al., 2008), score alignment (Bazzica et al., 2014), source separation (Parekh et al., 2017b), etc. An overview of related literature is available in (Duan et al., 2019).

Classical chamber music is performed by a small ensemble of instrumentalists, with *one player per score track* (Burkholder and Grout, 2014). In this paper, we study the relationship between the instrument players' body movements and sound events in classical chamber ensemble performances. The aim is to solve the *source association* problem, i.e., identifying the bijection between score tracks (MIDI or MusicXML format) and players in the video. The bijection, together with a score-informed audio source separation technique (Ewert et al., 2014), can allow users to separate the audio source for each particular player in the video.

Exploiting information in the video about instrument players' movements for source association is challenging because many body movements (e.g., head movement) are irrelevant to sound articulation (Godøy and Jensenius, 2009) and relevant movements (e.g., maneuver with fingers) can be subtle. In music ensembles, similar body movements can be observed among different musicians when they have similar rhythmic patterns. These challenges are especially pronounced when the video clip is short (e.g., from online streams) and when the ensemble is large. For a quintet, possible associations can be enumerated as 120 permutations, but only one is correct.

Source association enables novel research and applications. It is essential for leveraging the visual information to analyze individual sound sources in music performances. The related techniques include multi-pitch analysis (Dinesh et al., 2017), performance expressiveness analysis (Li et al., 2017b), source separation (Parekh et al., 2017a), etc. By exploiting source associations, one can envision an augmented video streaming service that allows users to click on a player in the video and isolate/enhance the corresponding source of the audio (Zhao et al., 2018). Based on SLAVE (Thomas et al., 2009), a music exploration system that manages multimedia music collections, one can envision an augmented sheet music display interface where on each score track, the visual performance of the specific player is retrieved and demonstrated. For music production, source association can help enable remixing of audio sources along with automatic video scene recomposition. An online source association system, which does not need to "look into the future", can be further useful in online video streaming of live concerts. For example, it enables an auto-whirling camcorder to focus on the soloist.

In this paper, we build upon our previous work on source association for string instruments using bowing motions (Li et al., 2017a) and vibrato motions (Li et al., 2017c), and propose the first universal system to address the problem for common melodic instruments in Western chamber ensembles such as string, woodwind, and brass instruments (barring polyphonic instruments such as piano and harp). This system does not require prior knowledge of instrumentation of the piece or pre-training of audio-visual correspondence. The system input is the audio/video performance and the corresponding music score as pianoroll representations, and the output is the association between audio/score tracks and video players, assuming audio and video tracks are synchronized and audio and score tracks are associated. After temporally aligning the score with the live performance from auditory cues, the system uses three modules to analyze different visual motion types that may be present in the performance, as shown in Figure 1. Because many performed motions are related to note onsets, the first two modules focus on the motion-onset correspondence. The first module extracts large-scale body motions, which mainly capture bowing motions of string instruments. The second module extracts subtle fingering motions and correlates these with note onsets. The correlation aids associations for woodwind/brass instruments, as pitch changes are mostly controlled by finger-operated keys. In addition to note onsets, variations of acoustic features throughout tone articulations also show correspondence with certain motions, for instance, for the vibrato articulation in string instruments. Therefore, the third module is designed to detect periodic fingering motions (if any) and to correlate them with the periodic pitch fluctuation estimated from audio. This module is primarily directed at string instruments, where vibrato articulations can be characterized from the visual modality. Note that the first and third mod-

ules are adapted from previously proposed systems by (Li et al., 2017a) and (Li et al., 2017c) respectively, and the second module is proposed in this paper as the first solution for wind/brass instruments. Finally, we also propose to integrate the output of the three modules through weighted voting according to motion salience. It is noted that the system does not need to detect the instrument type; it simply extracts the three kinds of motions (if any) for each player and integrates their correspondence with score/audio tracks, jointly.

The proposed system works in an online fashion: The audio-score alignment, the correlation between motion and audio/score, and the association output are all updated in a frame-by-frame fashion without "looking into the future". Associations in each frame are updated using the Hungarian algorithm (Kuhn, 1955), with a minimum computational cost. Experiments on 17,574 audio-visual clips generated from 44 chamber music pieces in the URMP dataset (Li et al., 2019) that spans a polyphony range from duets to quintets, show that: 1) Different modules are helpful for different instruments, and the system is able to integrate them automatically to achieve a high overall accuracy; 2) Accuracy increases as longer video streams are available, reaching an average accuracy of 90% for 5-second video excerpts of string instruments, and for 30-second excerpts of woodwind and brass instruments. In summary, the proposed system for audio-visual source association:

- works universally for all instruments common in Western chamber ensemble performances,
- does not require prior knowledge of instrumentation, and
- relies purely on motion information for association without modeling instrument characteristics; which allows it to also work for ensembles of the same instrument type, e.g., violin duets.

In the following, we first review existing work on multi-modal modeling in Section 2, and highlight challenges involved in source association in music performances. We then describe our proposed method in three modules for the different motion cues for associations in Section 3. In Section 4, we conduct systematic experiments to evaluate the proposed system. Finally, we conclude the paper in Section 5.

2. Related Work

2.1 Source Localization

When there is at most one active sound source at a time, the problem of audio-visual source association is also known as *source localization*, i.e., indicating the location of the sound source in the video. For audio-visual speech, source localization is helpful for speaker face segmentation (Liu and Sato, 2008). Early work on speaker localization correlates audio energy changes with pixel motions via non-linear diffusion (Casanovas and Vandergheynst, 2010) or with semantic regions via video segmentation and tracking (Li et al., 2014). Other methods include time-delayed neural networks (Cutler and Davis, 2000), probabilistic multi-modal generative models (Fisher and Darrell, 2004), and Canonical Correlation Analysis (CCA) (Kidron et al., 2007; Izadinia et al., 2013).

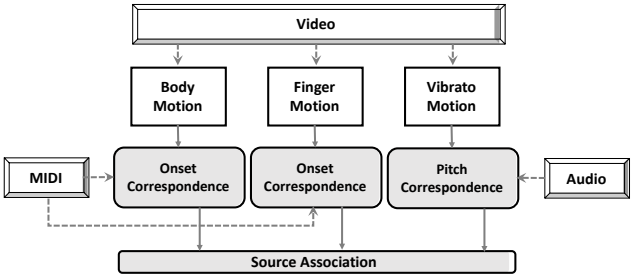

Figure 1: Outline of the proposed universal source association system for chamber ensemble performances. Three types of motion are modeled and correlated with the audio and score in three modules.

More recent work proposes to localize semantic objects in unconstrained videos by learning deep multi-modal representations. Owens and Efros (2018) propose a fused multi-sensory network to learn an audio-visual representation, which further localizes the sound objects on the video frames. Senocak et al. (2018) employ a similar two-stream network structure, where an attention mechanism is developed for sound source localization. A similar idea is adopted by Arandjelović and Zisserman (2018) for cross-modal retrieval and source localization, and by Tian et al. (2018) for both spatial and temporal localization.

2.2 Source Association for Separation

Other work deals with mixtures of active sources, where cross-modal association can be applied to isolate sounds that correspond to each visual object. Barzelay and Schechner (2007, 2010) detect drastic changes (i.e., onsets of events) in audio and video and then use their coincidence to associate audio-visual components that belong to the same source of harmonic sounds. Sigg et al. (2007) reformulate CCA by introducing non-negativity and sparsity constraints on the coefficients of the projection directions to locate and separate sound sources in movies. In (Casanovas et al., 2010), auditory and visual modalities are decomposed into relevant structures using redundant representations for source localization. Segments, where only one source is active, are used to learn a timbre model for the separation of the source. Ephrat et al. (2018) propose a deep network-based model to isolate single speech signals from a mixture of sounds given the target speaker from the video. Gao et al. (2018) map audio frequency bases to individual visual objects via an audio-visual object model, which further guides audio source separation. Most of these methods, however, ei-

ther deal with mixtures with at most two active sources or only focus on isolating one source from multiple active sources (e.g., background noises). The association problem for each individual source is not addressed.

2.3 Source Association for Chamber Ensembles

The source association problem for music ensembles is more challenging since all the available sound sources (the players) are active almost all the time, and the difficulty increases dramatically as the number of sources increases. Although each track is performed by one player in chamber music, the same kind of instruments are often used for different score tracks (e.g., a violin duet). Therefore, approaches aiming at learning a deep representation that maps audio features with visual appearances to localize each source (Owens and Efros, 2018; Senocak et al., 2018; Arandjelović and Zisserman, 2018) are not applicable. Instead, one needs to recognize the distinct motions of different players and correlate them with the music content to achieve association.

Bazzica et al. (2014) first propose to detect play/non-play conditions for each player in an orchestra, which are compared with each score track to solve the temporal alignment. In our previous work (Li et al., 2017a), we propose an approach to solving the association problem for string ensembles with up to five simultaneously active sources in a score-informed fashion. The approach analyzes the bowing motion and correlates it with note onsets in score tracks. The assumptions are that many note onsets correspond to the beginning of bowing strokes and that different instrumental parts often have different rhythmic patterns. When these assumptions are invalid, for example, when multiple notes are played within a single bow stroke (i.e., legato bowing) or when different parts show a similar rhythmic pattern, the approach becomes less robust. Later we propose a complementary approach (Li et al., 2017c) which correlates the fingering hand rolling motion with pitch fluctuations of vibrato notes for the association of string instruments. However, the method only works when vibrato notes are played. To our best knowledge, there is neither an existing work on integrating the bowing motion and vibrato motion for source association for string instruments, nor any extensions of the concept to deal with non-string instruments.

3. Method

The proposed system takes data in three modalities as the input: the audio recordings, the video recordings, and the music scores of the chamber music performances. As illustrated in Figure 1, the system uses three parallel modules to model three types of temporal correspondence between motions detected in the video and note events captured in other modalities for different instrumentalists. In this section, we present the system in detail.

3.1 Performance-Score Alignment

As the proposed approach is score informed, a preliminary step for the system is to temporally align the music score with the dynamic timing of the audio-visual ensemble performance (assuming audio and video are pre-synchronized). The temporal alignment is achieved through audio-score alignment on the harmonic content (Müller, 2015). To do so, the audio is first converted to short-time Fourier spectral magnitudes with a 42.7 ms frame length (2048 samples for a 48 kHz sampling rate), 10 ms hop size, Hamming window, and zero padding to produce 4 times the original length. The short-time Fourier spectral magnitudes are then mapped to 12-D chroma vectors, where each element represents a pitch class. Each chroma vector is normalized by its root mean square (RMS) value. A similar operation is applied to the score, which is segmented into non-overlapping frames of the same duration using the default tempo notated in the score. A 12-D binary chroma vector is calculated for each frame to indicate the presence (taking more than 50% of the frame) and absence of a pitch class. The chroma vector is then normalized by its RMS.

In offline scenarios where the entire performance is available beforehand, the alignment can be obtained by the dynamic time warping (DTW) algorithm (Müller, 2007), which is robust and efficient (Müller et al., 2006). In online scenarios where the performance data arrives as a live stream, one commonly used framework is an online DTW algorithm (Dixon, 2005), which provides options such as "forward-backward strategy" to reconsider the past decisions (Arzt et al., 2008), or a step to incorporate a tempo model (Arzt and Widmer, 2010) for robustness. An alternative framework employs a stochastic model (Grubb and Dannenberg, 1997; Duan and Pardo, 2011b), where the score position hypotheses are represented by a probability density function. In this paper, to deal with online video streaming scenarios, we apply the online method proposed by Duan and Pardo (2011b), which is based on a Hidden Markov Model with a 2D continuous state space to represent the score position and tempo. This framework is previously evaluated on the Bach10 dataset (Duan et al., 2010) showing decent results. Further qualitative check guarantees a good alignment performance on the URMP dataset used in our experiments.

3.2 Onset Correspondence with Body Motion

3.2.1 Body Motion Extraction

In music performances, body motion of performers conveys important musical expressions and ideas, e.g., the head nodding at leading notes. For some instruments, body motion directly articulates notes (e.g., strings, drums) or controls the pitch (e.g., trombones). To capture body motion from video recordings, one approach is optical flow estimation. In our previous

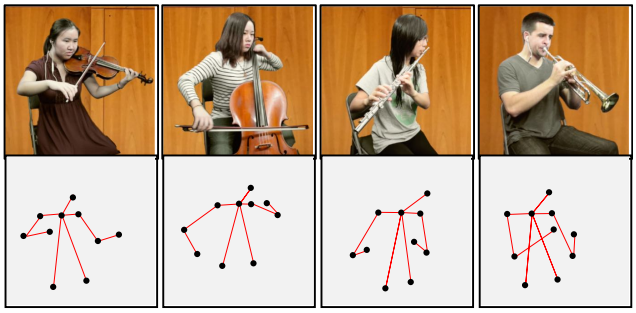

Figure 2: Body motion extraction. Upper body skeletons (second row) are extracted with OpenPose (Cao et al., 2017) in each video frame (first row) followed by temporal smoothing over time.

approach Li et al. (2017a) we apply optical flow estimation to extract bowing motion of string players. However, we argue that this pixel-level analysis may not be ideal for semantic-level understanding of body gestures and movements, and can be less robust to occlusions and camera viewpoint changes.

In this paper, we propose to apply OpenPose (Cao et al., 2017) on each frame, a multi-person pose estimation approach to extract body skeleton coordinates for all the players on stage without pre-segmentation on the video recording. A skeleton in each frame is represented as a 20-D vector $\mathbf{y}(t)$ corresponding to the horizontal and vertical coordinates of the 10 upper body joints, including nose, neck, shoulders, elbows, wrists, and hips. We do not include lower body joints as they are often less relevant to note events. Figure 2 shows video frames of several instrumentalists with the extracted body skeletons. To form a continuous skeleton sequence across time, we eliminate joint coordinates if the confidence score from OpenPose is smaller than 0.2 *and* the L_2 distance between a joint in consecutive frames is larger than 10% of the head-hip distance, which is considered the maximal regular movement in a ≈30-FPS video without shot transition. We also temporally smooth their coordinates using a moving average with a 5-frame window size. These post-processings are referred from (Li et al., 2018b) where the same approach is applied to extract skeletons for pianists. We then take the two hips as reference coordinates to align the body position across frames. Finally, we calculate motion velocities $\mathbf{z}(t)$ as the derivative of $\mathbf{y}(t)$ w.r.t. time. Compared to optical flow estimation, this gesture-based motion analysis approach is semantically more meaningful, less computationally expensive, and more robust to occlusions and camera viewpoint changes such as camera zooming or panning.

To extract motions related to note onsets in each video frame, for each player we denote the motion velocities of n frames in the past as $\mathbf{Z} = [\mathbf{z}_1(t), \mathbf{z}_2(t), \cdots, \mathbf{z}_n(t)]^T \in \mathbb{R}^{n\times 20}$ and apply principal component analysis (PCA) by eigen value decomposition $\mathbf{Z}^T\mathbf{Z} = \mathbf{V}\Sigma\mathbf{V}^T$, where $\mathbf{V}$ and Σ represent the matrix of

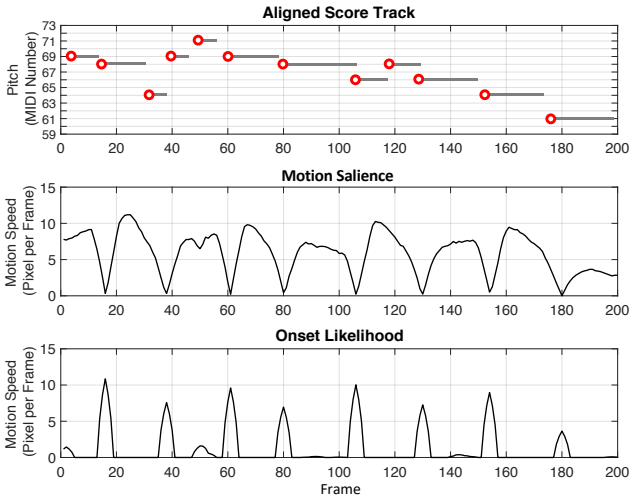

Figure 3: Example correspondence between body motion and note onsets. Top: temporally aligned score track with onsets marked by red circles. Middle: extracted motion salience (primarily bowing motion) from the visual performance of a violin player. Bottom: derived onset likelihood curve from the motion salience.

eigen vectors and the corresponding eigen values respectively. We then project the motion velocity $\mathbf{z}(t)$ onto the principal component direction (first column of $\mathbf{V}$) and take its absolute value as the motion salience $s(t)$. Choosing the salient motion discards the direction information of the motions (e.g., up/down-bow for violinists), which is less relevant to timings than the amplitude information. We set n to 150 frames, i.e., 5 seconds in time, assuming a player's pose stays consistent in this range. To reduce the computational cost, we update $\mathbf{V}$ every 1 second (assuming consistent motion patterns within a short period).

3.2.2 *Onset Likelihood*

From the motion salience $s(t)$, we infer the timings of the motion strokes that are potentially related to score note onsets. As a note onset often corresponds to the beginning or ending of a sound articulation motion (e.g., a bowing stroke for string instruments), the motion speed at the onset is often small. Therefore, local minima of the motion salience $s(t)$ is often indicative of note onsets. Let Ω be the set of all the local minima throughout a piece. For each local minimum $\tau \in \Omega$, we represent the likelihood of a note onset as $a(\tau) = \max_{\gamma\in[\tau,\tau+30]} s(\gamma) - s(\tau)$ that is determined by the maximum speed of the motion stroke considering the following 30 frames: the larger the value of $a(\tau)$ the more likely that a note onset is activated by the motion stroke. Here 30 frames are considered to span the high energy part of most notes. Therefore, we can define an *onset likelihood* curve $\phi_b(t)$ derived from body motion

analysis as

$$\phi_b(t) = \left(\sum_{\tau \in \Omega} a(\tau) \cdot \delta(t-\tau) \right) * \mathcal{N}(t), \quad (1)$$

where $\delta(t)$ is the Dirac delta function, $*$ is the convolution operation, and $\mathcal{N}(t)$ is a Gaussian function to give each predicted onset time a tolerance (width) with a standard deviation of 3 frames (30 ms) (considering some slight non-synchronization between different modalities in the recording file). It is noted that $\phi_b(t)$ can be calculated in an online fashion, with a delay of up to 1 second due to the search for the local maximum after each local minimum. Figure 3 plots the onset likelihood curve $\phi_b(t)$ along with the associated and temporally aligned score track as piano-roll, where the note onset timings are marked as red circles. We find that many of the note onsets can be associated with peaks of $\phi_b(t)$. The correspondence between the notes and peaks sets the basis for the association between score and motion, as described below.

3.2.3 Pair-wise Correspondence

We extract the motion-based onset likelihood curve for each player from the video performance as $\phi_b^{[p]}(t)$, where p is the player index. From each track of the temporally aligned score, we use a binary impulse train $\psi^{[q]}(t)$ to represent the note onsets, where q is the track index, $\psi^{[q]}(t) = 1$ if there is a note onset in the t-th frame of the q-th track and $\psi^{[q]}(t) = 0$ otherwise. Then the pair-wise *matching score* between the p-th player and the q-th score track, up to the t-th frame, can be calculated through inner product:

$$M_b^{[p,q]}(t) = \sum_{\tau=0}^{t} \phi_b^{[p]}(\tau) \cdot \psi^{[q]}(\tau). \quad (2)$$

This can be updated in an online fashion as new temporal frames arrive.

3.3 Onset Correspondence with Finger Motion

3.3.1 Finger Motion Extraction

While note articulation is visible on body movements for string instrumentalists, this is generally not the case for woodwind/brass instrumentalists, where notes are articulated by blowing to the reed/mouthpiece, showing a less visible motion around the mouth. However, pitch changes of these instruments are mostly controlled by finger-operated keys, which often result in synchronized events between finger movements and note onsets (Palmer et al., 2007). Compared to body motions, finger motions are more subtle and more prone to occlusion. In this section, we propose to extract finger motions and correlate them with note onsets.

We apply OpenPose again to extract the positions of all the finger joints from each player. Due to the limited video resolution and occlusion, the result is not robust enough to estimate the motion. Inspired by our previous work (Li et al., 2017c), we use optical flow estimation (Sun et al., 2010) to capture this subtle motion at the pixel level. To reduce the computational cost, we set a region of interest (ROI) around the detected finger joints from OpenPose for optical flow estimation. The ROI centers at the median of all the finger joints for each hand, and spans to cover all the joints. Similar to body skeletons, we smooth the joint coordinates using moving average filter with a window size of 5 frames. Then we compute the optical flow estimation inside the ROI. Again, to eliminate the rigid and affine motion, each optical flow vector is subtracted by the average motion vector of the ROI, resulting in a motion vector $\mathbf{u}^{(ij)}(t)$ at the pixel (i, j) and t-th frame. Figure 4 takes one flute player and one clarinet player as examples to visualize the optical flow estimation of one-hand finger motion in five consecutive frames, where the estimated finger joint positions are overlaid on the first video frames.

3.3.2 Onset Likelihood

On each frame we take the maximum value of pixel-wise motion magnitude $|\mathbf{u}^{(ij)}(t)|$ across all the pixels in the ROI as the motion flux, which captures the finger movements corresponding to pitch changes and is directly considered as *onset likelihood* $\phi_f(t)$ from finger motions. Figure 5 plots the onset likelihood curve $\phi_f(t)$ along with the associated and temporally aligned score track as piano-roll. We can observe salient motion flux around most note onset frames. Compared to Figure 3, the correspondence of note onsets to fingering motions for woodwind/brass instruments is not as robust as that to body motions for string instruments. The observation can be attributed to the fact that fine-grained motion is more sensitive to irrelevant motions. In addition, repeated notes for most woodwind/brass instruments are not reflected by finger maneuver on the keys.

Analogous to Eq. (2), the pair-wise matching score from finger motions can be calculated as:

$$M_f^{[p,q]}(t) = \sum_{\tau=0}^{t} \phi_f^{[p]}(\tau) \cdot \psi^{[q]}(\tau). \quad (3)$$

3.4 Pitch Correspondence with Vibrato Motion

In addition to the onset time, variations of acoustic features throughout the entire process of some note articulations show correspondence with specific motions. Vibrato is one such feature. Vibrato is a commonly used artistic note articulation method to color a tone and express emotions in music performances. Physically, vibrato is generated by pitch modulation of a note in a periodic fashion. For some instruments such as strings, vibrato is often visible as the left hand rolling motion on the fingerboard. The relationship between visible motion and vibrato motivates us to follow our previous

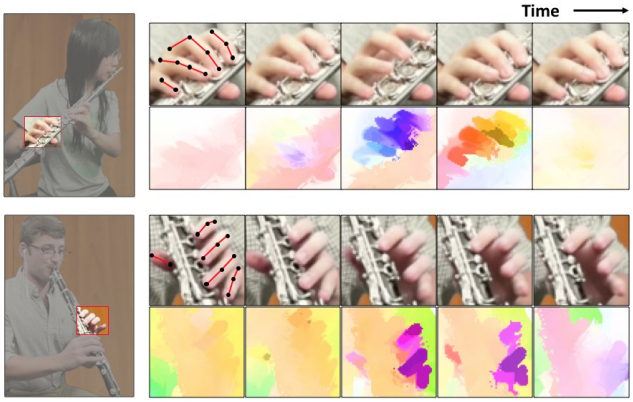

Figure 4: Optical flow visualization of finger motions in five consecutive frames corresponding to note changes. The color encoding scheme is adopted from Baker et al. (2011).

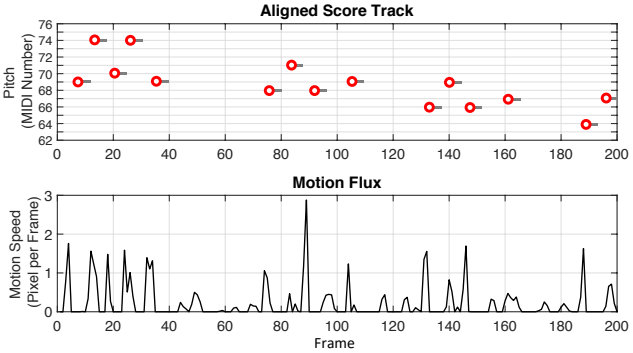

Figure 5: Example correspondence between finger motion and note onsets of a flute player. Top: temporally aligned score track with onsets marked by red circles. Bottom: extracted motion flux from finger movements.

work (Li et al., 2017c) to extract the fine motion and find the correspondence with pitch contours extracted from the audio modality.

3.4.1 *Vibrato Motion Extraction*

We retrieve the fingering motion $\mathbf{u}^{(ij)}(t)$ as computed from the previous section. Although the vibrato motion is mostly a rigid motion (fingers move together with little relative movements), it is periodic and very fast (usually about 4-7.5 Hz (Geringer et al., 2010)), and hence it is not removed as other slow rigid/affine motions. Figure 6 illustrates several frames of the optical flow estimation of the vibrato hand motions from the two players. For each frame t, we take the average motion vector across all pixels within the ROI as $\mathbf{u}(t) = [u_x(t), u_y(t)]^T$, where the motion direction is preserved for vibrato detection.

The vibrato detection module works as a binary classifier as proposed and trained by Li et al. (2017b). The classifier is implemented as a support vector machine (SVM) that takes the input of a 8-dimensional feature extracted from each sample, including the zero

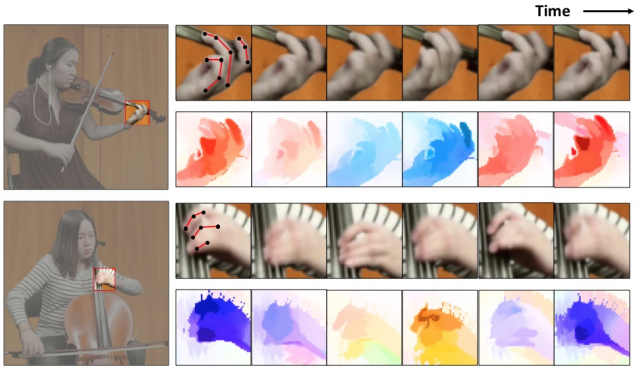

Figure 6: Optical flow visualization of the hand motions corresponding to vibrato articulation. Color encoding scheme is adopted from Baker et al. (2011).

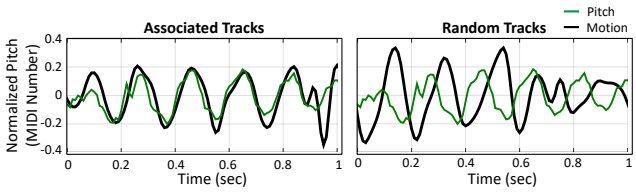

Figure 7: The same segment of normalized pitch contour $f(t)$ (green) overlaid with the motion displacement curve $d(t)$ (black) from the associated track (left) and another random track (right).

crossing rate of the x- and y- motion velocities and their auto-correlations, the energy of 3-9 Hz frequency range, and the auto-correlation peaks. According to Li et al. (2017b), this method achieves a vibrato detection accuracy of over 90% from visual motions regardless of the polyphony number and instrument type within the string instrument family. Here each input sample is a 1-second segment of $\mathbf{u}(t)$ (again introducing an average 0.5-second delay of the association system).

For each detected vibrato segment, we perform PCA on $\mathbf{u}(t)$ within this 1-second segment to obtain the 1-D principal motion velocity curve $v(t)$. We then integrate $v(t)$ over time to calculate a *motion displacement curve,* $d(t)$, which corresponds to the length fluctuation of the vibrating string, and hence the pitch fluctuation of the note. We normalize each vibrato segment of $d(t)$ to zero mean and unit variance. We set the non-vibrato segments of $d(t)$ to zero.

3.4.2 *Pitch Contour Extraction*

Utilizing the score information, we apply Soundprism (Duan and Pardo, 2011a), an online score-informed source separation system, to separate the polyphonic audio mixture into individual sources. Note that although audio recordings of individual instrumental tracks are available in the dataset, we do not use them as they are not generally available in real concert scenarios. To extract the pitch contour, we perform a score-informed pitch estimation step on each separated

audio source, as described by our previous work (Li et al., 2017c). The pitch contour of each note segment is normalized to have zero mean and unit variance, and is denoted as $f(t)$. The normalization operation discards the original pitch height information, and only preserves the pitch drift from the central frequency within each note. Figure 7 plots a 1-second segment of the normalized pitch contour overlaid with a motion displacement curve from the associated track (left) and a random track (right). Similar to Eqs. (2) and (3), we calculate the vibrato correspondence as:

$$M_v^{[p,q]}(t) = \sum_{\tau=0}^{t} d^{[p]}(\tau) \cdot f^{[q]}(\tau). \quad (4)$$

3.5 Integrating All Correspondences

We integrate the three modules to calculate the pair-wise correspondence between visual motion and score/audio events considering both onset timing and the entire note articulation process. The calculation is presented as

$$\begin{aligned} M^{[p,q]}(t) &= w_b(t) \cdot \bar{M}_b^{[p,q]}(t) \\ &+ w_f(t) \cdot \bar{M}_f^{[p,q]}(t) \\ &+ w_v(t) \cdot \bar{M}_v^{[p,q]}(t), \end{aligned} \quad (5)$$

where $\bar{M}_b^{[p,q]}(t)$, $\bar{M}_f^{[p,q]}(t)$, and $\bar{M}_v^{[p,q]}(t)$ represent the normalized correspondence across all of the pair-wise combinations between N players and N tracks as:

$$\bar{M}_i^{[p,q]}(t) = \frac{M_i^{[p,q]}(t)}{\sum_{p',q'=1}^{N} M_i^{[p',q']}(t)}, \quad i \in \{b, f, v\}, \quad (6)$$

and w_b, w_f, w_v represent the weighting parameters to re-scale the normalized correspondences from different modules. Weight w_v is set as $2w_f$, to place greater emphasis on finger motions with vibrato patterns. Weights w_b and w_f are linearly related to their motion salience/flux in the past frames as

$$\frac{w_b(t)}{w_f(t)} = \frac{\sum_{\tau=0}^{t} s(\tau)}{\sum_{\tau=0}^{t} \phi_f(\tau)}, \quad (7)$$

The linear relationship recovers the original scale of body and finger motion to weigh the correspondences $M_b(t)$ and $M_f(t)$. It allows the system to focus on the part with stronger motion cues, such as body motion for string instrumentalists, and finger motion for wind/brass instrumentalists. In Section 4, we test the components in isolation as well as some combinations of them.

For an ensemble with N players, the number of possible associations is the factorial of N. Let $\sigma(\cdot)$ be a permutation function from $p \in [1, N]$ to $q \in [1, N]$ that represents one association candidate, where the p-th player is associated with the $\sigma(p)$-th track. For each association candidate σ, we calculate an overall association score as the product of the N pair-wise correspondence values. The final association solution $\hat{\sigma}$ is returned to maximize the association score as:

$$\hat{\sigma} = \arg\max_{\sigma} \prod_{p=1}^{N} M^{[p,\sigma(p)]} = \arg\min_{\sigma} \sum_{p=1}^{N} -\log M^{[p,\sigma(p)]}. \quad (8)$$

The replacement of product with sum of negative logarithms makes the efficient Hungarian algorithm (Kuhn, 1955) directly applicable for finding the best association.

4. Experiments

4.1 Dataset

The proposed source association system is evaluated on the URMP dataset (Li et al., 2019). To our best knowledge, this is the only publicly available multi-track audio-visual music performance dataset that is suitable for our evaluations. It contains 44 classical chamber ensemble pieces ranging from duets to quintets, assembled from 149 individually recorded tracks. Each piece comes with an audio recording (48 kHz, 24 bits) of the ensemble performance along with the audio recording for each individual instrument track, an assembled video recording (1080P, 29.97 FPS) of all instrumentalists as a whole, pitch/note annotations for each track, and the corresponding MIDI file as music score. In the assembled video recording, players are arranged horizontally from left to right, with the right-front side exposed to camera. The video has a static view without camera panning/zooming or shot transitions during the whole performance. The whole dataset is accessible from (Li et al., 2018a).

We further expand the dataset by creating all possible track combinations within each piece. In the expanded set, audio is remixed from the provided individual audio tracks. For videos, we directly use the estimated pose of each player from the original video ensembles for augmented track combinations. This process gives equivalent results as if we first create the assembled videos of the augmented instrumental combinations and then run OpenPose on these assembled videos, but simply reduce computations in the experiments. For the example of a quartet, we further generate 6 duets and 4 trios from the 4 original tracks. Note that we do not combine tracks across pieces, to ensure the naturalness of the expanded set. The total expanded dataset comprises of 171 duets, 126 trios, 47 quartets, and 7 quintets. The number of pieces for different instrument arrangements are listed in Table 1.

To further understand the dataset, we calculate the *onset overlap rate* for each original piece. This statistic is defined as the percentage of onset positions that are shared by two or more tracks for each piece. This statistic is relevant to the performance of the proposed source association approach, as two out of the three motion analysis modules rely on onset patterns to as-

		String	Woodwind/Brass	Mixed	Total
Original	Duet	2	6	3	11
	Trio	2	6	4	12
	Quartet	5	6	3	14
	Quintet	2	4	1	7
Expanded	Duet	57	91	23	171
	Trio	41	65	20	126
	Quartet	15	25	7	47
	Quintet	2	4	1	7

Table 1: The number of pieces for different instrument arrangements from the original and expanded URMP dataset.

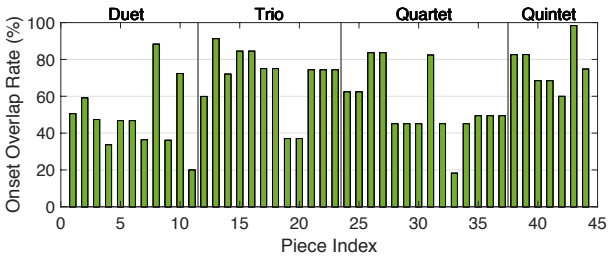

Figure 8: Onset overlap rate for each piece from the original URMP dataset.

sociate players with tracks. Figure 8 plots this statistic for all of the original 44 pieces. While the rate varies much from one piece to another, we see a general increasing trend as the polyphony increases.

4.2 System Setup

For implementation, the audio is processed with a frame length of 42.7 ms and a hop size of 10 ms for score following and pitch contour extraction. When calculating the vibrato correspondence, the motion curve extracted from the 29.97-frame-per-second (FPS) video is up-sampled to 100 FPS, enabling a synchronized time resolution between the audio and video. As vibrato detection is performed on 1-second segments and the onset likelihood curve from body motion is derived from a local maximum within future 30 frames (≈1 second), the system has a 1-second inherent delay when it runs for real-time applications. The past 5-second of body and finger motion velocities are stored in memory to apply PCA (described in Section 3.2.1) and to calculate the weighting parameters in Eq. (7).

For evaluation, we first address each track independently to investigate the quality of the extracted onset likelihood features, using the traditional onset detection measures. Then we evaluate the association performance on the expanded set of ensemble pieces. The result is grouped by different ensemble types, from duets to quintets, which directly correlate with the difficulty levels. Note that whatever number of tracks presented in the performance, only one association is correct. We do not include a quantitative evaluation of the score following and vibrato detection modules in this paper, since they have been fully evaluated in the previous work.

4.3 Onset Detection Evaluation

As two modules of the proposed system rely on the synchronization cues of onset timing between different modalities, we evaluate the quality of our proposed onset likelihood curves that are extracted from body motions and finger motions. To do so, we set up an onset detection task. We take the onset likelihood curve as the onset detection function (Bello et al., 2005), and perform peak-picking to retrieve the onsets. A true positive detection is counted when a detected onset is within a tolerance window of 3 video frames (100ms). This is wider than the standard 50ms in the literature, since the precise timing is not the main focus of the source association system.

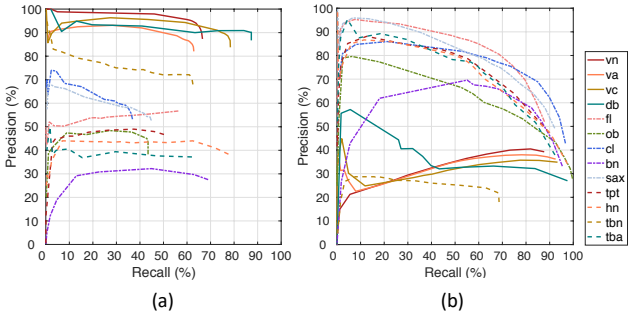

Figure 9: Onset detection evaluation results from body motions (a) and finger motions (b) for different instruments.

Figure 9 plots the precision versus recall by varying the peak-picking threshold on the onset likelihood curves extracted from body motions and finger motions respectively. Precision and recall are calculated for each instrument across all pieces in the original dataset. Observing Figure 9 reveals that the onset likelihood curve extracted from body motion shows better correlation with the ground-truth onset timings for string instruments, while, that from finger motion shows better correlation for woodwind/brass instruments. An exception is trombone, where the onset likelihood curve extracted from body motion shows better correlation than that from finger motions. The observation is reasonable as the trombone pitch change (hence note transition) is mainly performed by moving the slide using the right arm (body motion).

Another interesting observation is that although the onset likelihood curve $\phi_b(t)$ in Figure 3 is visually less noisy than $\phi_f(t)$ in Figure 5, the recall calculated from $\phi_b(t)$ for string instruments cannot reach as high value as that of woodwind/brass instruments calculated from $\phi_f(t)$. We argue that this is because legato bowing (i.e., articulating a sequence of notes from one sustained bowing action) is widely used in string instrument performances, where onset detection from bow motions misses some true positives. This explains the upper bound of recall rates (around 80% as in Figure 9 (a)) for string instruments. For woodwind/brass instruments, there are also onsets not vis-

String	excerpt duration (sec)					
	5	10	15	20	25	30
Duet	1323	642	420	303	236	200
Trio	1044	506	333	240	189	158
Quartet	355	172	114	82	65	54
Quintet	64	31	21	15	12	10
Woodwind/Brass	**excerpt duration (sec)**					
	5	10	15	20	25	30
Duet	1809	887	557	435	323	266
Trio	1275	626	391	309	229	187
Quartet	474	232	145	115	86	68
Quintet	66	32	20	16	12	9
Mixed	**excerpt duration (sec)**					
	5	10	15	20	25	30
Duet	441	203	141	96	82	60
Trio	380	174	121	82	70	51
Quartet	199	92	64	44	37	28
Quintet	22	10	7	5	4	3

Table 2: The number of evaluation samples with different length and instrumentation for source association.

ible such as repeated notes, but the amount is much smaller, which explains why the recall rates can reach closely to 100% in Figure 9 (b).

4.4 Source Association Evaluation

In this section we evaluate the source association performance, first for each module (corresponding to each component in Eq. (5) independently, then for the finally integrated approach. We use *association accuracy* as the evaluation measure, which is defined as the percentage of correctly associated pieces among all testing pieces. A piece is considered correctly associated if the exactly correct bijection between players and score/audio tracks is retrieved. Note that the difficulty of source association increases dramatically from small to large ensembles. In a quintet ensemble, there are in total $5! = 120$ bijection candidates, and only one is considered correct. Therefore, we divide our evaluation based on the size of ensembles.

Besides the ensemble size, the length of the performance also affects the difficulty of the association problem, assuming longer pieces provide richer cues. In an online setting, we hope that the proposed system can retrieve the correct association as quickly as possible. Therefore, in the experiments, we segment the testing pieces into non-overlapping excerpts for each of the following lengths: 5, 10, 15, 20, 25, and 30 seconds. When doing so, we first remove the beginning and the last 5 seconds of each piece as the performance may not cover the entire length of those segments. This segmentation further expands the testing pieces to a large number of evaluation samples, totaling 17,574 samples, as presented in Table 2.

4.4.1 Body Motion

We first evaluate the source association performance using the normalized onset correspondence $\bar{M}_b$ between score tracks and body motions (the first component of Eq. (5)). Figure 10 (a)-(c) shows the association accuracy for ensembles consisting of string, woodwind/brass, and mixed instruments

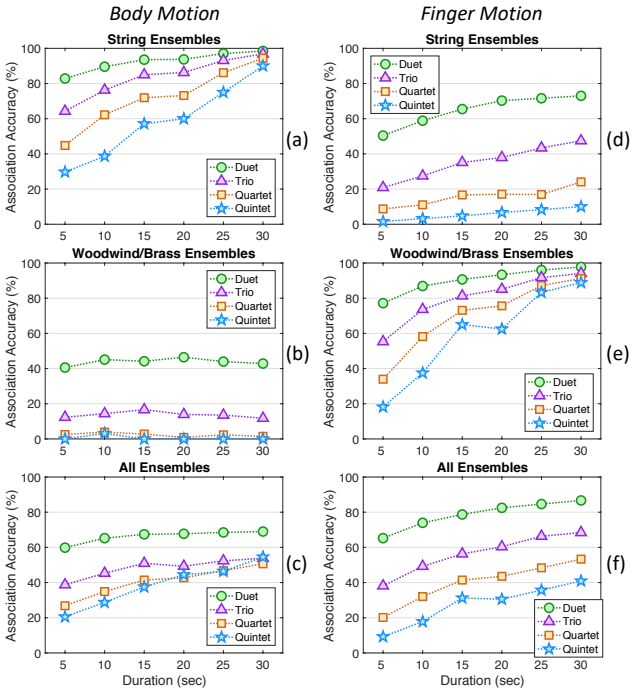

Figure 10: (a)-(c): Source association accuracy only using onset correspondence between score tracks and body motions (the first component $\bar{M}_b$ in Eq. (5)). (d)-(f): Source association accuracy only using onset correspondence between score tracks and finger motions (the second component $\bar{M}_f$ in Eq. (5)).

with different polyphony. Note that the "All Ensembles" evaluated in Figure 10 (c) and (f) contain all the instrument categories from Table 2, i.e., String+Woodwind/Brass+Mixed. For each piece, we plot how the association accuracy varies as the duration of the input stream increases from 5 to 30 seconds. Each marker in the figure is the association accuracy calculated from the number of excerpts shown in Table 2.

Comparing different ensemble sizes, the association accuracy decreases as the number of players/tracks increases. From Figure 10 (a), we find that correlating onsets with body motions is beneficial for string instruments. Note that this evaluation is reproduced from our previous work in (Li et al., 2017a) as one baseline system here. The accuracy increases as the duration of video stream increases, which provides more cues to solve the association. The accuracy reaches around 90% for all ensemble sizes when the video stream duration reaches 30 seconds. This strategy based on onset correspondence from body motion, however, is not effective for woodwind/brass instruments, where the association accuracy remains around random guess accuracy as shown in Figure 10 (b), e.g., 1/6 for trios. This observation is consistent with our expectations and the onset detection evaluations in Figure 9.

4.4.2 *Finger Motion*

We then evaluate the source association performance using the normalized onset correspondence $\bar{M}_f$ between score tracks and finger motions (the second component of Eq. (5)). The association accuracy is plotted in Figure 10 (d)-(f), with the same set of pieces used for evaluations plotted in Figure 10 (a)-(c). From Figure 10 (d)-(f) we can observe that finger motion is a more prominent cue for correspondence with note onsets for woodwind/brass instruments (except for trombone). When a 30-second video excerpt is available, the association accuracy reaches about 90% for all sizes of woodwind/brass ensembles. These observations are also consistent with our onset detection evaluations in Figure 9. For string instruments, however, the extracted finger motions are mostly vibrato motions, which are not relevant to note onsets.

Figure 10 also reveals some limitations of the source association solution based on onset-motion correspondence. First, there are many note onsets not revealed from body or finger motions, such as notes played with legato bowing for string instruments and repeated notes from woodwind/brass instruments, as analyzed in Section 4.3. Second, as note synchronization between players is at the foundation of music performances, note onsets between tracks have high chances to overlap with each other, as shown in Figure 8. The limitations restrict the association performance for approaches that only rely on onset-motion correspondence, especially from short video excerpts.

4.4.3 *Vibrato Motion*

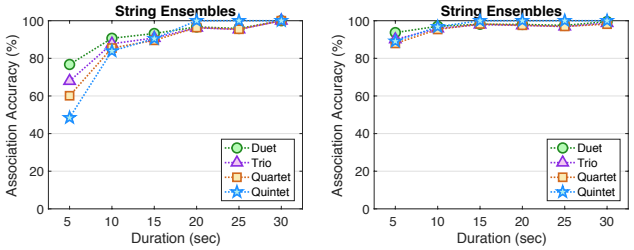

Figure 11: Source association accuracy of string ensembles by (a) only using vibrato correspondence between pitch fluctuation and hand motion ($\bar{M}_v$ in Eq. (5)), and (b) combining vibrato correspondence with onset correspondence from body motion in ($\bar{M}_b$ and $\bar{M}_v$ in Eq. (5)).

The correspondence between pitch fluctuations and vibrato motion (denoted as $\bar{M}_v$, the third component of Eq. (5)) helps to retrieve the source association on a finer level for string instrumentalists. The evaluation result is plotted in Figure 11 (a) for the same set of pieces performed by string ensembles used for evaluations plotted in Figure 10 (a). Note that this baseline is the same system as proposed in our previous work in (Li et al., 2017c). We do not include the woodwind/brass instrument group here since no vibrato pattern can be detected from finger motions. We can find that the source association can reach a high accuracy from shorter video clips, i.e., 90% after 10 seconds. The limitation of this approach is that vibrato articulation is not guaranteed to be always present in the performance. We thus combine this module with the onset correspondence from body motions, the two dominant cues to solve association for string instruments, to evaluate the association accuracy as shown in Figure 11 (b). The two components from $\bar{M}_b$ and $\bar{M}_v$ work together to reach a high association accuracy from a short video stream.

4.4.4 *The Integrated System*

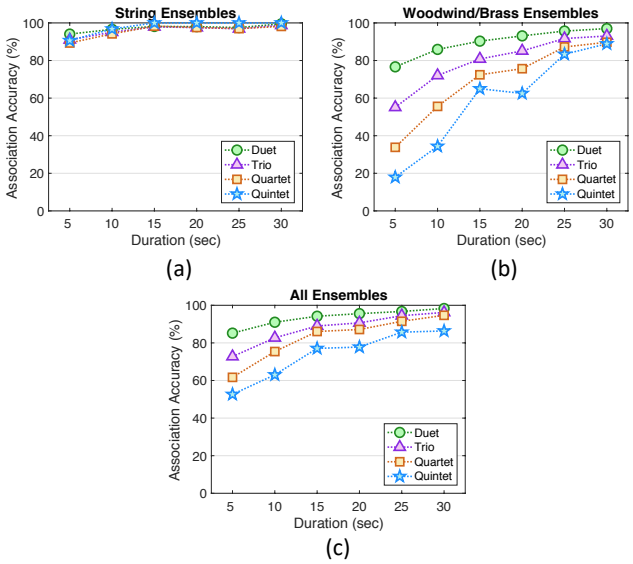

Figure 12: Source association accuracy of ensembles with different instrumentation using all of the three modules: onset correspondence from body motions, onset correspondence from finger motions, and vibrato correspondence from hand motions (Eq. (5)).

Finally, we evaluate the proposed complete source association system after integrating all the modules together, as presented in Eq. (5). The evaluated pieces are the same as the ones plotted in Figure 10. This presents a universal source association system for common melodic instruments. Overall woodwind/brass instruments have less chance to retrieve the correct association than string instruments, since only the subtle finger motions contribute to the correspondence with onset events. This correspondence is often inaccessible due to overlapping onsets across tracks or repeating onsets as analyzed in Section 4.4.2. Comparing Figure 12 (a) with Figure 11 (b), or Figure 12 (b) with Figure 10 (e), we observe that adding components with irrelevant association cues does not harm the system, thanks to the weighting strategy in Eq. (5) over different modules. Comparing Figure 12 (c) with Fig-

ure 10(c)/(f), the integrated system greatly improves the association accuracy for pieces with mixed types of instruments. The association accuracy for mixed ensembles is between that of string and woodwind/brass ensembles.

4.5 Discussions

The proposed source association system is designed and evaluated for the online scenario, where all the system components do not rely on the performance data after the current time instant. Note that due to the limitation of the dataset, we have not systematically evaluated the robustness of the system against camera viewpoint changes. However, we argue that this will not be a big problem for the proposed system, as all the rigid/affine motions are easy to eliminate by setting up reference points (e.g., players' hips) after extracting the skeleton data for each player. Another challenge in a real-world application is introduced by camera shot transitions in music video post-production. One suggested strategy is to clear the accumulated association scores and re-register the players when a shot transition is detected. But further experiments need to be conducted to validate this strategy. Another limitation of the experiments is that all the players in the dataset have their front-right side facing the camera with most finger motions visible. If this is not satisfied in real scenarios, only the first computation module (correspondence between body motion and note onsets) provides useful information, making the system only work for string ensembles. This conclusion, however, is also true for humans to recognize the association.

5. Conclusion

In this paper, we propose an online source association system for Western chamber ensembles, which aims to retrieve the association between players in the video and the audio/score tracks through the analysis of the cross-modal temporal correspondences. We designed three modules to model different correspondences between 1) body motions and note onsets, 2) finger motions and note onsets, and 3) vibrato motions and pitch fluctuations. Although these correspondences apply to different kinds of instruments, the proposed system automatically integrates them in an adaptive fashion, without the need for knowing the instrument types. This makes the system a universal framework for common instruments in Western chamber ensembles including strings, woodwind, and brass instruments. In addition, the system runs in an online fashion to update association results as the video stream progresses. Experiments with audio-visual recordings of performances with different polyphony and instrumentation demonstrate that the accuracy of the proposed system increases with the length of video streams, and high accuracy is achieved within a relatively short interval. The accuracy for string ensembles is generally better than that for woodwind, brass, and mixed-instrument ensembles because more correspondences are modeled for these instruments.

Acknowledgment

This work is supported by the National Science Foundation grant No. 1741472.

References

Arandjelović, R. and Zisserman, A. (2018). Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 1, pages 451–466. DOI: https://doi.org/10.1007/978-3-030-01246-5_27.

Arzt, A. and Widmer, G. (2010). Simple tempo models for real-time music tracking. In *Proceedings of the Sound and Music Computing Conference (SMC)*.

Arzt, A., Widmer, G., and Dixon, S. (2008). Automatic page turning for musicians via real-time machine listening. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, pages 241–245.

Askenfelt, A. (1989). Measurement of the bowing parameters in violin playing. II: Bow–bridge distance, dynamic range, and limits of bow force. *The Journal of the Acoustical Society of America*, 86(2):503–516. DOI: https://doi.org/10.1121/1.398230.

Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M. J., and Szeliski, R. (2011). A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 92(1):1–31. DOI: https://doi.org/10.1007/s11263-010-0390-2.

Barzelay, Z. and Schechner, Y. Y. (2007). Harmony in motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. DOI: https://doi.org/10.1109/CVPR.2007.383344.

Barzelay, Z. and Schechner, Y. Y. (2010). Onsets coincidence for cross-modal analysis. *IEEE Transactions on Multimedia*, 12(2):108–120. DOI: https://doi.org/10.1109/TMM.2009.2037387.

Bazzica, A., Liem, C. C., and Hanjalic, A. (2014). Exploiting instrument-wise playing/non-playing labels for score synchronization of symphonic music. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, pages 201–206.

Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., and Sandler, M. (2005). A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047. DOI: https://doi.org/10.1109/TSA.2005.851998.

Burkholder, J. P. and Grout, D. J. (2014). *A History of Western Music: Ninth International Student Edition*. WW Norton & Company, Inc.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 7291–7299. DOI: https://doi.org/10.1109/CVPR.2017.143.

Casanovas, A. L., Monaci, G., Vandergheynst, P., and Gribonval, R. (2010). Blind audiovisual source separation based on sparse redundant representations. *IEEE Transactions on Multimedia*, 12(5):358–371. DOI: https://doi.org/10.1109/TMM.2010.2050650.

Casanovas, A. L. and Vandergheynst, P. (2010). Nonlinear video diffusion based on audio-video synchrony. *IEEE Transactions on Multimedia*.

Cutler, R. and Davis, L. (2000). Look who's talking: Speaker detection using video and audio correlation. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, volume 3, pages 1589–1592. DOI: https://doi.org/10.1109/ICME.2000.871073.

Dahl, S. (2004). Playing the accent-comparing striking velocity and timing in an ostinato rhythm performed by four drummers. *Acta Acustica united with Acustica*, 90(4):762–776.

Dinesh, K., Li, B., Liu, X., Duan, Z., and Sharma, G. (2017). Visually informed multi-pitch analysis of string ensembles. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3021–3025. DOI: https://doi.org/10.1109/ICASSP.2017.7952711.

Dixon, S. (2005). Live tracking of musical performances using on-line time warping. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pages 92–97.

Duan, Z., Essid, S., Liem, C., Richard, G., and Sharma, G. (2019). Audiovisual analysis of music performances: Overview of an emerging field. *IEEE Signal Processing Magazine*, 36(1):63–73. DOI: https://doi.org/10.1109/MSP.2018.2875511.

Duan, Z. and Pardo, B. (2011a). Soundprism: An online system for score-informed source separation of music audio. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1205–1215. DOI: https://doi.org/10.1109/JSTSP.2011.2159701.

Duan, Z. and Pardo, B. (2011b). A state space model for online polyphonic audio-score alignment. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 197–200. DOI: https://doi.org/10.1109/ICASSP.2011.5946374.

Duan, Z., Pardo, B., and Zhang, C. (2010). Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):2121–2133. DOI: https://doi.org/10.1109/TASL.2010.2042119.

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W. T., and Rubinstein, M. (2018). Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics (TOG)*, 37(4). DOI: https://doi.org/10.1145/3197517.3201357.

Ewert, S., Pardo, B., Müller, M., and Plumbley, M. D. (2014). Score-informed source separation for musical audio recordings: An overview. *IEEE Signal Processing Magazine*, 31(3):116–124. DOI: https://doi.org/10.1109/MSP.2013.2296076.

Fisher, J. W. and Darrell, T. (2004). Speaker association with signal-level audiovisual fusion. *IEEE Transactions on Multimedia*, 6(3):406–413. DOI: https://doi.org/10.1109/TMM.2004.827503.

Gao, R., Feris, R., and Grauman, K. (2018). Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 3, pages 36–54. DOI: https://doi.org/10.1007/978-3-030-01219-9_3.

Geringer, J. M., MacLeod, R. B., and Allen, M. L. (2010). Perceived pitch of violin and cello vibrato tones among music majors. *Journal of Research in Music Education*, 57(4):351–363. DOI: https://doi.org/10.1177/0022429409350510.

Godøy, R. I. and Jensenius, A. R. (2009). Body movement in music information retrieval. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, pages 45–50.

Grubb, L. and Dannenberg, R. (1997). A stochastic method of tracking a vocal performer. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 301–308.

Izadinia, H., Saleemi, I., and Shah, M. (2013). Multimodal analysis for identification and segmentation of moving-sounding objects. *IEEE Transactions on Multimedia*, 15(2):378–390. DOI: https://doi.org/10.1109/TMM.2012.2228476.

Kidron, E., Schechner, Y. Y., and Elad, M. (2007). Cross-modal localization via sparsity. *IEEE Transactions on Signal Processing*, 55(4):1390–1404. DOI: https://doi.org/10.1109/TSP.2006.888095.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 2(1-2):83–97. DOI: https://doi.org/10.1002/nav.3800020109.

Li, B., Dinesh, K., Duan, Z., and Sharma, G. (2017a). See and listen: Score-informed association of sound tracks to players in chamber music performance videos. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2906–

2910. DOI: https://doi.org/10.1109/ICASSP.2017.7952688.

Li, B., Dinesh, K., Sharma, G., and Duan, Z. (2017b). Video-based vibrato detection and analysis for polyphonic string music. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, pages 123–130.

Li, B., Liu, X., Dinesh, K., Duan, Z., and Sharma, G. (2018a). Data from: "Creating a multi-track classical music performance dataset for multi-modal music analysis: Challenges, insights, and applications.". *Dryad Digital Repository*. https://doi.org/10.5061/dryad.ng3r749.

Li, B., Liu, X., Dinesh, K., Duan, Z., and Sharma, G. (2019). Creating a music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia*, 21(2):522–535. DOI: https://doi.org/10.1109/TMM.2018.2856090.

Li, B., Maezawa, A., and Duan, Z. (2018b). Skeleton plays piano: online generation of pianist body movements from MIDI performance. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.

Li, B., Xu, C., and Duan, Z. (2017c). Audiovisual source association for string ensembles through multi-modal vibrato analysis. In *Proceedings of the Sound and Music Computing (SMC)*, pages 159–166.

Li, K., Ye, J., and Hua, K. A. (2014). What's making that sound? In *Proceedings of the ACM International Conference on Multimedia*, pages 147–156. DOI: https://doi.org/10.1145/2647868.2654936.

Liu, Y. and Sato, Y. (2008). Finding speaker face region by audiovisual correlation. In *Proceedings of the Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications (M2SFA2)*.

Müller, M. (2007). Dynamic time warping. In *Information retrieval for music and motion*, chapter 4, pages 69–84. Springer. DOI: https://doi.org/10.1007/978-3-540-74048-3_4.

Müller, M. (2015). Music synchronization. In *Fundamentals of music processing*, chapter 3, pages 115–166. Springer. DOI: https://doi.org/10.1007/978-3-319-21945-5_3.

Müller, M., Mattes, H., and Kurth, F. (2006). An efficient multiscale approach to audio synchronization. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.

Obata, S., Nakahara, H., Hirano, T., and Kinoshita, H. (2009). Fingering force in violin vibrato. In *Proceedings of the International Symposium on Performance Science*, volume 429.

Owens, A. and Efros, A. A. (2018). Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 6, pages 639–658. DOI: https://doi.org/10.1007/978-3-030-01231-1_39.

Paleari, M., Huet, B., Schutz, A., and Slock, D. (2008). A multimodal approach to music transcription. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 93–96. DOI: https://doi.org/10.1109/ICIP.2008.4711699.

Palmer, C., Carter, C., Koopmans, E., and Loehr, J. D. (2007). Movement, planning, and music: Motion coordinates of skilled performance. In *Proceedings of the International Conference on Music Communication Science*, pages 119–122. University of New South Wales Sydney, NSW.

Parekh, S., Essid, S., Ozerov, A., Duong, N., Perez, P., and Richard, G. (2017a). Motion informed audio source separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6–10. DOI: https://doi.org/10.1109/ICASSP.2017.7951787.

Parekh, S., Essid, S., Ozerov, A., Duong, N. Q., Pérez, P., and Richard, G. (2017b). Guiding audio source separation by video object information. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 61–65. DOI: https://doi.org/10.1109/WASPAA.2017.8169995.

Parncutt, R. and McPherson, G. (2002). *The science and psychology of music performance: Creative strategies for teaching and learning*. Oxford University Press. DOI: https://doi.org/10.1177/1321103X020190010803.

Senocak, A., Oh, T.-H., Kim, J., Yang, M.-H., and Kweon, I. S. (2018). Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4358–4366. DOI: https://doi.org/10.1109/CVPR.2018.00458.

Sigg, C., Fischer, B., Ommer, B., Roth, V., and Buhmann, J. (2007). Nonnegative CCA for audiovisual source separation. In *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing*, pages 253–258. DOI: https://doi.org/10.1109/MLSP.2007.4414315.

Sörgjerd, M. (2000). *Auditory and Visual Recogniton of Emotional Expression in Performance of Music*. PhD thesis, Uppsala universitet, Institutionen för psykologi.

Sun, D., Roth, S., and Black, M. J. (2010). Secrets of optical flow estimation and their principles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2439. DOI: https://doi.org/10.1109/CVPR.2010.5539939.

Thomas, V., Fremerey, C., Damm, D., and Clausen, M. (2009). SLAVE: a score-lyrics-audio-video-explorer. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.

Tian, Y., Shi, J., Li, B., Duan, Z., and Xu, C. (2018). Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 2, pages 252–268. DOI: https://doi.org/10.1007/978-3-030-01216-8_16.

Tsay, C.-J. (2014). The vision heuristic: Judging music ensembles by sight alone. *Organizational Behavior and Human Decision Processes*, 124(1):24–33. DOI: https://doi.org/10.1016/j.obhdp.2013.10.003.

Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., and Torralba, A. (2018). The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 1, pages 587–604. DOI: https://doi.org/10.1007/978-3-030-01246-5_35.

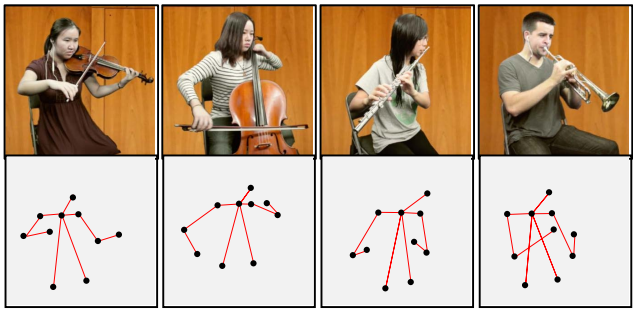

Figure 2: Body motion extraction. Upper body skeletons (second row) are extracted with OpenPose (Cao et al., 2017) in each video frame (first row) followed by temporal smoothing over time.

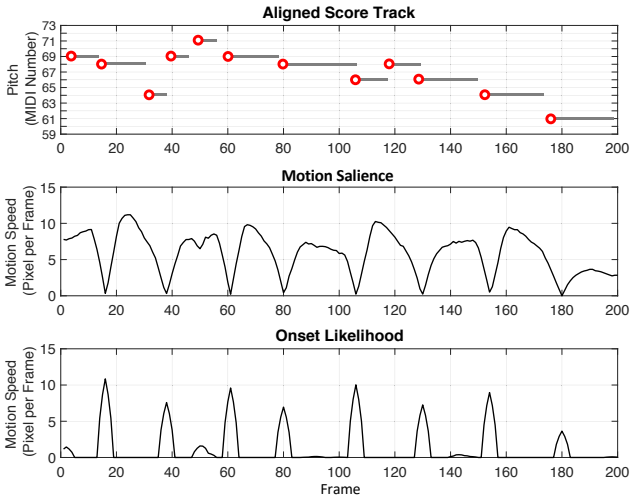

Figure 3: Example correspondence between body motion and note onsets. Top: temporally aligned score track with onsets marked by red circles. Middle: extracted motion salience (primarily bowing motion) from the visual performance of a violin player. Bottom: derived onset likelihood curve from the motion salience.

approach Li et al. (2017a) we apply optical flow estimation to extract bowing motion of string players. However, we argue that this pixel-level analysis may not be ideal for semantic-level understanding of body gestures and movements, and can be less robust to occlusions and camera viewpoint changes.

In this paper, we propose to apply OpenPose (Cao et al., 2017) on each frame, a multi-person pose estimation approach to extract body skeleton coordinates for all the players on stage without pre-segmentation on the video recording. A skeleton in each frame is represented as a 20-D vector $\mathbf{y}(t)$ corresponding to the horizontal and vertical coordinates of the 10 upper body joints, including nose, neck, shoulders, elbows, wrists, and hips. We do not include lower body joints as they are often less relevant to note events. Figure 2 shows video frames of several instrumentalists with the extracted body skeletons. To form a continuous skeleton sequence across time, we eliminate joint coordinates if the confidence score from OpenPose is smaller than 0.2 *and* the L_2 distance between a joint in consecutive frames is larger than 10% of the head-hip distance, which is considered the maximal regular movement in a ≈30-FPS video without shot transition. We also temporally smooth their coordinates using a moving average with a 5-frame window size. These post-processings are referred from (Li et al., 2018b) where the same approach is applied to extract skeletons for pianists. We then take the two hips as reference coordinates to align the body position across frames. Finally, we calculate motion velocities $\mathbf{z}(t)$ as the derivative of $\mathbf{y}(t)$ w.r.t. time. Compared to optical flow estimation, this gesture-based motion analysis approach is semantically more meaningful, less computationally expensive, and more robust to occlusions and camera viewpoint changes such as camera zooming or panning.

To extract motions related to note onsets in each video frame, for each player we denote the motion velocities of n frames in the past as $\mathbf{Z} = [\mathbf{z}_1(t), \mathbf{z}_2(t), \cdots, \mathbf{z}_n(t)]^T \in \mathbb{R}^{n\times 20}$ and apply principal component analysis (PCA) by eigen value decomposition $\mathbf{Z}^T\mathbf{Z} = \mathbf{V}\Sigma\mathbf{V}^T$, where $\mathbf{V}$ and Σ represent the matrix of eigen vectors and the corresponding eigen values respectively. We then project the motion velocity $\mathbf{z}(t)$ onto the principal component direction (first column of $\mathbf{V}$) and take its absolute value as the motion salience $s(t)$. Choosing the salient motion discards the direction information of the motions (e.g., up/down-bow for violinists), which is less relevant to timings than the amplitude information. We set n to 150 frames, i.e., 5 seconds in time, assuming a player's pose stays consistent in this range. To reduce the computational cost, we update $\mathbf{V}$ every 1 second (assuming consistent motion patterns within a short period).

3.2.2 *Onset Likelihood*

From the motion salience $s(t)$, we infer the timings of the motion strokes that are potentially related to score note onsets. As a note onset often corresponds to the beginning or ending of a sound articulation motion (e.g., a bowing stroke for string instruments), the motion speed at the onset is often small. Therefore, local minima of the motion salience $s(t)$ is often indicative of note onsets. Let Ω be the set of all the local minima throughout a piece. For each local minimum $\tau \in \Omega$, we represent the likelihood of a note onset as $a(\tau) = \max_{\gamma\in[\tau,\tau+30]} s(\gamma) - s(\tau)$ that is determined by the maximum speed of the motion stroke considering the following 30 frames: the larger the value of $a(\tau)$ the more likely that a note onset is activated by the motion stroke. Here 30 frames are considered to span the high energy part of most notes. Therefore, we can define an *onset likelihood* curve $\phi_b(t)$ derived from body motion

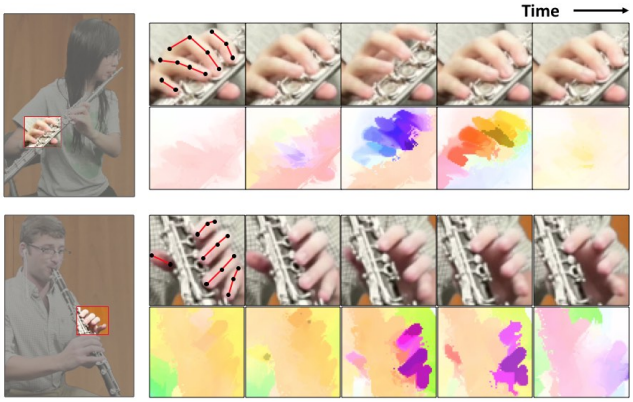

Figure 4: Optical flow visualization of finger motions in five consecutive frames corresponding to note changes. The color encoding scheme is adopted from Baker et al. (2011).

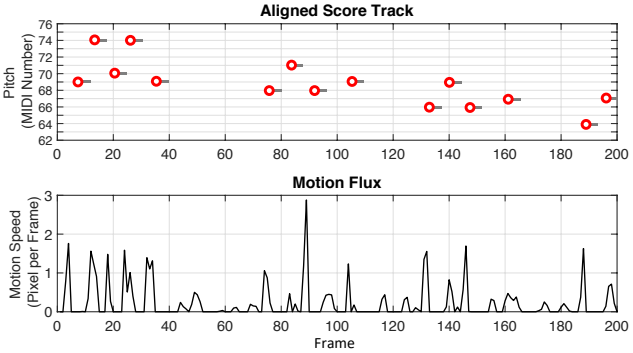

Figure 5: Example correspondence between finger motion and note onsets of a flute player. Top: temporally aligned score track with onsets marked by red circles. Bottom: extracted motion flux from finger movements.

work (Li et al., 2017c) to extract the fine motion and find the correspondence with pitch contours extracted from the audio modality.

3.4.1 *Vibrato Motion Extraction*

We retrieve the fingering motion $\mathbf{u}^{(ij)}(t)$ as computed from the previous section. Although the vibrato motion is mostly a rigid motion (fingers move together with little relative movements), it is periodic and very fast (usually about 4-7.5 Hz (Geringer et al., 2010)), and hence it is not removed as other slow rigid/affine motions. Figure 6 illustrates several frames of the optical flow estimation of the vibrato hand motions from the two players. For each frame t, we take the average motion vector across all pixels within the ROI as $\mathbf{u}(t) = [u_x(t), u_y(t)]^T$, where the motion direction is preserved for vibrato detection.

The vibrato detection module works as a binary classifier as proposed and trained by Li et al. (2017b). The classifier is implemented as a support vector machine (SVM) that takes the input of a 8-dimensional feature extracted from each sample, including the zero

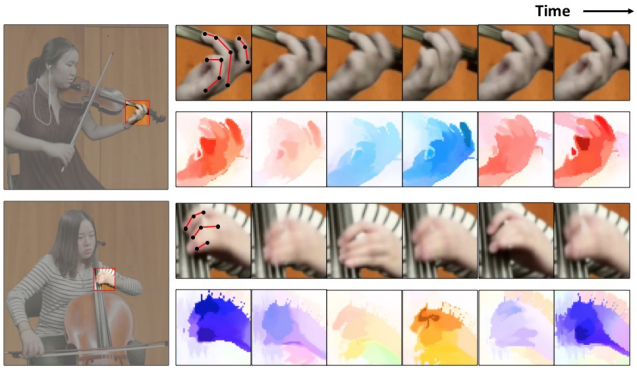

Figure 6: Optical flow visualization of the hand motions corresponding to vibrato articulation. Color encoding scheme is adopted from Baker et al. (2011).

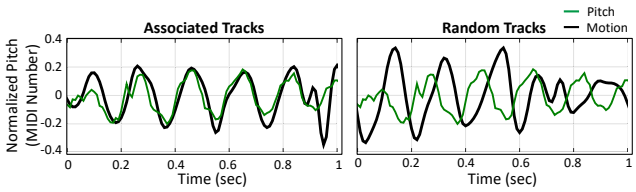

Figure 7: The same segment of normalized pitch contour $f(t)$ (green) overlaid with the motion displacement curve $d(t)$ (black) from the associated track (left) and another random track (right).

crossing rate of the x- and y- motion velocities and their auto-correlations, the energy of 3-9 Hz frequency range, and the auto-correlation peaks. According to Li et al. (2017b), this method achieves a vibrato detection accuracy of over 90% from visual motions regardless of the polyphony number and instrument type within the string instrument family. Here each input sample is a 1-second segment of $\mathbf{u}(t)$ (again introducing an average 0.5-second delay of the association system).

For each detected vibrato segment, we perform PCA on $\mathbf{u}(t)$ within this 1-second segment to obtain the 1-D principal motion velocity curve $v(t)$. We then integrate $v(t)$ over time to calculate a *motion displacement curve,* $d(t)$, which corresponds to the length fluctuation of the vibrating string, and hence the pitch fluctuation of the note. We normalize each vibrato segment of $d(t)$ to zero mean and unit variance. We set the non-vibrato segments of $d(t)$ to zero.

3.4.2 *Pitch Contour Extraction*

Utilizing the score information, we apply Soundprism (Duan and Pardo, 2011a), an online score-informed source separation system, to separate the polyphonic audio mixture into individual sources. Note that although audio recordings of individual instrumental tracks are available in the dataset, we do not use them as they are not generally available in real concert scenarios. To extract the pitch contour, we perform a score-informed pitch estimation step on each separated