Zhiyao Duan, Slim Essid, Cynthia C.S. Liem,
Gaël Richard, and Gaurav Sharma

Audiovisual Analysis of Music Performances

Overview of an emerging field

©ISTOCKPHOTO.COM/TRAFFIC_ANALYZER

In the physical sciences and engineering domains, music has traditionally been considered an acoustic phenomenon. From a perceptual viewpoint, music is naturally associated with hearing, i.e., the audio modality. Moreover, for a long time, the majority of music recordings were distributed through audio-only media, such as vinyl records, cassettes, compact discs, and mp3 files. As a consequence, existing automated music analysis approaches predominantly focus on audio signals that represent information from the acoustic rendering of music.

Music performances, however, are typically multimodal [1], [2]: while sound plays a key role, other modalities are also critical to enhancing the musical experience. In particular, the visual aspects of music—be they disc cover art, videos of live performances, or abstract music videos—play an important role in expressing musicians' ideas and emotions. With the popularization of video-streaming services over the past decade, such visual representations also are increasingly available with distributed music recordings. In fact, video-streaming platforms have become one of the preferred music distribution channels, especially among the younger generation of music consumers.

Simultaneously seeing and listening to a musical performance often provides a richer experience than pure listening. Researchers have found that "the visual component is not a marginal phenomenon in music perception, but an important factor in the communication of meanings" [3]. Even for prestigious classical music competitions, studies have revealed that visually perceived elements of the performance, such as the musician's gestures, motions, and facial expressions, affect the evaluations of judges (experts or novices alike) even more significantly than the sound [4].

Symphonic music provides another example of visible communicated information where large groups of orchestra musicians play simultaneously in close coordination. For expert audiences familiar with the genre, both the visible coordination between musicians and the ability to closely watch individuals within the group add to the attendee's emotional experience of a concert [5]. Attendees unfamiliar with the genre can also be

Digital Object Identifier 10.1109/MSP.2018.2875511
Date of publication: 24 December 2018

better engaged via enrichment, i.e., offering supporting information in various modalities (e.g., visualizations or textual explanations) beyond the stimuli that the event naturally triggers in the physical world.

In addition to the audiences at music presentations, others also gain from information obtained through audiovisual rather than audio-only analysis. In educational settings, instrument learners benefit significantly from watching demonstrations by professional musicians, where the visual presentation provides deeper insight into specific instrument-technical aspects of the performance (e.g., fingering or choice of strings). Generally, when broadcasting audiovisual productions involving large ensembles captured with multiple recording cameras, it is also useful for the producer to be aware of which musicians are visible in which camera stream at each point in time. For such analyses to be done, relevant information needs to be extracted from the recorded video signals and coordinated with recorded audio. As a consequence, there has recently been growing interest in the visual analysis of musical performances, even though such analysis was largely overlooked in the past.

Aim and focus

In this article, we aim to introduce this emerging area to the music signal processing community and the broader signal processing community. To our knowledge, this article is the first overview of research in this area. For conciseness, we restrict our attention to the analysis of audiovisual music performances, which is an important subset of audiovisual music productions that is also representative of the main challenges and techniques of this field of study. Other specific applications, such as the analysis of music video clips or other types of multimodal recordings not involving audio and visuals (e.g., lyrics or music score sheets), although important in their own right, are not covered here to maintain a clear focus and a reasonable length.

Significance and challenges

Significance

Figure 1 illustrates some examples of how visual and aural information in a musical presentation complement each other, and how they offer more information on the performance than what can be obtained by considering only the audio channel and a musical score. In fact, while the musical score is often considered to be the ground truth of a musical presentation, significant performance-specific expressive information, such as the use of vibrato, is not indicated in the score and is instead evidenced in the audiovisual performance signals.

Compared to audio-only music performance analysis, the visual modality offers extra opportunities to extract musically meaningful cues out of recorded performance signals.

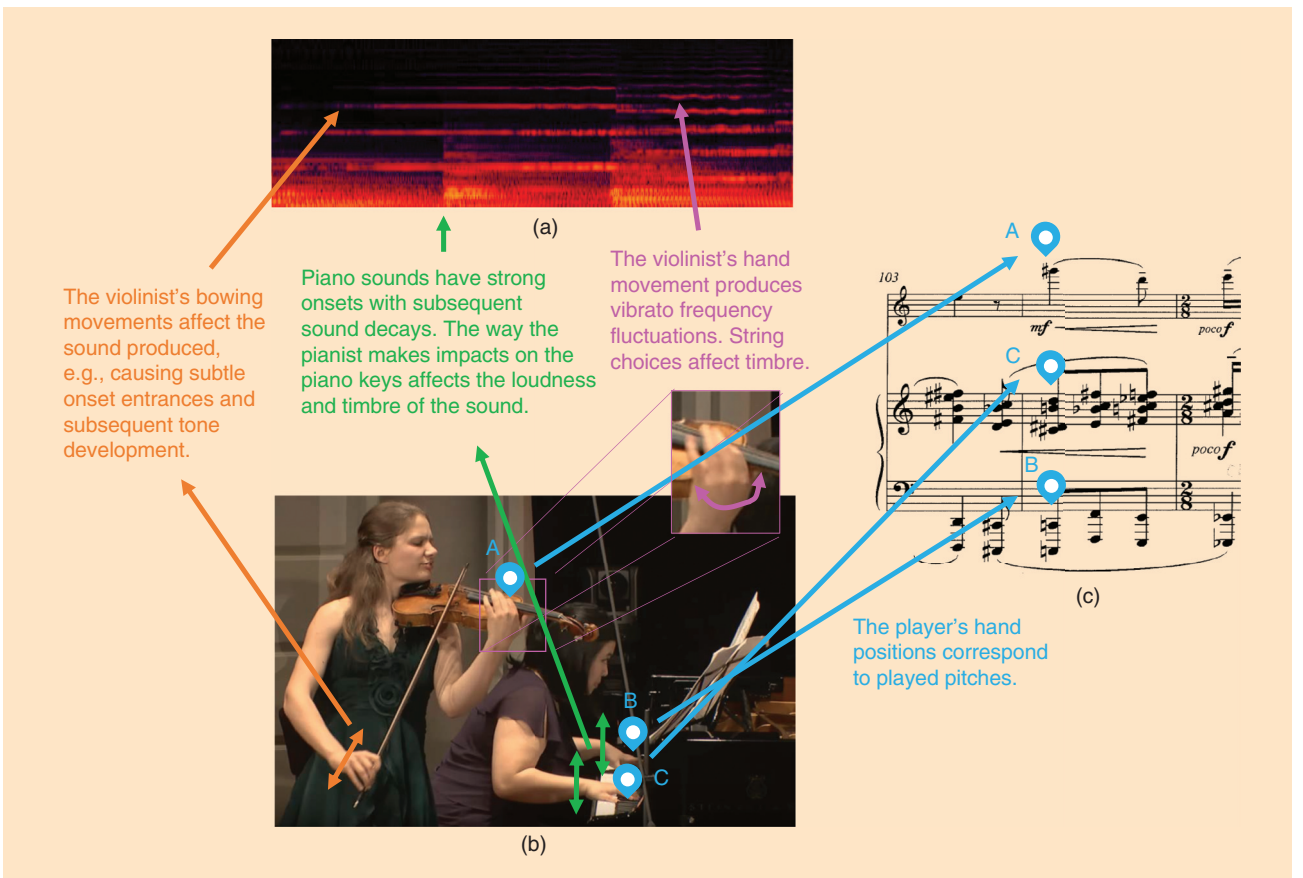

FIGURE 1. Examples of the information present in three parallel representations of a music performance excerpt: (a) a spectrogram of a recorded audio signal, (b) a video recording of performing musicians, and (c) a score of the performed music.

In some cases, the visual modality allows for addressing tasks that would not be possible in audio-only analysis, e.g., tracking a musician's fingerings or a conductor's gestures and analyzing individual players in the same instrumental section of an orchestra. In other cases, the visual modality provides significant help in task solving, e.g., in source separation and in the characterization of expressive playing styles. In the "Overview of Existing Research" section, we discuss several representative tasks along these lines.

Audiovisual analysis of musical performances broadens the scope of music signal processing research, connecting the audio signal processing area with other areas, i.e., image processing, computer vision, and multimedia. The integration of the audio and visual modalities also naturally creates a connection with emerging research areas, such as virtual reality and augmented reality, and extends music-related human–computer interaction. It serves as a controlled test bed for research on multimodal data analysis, which is critical for building robust and universal intelligent systems.

Challenges

The multimodal nature of audiovisual analysis of music poses new research challenges. First, the visual scenes of music presentations present new problems for image processing and computer vision. Indeed, the visual scene is generally cluttered, especially when multiple musicians are involved, who additionally may be occluded by each other and by music stands. Also, musically meaningful motions may be subtle (e.g., fingering and vibrato motion), and camera views may be complex (e.g., musicians not facing toward cameras, zoom-in/out, and changes of views).

Second, the way to integrate audio and visual processing in the modeling stage of musical scene analysis is a key challenge. In fact, independently tackling the audio and visual modalities to merely fuse the output of the corresponding (unimodal) analysis modules at a later stage is generally not an optimal approach. To take advantage of potential cross-modal dependencies, it is better to combine low-level audiovisual representations as early as possible in the data analysis pipeline. This is, however, not always straightforward. Certain visual signals (e.g., the bowing motion of string instruments) and audio signals (e.g., note onsets) of a sound source are often highly correlated, yet some performer movements (e.g., head nodding) are not directly related to sound [6]. How to discover and exploit audiovisual correspondence in a complex audiovisual scene of music performances is thus a key question.

Third, the lack of annotated data is yet another challenge. While commercial recordings are abundant, they are usually not annotated and are also subject to copyright restrictions that limit their distribution and use. Annotated audio data sets of musical performances are already scarce because of the complexities of recording and ground-truth annotation. Audiovisual data sets are even scarcer, and their creation requires more effort. The lack of large-scale annotated data sets limits the application of many supervised learning techniques that have proven successful for data-rich problems. We note that available music data sets were surveyed in a recent paper [7] that detailed the creation of a new multitrack audiovisual classical music data set. The data set provided in [7] was relatively small, with only 44 short pieces, but was richly annotated, providing individual instrument tracks to allow the assessment of source separation methods and associated music score information in a machine-readable format.

At the other end of the data spectrum, the YouTube-8M data set [8] provides a large-scale labeled video data set (with embedded audio) that also includes many music videos. However, the YouTube-8M data set is currently annotated only with overall video labels and therefore is suited primarily for video/audio classification tasks.

Overview of existing research

It is not an easy task to give a well-structured overview of an emerging field, yet here we make a first attempt from two perspectives. The following section categorizes the existing work into different analysis tasks for different instruments, while the section after that provides a perspective on the type of audiovisual correspondence that is exploited during the analysis.

Categorization of audiovisual analysis tasks

Table 1 organizes existing work on audiovisual analysis of musical presentations along two dimensions: 1) the type of musical instrument and 2) the analysis task.

The first dimension is not only a natural categorization of musicians in a music performance but is also indicative of the types of audiovisual information revealed during the performance. For example, percussionists show large-scale motions that are almost all related to sound articulation. Pianists' hand and finger motions are also related to sound articulation, but they are much subtler and also indicative of the notes being played (i.e., the musical content). For guitars and strings, the left-hand motions are indicative of the notes being played, while the right-hand motions tell us how the notes are articulated (e.g.,

Table 1. A categorization of existing research on audiovisual analysis of music performances according to the type of instrument and the analysis task.

Visual	Is Critical		Is Significant				
Tasks	**Fingering**	**Association**	**Play/Nonplay**	**Onset**	**Vibrato**	**Transcription**	**Separation**
Percussion	N/A	–	[9]	–	N/A	[10]	–
Piano	[11], [12]	–	–	–	N/A	–	–
Guitar	[13]–[16]	–	–	–	–	[16]	–
Strings	[17]	[18], [19]	[9], [20]	[19]	[21]	[17], [20]	[22]
Wind	–	–	[9]	[23]	–	–	–
Singing	N/A	–	–	–	–	–	–

Certain combinations of instruments and tasks do not make sense, and are marked N/A. Various techniques and their combinations have been employed, including support vector machines, hidden Markov models, nonnegative matrix factorization, and deep neural networks.

legato or staccato). For wind instruments, note articulations are difficult to see, and almost all visible motions (e.g., the fingering of a clarinet or the hand positioning of a trombone) are about notes. Finally, singers' mouth shapes reveal only the syllables being sung but not the pitch; also, their body movements can be correlated with the musical content but are not predictive enough for the details.

The second dimension is about tasks or aspects that the audiovisual analysis focuses on. The seven tasks/aspects are further classified into two categories: tasks in which visual analysis is critical and those in which visual analysis provides significant help. In the first category, there are the following tasks:

- *Fingering analysis*: It is very difficult to infer the fingering purely from audio, while it becomes possible by observing the finger positions. There has been research on fingering analysis from visual analysis for guitar [13]–[16], violin [17], and piano [11], [12]. Fingering patterns are mostly instrument specific, but the common idea is to track hand and finger positions relative to the instrument body.
- *Audiovisual source association*: This is a task that determines which player in the visual scene corresponds to which sound source in the audio mixture. The problem is addressed for string instruments by modeling the correlation between visual features and audio features, such as the association between bowing motions and note onsets [18] and that between vibrato motions and pitch fluctuations [19].

The second category contains more tasks. They can be listed as follows:

- *Playing/nonplaying (P/NP) activity detection*: In an ensemble or orchestral setting, it is extremely difficult to detect from the audio mixture whether a certain instrument is being played, yet the visual modality, if not occluded, offers a direct observation of the playing activities of each musician. Approaches based on image classification and motion analysis [9], [20] have been proposed.
- *Vibrato analysis*: This is for string instruments. The periodic movement of the fingering hand detected from visual analysis has been shown to correlate well with the pitch fluctuation of vibrato notes and has been used to detect vibrato notes and analyze the vibrato rate and depth [21].
- *Automatic music transcription*: This and its subtasks, such as multipitch analysis, are very challenging if only audio signals are available. Studies have found that audiovisual analysis is beneficial for monophonic instruments like the violin [17], polyphonic instruments like the guitar [16] and drums [10], and musical bodies like string ensembles [20]. The common underlying idea is to improve audio-based transcription results with play/nonplay activity detection and fingering analysis.
- *Audio source separation*: This is a task that can be significantly improved by audiovisual analysis. The motions of players are often highly correlated with the characteristics of the sound sources [6]. There has been work on modeling such correlations for audio source separation [22].

Besides instrumental players, conductor gesture analysis has also been investigated in audiovisual music performance analysis. Indeed, conductors do not directly produce sounds (besides occasional noises), but they are critical in musical presentations. Under the direction of different conductors, the same orchestra can produce significantly different renditions of the same musical piece. One musically interesting research problem is comparing the conducting behaviors of different conductors and analyzing their influences on the sound production of the orchestra. There has been work on conductor baton tracking [24] and gesture analysis [25] using visual analysis.

Different levels of audiovisual correspondence

Despite the various forms of music performances and analysis tasks, the common underlying idea of audiovisual analysis is to find and model the correspondence between audio and visual modalities. This correspondence can be static, i.e., between a fixed image and a short time frame of audio. For example, a certain posture of a flutist is indicative of whether the musician is playing or not; a static image of a fingering hand is informative regarding the notes being played.

This correspondence can also be dynamic, i.e., between a dynamic movement observed in the video and the fluctuation of audio characteristics. For example, a strumming motion of a guitar player's right hand is a strong indicator of the rhythmic pattern of the music passage; the periodic rolling motion of a violin player's left hand corresponds well to the pitch fluctuation of vibrato notes. Because of the large variety of instruments and their unique playing techniques, this dynamic correspondence is often instrument specific. The underlying idea of dynamic correspondence, however, is universal among different instruments. Therefore, it is appealing to build a unified framework for capturing this dynamic correspondence. If such correspondence can be captured robustly, the visual information can be better exploited to stream the corresponding audio components into sources, leading to visually informed source separation.

In the following three sections, we further elaborate upon these different levels of audiovisual correspondence by summarizing existing works and presenting concrete examples.

Static audiovisual correspondence

In this section, we first discuss works focusing on the modeling of static audiovisual correspondence in musical performances. *Static* here refers to correspondences between sonic realizations and their originating sources that remain stable over the course of a performance and for which the correspondence analysis does not rely on short-time dynamic variations. After giving a short overview with more concrete examples, a more extended case study discussion will be given on P/NP detection in instrument ensembles.

Overview

Typical static audiovisual correspondences have to do with positions and poses: which musician sits where, at what parts of the instrument the interaction occurs that leads to sound production, and how the interaction with the instrument can be characterized.

Regarding musicians' positions, when considering large ensemble situations, it is too laborious for a human to annotate

every person in every shot, especially when multiple cameras record the performance at once. At the same time, because of the typically uniform concert attire worn by ensemble members and musicians being part of large player groups that will actively move and occlude one another, recognizing individual players purely by computer vision methods is again a nontrivial problem, for which it also would be unrealistic to acquire large amounts of training data. However, within the same piece, orchestra musicians will not change positions relative to one another. Therefore, the orchestra setup can be considered as a quasi-static scene.

The work in [26] proposed to identify each musician in each camera over a full-recording timeline by combining partial visual recognition with knowledge of the scene's configuration and a human-in-the-loop approach in which humans were strategically asked to indicate the identities of performers in visually similar clusters. With minimal human interaction, a scene map was built up, and the spatial relations within this scene map assisted face clustering in crowded quasi-static scenes.

Regarding positions of interest on an instrument, work has been performed on the analysis of fingering. This can be seen as static information, as the same pressure action on the same position of the instrument will always yield the same pitch realization. Visual analysis has been performed to analyze fingering actions on pianos [11], [12], guitars [13]–[16], and violins [16], [17]. The main challenges involve the detection of the fingers in unconstrained situations and without the need to add markers to the fingers.

Case study: P/NP detection in orchestras

Whether individual musicians in large ensembles are playing their instrument or not seems to be unimportant; however, this information can be significant to critical in audiovisual analysis. Within the same instrument group, not all players may be playing at once. If this occurs in a multichannel audio recording, it is not trivial to distinguish which subset of individuals is playing, while this will visually be obvious. Furthermore, having a global overview of which instruments are active and visible in performance recordings provides useful information for audiovisual source separation.

In [9], a method was proposed to detect P/NP information in multicamera recordings of symphonic concert performances in which unconstrained camera movements and varying shooting perspectives occur. As a consequence, performance-related movement may not always be easily observed from the video, although coarser P/NP information can still be inferred through face and pose clustering.

A hierarchical method was proposed, which is illustrated in Figure 2 and that focuses on employing clustering techniques rather than learning sophisticated human–object interaction models. First, musician diarization is performed to annotate which musician appears when and where in a video. For this, keyframes are extracted at regular time intervals. In each keyframe, face detection is performed, including an estimation of the head pose angle and an inference of bounding boxes for the hair and upper body of the player. Subsequently, segmentation is performed on the estimated upper body of the musician, taking into account the gaze direction of the musician, as the instrument is expected to be present in the same direction.

After this segmentation step, face clustering methods are applied, including several degrees of contextual information (e.g., on the scene and upper body) and different feature sets, the richest ones consisting of a pyramid histogram of oriented gradients, the Joint Composite Descriptor, Gabor texture, edge histogram, and auto color correlogram.

Upon obtaining per-musician clusters, a renewed clustering is performed per musician, aiming to generate subclusters that contain images of only the same musician, performing one particular type of object interaction, recorded from one particular camera viewpoint. Finally, a human annotator action completes the labeling step: an annotator has to indicate who the musician is and whether a certain subcluster contains a playing or nonplaying action. As the work in [9] investigated various experimental settings (e.g., clustering techniques and feature sets), yielding thousands of clusters, the expected annotator action at various levels of strictness is simulated by setting various thresholds on how dominant a class within a cluster should be.

An extensive discussion of evaluation outcomes per framework module is given in [9]. Several takeaway messages can be derived from this work. First of all, the face and upper body regions are most informative for clustering. Furthermore,

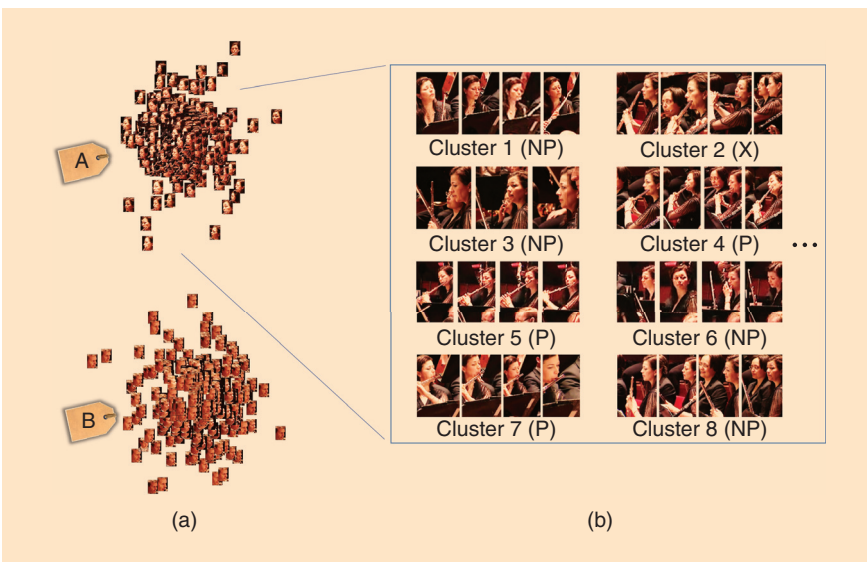

FIGURE 2. An example of hierarchical clustering steps for P/NP detection: (a) Diarization is performed on global face clustering results to identify a musician's identity. (b) Then, within each global artist cluster, subclusters are assigned with a P/NP label.

the proposed method can effectively discriminate playing versus nonplaying action, while generating a reasonable number of subclusters (i.e., enough to yield informative subclusters, but not too many, which would cause a high annotator workload). Face information alone may already be informative, as it indirectly reveals pose. However, in some cases, clustering cannot yield detailed, relevant visual analyses (e.g., subtle mouth movements for a wind player), and the method has a bias toward false positives, which can be caused by playing-anticipation movement. The application of merging strategies per instrumental part helps in increasing timeline coverage, even if a musician is not always detected. Finally, high annotator rejection thresholds (demanding clear majority classes within clusters) effectively filter out nonpure clusters.

One direct application of P/NP activity detection is in automatic music transcription. In particular, for multipitch estimation (MPE), P/NP information can be used to improve the estimation of instantaneous polyphony (i.e., the number of pitches at a particular time) of an ensemble performance, assuming that each active instrument produces only one pitch at a time. Instantaneous polyphony estimation is a difficult task from the audio modality itself, and its errors constitute a large proportion of music transcription errors. Furthermore, P/NP is also helpful for multipitch streaming (MPS), i.e., assigning pitch estimates to pitch streams corresponding to instruments: a pitch estimate should be assigned only to an active source. This idea has been explored in [20], and it was shown that both MPE and MPS accuracies are significantly improved by P/NP activity detection for ensemble performances.

Dynamic audiovisual correspondence

In a music performance, a musician makes many movements [6]. Some of these (e.g., bowing and fingering) are the articulation sources of sound while others (e.g., head shaking) are responses to the performance. In both cases, the movements show a strong correspondence with certain feature fluctuations in the music audio. Capturing this dynamic correspondence is important for the analysis of musical presentations.

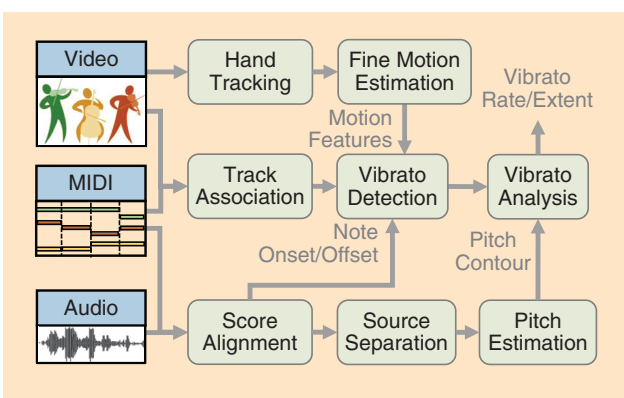

FIGURE 3. An overview of an audiovisual vibrato detection and analysis system for string instruments in ensemble performances that was proposed in [21].

Overview

Because of the large variety of musical instruments and their playing techniques, dynamic audiovisual correspondence displays different forms. In the literature, researchers have investigated the correspondence between bowing motions and note onsets of string instruments [18], between hitting actions and drum sounds of percussion instruments [10], and between left-hand rolling motions and pitch fluctuations of string vibrato notes [19], [21]. On the visual modality, object tracking and optical flow techniques have been adopted to track relevant motions, while on the audio modality, different audio features have been considered.

The main challenge lies in determining where to look for the dynamic correspondence and what to look for. This is challenging not only because the correspondence is dependent on the instrument and playing technique, but also because there are many irrelevant motions in the visual scene [6] and interferences from multiple, simultaneous sound sources in the audio signal. Almost all existing methods rely on prior knowledge of the instrument type and playing techniques to attend to relevant motions and sound features. For example, for the association between string players and score tracks, [18] captured the correspondence between bowing motions and some note onsets. This is informed by the fact that many string instrument notes are started with a new bow stroke and that different tracks often show different onset patterns. For the association of wind instruments, the onset cue is still useful, but the motion capture module would need to be revised to capture the more subtle and diverse finger movements.

Case study: Vibrato analysis of string instruments

Vibrato is an important musical expression, and vibrato analysis is important for musicological studies, music education, and music synthesis. Acoustically, vibrato is characterized by a periodic fluctuation of pitch, with a rate between 5 and 10 Hz. Audio-based vibrato analysis methods rely on the estimation of the pitch contour. In an ensemble setting, however, multipitch estimation is very challenging because of the interference of other sound sources. For string instruments, vibrato is the result of periodic change of the length of the vibrating string, which is effectuated by the rolling motion of the left hand. If the rolling motion is observable, then vibrato notes can be detected and studied with the help of visual analysis. Because such analysis does not suffer from the presence of other sound sources (barring occlusion), it offers a tremendous advantage for vibrato analysis of string instruments in ensemble settings.

In [21], an audiovisual vibrato detection and analysis system was proposed. As shown in Figure 3, this approach integrates audio, visual, and score information and contains several modules to capture the dynamic correspondence among these modalities.

The first step is to detect and track the left hand for each player using the Kanade–Lucas–Tomasi tracker. This results in a dynamic region of the tracked hand, shown as the green box in Figure 4(a). Optical flow analysis is then performed to

calculate motion velocity vectors for each pixel in this region in each video frame. Motion vectors in frame t are spatially averaged as $\mathbf{u}(t) = [u_x(t), u_y(t)]$, where u_x and u_y represent the mean motion velocities in the x and y directions, respectively. It is noted that these motion vectors may also contain the slower large-scale body movements that are not associated with vibrato. Therefore, to eliminate the body movement effects, the moving average of the signal $\mathbf{u}(t)$ is subtracted from itself to obtain a refined motion estimation $\mathbf{v}(t)$. Figure 4(c) shows the distribution of all $\mathbf{v}(t)$ across time, from which the principal motion direction can be inferred through principal component analysis, which aligns well along the fingerboard. The projection of the motion vector $\mathbf{v}(t)$ onto the principal direction is defined as the one-dimensional (1-D) motion velocity curve $V(t)$. Taking an integration over time, one obtains a 1-D hand displacement curve $X(t) = \int_0^t V(\tau)\,d\tau$, which corresponds directly to the pitch fluctuation.

To use the motion information to detect and analyze vibrato notes, one needs to know to which note the hand motion corresponds. This is solved by audiovisual source association and audio–score alignment. In this work, audiovisual source association is performed through the correlation between bowing motions and note onsets, as described in [18]. Audio–score alignment [27] synchronizes the audiovisual performance (assuming perfect audiovisual synchronization) with the score, from which onset and offset times of each note are estimated. This can be done by comparing the harmonic content of the audio and the score and dynamic time warping. Score-informed source separation is then performed, and the pitch contour of each note is estimated from the separated source signal.

Given the correspondence between the motion vectors and the sound features (pitch fluctuations) of each note, vibrato detection is performed with two methods. The first uses a support vector machine to classify each note as vibrato or nonvibrato using features extracted from the motion vectors. The second technique simply sets a threshold on the autocorrelation of the 1-D hand displacement curve $X(t)$.

For vibrato notes, the vibrato rate can also be calculated from the autocorrelation of the hand displacement curve $X(t)$. However, the vibrato extent (i.e., the dynamic range of the pitch contour) cannot be estimated by capturing the motion extent. This is because it varies based upon the camera distance and angle as well as the vibrato articulation style, hand position, and instrument type. To address this issue, the hand displacement curve is scaled to match the estimated noisy pitch contour from score-informed audio analysis. Specifically, assuming $F(t)$ is the estimated pitch contour [in a Musical Instrument Digital Interface (MIDI) number] of the detected vibrato note from audio analysis after subtracting its dc component, the vibrato extent v_e (in musical cents) is estimated as $\hat{v}_e$, with

$$\hat{v}_e = \arg\min_{v_e} \sum_{t=t^{\text{on}}}^{t^{\text{off}}} \left| 100 \cdot F(t) - v_e \frac{X(t)}{\hat{w}_e} \right|^2, \tag{1}$$

where $100 \cdot F(t)$ is the pitch contour in musical cents and $\hat{w}_e$ is the dynamic range of $X(t)$.

Music source separation using dynamic correspondence

Audio source separation in music recordings is a particularly interesting task, where audiovisual matching between the visual events of a performer's actions and their audio rendering can be of great value. Notably, such an approach enables addressing audio separation tasks that could not be performed in a unimodal fashion (solely analyzing the audio signal), as when considering two or more instances of the same instruments, say, a duet of guitars or violins, as done in the work of Parekh et al. [22]. Knowing whether a musician is playing or not at a particular point in time gives important cues for source allocation. Seeing the hand and finger movements of a cellist

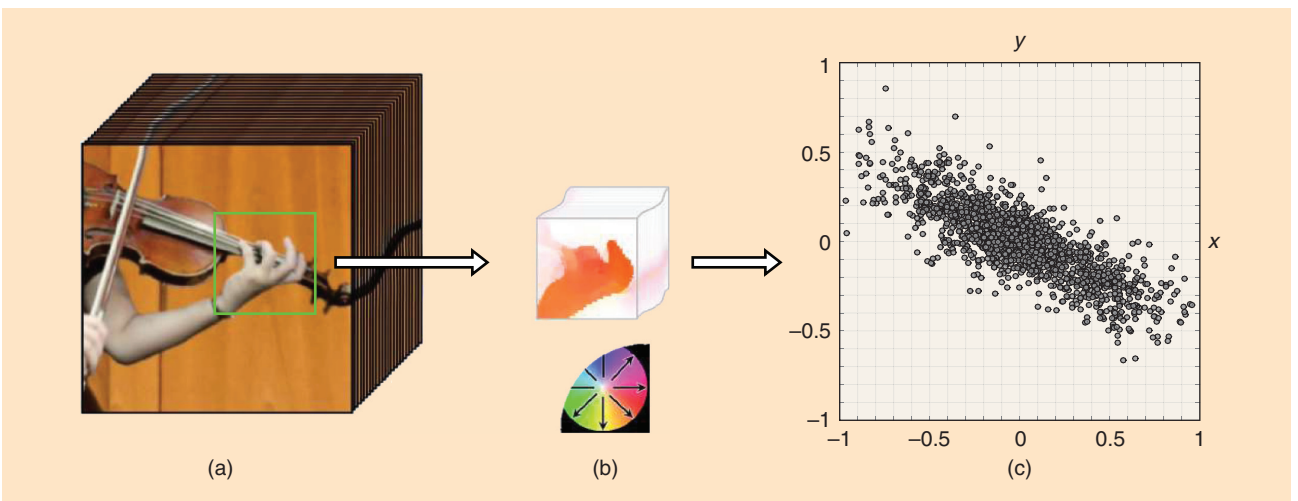

FIGURE 4. The motion capture results from (a) left-hand tracking, (b) color-encoded pixel velocities, and (c) a scatter plot of frame-wise refined motion velocities.

helps us attend to the cello's section sound in an orchestral performance. The same idea applies to visually informed audio source separation.

Overview

There is a large body of work in multimodal (especially audiovisual) source separation for speech signals, but much less effort has been dedicated to audiovisual music performance analysis for source separation. It was shown in the work of Godoy et al. [6], however, that there are certain players' motions that are highly correlated to the sound characteristics of audio sources. In particular, by analyzing a solo piano performance, the authors highlighted the correlation that may exist between music and hand movements or the sway in the upper body. An earlier work by Barzelay and Shechner [28] exploited such a correlation in introducing an audiovisual system for individual musical source enhancement in violin–guitar duets. The authors isolated audio-associated visual objects by searching for cross-modal temporal incidences of events and then used these to perform musical source separation.

Case study: Motion-driven source separation in a string quartet

The idea that motion characteristics obtained from visual analysis encode information about the physical excitation of a sounding object is also exploited in more recent studies. As an illustration, we detail a model in which it is assumed that the characteristics of a sound event (e.g., a musical note) is highly correlated with the speed of sound-producing motion [22]. More precisely, the proposed approach extends the popular nonnegative matrix factorization (NMF) framework using visual information about objects' motion. Applied to string quartets, the motion of interest is mostly carried by the bow speed. The main steps of this method are the following (see Figure 5):

1) Gather motion features, i.e., average motion speeds (further described later), in a data matrix $\mathbf{M} \in \mathbb{R}_+^{N \times C}$ that summarizes the speed information of the coherent motion trajectories within predefined regions. In the simplest case, there is one region per musician (i.e., per source). $C = \sum_j C_j$ is the number of motion clusters, where C_j is the number of clusters per source j, and N is the frame size of the short-time Fourier transform (STFT) used for computing the audio signal's spectrogram.
2) Ensure that the typical motion speeds (such as the bow speed) are active synchronously with the typical audio events. This is done by constraining the audio spectrogram decomposition obtained by NMF $\mathbf{V} \approx \mathbf{WH}$ and the motion data decomposition $\mathbf{M} \approx \mathbf{H}^\top \mathbf{A}$ to share the same activity matrix $\mathbf{H} \in \mathbb{R}_+^{K \times N}$, where $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ is the matrix collecting the so-called nonnegative audio spectral patterns (column-wise), and where $\mathbf{A} = [\alpha_1, \ldots, \alpha_C]$ gathers nonnegative linear regression coefficients for each motion cluster, with $\alpha_c = [\alpha_{1c}, \ldots, \alpha_{Kc}]^T$.
3) Ensure that only a limited number of motion clusters is active at a given time. This can be done by imposing a sparsity constraint on $\mathbf{A}$.
4) Assign an audio pattern to each source for separation and reconstruction. This is done by assigning the kth basis vector (column of $\mathbf{W}$) to the jth source, if $\text{argmax}_c \alpha_{kc}$ belongs to the jth source cluster. The different sources are then synthesized by element-wise multiplication between the soft mask, given by $(\mathbf{W}_j \mathbf{H}_j)./(\mathbf{WH})$, and the mixture spectrogram, followed by an inverse STFT, where ./ stands for element-wise division, and $\mathbf{W}_j$ and $\mathbf{H}_j$ are the submatrices of spectral patterns $\mathbf{w}_k$ and their activations $\mathbf{h}_k$ assigned to the jth source (see Figure 6).

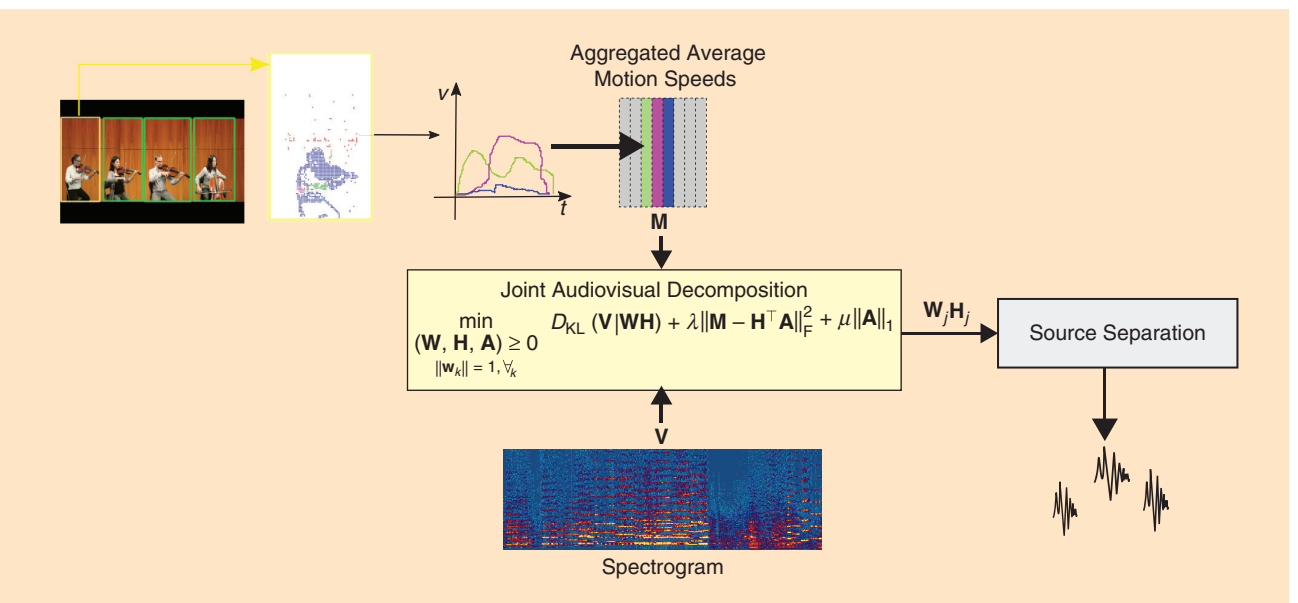

FIGURE 5. A joint audiovisual music source separation system.

A possible formulation for the complete model can then be written as the following optimization problem:

$$\underset{\substack{(\mathbf{W},\mathbf{H},\mathbf{A})\geq 0\\ \|\boldsymbol{w}_k\|=1,\ \forall k}}{\text{minimize}} \quad D_{KL}(\mathbf{V}|\mathbf{WH}) + \lambda\left\|\mathbf{M}-\mathbf{H}^{\top}\mathbf{A}\right\|_F^2 + \mu\|\mathbf{A}\|_1, \quad (2)$$

where D_{KL} is the Kullback–Leibler divergence, λ and μ are positive hyperparameters (to be tuned), and $\|\,.\,\|_F$ is the Frobenius norm.

More details can be found in [22], but for most situations, this joint audiovisual approach significantly outperformed the corresponding sequential approach proposed by the same authors and the audio-only approach introduced in [29]. For example, for a subset of the University of Rochester Multimodal Music Performance data set [7], the joint approach obtained a signal-to-distortion ratio of 7.14 dB for duets and 5.14 dB for trios, while the unimodal approach of [29] obtained signal-to-distortion ratios of 5.11 dB and 2.18 dB, respectively. It is worth mentioning that, in source separation, a difference of +1 dB is usually acknowledged as significant.

The correlation between motion in the visual modality and audio is also at the core of some other recent approaches. While bearing some similarities to the system detailed previously, the approach explained in [18] further exploits the knowledge of the MIDI score to well align the audio recording (e.g., onsets) and video (e.g., bow speeds). An extension of this work is presented in [19], where the audiovisual source association is performed through a multimodal analysis of vibrato notes. It is in particular shown that the fine-grained motion of the left hand is strongly correlated with the pitch fluctuation of vibrato notes and that this correlation can be used for audiovisual music source separation in a score-informed scenario.

Current trends and future work

This article provides an overview of the emerging field of audiovisual music performance analysis. We used specific case studies to highlight how techniques from signal processing, computer vision, and machine learning can jointly exploit the information contained in the audio and visual modalities to effectively address a number of music analysis tasks.

Current work in audiovisual music analysis has been constrained by the availability of data. Specifically, the relatively small size of current annotated audiovisual data sets has precluded the extensive use of data-driven machine-learning approaches, such as deep learning. Recently, deep learning has been utilized for vision-based detection of acoustic timed music events [23]. Specifically, the detection of onsets performed by clarinet players was addressed in this work by using a three-dimensional convolutional neural network (CNN) that relied on multiple streams, each based on a dedicated region of interest (ROI) from the video frames that was relevant to sound production. For each ROI, a reference frame was examined in the context of a short surrounding frame sequence, and the desired target was labeled as either an *onset* or *not an onset*. Although state-of-the-art audio-based onset detection methods outperform the model proposed in [23], the data set, task setup, and architecture setup gave rise to interesting research questions, especially on how to deal with significant events in temporal multimedia streams that occur at fine temporal and spatial resolutions.

Interesting ideas exploiting deep-learning models can also be found in related fields. For example, in [30] a promising strategy in the context of emotional analysis of music videos was introduced. The approach consisted in fusing learned audiovisual midlevel representations using CNNs. Another important promising research direction is transfer learning, which could better cope with the limited size of annotated

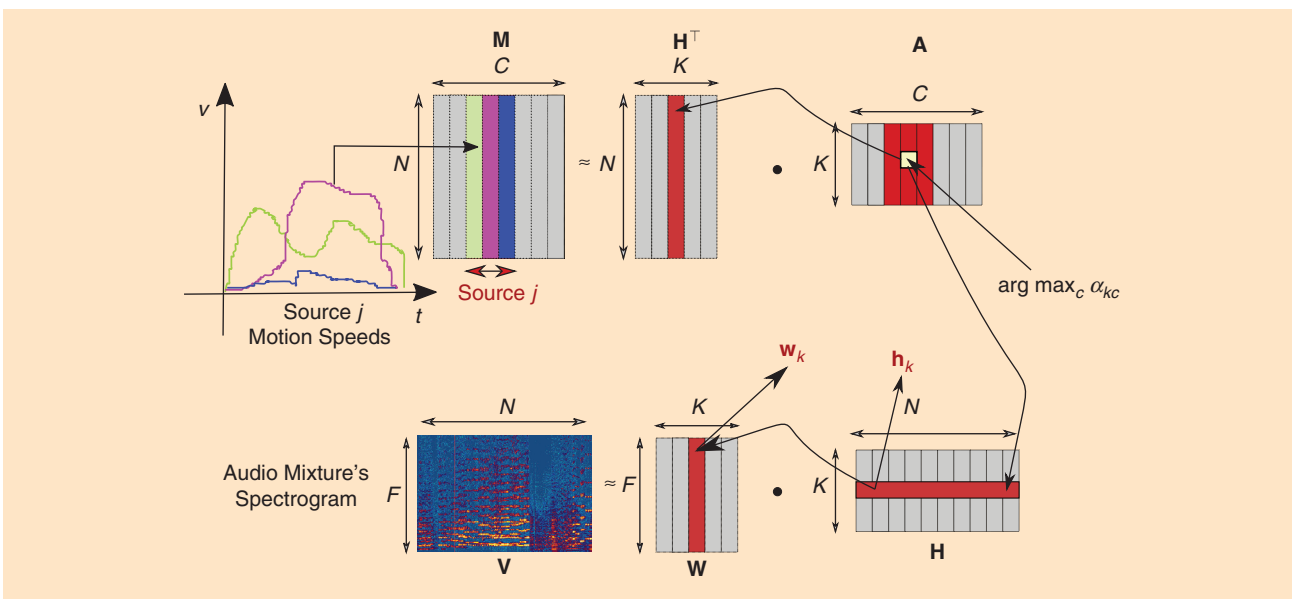

FIGURE 6. A joint audiovisual source separation—an illustration of the audio pattern assignment to source *j* (an example for the *k*th basis vector).

audiovisual musical performance data sets. As highlighted in [31], it is possible to learn an efficient audio feature representation for an audio-only application, specifically audio event recognition, by using a generic audiovisual database.

The inherent mismatch between the audio content and the corresponding image frames in a large majority of video recordings remains a key challenge for audiovisual music analysis. For instance, at a given point in time, edited videos of live performances often show only part of the performers' actions (think of live orchestra recordings). In such situations, the audiovisual analysis systems need to be flexible enough to effectively exploit the partial and intermittent correspondences between the audio and visual streams. Multiple-instance learning techniques already used for multimodal event detection in the computer vision community may offer an attractive option for addressing this challenge.

As new network architectures are developed for dealing with multimodal temporal signals and as significantly larger annotated data sets become available, we expect that deep learning-based data-driven approaches will lead to rapid progress in audiovisual music analysis, mirroring the deep-learning revolution in computer vision, natural-language processing, and audio analysis.

Beyond the immediate examples included in the case studies presented in this article, audiovisual music analysis can be extended toward other music genres, including pop, jazz, and world music. It can also help improve a number of applications in various musical contexts. Video-based tutoring for music lessons is already popular (e.g., guitar lessons on YouTube). The use of audiovisual music analysis can make such lessons richer by better highlighting the relations between the player's actions and the resulting musical effects. Audiovisual music analysis can similarly be used to enhance other music understanding/learning activities, including score-following, auto-accompaniment, and active listening.

Better tools for modeling the correlation between visual and audio modalities can also enable novel applications beyond the analysis of music performances. For example, in recent work on cross-modal audiovisual generation [32], sound-to-image sequence generation and video-to-sound spectrogram generation have been demonstrated using deep generative adversarial networks. Furthermore, the underlying tools and techniques can also help address other performing arts that involve music. Examples of such work include dance movement classification [33] and alignment of different dancers' movements within a single piece [34] by using (visual) gesture tracking and (audio) identification of stepping sounds.

Acknowledgments

Zhiyao Duan is partially supported by National Science Foundation grant 1741472.

Authors

Zhiyao Duan (zhiyao.duan@rochester.edu) received his B.S. degree in automation and his M.S. degree in control science and engineering from Tsinghua University, Beijing, in 2004 and 2008, respectively, and his Ph.D. degree in computer science from Northwestern University, Evanston, Illinois, in 2013. He is an assistant professor in the Department of Electrical and Computer Engineering and the Department of Computer Science, University of Rochester, New York. His research interest is in the broad area of computer audition, i.e., designing computational systems that are capable of understanding sounds, including music, speech, and environmental sounds. He copresented a tutorial on automatic music transcription at the International Conference on Music Information Retrieval 2015. He received a best paper award at the 2017 Sound and Music Computing Conference and a best paper nomination at the International Conference on Music Information Retrieval 2017. He is a Member of the IEEE.

Slim Essid (slim.essid@telecom-paristech.fr) received his state engineering degree from the École Nationale d'Ingénieurs de Tunis, Tunisia, in 2001, his M.Sc. (D.E.A.) degree in digital communication systems from the École Nationale Supérieure des Télécommunications, Paris, France, in 2002, his Ph.D. degree from the Université Pierre et Marie Curie (UPMC), Paris, France, in 2005, and his Habilitation à Diriger des Recherches degree from UPMC in 2015. He is a professor in Telecom ParisTech's Department of Images, Data, and Signals and the head of the Audio Data Analysis and Signal Processing team. His research interests are machine learning for audio and multimodal data analysis. He has been involved in various collaborative French and European research projects, among them Quaero, Networks of Excellence FP6-Kspace, FP7-3DLife, FP7-REVERIE, and FP-7 LASIE. He has published over 100 peer-reviewed conference and journal papers, with more than 100 distinct coauthors. On a regular basis, he serves as a reviewer for various machine-learning, signal processing, audio, and multimedia conferences and journals, e.g., a number of IEEE transactions, and as an expert for research funding agencies.

Cynthia C.S. Liem (c.c.s.liem@tudelft.nl) received her B.Sc., M.Sc., and Ph.D. degrees in computer science from Delft University of Technology, The Netherlands, in 2007, 2009, and 2015, respectively, and her B.Mus. and M.Mus. degrees in classical piano performance from the Royal Conservatoire, The Hague, The Netherlands, in 2009, and 2011, respectively. Currently, she is an assistant professor in the Multimedia Computing Group of Delft University of Technology. Her research focuses on music and multimedia content discovery through search and recommendation techniques. She gained industrial experience at Bell Labs Netherlands, Philips Research, and Google and is a recipient of several major grants and awards, including the Lucent Global Science Scholarship, Google European Doctoral Fellowship, and Netherlands Organisation for Scientific Research Veni grant. She is a Member of the IEEE.

Gaël Richard (gael.richard@telecom-paristech.fr) received his State Engineering degree from Télécom ParisTech, France, in 1990, his Ph.D. degree from the University of Paris XI, France, in 1994 in speech synthesis, and his Habilitation à

Diriger des Recherches degree from the University of Paris XI in 2001. After receiving his Ph.D. degree, he spent two years at Rutgers University, Piscataway, New Jersey, in the Speech Processing Group of Prof. J. Flanagan, where he explored innovative approaches for speech production. In 2001, he joined Télécom ParisTech, where he is now a professor in audio signal processing and the head of the Image, Data, and Signal Department. He is coauthor of more than 200 papers, His research interests are mainly in the field of speech and audio signal processing and include topics such as signal representations and signal models, source separation, machine-learning methods for audio/music signals, music information retrieval, and multimodal audio processing. He is a Fellow of the IEEE.

Gaurav Sharma (gaurav.sharma@rochester.edu) received his B.S. degree in electronics and communication engineering from the Indian Institute of Technology, Roorkee (formerly the University of Roorkee), in 1990 and his Ph.D. degree in electrical and computer engineering from North Carolina State University, Raleigh, in 1996. He is a professor in the Department of Electrical and Computer Engineering, Department of Computer Science, and Department of Biostatistics and Computational Biology, University of Rochester, New York. From 1996 to 2003, he was with Xerox Research and Technology, Webster, New York, initially as a member of the research staff and subsequently in the position of principal scientist. His research interests include multimedia signal processing, media security, image processing, computer vision, and bioinformatics. He is the editor-in-chief of *IEEE Transactions on Image Processing*. From 2011 to 2015, he was the editor-in-chief of *Journal of Electronic Imaging* and is the editor of *Color Imaging Handbook* (CRC Press, 2003). He is a fellow of SPIE and of the Society of Imaging Science and Technology and a member of Sigma Xi. He is a Fellow of the IEEE.

References

[1] C. C. S. Liem, M. Müller, D. Eck, G. Tzanetakis, and A. Hanjalic, "The need for music information retrieval with user-centered and multimodal strategies," in *Proc. Int. ACM Workshop Music Information Retrieval User-Centered and Multimodal Strategies at ACM Multimedia*, 2011, pp. 1–6.

[2] S. Essid and G. Richard. (2012). Fusion of multimodal information in music content analysis. Multimodal Music Processing. Dagstuhl, Germany: Schloss Dagstuhl-Leibniz-Zentrum für Informatik. [Online]. *3*, pp. 37–52. Available: http://drops.dagstuhl.de/opus/volltexte/2012/3465

[3] F. Platz and R. Kopiez, "When the eye listens: A meta-analysis of how audio-visual presentation enhances the appreciation of music performance," *Music Perception: Interdisc. J.*, vol. 30, no. 1, pp. 71–83, 2012.

[4] C.-J. Tsay, "Sight over sound in the judgment of music performance," *Nat. Acad. Sci.*, vol. 110, no. 36, pp. 14,580–14,585, 2013.

[5] M. S. Melenhorst and C. C. S. Liem, "Put the concert attendee in the spotlight. a user-centered design and development approach for classical concert applications," in *Proc. Int. Society Music Information Retrieval Conf.*, 2015, pp. 800–806.

[6] R. I. Godøy and A. R. Jensenius, "Body movement in music information retrieval," in *Proc. Int. Society Music Information Retrieval Conf.*, 2009, pp. 45–50.

[7] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a multi-track classical music performance dataset for multi-modal music analysis: Challenges, insights, and applications," *IEEE Trans. Multimedia*, 2018. doi: 10.1109/TMM.2018.2856090.

[8] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. (2016). YouTube-8M: A large-scale video classification benchmark. arXiv. [Online]. Available: http://arxiv.org/abs/1609.08675

[9] A. Bazzica, C. C. S. Liem, and A. Hanjalic, "On detecting the playing/non-playing activity of musicians in symphonic music videos," *Comput. Vision Image Understanding*, vol. 144, pp. 188–204, Mar. 2016.

[10] K. McGuinness, O. Gillet, N. E. O'Connor, and G. Richard, "Visual analysis for drum sequence transcription," in *Proc. IEEE European Signal Processing Conf.*, 2007, pp. 312–316.

[11] D. Gorodnichy and A. Yogeswaran, "Detection and tracking of pianist hands and fingers," in *Proc. Canadian Conf. Computer and Robot Vision*, 2006. doi: 10.1109/CRV.2006.26.

[12] A. Oka and M. Hashimoto, "Marker-less piano fingering recognition using sequential depth images," in *Proc. Korea-Japan Joint Workshop Frontiers Computer Vision*, 2013. doi: 10.1109/FCV.2013.6485449.

[13] A.-M. Burns and M. M. Wanderley, "Visual methods for the retrieval of guitarist fingering," in *Proc. Int. Conf. New Interfaces Musical Expression*, 2006, pp. 196–199.

[14] C. Kerdvibulvech and H. Saito, "Vision-based guitarist fingering tracking using a Bayesian classifier and particle filters," in *Proc. Pacific Rim Conf. Advances in Image and Video Technology*, 2007, pp. 625–638.

[15] J. Scarr and R. Green, "Retrieval of guitarist fingering information using computer vision," in *Proc. Int. Conf. Image and Vision Computing New Zealand*, 2010. doi: 10.1109/IVCNZ.2010.6148852.

[16] M. Paleari, B. Huet, A. Schutz, and D. Slock, "A multimodal approach to music transcription," in *Proc. Int. Conf. Image Processing*, 2008, pp. 93–96.

[17] B. Zhang and Y. Wang, "Automatic music transcription using audio-visual fusion for violin practice in home environment," Nat. Univ. Singapore, Tech. Rep. TRA7/09, 2009.

[18] B. Li, K. Dinesh, Z. Duan, and G. Sharma, "See and listen: Score-informed association of sound tracks to players in chamber music performance videos," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2017, pp. 2906–2910.

[19] B. Li, C. Xu, and Z. Duan, "Audiovisual source association for string ensembles through multi-modal vibrato analysis," in *Proc. Sound and Music Computing*, 2017, pp. 159–166.

[20] K. Dinesh, B. Li, X. Liu, Z. Duan, and G. Sharma, "Visually informed multi-pitch analysis of string ensembles," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2017, pp. 3021–3025.

[21] B. Li, K. Dinesh, G. Sharma, and Z. Duan, "Video-based vibrato detection and analysis for polyphonic string music," in *Proc. Int. Society Music Information Retrieval Conf.*, 2017, pp. 123–130.

[22] S. Parekh, S. Essid, A. Ozerov, N. Q. K. Duong, P. Perez, and G. Richard, "Guiding audio source separation by video object information," in *Proc. IEEE Workshop Applications Signal Processing Audio and Acoustics*, 2017, pp. 61–65.

[23] A. Bazzica, J. C. van Gemert, C. C. S. Liem, and A. Hanjalic. (2017). Vision-based detection of acoustic timed events: A case study on clarinet note onsets. arXiv. [Online]. Available: http://arxiv.org/abs/1706.09556

[24] D. Murphy, "Tracking a conductor's baton," in *Proc. Danish Conf. Pattern Recognition and Image Analysis*, 2003, pp. 1–8.

[25] Á. Sarasúa and E. Guaus, "Beat tracking from conducting gestural data: A multi-subject study," in *Proc. ACM Int. Workshop Movement and Computing*, 2014, pp. 118–123.

[26] A. Bazzica, C. C. S. Liem, and A. Hanjalic, "Exploiting scene maps and spatial relationships in quasi-static scenes for video face clustering," *Image Vision Comput.*, vol. 57, pp. 25–43, Jan. 2017.

[27] R. B. Dannenberg and C. Raphael, "Music score alignment and computer accompaniment," *Commun. ACM*, vol. 49, no. 8, pp. 38–43, 2006.

[28] Z. Barzelay and Y. Y. Schechner, "Harmony in motion," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[29] M. Spiertz and V. Gnann, "Source-filter based clustering for monaural blind source separation," in *Proc. Int. Conf. Digital Audio Effects*, 2009, pp. 1–7.

[30] E. Acar, F. Hopfgartner, and S. Albayrak, "Fusion of learned multi-modal representations and dense trajectories for emotional analysis in videos," in *Proc. 13th Int. Workshop Content-Based Multimedia Indexing*, 2015, pp. 1–6.

[31] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning sound representations from unlabeled video," in *Proc. Advances Neural Information Processing*, 2016, pp. 1–9.

[32] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *Proc. Thematic Workshops ACM Multimedia*, 2017, pp. 349–357.

[33] A. Masurelle, S. Essid, and G. Richard, "Multimodal classification of dance movements using body joint trajectories and step sounds," in *Proc. IEEE Int. Workshop Image Analysis Multimedia Interactive Services*, 2013, pp. 1–4.

[34] A. Drémeau and S. Essid, "Probabilistic dance performance alignment by fusion of multimodal features," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2013, pp. 3642–3646.

SP