SEE AND LISTEN: SCORE-INFORMED ASSOCIATION OF SOUND TRACKS TO PLAYERS IN CHAMBER MUSIC PERFORMANCE VIDEOS

Bochen Li, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma

University of Rochester, Department of Electrical and Computer Engineering

ABSTRACT

Both audio and visual aspects of a musical performance, especially their association, are important for expressing players' ideas and for engaging the audience. In this paper, we present a framework for combining audio and video analyses of multi-instrument chamber music performances to associate players in the video to the individual separated instrument sources from the audio, in a score-informed fashion. The instrument sources are first separated using a score-informed source separation techniques. The individual sources are then associated with different players in the video by correlating the onset instants of their aligned score tracks with the players' motion detected using optical flow. Experiments on 19 musical pieces with varying polyphony show that the proposed method obtains the correct association for 17 pieces, and an accuracy of 89.2% of the association of all individual tracks. The approach enables novel music enjoyment experiences by allowing users to target an audio source by clicking on the player in the video to separate/enhance it.

***Index Terms*—** Multi-modal music analysis, motion analysis, source separation, source association, audio-score alignment

1. INTRODUCTION

Music is not just an art of sound. The visual aspects of musical performances play an important role in expressing performers' ideas and emotions and in engaging the audience in live concerts and music videos. With the popularization of video streaming services, more people like to see and listen to musical performances at the same time. The coordination between the two senses enables a richer and more enjoyable experience. Current Music Information Retrieval (MIR) research, however, focuses mainly on the audio modality of musical performances, ignoring the visual aspects.

The analysis of visual aspects of musical performances can significantly advance MIR research. Challenging tasks such as instrument playing activity detection [1] and music transcription [2] in polyphonic music can be much easier to tackle by analyzing the visual aspects of the performance. It also opens up new frontiers in MIR research such as performance expressiveness analysis [3], fingering investigation [4], and conductor following [5].

A key problem in MIR is source separation and association. *Separation* is to split the audio mixture into sound source signals, while *association* is to match sound sources with players. When the visual modality is available, source association connects the audio and visual modalities, and is essential for leveraging the visual information to analyze sound sources. On the application side, 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 and to also display auxiliary information about the player and the performance. Such association can also help music editors to remix audio sources and recompose video scenes.

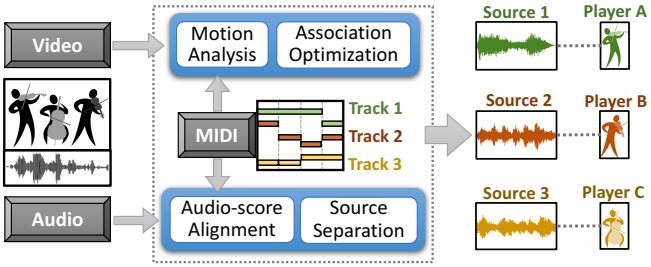

Fig. 1. Framework of the proposed system. Separated sources are associated with players in the video, encoded with different colors.

In this paper, we address the source association problem in video recordings of multi-instrument music performances by synergistic combination of analysis of the audio and video modalities. The key idea and contribution in this work lies in recognizing that *sound sources corresponding to individual instruments exhibit strong temporal interdependence with motion observed in the video in spatial regions that contain the corresponding instrument players*. Motivated by this interdependence, we propose a framework for combining source separation via audio analysis and motion video analysis to associate players in the video with corresponding audio sources. For the work we present here, we focus specifically on classical chamber music performances, where a score is typically available, allowing us to use the score to inform the analysis and the association. Our system framework is illustrated in Fig. 1. The lower half of Fig. 1 illustrates the score-informed source separation, which is adapted from our prior work [6]. The MIDI score is first aligned with the audio mixture (thus the video as well) and then used to guide the separation of audio sources. The upper part of the figure illustrates the video analysis and the exploitation of temporal interdependence to achieve the association between score tracks (thereby audio sources) and players in the video.

While our framework is general, for the initial demonstration that we present here, we focus primarily on chamber music with string instruments, where we can clearly motivate our choice of motion features and demonstrate the intuition for the association. For motion analysis in this setting, we employ a state-of-the-art optical flow technique [7] combined with a principle axes projection to capture the predominant motion of each string player as a 1D signal that we refer to as the *principal motion velocity curve*. We then find the bijection between the players and the score tracks that maximizes the coincidence between players bowing motion onset in the video, identified by zero crossings of the principal motion velocity curve, and the score tracks' note onsets. This simple method for matching associations is based on the observation that many notes in string instrument performances start with a bowing stoke. We evaluate the proposed method on 19 audio-visual musical performance

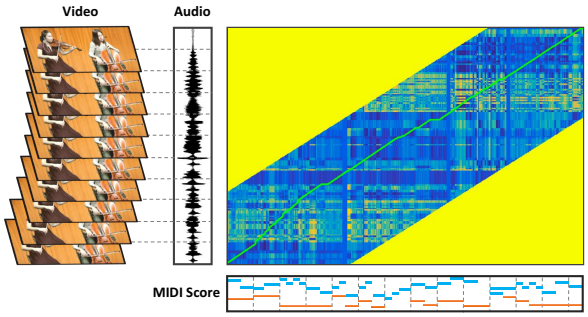

Fig. 2. Audio-score alignment which naturally results in video-score alignment. The alignment path (green line) is searched within a fixed range around the diagonal of the distance.

recordings, each containing at most one non-string instrument. The method is effective and correctly estimates the association between the separated sound sources and the players in the 16 video pieces, with 89.2% of individual tracks are correctly associated.

Results obtained from this work were previously featured in a demonstration in [8] (without an accompanying paper). This paper describes the technical approach and presents quantitative results from systematic experiments. In the following, we first briefly describe the audio analysis modules (audio-score alignment and audio source separation) in Section 2. We then describe in detail the video analysis modules that achieve source association through the bowing motion analysis of string players in Section 3.

2. AUDIO ANALYSIS

The main goal of audio analysis is to align the score with the performance. Since audio has been naturally synchronized with video in video recordings, the alignment between audio and score naturally results in the alignment between video and score as well, which is necessary for the player-source association problem in Section 3.3.

We adopt a commonly used offline Dynamic Time Warping (DTW) approach [9] based on the chroma representation [10]. We calculate a 12-D chroma feature vector sequence for both audio and score. We then calculate the Euclidean distance between each pair of the audio and score chroma vectors to derive a local distance cost matrix, as shown in Fig. 2, and use dynamic programming to search for the alignment path with the smallest overall cost. To speed up the computation, we only search alignment paths within 5 seconds around the diagonal of the distance matrix [11]. We also require the paths to be monotonic as in [9].

Another goal of audio analysis is to separate audio sources. The separated sources, together with the player-source association identified in Section 3, allows users to isolate/enhance sound sources of the players that they select in the video. With the score available and aligned well with the audio, score-informed source separation usually achieves better separation results. Various approaches to score-informed source separation have been proposed, including Non-negative Matrix Factorization [12, 13, 14], Gaussian Mixture Model [15], and adaptive instrument model [16]. We adopt the pitch-based approach of [6]. It first estimates the actually performed pitches around score-notated pitches in each frame, and then build harmonic masks that consider different overlapping harmonics cases to separate sources. Note that this score-informed source separation module does not serve for player-source association problem in this paper, as the association is handled between video and score tracks

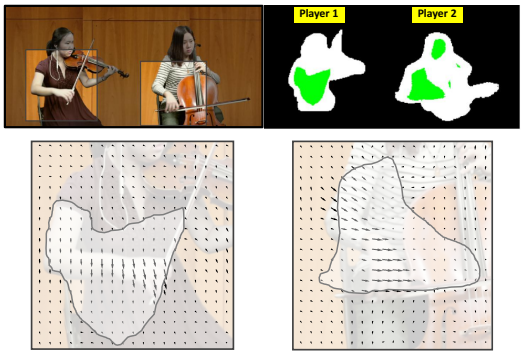

Fig. 3. A video frame (top-left), detected players in white and their high motion regions in green (top-right), and the optical flow vectors (bottom).

in Section 3. In future work, characteristics such as vibrato in the separated sources could also be used to associate with the fine motions of players in the video.

3. VIDEO ANALYSIS

The goal of the video analysis modules is to associate score tracks (hence sound sources) with players in the video. In musical performances, there is often a natural correlation between a player's motion (fingers, hands, body) and the music content (notes, beats, phrases). This correlation is especially prominent for string instruments, as many notes are started by a bow stroke, that is, many note onsets can be matched to onsets of bow strokes. We extract the bowing motion of string instrument players via optical flow estimation, and then match bow stroke onsets with note onsets in the aligned score tracks to achieve source association.

3.1. Optical Flow Estimation

Optical flow is a powerful approach to compute pixel-level motion velocities in a motion field of an image sequence [17]. We adopt a state-of-the-art method proposed by Sun *et al.* [7] that ranked first on the Middlebury optical flow benchmark [18] at the time of publication. For the audio-visual source association we need to characterize the bowing motion of the string players which often shows a high motion magnitude than other kinds of motion. For this, we first calculate the optical flow vectors for each pixel for the entire video sequence. We then sum the magnitude of the flow vectors across all frames on each pixel. This gives us a 2-D motion magnitude function of the video scene. We approximate this function using the Gaussian Mixture Model (GMM), with the assumption that each Gaussian component corresponds to one player. This provides a rough detection of the players, as shown in the white region of Fig. 3. Then we obtain a high motion region by computing a histogram of the 2-D motion magnitude function values within each player and selecting pixels with the values higher than two standard deviations above the mean, as shown in the highlighted green region in Fig. 3. Optical flow estimation result is also displayed as vectors.

3.2. Bowing Motion Capture

To capture the bowing motion of each player, we average the estimated motion velocities of all pixels in the high motion region to obtain a *global motion vector* in each video frame as $\mathbf{u}(n) =$

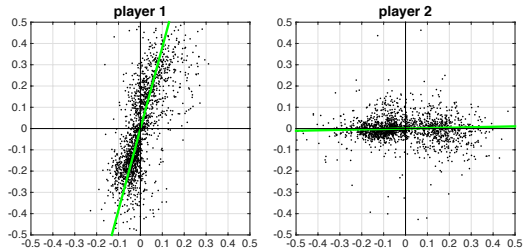

Fig. 4. Global motion vectors (dots) across all frames for the two players, showing different principal motion directions (green lines).

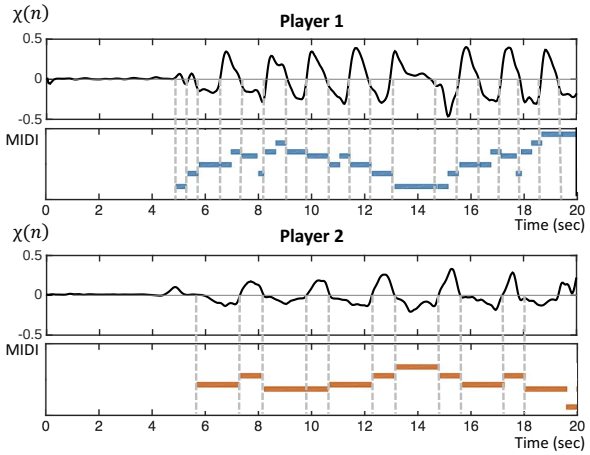

Fig. 5. Principal motion velocity curve $\chi(n)$ for the two players with aligned note sequences from the two corresponding score tracks. Dash lines show prominent correlations between their onsets.

$[u_x(n), u_y(n)]^T$, where n is the frame index. Fig. 4 shows the distribution of all the global motion vectors in the 2-D plane throughout all frames of the video of the two players in the example shown in Fig. 3. Note that for string instruments, bowing is mainly a 1-D motion. To investigate the principal motion direction of the bowing actions, we perform principal component analysis (PCA) on the global motion vectors and use the eigenvector $\tilde{\mathbf{u}} = (\tilde{u}_x, \tilde{u}_y)^T$ to represent the principal direction, plotted as the green lines in Fig. 4.

We then project the global motion vectors $\mathbf{u}(n)$ onto the principal direction to define a *principal motion velocity curve* $\chi(n)$ as

$$\chi(n) = \frac{\mathbf{u}(n)^T \tilde{\mathbf{u}}}{\|\tilde{\mathbf{u}}\|}. \tag{1}$$

This helps to suppress motions along other directions which might be due to the irrelevant body swings and instrument movements. Fig. 5 shows the principal motion velocity curves $\chi(n)$ of the two players.

3.3. Video-Score Source Association

Figure 5 also plots the notes in the corresponding score tracks of the two players, which have been aligned to the music performance during audio-score alignment. Looking at the vertical dash lines, we find that most note onsets occur around the zero crossing points of $\chi(n)$, which denotes the time instants of bowing direction changes, i.e., onsets of bow strokes. Exceptions are the legato bowing technique where a sequence of notes are played in a single bow stroke.

To reliably obtain bow stroke onsets, we firstly detect all the zero crossing points $\mathscr{Z}$ of $\chi(n)$. We then calculate the moving average of the magnitude of the principal motion velocity curve $\bar{\chi}(n)$ using a sliding window with a radius of 5 frames. A zero-crossing of $\chi(n)$ is more likely to represent a real bow stroke onset if the motion velocity magnitude around it is high. Based on this idea, we define the *bow stroke onset likelihood curve* $f(n)$ as

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

where $\delta(n, m)$ is the Kronecker delta function that equals to 1 when $n = m$ and equals to 0 otherwise. The expression in the parentheses denotes an impulse train located at the zero crossings of the principal motion velocity curve $\chi(n)$, modulated by the average velocity magnitude $\bar{\chi}(n)$. It is then convolved with a Gaussian smoothing function $\mathcal{N}(n; 0, \sigma^2)$ to tolerate timing errors of the zero crossings. The standard deviation σ is set to 3 frames.

As described above, the stroke onset function $f(n)$ often shows a good correspondence with the note onset activities for string players. We use a binary impusle train $g(n)$ to represent the note onset activities for each player, where each impulse represents a note onset from score tracks. To associate players in the video with score tracks, we match the stroke onset functions of all players with note onset functions of all score tracks and consider all their permutations. We calculate a matching score for each permutation. To define the matching score, we need to consider two kinds of mismatches between stroke onset functions and note onset functions: 1) note onsets not matching with stroke onsets due to legato bowing, and 2) stroke onsets not matching with note onsets due to irrelevant motion such as large body swing. Therefore, we define two matching functions $M^-_{p,q}$ and $M^+_{p,q}$ to pair the stroke onset function of the p-th player $f_p(n)$ with the note onset function of the q-th score track $g_q(n)$, and use their geometric mean as the final matching function:

$$\begin{cases} M^-_{p,q} = f_p(n)^T g_q(n) / \sum_m g_q(m) \\ M^+_{p,q} = f_p(n)^T g_q(n) / \sum_m f_p(m) \\ M_{p,q} = \sqrt{M^-_{p,q} \cdot M^+_{p,q}}, \end{cases} . \tag{3}$$

where $M^-_{p,q}$ resembles the recall rate of note onsets if we view the stroke onset function as an estimate of note onset activities, while $M^+_{p,q}$ resembles the precision rate.

For N players and N score tracks, there are $N!$ bijective associations (permutations). Let $\sigma(\cdot)$ be one association, then the p-th player will be associated with the $\sigma(p)$-th score track. We define an *association score* as the product of all pair-wise matching scores:

$$S_\sigma = \prod_{p=1}^{N} M_{p,\sigma(p)}, \tag{4}$$

The association σ that maximizes S_σ is selected. Although this can be done more efficiently by the Hungarian algorithm [19], we simply enumerate all associations considering the small value of N.

4. EXPERIMENTS

We evaluate the proposed score-informed audio-visual association approach on 19 pieces from the URMP dataset[1] [20] including 5 duets, 4 trios, 7 quartets, and 3 quintets. The 19 pieces were selected with a criterion that each piece contains no more than one non-string

[1] http://www.ece.rochester.edu/projects/air/projects/datasetproject.html

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

Table 1. Evaluation of the proposed player-score association method on 19 music performances. Abbreviations of instrument types are: violin (Vn.), viola (Va.), cello (Vc.), bass (D.B.), flute (Fl.), oboe (Ob.), clarinet (Cl.), saxphone (Sax.), trumpet (Tp.).

instrument. Each piece was assembled (mixed for audio and composed for video) from separately recorded but well coordinated performances of individual instrumental tracks. The background of the composed videos were replaced with a concert hall background as shown in Figure 3. Audio is sampled at 48 KHz. Video frame rate is 29.97 fps and resolution is 1080P. Multi-track MIDI scores of these pieces are also provided in the URMP dataset.

For audio analysis, we used a frame length of 42.7 ms and a hop size of 9.3 ms for the Short-Time Fourier Transform (STFT). Alignment paths are searched within 5 seconds around the diagonal alignment path. For video analysis, we downsampled the resolution to 240P. All parameters for optical flow were set the same as [7].

Table 1 shows the association results on the 19 pieces together with their metadata. The number of possible permutations increases with polyphony (2 for duets, 6 for trios, 24 for quartets, and 120 for quintets) and the association becomes more difficult. Since this is the first approach on this association problem, we do not have baseline methods to compare with. To make the evaluation more meaningful, we not only output the association that achieves the highest matching score in Eq. (4), but also rank all permutations according to their association score to see if the correct association, when it is not ranked the first, fails gracefully or with a big gap.

We propose three measures to evaluate out method's performance on each piece: 1) The number of correctly associated sources, 2) the rank of the correct association among all possible permutations, and 3) the ratio between the association score of the correct association and the highest association score of all wrong associations (named *Metric Ratio* in the table). This value will be ≥ 1 if the correct association is ranked the first by the proposed method, and a higher value indicates a larger margin of the correct association over all wrong ones. This value will be ≤ 1 if it is not ranked the first by the proposed method, and a higher value indicates a smaller gap of the correct association from the firstly ranked association.

We can see that for 17 out of the 19 pieces, all sources are correctly associated by the proposed method, i.e., the correct association is ranked the first. Overall, 89.2% of all (58 out of 65) sources of all the pieces are correctly associated. For many pieces like 2, 4-5, and 14, high metric ratios are achieved. Take the 14th piece as an example, the ratio of 9.11 far exceeds 1. This is thanks to the alternated entries of different instruments at the beginning, i.e., the Fugue style, making the players' motions quite different from each other.

Among all these results, the two pieces (No. 13 and 19) fail to output the correct association at the first rank. The main reason for the association error here in 13th piece, we argue, is the usage of wind instrument, for which the prominent motion captured by the proposed method is likely body swings (instead of finger movements) that are not highly correlated with score note onsets. This is also true for the 3rd piece, which almost fails with the metric ratio close to 1. For the 19th piece, the correct association ranks the 19th out of 120 permutations even though they are all string instruments. Further investigation reveals that it contains too much legato bowing, where most note onsets cannot be matched with bowing strokes. Note that the proposed method also almost fails on the 7th piece where all are string instruments too. This is because the three sources share very similar rhythmic patterns. For all of the above pieces, we find that the correct association is difficult to identify even for amateurish musicians.

5. CONCLUSIONS

We proposed a methodology for audio-visual source association by synergistic analyses of the audio and video modalities and demonstrated the methodology for the analysis of chamber music performance videos. Specifically, we developed a score-informed approach to model the temporal interdependence between the aligned score tracks of the instruments and the motion observed in the video of the instrument players. Experiments showed a high success rate on pieces with different polyphony. The technique enables novel and richer music enjoyment experiences that allow users to isolate/enhance sound sources by clicking on the players in the video.

6. REFERENCES

[1] A. Bazzica, C. C. Liem, and A. Hanjalic, "Exploiting instrument-wise playing/non-playing labels for score synchronization of symphonic music." in *Proc. Intl. Society for Music Information Retrieval (ISMIR)*, 2014.

[2] M. Paleari, B. Huet, A. Schutz, and D. Slock, "A multimodal approach to music transcription," in *Proc. Intl. Conf. Image Process. (ICIP)*, 2008.

[3] D. Radicioni, L. Anselma, and V. Lombardo, "A segmentation-based prototype to compute string instruments fingering," in *Proc. Conf. Interdisciplinary Musicology (CIM)*, 2004.

[4] J. Scarr and R. Green, "Retrieval of guitarist fingering information using computer vision," in *Proc. Intl. Conf. Image and Vision Computing New Zealand (IVCNZ)*, 2010.

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

[6] Z. Duan and B. Pardo, "Soundprism: An online system for score-informed source separation of music audio," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1205–1215, 2011.

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

[8] B. Li, Z. Duan, and G. Sharma, "Associating players to sound sources in musical performance videos," in *Late Breaking Demo, Intl. Society for Music Information Retrieval (ISMIR)*, 2016.

[9] N. Orio and D. Schwarz, "Alignment of monophonic and polyphonic music to a score," in *Proc. Intl. Computer Music Conf. (ICMC)*, 2001.

[10] T. Fujishima, "Realtime chord recognition of musical sound: A system using common lisp music," in *Proc. Intl. Computer Music Conf. (ICMC)*, 1999.

[11] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Process*, vol. 26, no. 1, pp. 43–49, 1978.

[12] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, 2011.

[13] J. Fritsch and M. D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, 2013, pp. 888–891.

[14] S. Ewert and M. Müller, "Using score-informed constraints for nmf-based source separation," in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, 2012, pp. 129–132.

[15] P. Sprechmann, P. Cancela, and G. Sapiro, "Gaussian mixture models for score-informed instrument separation," in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, 2012, pp. 49–52.

[16] F. J. Rodriguez-Serrano, Z. Duan, P. Vera-Candeas, B. Pardo, and J. J. Carabias-Orti, "Online score-informed source separation with adaptive instrument models," *Journal of New Music Research*, vol. 44, no. 2, pp. 83–96, 2015.

[17] D. Fleet and Y. Weiss, "Optical flow estimation," in *Handbook of mathematical models in computer vision*. Springer, 2006, pp. 237–257.

[18] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *Intl. Journal of Computer Vision*, vol. 92, no. 1, pp. 1–31, 2011.

[19] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[20] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a musical performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Trans. Multimedia*, submitted. Available: https://arxiv.org/abs/1612.08727.