VISUALLY INFORMED MULTI-PITCH ANALYSIS OF STRING ENSEMBLES

Karthik Dinesh⋆*, Bochen Li*⋆*, Xinzhao Liu*†*, Zhiyao Duan, and Gaurav Sharma*

Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY, USA

ABSTRACT

Multi-pitch analysis of polyphonic music requires estimating concurrent pitches (estimation) and organizing them into temporal streams according to their sound sources (streaming). This is challenging for approaches based on audio alone due to the polyphonic nature of the audio signals. Video of the performance, when available, can be useful to alleviate some of the difficulties. In this paper, we propose to detect the play/non-play (P/NP) activities from musical performance videos using optical flow analysis to help with audio-based multi-pitch analysis. Specifically, the detected P/NP activity provides a more accurate estimate of the instantaneous polyphony (i.e., the number of pitches at a time instant), and also helps with assigning pitch estimates to only active sound sources. As the first attempt towards audio-visual multi-pitch analysis of multi-instrument musical performances, we demonstrate the concept on 11 string ensembles. Experiments show a high overall P/NP detection accuracy of 85.3%, and a statistically significant improvement on both the multi-pitch estimation and streaming accuracy, under paired t-tests at a significance level of 0.01 in most cases.

***Index Terms*—** Multi-pitch estimation, streaming, audio-visual analysis, source separation, constrained clustering, SVM classifier

1. INTRODUCTION

Multi-pitch analysis of polyphonic music is important in many music information retrieval (MIR) tasks including automatic music transcription, music source separation, and audio-score alignment. It can be performed at different levels: *Multi-pitch Estimation (MPE)* is to estimate concurrent pitches and the number of pitches (polyphony) in each time frame; *Multi-pitch Streaming (MPS)* goes one step further to also assign the pitch estimates to different sound sources.

There exist various audio-based methods for multi-pitch analysis. For MPE, methods include auto-correlation [1] and Bayesian inference [2] in the time-domain, and harmonic amplitude summation [3] and peak/non-peak modeling [4] in the frequency domain. For MPS, methods often rely on modeling the timbre of sound sources to organize pitch estimates. Supervised methods, which learn timbre models from isolated training excerpts of sources, employ Bayesian models [5], hidden Markov models [6], and probabilistic latent component analysis (PLCA) [7]. Unsupervised methods that infer timbre models of sound sources directly from the mixture audio are also proposed [8, 9, 10]. The common idea is to cluster pitch estimates that have similar timbre features into the same stream, while the clustering process is often aided by constraints that model the locality relations between pitches.

These state-of-the-art audio-based methods, however, cannot achieve satisfactory performance for many applications as yet. This is due to the core challenge that polyphonic audio signals have: signals of different sound sources mix together and interfere with each other. More specifically, multi-pitch estimation needs to estimate the number of mixed sound sources at each time instant (instantaneous polyphony). This is difficult for audio-based approaches due to large variety of harmonic relations and timbre combinations of concurrent pitches. Furthermore, even if the instantaneous polyphony were correctly estimated in each frame, identifying which sources are active in these frames for the estimated pitches to assign to is also challenging purely from audio. These issues, however, could be alleviated when videos are available. Specifically, availability of video can help identify Play/Non-play (P/NP) activities of instrument players, helping with the estimation of the instantaneous polyphony and the detection of active sound sources for pitches to be assigned.

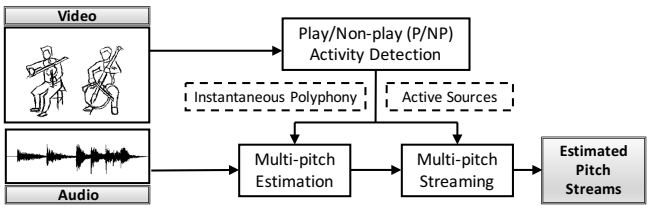

Fig. 1. Proposed framework for enhancing multi-pitch analysis using video-based play/non-play activity detection.

Advances in the field of multimodal signal analysis have propelled the use of visual features along with audio features to solve a variety of problems like information retrieval [11], multimedia content authoring [12], sentiment analysis [13], shot change detection [14], and audio-visual feature extraction [15]. In the field of music performance analysis, visual information has been exploited to detect instrument playing activities in an orchestra for audio-score alignment [16]. Video analysis has also been employed to track the fret-board and movement of hands to transcribe guitar performances [17]. However, to date, there is remarkably little visually informed work on the fundamental problem of multi-pitch analysis.

In this paper, we build upon our prior work on audio-based MPE [4] and MPS [9] to propose the first method that leverages visual information for multi-pitch analysis of string ensembles. Fig.1 shows the system overview. The video analysis module detects players as well as their P/NP activity through an optical-flow-based motion analysis in each video frame. The instantaneous polyphony of each audio frame is then derived from the P/NP activity and is used to inform the audio-based MPE module. The P/NP information is also passed to the MPS module so that estimated pitches are only allowed to be assigned to active players in each frame. Experiments on 11 string ensembles show that the proposed video-based P/NP detection achieves a high overall accuracy of 85.3%. The incorporation of these detected P/NP results to our audio-based baselines results in a statistically significant improvement on both the MPE and the MPS accuracy at a significance level of 0.01 in most cases.

⋆Karthik Dinesh and Bochen Li made equal contribution to this paper.
†Xinzhao Liu is currently with Listent American Corp. We thank CIRC, University of Rochester for providing computational resources for the project.

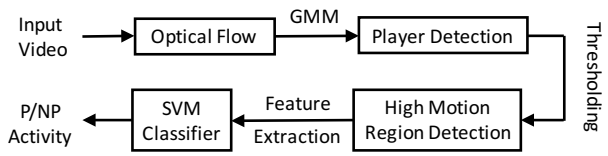

Fig. 2. Video analysis for P/NP activity detection.

2. PROPOSED METHOD

2.1. Play/Non-play Activity Detection

We employ optical flow estimation and supervised classification to detect P/NP activities of players in each video frame. Figure 2 summarizes the analysis workflow.

2.1.1. Optical Flow Estimation

Optical flow [18], which estimates the motion field using the observed pattern of brightness displacements from frame to frame, forms the basis of our motion analysis. The assumption that the brightness is preserved as the pixels get displaced due to motion in the scene, yields the classic optical flow equation [19]

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0, \tag{1}$$

where (x, y) represent the spatial (pixel) coordinates, t denotes time, and $I = I(x, y, t)$ denotes the observed spatio-temporal pattern of image intensity and $(u, v) = [\frac{dx}{dt}, \frac{dy}{dt}]$ represents the flow vector in horizontal (x) and vertical (y) directions. The collection of flow vectors over spatial extent of the frame forms the motion field $u_t(x, y), v_t(x, y)$, where t indexes the frames. Optical flow techniques estimate the flow field for each frame by minimizing an energy function that combines a data term based on (1) with regularization terms that ensure smoothness of the flow field. In this paper, we adopt the approach of [20] which improves upon the classical objective function of [19, 21] by incorporating flow and image boundary information in the regularization function and provides highly accurate motion field estimations.

2.1.2. Player Detection

Instrument play movements are typically the dominant motion in the video. For a given camera viewpoint, these movements localize in spatial regions corresponding to individual players across the span of temporal frames. We therefore identify distinct regions of significant motion in the video to estimate the locations of the players. Specifically, we compute a temporally aggregated motion magnitude function for each spatial location as the sum of the optical flow motion field magnitudes over the frames. The motion magnitude function is modeled as a mixture of Gaussians, and a rough estimate of the locations of the players is obtained by identifying the spatial locations that associate with individual components (with high probability).

2.1.3. High Motion Region Detection

Within the spatial regions corresponding to a string player's movements, pixels and time intervals with high motion (e.g., the bowing hand) are indicative of the P/NP activity. Therefore, we detect high motion regions (pixels) from the initial estimate of individual player locations obtained in the previous step. Given that different players (instruments) exhibit different degrees of motion in the video, we use an adaptive threshold on the motion magnitude for each player, using a threshold equal to the temporal mean + twice standard deviation of the histogram of the flow vector magnitude. A sample frame

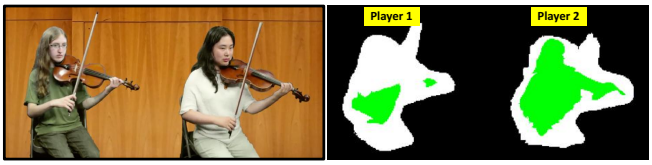

Fig. 3. Sample frame from the performance video (left) and the player detection results (right) with the detected high motion regions in green, detected players in white and the background in black.

is shown in Fig. 3 where we can see clusters of high motion regions (green) within the detected players (white) who are detected from the background (black).

2.1.4. Feature Extraction and Classification

We train a support vector machine classifier (SVM) to classify P/NP activities of each player in each video frame. A 20-dimensional feature vector is extracted from the motion vectors of the detected high motion pixels for each player. These features include: (a) Mean, variance, and standard deviation of the flow vectors in x and y directions separately. (b) Mean, variance, and standard deviation of the flow vector magnitude. (c) Sum of the motion vector magnitude in each region of the frame characterizing the total amount of motion. (d) The major directions of the motion present in each region, characterized by the eigenvectors and eigenvalues of the Principal Component Analysis. (e) Statistics from the Gray Level Co-occurrence Matrix (GLCM), which is obtained from the flow vector magnitude in the high motion region detected in the previous step. The statistics include (1) Energy, measuring the orderliness or regularity of flow vector magnitude, (2) Correlation, measuring the joint probability of the occurrence of flow vector magnitudes, (3) Contrast, measuring local variation of flow vector magnitude, and (4) Homogeneity, measuring similarity of flow vector magnitudes.

To train the SVM, we collect solo string performances that are distinct from the test set. The ground-truth P/NP labels for the training pieces are obtained from audio-based single-pitch detection results followed by manual corrections: If a pitch is detected in the audio of a video frame, then the frame is annotated as Play; otherwise, it is annotated as Non-play. To parameterize the SVM training algorithm we, (a) set the kernel function parameter to radial basis function kernel (RBF), (b) set the kernel scale parameter to automatic scaling. The relative amount of play/non-play classes in the training data varied from 76%-80% for play labels and 20%-24% for non-play labels.

2.2. Multi-pitch Estimation

2.2.1. Audio-based MPE

The proposed method is built upon an audio-based method proposed in [4]. It is a maximum-likelihood approach modeling both spectral peaks and non-peak regions of the audio frame to be analyzed. Assume that the audio frame has N monophonic sound sources and let θ be a set of N fundamental frequencies. Fundamental frequency of each source is estimated by maximum likelihood estimation,

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \left(O|\theta \right), \tag{2}$$

where Θ denotes the space of all possible sets of fundamental frequencies, and O is the observation, i.e., magnitude spectrum. This method estimates pitches in an iterative way from more prominent pitches to less prominent ones using a greedy strategy. After pitches and polyphony are estimated in each frame, a post-processing step is

employed to smooth the estimates within several neighboring frames to remove inconsistent estimation errors.

2.2.2. Visually Informed MPE

An important disadvantage of this audio-based MPE method (and other audio-based methods as well) is that the instantaneous polyphony estimation is not that accurate. For low-polyphony pieces (e.g., duets and trios) it tends to overestimate, while for high-polyphony pieces (e.g., quartets and above) it tends to underestimate. This is due to the polyphonic nature of string ensembles and the harmonic relations among the sources. The P/NP labels detected from the visual scene, on the other hand, can provide a more accurate estimation for the instantaneous polyphony. We therefore count the number of active players in each video frame and use it to replace the audio-based polyphony estimates in the corresponding audio frames. To account for the possible errors in the P/NP detection in individual frames, we still adopt the post-processing module to refine the pitch and polyphony estimates within neighboring frames.

2.3. Multi-pitch Streaming

2.3.1. Audio-based MPS

We also build the visually informed MPS algorithm upon our prior audio-based MPS framework [9]. It formulates the MPS problem as a constrained clustering problem. It takes pitch estimates in individual frames (MPE results) as input and clusters them into different pitch streams. Two kinds of constraints are considered: must-links and cannot-links. Must-link constraints are added to pitches that are close in both time and frequency. Cannot-link constraints are used to prevent assigning pitches in the same time frame to the same source. Pitches from the same source have similar timbre features, so the objective function is designed to minimize the timbre inconsistency of pitches within the same stream as

$$f\left(\Pi\right)=\sum_{m=1}^{M}\sum_{t_i\in S_m}\left\|t_i-c_m\right\|^2, \tag{3}$$

where Π is the clustering of pitches, M represents the number of monophonic sound sources, t_i denotes the timbre feature vector of the i-th pitch, and c_m is the centroid of timbre features in stream S_m.

An iterative algorithm was proposed to solve this constrained clustering problem in [9]. After an initialization, in each iteration the clustering is updated to decrease the objective function while satisfying all the constraints that have already been satisfied. The new clustering is found through a *swap operation*: swapping cluster labels of two streams within a *swap set*. The swap set is defined as a connected subgraph of the pitches between two clusters using the already satisfied constraints as edges. If the swap operation is accepted (i.e., it decreases the objective function), then the set of already satisfied constraints is updated to all constraints satisfied by the new clustering. The algorithm was proven to converge.

2.3.2. Visually Informed MPS

To design the visually informed MPS system, we inject the P/NP information obtained from the video into the audio-based framework to prevent assigning pitches to non-playing players. The algorithm is described in Algorithm 1, where '*' indicates the changes from the audio-based algorithm [9] for incorporation of the P/NP information.

As shown in the algorithm, the P/NP information is incorporated at two places. First, during the clustering initialization, estimated pitches are sorted in a descending order and are assigned to only the active performers from high-pitched instruments to low-pitched instruments. Second, when updating the clustering through the swap operations, only swaps that satisfy the P/NP constraints are accepted. The satisfaction criterion is that for each source in the swap set, among the frames that the source has a pitch after the swap operation, at least 50% of the frames are labeled as Playing according to the P/NP information. This criterion prevents the algorithm from assigning too many pitches to an inactive source during the clustering update processing. We chose this threshold to account for possible errors in the P/NP detection results. As a preliminary study, we did not investigate the effect of this parameter on the MPS results.

Algorithm 1: Visually informed MPS algorithm. '$*$' indicates places of the incorporation of the P/NP information.

M - the number of monophonic sound sources
PNP - the binary P/NP matrix indicating which player is playing at which frame (1-playing, 0-not playing)

begin
  *** Initialization:**Assign pitches to only active players in the pitch-descending order; $t \leftarrow 0$;
  repeat
    $t \leftarrow t+1$;
    $f_{max} \leftarrow f(\Pi_{t-1})$;
    $\Pi_t \leftarrow \Pi_{t-1}$;
    while $f_{max} == f(\Pi_{t-1})$ & *not all pitches* $p_1, ..., p_N$ *are traversed* **do**
      Randomly pick p_n which is in stream S_m and not be replaced;
      for $j = 1 : M$ **do**
        Find the swap set of p_n between S_m and S_j;
        *** if** *PNP is satisfied in the swap set* **then**
          Do the swap to get a new clustering Π_s;
          if $f(\Pi_s) < f_{max}$ **then**
            $f_{max} \leftarrow f(\Pi_s)$;
            $\Pi_t \leftarrow \Pi_s$;
          end
        end
      end
    end
    C_T = constraints satisfied by Π_t;
  until $\Pi_t = \Pi_{t-1}$;
  Return Π_t and C_t;
end

3. EXPERIMENTS

3.1. Dataset

We evaluate the proposed system on the URMP dataset[1] [22]. Each piece was assembled (mixed for audio and composed for video) from isolated recorded but well coordinated performances of individual instrumental tracks. We selected all the string-instrument (violin, viola, cello, and bass) pieces which include 3 duets, 2 trios, 4 quartets and 2 quintets. Video files are downsampled to 240P for optical flow estimation. Note that as an initial demonstration here we only evaluate on a few pieces due to the lack of large audio-visual datasets.

3.2. Evaluation of Play/Non-play Activity Detection

Since we have only 11 videos, we adopt a training strategy where the piece to be evaluated is considered a test piece and the remaining

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

Piece No.	P/NP Detection Accuracy (%)					MPE Accuracy (%)		
	P1	P2	P3	P4	P5	Audio	Video PNP	GT PNP
# 1	97.4	91.5	-	-	-	70.2	83.6	85.1
# 2	93.6	93.3	-	-	-	68.7	72.2	74.2
# 3	81.1	71.3	-	-	-	58.5	62.7	70.0
# 4	92.5	91.4	78.4	-	-	59.8	65.9	68.6
# 5	93.9	92.9	89.4	-	-	75.0	76.7	79.0
# 6	83.4	88.4	78.6	73.4	-	49.5	52.3	56.3
# 7	69.3	73.6	75.1	70.1	-	52.1	52.0	59.0
# 8	90.0	90.9	84.6	86.4	-	62.2	62.3	66.6
# 9	93.1	95.5	82.4	91.5	-	62.2	63.3	65.7
# 10	91.9	92.3	88.5	94.1	91.2	47.4	52.3	53.3
# 11	74.2	75.1	70.0	75.3	62.5	46.4	44.0	48.8

Table 1. Results of video-based Play/Non-play detection and MPE accuracy of the 11 test pieces.

videos are considered as the training set from which the features are extracted to form training and test feature matrix whereas for the training and test labels we use the ground truth P/NP information from the annotated audio file. The training feature matrix with the training label is fed into the SVM training algorithm to develop a model which is used on the test feature matrix to get the predicted labels and the predicted labels are compared with test labels to find a match which is used as a measure of accuracy. The left half of Table 1 shows the video-based P/NP detection accuracy for all 11 videos. We can see good match between predicted labels and the ground truth test labels which has resulted in an average accuracy level 85.3% for the pieces. For piece #7 and #11, as we can observe, the accuracy has decreased because of the limited bowing motion due to the nature of the composition of the two pieces. Higher the accuracy, higher is the probability of an improvement on multi-pitch estimation and streaming accuracy.

3.3. Evaluation of Multi-pitch Estimation

For audio analysis, we first evaluate the multi-pitch estimation results using the MPE accuracy measure proposed in [23] with the error tolerance of one quarter tone. The right half of Table 1 lists all of the 11 testing pieces and compares the audio-based baseline method with the proposed visually informed method. For almost all of the testing pieces, the proposed method achieves prominent improvements (13% on the first piece) based on the audio-based method. The MPE accuracy drops when the polyphony number increases, and the improvements become less pronounced when the P/NP detection accuracy decreases. For 7th and 11th pieces the proposed approach even drops from baseline method due to the relatively low accuracy of P/NP detection, which supports our analysis in Sec. 3.2. We further add another testing group where the ground-truth P/NP labels are incorporated into the MPE process. This gives a upper bound of the MPE accuracy improvement by using a perfect visual activity detection module.

To further prove the effectiveness of the proposed approach, we also create more subsets using the 11 pieces for a statistical evaluation. We arrange all possible track combinations within each piece. For the example of a quartet, we can further arrange 6 duets and 4 trios using the 4 original tracks. This operation on all of the 11 pieces totally results in 53 duets, 38 trios, 14 quartets and 2 quintets, on which the average increase of the MPE accuracy from the baseline method to the proposed visual-based method is 3.71%. We group all these subsets by polyphony (excluding the 2 quintets) and show the boxplot of MPE accuracy in Fig. 4. The improvements of the first two polyphony groups are statistically significant under a paired t-test with $p < 10^{-19}$ and $p < 10^{-3}$, respectively. The improvement for quartets is not statistically significant under the same test at the significance level of 0.05.

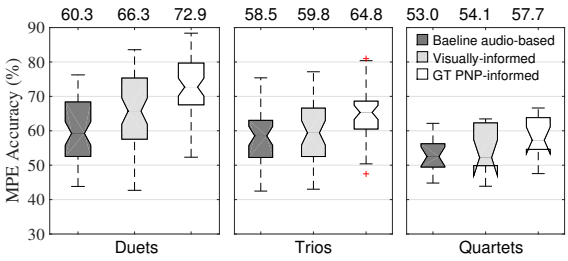

Fig. 4. Boxplot of MPE accuracy grouped by polyphony on all subsets derived from the 11 pieces, comparing the baseline audio-based method (dark gray), proposed visually informed method (light gray), and the incorporation of ground-truth PNP labels (white). The number above each box shows the mean value of the box.

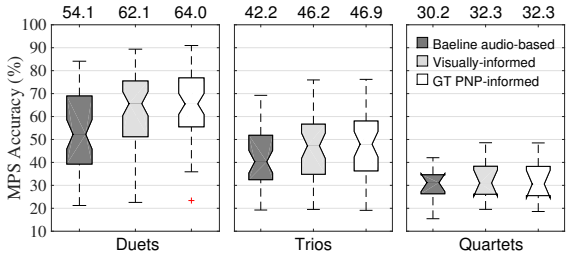

Fig. 5. Boxplot of MPS accuracy grouped by polyphony on all subsets derived from the 11 pieces, comparing the baseline audio-based method (dark gray), proposed visually informed method (light gray), and the incorporation of ground-truth PNP labels (white). The number above each box shows the mean value of the box.

3.4. Evaluation of Multi-pitch Streaming

We evaluate the proposed visually informed MPS system on the same derived track combinations from the 11 pieces. The criterion for a pitch to be considered correctly streamed needs to satisfy both the frequency deviation condition and the stream assignment condition: it should deviate less than a quarter tone from the ground-truth pitch in the stream that it is assigned to [9]. Fig. 5 shows the boxplot of the MPS accuracy of the three comparison methods on three polyphony groups. It can be seen that the visually informed approach improves over the audio-based baseline consistently over all three groups, reaching close to the ground-truth P/NP-informed upper bound. A paired t-test shows that the improvement is statistically significant for all groups at a significance level of 0.01. Further analyses show that the improvement is more pronounced when the pieces have a layered structure (up to 30% improvement), i.e., different tracks come and go at different times. This is intuitive as the this is when the the video-based P/NP detection is most informative for streaming.

4. CONCLUSION AND DISCUSSION

In this paper, we propose and demonstrate a framework for visually-informed multi-pitch analysis of string ensembles. The play/non-play activity of different players is detected from analysis of the video and incorporated into the techniques used for multi-pitch estimation (MPE) and multi-pitch streaming (MPS). Our results demonstrate that, in most cases, the proposed incorporation of visual information offers statistically significant improvements (under a paired t-test) in pitch analysis accuracy over purely audio-based approaches.

5. REFERENCES

[1] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 708–716, 2000.

[2] M. Davy, S. Godsill, and J. Idier, "Bayesian analysis of polyphonic western tonal music," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2498–2517, 2006.

[3] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in *Proc. Intl. Society for Music Information Retrieval (ISMIR)*, 2006, pp. 216–221.

[4] Z. Duan, B. Pardo, and C. Zhang, "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 8, pp. 2121–2133, 2010.

[5] E. Vincent, "Musical source separation using time-frequency source priors," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 1, pp. 91–98, 2006.

[6] M. Wohlmayr, M. Stark, and F. Pernkopf, "A probabilistic interaction model for multipitch tracking with factorial hidden markov models," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 799–810, 2011.

[7] M. Bay, A. F. Ehmann, J. W. Beauchamp, P. Smaragdis, and J. S. Downie, "Second fiddle is important too: Pitch tracking individual voices in polyphonic music." in *Proc. Intl. Society for Music Information Retrieval (ISMIR)*, 2012, pp. 319–324.

[8] K. Hu and D. Wang, "An unsupervised approach to cochannel speech separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 1, pp. 122–131, 2013.

[9] Z. Duan, J. Han, and B. Pardo, "Multi-pitch streaming of harmonic sound mixtures," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 138–150, 2014.

[10] V. Arora and L. Behera, "Multiple F0 estimation and source clustering of polyphonic music audio using PLCA and HMRFs," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 2, pp. 278–287, 2015.

[11] F.-F. Kuo, M.-K. Shan, and S.-Y. Lee, "Background music recommendation for video based on multimodal latent semantic analysis," in *Proc. IEEE Intl. Conf. Multimedia and Expo (ICME)*, 2013, pp. 1–6.

[12] O. Gillet, S. Essid, and G. Richard, "On the correlation of automatic audio and visual segmentations of music videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 3, pp. 347–355, 2007.

[13] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, Part A, pp. 50–59, 2016.

[14] S. Essid and G. Richard, "Fusion of multimodal information in music content analysis," in *Multimodal Music Process.*, ser. Dagstuhl Follow-Ups, 2012, vol. 3, pp. 37–52.

[15] E. Acar, F. Hopfgartner, and S. Albayrak, "Understanding affective content of music videos through learned representations," in *Proc. Intl. Conf. Multimedia Modeling*. Springer, 2014, pp. 303–314.

[16] 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, pp. 201–206.

[17] M. Paleari, B. Huet, A. Schutz, and D. Slock, "A multimodal approach to music transcription," in *Proc. IEEE Intl. Conf. Image Process. (ICIP)*, 2008, pp. 93–96.

[18] Y. Wang, J. Ostermann, and Y.-Q. Zhang, *Video processing and communications*. Prentice Hall Upper Saddle River, 2002, vol. 5.

[19] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[20] 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, pp. 2432–2439.

[21] M. J. Black and P. Anandan, "A framework for the robust estimation of optical flow," in *Proc. Intl. Conf. Computer Vision*, 1993, pp. 231–236.

[22] 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.

[23] S. Dixon, "On the computer recognition of solo piano music," in *Proc. Australasian Computer Music Conf.*, 2000, pp. 31–37.