LOOKING BEYOND SOUND: UNSUPERVISED ANALYSIS OF MUSICIAN VIDEOS

Cynthia C. S. Liem, Alessio Bazzica and Alan Hanjalic

Multimedia Information Retrieval Lab
Delft University of Technology, The Netherlands

ABSTRACT

In this work, we focus on visual information conveyed by performing musicians. While musicians are playing, their movement relates to their musical performance. As such, analysis of this information can support structural characterization and timeline indexing of a recorded performance, especially in cases when such analyses are not trivially computed from the musical audio. We propose an unsupervised visual analysis method, in which visual novelty is inferred from motion orientation histograms of regions of interest. Considering our method in a case study on audiovisually recorded jam sessions, we show that our analysis of the visual channel yields promising and meaningful performance-related information, including information complementary to the audio channel.

1. INTRODUCTION

When making music, a musician will move. In general, muscular action is needed in order to have an instrument producing the desired musical sounds. On top of this, the musician's experience of the played piece may trigger additional movement, or influence the movement necessary for sound production. Hence, when looking at a performing musician, there will be visual cues regarding developments over the course of the performed musical piece, which will influence an audience member's perception of the musical interpretation [1].

Traditionally, audio analysis is employed to characterize the timeline of a recorded musical piece. A common way to detect new or novel events over time, such as the occurrence of structural boundaries, is to compute frame-level features such as Mel-Frequency Cepstral Coefficients (MFCCs) or chromagrams, followed by a self-similarity analysis [2]. However, in order for these features to give convincing results, they should show sufficient variation throughout the recording. This is not the case for music in which timbre and harmonic content do not develop much throughout a piece. This e.g. frequently happens in jam sessions, when a fixed chord scheme is followed and differentiation between musical sections is based on alternating improvised solo parts.

In this work, we aim at taking a step forward regarding this problem by considering movement of musicians over the course of their performance. We study this information based on video data, since in practical situations, the visual channel is a straightforward modality to record in a non-obtrusive way.

Our approach is driven by the interest to find generally applicable movement descriptors, allowing for overall unsupervised timeline indexing of a performance, characterizing highlights and sectional changes over the course of the performance, and supporting or complementing information on this as obtained from audio analysis. As such, our techniques are meant to ultimately support non-linear access scenarios.

This paper is outlined as follows: in Section 2, we discuss related work. We then present our visual analysis approach and its rationale in Section 3. After this, we present the data used for our current case study in Section 4, after which results are discussed in Section 5. Finally, general conclusions and an outlook to future work are presented in Section 6.

2. RELATED WORK

Many existing studies on characteristics of music-related movement have involved gestural analysis. For example, Wanderley [3] and Caramiaux et al. [4] investigate the consistency and parsing of ancillary gestures (i.e., gestures related to the instrument which are not caused because of sound production) by instrumental musicians. Studies on music-induced motion in listeners rather than performers have e.g. been conducted by Nymoen et al. [5]. Typically, for gesture-oriented work 3D motion capture data is used, and due to the generated large amount of sensor data, detailed analysis can often just be feasibly performed on short excerpts. As pointed out by Godøy and Jensenius [6], in this direction of work, video processing methods for extracting features of music-related body movement still are generally lacking.

An exception is the work started by Gillet and Richard in [7], and expanded in McGuinness et al. [8], where the aim was to transcribe drum sequences from video recordings of performers. Given this goal, the focus was on classification of highly instrument-specific events. As mentioned before, in our current work, we aim at taking a more general perspective on visual information conveyed by performing musicians. Therefore, we will not focus on classification of specific events, nor will we explicitly strive to only analyze ancillary or expressive movements.

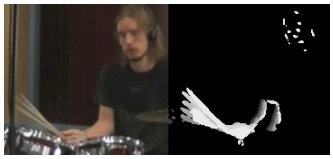


Fig. 1. Motion History Image of a drummer in performance.

3. VISUAL ANALYSIS

Given a set of video recordings of performing musicians, we aim to extract a series of visual novelty points over time, relating to the temporal development of the musical performance. We do not wish to depend on a specific set of instruments, and wish to be as flexible as possible regarding characteristic individual motion patterns of musicians. Therefore, we focus on analyzing motion patterns rather than the shape of objects, without restricting to a pre-defined vocabulary of motion patterns, nor attempting to establish such an explicit vocabulary. Regarding the video setup, we require that the video was recorded with a stationary camera, but do not put any specific restriction on positioning of players, instruments and cameras otherwise. The only further requirements are that moving objects are not completely occluded, and that the motion is not uniquely occurring along a spectator's line of sight.

We aim at efficiently detecting any moving object with relation to the music (thus, objects associated to the player or his instrument). Many such objects may move at the same time, and each object can move in a specific direction. For instance, a drummer can hit a snare drum while triggering the hi-hat with the pedal. In order to encode a variable number of moving objects moving towards any direction, we choose to detect region of interests (ROIs) adopting the approach of Bradski et al. [9]. First, recent motion is encoded as a *motion-history image* (MHI) accumulating thresholded frame differences (e.g. see Fig. 1). Then, each MHI is segmented according to an iterative algorithm called *downward stepping floodfill*: the most recent motion is progressively connected to the older through a sequence of gradient descent steps. When two different objects are moving, they usually lead to two different floodfill regions. Each of these is used as ROI and encoded as a silhouette mask moving towards a specific direction. Inspired by Davis [10], we iterate over the extracted ROIs to build a histogram of motion orientations per frame (see Fig. 2). The bins are assigned quantizing the orientation in 12 sectors, and the area is measured as the number of silhouette pixels within the ROI. To encode temporal development, we then apply a 2-sec. sliding window, summing together the histograms of frames within the window into a single 12-bin vector. The resulting vector is expanded adding a 'no motion flag' which is set to 1 if all the bins are zero, i.e. no ROI has been extracted in the past 2 sec. (see Fig. 3).

Based on the summed histogram features, we wish to detect significant visual motion variations over time. For this,

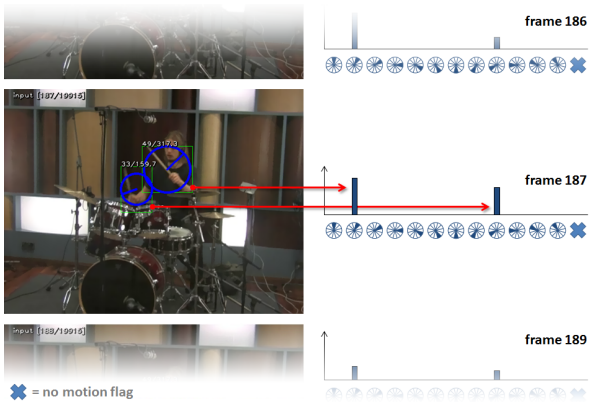


Fig. 2. Motion Orientation Histograms.

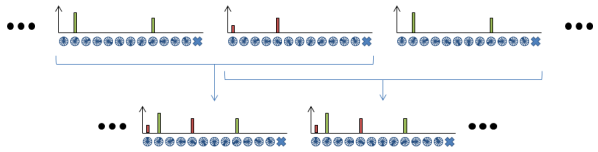


Fig. 3. Encoding sequences of motion pattern descriptors

two properties are particularly useful, the first being 'classical' novelty: the measure of temporal change as proposed by Foote [2], obtained from self-similarity matrix analysis employing a checkerboard kernel. For this, we employ the cosine distance, and use a Gaussian checkerboard kernel of 20 sec. to compare motion patterns in the near past and future[1]. The second useful property is the overall amount of motion, which simply can be computed by summing together the contributions of the 12 histogram bins at each point in time.

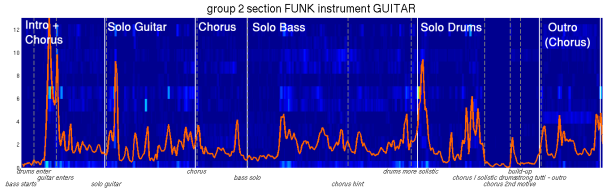
Finally, we compute our visual novelty curve by combining the degree of 'classical' novelty with the degree of motion. For this, we normalize both measures, and simply multiply them frame-by-frame. This results in a function which peaks when there is a lot of novelty and a lot of motion.

4. DATA

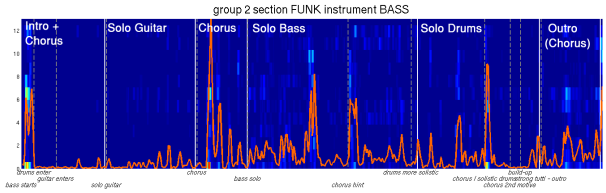
For our current study, we use a dataset with multi-track studio recordings of live performances in swing, blues and funk styles, released by Abeßer et al. [11]. The dataset consists of multi-track recordings of 3 combos of 3 musicians, playing guitar, bass guitar and drums. Each of the combos is recorded during a session in which swing, blues and funk styles are performed. For each of the styles, improvised solo parts occur[2], which are annotated in the dataset. Together with the multi-track audio recordings, video material is released with

[1] The choice for this wide kernel is chosen to favor coarse development over short-time details.

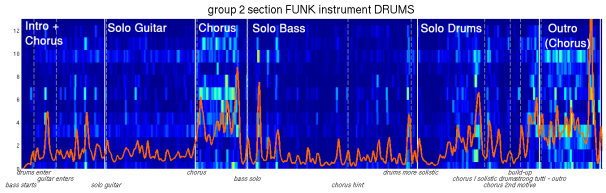
[2] As such, there will not be a notated score, and every performance will create a new piece that was not played before.



(a) Motion histograms and visual novelty: guitar



(b) Motion histograms and visual novelty: bass



(c) Motion histograms and visual novelty: drums

Fig. 4. Motion features (histograms with visual novelty overlayed) for group 2, Funk style session. Vertical white lines indicate structural boundaries as annotated in the original dataset. Vertical dashed grey lines indicate additional notes as described in Footnote 4.

the dataset, showing each of the musicians during their performance in a single, static shot.

Regarding the recorded data, we cut out the excerpts from the sessions which actually corresponded to the featured styles, removing breaks and intermediate talking, thus retaining 70 min. of recorded material. We then manually synchronized the video and audio streams in the dataset, ensuring that any possible temporal deviation remained under 0.5 sec.

The multi-track audio recordings consisted of many separate audio tracks: 1 for the bass, 2 for the guitar, and 6 for the drums. As we do not assume that so many tracks per instrument will be available in future work beyond this case study, we mixed[3] these together for each instrument, and as a full mix involving all instruments.

The original dataset provided structural session annotations at the full-second resolution. Respecting this resolution, we corrected annotated boundaries to have them start and end with an acoustic event. To allow deeper analysis, we made additional manual annotations, marking every 4 beats ('a bar') and the starts of repeated chord schemes or cells ('a cycle')[4].

[3] In all cases, we mixed together the tracks by simply adding them up, and correcting the peak level to be at 0.0 dB.

[4] We release these annotations at `http://homepage.tudelft.nl/04d13/wiamis2013.html`.

		Blues	Funk	Swing
Group 1	Bass	0.073	-0.088	-0.245
	Guitar	0.350	0.251	-0.081
	Drums	0.280	0.382	0.475
Group 2	Bass	0.151	-0.088	-0.179
	Guitar	0.248	0.184	0.047
	Drums	0.325	0.518	0.538
Group 3	Bass	-0.0507	-0.012	-0.201
	Guitar	0.095	0.210	-0.052
	Drums	0.204	0.196	0.612

Table 1. Pearson's correlation coefficient for visual novelty with onset intensity of instrument audio mix.

5. RESULTS AND DISCUSSION

While the dataset in our current study is small (but rich), the observed behavior of our proposed analysis method with regard to this dataset is promising. Over time, the motion orientation histograms and their derived visual novelty show explainable patterns with respect to the structural annotations. Furthermore, they give indications of internal development throughout a performance, even if timbre and instrumentation will not vary much over the course of the piece.

A good illustration of this can be seen in Fig. 4, which shows motion histograms and overlayed visual novelty graphs for the instrumentalist videos of the second combo in the dataset, playing in Funk style[5]. This particular Funk session was striking, since it was entirely based on a one-bar, continuously repeated cell in the bass guitar. Despite this constant foundation, the movement behavior of the instrumentalists is not uniform. Peaking behavior in the novelty curves intensifies when an instrumentalist has the solo role.

In order to verify to what extent our visual novelty curve reflects information already present in the audio channel, we wished to compare our visual novelty curves to a representative audio-based descriptor. As, due to our data genre, timbre- or harmony-based descriptors would not be as suitable as usual, we chose a more low-level feature. This feature was computed by running an onset detector on each mixed audio track, and then summing the energy contributions of every detected onset peak per second in the recording. We considered the resulting onset intensity feature to be a reasonable approximation of the auditory event density in the tracks.

For every video recording, we computed Pearson's correlation coefficient between the visual novelty curve and two onset intensity vectors: the vector computed from the instrument-specific mixed audio track, and the vector computed from the full ensemble-mixed audio track. Results for instrument-specific tracks are shown in Table 1, while those for full ensemble mixes are shown in Table 2.

From the correlation values, we can conclude that the visual novelty information is largely complementary to onset

[5] Additional illustrations and examples for other recordings in the dataset are given at the website mentioned in Note 4.

		Blues	Funk	Swing
Group 1	Bass	0.113	-0.103	-0.138
	Guitar	0.367	0.130	-0.022
	Drums	0.263	0.299	0.232
Group 2	Bass	0.170	0.034	0.098
	Guitar	0.315	0.140	0.015
	Drums	0.376	0.508	0.492
Group 3	Bass	-0.068	0.111	0.117
	Guitar	0.195	0.170	0.048
	Drums	0.213	0.250	0.562

Table 2. Pearson's correlation coefficient for visual novelty with onset intensity of full audio mix.

intensity information, with the exception of the drums player. This is explainable, since the drums player cannot move a lot beyond direct interaction with the instrument. From similar reasoning, the generally poor correlation of the bass guitar player with the onset intensity information can be explained: the truly instrument-related movement on a bass guitar is more subtle than other movement made by the player, such as foot-tapping along with the music. While the latter action is not causing sound production, it is synchronized to the music and a sign of entrainment, and as such still related to the music jointly made by the ensemble. While it needs further investigation, we conjecture that this can be an explanation that correlation coefficients with the onset intensities of full ensemble mixes are generally higher than those computed for the individual instrumental mixes.

We noted over multiple recordings that novelty peak maxima indicate major body movement, such as a posture change. Once again, these changes are often in sync with the music, and are related to events in the performance (e.g. picking up a plectrum for an intensified solo part). However, they cannot be fully discerned from truly incidental movement yet, so this needs further consideration in follow-up work.

6. CONCLUSIONS AND FUTURE WORK

We presented unsupervised visual analysis techniques for videos of performing musicians. Our initial observations show that explainable results are yielded, which can characterize events and entrainment throughout a performance for different players, and in certain cases complement audio channel information.

One of our priorities in future work will be to establish more quantitative strategies to evaluate feature performance. A clear-cut ground truth does not exist for this type of performance characterization: we do not aim to detect exact structural boundaries, but to clarify development and variation, and find novel events within these boundaries. It is challenging but very interesting to devise appropriate measures for this.

Having seen promising initial results on jam session data, we now plan to expand our analysis to performances in other genres, with larger numbers of musicians. An attractive property of using video data is that it takes physical space into account: even in a video featuring multiple musicians at the same time, it is straightforward to study sub-groups or individuals, by just defining an appropriate subset of pixels. This is less trivial for audio data: if contributions of sub-groups within an ensemble are to be studied there, these need to be have been recorded in separate tracks, or source separation techniques will have to be applied. We therefore hope to find more opportunities to use visually-based analysis to support and enhance musical performance analysis.

Acknowledgements Cynthia Liem is a recipient of the Google European Doctoral Fellowship in Multimedia, and this research is supported in part by this Google Fellowship. Additionally, the research leading to these results has received funding from the European Union Seventh Framework Programme FP7 / 2007–2013 through the PHENICX project under Grant Agreement no. 601166.

7. REFERENCES

[1] B. W. Vines, M. M. Wanderley, C. L. Krumhansl, R. L. Nuzzo, and D. J. Levitin, "Performance gestures of musicians: What structural and emotional information do they convey?," in *Gesture-Based Communication in Human-Computer Interaction*, A. Camurri and G. Volpe, Eds., vol. 2915 of *Springer LNCS*, pp. 468–478. Springer, 2004.

[2] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *Proc. IEEE ICME*, 2000, pp. 452–455.

[3] M. M. Wanderley, "Quantitative analysis of non-obvious performer gestures," in *Gesture and Sign Language in Human-Computer Interaction*, I. Wachsmuth and T. Sowa, Eds., vol. 2298 of *Springer LNCS*, pp. 241–253. Springer, 2002.

[4] B. Caramiaux, M. M. Wanderley, and F. Bevilacqua, "Segmenting and parsing instrumentalist's gestures," *J. New Music Research*, vol. 41, no. 1, pp. 13–29, April 2012.

[5] K. Nymoen, B.Caramiaux, M. Kozak, and J. Torresen, "Analyzing sound tracings: a multimodal approach to music information retrieval," in *Proc. ACM MIRUM*, Scottsdale, Arizona, 2011, pp. 39–44.

[6] R. I. Godøy and A. R. Jensenius, "Body movement in music information retrieval," in *Proc. ISMIR*, Kobe, Japan, October 2009, pp. 45–50.

[7] O. Gillet and G. Richard, "Automatic transcription of drum sequences using audiovisual features," in *Proc. ICASSP*. IEEE, 2005, vol. 3, pp. iii–205.

[8] K. McGuinness, O. Gillet, N.E. O'Connor, and G. Richard, "Visual analysis for drum sequence transcription," in *Proc. EUSIPCO*, 2007, pp. 312–316.

[9] G. R. Bradski and J. W. Davis, "Motion segmentation and pose recognition with motion history gradients," *Mach. Vis. Appl.*, vol. 13, no. 3, pp. 174–184, 2002.

[10] J. W. Davis, "Recognizing movement using motion histograms," *Technical Report 487, MIT Media Lab*, 1999.

[11] J. Abeßer, O. Lartillot, C. Dittmar, T. Eerola, and G. Schuller, "Modeling musical attributes to characterize ensemble recordings using rhythmic audio features," in *Proc. ICASSP*, Praha, Czech Republic, May 2011, pp. 189–192.