A multimodal approach for percussion music transcription from audio and video

Bernardo Marenco, Magdalena Fuentes, Florencia Lanzaro, Martín Rocamora, and Alvaro Gómez *

Facultad de Ingeniería, Universidad de la República
{bmarenco,mfuentes,carla.florencia.lanzaro,rocamora,agomez}@fing.edu.uy

Abstract. A multimodal approach for percussion music transcription from audio and video recordings is proposed in this work. It is part of an ongoing research effort for the development of tools for computer-aided analysis of Candombe drumming, a popular afro-rooted rhythm from Uruguay. Several signal processing techniques are applied to automatically extract meaningful information from each source. This involves detecting certain relevant objects in the scene from the video stream. The location of events is obtained from the audio signal and this information is used to drive the processing of both modalities. Then, the detected events are classified by combining the information from each source in a feature-level fusion scheme. The experiments conducted yield promising results that show the advantages of the proposed method.

Keywords: multimodal signal processing, machine learning applications, music transcription, percussion music, sound classification

1 Introduction

Although music is mainly associated with an acoustic signal, it is inherently a multimodal phenomenon in which other sources of information are involved, for instance visual (images, videos, sheet music) and textual (lyrics, tags). In the recent decades this has lead to research on multimodal music processing, in which signal processing and machine learning techniques are applied for automatically establishing semantic relationships between different music sources [12]. This has several applications, such as sheet music to audio synchronization or lyrics to audio alignment [12], audiovisual musical instrument recognition [11] and cross-modal correlation for music video analysis [6]. Even thought the problem of music transcription has received a lot of attention [10], most existing research focus on audio signals. Among the few works that take multimodal information into account, drum transcription is tackled in [7] using audio and video.

In this work, a multimodal approach for percussion music transcription is proposed. It is tailored to the analysis of audio and video recordings of Candombe drumming performances, with the aim of determining the location of sound

* This work was supported by funding agencies CSIC and ANII from Uruguay.

events and their classification into a set of different stroke types. A comparison between the performance attained by monomodal classifiers (i.e. audio or video) and the multimodal approach is considered, so as to evaluate the impact of including different sources of information. The ultimate goal of this research is to contribute with automatic software tools for computer-aided music studies.

Candombe drumming is the essential component of an afro-rooted tradition which has evolved for almost two centuries and is at present one of the most characteristic features of Uruguayan popular culture [1]. The rhythm is played marching on the street by groups of drums of which there are three different types: *chico*, *repique* and *piano* [4]. The drumhead is hit with one hand and with a stick in the other, producing different types of sound. The stick is also used to hit the shell of the drum which is made of wood.

The rest of this paper is organized as follows. Next section outlines the proposed multimodal approach for percussion transcription. Then, Section 3 introduces the datasets used for developing and evaluating the system. Section 4 is devoted to explaining the signal processing techniques applied to the audio and video streams. The feature selection techniques and the classification scheme adopted are described in Section 5. Experiments and results are presented in Section 6. Finally, the paper ends with some concluding remarks.

2 Proposed multimodal approach

The information coming from the different modalities can be integrated in various ways, for instance by combining the outputs of monomodal classifiers (i.e. decision-level fusion) or by considering the features from each mode into an unified set for classification (i.e. feature-level fusion). In this work, high-level information obtained from one modality is used to drive the processing of both modalities, which as noted in [3] can be a better strategy than merely relying on direct feature-level or decision-level fusion. This is based on the fact that the location of events can be obtained quite reliably using the audio signal alone (known as onset detection [2]), whereas the classification of the type of stroke can be better attained by exploiting the information from both modalities. This is illustrated in Figure 1, where a stroke performed with the stick is shown. The detection of events relies on the Spectral Flux (see Sec. 4.1) of the audio signal and exhibits a clear maximum, while the position of the stick extracted from the video shows a corresponding minimum indicating a stick stroke.

The proposed system is depicted in the diagram of Figure 2. Onset detection guides the feature extraction in both modalities. In order to obtain meaningful information from the video, relevant objects in the scene have to be segmented, namely the stick, the drumhead and the hand. After extracting features from each modality, a feature selection stage is introduced separately, which proved to be more effective than a single selection over the whole feature set. Finally, detected events are classified by combining the information from both modalities in a feature-level fusion scheme.

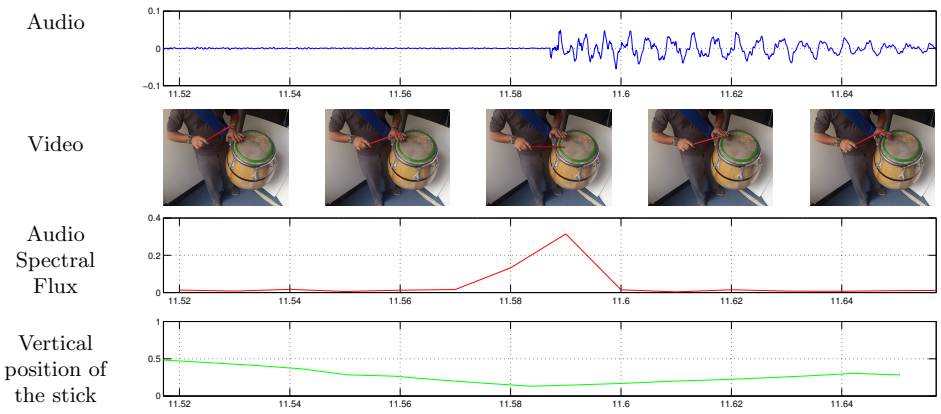

Fig. 1. Example of multimodal information extracted. Event detection is based on the audio signal, but classification exploits information from both modalities.

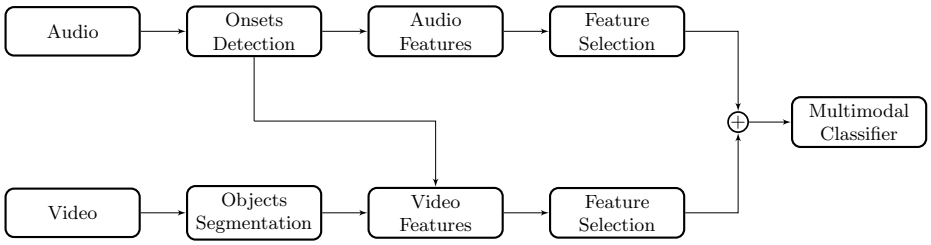

Fig. 2. Block diagram of the proposed multimodal transcription system. The symbol ⊕ represents the fusion of the features from each modality into an unified set.

3 Datasets

Two recording session were conducted in order to generate the data for this research. High-quality equipment was used for recording audio. Since the performances involve very fast movements, video acquisition was carried out at high-speed rates (120 and 240 fps). Four percussionists took part in the first session, and each of them recorded two improvised performances of the *repique* drum. In the second session, several different configurations of solo drums and ensembles were recorded, performed by five renowned Candombe players.

Recordings from the first dataset were manually labeled by an expert, indicating the location of the strokes and their corresponding type. This data was used for training and evaluating the classification system. Recordings of the second session were mainly used for developing the object segmentation and feature extraction algorithms.

The following six different stroke classes were considered for a *repique* performance: *wood*, in which the stick hits the wooden shell of the drum; *hand*, when

the bare hand hits the drumhead; *stick*, that corresponds to a single hit of the drumhead with the stick; *bounce*, in which the stick hits the drumhead several times in a short time interval; *rimshot*, with the stick hitting tangentially to the drumhead; and *flam*, which consists of two single strokes played almost together by alternating the hand and stick (in any order).

4 Signal processing

4.1 Audio signal processing

Onsets detection The first step in the proposed system is to automatically find the location of sound events. A typical approach is adopted in this work, namely the Spectral Flux, which emphasizes changes in the energy content of the audio signal in different frequency bands [2]. The Short-Time Fourier Transform of the signal is calculated and mapped to the MEL scale (approximately linear in log frequency) for sequential 20 -ms duration windows in hops of 10 ms. The resulting sequences are time-differentiated (via first-order difference), half-wave rectified, and summed along the MEL sub-bands. Finally, a global threshold of 10% of the maximum value is applied to the obtained detection function to determine the location of onsets.

Audio features The type of sound, i.e. the *timbre*, is mainly related to the energy distribution along its spectrum. Therefore, two different sets of features commonly used for describing spectral timbre of audio signals were adopted. In addition, two features were proposed to capture the behaviour of strokes which involve several events in a short time interval. A temporal window of 90 ms centered at the detected onset is considered for computing each type of feature.

Spectral shape features A feature set commonly used for general-purpose musical instruments classification was considered, comprising several measures describing the shape of the spectral energy distribution, namely: spectral centroid, spectral spread, spectral skewness, spectral kurtosis, spectral decrease, spectral slope and spectral crest [10].

Mel-frequency Cepstral Coefficients (MFCC) These are the most widespread features in speech and music processing for describing spectral timbre. A filter bank of 160 bands is applied to the signal frame, whose center frequencies are equally-spaced according to the MEL scale. Then an FFT is calculated and the log-power on each band is computed. The elements of these vectors are highly correlated so a Discrete Cosine Transform is applied and the 40 lowest order coefficients are retained.

Spectral Flux features The *bounce* and *flam* strokes involve several events in a short time interval. Therefore, the number of Spectral Flux peaks and the amplitude difference between the first and second peak (set to zero if no second peak exist) are also computed to capture the behaviour of these type of strokes.

4.2 Video signal processing

In order to analyze a video recording and to extract features for describing the performance, it is necessary to automatically detect the most important objects that appear in the scene. These are the drumhead, the stick and the hand of the performer. The scene was slightly prepared to simplify the detection and the evaluation process, but taking care not to alter the sound of the drums or disturb the performer. The stick and the contour of the drumhead were painted to ease their segmentation. Besides, in the second recording session some fiducial paper markers were pasted to the drum and the body of the performer for evaluation purposes. This can be seen in Fig. 3(a), which shows a video frame that is used for illustrating the objects segmentation and detection process hereafter.

Objects segmentation and detection

Drumhead detection While the drumhead is a circle, it becomes an ellipse in the image due to the camera perspective projection. Detecting ellipses in images is a classical problem [13] usually tackled in two steps: (a) edge detection on the image and (b) estimation of the ellipse that best fits the obtained edges. As mentioned before, the drumhead contour was painted, so the first step is achieved by color filtering the image. The fitting of the best ellipse was performed based on [5], using the OpenCV implementatin. Color filtering and ellipse fitting for drumhead detection are shown in Fig. 3(b) and Fig. 3(c), respectively.

Stick detection The detection of the stick is also carried out in a two-step way. Only relying in color filtering turned out to be insufficient for segmenting the stick due to the lighting conditions. Since the stick is one of the fast moving objects in the scene, a background subtraction algorithm [14] was used together with the color cue to point out pixel candidates. Then, the stick is identified from the filtered pixels by a line segment detector [8], also implemented in OpenCV. To assure a continuous detection of the stick, coherence is imposed by restricting the segment detection in a frame to a window determined by the movement in previous frames. The moving objects mask is depicted in Fig. 3(d) and the detected stick in Fig. 3(e).

Hand detection The detection of the hand that hits the drum is addressed by segmenting the main skin blob above the drumhead. Skin segmentation cannot be accomplished by a simple thresholding of the color space, because there are other regions with similar chromaticity, such as the wooden floor and the drum. To overcome this difficulty, a permissive color filtering is followed by a tree classifier which recognizes skin pixels in the YCbCr color space. Once the skin pixels are identified, some morphological operations are applied, and the bounding box of the largest blob within a certain region above the previously detected drumhead is selected. Kalman filtering of the detections is also applied to impose temporal coherence. The hand detection process is illustrated in Fig. 3: output of the color filter (f), skin detection based on tree classifier (g), mask above the drumhead (h), and the bounding box of the hand blob (i).

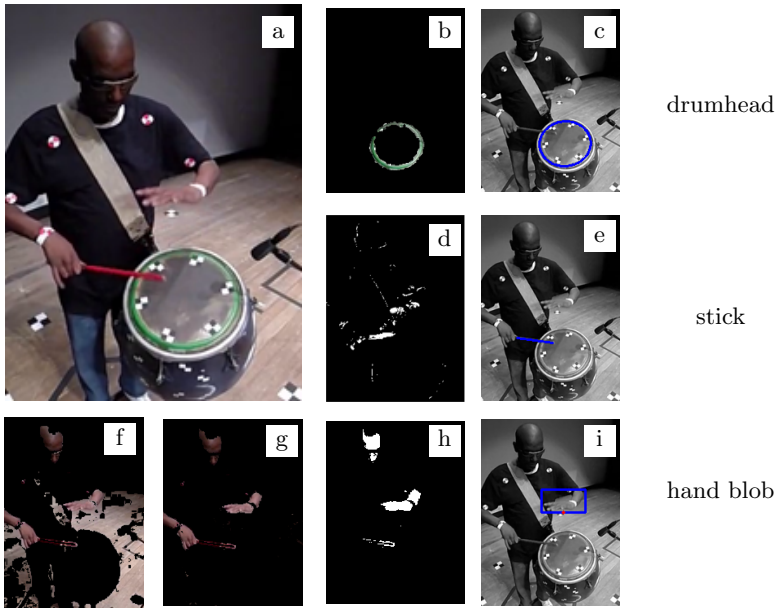

Fig. 3. Objects detection for a video frame (a). Drumhead: color filter (b) and ellipse fitting (c). Stick: moving objects mask (d) and linear segments detection (e). Hand blob: color filter (f), skin detection (g), mask above drumhead (h), bounding box (i).

Video features Based on the position of the drumhead, the stick and the hand blob, several features are devised to describe the type of movement of these objects during an audio event. Features are computed within a time window centered at the onset, as shown in Fig. 1. The extracted features for both the stick and hand blob are: the normalized distance to the drumhead, the maximum and minimum value of vertical speed, the zero crossings of vertical speed, and the first ten coefficients of the Discrete Cosine Transform of the vertical position.

5 Multimodal classification

Feature selection Feature selection was carried out within each modality independently, before the feature-level fusion of both sets. This turned out to be more effective than a single selection over the whole feature set. To do that, a correlation-based feature selection method was adopted [9], considering 10-fold cross-validation (CV). For the audio modality 37 features where selected out of 49, and for the video modality 14 out of 30, for a complete set of 51 features.

Classification Based on the performance attained by different classification techniques in preliminary tests, a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel was selected for the implemented system. Optimal values for parameters C and γ were grid-searched in a CV scheme.

6 Experiments and results

The database of labeled *repique* performances was used for stroke classification experiments. There are two recordings for each of the four performers, for a total of 4132 strokes. To test the generalization ability of the system, three performers were considered for training (6 recordings) and one for testing (2 recordings). This was repeated in a CV scheme in which each performer is considered once for testing. As shown in Table 1, the multimodal approach outperforms the monomodal systems in each fold of the CV. Besides, the advantage is also noticeable for each type of stroke, which is presented in Table 2-left. Notice that the poor classification rate attained by the audio modality for performer 1 (Table 1), resulting from a different tuning of the drum, is effectively compensated in the multimodal method. In addition, confusion matrix of the multimodal approach is shown in Table 2-right. The most troublesome stroke was *flam*, probably influenced by the short number of instances of this type in the database.

train data	test data	multimodal	audio	video
performers 1, 2, 3	performer 4	89.5	83.7	74.3
performers 1, 3, 4	performer 2	95.9	88.2	77.7
performers 1, 2, 4	performer 3	91.2	87.6	75.8
performers 2, 3, 4	performer 1	92.7	60.1	88.0

Table 1. Percentage of correctly classified instances in each fold of the CV.

stroke	multimodal	audio	video
wood	98.2	97.2	91.2
hand	99.2	86.6	98.9
stick	89.9	71.3	74.8
bounce	76.2	71.8	27.0
rimshot	93.7	83.9	57.8
flam	45.9	6.2	23.5

a	b	c	d	e	f	← classified as
556	0	0	0	10	0	a wood
1	1468	2	5	3	2	b hand
0	44	1019	49	20	2	c stick
0	2	76	224	1	4	d bounce
1	4	23	4	524	3	e rimshot
0	13	5	24	4	39	f flam

Table 2. Percentage of correctly classified instances for each type of stroke (left) and confusion matrix for the multimodal approach (right), averaged over all the CV folds.

7 Concluding remarks

A multimodal approach for percussion music transcription was presented, which focuses on the analysis of audio and video recordings of Candombe drumming. This work is part of an ongoing interdisciplinary research effort for the development of tools for computer-aided music analysis. Recording sessions were conducted in order to generate the data for this research. Due to the fast movements involved in real performances, high video frame rates are mandatory. This generates huge amounts of data, calling for automatic analysis methods. To that

end, several signal processing techniques are applied for automatically extracting meaningful information from audio and video recordings.

In the proposed approach, multimodality is exploited two-fold: (a) onsets detected on the audio source are used to drive the processing of both modalities, and (b) classification of the detected events is performed by combining the information from audio and video in a feature-level fusion scheme. Results show that the method is able to improve the performance attained by each modality on its own, which will be further explored in future research.

References

1. G. Andrews. *Blackness in the White Nation: A History of Afro-Uruguay.* The University of North Carolina Press, Chapel Hill, 2010.
2. S. Dixon. Onset detection revisited. In *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06)*, pages 133–137, Montreal, Canada, September 2006.
3. S. Essid and G. Richard. Fusion of Multimodal Information in Music Content Analysis. In M. Müller, M. Goto, and M. Schedl, editors, *Multimodal Music Processing*, pages 37–52. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Germany, 2012.
4. L. Ferreira. An afrocentric approach to musical performance in the black south atlantic: The candombe drumming in Uruguay. *TRANS-Transcultural Music Review*, 11:1–15, July 2007.
5. A. Fitzgibbon and R. B Fisher. A buyer's guide to conic fitting. In *British Machine Vision Conference BMVC95*, pages 513–522, Birmingham, September 1995.
6. O. Gillet, S. Essid, and G. Richard. On the correlation of automatic audio and visual segmentations of music videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(3):347–355, March 2007.
7. O. Gillet and G. Richard. Automatic transcription of drum sequences using audiovisual features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '05)*, pages 205–208, March 2005.
8. R. Grompone von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall. LSD: a Line Segment Detector. *Image Processing On Line*, 2:35–55, 2012.
9. Mark A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 359–366, San Francisco, CA, USA, 2000.
10. A. Klapuri and M. Davy, editors. *Signal Processing Methods for Music Transcription.* Springer, New York, 2006.
11. A. Lim, K. Nakamura, K. Nakadai, T. Ogata, and H. Okuno. Audio-visual musical instrument recognition. In *National Convention of Audio-Visual Information Processing Society*, March 2011.
12. M. Müller, M. Goto, and M. Schedl, editors. *Multimodal Music Processing.* Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Germany, 2012.
13. Saburo Tsuji and Fumio Matsumoto. Detection of ellipses by a modified Hough Transformation. *IEEE Transactions on Computers*, 27(8):777–781, 1978.
14. Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004)*, volume 2, pages 28–31. IEEE, 2004.