AUTOMATIC TRANSCRIPTION OF DRUM SEQUENCES USING AUDIOVISUAL FEATURES

Olivier Gillet and Gaël Richard

GET-ENST (TELECOM Paris)
Signal and Image Processing department
46, rue Barrault, 75013 Paris, France
[olivier.gillet, gael.richard]@enst.fr

ABSTRACT

The transcription of a music performance from the audio signal is often problematic, either because it requires the separation of complex sources, or simply because some important high-level music information cannot be directly extracted from the audio signal. In this paper, we propose a novel multimodal approach for the transcription of drum sequences using audiovisual features. The transcription is performed by Support Vector Machines (SVM) classifiers, and three different information fusion strategies are evaluated. A correct recognition rate of 85.8% can be achieved for a detailed taxonomy and a fully automated transcription.

1. INTRODUCTION

As a consequence of the exponentially growing amount of available digital data, automatic indexing and retrieval of information based on content is becoming more and more important and represent very challenging research areas. Automatic indexing of digital information allows to extract a textual description of this information (i.e. meta data). In the context of music signals, or audiovisual signals of music performances, such a description would ultimately be a complete transcription - in the form of a detailed musical score. Even if promising results have been achieved in the field of music transcription, several problems still need to be addressed in order to design systems powerful enough to obtain a complete and perfect representation of high-level musical information. The transcription task becomes very complex when the problem of source separation arises, especially because the number of sounds played simulatenously remains unknown. Moreover, many parameters related to expressiveness, style or playing technique cannot be easily extracted from the audio signals, but are easier to extract from a video signal of the instrumentist.

In this paper, we describe and evaluate a novel multimodal approach in which video signals recorded by a camera filming a drummer are analyzed in order to enhance the transcription of the performance. This work is a follow-up of a previous study conducted on drum loops transcription [1] where only audio features were used. It is important to note that we ultimately aim at the indexing of existing audiovisual recordings of music performances, a task for which it is impossible to use specific instrumentation such as sensors, or to control the recording conditions in such a way that scene recognition will be performed more easily (for example by using coloured sticks or gloves, or a neutral background). To our knowledge, there is no prior works related to the transcription of music using directly a multimodal approach. However, researches have been carried out in the analysis of the correlation between video and audio sources, for various purposes such as computer human interaction, biometrics, or video indexing. In [2], Smaragdis and Casey present an application of Independent Component Analysis to the extraction of audiovisual features from a video stream, and give a simplified musical example of fingers on a piano keyboard. In [3] Fisher and Darell present various statistical model for joint audio/video analysis, especially for tasks such as speaker localization in video scenes. The computer-vision part our problem has a few similarities with the problem of gesture analysis [4]. In [5], Murphy presents a computer-vision system for tracking a conductor's baton. In [6], Wanderley shows how an expressiveness parameter can be derived from the angle of a clarinet with respect to the performer. Finally, Dahl conducted numerous multimodal experiments showing the relationship between body movements and emotions in marimba performances or the correlation between video features and musical accent [7] in drumming.

The paper is organized as follows. The next section describes the overall system architecture. Section 3 presents the database specifically recorded for this work. Then, section 4 is dedicated to the description of the video features extraction. The different statistical classification approaches tested are presented in section 5. Section 6 discusses the results obtained and, finally, section 7 suggests some conclusions and future directions.

2. SYSTEM ARCHITECTURE

The system aims at transcribing audiovisual drum sequences into a higher level representation consisting of a list of pairs (onset time, instrument of the drum kit played). It is built on a previously developed audio-only transcriber presented in [1].

2.1. Previous audio transcription system

The audio-only transcription system on the top of which the audiovisual extension was built incorporates 3 modules, namely:

- **A segmentation and tempo extraction module.** These parameters were obtained by applying an onset detection algorithm based on sub-band decomposition [8].
- **A features extraction module.** The features extracted from the audio signals include: The **mean of 13 Mel Frequency Cepstral Coefficients** including c_0, calculated on 20 ms frames with an overlap of 50 % and averaged over the stroke duration ; **4 spectral shape parameters** defined from the

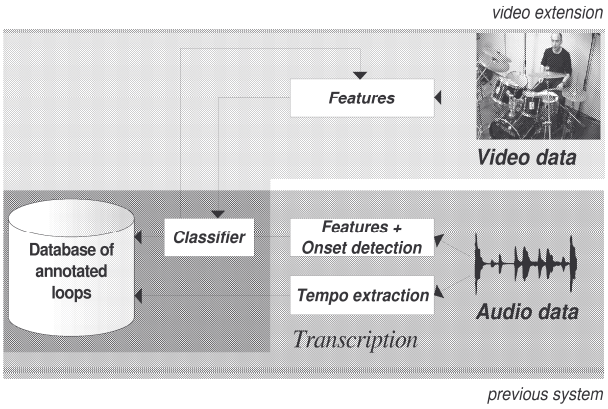

Fig. 1. System architecture

first four order moments ; and **6 Band-wise frequency contents parameters** corresponding to the log-energy in six pre-defined bands (in Hertz: [10-70] Hz, [70-130] Hz, [130-300] Hz, [300-800] Hz, [800-1500] Hz, [1500-5000] Hz).

- **A classification module** for which several classifiers (Hidden Markov Models, Support Vector Machines) were tested.

2.2. Audiovisual transcription system

The extensions and improvements of the previous system which are presented in this work include:

- **A new audiovisual database,** detailed in the next section.
- **A new set of features extracted from the video track.** Because the computation of the video features requires a calibration of the scene, the output of a transcription carried out on the sole audio signal can be used to derive a set of video features that will subsequently enhance the transcription. Alternatively, the user can manually calibrate the system.
- **New classification approaches.** Some of the classifiers presented in our previous work are no longer suitable to the taxonomy and size of the new database. Moreover, several classification and information fusion schemes to deal with the availability of the two audio and video information sources were to be evaluated.

Because audio signals of drum instruments have very sharp onsets, it is easier to detect the start time and duration (T, d) of each stroke in the audio domain than in the video domain.

The overall architecture of the resulting system is depicted in figure 1.

3. DATABASE

Since no audio/video database of drum performances was available, we recorded our own database which consists of 35 sequences containing 2170 strokes. The sequences were played on a drum kit made up of 9 instruments: a bass drum, a snare drum, three toms (high, medium, low), one hi-hat cymbal, two crash cymbals and one ride cymbal. In order to increase the variability of the recorded data, the sequences were performed with two sets of sticks: classic sticks and "bundle sticks" - small wood rods bundled together. Four studio-quality microphones were used: one for the bass drum, one for the snare drum, and two overhead microphones. In the scope of this work, the audio signals were recorded at the stereo output of the mixing desk, at a sample rate of 48 kHz, and converted into mono by combining the right and left channels.

The video signals were recorded with a Canon XL1 professional DV camera. The camera was fixed on a tripod and remained steady during the whole recording. The video was recorded in DV format with a resolution of 720x576, at 25 frames per second. For the purpose of this work, only the luminosity channel of the video was processed. Moreover, since the DV format is interleaved, scanline artifacts were removed with simple spatial filtering. As our goal is the indexing of pre-recorded material, we avoided using any specific sensor or, visual clues such as coloured gloves, sticks or backgrounds to improve the detection, even if the recording conditions for this database were well controlled.

An intermediate annotation was at first obtained with our previous audio based transcription system ; and secondly, this annotation was corrected and refined. It is worth precising that despite the similar instrument set used, the taxonomy used in this work is slightly different and detailed than in [1]. For example, a tom (resp. cymbal) stroke will not be labelled as **tom** (resp. **cymb**) but as **low tom**, **mid tom**, **high tom** (resp. **crash cymbal 1**, **crash cymbal 2**, **ride cymbal**).

As a result, each acoustic event is labelled with the corresponding instrument or combination of instruments when several instruments are played at the same time (for example if the bass drum and the ride cymbal are hit simultaneously, both labels are attached to the corresponding stroke).

4. VIDEO FEATURES

4.1. Masks

We observed that when an instrument of the drum kit is played, two kinds of visual clues can be derived from the video: the motion of the sticks, or any specific gesture the drummer has to perform to hit the instrument (for example, kicking the pedal of the bass drum) ; and the motion of the instrument itself, or the vibration of its membrane.

Thus, two areas of the video images are defined for each instrument: an area in which motion is associated to the gesture performed by the drummer to hit the instrument, and an area in which motion is associated to the vibration of the instrument itself once hit. We subsequently use two 2D weighting masks $M_{gesture}(x, y)$ and $M_{instr}(x, y)$ to represent these areas.

The thresholded difference sequence was used as a simple motion estimator. If $V(x, y, t)$ is the sequence of video image, the thresholded difference sequence $D(x, y, t)$ is given by:

$$D'(x,y,t) = |V(x,y,t) - V(x,y,t-1)| \quad (1)$$

$$D(x,y,t) = \begin{cases} D'(x,y,t) & \text{if } D'(x,y,t) > S, \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

For each instrument, and each stroke starting at frame T, the duration of which is d frames, two features are computed from the thresholded difference sequence and the weighting masks:

- The intensity of motion in the gesture mask, accross a short time interval centered on the beginning of the stroke.
  $I_{gesture} = \sum_{t\in[T-\delta,T+\delta]} \sum_{x,y} M_{gesture}(x,y)D(x,y,t)$
  Typical value for δ is $\delta = 2$.
- The intensity of motion in the instrument mask, accross the whole duration of the stroke.
  $I_{instr} = \sum_{t\in[T+\delta,T+d-\delta]} \sum_{x,y} M_{instr}(x,y)D(x,y,t)$

This results in a set of 18 features computed for each stroke: The $I_{gesture}$ and I_{instr} features for each of the 9 instruments of the kit.

4.2. Calibration

The system is calibrated by defining the 18 masks. Different calibration schemes are devised:

- *Manual*. A human operator manually defines the image regions corresponding to each instrument of the kit.
- *Automatic*. A transcription is obtained using the audio-only transcription system. This transcription is used to generate a mask, by averaging the difference sequence accross the appropriate interval and all the recognized occurences of each instrument of the kit.

5. CLASSIFICATION

5.1. Information fusion

The fusion of video and audio information is performed by three different fusion approaches:

- **Joint features vectors.** Let x_{audio} (resp. x_{video}) be the audio (resp. video) features vector. Classifiers are trained with joint features vectors:
  $$x_{joint} = [x_{audio}(1)...x_{audio}(25)x_{video}(1)...x_{video}(18)]$$
- **Best of unimodal experts.** Two classifiers are trained, one using the audio features, the other the video features. For each stroke, the output of the classifier giving the best confidence score is kept. For instance, the video classifier is used only when the audio classifier produces an uncertain result. The advantage of this approach is that it allows the use of a larger database for audio transcription, and a smaller, specific database adapted to the current scene and camera angle for the video transcription.
- **Fusion.** As above, two classifiers are trained except that these classifiers produce for each class 2 probabilities: $P(class|x_{audio})$, $P(class|x_{video})$.
  Each stroke is labelled with the class that maximizes the product of these two probabilities.

As some of the parameters are correlated, especially when joining video and audio features, a Principal Component Analysis is performed on the fused data set when the joint feature vectors approach is chosen, or on the separate audio and video datasets when another approach is chosen.

5.2. SVM classification

It was shown in [1] that Support Vector Machines (SVM) were well suited for drum loops transcription and are therefore used in this study.

In our work, we use the "one versus one" approach, in which $\frac{n(n-1)}{2}$ binary SVM classifiers are trained, each discriminating between a pair of classes. If x is the input vector, (i,j) a pair of classes, (x_{ijk}) (resp. (v_{ijk})) the support vectors (resp. the weights), c_{ij} the parameter of the binary SVM classifier trained to discriminate the classes i and j, the decision function commonly used is :

$$f_{ij}(x) = \sum_k w_{ijk}K(x,x_{ijk}) + c_{ij} \quad (3)$$

$$D_{ij}(x) = \text{sgn} f_{ij}(x) \quad (4)$$

The input vector x will be classified as i (resp. j) if $f_{ij}(x)$ is positive (resp. negative).

However, to obtain a confidence measure, a specific decision function is defined: the output f_{ij} is mapped to the interval $]0,1[$ with a sigmoid function: $D'_{ij}(x) = \frac{1}{1+e^{Af_{ij}(x)+B}}$

Provided that appropriate values of the parameters A, B are chosen [9], this quantity can be interpreted as an a-posteriori probability $P_{ij}(\text{class} = i|x) = D'_{ij}(x)$. The final output of the classifier is a probability for each class, computed by coupling the pairwise probabilities using the algorithm proposed by Hastie and Tibshirani in [10]. The class assigned to the input x is the one that maximizes the quantity $P(\text{class} = i|x)$, which can be used itself as a probabilistic measure of the accuracy of the classification. This method gives similar results, and a much better ranking function, than more classic approaches using voting and vote counting.

In the scope of this study, a radial basis kernel was chosen: $K(x,y) = \exp^{-\gamma||x-y||^2}$ where γ is equal to the inverse of the number of features. The library LibSVM [11] allowed an easy implementation of these SVM classifiers with a modified output.

6. RESULTS

6.1. Evaluation protocol

Two main experiments were conducted on our dataset. In the first experiment, the video features were computed with a mask manually drawn on the picture. In the second experiment, the video features were automatically computed from an automatic audio-only annotation of the database. Example of computed masks are provided in figure 2. One can also check and correct the automatic transcription used as a preliminary step for the calibration in this second experiment.

For each of these experiments, we compare the recognition rate obtained with different feature sets and fusion schemes. **Blind** is the recognition rate obtained using only audio features. **Deaf** is the recognition rate obtained using only video features. **Joint features**, **Fusion** and **Best expert** are the recognition rates obtained using a combination of video and audio features.

A K-fold cross-validation approach was followed. It consists in splitting the whole database in $K = 5$ subsets, training the classifier on four of them, and keeping the last subset for evaluation. The procedure is then iterated by rotating the 5 subsets used for training and testing.

Fig. 2. Examples of computed masks: gesture for bass drum (the pedal is kicked by the right foot), gesture for the cymbal at the right of the drummer, gesture for the low tom at the right of the drummer, and reference image.

	Manual	Automatic
Deaf	67.7%	64.0%
Best expert	82.7%	82.1%
Fusion	84.3%	82.7%
Joint features	**86.7%**	**85.8%**
Blind	81.5%	81.5%

Table 1. Drum instruments recognition results

6.2. Results and discussion

Our classifier using only audio features as presented in [1] managed to cope with a lot of variability in the dataset and complex situations like effects or overlapping strokes. Not surprisingly, it performs well on this simpler dataset, in which only one drum kit is used. Another interesting point is that the set of audio features chosen in our previous work is still relevant for this classification task which uses a more detailed taxonomy.

The increased recognition rate obtained with a combination of audio and video features validates our multimodal approach, however, the **Best expert** strategy in which the most reliable of the information sources is used does not give the best results. This can be explained by the fact that processing the audio and video data in the same classifier allows to take advantage of their correlation. Especially, the PCA step is very important since it forges truly multimodal features.

It is worth precising that these comparisons are relevant only if the variance of the K-fold cross-validation is small enough. However, estimating this variance is difficult. More precisely, because of our limited dataset, there was a high variability in the estimations obtained by the different estimators presented in [12] ; using the estimator $\hat{\theta}_3$, the standard deviation is 2.1.

7. CONCLUSION AND FUTURE WORK

This paper presented a novel approach to enhance the transcription of drum sequences using audio and video features. The system can work without calibration, even if the best results, a correct recognition rate of 86.7%, are obtained with manual calibration. The overall gain of our multimodal approach, is still limited in the context of the well controlled database used. Future work will in fact consider more complex situations including the transcription of drum signals when other instruments are playing along with the drummer. This could validate the hypothesis that video features will drastically improve the transcription results, in situations when separating the audio sources will become impossible. More robust video features will also have to be tested, as well as sequence models (Hidden Markov Models) based on joint video/audio features.

8. ACKNOWLEDGEMENTS

The authors wish to thank Michel Desnoues for having performed and recorded the sequences used in this work.

9. REFERENCES

[1] O. Gillet and G. Richard, "Automatic transcription of drum loops," in *Proceedings of the IEEE ICASSP 2004 Conference*, May 2004.

[2] P. Smaragdis and M. Casey, "Audio/visual independent components," in *Proceedings of International Symposium on ICA and Blind Source Separation*, april 2003.

[3] J. W. Fisher and T. Darrell, "Signal level fusion for multimodal perceptual user interface," in *Proceedings of Workshop on Perceptive User Interfaces*, october 2001.

[4] M.M. Wanderley and M. Battier, *Trends in Gestural Control of Music*, Ircam - Centre Georges Pompidou, 2000.

[5] D. Murphy, "Tracking a conductor's baton," in *Proceedings of 12th Danish Conference on Pattern Recognition and Image Analysis 2003*, 2003.

[6] M. M. Wanderley and P. Depalle, "Gesturally-controlled digital audio effects," in *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01)*, December 2001.

[7] S. Dahl, "The playing of an accent - preliminary observations from temporal and kinematic analysis of percussionists," in *Journal of New Music Research*, 2000, vol. 29(3), pp. 225–234.

[8] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1999.

[9] J. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," in *Advances in Large Margin Classiers*, 2000, pp. 61–74.

[10] Trevor Hastie and Robert Tibshirani, "Classification by pairwise coupling," in *Advances in Neural Information Processing Systems*, 1998, vol. 10.

[11] C.C. Chang and C.J. Lin, *LIBSVM: a library for support vector machines*, 2001, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[12] Yoshua Bengio and Yves Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," CIRANO Working Papers 2003s-22, CIRANO, May 2003, available at http://ideas.repec.org/p/cir/cirwor/2003s-22.html.