DETECTING ARTIST PERFORMANCES IN A TV SHOW

Félicien Vallet[1,2]*, Gaël Richard*[1]*, Slim Essid*[1]*, Jean Carrive*[2]

[1]TELECOM ParisTech
37, rue Dareau
75014 Paris, FRANCE
{firstname.lastname@telecom-paristech.fr}

[2]Institut National de l'Audiovisuel
4, Avenue de l'Europe
94366 Bry-sur-Marne Cedex, FRANCE
{jcarrive@ina.fr}

ABSTRACT

TV show structuring consists in breaking down a program into several sequences (interviews, musical performances, film excerpts, etc.) and in retrieving for each of these segments high-level knowledge often refered to semantic content. The focus here is on two particular tasks: the detection of musical performances, i.e. the segments where artists are performing live music, and the identification of the artist for each of these segments. The corpus used in this study is "Le Grand Échiquier", a French talk show from the 1970s-1980s provided by INA, each show lasting around three hours. This corpus contains on top of videos, annotations made by professional documentalists, providing a list of participants and a summary for each show.

1. INTRODUCTION

This paper presents a preliminary study based on only one show (CPB84052346). The approach relies on the fusion of data gathered from partners. A major aspect of the work resides in the use of multimodal features. In [1], Cheng et al. propose a semantic-event segmentation of wedding ceremony videos. Similarly, in our case, to build a robust musical performance segmentation we combine basic musical descriptors carried by the audio track, with video descriptors. This fusion step helps for the disambiguation of complex situations likely to contain music, such as film excerpts. Once the musical performance segmentation obtained, the next task is the labeling of each segment with the name of the performing artist.

2. SEGMENTATIONS AND DESCRIPTORS

We use segmentations gathered from partners of the projects *K-Space* [2] and *Infom@gic*:

- **Audio segmentations** includes general sound classification discriminating music, speech and applause by TELECOM ParisTech and TUB.
- **Video segmentations** consists in face detection by TELECOM ParisTech and film excerpts detection by EADS.
- **Speech segmentation** consists in automatic transcription by VECSYS.

Figure 1 displays the general scheme (yellow boxes contain the available segmentations and descriptors).

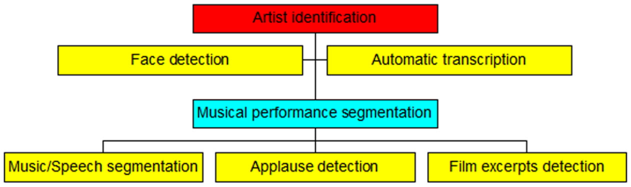

Fig. 1: *Overall fusion process.*

3. MUSICAL PERFORMANCE SEGMENTATION

In this section we fuse various segmentations. An important issue is the integration of segmentations of different natures. While some of the data present, at a fixed time step, probability values to belong to a class, some others are merely rough boundaries indicating the detection of a concept, like a film excerpt for instance.

3.1. Class probability fusion

The general sound classification provided by TUB uses Gaussian Multivariate Model (GMM) while TELECOM ParisTech used Support Vector Machines (SVM). These segmentations are to be fused in order to produce a common temporal representation. For this, we average the probability/sec of each class. This provides, for the music and speech classes, probability curves of the whole show.

3.2. Heuristic rules

For the film excerpts and applause detections, averaging is not an issue since these segmentations present rough boundaries and no class probabilities. However, it is crucial to proceed to the fusion. Indeed, fusing the music/non-music segmentation obtained earlier with applause and film excerpts detections helps for the disambiguation of sequences containing music, but not considered as live music performances. To rule out such cases we use heuristic hypotheses:

- a musical segment last at least 90 sec.
- a musical segment is followed by applause from the audience.
- a film excerpt last at least 30 sec.
- a film excerpt is followed by applause from the audience.
- musical segments over film excerpts are ignored.

The result obtained after the application of these rules is a refined segmentation that can be used for the artist identification task.

4. ARTIST IDENTIFICATION

The aim here is to bridge the gap between low-level descriptions and semantic content. For this, two tools are of great use: the annotation by documentalists (particularly the participants list) and the automatic transcription. The idea is to retrieve first and last names of the artists which are expected to perform on stage. Again several heuristic rules are used:

- we suppose that the name of the artist is pronounced before or after each music segment.
- the time window for the names retrieval is 90 sec before and 90 sec after each music segment.
- we associate the biggest face detected during the musical segment with the main artist.

5. RESULTS

5.1. Musical performance segmentation

With the heuristic rules, we obtained a rather robust segmentation for live musical appearance for the show CPB84052346. The duration of the show is 3h 25min 57sec and contains 42min 31sec of musical performances. In the following table, scores are given in percent with respect to the total number of live music segments:

Accuracy	False Alarm	Non-Detection
86%	6%	8%

Results for musical appearance segmentation for the show CPB84052346.

From a user-oriented point of view, the test show contains twelve segments of artists performing live. The non-detection rate comes from two musical segments too short to be detected while the false alarm rate can be explained by the detection of a small documentary that contains a lot of music.

5.2. Artist identification

It is much more difficult to get results for this aspect. First of all, all names cannot be retrieved because all of them were not in the database used by VECSYS for the automatic transcription. So, it may happen that only first names are detected. Also, for a given musical segment, several names can be proposed. Figure 1 shows the biggest face detected for the first eight segments while the table displays the results for the artist identification:

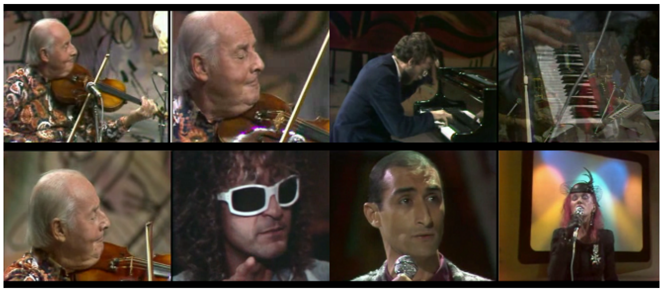

Fig. 2: *Face retrieval for musical segments*

Excerpt	First Names	Last Names	Truth
1	Gérard	Oury	Stéphane Grappelli
2	Gérard, Stéphane	Oury, Grappelli	Stéphane Grappelli
3	François	-	François Duchable
4	François, Jacques	Higelin	Jacques Higelin
5	Stéphane	Grappelli	Stéphane Grappelli
6	Michel	-	Michel Polnareff
7	Gérard, Farid	-	Farid Chopel
8	Diane, Jacques	Higelin	Diane Dufresne

Results of artist identification for the show CPB84052346 (Truth being the actual name of the artist).

Our method provides a good indication of the identity of the performing artist. Besides, the assumption about the biggest face detected during the music segment belonging to the artist seems quite justified despite one error (excerpt 4).

6. PERSPECTIVES

With this study, it seems possible to build an automatic approach to analyse sentences from the documentalists' annotations and then provide semantic knowledge to low-level segmentations, like the automatic labeling of musical excerpts.

7. REFERENCES

[1] Wen-Huang Cheng et al., "Semantic-event based analysis and segmentation of wedding ceremony videos" in proceedings *MIR '07*, Augsburg, Germany, 2007.

[2] K-Space, Network of Excellence http://www.k-space.eu/