INTERSPEECH 2009 BRIGHTON

Robust Audio-based Classification of Video Genre

Mickael Rouvier, Georges Linarès, Driss Matrouf

LIA, University of Avignon

{mickael.rouvier, georges.linares, driss.matrouf}@univ-avignon.fr

Abstract

Video genre classification is a challenging task in a global context of fast growing video collections available on the Internet. This paper presents a new method for video genre identification by audio analysis. Our approach relies on the combination of low and high level audio features. We investigate the discriminative capacity of features related to acoustic instability, speaker interactivity, speech quality and acoustic space characterization. The genre identification is performed on these features by using a SVM classifier. Experiments are conducted on a corpus composed from cartoons, movies, news, commercials and musics on which we obtain an identification rate of 91%.

Index Terms: video genre classification, audio-based video processing

1. Introduction

The last years, the amount of video available on the Internet or digital TV have increased considerably and users need efficient tools to crawl these large collections. This point motivated many works on structuring audiovisual databases by content analysis, mostly by visual-based approaches [1]. Some authors investigated text-based categorisation [2], usually based on viewable text or speech contents automatic transcription. Nevertheless, web-data are strongly variable and recognition rates are generally too high to perform a correct analysis of automatic transcriptions. Audio-only approaches could be more robust to both acoustic context variability and low speech quality, and some authors proposed audio-only based methods for video genre categorization. Rather than the classification strategies, research efforts focused mainly on the features used for categories characterization. Conventional approach consists in acoustic-space characterization by using statistic classifier like Gaussian Mixture Model (GMM), neural nets or support vector machines (SVM) on cepstral domain features [3, 4, 5]. [6, 5, 3] propose time-domain features like zero crossing rates or energy distributions. [7] investigate higher level analysis like tracking of audiovisual events.

In this paper, we focus on video genre classification by using an audio-only method. We propose to combine low and high level descriptors based not only on cepstral analysis, but also on the audio structure of the targeted videos. Finally, we compare various combination schemes based on statistic classifiers.

The paper is organized as follows : the next section describes precisely the targeted task and the corpora involved in this study. Section 3 presents the principle and the architecture of the proposed categorization system. In Section 4, we discuss about relevance and tractability of various acoustic features, and we propose a set of high-level descriptors. We first evaluate the discriminative capacity of each and we estimate its complementary to others. Section 5 focuses on the upper level, where classification is performed : a combination scheme operating in probability space is compared to the classical scheme where classifiers operate directly in feature space. Section 6 concludes and draws some perspectives.

This research was supported by the ANR agency (Agence Nationale de la Recherche), RPM2 project (ANR-07-AM-008)

2. Task and corpus

We have selected 5 categories that are commonly targeted by video genre classification tasks : news, movies, cartoons, musics and commercials. The corpus is composed of 1200 videos indexed by *Dailymotion* and *Youtube*, with duration from 2 to 5 minutes. 1000 of them are used for the training of the various components of our system, 200 composing the test set. The 5 genres are equally represented in this database (about 200 video per genre for training, 40 for testing). The speech contents are systematically in French language.

3. System Overview

The system proposed is a 2-level architecture, where the first level consists in features extraction that are combined at the second level. We identify the following 4 feature groups, that are described in-depth in the next section :

- acoustic space : this is the most frequently used descriptor for categorization by audio only. The general idea is to distinguish genres by statistical modeling of their cepstral patterns.
- speaker interactivity : video genre could differ from their interactivity profile; for example, speaker turns and speaking time are probably differents beetwen cartoons and news.
- speech quality : most of the speech-quality related features rely on speech recognition methods; we estimate the quality of speech contents by acoustic analysis and by an a posteriori evaluation of a speech recognition process that is applied to the speech segments.
- instability : these features represent the regularity of the acoustic stream in the time domain.

These feature groups may share some mutual information; for example, speaker interactivity and acoustic instability are probably partially correlated. Nevertheless, each of them is supposed to offer a particular view of the targeted document that may be useful to the upper level genre identification system.

In order to evaluate the discriminative capacity of each of these feature groups, we train a set of genre-dependent GMM models on each one. The decision process follows the classical scheme of bayesian classification, the identified genre being the one that maximizes the conditional probability $P(X^k|G_i^k)$, where X is the feature vector, G_i^k the model for the genre and the feature group k.

The second level combines these level-1 classifiers. All probabilities are grouped in a vector, and a Support Vector Machine (SVM) classifier is trained on these probability vectors. Here, we used the linear kernel SVMs that are trained by a leave-one-out strategy : considering that we use SVM for combination of GMM outputs, the SVM training cannot be achieved

10.21437/Interspeech.2009-337

on the data that were used for GMM classifiers. Therefore, each example is extracted from the training corpus and the set of GMMs is trained on the remaining examples. The single example is then submitted to these GMM classifiers, resulting probabilities being saved as an input vector for SVM training. By applying this process on all examples, we collect a SVM training corpus containing probabilities estimated on data that are out of the GMM training corpora.

Complementarity of features is estimated by following a step-by-step protocol that consists in adding, at each step, a new feature group to the previous set. Then, we train a new SVM classifier and study how the classification scores take benefits from the additional information provided to the lastly added features. Final performance will be evaluated on the full feature vector that integrates the four descriptor groups.

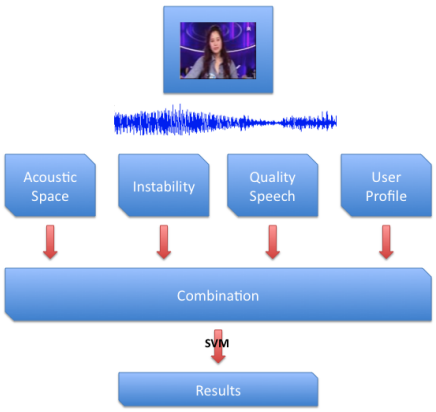

Figure 1: *Principle scheme of genre classifier : features related to acoustic space characterization, speaker interactivity, speech quality and Acoustic Instability are extracted and combined by a SVM classifier.*

4. Low and High level features for genre identification

4.1. Acoustic space characterization

One of the most popular approach for genre identification by acoustic space characterization relies on MFCC (Mel Frequency Cepstral Coefficients) analysis and GMM or SVM classifiers [6, 5, 3]. [6] demonstrates the efficiency of 14 MFCC coefficients with delta and delta-2 coefficients. By using a classical GMM classifier, this system obtain 52% correct identification of 5 video genres (sports, cartoons, commercials, news and musics). This approach was motivated by the performance of MFCC/GMM on speaker identification tasks.

Here, acoustic space characterization is achieved by following this classical scheme, but we propose to explore two different ways. First, PLP (Perceptual Linear Predictive) coefficients are known to be more robust than MFCC on speech recognition tasks. We compare them on our identification task.

Second, one of the main difficulty of genre categorization is due to the diversity of the videos that are similarly labelled. Some techniques for intra-class variability reduction have been proposed in the speaker-identification field, especially factor analysis that recently demonstrated a strong efficiency ([8]). Here, we evaluate this method for reduction of intra-genre variability.

4.1.1. Factor Analysis for genre classification

Gaussian Mixture Models (GMM) and GMM-UBM (Universal Background Model) constitute a classical approach to speaker verification. The world model (UBM-GMM) represents all genres, the genre-specific GMMs resulting from a world model adaptation. Only means vectors are adapted, the weights and the variances remaining constant with respect to those of the UBM.

The goal of Factor Analysis (FA) is to decompose the genre-specific model into three different components: a genre-session-independent component, a genre-dependent component and a session-dependent component (each recording corresponding to one session). A GMM mean supervector is defined as the concatenation of the means of the gaussian components of the GMM. Let D be the dimension of the feature space, the dimension of a supervector mean is $M.D$, where M is the number of components in the GMM. For a genre GE belonging in session h, the factor analysis model can be formulated as:

$$\mathbf{m}_{(h,GE)} = \mathbf{m} + \mathbf{D}\mathbf{y}_{GE} + \mathbf{U}\mathbf{x}_{(h,GE)}, \quad (1)$$

where $\mathbf{m}_{(h,GE)}$ is the session-genre dependent supervector mean, $\mathbf{D}$ is $MD \times MD$ diagonal matrix, $\mathbf{y}_{GE}$ the genre vector (a MD vector), $\mathbf{U}$ is the session variability matrix of low rank R (a $MD \times R$ matrix) and $\mathbf{x}_{(h,GE)}$ are the channel factors, a R vector. All parameters of the FA model are estimated by using Maximum Likelihood criterion and algorithm EM. Several sessions corresponding to each genre have to be used for accurate estimate of FA parameters. The session compensated model is simply obtained from 1, as follows:

$$\mathbf{m}_{GE} = \mathbf{m} + \mathbf{D}\mathbf{y}_{GE}, \quad (2)$$

4.1.2. Experiments

PLP and MFCC parametrization are first compared on a baseline GMM classifier. For both, we used 512 mixture models (one GMM per genre) trained by likelihood maximization swith expectation-maximization algorithm. Results are reported in table 1. The performance obtained by MFCC analysis are close to the ones reported in [6], on a similar task. PLP provides better results, with a relative improvement of 42% of identification rate, in comparison to the baseline MFCC/GMM.

Performances are strongly improved by factor analysis. On PLP-based system, the relative error rate reduction is about 29%, absolute identification rates increasing from 76% to 86%.

Table 1: *Correct genre identification rates by GMM-classification on acoustic-space features : PLP and MFCC based parameters are compared, without (MFCC and PLP) and with variability reduction by factor analysis (FA-MFCC and FA-PLP).*

	Mus	New	Com	Car	Mov	Total
MFCC	0.58	0.84	0.17	0.31	0.73	0.52
PLP	0.95	0.92	0.46	0.78	0.70	0.76
FA-MFCC	0.95	0.84	0.58	0.85	0.92	0.83
FA-PLP	0.97	0.97	0.56	0.87	0.95	0.86

Finally, the best system obtain 86% of correct identification, all genre except commercials being recognized at more than 86%. Commercials obtain the worse result for all methods, this last result being probably due to its natural similarity to other genres : we observed confusion rates with musics, news, cartoons and movies respectively of 22%, 10%, 10%, and only 2% with cartoons.

4.2. Interactivity Features

The number of speakers and how they communicate together could differ according to the genre. For example, there is usually only one main speaker in news, when cartoons or movies contain generally many speakers with highly variable speaking times and speaker turns. The interactivity features aim at represent these speaker-related profiles. The feature vector is composed of the 3 following parameters : the density of speaker turns, number of speakers, and speaking time of the main speaker.

These data are extracted by using a speaker diarization system based on a 3 stages segmentation and clustering process. The first one performs a Viterbi segmentation based on the 3 following classes : *speech*, *speech over music* and *music*. Each of them is modeled by a GMM of 64 mixtures. Acoustic vectors are composed by 12 MFCC coefficients, their first and second order derivatives and the delta and delta2 energy. This system is fully described in ([10]). The last 2 stages perform speaker turn detection and clustering. We used the system described in [11] based on Bayesian Information Criterion (BIC) and agglomerative clustering strategy. This techniques allow to estimate the number of speaker and speaker turns for each document.

We first tested a GMM-based classification with the 3 interactivity descriptors. Then, GMM outputs are combined to the acoustic classifiers described in the previous section : all probabilities are grouped in a probability vector which is processed by a SVM classifier. As described in section 3, SVM is trained by a leave-one-out strategy that allows to estimate SVM parameters without requirement of additional training data.

Table 2: *Interactivity features for genre classification : correct identification rates by genre, for GMM-classification based on acoustic space characterization only (AS), interactivity only (Int), and interactivity combined with acoustic space characterization (AS+Int) in a SVM classifier.*

	Mus	New	Com	Car	Mov	Total
AS	0.97	0.97	0.56	0.87	0.95	0.86
Int.	0.29	0.52	0.90	0.95	0.56	0.64
AS+Int	0.95	0.84	0.85	0.85	0.92	0.88

Globally, results shows that interactivity is significantly less discriminative than cepstral parameters; nevertheless, it provides complementary information to improve significantly accuracy on commercials and cartoons that were badly recognized by acoustic space method. Globally, the absolute gain is about 2% since performance are well balanced, recognition rates being greater than 84% for all genres.

4.3. Speech quality

The basic idea is that the speech quality could provide some relevant information about genres. For example, speech is usually clean in news, whose the linguistic domain is well covered by speech recognition systems, since linguistic domain of commercials may be unexpected due to the product specificities and speaking styles.

We use 3 features in this group, all based on the LIA broadcast news transcription system ([12]). This system is based on n-gam language models and state tyied context-dependent HMM for acoustic modeling.

The first descriptor is the posterior probability of the one-best hypothesis. We use posteriors as a confidence measure that integrates not only the acoustic and the linguistic scores, but also some information related to the decoding graph that was effectively developed by the search algorithm. The second is the linguistic probability of the one best hypothesis, language models being trained on materials that are extracted from transcription of French broadcast news and newspapers. The last feature is based on phonetic entropy. This descriptor was introduced by [13] for speech/music separation. It is computed as the entropy of acoustic probabilities :

$$H(n) = -\frac{1}{N}\sum_{m=1}^{N}\sum_{k=1}^{K} P(q_k|x_m)\log_2 P(q_k|x_m) \quad (3)$$

where the frame values are averaged over a temporal window of size N and K represents a phonetic model. This measure is supposed to be high on low-quality speech, and to decrease on clean speech.

Table 3: *Speech quality for genre classification : identification rates by genre, for GMM-classification based on Speech quality only (Q), and combined with other features (AS+Int+Q).*

	Mus	New	Com	Car	Mov	Total
AS+Int	0.95	0.84	0.85	0.85	0.92	0.88
Q	0.63	0.94	0.46	0.41	0.39	0.56
As+Int+Q	1.0	0.84	0.82	0.78	0.95	0.88

Speech quality seems to be significantly less accurate than the previously evaluated features, except for news that are correctly recognized at 94%. By combining all features, no gain are observed, music recognition being perfect since the movies identification rate decreases. These results indicate a relatively poor complementarity with others features. This dispointing result may be due to the fact that the combination scheme we use (GMM modeling followed by SVM-based combination) is poorly effective on features that are posteriors probabilities. This point is discussed in section 5, where we focus on classification strategies.

4.4. Acoustic Instability

This feature group aims to extract acoustic variability in time-domain. We implement 5 descriptors related to short term energy (STE), zero crossing rates (ZCR) and density of acoustic model breaking (AMB).

STE is computed classically in a sliding window. Considering that the absolute mean energy level does not provide any significant information (it depends from a lot of unrelevant factors), we first normalize energy according to the maximum energy level observed in the document. Then, the STE mean and variance are extracted from the whole document and used as instability descriptors.

ZCR is the number of times that the audio waveform crosses the zero axis each second. ZCR is mostly used in the speech/music classification algorithms, but, in a more general way, it may be representative of the variability of patterns that are found in the document. Here, we use mean and variance of ZCR as an instability index.

The last descriptor is the density of acoustic ruptures. This value is calculated by a BIC-based rupture detector similar to the one used in speaker diarization systems. Mono-gaussian model are estimated on each $30ms$ window. Then, each signal window is splitted in 2 parts on which 2 mono-gaussian models are trained. The likelihood lost between the whole window model and the 2 sub-models is considered as a breaking index. If this index overcomes a fixed threshold, we consider than a rupture is detected. The number of detected breaking (normalized by the document duration) is used as instability index.

Results shows that, even if the instability alone is not highly efficient, this feature group provides a complementary informa-

Table 4: *Acoustic instability for genre classification : identification rates by genre, for GMM-classification based on instability only (Un), and combined with other features (AS+Int+Q+Un).*

	Mus	New	Com	Car	Mov	Total
As+Int+Q	1.0	0.84	0.82	0.78	0.95	0.88
Un.	0.60	0.68	0.36	0.73	0.41	0.55
AS+Int+Q+Un	0.97	0.86	0.85	0.92	0.92	0.91

tion that allows of a relative reduction of error rates of about 25% (from 88% to 91% correct).

5. Model combination versus feature combination

In previous experiments, we choose to combine GMM classifiers since the classical way consists in grouping all features in multi-dimensional feature vectors that are directly processed by classifiers. This approach is motivated by the fact that classification difficulty may be increased by highly heterogeneous data. The mapping of observations into probability space allows to work in an unified and well defined framework, where various combination strategies could be applied. In other hand, some relevant information may be lost by the intermediate classification, especially features correlation; moreover, direct classification is simpler and faster. We compare the 2 by performing a SVM-based classification directly on aggregated features, estimate of acoustic space descriptors remaining similarly computed by accumulation of GMM-based frame-level statistics. Due to this intermediate classification, we use, for SVM training, the leave-one-out strategy previously described. Comparison of the 2 are reported in table 5. They show that the 2 approaches obtain similar averaged performances, even if they differ according to the genre: cartoons and movies take benefits from feature combination since music detection is significantly improved by model-based approach. Another point is that speech quality features provide additional gains which were not observed with model combination, at contrary to interactivity features which do not improve the global performance. As previously discussed, these changes are probably due to the fact that feature correlations are smoothed by model-based combination. It also suggests that genre specific combination strategies could improve the classification accuracy.

Table 5: *Performances of* **feature-level** *combination of cepstral features (AS), interactivity (I), speech quality (Q) and instability (U). Combination is performed by SVM on vectors including successively each group of feature. Baseline consists in a classical MFCC/GMM classification.*

	Mus	New	Com	Car	Mov	Total
Baseline	0.58	0.84	0.17	0.31	0.73	0.52
AS	0.97	0.97	0.56	0.87	0.95	0.86
AS+I	0.95	0.84	0.85	0.85	0.92	0.88
AS+I+Q	1.0	0.84	0.82	0.78	0.95	0.88
AS+I+Q+U	0.97	0.86	0.85	0.92	0.92	0.91

6. Conclusion and Perspectives

We have studied low and high level features for audio-only based classification of video genres. Classification on each feature group show clearly that acoustic space characterization remains the most discriminative descriptor of genres, especially with PLP acoustic features and variability reduction by factor analysis. Nevertheless, high level descriptors such interactivity and speech quality provide complementary information : we finally obtain an identification rate of about 91%, significantly better than audio-only based reported in the literature and similar to the one obtained by audio-video combination. We plan now to generalize this method to a larger set of genres and to evaluate various combination strategies.

Table 6: *Performances of* **model-level** *combination of cepstral features (AS), interactivity (I), speech quality (Q) and instability (U)*. Combination is performed by SVM on vector of probabilities which results from first GMM-based classification by genre.

	Mus	New	Com	Car	Mov	Total
Baseline	0.58	0.84	0.17	0.31	0.73	0.52
AS	0.97	0.97	0.56	0.87	0.95	0.86
AS+I	0.82	0.78	0.90	0.90	0.90	0.86
AS+I+Q	0.80	0.92	0.87	0.97	0.87	0.89
AS+I+Q+U	0.87	0.84	0.85	0.97	1.0	0.91

7. References

[1] D. Brezeale and D. J. Cook, "Automatic video classification : A survey of the literature," in *Systems, Man, and Cybernetics*, 2008.

[2] W. Zhu, C. Toklu, and S.-P. Liou, "Automatic news video segmentation and categorization based on closed-captioned text," in *Multimedia and Expo, ICME*, 2001.

[3] M. Roach, L.-Q. Xu, and J. Mason, "Classification of non-edited broadcast video using holistic low-level features," in *IWDC'2002*, 2002.

[4] R. Jasinschi and J. Louie, "Automatic tv program genre classification based on audio patterns," in *Euromicro Conference, 2001*, 2001.

[5] L.-Q. Xu and Y. Li, "Video classification using spatial-temporal features and pca," in *Multimedia and Expo, 2003. ICME '03*, 2003.

[6] M. Roach and J. Mason, "Classification of video genre using audio," in *European Conference on Speech Communication and Technology*, 2001.

[7] S. Moncriefi, S. Venkatesh, and C. Dorai, "Horror film genre typing and scene labeling via audio analysis," in *Multimedia and Expo, 2003*, 2003.

[8] D. Matrouf and al., "A straightforward and efcient implementation of the factor analysis model for speaker verication," in *InterSpeech 2007*, 2007.

[9] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing, Special issue on biometric signal processing*, 2004.

[10] D. Istrate, N. Scheffer, C. Fredouille, and J.-F. Bonastre, "Broadcast news speaker tracking for ester 2005 campaign," in *Interspeech 2005*, 2005.

[11] X. Zhu, C. Barras, S. Meignier, and J.-L. Gauvain, "Combining speaker identification and bic for speaker diarization," in *Interspeech 2005*, 2005.

[12] G. Linarès, P. Nocéra, D. Massonie, and D. Matrouf, "The lia speech recognition system: from 10xrt to 1xrt," in *Lecture Notes in Computer Science*, 2007.

[13] J. Ajmera, I. A. McCowan, and H. Bourlard, "Robust hmm-based speech/music segmentation," in *ICASSP 2002*, 2002.