ABSTRACT

Video genre identification methods are frequently based on image or motion analysis, which are relatively time-consuming processes. Since such approaches are tractable by batch processing, as-soon-as-possible identification requires faster methods. In this paper, we investigate the use of audio-only methods for on-the-fly video classification. We propose to use several acoustic feature streams and we evaluate various combination schemes at the frame or at the score level. Results are compared to those obtained by humans, according to the listening duration. Although the system based on model combination slightly outperforms the humans on very soon detection, the latter remain significantly more accurate on long sessions.

Index Terms— Video genre identification, video classification, audio processing, speech processing

1. INTRODUCTION

In the last few years, the amount of video data accessible on the Internet increased considerably, making browsing and searching impossible without indexing. In most of the situations, the metadata attached to the videos are unavailable, and video indexing requires content extraction and analysis. Various descriptors may be used for structuring the video databases. They are usually related to the context of the documents, the content type (music, speech, ...), the topic, etc. One of them is the video genre, which refers to the editorial style of the video. A complete taxonomy of genres can be found in [1], but most of the identification systems try to identify usual genres, such as commercials, movies, cartoons, news, music, etc.

In this paper, we focus on a particular usage scenario where the user needs an on-the-fly tagging of a continuous video stream, for example to remove commercials or to select the kind of video she or he requested for watching on TV. Two main constraints have to be satisfied in the on-the-fly identification scenario. First, global descriptors that are estimated on the whole video cannot be used, due to the fact that identification must be synchronously performed. Second, as-soon-as possible (ASAP) identification requires fast methods since video processing is frequently time-consuming.

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

Many research reports addressed the issue of video classification. Video-based systems obtained a good performance, but usually rely on features that are costly to extract in terms of CPU-load. Text-based approaches apply text classification methods to close captions or Teletext streams, despite the fact that they are frequently unavailable.

Audio-based identification has been generally developed in two different ways. High-level approaches consist in tracking audio events or in analyzing the outputs of a speech recognizer [2]. These methods rely on a priori knowledge of audio contents (for example jingles) and speech recognition is a heavy process that requires large CPU and memory resources. Moreover, in the context of TV or Web data, high word error rates can be expected and any linguistic analysis would be negatively affected by recognition errors.

Low-level approaches use classifiers that operate in cepstral [3, 4, 5] or temporal domains [6]. Various features and classification strategies were investigated in the last few years. The next section presents an overview of these methods, and discusses about their tractability for the on-the-fly classification task.

The remainder of the paper is organized as follows: the next section presents the acoustic features used by video genre identification systems. In Section 3, we detail the genre identification task and the corpus involved in our experiments. Section 4 details our proposal for frame-level and score-level system combination. Finally, we present an experiment aiming to estimate the human performance on the on-the-fly categorization task. The last section concludes the paper and proposes some further developments and experiments.

2. AUDIO FEATURES FOR VIDEO GENRE IDENTIFICATION

Cepstral and time-domain features are the mostly used low-level acoustic descriptors for video classification. Moreover, cepstral parameters have been extensively used, not only in the speech processing field, but also in various audio processing tasks. In [3], the authors evaluate an MFCC/Neural Network system for Video Genre Identification (VGI). This system exhibits a correct classification rate of 51% in a 5-genre task. In [6], the relevance time-domain is studied. These features represent how the document is structured on the temporal axis, typically by using Zero-Crossing Rates (ZCR) and energy variances. In a previous paper [7], we combined these
low-level features to others, related to speaker interactivity and to speech quality. These two last types of features involve speaker diarization systems or ASR systems, and a global analysis of the whole video, avoiding on-the-fly identification.

Therefore, we focus on short-term cepstral analysis of the speech stream. Many speech processing systems rely on multiple acoustic features to improve the accuracy [8, 9]. We investigate the effectiveness of integrating complementary acoustic features for on-the-fly VGI. In section 4, different systems are studied to take full advantage of the complementarity between the three acoustic features PLP (Perceptual Linear Prediction), Rasta-PLP and Mel-Frequency Cepstral Coefficients (MFCC).

3. TASK AND CORPUS

The video genre classification task consists in identifying video genre in 7 categories (news, sport, cartoons, music, documentary, movie and news). The corpus is composed of 1680 videos indexed manually, with durations from 2 to 5 minutes. 1400 of them are used for training the various components of our system, 280 composing the test set. The seven genres are evenly represented in this database (about 200 videos per genre for training, 40 for testing). The speech contents are systematically in the French language, although the music category sometimes contains English songs.

4. PROPOSED SYSTEM

4.1. System Overview

Fig. 1. Architecture of on-the-fly video genre classification system.

The system proposed is a two-level architecture, where the first level extracts cepstral features and the second level combines features and classifies the videos. Acoustic frames are computed every 10 ms in a Hamming window of 20 ms large. PLP and Rasta-PLP vectors are composed of 12 coefficients and first and second order derivatives. MFCC-based vectors are composed of 14 MFCC coefficients, the energy and first and second order derivatives of these 15 features [10].

At the second level, the classifiers are genre-dependent GMM-UBM (Gaussian Mixture Model - Universal Background Model) with Factor Analysis (FA). Lastly, previous researches in the field of speech recognition demonstrated the interest of complementary features combination. We evaluate and compare two combination methods that operate respectively at the frame and the score levels. These points are further developed in the next two sections.

4.2. Factor analysis for Video classification

A popular audio genre classification approach consists in using statistical classifiers, such as GMM on cepstral features [11]. Then, one of the main difficulties of genre categorization stems from the diversity of the videos that are similarly labeled. This problem has been recently tackled by applying FA to intra-class variability reduction, on various tasks such as speaker or language identification. We successfully applied this technique to VGI in a recent paper [12]. Here, FA analysis is applied to genre-dependent GMM-UBM models, independently from the used features.

4.3. Cepstral Feature Combination

The goal of acoustic feature combination is to exploit complementary information provided by different acoustic features. Two approaches are proposed in the literature for combining multiple acoustic features:

- score level: this approach consists in combining the scores of genre-dependent GMMs for all the features. This is achieved by training a neuromimetic classifier on feature vectors that group all GMM-based classifier outputs.

- frame level: in this technique, all the acoustic features are grouped in a super-vector on which the classifiers operate.

We tested these two different approaches and to compare the results on our video test set.

4.3.1. Score level

In this approach, combination is applied at the score level. The basic idea is to estimate the posterior probability of a genre by combining the scores provided by feature-dependent GMMs. This combination is achieved by a Multi-Layer Perceptron (MLP), this choice being motivated by empirical experiments where MLP emerged as better meta-classifiers than vote-based techniques or log-linear combination of posterior probabilities [13]. MLP is trained by the classical gradient retro-propagation algorithm.

Thus, the posterior probabilities are estimated on each feature stream. The three estimations are then grouped in
a score vector, which is used as input for the MLP meta-classifier.

MLP is a 3 layer network with respectively 21, 11 and 7 cells. The 21 inputs are the scores selected from the three acoustic feature classifiers. Each of the 7 network outputs corresponds to a video class.

**Table 1.** Classification rates according to the acoustic features on the 7-genre identification task.

<table>
<thead>
<tr>
<th></th>
<th>Rasta-PLP</th>
<th>PLP</th>
<th>MFCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results</td>
<td>85.81</td>
<td>85.46</td>
<td>85.46</td>
</tr>
</tbody>
</table>

The accuracy for the three acoustic features are very close. We obtain an identification rate of about 85% from all of them. In Table 2, we combine these different acoustic features for estimating the complementarity of the three acoustic features. The results are compared to the Oracle performances, obtained by choosing the best classification hypothesis, knowing the targeted video genre.

**Table 2.** Oracle performance and MLP-based combination of acoustic features. Results demonstrate a relatively high level of complementarity, and the efficiency of MLP-based combination that obtains results very close to the Oracle.

<table>
<thead>
<tr>
<th></th>
<th>MLP-based combination</th>
<th>Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPLP+PLP</td>
<td>0.91</td>
<td>0.92</td>
</tr>
<tr>
<td>PLP+MFCC</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>RPLP+MFCC</td>
<td>0.90</td>
<td>0.91</td>
</tr>
<tr>
<td>RPLP+MFCC+PLP</td>
<td>0.93</td>
<td>0.94</td>
</tr>
</tbody>
</table>

By studying the results obtained by combining two acoustic features for no matter which set of parameters (RASTA-PLP, MFCC, or PLP), we observe a relative error rate reduction of about 33% (from 85% to 90% in absolute value). By combining three acoustic features, we observe another relative reduction of error rates of 30% (from 90% to 93% in absolute value). Although the identification rates for each acoustic feature type are very close, the results tend to confirm that these three acoustic features are complementary.

**4.3.2. Frame level combination**

We propose to combine directly the different acoustic features in a large feature. The resulting feature can be as long as $39 + 39 + 45 = 123$, where 39 is the dimension of the PLP and Rasta-PLP features, and 45 is that of the MFCC feature. However, the resulting feature vector (super-vector) could contain significant redundancies, and hence dimensionality reduction was considered. Including all the dimensions does not always improve video genre classification, but a truncated feature vector often gave better results.

In order to reduce the feature vector dimensionality, we use Heteroscedastic Linear Discriminant Analysis (HLDA). HLDA is a technique that aims to estimate a subspace where classes are easily separable. One of the interests of this method is that the data space dimensionality depends on the number of classes, and may be optimally reduced without strong loss of useful information.

HLDA generalized LDA by removing the restriction of a common within-class covariance matrix. HLDA has been recently used with great success for finding an optimal linear combination of successive vectors of multiple feature streams [14]. The theory of HLDA is fully described in detail in [15].

In Figure 2 we try to assess the impact of dimensionality reduction on the super-vector. These results are compared to the classical classification of the acoustic feature Rasta-PLP by using GMM/FA.

**Fig. 2.** HLDA dimensionality impact on correct classification rates.

Finally, the best system (with a dimension reduction of 60 coefficients) obtains 87% of correct identification and 84% for Rasta-PLP. The absolute gain is about 3%, which is very poor compared to ANN where the absolute gain is about 9%. The results show that HLDA seems to be significantly less accurate than the ANN combination.

**5. COMPARISON WITH HUMAN PERFORMANCE**

In previous experiments, we chose different ways for combining the acoustic features on the full-length video test corpus. Here, we compare the best automatic genre identification systems to human performances. Results are reported only for the two hybrid systems (MLP and HLDA based) and for the best single feature system, based on the RASTA-PLP parametrization.

We tested how a panel of people would classify the videos. The panel consisted of 14 people, from 14 to 53 years old. The test generated 28 trials (randomly chosen) per participant for a total of 392 responses. For each video, they stopped every 5 seconds and annotated them.

**5.1. Results**

As expected, the combination of the three acoustic feature streams outperforms the classification based on single feature streams. In spite of very close individual performances, all
combination schemes benefit from the feature complementarity. Nevertheless, the behaviors differ significantly: although ANN outperforms other approaches on long segments (with durations up to 15s), the HLDA performs better for very soon identification, where only the first few seconds of the document are available. In the very soon identification case, the HLDA-based combination outperforms also the humans as well: on 5-second identification, we observe accuracies of 53%, 46%, 45% and 40% respectively for HLDA, Human, ANN and Rasta-PLP methods. Nevertheless, this ranking rapidly changes: the ANN based combination obtains very good results (79%) on longer durations, the observed performance being very close to the Oracle.

Fig. 3. Classification rates of all the identification methods with respects to the listening duration.

6. CONCLUSION

We have studied the effectiveness of integrating complementary acoustic features for on-the-fly video genre identification. Different combination strategies were proposed, based on direct (HLDA) or indirect (ANN) combination. The results clearly demonstrate the interest in combining features for synchronous classification. The comparison with human performances suggests that people need a few seconds to correctly identify the video genre, probably because they use higher level information, related to linguistic and semantic contents.

Our experiments are conducted on single-story videos, which were extracted from Web-platforms. Nevertheless, many applications operate on more structured videos, where not only classification but also segmentation is required. We now plan to develop such a genre segmentation system and to integrate low level video features in the identification process.

7. REFERENCES