ABSTRACT

In this paper, we present a new method for video genre identification based on the linguistic content analysis. This approach relies on the analysis of the most frequent words in the video transcriptions provided by an automatic speech recognition system. Experiments are conducted on a corpus composed of cartoons, movies, news, commercials, documentary, sport and music. On this 7-genre identification task, the proposed transcription-based method obtains up to 80% of correct identification. Finally, this rate is increased to 95% by combining the proposed linguistic-level features with low-level acoustic features.

Index Terms—video genre classification, audio-based video processing, linguistic feature extraction

1. INTRODUCTION

Video indexing is a challenging task in a context of fast growing of digital TVs and video collections on the Internet. One of the most helpful descriptors for video browsing and retrieval is the video genre, which refers to editorial styles, such as movies, news, commercials, etc. Recent papers address the video genre identification issue. Most of them rely on the extraction of video features, such as color, brightness, or motion-based features [1]. These low-level descriptors are combined by statistical classifiers, typically Support vector machines (SVM) or Gaussian mixture models (GMM). Some authors proposed to use only the audio channel, mainly by classification on cepstral features [1]. In [2], we studied the discriminative capacity of low-level and high-level acoustic features, and we compared various combination schemes.

Although discriminative features of video genres could be extracted from linguistic contents, only a few authors proposed text-only approaches for genre identification. One of the major issues in using text modality for video classification is the lack of textual information associated to the videos. [3, 4] use close captions that are not systematically available. A solution could rely on an automatic speech recognition (ASR) system for linguistic content extraction [5]. Nevertheless, ASR systems frequently fail on Web data, because of adverse acoustic conditions or unexpected linguistic domains.

In this paper, we propose an ASR-based video genre identification approach that is tolerant to weak lexical coverage and high word error rates. This method relies on linguistic-level features that are extracted from the automatic transcriptions.

The paper is organized as follows: the next Section presents the transcription-based features. In Section 3, we describe the experimental framework. Section 4 presents the various acoustic features that are compared to and combined with the transcription-based descriptors.

2. LINGUISTIC-LEVEL GENRE IDENTIFICATION

2.1. Overview

Most of the linguistic-level methods for video genre classification rely on extracting relevant words from the available video meta-data (close captions, tags, etc.), by removing stopwords or by using a term frequency-inverse document frequency (TF-IDF) ranking [6]. Such keyword-based methods aim to capture the topics that occur in the different video genres. This approach is poorly robust for classifying genres like news or documentary, where topics are frequently unexpected. Moreover, video genre refers to the editorial style that is not directly related to the topic or, more generally, to the semantic content. We think that editorial style can be characterized by analysing the stopwords used, unlike in the classical approaches where stopwords are removed from the text features [7].

Transcriptions are provided by an ASR system that processes the audio channel of the videos. The ASR system uses a closed lexicon and a trigram language model (LM) that is estimated on a huge amount of textual data. Learning such a model for each video genre is not tractable, due to the lack of available text materials. We thus use a standard language model having a weak lexical coverage for transcribing all the videos. Therefore, the transcription contains errors, especially on infrequent words, which are also often meaningful. However, the model should be sufficient for correctly transcribing the stopwords, since they are found in all video genres.

This research was supported by the ANR agency (Agence Nationale de la Recherche), RPM2 project (ANR-07-AM-008)
Document modeling relies on the classical bag-of-words model [8]. In this approach, each dimension of the feature space represents a word, documents being represented by word frequency vectors.

We use the classifier architecture proposed in [2], that is composed of two levels of classifiers: low-level genre-dependent classifiers, which pre-process specific groups of features, and a top-level classifier that makes the final decision. This meta-classifier operates on the outputs of the level-1 classifiers.

2.2. The Baseline TF-IDF Features

Our baseline consists in extracting keywords from the videos by using the TF-IDF metric:

\[
\text{TF-IDF}(d, t) = \text{TF}(d, t) \times \text{IDF}(w) \tag{1}
\]

where \(\text{TF}(d, w)\) is the frequency of the word \(w\) in the document \(d\) and \(\text{IDF}(w)\) represents the discriminative strength of the word \(w\) and is defined as:

\[
\text{IDF}(w) = \log \left( \frac{N}{df(w)} \right) \tag{2}
\]

where \(N\) is the total number of documents in the database and \(df(w)\) is the number of documents that contain the word \(w\). The higher the TF-IDF value is, the more the word is representative for the document. The words with a high TF-IDF value are generally meaningful, topic-bearing words.

For each genre, we build a feature vector \(V_g\) with the \(n\) best TF-IDF-ranked words. These vectors are then grouped in a supervector \(\hat{V}\), which is given to the low-level classifier.

2.3. The Stopword Features

Unlike in the classical TF-IDF approach, we propose to use the stopwords as features. These words are characterized by a high frequency and by the fact that they are not meaningful. They generally exhibit a low TF-IDF value, because they are present in all the documents.

We think that the frequency of these stopwords is characteristic of the video genre. Unlike the classical TF-IDF approach, the proposed method is topic-independent, since the stopwords are not topic-related.

Moreover, the automatic transcription is obtained with an LM that is not adapted to the documents; the words that are not in the lexicon of the LM will be erroneous. Thus, the transcription contains errors. We think that the stopwords are more robust than the TF-IDF keywords to the mismatch between the LM and the documents.

Thus, we propose to construct a feature vector \(V_s\) with the \(n\) most frequent words in the transcriptions of the training corpus. The vector \(V_s\) is then given to the low-level classifier.

3. EVALUATION

3.1. The Corpus

We evaluate our proposal on a seven-genre classification task. The genres are music, commercial, cartoon, documentary, news, sport and film.

The corpus is built with 1680 videos (with duration from 2 to 5 minutes), where genre labels have been manually assigned. 1400 of them are used for training various components of our system. 280 compose 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 in the French language. For this corpus, neither reference transcriptions, nor close captions are available.

By using a classical approach based on Mel Frequency Cepstral Coefficients and GMM classifiers, we obtain, on this corpus, a correct classification rate of 52%, corresponding to the results reported in [9] on a similar task.

3.2. The Linguistic Features Evaluation

For transcribing the videos, we used the LIA broadcast news ASR system, SPEERAL [10]. This system is based on an A* decoder using state-dependent hidden Markov models for acoustic modeling. The baseline LM is a 65k word broadcast news 3-gram, estimated on 200M words from the French newspaper “Le Monde” and from the ESTER broadcast news training corpus of about 1M words.

Concerning the TF-IDF feature extraction, we tried almost all the values for the number of the best TF-IDF-ranked words used as input features for the low-level classifier, from 1 to the total number of different words in the whole corpus, and we found that the optimal value in our case is about 6000. We tested three types of low-level classifiers: SVM, Boosting and artificial neural networks (ANN). The word frequencies in the feature vectors of each document are normalized with respect to the size of the document. The total number of words in the document is added as a feature. The Boosting classifier obtains the best results: it achieves a 72.1\% correct classification rate (CCR). This result constitutes our baseline and is reported in Figure 1.

For the stopword features, the CCR of the two best classifier types are presented in Figure 1, depending on the number of the most frequent words used as input features, from 1 to 1000. Beyond 1000 words, the performances of the classifiers slightly decrease, because of the noise. Note that the ANN is a multi-layer perceptron with one hidden layer; the size of the hidden layer is optimized on the training data. The presented performances of the ANN stopword classifier are obtained by using raw frequencies in the feature vectors, which proved to work better than using normalized frequencies.

It is worth noting that the classifier performances grow quickly with the number of stopword features. Performances comparable to the baseline of 6000 feature words are obtained
with only 23 feature words. The best CCR is 80.4%, obtained with the ANN classifier using the 100 most frequent words. This last result represents an absolute improvement of 8.3%, compared to the baseline TF-IDF, and a reduction of the feature space dimensionality of about 98%.

With only the most frequent word, $\text{<sil>}$, which represents a silence, as input, the best classifier achieves a 48.6% correct classification rate. Adding the second word gives a 63.9% correct classification rate. The performance gain of adding words follows an inverse logarithmic law. The less frequent the added words are, the lower the gain obtained is. We can conclude that the more a word is frequent, the more important the word is for classifying genres. Table 1 contains the nine most frequent words in the training corpus, associated with their frequencies.

These performances validate our initial hypothesis: the stopword frequencies contain information that is characteristic to the video genre. Moreover, the proposed approach yielded an absolute CCR gain of about 8% with respect to the baseline TF-IDF, while the feature space is reduced by 98%. In the next section we present our experiments on combining these features with several other audio features.

### Table 1. Frequency of the nine most frequent words found in the automatic transcriptions of the training corpus.

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
<th>Word</th>
<th>Frequency</th>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{&lt;sil&gt;}$</td>
<td>146100</td>
<td>et</td>
<td>12236</td>
<td>est</td>
<td>9385</td>
</tr>
<tr>
<td>de</td>
<td>20093</td>
<td>le</td>
<td>10961</td>
<td>des</td>
<td>8682</td>
</tr>
<tr>
<td>les</td>
<td>12526</td>
<td>la</td>
<td>10819</td>
<td>il</td>
<td>7628</td>
</tr>
</tbody>
</table>

### 4. ACOUSTIC AND LINGUISTIC FEATURE COMBINATION

The linguistic descriptor proposed in this paper is based on spoken contents. In a previous study, we evaluated acoustic features that are not directly dependent on the linguistic contents [2]. In this section, we evaluate how much these fundamentally different features bring complementary information.

#### 4.1. Short-term Cepstral features

In [2], we demonstrate the efficiency, for genre modeling, of 12 perceptual linear prediction (PLP) coefficients with their first and second order derivatives. Then, factor analysis (FA) is used to cope with the intra-genre variability. The FA models estimate super-vectors that are used as input to an SVM low-level classifier. The approach is fully described in [11].

The results reported in Table 2 show that the acoustic space descriptor alone (the AS row) performs better than the linguistic descriptor; nevertheless, combining them (the AS+L row) allows one to improve significantly the accuracy on commercials and cartoons and the global accuracy increases by 2% in absolute value.

#### 4.2. Speaker Interactivity

The number of speakers and the way they communicate is strongly different according to the genre of the video. For example, in news there is usually one main speaker (anchor), unlike in cartoons or movies. The interactivity feature aims to represent these speaker-related profiles. As described in [2], a speaker diarization system is used for obtaining an estimation of the number of speakers and speaker turns for each document. The feature vector is built with the following three parameters: the density of the speaker turns, the number of speakers, and the speaking time of the main speaker.

The results yielded by the speaker interactivity descriptor alone (the I row) and in combination with the linguistic descriptor (the I+L row) are reported in Table 2. By adding the linguistic descriptor to the speaker interactivity features, we achieve an absolute gain in accuracy of about 25%. The linguistic descriptor improves accuracies on documentary, news and cartoon. All the genres except music are recognized in more than 73% of the cases.

#### 4.3. Speech Quality

The basic idea is that the speech quality could provide some relevant information about the genres. For example, speech is usually clean in news, where the linguistic domain is well covered by ASR systems.

We use three features in this group, all based on the SPEERAL system, described in Subsection 3.2.

The first feature is the posterior probabilities of the one-best hypothesis words. We use posteriors as a confidence
measure that contains not only the acoustic and linguistic scores, but also some information pertaining to the decoding graph that was explored. The second feature is the linguistic probability of the whole one-best hypothesis. The last feature is the phonetic entropy, that was introduced by [12] for speech/music separation. It consists of the entropy of the acoustic probabilities, which is high on low-quality speech, and decreases on clean speech.

The results are shown for the descriptor alone (the Q row) and in combination with the linguistic descriptor (the Q+L row) in Table 2. We observe an absolute gain of about 28% by adding the linguistic descriptor. Although the two descriptors considered pertain to linguistic elements, combining them improves the overall accuracy from about 80% for the linguistic descriptor alone, to 81% with the speech quality. This last result shows that the two descriptors contain complementary information.

4.4. Combination

In the previous experiments, we saw that the linguistic descriptor improved the accuracy for each of the other groups of features, and particularly on the document, news, commercial, and cartoon genres. We now combine all the groups by creating a super-vector that contains all the proposed features. The result of this global combination is shown on the last row of Table 2. We observe an absolute accuracy gain of 4%, compared to the best descriptor alone, and a gain of 2% compared to the best previous combination. These results show that all the proposed features are complementary and relevant for genre classification.

Table 2. The correct classification rate [%] of the proposed descriptors.

<table>
<thead>
<tr>
<th></th>
<th>Doc</th>
<th>Mus</th>
<th>New</th>
<th>Com</th>
<th>Car</th>
<th>Mov</th>
<th>Spo</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS</td>
<td>97</td>
<td>97</td>
<td>78</td>
<td>82</td>
<td>92</td>
<td>92</td>
<td>100</td>
<td>91</td>
</tr>
<tr>
<td>AS+L</td>
<td>97</td>
<td>97</td>
<td>83</td>
<td>90</td>
<td>97</td>
<td>87</td>
<td>100</td>
<td>93</td>
</tr>
<tr>
<td>I</td>
<td>62</td>
<td>41</td>
<td>21</td>
<td>85</td>
<td>48</td>
<td>53</td>
<td>77</td>
<td>56</td>
</tr>
<tr>
<td>I+L</td>
<td>97</td>
<td>56</td>
<td>84</td>
<td>87</td>
<td>92</td>
<td>73</td>
<td>75</td>
<td>81</td>
</tr>
<tr>
<td>Q</td>
<td>60</td>
<td>48</td>
<td>52</td>
<td>48</td>
<td>24</td>
<td>39</td>
<td>92</td>
<td>53</td>
</tr>
<tr>
<td>Q+L</td>
<td>97</td>
<td>68</td>
<td>81</td>
<td>79</td>
<td>92</td>
<td>68</td>
<td>85</td>
<td>81</td>
</tr>
<tr>
<td>AS+L+I+Q</td>
<td>97</td>
<td>97</td>
<td>84</td>
<td>97</td>
<td>100</td>
<td>90</td>
<td>100</td>
<td>95</td>
</tr>
</tbody>
</table>

5. CONCLUSION

In this paper we proposed a new way of using the linguistic content in video genre classification when no video meta-data are available. The linguistic content of the videos is obtained by using an ASR system. The classification is based on the analysis of the stopword frequencies and is thus tolerant to weak lexical coverage. Results show that this approach outperforms significantly the more classical TF-IDF based methods. Moreover, the proposed linguistic features offer a high level of complementarity to acoustic-level features: by combining the two descriptor types, we observe a relative improvement of the classification error rate of about 25%, reaching 95% of correct identification.

We now plan to extract more information from the ASR system used for transcribing the document, especially descriptors of the word lattice topology.

6. REFERENCES