LIA @ MediaEval 2011 : Compact Representation of Heterogeneous Descriptors for Video Genre Classification

Mickael Rouvier
LIA - University of Avignon
Avignon, France
mickael.rouvier@univ-avignon.fr

Georges Linarès
LIA - University of Avignon
Avignon, France
georges.linares@univ-avignon.fr

ABSTRACT

This paper describes our participation in Genre Tagging Task @ MediaEval 2011, which aims at predicting the Genre of Internet videos. We propose a method that extracts low dimensional feature space based on text, audio and video information. In the best configuration, our system yields a 0.56 MAP (Mean Average Precision) on the test corpus.

Categories and Subject Descriptors

H.3.1 [**Information Search and Retrieval**]: Content Analysis and Indexing—*Indexing method*

General Terms

Algorithms, Measurement, Performance, Experimentation

1. INTRODUCTION

The Genre Tagging Task held as part of MediaEval 2011 required participants to automatically predict tags of videos from Internet. The task consists in associating each video to one and only one of the 26 provided genres [3]. The videos are provided with some additional information like metadata (title and description), speech recognition transcripts [2] and social network information (gathered from Twitter).

One of the main difficulty of video genre categorization is due to the diversity of the information sources that are genre-dependent (spoken contents, video structure, audio and video patterns, etc.) and to the variability of video classes. In order to deal with this problem, we propose to combine various features from audio and video channels and to reduce the resulting large-dimensional input data to a low dimensional feature vector while retaining most of the relevant information (reducing redundancies and minimising the useless information).

The paper is organized as follows. Section 2 explains how we collect our training data. Section 3 describes our system. Section 4 present the features used by our system and Section 5 summarizes the results.

2. CORPUS

This task is seen as a classification task/problem. We choose to follow a slightly supervised approach. The training dataset is collected from the Web. A simple and effective way to obtain a corpus is to download documents given by a web search engine using useful queries. For example, for the *Health* genre we download all the videos on Youtube using the query : health. But using only the genre like query does not allow to download a training set that represents the class variability. We propose to expand our query to other terms revolving around the genre. For example, for the *Religion* genre, we need to expand our query to related terms like : ethnic, belief, freedom, practice, etc. In order to find terms closely related to the genre, we propose to use Latent Dirichlet Allocation (LDA). LDA is a unsupervised word clustering method that relies on word co-occurrence analysis. The 5000-classes LDA model was estimated on the Gigaword corpus. Each cluster is composed of the 10 best words. Queries are expanded by adding all words from the best cluster containing the genre tag.

We propose the use of different information sources extracted from audio, speech and video channels. Consequently, our training set should include all these sources, especially transcription of spoken contents. ASR performance on Web data are usually high, however the ASR system used on the test set is not freely distributed. To overcome this problem, we propose to collect text materials by downloading web pages from the Internet. Our training corpus consists of web pages collected from Google (60 documents per class, 1560 documents) and videos collected from Youtube and Dailymotion (120 documents per class, 3120 documents). There are more videos than web pages because of technical restrictions imposed by Google. The collected documents (web pages and videos) are only in English.

3. SYSTEM DESCRIPTION

The proposed system has a 2-level architecture. The first level consists of extracting low dimensional features from speech, audio and video. Each feature is then given to a SVM (Support Vector Machine) classifier. The second level combines the scores of three SVM models. This combination is achieved by linear interpolation whose coefficients are determined on the development corpus.

4. FEATURES

4.1 Text

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. Our system consists of extracting relevant keywords from the documents by using the TF-IDF metric.

MediaEval 2011 Workshop, September 1-2, 2011, Pisa, Italy

The words with a high TF-IDF value are generally meaningful, topic-bearing words. Thus, we propose to construct a feature vector with the n ($n = 600$ in our experiments) most frequent words in the documentary of the training corpus.

4.2 Audio

One of the most popular approach for genre identification by acoustic space characterization relies on MFCC (Mel Frequency Cepstral Coefficients) analysis and GMM (Gaussian Mixture Model) or SVM classifiers.

However, audio features include both useful and useless information.

Unfortunately, separating useful and useless information is a heavy process, and the useless space contains some information that can be used to distinguish between genres. For this reason, [1] proposed a single space (the *total variability space*) that models the two variabilities. In this new space, a given audio utterance is represented by a new vector named total factors (we also refer to this vector as *i-vector*), that allows to reduce redundancies and to enhancing useful information.

Acoustic frames of MFCC are computed every 10 ms in a Hamming window of 20 ms large. MFCC vectors are composed of 12 coefficients, energy and first and second order derivatives of these 13 features. For these experiments the UBM is composed of 512 gaussians and the i-vector is a 400 dimension vector.

4.3 Video

We used features based on the color like : *Color Structure Descriptor* or *Dominant Color Structure* or features based on the texture like : *Homogeneous Texture Descriptor (HTD)* or *Edge Histogram Descriptor*. On this task, it seems that texture was the best feature and specially the HTD.

HTD is an efficient feature not only for computing texture features but also for representing texture information. HTD provides a quantitative characterization of homogeneous texture regions for similarity retrieval. HTD consists in the mean, the standard deviation value of an image, energy, and energy deviation values of Fourier transform of the image.

Similarly to the audio feature processing, we extract, for each video feature, an i-vector. For these experiments the UBM is composed of 128 gaussians and the i-vector is a 50 dimension vector.

5. RESULTS

We submitted five runs for the Genre Tagging Task, combining the results, presented in the section above. In detail, the configuration for each run was as follows :

Run 1: We use only text feature. The text feature is built on the speech transcript.

Run 2: We use text, audio and video features. The text feature is built on speech transcript and description of the video given in the metadata.

Run 3: We use text, audio and video features. The text feature is built on speech transcript, description of the video given in the metadata and tags.

Run 4: In the previous run, the SVM classifier has been learned on the features given by the training corpus. We notice that the training corpus has been downloaded by using a slightly supervised method, and some of the video may be uncorrectly tagged. To improve the performance, we propose to integrate the development corpus in the training corpus. Development corpus was provided by Mediaeval in which the genres were manually checked. This run uses exactly the same features as run 3, but the SVM classifier has been learned on the features given by the training and development corpus.

Run 5: In the development corpus, we observed that the username of the video (present in the metadata) can give some interesting information to predict the video genre. Indeed a user often uploads multiple videos of the same genre. For example, the users *Anglicantv* or *Aabbey1* often upload videos of the genre *Religion*. Here, we use the dev set as a knowledge base, where the favorite genre of people is known. For each video, we search if the username is present in the dev corpus and increase the score of the genre in which the user uploaded the videos. Here, we boost scores from the run 4 according to this new information. We conducted a post-campaign experiment that show that, by using only this information, the system performs 51% MAP.

Table 1: Results of the submitted runs

Team	Run	Id	MAP
LIA	1	run1	0.1179
LIA	2	run2	0.17
LIA	3	run3	0.1828
LIA	4	run4	0.1964
LIA	5	run5	0.5626

In run 2, we observe that the audio and video features provide interesting information to predict the video genre. The runs 2, 3 and 4 achieved a similar performance which means that the different configurations did not strongly contribute to the global results. The use of the owner id strongly improves the results.

6. CONCLUSION

We have described in this paper an approach based on the use of audio, video, text (transcription and metadata) features for Video Genre Classification. According to the results, username seems to be simple and strongly efficient information to predict the video genre.

7. REFERENCES

[1] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In *INTERSPEECH*, pages 1559–1562, 2009.

[2] J.-L. Gauvain, L. Lamel, and G. Adda. The limsi broadcast news transcription system. *Speech Communication*, 37(1-2):89 – 108, 2002.

[3] M. Larson, M. Eskevich, R. Ordelman, C. Kofler, S. Schmiedeke, and G. Jones. Overview of MediaEval 2011 Rich Speech Retrieval Task and Genre Tagging Task. In *MediaEval 2011 Workshop*, Pisa, Italy, September 1-2 2011.